Mean Estimation from Adaptive One-bit Measurements

Alon Kipnis
Department of Electrical Engineering
Stanford University
Stanford, CA

John C. Duchi
Department of Electrical Engineering
and Department of Statistics
Stanford University
Stanford, CA

***Abstract*—We consider the problem of estimating the mean of a normal distribution under the following constraint: the estimator can access only a single bit of each sample from this distribution. We study the squared error risk in this estimation as a function of the number of samples and one-bit measurements n. We show that, if at step n the single bit is a function of both the new sample and the previous $n-1$ acquired bits, no estimator can attain asymptotic mean squared error smaller than $\pi/(2n)$ times the variance. In other words, one-bit restriction increases the number of samples required for a prescribed accuracy of estimation by a factor of at least $\pi/2$ compared to the unrestricted case. In addition, we provide an explicit estimator that attains this asymptotic error, showing that, rather surprisingly, only $\pi/2 < 1.571$ times more samples are required in order to attain estimation performance equivalent to the unrestricted case.**

I. Introduction

Processing and estimating information from data collected at multiple sensors and physical locations is subject to communication constraints. For example, consider large-scale sensor arrays where information is collected at multiple physical locations and transmitted to a central estimation unit. In this scenario, the ability to estimate a particular parameter from data collected by the sensors is dictated not only by quality of observations and their numbers, but also by limited communication rates between the sensors and the central estimator. The question that we ask is to what extend a parametric estimation task is affected by communication constraints, and what are the fundamental performance limits in estimating a parameter subject to these restrictions. In this paper we answer this question in a particular simple 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 on each sample X_n from this distribution. As it turns out, the ability to share information among different samples before sending each single-bit message dramatically affects the performance in estimating θ. We distinguish in particular among three different settings:

(i) *Centralized* encoding: all n encoders confer and produce a single n bit message which is a function of $X_1, \ldots, X_n$.
(ii) *Adaptive* or *sequential* encoding: the nth encoder observes X_n and the $n-1$ previous single bit messages.
(iii) *Distributed* encoding: the output of the nth encoder is a single bit that is only a function of X_n.

Clearly, as far as information sharing is concerned, settings (iii) is a more restrictive version of (ii) which is more restrictive than (i). We measure the estimation performance by the mean squared error (MSE) risk. Since without bit constraint this risk is equals σ^2/n, we are interested in particular in the *asymptotic relative efficiency* (ARE) of estimators in the constrained setting, defined by the ratio between their MSE to σ^2/n as n goes to infinity.

In setting (i) the estimator can evaluate the sample mean and then communicate it using n bits. This strategy leads to MSE behavior of $\sigma^2/n + o(2^{-n})$. Therefore, the ARE in this setting is 1. Namely, asymptotically, there is no loss in performance due to the communication constraint under centralized encoding. In this work we show that a similar result does not hold even in setting (ii): the ARE of any adaptive estimation scheme is at least $\pi/2$. Namely, the single-bit per sample constraint incurs a minimal penalty of at least 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 that almost surely attains this ARE. In other words, we show that the lower bound of $\pi/2$ on the asymptotic relative efficiency is tight, and that it is attained regardless of the particular realization of θ or the size of the parameter space from which it is taken. Clearly, the minimal penalty on the efficiency of $\pi/2$ also holds under setting (iii), although the question whether such efficiency is achievable (or otherwise, what is the minimal ARE) remains open.

As the variance σ^2 goes to zero, the task of finding θ using one-bit queries in the adaptive setting (ii) is easily solved by a bisection style method over the parameter space. Therefore, the general case of non-zero variance is a reminiscent of the noisy binary search problem with possibly infinite number of unreliable tests [1], [2]. However, since we assume a continuous parameter space, a more closely related problem is that of one-bit analog-to-digital conversion of a noisy signal. For example, sigma-delta modulator (SDM) analog-to-digital conversion [3] uses one-bit threshold detector combined with a feedback loop to update an accumulated error state, and therefore falls under setting (ii). A SDM with a constant input θ corrupted by a Gaussian noise was studied in [4], where it was shown that the output of the modulator converges to the true constant input almost surely. In other words, the SDM provides a consistent estimator for setting (ii). The rate of this convergence, however, was not analyzed and cannot be derived from the results of [4]. Our results imply that the rate of convergence of a SDM to a constant input is at most $\sigma^2\pi/2$ over the number of feedback iterations.

