Binary and Multi-Bit Coding for Stable Random Projections

Ping Li
Department of Statistics and Biostatistics
Department of Computer Science
Rutgers University
Piscataway, NJ 08854, USA
pingli@stat.rutgers.edu

Abstract

The recent work [17] developed a 1-bit compressed sensing (CS) algorithm based on α-stable random projections. Although it was shown in [17] that the method is a strong competitor to other existing 1-bit CS algorithms, the procedure requires knowing K, the sparsity, which is the l_0 norm of the signal. Other existing 1-bit CS algorithms require the l_2 norm of the signal.

In this paper, we develop an estimation procedure for the l_α norm of the signal, where $0 < \alpha \leq 2$ from binary or multi-bit measurements. We demonstrate that using a simple closed-form estimator with merely 1-bit information does not result in a significant loss of accuracy if the parameter is chosen appropriately. Theoretical tail bounds are also provided. Using 2 or more bits per measurement reduces the variance and importantly, stabilizes the estimate so that the variance is not too sensitive to chosen parameters.

1 Introduction

Compressed sensing (CS) [10, 5] aims at recovering sparse signals from linear measurements. Consider a K-sparse signal of length N, denoted by x_i, $i = 1$ to N, with $\sum_{i=1}^N 1\{x_i \neq 0\} = K$. In the CS framework, we first collect measures $y_j = \sum_{i=1}^N x_i s_{ij}$, where s_{ij} is the (i, j)-th entry of the design matrix. Classical CS methods sample the sensing matrix s_{ij} from a Gaussian distribution. More recently, [17] showed that one can also sample s_{ij} from an α-stable distribution [25, 22]. Note that the case of $\alpha = 2$ corresponds to Gaussian.

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

In the case of Gaussian (i.e., $\alpha = 2$), the measurements are also normally distributed because $y_j = \sum_{i=1}^N x_i s_{ij} \sim N(0, \sum_{i=1}^N |x_i|^2)$ if $s_{ij} \sim N(0, 1)$. The α-stable distribution generalizes the Gaussian. Suppose a random variable s is α-stable with scale 1, denoted as $s \sim S(\alpha, 1)$. The characteristic function is $E(e^{\sqrt{-1}st}) = e^{-|t|^\alpha}$. This is also one definition of stable distributions. Consequently, if $s_{ij} \sim S(\alpha, 1)$, then the measurements are also stable $y_j \sim S(\alpha, \sum_{i=1}^N |x_i|^\alpha)$, where the scale $\sum_{i=1}^N |x_i|^\alpha$ corresponds to the l_α norm of the signal.

1.1 1-Bit Compressed Sensing (1-Bit CS)

The problem of 1-bit CS has been studied in the literature of statistics, information theory and machine learning, e.g., [4, 15, 13, 21, 8, 24], due to its many advantages by using only 1-bit (i.e., the sign) of the measurements. The hardware will always have to quantize the measurements in some way; and storing only the signs is the simplest quantization scheme. Also, using only the signs will potentially reduce the cost of storage and transmission. Of course, 1-CS only makes sense if the number of measurements does not increase too much compared to using full information, and the decoding procedure should not be too complicated.

1.2 1-Bit CS with Stable Designs

Prior to [17], existing 1-bit CS methods use the Gaussian design. [17] proposed sampling $s_{ij} \sim S(\alpha, 1)$. They also developed an efficient one-scan procedure for the decoding, as summarized in Algorithm 1. This particular algorithm requires using a small α and it needs to know K, the sparsity of the signal.

In [17], the empirical comparisons with 1-bit marginal regression [21, 24] illustrate that their proposed method needs orders of magnitude fewer measurements. Compared to 1-bit Iterative Hard Thresholding (IHT) [15], the algorithm is still significantly more accurate. Furthermore, while the method in [17] is reasonably robust against random sign flipping, IHT

Algorithm 1 Stable measurement collection and the one scan 1-bit algorithm for sign recovery. [17]

Input: K-sparse signal $\mathbf{x} \in \mathbb{R}^{1\times N}$, design matrix $\mathbf{S} \in \mathbb{R}^{N\times M}$ with entries sampled from $S(\alpha,1)$ with small α (e.g., $\alpha = 0.05$). We sample $u_{ij} \sim uniform(-\pi/2,\ \pi/2)$ and $w_{ij} \sim exp(1)$ and compute s_{ij} by (2).

Collect: measurements: $y_j = \sum_{i=1}^N x_i s_{ij}$, $j = 1$ to M.

Compute: For each coordinate $i = 1$ to N, compute

$$Q_i^+ = \sum_{j=1}^M \log\left(1 + sgn(y_j)sgn(u_{ij})e^{-(K-1)w_{ij}}\right),$$
$$Q_i^- = \sum_{j=1}^M \log\left(1 - sgn(y_j)sgn(u_{ij})e^{-(K-1)w_{ij}}\right)$$

Output: For $i = 1$ to N, report the estimated sign:

$$\hat{sgn}(x_i) = \begin{cases} +1 & \text{if } Q_i^+ > 0 \\ -1 & \text{if } Q_i^- > 0 \\ 0 & \text{if } Q_i^+ < 0 \text{ and } Q_i^- < 0 \end{cases}$$

is known to be very sensitive to that kind of noise.

Thus, we are left with an interesting unsolved estimation problem from quantized stable measurements. In particular, we need to estimate the l_α norm of the signal. Estimating the l_α norm from stable measurements using full information have been studied in machine learning [18, 19] and theory literature [14, 16].

1.3 Problem Formulation

Given samples $y_i \sim N(0,\sigma^2)$, we would like to estimate σ^2 from only 1-bit information of $|y_i|$. More generally, the problem of interest is about bit-efficient estimation of the scale parameter of the α-stable distribution. That is, given n i.i.d. samples,

$$y_j \sim S(\alpha, \Lambda_\alpha), \qquad j = 1,2,...,n \tag{1}$$

from an α-stable distribution $S(\alpha,\Lambda_\alpha)$, we hope to estimate the scale parameter Λ_α by using only 1-bit or multi-bit information of $|y_j|$. Here we adopt the parameterization [25, 22] such that, if $y \sim S(\alpha,\Lambda_\alpha)$, then the characteristic function is $E\left(e^{\sqrt{-1}yt}\right) = e^{-\Lambda_\alpha|t|^\alpha}$. Under this parameterization, when $\alpha = 2$, $S(2,\Lambda_2)$ is equivalent to a Gaussian distribution $N(0,\sigma^2 = 2\Lambda_2)$. When $\alpha = 1$, $S(1,1)$ is the standard Cauchy.

1.4 Sampling from α-stable Distribution

Although in general there is no closed-form density of $S(\alpha,1)$, we can sample from the distribution using a procedure provided by [6]. That is, one can first sample an exponential $w \sim exp(1)$ and a uninform $u \sim unif(-\pi/2,\pi/2)$, and then compute

$$s_\alpha = \frac{\sin(\alpha u)}{(\cos u)^{1/\alpha}}\left[\frac{\cos(u-\alpha u)}{w}\right]^{(1-\alpha)/\alpha} \sim S(\alpha,1) \tag{2}$$

This paper will heavily use the distribution of $|s_\alpha|^\alpha = \frac{|\sin(\alpha u)|^\alpha}{\cos u}\left[\frac{\cos(u-\alpha u)}{w}\right]^{(1-\alpha)}$. Intuitively, as $\alpha \to 0$, $1/|s_\alpha|^\alpha$ converges to $exp(1)$ in distribution [9].

