Efficient Reconstruction of Sparse Vectors from Quantized Observations

Amine Mezghani and Josef A. Nossek
Institute for Circuit Theory and Signal Processing
Munich University of Technology, 80290 Munich, Germany
E-Mail: {Mezghani, Nossek}@nws.ei.tum.de

Abstract—Compressive sensing is a recent technique for estimating a sparse vector from a reduced number of observations. Several algorithms have been developed and studied in this context. However, the reduction of number of samples (or the sampling rate) is one aspect of compressive sensing. In fact, the quantization of the observations, which is generally unavoidable in practice (due to the analog-to-digital conversion), could be also understood as another aspect of compressed sensing aiming at reducing the complexity of the sampling device. Moreover, reducing the ADC resolution might be much more beneficial in terms of circuit complexity and power consumption than decreasing the sampling rate. Therefore, we investigate this further aspect of compressive sensing related to the resolution of measurements instead of their number. To this end, the problem of reconstructing a signal vector from quantized measurements is formulated as $\ell_1$ and $\ell_2$ regularized problems and a low complexity iterative algorithm is derived, which does not require information about the input distribution. The presented algorithm, can be considered as a generalization to the thresholding algorithm proposed in [4] for the unquantized case. In other words, we propose a general formulation and solution for the sparse signal estimation problem with quantized observations. Experimental results are provided to demonstrate the validity and performance of the presented algorithm quite well.

By simulation, we also study the effect of quantization and noise on the estimation performance of the algorithm. Additionally, we carry out a state evolution analysis similarly to [3], [1] and [5] to evaluate the compromise between resolution and the number of observations and study the possibility of optimizing the quantizer characteristic. We note that the state evolution analysis is not mathematically rigorous since it is based on some heuristic assumptions. Nevertheless, simulations demonstrate the validity of the state evolution analysis, since it predicts the performance of the proposed algorithm.

I. INTRODUCTION

Most of the contributions on reconstructing sparse vectors assume that the observation vector is available with infinite precision. In practice, however, a quantizer (e.g. A/D-converter) is applied to the observed analog signal, so that the data can be processed in the digital domain. Recently, there are several works that deal with the problem of reconstructing a sparse vector from quantized input. In [2], a fixed point iterative reconstruction algorithm was proposed based on a maximum likelihood approach. A generalized approximate message passing was derived in [1] based on the work [3], and a quantizer optimization by means of a state evolution analysis was carried out. In fact, approximate message passing algorithms can achieve very good reconstruction performance while having a low complexity compared to other algorithms. However, the proposed message passing-like algorithm in [1] still require the knowledge of the input statistic since it is based on a Bayesian formulation.

There are many applications where severe quantization may be unavoidable or even preferred. In fact, the quantization of the observations which is generally unavoidable in practice (due to the analog-to-digital conversion) could be also understood as an other aspect of compressed sensing aiming at reducing the complexity of the sampling device. Moreover, reducing the ADC resolution might be much more beneficial in terms of circuit complexity and power consumption than decreasing the sampling rate. Therefore, we investigate this further aspect of compressive sensing related to the resolution of measurements instead of their number.

To this end, the problem of reconstructing a signal vector from quantized measurements is formulated as $\ell_1$ and $\ell_2$ regularized problems and a low complexity iterative algorithm is derived, which does not require information about the input distribution. The presented algorithm, can be considered as a generalization to the thresholding algorithm proposed in [4] for the unquantized case. In other words, we propose a general formulation and solution for the sparse signal estimation problem with quantized observations. Experimental results are provided to demonstrate the validity and performance of the presented algorithm quite well.

By simulation, we also study the effect of quantization and noise on the estimation performance of the algorithm. Additionally, we carry out a state evolution analysis similarly to [3], [1] and [5] to evaluate the compromise between resolution and the number of observations and study the possibility of optimizing the quantizer characteristic. We note that the state evolution analysis is not mathematically rigorous since it is based on some heuristic assumptions. Nevertheless, simulations demonstrate the validity of the state evolution analysis, since it predicts the performance of the proposed algorithm.