We note that even the centralized encoding of setting (i) poses a non-trivial challenge for the design and analysis of optimal encoding and estimation scheme. Indeed, the standard technique to encode an unknown random quantity using n bits is attained by a scalar quantizer. However, the optimal design of this quantizer depends on the distribution of its input [5], which is the goal of our estimation task. Therefore a non-trivial exploration exploitation tradeoff arises. In our setting, the only missing parameter is the mean, which, under setting (i), is known to the encode with uncertainty interval proportional to $\sigma/\sqrt{n}$. Therefore, while it is clear that uncertainty due to quantization decreases exponentially in the number of bits, an exact expression for the MSE in this setting is not easy to derive. The situation is even more involved in the adaptive encoding of setting (ii): an encoding and estimation strategy that is globally optimal for $n-1$ adaptive one-bit messages of a sample of size $n-1$, may not lead to a globally optimal strategy upon the recipient of the nth sample. Conversely, any one-step optimal strategy, in the sense that it finds the best one-bit message as a function of the current sample and the previous 2^{n-1} messages, is not guaranteed to be globally optimal. Therefore, the main contribution of this paper is a lower bound on the MSE risk of any globally optimal strategy. This lower bound is obtained by showing that the Fisher information of any n steps strategy is not larger than $2n/(\pi\sigma^2)$. The desired bound of $\pi\sigma^2/(2n)$ on the MSE follows from the van Trees version of the information inequality [6]. Finally, we show that an estimator that attains asymptotic MSE of $\sigma^2\pi/(2n)$ is obtained as a special case of [7, Thm. 4]. In addition to these two results, we also derive the one-step optimal strategy and demonstrates numerically that the MSE under this strategy converge to $\pi\sigma^2/(2n)$. A proof for this converges remains elusive.

Our results imply that even under coarse quantization constraints, it is possible to achieve MSE in parametric estimation within only a relatively small penalty compared to the unconstrained estimator. A possible clue for this non-intuitive result is obtained from drawing the connection between our setting and the remote multiterminal source coding problem, also known as the CEO problem [8], [9], [10], [11]. This connection, explained in details in Section III, immediately leads to a lower bound of $4/3$ on the ARE in the distributed encoding of setting (iii). While this lower bound provides no new information compared to the lower bound of $\pi/2$ we derive here for setting (ii), it shows that the distributed nature of the problem is not a limiting factor in achieving MSE close to optimal even under one-bit quantization of each sample.

Setting (ii) and (iii) were considered in [12] under the more general assumption of m machines each has access to n/m independent samples. The main result of [12] are bounds on the estimation error as a function of the number of bits

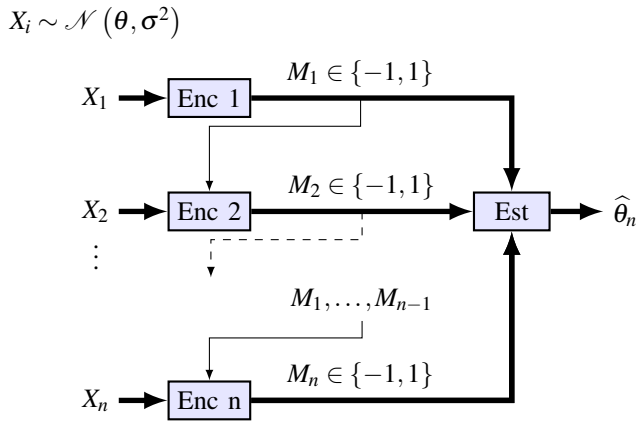

Fig. 1: Adaptive one-bit encoding: the ith encoder delivers a single bit message that is a function of its private sample X_i and the previous $i-1$ messages $M_1, \ldots, M_{i-1}$.

R with which each machine can communicate. Specializing their result to our setting, by taking $m = n$ and $R = 1$, leads to looser bounds then $\sigma^2\pi/(2n)$ for case (ii) and (iii). Other related works include statistical inference under multiterminal data compression [13], [14], although their results apply only to setting (i). The counterpart of setting (iii) in the case of hypothesis testing was considered in [15]. Estimation problem subject to the one-bit quantization constraint were also considered recently in compressed sensing [16], [17] and in MIMO detection in wireless communication [18].