2 Estimation of Λ_α Using Full (Infinite-Bit) Information

Given n i.i.d. samples $y_j \sim S(\alpha,\Lambda_\alpha)$, $j = 1$ to n, we review various estimators of Λ_α using full information. When $\alpha = 2$ (i.e., Gaussian), the *arithmetic mean* estimator is statistically optimal (i.e., the asymptotic variance reaches the reciprocal of Fisher Information):

$$\hat\Lambda_{2,f} = \frac{1}{n}\sum_{j=1}^n |y_j|^2, \qquad Var\left(\hat\Lambda_{2,f}\right) = \frac{\Lambda_2^2}{n}2$$

When $\alpha = 1$, the optimal estimator $\hat\Lambda_{1,f}$ is solved from a nonlinear quation

$$\sum_{j=1}^n \frac{\hat\Lambda_{1,f}^2}{\hat\Lambda_{1,f}^2 + y_j^2} = \frac{n}{2}, \qquad Var\left(\hat\Lambda_{1,f}\right) = \frac{\Lambda_1^2}{n}2 + O\left(\frac{1}{n^2}\right)$$

The *harmonic mean* estimator [16] is suitable for small α and becomes optimal as $\alpha \to 0+$:

$$\hat\Lambda_{\alpha,f,hm} = \frac{-\frac{2}{\pi}\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)}{\sum_{j=1}^n |y_j|^{-\alpha}} \times \left(n - \left(\frac{-\pi\Gamma(-2\alpha)\sin(\pi\alpha)}{\left[\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)\right]^2} - 1\right)\right) \tag{3}$$

$$Var\left(\hat\Lambda_{\alpha,f,hm}\right) = \frac{\Lambda_\alpha^2}{n}\left(\frac{-\pi\Gamma(-2\alpha)\sin(\pi\alpha)}{\left[\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)\right]^2} - 1\right) + O\left(\frac{1}{n^2}\right)$$

where $\Gamma(.)$ is the gamma function. When $\alpha \to 0+$, the variance becomes $\frac{\Lambda_{0+}^2}{n} + O\left(\frac{1}{n^2}\right)$.

[18] also proposed a "fractional-power" estimator which is nearly optimal for entire range of $0 < \alpha \le 2$.

In summary, the optimal variances for $\alpha = 0+$, 1, and 2, are respectively

$$\frac{\Lambda_{0+}^2}{n}1, \qquad \frac{\Lambda_1^2}{n}2, \qquad \text{and} \qquad \frac{\Lambda_2^2}{n}2 \tag{4}$$

Our goal is to develop 1-bit and multi-bit schemes to achieve variances which are close to be optimal.

3 1-Bit Coding and Estimation

Consider n i.i.d. samples $y_j \sim S(\alpha,\Lambda_\alpha)$, $j = 1$ to n. To estimate Λ_α using one bit information of each $|y_j|$, we choose a threshold C and compare it with $|y_j|^\alpha$, $j = 1,2,...,n$. In other word, we store a "0" if $|y_j|^\alpha \le C$ and a "1" if $|y_j|^\alpha > C$. Note that we can express $|y_j|^\alpha$ as

$$z_\alpha = |y_j|^\alpha \sim \Lambda_\alpha |s_\alpha|^\alpha, \qquad s_\alpha \sim S(\alpha,1).$$

Let f_α and F_α be the pdf and cdf of $|s_\alpha|^\alpha$, respectively. Then we can define p_1 and p_2 as follows

$$\begin{aligned}p_1 = \mathbf{Pr}\left(z_\alpha \leq C\right) = F_\alpha\left(C/\Lambda_\alpha\right), \\ p_2 = \mathbf{Pr}\left(z_\alpha > C\right) = 1 - p_1 = 1 - F_\alpha\left(C/\Lambda_\alpha\right)\end{aligned}$$

which are needed for computing the likelihood. Denote

$$n_1 = \sum_{j=1}^n 1\{z_j \leq C\}, \qquad n_2 = \sum_{j=1}^n 1\{z_j > C\}$$

The log-likelihood of the $n = n_1 + n_2$ observations is

$$\begin{aligned}l =& n_1 \log p_1 + n_2 \log p_2 \\ =& n_1 \log F_\alpha\left(C/\Lambda_\alpha\right) + n_2 \log\left[1 - F_\alpha\left(C/\Lambda_\alpha\right)\right]\end{aligned}$$

To seek the MLE (maximum likelihood estimator) of Λ_α, we compute the first derivative $l' = \frac{\partial l}{\partial \Lambda_\alpha}$:

$$l' = n_1 \frac{f_\alpha\left(C/\Lambda_\alpha\right)}{F_\alpha\left(C/\Lambda_\alpha\right)}\left(-\frac{C}{\Lambda_\alpha^2}\right) + n_2 \frac{-f_\alpha\left(C/\Lambda_\alpha\right)}{1 - F_\alpha\left(C/\Lambda_\alpha\right)}\left(-\frac{C}{\Lambda_\alpha^2}\right)$$

Setting $l' = 0$ yields the MLE solution denoted by $\hat{\Lambda}_\alpha$:

$$F_\alpha^{-1}\left(n_1/n\right) = C/\Lambda_\alpha \Longrightarrow \hat{\Lambda}_\alpha = C/F_\alpha^{-1}\left(n_1/n\right)$$

To assess the estimation variance of $\hat{\Lambda}_\alpha$, the classical statistics estimation theory says

$$Var\left(\hat{\Lambda}_\alpha\right) = \frac{1}{-E\left(l''\right)} + O\left(\frac{1}{n^2}\right)$$

After some algebra, we have $E\left(l''\right) = -n\frac{C^2}{\Lambda_\alpha^4}\frac{f_\alpha^2}{F_\alpha(1-F_\alpha)}$. For convenience, we introduce $\eta = \frac{\Lambda_\alpha}{C}$ and summarize the results in Theorem 1, which also provides the $O\left(\frac{1}{n}\right)$ bias term [2, 23].

Theorem 1 *Given n i.i.d. samples $y_j \sim S(\alpha, \Lambda_\alpha)$, $j = 1$ to n, a threshold C, and $n_1 = \sum_{j=1}^n 1\{z_j \leq C\}$, the maximum likelihood estimator (MLE) of Λ_α is*

$$\hat{\Lambda}_\alpha = C/F_\alpha^{-1}\left(n_1/n\right) \tag{5}$$

Denote $\eta = \frac{\Lambda_\alpha}{C}$. The asymptotic bias of $\hat{\Lambda}_\alpha$ is

$$\begin{aligned}E\left(\hat{\Lambda}_\alpha - \Lambda_\alpha\right) =& \frac{\Lambda_\alpha}{n}\frac{n_1}{n}\left(1 - \frac{n_1}{n}\right) \\ &\times \left(\frac{\eta^2}{f_\alpha^2(1/\eta)} + \frac{\eta f'_\alpha(1/\eta)}{2f_\alpha^3(1/\eta)}\right) + O\left(\frac{1}{n^2}\right)\end{aligned} \tag{6}$$

and the asymptotic variance of $\hat{\Lambda}_\alpha$ is

$$Var\left(\hat{\Lambda}_\alpha\right) = \frac{\Lambda_\alpha^2}{n}V_\alpha\left(\eta\right) + O\left(\frac{1}{n^2}\right) \tag{7}$$

$$V_\alpha\left(\eta\right) = \eta^2\frac{F_\alpha(1/\eta)(1 - F_\alpha(1/\eta))}{f_\alpha^2(1/\eta)}, \tag{8}$$

where f_α and F_α are the pdf and cdf of $|S(\alpha,1)|^\alpha$, respectively, and $f'_\alpha(z) = \frac{\partial f_\alpha(z)}{\partial z}$. □