The rest of this paper is organized as follows: the main problem and notation are defined in Section II. In Section III we illustrate a connection between our parametric estimation problem and the remote multiterminal lossy compression problem. Our main results are given in Section IV, where we present a lower bound on estimation from one-bit samples in the adaptive scheme, as well as an achievable scheme that attains this bound. Concluding remarks are given in Section V. For sake of fluency, long proofs are deferred to the Appendix.

II. Problem Formulation

Let X_i, $i = 1, \ldots, n$, be n independent samples from the normal distribution with mean θ and variance σ^2. We assume that the mean θ is drawn once from a prior distribution $\pi(\theta)$ on 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 density $\pi(d\theta)$. 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 $M^n = (M_1, \ldots, M_n)$.

(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}.

(iii) The ith message M_i takes only two possible values, say 1 and -1.

In other words, the ith message is defined by a function from the real line to $\{-1,1\}$ that is measurable with respect to the sigma algebra generated by M^{i-1} and X_i, and the n messages M^n are the only available to the estimator. Upon observing M^n, the estimator produces an estimate $\widehat{\theta}_n(M^n)$ of θ. A system describing the above scheme is illustrated in Fig. 1.

In this work we are concerned with the MSE risk defined as

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

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 minimization of (1) over all encoding and estimation strategies, and the characterization of its minimal value as a function of n. This minimization is the combination of the following two procedures: (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 ARE of estimators of 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 ARE is defined as

$$\lim_{n\to\infty}\frac{n}{\sigma^2}\mathbb{E}\left(\widehat{\theta}_n-\theta\right)^2. \tag{2}$$

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.

Before deriving our main results, we comment on the relation between our setting and the remote multiterminal source coding problem.

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 [8], 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 [11]. 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 [19, 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 [20, 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. Results

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. 1. 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 [21, Thm. 2.13],[6] 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 [7, 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

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

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

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

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

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

is given by

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

where τ is the unique solution to

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

□

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

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

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

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. 2. Also shown in Fig. 2 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.

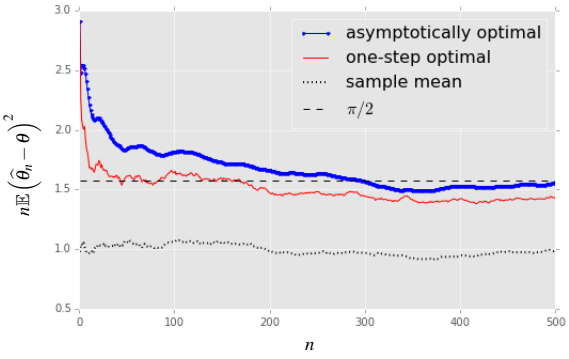

Fig. 2: Normalized empirical risk $n\left(\widehat{\theta}_n - \theta\right)^2$ versus number of samples n for 500 Monte Carlo trials. In each trial, θ is chosen uniformly in the interval $(-3,3)$.

V. Conclusions

We considered the MSE risk and asymptotic relative efficiency in estimating the mean of a normal distribution from a single-bit encoding of each sample from this distribution. In the adaptive scenario where each one-bit message is a function of the previously seen messages and current sample, we showed that the minimal relative efficiency is $\pi/2$. Namely, there is a penalty factor of at least $\pi/2$ on the asymptotic MSE 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. Another open question that follows from our setting is the asymptotic MSE and relative efficiency of a fully distributed one-bit scheme.

Appendix

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

Proof of Thm. 2

We first prove the following two lemmas:

Lemma 7: For any $x_1 \geq \ldots \geq x_n \in \mathbb{R}$, we have

$$\frac{\left(\sum_{k=1}^n (-1)^{k+1}\phi(x_k)\right)^2}{\left(\sum_{k=1}^n (-1)^{k+1}\Phi(x_k)\right)\left(1-\sum_{k=1}^n (-1)^{k+1}\Phi(x_k)\right)} \leq \frac{2}{\pi}. \tag{17}$$

Lemma 8: 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 Lem. 7: We use induction on $n \in \mathbb{N}$. For the base case $n = 1$ we have

$$\frac{\phi^2(x)}{\Phi(x)\left(1-\Phi(x)\right)}. \tag{18}$$

Taking the logarithm of (18) and differentiating, we conclude that any point x that maximizes (18) also 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 $x = 0$, in which (18) equals $2/\pi$.