3.1 $\alpha \to 0+$

As $\alpha \to 0+$, we have $1/[s_\alpha]^\alpha \sim exp(1)$. Thus

$$F_{0+}(z) = e^{-1/z}, f_{0+}(z) = \frac{1}{z^2}e^{-1/z}, F_{0+}^{-1}(z) = \frac{1}{\log 1/z}$$

We can then derive the estimator and its variance:

$$\begin{aligned}\hat{\Lambda}_{0+} =& \frac{C}{F_{0+}^{-1}\left(n_1/n\right)} = C\log n/n_1, \\ Var\left(\hat{\Lambda}_{0+}\right) =& \frac{\Lambda_{0+}^2}{n}V_{0+}(\eta) + O\left(\frac{1}{n^2}\right) \\ V_{0+}\left(\eta\right) =& \eta^2\frac{F_\alpha(1/\eta)(1 - F_\alpha(1/\eta))}{f_\alpha^2(1/\eta)} = \frac{e^\eta - 1}{\eta^2}\end{aligned}$$

Numerically, one can compute the minimum of $V_{0+}\left(\eta\right)$ to be 1.544..., attained at $\eta = 1.594...$

3.2 $\alpha = 1$

By properties of Cauchy distribution, we know

$$F_1(z) = \frac{2}{\pi}\tan^{-1} z, f_1(z) = \frac{2}{\pi}\frac{1}{1+z^2}, F_1^{-1}(z) = \tan\frac{\pi}{2}z$$

Thus, we can derive the estimator and variance

$$\hat{\Lambda}_1 = \frac{C}{\tan\frac{\pi}{2}\frac{n_1}{n}}, \qquad Var\left(\hat{\Lambda}_1\right) = \frac{\Lambda_1^2}{n}V_1(\eta) + O\left(\frac{1}{n^2}\right)$$

The minimum of $V_1(\eta)$ is $\frac{\pi^2}{4}$, attained at $\eta = 1$. To see this, let $t = 1/\eta$. Then $V_1\left(\eta\right) = \frac{1}{t^2}\frac{F_1(t)(1-F_1(t))}{f_1^2(t)}$ and

$$\begin{aligned}\frac{\partial \log V_1\left(\eta\right)}{\partial t} =& -\frac{2}{t} + \frac{f_1(t)}{F_1(t)} + \frac{-f_1(t)}{1 - F_1(t)} - 2\frac{f'_1(t)}{f_1(t)} \\ =& -\frac{2}{t} + \frac{4t}{1+t^2} + \frac{\frac{1}{1+t^2}}{\tan^{-1} t} - \frac{\frac{2}{\pi}\frac{1}{1+t^2}}{1 - \frac{2}{\pi}\tan^{-1} t} \\ =& \frac{1}{1+t^2}\left[t^2 - 1 + \frac{1}{\tan^{-1} t} - \frac{1}{\frac{\pi}{2} - \tan^{-1} t}\right]\end{aligned}$$

Setting $\frac{\partial \log V_1(\eta)}{\partial t} = 0$, the solution is $t = 1$. Hence the optimum is attained at $\eta = 1$.

3.3 $\alpha = 2$

Since $S(2,1) \sim \sqrt{2} \times N(0,1)$, i.e., $|s_\alpha|^2 \sim 2\chi_1^2$, we have

$$F_2(z) = F_{\chi_1^2}(z/2), \qquad f_2(z) = f_{\chi_1^2}(z/2)/2,$$

where $F_{\chi_1^2}$ and $f_{\chi_1^2}$ are the cdf and pdf of a chi-square distribution with 1 degree of freedom, respectively. The MLE is $\hat{\Lambda}_2 = \frac{C}{F_2^{-1}(n_1/n)}$. Numerically, the optimal $V_2(\eta) = 3.066...$, attained at $\eta = \frac{\Lambda_2}{C} = 0.228...$

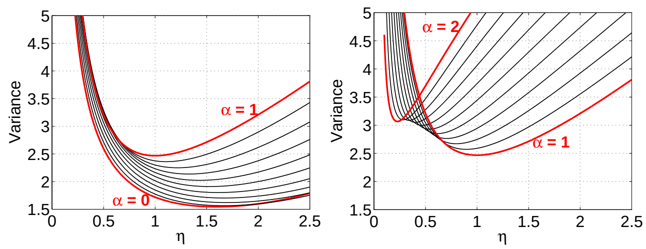

Figure 1: The variance factor $V_\alpha(\eta)$ in (8) for $\alpha \in [0,\ 2]$, spaced at 0.1. The lowest point on each curve corresponds to the optimal variance at that α value.

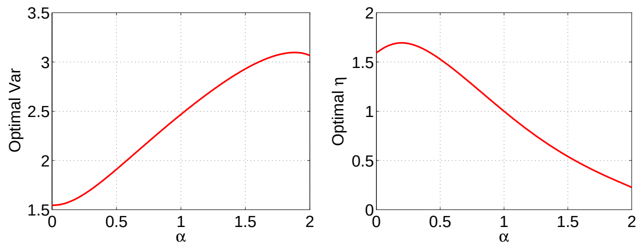

Figure 2: Optimal variance values $V_\alpha(\eta)$ (left panel) and the corresponding optimal η values (right panel). Each point on the curve corresponds to the lowest point of the curve for that α as in Figure 1.

3.4 General $0 < \alpha \leq 2$

For general $0 < \alpha \leq 2$, the cdf F_α and pdf f_α can be computed numerically. Figure 1 plots $V_\alpha(\eta)$ for α from 0 to 2. The lowest point on each curve corresponds to the optimal (smallest) $V_\alpha(\eta)$. Figure 2 plots the optimal V_α values and optimal η values.

Figure 1 suggests that the 1-bit scheme performs reasonably well. The optimal variance coefficient V_α is not much larger than the variance using full (infinite-bit) information . For example, when $\alpha = 1$, the optimal variance coefficient using full information is 2 (i.e., see (4)), while the optimal variance coefficient of the 1-bit scheme is just $\frac{\pi^2}{4} = 2.467...$ which is only 20% larger. Furthermore, we can see that, at least when $\alpha \leq 1$, $V_\alpha(\eta)$ is not very sensitive to η in a wide range of η values, a property which is practically important.

3.5 Error Tail Bounds

Theorem 2

$$\mathbf{Pr}\left(\hat{\Lambda}_\alpha \geq (1+\epsilon)\Lambda_\alpha\right) \leq \exp\left(-n\frac{\epsilon^2}{G_{R,\alpha,C,\epsilon}}\right),\ \ \epsilon \geq 0$$

$$\mathbf{Pr}\left(\hat{\Lambda}_\alpha \leq (1-\epsilon)\Lambda_\alpha\right) \leq \exp\left(-n\frac{\epsilon^2}{G_{L,\alpha,C,\epsilon}}\right),\ \ 0 \leq \epsilon \leq 1$$

where $G_{R,\alpha,C,\epsilon}$ and $G_{L,\alpha,C,\epsilon}$ are computed as follows:

$$\frac{\epsilon^2}{G_{R,\alpha,C,\epsilon}} = -F_\alpha(1/(1+\epsilon)\eta)\log\left[\frac{F_\alpha(1/\eta)}{F_\alpha(1/(1+\epsilon)\eta)}\right] \quad (9)$$
$$-(1-F_\alpha(1/(1+\epsilon)\eta))\log\left[\frac{1-F_\alpha(1/\eta)}{1-F_\alpha(1/(1+\epsilon)\eta)}\right]$$

$$\frac{\epsilon^2}{G_{L,\alpha,C,\epsilon}} = -F_\alpha(1/(1-\epsilon)\eta)\log\left[\frac{F_\alpha(1/\eta)}{F_\alpha(1/(1-\epsilon)\eta)}\right] \quad (10)$$
$$-(1-F_\alpha(1/(1-\epsilon)\eta))\log\left[\frac{1-F_\alpha(1/\eta)}{1-F_\alpha(1/(1-\epsilon)\eta)}\right] \quad \square$$

The tail bounds provide a precise probabilistic guarantee. That is, to ensure error $\mathbf{Pr}\left(\hat{\Lambda}_\alpha \geq (1+\epsilon)\Lambda_\alpha\right) + \mathbf{Pr}\left(\hat{\Lambda}_\alpha \leq (1-\epsilon)\Lambda_\alpha\right) \leq \delta,\ \ 0 \leq \delta \leq 1$, it suffices that

$$\exp\left(-n\frac{\epsilon^2}{G_{R,\alpha,C,\epsilon}}\right) + \exp\left(-n\frac{\epsilon^2}{G_{L,\alpha,C,\epsilon}}\right) \leq \delta \quad (11)$$

for which it suffices $n \geq \frac{G_{\alpha,C,\epsilon}}{\epsilon^2}\log 2/\delta$, where $G_{\alpha,C,\epsilon} = \max\{G_{R,\alpha,C,\epsilon},\ G_{L,\alpha,C,\epsilon}\}$. Figure 3 provides the tail bound constants for $\alpha = 0+$, at selected η values.

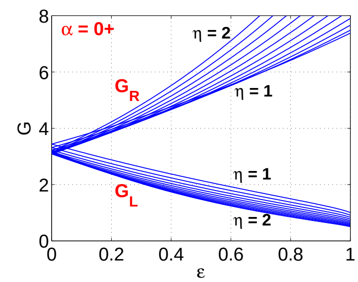

Figure 3: The tail bound constants $G_{R,0+,C,\epsilon}$ (9) (upper group) and $G_{L,0+,C,\epsilon}$ (10) (lower group), for $\eta = 1$ to 2 spaced at 0.1. Recall $\eta = \frac{\Lambda_\alpha}{C}$.

3.6 Bias-Correction

Bias-correction for MLE is important for "small" sample size n. In Theorem 1, Eq. (6) naturally provides a bias-correction for $\hat{\Lambda}_\alpha$, known as the "Bartlett correction" in statistics. To do so, we will need to use the estimate $\hat{\Lambda}_\alpha$ to compute the η. Since $\hat{\Lambda}_\alpha = C/F_\alpha^{-1}(n_1/n)$, we have $\hat{\Lambda}_\alpha/C = 1/F_\alpha^{-1}(n_1/n)$. The bias-corrected estimator, denoted by $\hat{\Lambda}_{\alpha,c}$ is

$$\hat{\Lambda}_{\alpha,c} = \frac{\hat{\Lambda}_\alpha}{1+\frac{1}{n}\frac{n_1}{n}\left(1-\frac{n_1}{n}\right)\left(\frac{\hat{\eta}^2}{f_\alpha^2(1/\hat{\eta})}+\frac{\hat{\eta}f_\alpha'(1/\hat{\eta})}{2f_\alpha^3(1/\hat{\eta})}\right)},$$
$$\text{where}\ \ \hat{\eta} = 1/F_\alpha^{-1}(n_1/n)$$

which, when $\alpha = 0+$, 1, and 2, becomes respectively

$$\hat{\Lambda}_{0+,c} = \frac{C\log n/n_1}{1+\frac{1/n_1-1/n}{2\log n/n_1}},$$

$$\hat{\Lambda}_{1,c} = \frac{\frac{C}{\tan\frac{\pi}{2}\frac{n_1}{n}}}{1+\frac{1}{n}\frac{\pi^2}{4}\frac{n_1}{n}\left(1-\frac{n_1}{n}\right)\left(1+\frac{1}{\tan^2\frac{\pi}{2}\frac{n_1}{n}}\right)}$$

$$\hat{\Lambda}_{2,c} = \frac{\frac{C}{2F_{\chi_1^2}^{-1}(n_1/n)}}{1+\frac{\pi}{n}\frac{n_1}{n}\left(1-\frac{n_1}{n}\right)\left(\frac{3}{F_{\chi_1^2}^{-1}(n_1/n)}-1\right)e^{F_{\chi_1^2}^{-1}(n_1/n)}}$$

Bias-correction for the MLE is an important topic in classical statistics theory [2, 23]. The calculation is in general tedious. In Theorem 1, we are able to compute the bias of order $O\left(\frac{1}{n}\right)$ and hence it is easy to remove it once it is calculated. However, the $O\left(\frac{1}{n}\right)$ term in the variance will still be the same. To see this, we can write $\hat{\Lambda}_{\alpha,c} = \hat{\Lambda}_\alpha\left(1 - \frac{1}{n}e + O\left(\frac{1}{n^2}\right)\right)$, where we use e for the sophisticated constant. Then

$$\begin{aligned} Var\left(\hat{\Lambda}_{\alpha,c}\right) =& Var\left(\hat{\Lambda}_\alpha\right)\left(1 - \frac{2}{n}e + O\left(\frac{1}{n^2}\right)\right) \\ =& \frac{\Lambda_\alpha^2}{n}V_\alpha(\eta) + O\left(\frac{1}{n^2}\right) \end{aligned}$$

Of course, the $O\left(\frac{1}{n^2}\right)$ term of the variance is smaller (i.e., bias-correction also reduces variance), but we will not be able to see the exact expression unless we carry out some even more sophisticated calculations [2, 23].

Interestingly (perhaps also surprisingly), for the problem we study here, the bias-correction step is actually in a sense crucial, as demonstrated in the next section.

4 Experiments on 1-Bit Coding

We conduct simulations to (i) verify the 1-bit variance formulas of the MLE, and (ii) apply the 1-bit estimator to 1-bit compressed sensing [17].

4.1 Bias and Variance

Figure 4 provides the simulations for verifying the 1-bit estimator $\hat{\Lambda}_{0+}$ and its bias-corrected version $\hat{\Lambda}_{0+,c}$ using small α (i.e., 0.05). For each sample size n, we generate 10^6 samples from $S(\alpha, 1)$, which are quantized according a pre-selected threshold C. Then we apply both $\hat{\Lambda}_{0+}$ and $\hat{\Lambda}_{0+,c}$ and report the empirical mean square error (MSE = variance + bias^2) from 10^6 repetitions. For thorough evaluations, we conduct simulations for a wide range of $n \in [5, 1000]$.

The results are presented in log-log scale, which exaggerates the portion for small n and the y-axis for large n. The plots confirm that when n is not too small (e.g., $n > 100$), the bias of MLE estimate varnishes and the asymptotic variance formula (8) matches the mean square error. For small n (e.g., $n < 100$), the bias correction becomes important.