Assume now that (17) holds for all integers up to some $n = N-1$ and consider the case $n = N$. The maximal value of (17) 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 the LHS of (17) equals

$$\frac{\phi(x_1)^2}{\Phi(x_1)(1-\Phi(x_1))},$$

which was shown to be smaller than $\pi/2$. If N is even, then for any constant c the limit of the LHS of (17) exits as $(x_1, \ldots, x_N) \to (c, \ldots, c)$ and equals zero. Therefore, the maximum of the LHS of (17) is not attained at this line. We now consider the possibility that the LHS of (17) is maximized at the borders, as one or more of the coordinates of $(x_1, \ldots, x_N)$ approaches plus or minus infinity. For simplicity we only consider the cases where x_N goes to minus infinity or x_1 goes to plus infinity (the general case where the first m coordinates goes to infinity or the last m to minus infinity is obtained using similar arguments). Assume first $x_N \to -\infty$. Then the LHS of (17) 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 the LHS of (17) 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 Lem. 7.

Proof of Lem. 8: The Fisher information of M with respect to θ is given by

$$\begin{aligned} I_\theta &= \mathbb{E}\left[\left(\frac{d}{d\theta}\log P(M|\theta)\right)^2 |\theta\right] \\ &= \frac{\left(\frac{d}{d\theta}P(M=1|\theta)\right)^2}{P(M=1|\theta)} + \frac{\left(\frac{d}{d\theta}P(M=-1|\theta)\right)^2}{P(M=-1|\theta)} \\ &\overset{(a)}{=} \frac{\left(-\int_A \phi'\left(\frac{x-\theta}{\sigma}\right)dx\right)^2}{\sigma^2 P(M=1|\theta)} + \frac{\left(\int_A \phi'\left(\frac{x-\theta}{\sigma}\right)dx\right)^2}{\sigma^2 P(M=-1|\theta)} \\ &= \frac{\left(\int_A \phi'\left(\frac{x-\theta}{\sigma}\right)dx\right)^2}{\sigma^2 P(M=1|\theta)\left(1-P(M=1|\theta)\right)}, \\ &= \frac{\left(\int_A \phi'\left(\frac{x-\theta}{\sigma}\right)dx\right)\left(\int_A \phi'\left(\frac{x-\theta}{\sigma}\right)dx\right)}{\sigma^2\left(\int_A \phi\left(\frac{x-\theta}{\sigma}\right)dx\right)\left(1-\int_A \phi\left(\frac{x-\theta}{\sigma}\right)dx\right)}, \end{aligned} \tag{19}$$

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 (19) 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 (19) can be rewritten as

$$= \frac{\left(\sum_{j=1}^{k} \phi\left(\frac{a_j-\theta}{\sigma}\right) - \phi\left(\frac{b_j-\theta}{\sigma}\right)\right)^2}{\sigma^2\left(\sum_{j=1}^{k} \Phi\left(\frac{b_j-\theta}{\sigma}\right) - \Phi\left(\frac{a_j-\theta}{\sigma}\right)\right)} \times \frac{1}{1-\left(\sum_{j=1}^{k} \Phi\left(\frac{b_j-\theta}{\sigma}\right) - \Phi\left(\frac{a_j-\theta}{\sigma}\right)\right)} \quad (20)$$

The proof of Lem. 8 is completed since it follows from 7 that for any $\theta \in \mathbb{R}$ and any choice of the intervals endpoints, (20) 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

$$M_i = \begin{cases} X_i \in A_i & 1, \\ X_i \notin A_i & -1, \end{cases} \quad (21)$$

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 (21) with $A_i = M^{-1}(1)$. Consider the conditional distribution $P(M^n|\theta)$ of M^n given θ. We have

$$P(M^n|\theta) = \prod_{i=1}^{n} P\left(M_i|\theta, M^{i-1}\right), \quad (22)$$