Note that when n is large (i.e., when errors are very small), the plots show some discrepancies. This is due to the fact that we have to use small α for the simulations but the estimators $\hat{\Lambda}_{0+}$ and $\hat{\Lambda}_{0+,c}$ are based on $\alpha = 0+$. The differences are very small and only become visible when the estimation errors are so small (due to the exaggeration of the log-scale). To remove this effect, we conduct similar simulations for $\alpha = 1$ and present the results in Figure 5, which does not show the discrepancies at large n. We can see that the bias-correction step is also important for $\alpha = 1$.

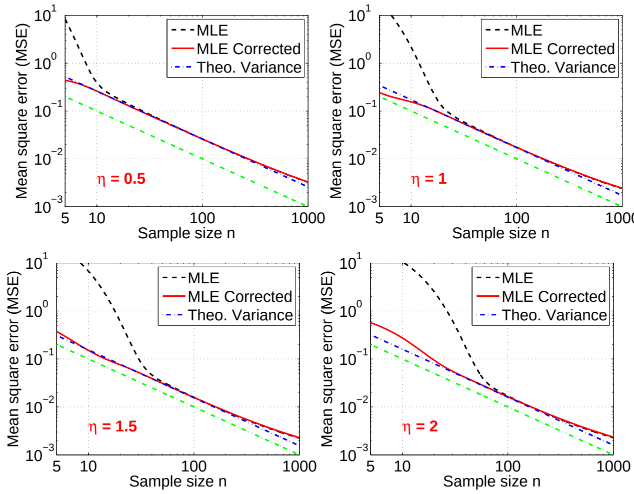

Figure 4: Empirical Mean square errors of $\hat{\Lambda}_{0+}$ (dashed curves) and $\hat{\Lambda}_{0+,c}$ (solid curves) from 10^6 simulations of $S(\alpha, 1)$ for $\alpha = 0.05$, at each sample size n. Each panel present results for a different $\eta = \frac{\Lambda_\alpha}{C}$. For both estimators, the empirical MSEs converge to the theoretical asymptotic variances (8) (dashed dot curves and blue if color is available) when n is large enough. In each panel, the lowest curve (dashed dot and green if color is available) represents the theoretical variance using full (infinite-bit) information, i.e., $1/n$ in this case. For small n, the bias-correction is important.

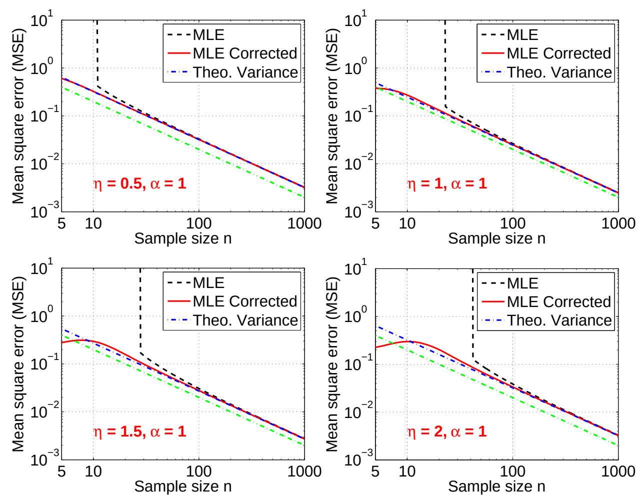

Figure 5: Mean square errors of $\hat{\Lambda}_1$ (dashed curves) and $\hat{\Lambda}_{1,c}$ (solid curves) for $\alpha = 1$. Note that the lowest curve (dashed dot and green if color is available) in each panel represents the optimal variance using full (i.e., infinite-bit) information, which is $2/n$ for $\alpha = 1$.

4.2 One Scan 1-Bit Compressed Sensing (CS)

Algorithm 1 summarizes the recently proposed one-scan 1-bit CS method based on α-stable designs for small α [17], which requires knowing $K = \Lambda_{0+}$ (the sparsity). Note that in their paper, they also used a refined method by making more aggressive use of K.

Here, we replace K in Algorithm 1 with $\hat{\Lambda}_{0+,c}$ for computing Q_i^+ and Q_i^-. We report the sign recovery errors $\sum_i |\widehat{sgn}(x_i) - sgn(x_i)|/K$ from 10^4 simulations. We let $N = 1000$, $K = 20$, and sample the nonzero coordinates from $N(0, 5^2)$. For estimating K, we use $n \in \{50, 100\}$ samples with $\eta \in \{0.2,\ 0.5,\ 1.5,\ 2,\ 3\}$. Recall $\eta = 1.5$ is close to be optimal (1.594...) for $\hat{\Lambda}_{0+}$.

Figure 6 reports the sign recovery errors at 75% quantile (upper panels) and 95% quantile (bottom panels). The number of measurements for sparse recovery is chosen as $M = \zeta K \log(N/0.01)$, although we only use $n \in \{50, 100\}$ samples to estimate K. For comparison, Figure 6 also reports the results for estimating K using n full (i.e., infinite-bit) samples.

When $n = 100$, except for $\eta = 0.2$, the performance of $\hat{\Lambda}_{0+,c}$ is stable with no essential difference from the estimator using full information. The performance for $n = 50$ is less stable.

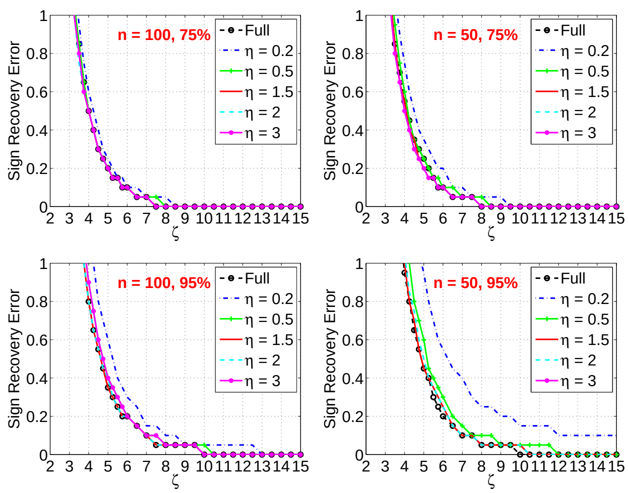

Figure 6: Sign recovery error: $\sum_i |\widehat{sgn}(x_i) - sgn(x_i)|/K$, using Algorithm 1 and estimated K in computing Q_i^+ and Q_i^- in Algorithm 1, for $N = 1000$, $K = 20$. The number of measurements for recovery is $M = \zeta K \log(N/0.01)$ and we use n samples to estimate K for $n \in \{50,\ 100\}$. We report 75% (upper panels) and 95% (bottom panels) quantiles of the sign recovery errors, from 10^4 repetitions. We estimate K using the full information (i.e., the estimator (3)) as well as 1-bit estimator $\hat{\Lambda}_{0+,c}$ with selected values of $\eta \in \{0.2,\ 0.5,\ 1.5,\ 2,\ 3\}$. When $n = 100$, except for $\eta = 0.2$ (which is too small), the performance of $\hat{\Lambda}_{0+,c}$ is fairly stable with no essential difference from the estimator using full information. The performance of $\hat{\Lambda}_{0+,c}$ is not as stable when $n = 50$. Note that, when a curve does not show in the panel (e.g., $n = 50$, $\eta = 3$, and 95%), it basically means the error is too large.

Significance: We claim that the empirical results as shown in Figure 6 are significant for 1-bit CS. It basically says the number of required measurements is $\zeta K \log N$ where ζ is merely a small number like 9. Interested readers may consult [17], which also presented recovery results using 1-bit marginal regression and 1-bit IHT. One can see that with estimated K, Algorithm 1 is still significantly more accurate.

5 2-Bit Coding and Estimation

It is desirable to further stabilize the estimates (and lower the variance) by using more bits. With the 2-bit scheme, we need to introduce 3 threshold values.

Theorem 3 *Given n i.i.d. samples $y_j \sim S(\alpha, 1)$, $j = 1$ to n, three thresholds $0 < C_1 \leq C_2 \leq C_3$, $n_1 = \sum_{j=1}^n 1\{z_j \leq C_1\}$, $n_2 = \sum_{j=1}^n 1\{C_1 < z_j \leq C_2\}$, $n_3 = \sum_{j=1}^n 1\{C_2 < z_j \leq C_3\}$, $n_4 = \sum_{j=1}^n 1\{z_j > C_3\}$, and*

$$\eta_1 = \frac{\Lambda_\alpha}{C_1}, \qquad \eta_2 = \frac{\Lambda_\alpha}{C_2}, \qquad \eta_3 = \frac{\Lambda_\alpha}{C_3}$$

The MLE, $\hat{\Lambda}_\alpha$, is the solution to the equation:

$$\begin{aligned}0 =& n_1 \frac{C_1 f_\alpha(1/\eta_1)}{F_\alpha(1/\eta_1)} + n_2 \frac{C_2 f_\alpha(1/\eta_2) - C_1 f_\alpha(1/\eta_1)}{F_\alpha(1/\eta_2) - F_\alpha(1/\eta_1)}\\ &+ n_3 \frac{C_3 f_\alpha(1/\eta_3) - C_2 f_\alpha(1/\eta_2)}{F_\alpha(1/\eta_3) - F_\alpha(1/\eta_2)} + n_4 \frac{-C_3 f_\alpha(1/\eta_3)}{1 - F_\alpha(1/\eta_3)}\end{aligned}$$

The asymptotic variance of the MLE is

$$Var\left(\hat{\Lambda}_\alpha\right) = \frac{\Lambda_\alpha^2}{n} V_\alpha(\eta_1, \eta_2, \eta_3) + O\left(\frac{1}{n^2}\right)$$

where the variance factor can be expressed as

$$\begin{aligned}\frac{1}{V_\alpha(\eta_1, \eta_2, \eta_3)} =& \frac{1}{\eta_1^2} \frac{f_\alpha^2(1/\eta_1)}{F_\alpha(1/\eta_1)} + \frac{1}{\eta_3^2} \frac{f_\alpha^2(1/\eta_3)}{1 - F_\alpha(1/\eta_3)}\\ &+ \frac{\left[f_\alpha(1/\eta_2)/\eta_2 - f_\alpha(1/\eta_1)/\eta_1\right]^2}{F_\alpha(1/\eta_2) - F_\alpha(1/\eta_1)}\\ &+ \frac{\left[f_\alpha(1/\eta_3)/\eta_3 - f_\alpha(1/\eta_2)/\eta_2\right]^2}{F_\alpha(1/\eta_3) - F_\alpha(1/\eta_2)}\end{aligned}$$

The asymptotic bias is

$$E\left(\hat{\Lambda}_\alpha\right) = \Lambda_\alpha \left(1 + \frac{1}{nB} - \frac{D}{2nB^2}\right) + O\left(\frac{1}{n^2}\right)$$

where

$$\begin{aligned}B =& \frac{\left(-\frac{C_1}{\Lambda_\alpha}\right)^2 f_1^2}{F_1} + \frac{\left[\left(-\frac{C_2}{\Lambda_\alpha}\right) f_2 - \left(-\frac{C_1}{\Lambda_\alpha}\right) f_1\right]^2}{F_2 - F_1}\\ &+ \frac{\left[\left(-\frac{C_3}{\Lambda_\alpha}\right) f_3 - \left(-\frac{C_2}{\Lambda_\alpha}\right) f_2\right]^2}{F_3 - F_2} + \frac{\left(-\frac{C_3}{\Lambda_\alpha}\right)^2 f_3^2}{1 - F_3}\end{aligned}$$

$$\begin{aligned}D =& \frac{\left(-\frac{C_1}{\Lambda_\alpha}\right)^3 f_1 f_1'}{F_1} + \frac{\left(-\frac{C_3}{\Lambda_\alpha}\right)^3 f_3 f_3'}{1 - F_3}\\ &+ \frac{\left[\left(-\frac{C_2}{\Lambda_\alpha}\right) f_2 - \left(-\frac{C_1}{\Lambda_\alpha}\right) f_1\right]\left[\left(-\frac{C_2}{\Lambda_\alpha}\right)^2 f_2' - \left(-\frac{C_1}{\Lambda_\alpha}\right)^2 f_1'\right]}{F_2 - F_1}\\ &+ \frac{\left[\left(-\frac{C_3}{\Lambda_\alpha}\right) f_3 - \left(-\frac{C_2}{\Lambda_\alpha}\right) f_2\right]\left[\left(-\frac{C_3}{\Lambda_\alpha}\right)^2 f_3' - \left(-\frac{C_2}{\Lambda_\alpha}\right)^2 f_2'\right]}{F_3 - F_2}\end{aligned}$$ □

The asymptotic bias formula in Theorem 3 leads to a bias-corrected estimator

$$\hat{\Lambda}_{\alpha,c} = \frac{\hat{\Lambda}_\alpha}{1 + \frac{1}{nB} - \frac{D}{2nB^2}} \tag{12}$$

Note that, with a slight abuse of notation, we still use $\hat{\Lambda}_\alpha$ to denote the MLE of the 2-bit scheme and we rely on the number of parameters (e.g., η_1, η_2, η_3) to differentiate V_α for different schemes.

It is intuitive to see why the 2-bit scheme improves the 1-bit scheme (which only needs one threshold η). Suppose we have the best guessed η for the 1-bit scheme, we can always let $\eta_2 = \eta$ and choose $\eta_3 < \eta < \eta_1$.

5.1 $\alpha \to 0+$

In this case, we can slightly simplify the expression:

$$V_{0+}(\eta_1, \eta_2, \eta_3) = \frac{1}{\frac{(\eta_1-\eta_2)^2}{e^{\eta_1}-e^{\eta_2}} + \frac{(\eta_2-\eta_3)^2}{e^{\eta_2}-e^{\eta_3}} + \frac{\eta_3^2}{e^{\eta_3}-1}}$$

Numerically, the minimum of $V_{0+}(\eta_1, \eta_2, \eta_3)$ is 1.122..., attained at $\eta_1 = 3.365..., \eta_2 = 1.771..., \eta_3 = 0.754...$. The value 1.122... is much smaller than 1.544..., the minimum variance coefficient of the 1-bit scheme.

Figure 7 illustrates that, with the 2-bit scheme, the variance is less sensitive to the choice of the thresholds.

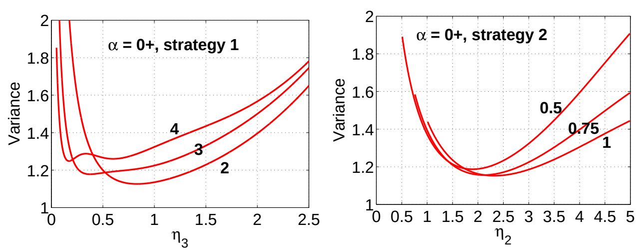

Figure 7: **Left** (strategy 1): $V_{0+}(\eta_1, \eta_2, \eta_3)$ for $\eta_2 = t\eta_3$, $\eta_1 = t\eta_2$, at $t = 2, 3, 4$, with varying η_3. **Right** (strategy 2): V_{0+} for fixed $\eta_1 = 5$, $\eta_3 \in \{0.5,\ 0.75,\ 1\}$, and η_2 varying between η_3 and η_1.