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

$$I_\theta(M^n) = \sum_{i=1}^{n} I_\theta(M_i|M^{i-1}), \quad (23)$$

where $I_\theta(M_i|M^{i-1})$ is the Fisher information of the distribution of M_i given M^{i-1}, where it follows from Lem. 8 that $I_\theta(M_i|M^{i-1}) \leq 2/(\pi\sigma^2)$. We now use the following theorem from [21, Thm. 2.13] (see also [22], [6]):

Theorem 9 (The van Trees inequality [21]): Denote by $p(\cdot,\theta)$ the density of P_θ with respect to the Lenesgue measure. Assume that: (i) the density $p(x,\theta)$ is measureable in (x,θ) and absolutely continuous in t for almost all x with respect to the Lebesgue measure. (ii) The Fisher information

$$I(\theta) = \int \left(\frac{p'(x,\theta)}{p(x,\theta)}\right)^2 p(x,\theta)dx,$$

where $p'(x,t)$ denotes the derivative of $p(x,\theta)$ in t, is finite and integrable on Θ. (iii) The prior density $\pi(\theta)$ is absolutely continuous on its support Θ with zero mass at the boundries of Θ, and has a finite Fisher information

$$I_0 = \int_\Theta \frac{(\pi'(\theta))^2}{\pi(\theta)} d\theta.$$

Then, for any estimator $\widehat{t}(\mathbf{X})$, the Bayes risk is bounded as follows:

$$\int_\theta \mathbb{E}\left[\left(\widehat{t}(\mathbf{X}) - \theta\right)^2\right] \pi(d\theta) \geq \frac{1}{\int I(\theta)\pi(d\theta) + I_0}.$$

Theorem 9 applied to our problem with $p(x,\theta) = P(M^n|\theta)$ implies

$$\mathbb{E}\left(\widehat{\theta}_n - \theta\right)^2 \geq \frac{1}{\mathbb{E}I_\theta(M^n) + I_0} = \frac{1}{\sum_{i=1}^{n} I_\theta(M_i|M^{i-1}) + I_0} \geq \frac{1}{2n/(\pi\sigma^2) + I_0}.$$

□

Proof of Thm. 3

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

Theorem 10: [7, Thm. 4] Let

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

where the Z_is are i.i.d. from $\mathcal{N}(0,\sigma^2)$ and independent of each other. Define

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

where the sequence $\{\gamma_i\}_{i=1}^{\infty}$ satisfies conditions (i) and (ii) in Thm. 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 Thm. 10, we set $\varphi(x) = \mathrm{sgn}(x)$ and $Z_i = X_i - \theta$. We have:

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

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

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

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 Thm. 10 are fulfilled for any $\lambda > 0$. Since

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

Thm. 3 follows from Thm. 10. □

Proof of Thm. 4

In this subsection we prove Lemmas 6 and 5 that lead to Thm. 4.

Proof of Lem. 6: Since any single-bit message $M(u) \in \{-1,1\}$ 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:

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

so clearly, the partition $A_1' = A_1 \setminus B$, $A_{-1}' = A_{-1} \cup B$ attains lower error variance, what 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 [23] for two quantization regions) we conclude that τ must satisfy the equation

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

□

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

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

and

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

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

$$1 - \frac{f(\tau)\int_{-\infty}^\tau f(x)(\tau - x)dx}{\left(\int_{-\infty}^\tau f(x)dx\right)^2}. \tag{24}$$

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 (24) is non-negative and thus $h^-(x)$ is monotonically increasing. Since

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

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

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

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

The work of A. Kipnis is partially supported by NSF Center for Science of Information (CSoI) under grant CCF-0939370 and NSF-BSF under grant 1609695.

REFERENCES

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

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

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

[4] 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.

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

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

[7] 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.

[8] 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.

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

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

[11] 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.

[12] 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.

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

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

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

[16] P. T. Boufounos and R. G. Baraniuk, "1-bit compressive sensing," in *Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on*. IEEE, 2008, pp. 16–21.

[17] 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.

[18] J. Singh, O. Dabeer, and U. Madhow, "On the limits of communication with low-precision analog-to-digital conversion at the receiver," *IEEE Transactions on Communications*, vol. 57, no. 12, 2009.

[19] 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.

[20] 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

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

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

[23] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar 1982.