In practice, there are at least two simple strategies for selecting the parameters $\eta_1 \geq \eta_2 \geq \eta_3$:

- *Strategy 1*: First select a "small" η_3, then let $\eta_2 = t\eta_3$ and $\eta_1 = t\eta_2$, for some $t > 1$.
- *Strategy 2*: First select a "small" η_3 and a "large" η_1, then select a "reasonable" η_2 in between.

See the plots for examples of the two strategies in Figure 7. We re-iterate that for the task of estimating Λ_α using only a few bits, we must choose parameters (thresholds) beforehand. While in general the optimal results are not attainable, as long as the chosen parameters fall in a reasonable (and wide) range, the estimation variance will not be far away from optimal.

5.2 $\alpha = 1$

The minimum of $V_1(\eta_1, \eta_2, \eta_3)$ is 2.087..., attained at $\eta_1 = 1.927...$, $\eta_2 = 1$, $\eta_3 = 0.519...$. The value 2.087... is very close to the optimal variance coefficient 2 using full information. Figure 8 plots examples of $V_1(\eta_1, \eta_2, \eta_3)$ for both "strategy 1" and "strategy 2".

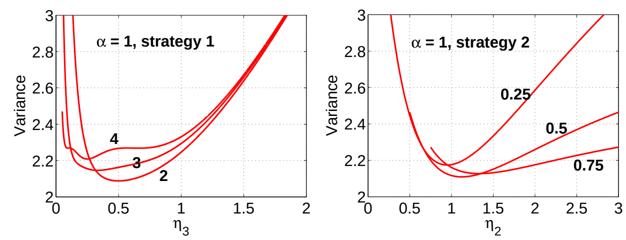

Figure 8: **Left** (strategy 1): $V_1(\eta_1, \eta_2, \eta_3)$ for $\eta_2 = t\eta_3$, $\eta_1 = t\eta_2$, at $t = 2, 3, 4$, with varying η_3. **Right** (strategy 2): V_1 for fixed $\eta_1 = 3$, $\eta_3 \in \{0.25,\ 0.5,\ 0.75\}$, and η_2 varying between η_3 and η_1.

5.3 $\alpha = 2$

Numerically, the minimum of $V_2(\eta_1, \eta_2, \eta_3)$ is 2.236..., attained at $\eta_1 = 0.546...$, $\eta_2 = 0.195...$, $\eta_3 = 0.093...$. Figure 9 presents examples of $V_2(\eta_1, \eta_2, \eta_3)$ for both strategies for choosing η_1, η_2, and η_3.

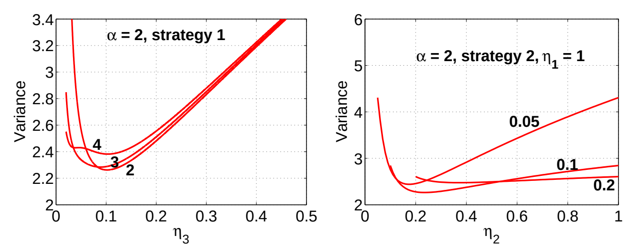

Figure 9: **Left** (strategy 1): $V_1(\eta_1, \eta_2, \eta_3)$ for $\eta_2 = t\eta_3$, $\eta_1 = t\eta_2$, at $t = 2, 3, 4$, with varying η_3. **Right** (strategy 2): V_1 for fixed $\eta_1 = 1$, $\eta_3 \in \{0.05,\ 0.1,\ 0.2\}$, and η_2 varying between η_3 and η_1.

5.4 Simulations

Figure 10 presents the simulation results for verifying the 2-bit estimator $\hat{\Lambda}_{0+}$ and its bias-corrected version $\hat{\Lambda}_{0+,c}$. We choose $\eta_3 \in \{0.05,\ 0.25,\ 0.75,\ \ 2\}$ and fix $\eta_2 = 3\eta_3$, $\eta_1 = 3\eta_2$. Although these choices are not optimal, Figure 10 shows the estimators still perform well for such a wide range of η_3 values. Compared to 1-bit estimators, the 2-bit estimators are noticeably more accurate and less sensitive to parameters. Again, the bias-correction step is useful when the sample size n is not large. Similar to Figure 4, we can observe discrepancies at large n (as magnified by the log-scale), because we simulate data using $\alpha = 0.05$ and we use estimators based on $\alpha = 0+$. To remove this effect, we also provide simulations for $\alpha = 1$ in Figure 11.

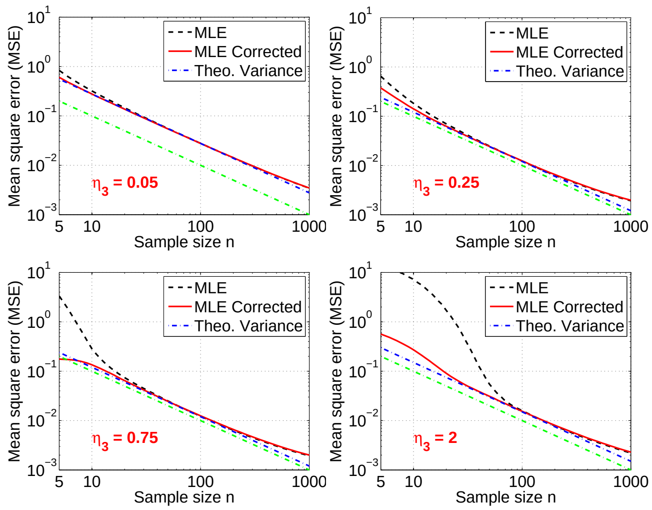

Figure 10: Empirical Mean square errors of the 2-bit estimators: $\hat{\Lambda}_{0+}$ (dashed curves) and $\hat{\Lambda}_{0+,c}$ (solid curves). We use $\alpha = 0.05$ to generate stable samples $S(\alpha, 1)$ and we consider 4 different $\eta_3 = \frac{\Lambda_\alpha}{C_3}$ values. We let $\eta_2 = 3\eta_3$ and $\eta_1 = 3\eta_2$. For both estimators, the empirical MSEs converge to the theoretical asymptotic variances (8) (dashed dot curves and blue if color is available) when n is not small. In each panel, the lowest curve (dashed dot and green if color is available) represents the theoretical variances using full (infinite-bit) information, i.e., $1/n$ in this case.

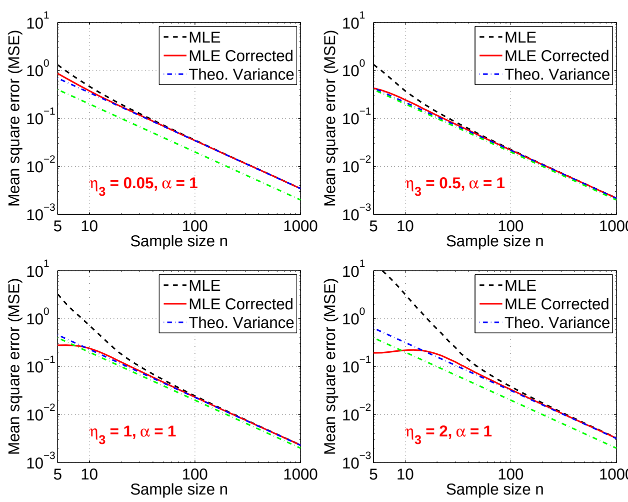

Figure 11: Mean square errors of the 2-bit estimator $\hat{\Lambda}_1$ (dashed curves) and its bias-corrected version $\hat{\Lambda}_{1,c}$ (solid curves), for $\alpha = 1$, by using 6 different η_3 values (one for each panel) and fixing $\eta_2 = 3\eta_3$, $\eta_1 = 3\eta_2$. The lowest curve (dashed dot and green if color is available) in each panel represents the optimal variance using full information, which is $2/n$ for $\alpha = 1$.

5.5 Efficient Computational Procedure

With the 1-bit scheme, the cost for computing the MLE is negligible. With the 2-bit scheme, however, the computational cost might be a concern if we try to find the MLE solution numerically every time (at run time). A tabulation-based efficient procedure is presented in the full paper (arXiv:1503.06876).

6 Multi-Bit (Multi-Partition) Coding

With more bits, it is more flexible to consider schemes based on $(m+1)$ partitions, e.g., $m = 1$ for the 1-bit scheme, $m = 3$ for the 2-bit scheme, and $m = 7$ for the 3-bit scheme. After some algebra, the asymptotic variance of the MLE $\hat{\Lambda}_\alpha$ is derived as

$$Var\left(\hat{\Lambda}_\alpha\right) = \frac{\Lambda_\alpha^2}{n} V_\alpha(\eta_1, ..., \eta_m) + O\left(\frac{1}{n^2}\right), \qquad \text{where}$$

$$\frac{1}{V_\alpha(\eta_1, ..., \eta_m)} = \frac{1}{\eta_1^2}\frac{f_\alpha^2(1/\eta_1)}{F_\alpha(1/\eta_1)} + \frac{1}{\eta_m^2}\frac{f_\alpha^2(1/\eta_m)}{1 - F_\alpha(1/\eta_m)} + \sum_{s=1}^{m-1}\frac{\left[f_\alpha(1/\eta_{s+1})/\eta_{s+1} - f_\alpha(1/\eta_s)/\eta_s\right]^2}{F_\alpha(1/\eta_{s+1}) - F_\alpha(1/\eta_s)}$$

7 Conclusion

The method of random projections has become a routine technique in machine learning and compressed sensing (CS). In many situations, quantized projections will be desirable. For example, 1-bit compressed sensing (1-bit CS) has been popular. When we only need 1-bit per measurement, it becomes convenient/economical to store and transmit the data.

For example, we can have compressed sensing functionality in a handheld device (or remote sensors with very limited storage and bandwidth) [11]. We collect compressed signals (e.g., images or videos) and transmit them to a central sever for processing. For this kind of applications, an effective algorithm for 1-bit compressed sensing will be essential. Note that the remote signals can also come from the space [3].

Existing 1-bit CS algorithms require knowing the scale of the signal, such as the l_0 or the l_2 norm. This motivates estimating the scale parameter of an α-stable distribution from quantized samples. In this paper, we develop 1-bit and multi-bit coding schemes, which perform well (even with just 1-bit) in that, if the parameters are chosen appropriately, their variances are actually not much larger than the variances of using full (i.e., infinite-bit) information. In general, using more bits increases the computational and/or storage cost, with the benefits of stabilizing the estimates when the chosen parameters are not much away from optimal.

One interesting and in fact surprising observation is that the classical "Bartlett correction" for MLE bias-correction is particularly effective in our case.

Finally, we should mention that, although this paper as presented has focused on the motivation about 1-bit CS, the work also provides an efficient mechanism for data stream computations [1, 12, 7, 20, 14, 16].

Acknowledgement: The work is supported in part by NSF-Bigdata-1419210 and NSF-III-1360971.

References

[1] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *STOC*, pages 20–29, Philadelphia, PA, 1996.

[2] Maurice S. Bartlett. Approximate confidence intervals, II. *Biometrika*, 40(3/4):306–317, 1953.

[3] Jrme Bobin, Jean-Luc Starck, and Roland Ottensamer. Compressed sensing in astronomy. *IEEE Journal OF Selected Topics in Signal Processing*, 2(5):718–726, 2008.

[4] P.T. Boufounos and R.G. Baraniuk. 1-bit compressive sensing. In *Information Sciences and Systems, 2008.*, pages 16–21, March 2008.

[5] Emmanuel Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, Feb 2006.

[6] John M. Chambers, C. L. Mallows, and B. W. Stuck. A method for simulating stable random variables. *Journal of the American Statistical Association*, 71(354):340–344, 1976.

[7] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. *Theor. Comput. Sci.*, 312(1):3–15, 2004.

[8] Sheng Chen and Arindam Banerjee. One-bit compressed sensing with the k-support norm. In *AISTATS*, 2015.

[9] N. Cressie. A note on the behaviour of the stable distributions for small index. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 31(1):61–64, 1975.

[10] David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, April 2006.

[11] JAlbert C. Fannjiang, Thomas Strohmer, and Pengchong Yan. Compressed remote sensing of sparse objects. *SIAM J. Imaging Sci.*, 3(3):595–618, 2000.

[12] Joan Feigenbaum, Sampath Kannan, Martin Strauss, and Mahesh Viswanathan. An approximate l_1-difference algorithm for massive data streams. In *FOCS*, pages 501–511, New York, 1999.

[13] Sivakant Gopi, Praneeth Netrapalli, Prateek Jain, and Aditya Nori. One-bit compressed sensing: Provable support and vector recovery. In *ICML*, 2013.

[14] Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *Journal of ACM*, 53(3):307–323, 2006.

[15] Laurent Jacques, Jason N. Laska, Petros T. Boufounos, and Richard G. Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE Transactions on Information Theory*, 59(4):2082–2102, 2013.

[16] Ping Li. Estimators and tail bounds for dimension reduction in l_α ($0 < \alpha \leq 2$) using stable random projections. In *SODA*, pages 10 – 19, San Francisco, CA, 2008.

[17] Ping Li. One scan 1-bit compressed sensing. In *AISTATS*, 2016.

[18] Ping Li and Trevor J. Hastie. A unified near-optimal estimator for dimension reduction in l_α ($0 < \alpha \leq 2$) using stable random projections. In *NIPS*, Vancouver, BC, Canada, 2007.

[19] Ping Li, Trevor J. Hastie, and Kenneth W. Church. Nonlinear estimators and tail bounds for dimensional reduction in l_1 using cauchy random projections. *Journal of Machine Learning Research*, 8:2497–2532, 2007.

[20] S. Muthukrishnan. Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, 1:117–236, 2005.

[21] Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2013.

[22] Gennady Samorodnitsky and Murad S. Taqqu. *Stable Non-Gaussian Random Processes.* Chapman & Hall, New York, 1994.

[23] Leonard R. Shenton and Kimiko O. Bowman. Higher moments of a maximum-likelihood estimate. *Journal of Royal Statistical Society B*, 25(2):305–317, 1963.

[24] Martin Slawski and Ping Li. b-bit marginal regression. In *NIPS*, Montreal, CA, 2015.

[25] Vladimir M. Zolotarev. *One-dimensional Stable Distributions.* American Mathematical Society, Providence, RI, 1986.