II. PROBLEM FORMULATION

We consider the compressed sensing reconstruction problem, where we aim at estimating a sparse $M$-dimensional vector $x \in \mathbb{R}^M$ according to the linear model

$$ y = Ax + \eta, $$

where $y$ is the noisy unquantized measurement vector, $A \in \mathbb{R}^{N \times M}$ is the random measurement matrix, and $\eta$ refers to Gaussian noise vector with covariance $R_\eta = E[\eta\eta^T] = \sigma_\eta^2 I$. Usually the random measurement matrix $A$ is normalized to have unit column variance, which ensures that $E_A[\|Ax\|^2] = \|x\|^2$ for any input $x$. Additionally, we define

$$ \beta = \frac{N}{M}, $$

as the number of measurements per vector coefficient, i.e., the oversampling factor.

In a practical system, each receive signal component $y_j$, $1 \leq j \leq N$, is quantized by a $b$-bit resolution scalar quantizer...
(A/D-converter). Thus, the resulting quantized signals read as
\[ r_j = Q(y_j), \]  
where \( Q(\cdot) \) denotes the quantization operation. For the case that we use a uniform symmetric mid-riser type quantizer, the quantized receive alphabet for each dimension is given by
\[ r_j \in \left\{ \left( -\frac{b}{2} - 1, 2 + k \right) \Delta; \ k = 1, \ldots, 2^b \right\} = \mathcal{R}, \]  
where \( \Delta \) is the quantizer step-size and \( b \) the number of quantizer bits, which are set the same for all the quantizers.

With these definitions, the conditional probability of the quantized output given an input \( x \) reads as
\[ Pr(r|\mathbf{x}) = \sum_{j=1}^{N} \rho(r_j|a_j^{T}\mathbf{x}), \]  
where \( a_j^{T} \) is the \( j \)-th row of \( A \) and
\[ \rho(r_j|a_j^{T}\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma_0} \int_{r_j}^{r_{j_0}} e^{-\frac{(y-a_j^{T}\mathbf{x})^2}{2\sigma_0^2}} dy = \Phi\left( \frac{r_{j_0} - a_j^{T}\mathbf{x}}{\sigma_0} \right) - \Phi\left( \frac{r_j - a_j^{T}\mathbf{x}}{\sigma_0} \right), \]  
with \( \Phi(x) \) represents the cumulative Gaussian distribution given by
\[ \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt. \]

Herewith, the lower and upper quantization boundaries are (for uniform symmetric quantizers)
\[ r_{j_0} = \begin{cases} &r_j - \frac{\Delta}{2} \quad \text{for } r_j \geq -\frac{\Delta}{2}(2^b - 2) \\ &-\infty \quad \text{otherwise}, \end{cases} \]
and
\[ r_{j_0}^{\uparrow} = \begin{cases} &r_j + \frac{\Delta}{2} \quad \text{for } r_j \leq \frac{\Delta}{2}(2^b - 2) \\ &+\infty \quad \text{otherwise}. \end{cases} \]

III. APPROXIMATIVE MAX-SUM ALGORITHM

To estimated the vector \( \mathbf{x} \) from the quantized output vector \( r \), we aim at solving the following \( \ell_p \) (\( p \in \{1, 2\} \)) regularized optimization problem
\[ \hat{x} = \operatorname{argmax}_{x} \left[ \frac{1}{\lambda'} \sum_{j=1}^{N} \ln(\rho(r_j|a_j^{T}\mathbf{x})) - \frac{1}{p} \|a_j^{T}\mathbf{x}\|_{p}^{p} \right], \]  
with a certain regularization parameter \( \lambda' \) that can be chosen depending on the noise variance and the sparsity of \( x \). Using the generalized approximate message passing (GAMP) approach [3], [6], the following iterative approximative MAX-SUM algorithm, which will be derived in the next subsections, has been obtained to solve both regularized problems:

**Initialization:** \( z^0 = 0 \), \( x^0 = 0 \), \( \theta^0 = \text{const} \neq 0 \)

**Repeat until convergence:**

**Output Step:**
\[ z^t = g_t(-A\mathbf{x}^{t-1} + \theta^{t-1}, r) \]
\[ \xi^t = \langle g_t(-A\mathbf{x}^{t-1} + \theta^{t-1}, z^{t-1}, r) \rangle \]

**Input Step:**
\[ x^t = f_t(A^{T}z^t + \xi^t x^{t-1}) \]
\[ \theta^t = \beta \cdot \langle f_t(A^{T}z^t + \xi^t x^{t-1}) \rangle \]

The scalar functions \( f_t(v) \) and \( g_t(u, r) \) introduced above are applied to their arguments component-by-component and they are given by the expressions
\[ f_t(v) = \operatorname{argmax}_{x} \left[ -\frac{1}{p} \|x\|_{p} - \frac{1}{2\sigma_t^2} (\xi^t x - v)^2 \right], \]
\[ g_t(u, r) = \frac{1}{\theta^t} \operatorname{argmax}_{w} \left[ \lambda' \ln(\rho(r|w) - \frac{1}{2\theta^t} (w - u)^2) \right] - \frac{u}{\theta^t}, \]
while \( f_t(v) \) and \( g_t(u, r) \) are their respective derivative with respect to the first argument. Further, we have introduced the notation \( (v) = \sum_{i=1}^{N} v_i/N \) for any vector \( v \in \mathbb{R}^{N} \).

Interestingly, the function \( g_t(u, r) \) in (14) is quite simple for the noiseless case and takes the following form regardless of the considered vector norm
\[ g_t(u, r)|_{\sigma_0 = 0} = \begin{cases} 0 & \text{if } r_{lo} \leq u \leq r_{up} \\ \frac{1}{\pi}(r_{lo} - u) & \text{if } u \leq r_{lo} \\ \frac{1}{\pi}(r_{up} - u) & \text{if } u \geq r_{up} \end{cases}, \]  

For the noisy case, there is no closed form expression for \( g_t(u, r) \). Nevertheless, we can use the following approximation to solve the optimization in (14) in closed form
\[ \frac{1}{N} \ln \left[ \Phi\left( \frac{r_{up}^{\uparrow} - x}{\sigma_0} \right) - \Phi\left( \frac{r_{lo}^{\downarrow} - x}{\sigma_0} \right) \right] \approx \begin{cases} 0 & \text{if } r_{lo} + c_3 \leq u \leq r_{up} - c_3 \\ -\frac{1}{\lambda}(r_{lo}^{\downarrow} + c_3 - x)^2 & \text{if } x \leq r_{lo} + c_3 \\ -\frac{1}{\lambda}(r_{up}^{\uparrow} - c_3 - x)^2 & \text{if } x \geq r_{up} - c_3 \end{cases}, \]

with appropriately chosen constant \( c_3 \) and \( \lambda \). A low complexity approximation of the output function is then given by
\[ g_t(u, r) = \begin{cases} 0 & \text{if } r_{lo} + c_3 \leq u \leq r_{up} - c_3 \\ \frac{1}{\theta^t + \lambda}(r_{lo}^{\downarrow} + c_3 - u) & \text{if } u \leq r_{lo} + c_3 \\ \frac{1}{\theta^t + \lambda}(r_{up}^{\uparrow} - c_3 - u) & \text{if } u \geq r_{up} - c_3 \end{cases}. \]

We have found by trying several values that choosing
\[ c_3 = \min\{1.8 \cdot \sigma_0, (r_{up} - r_{lo})/2\}, \]
provides good results (c.f. Fig. 1), while the regularization parameter \( \lambda \) has to be adapted depending on the noise variance \( \sigma_0^2 \) and the sparsity of the vector \( x \). Also, we see that (17) is equivalent to (15) when \( \lambda = c_3 = 0 \).

On the other hand, the function \( f_t(v) \) defined by the optimization problem (13) depends on the considered norm. For the \( \ell_1 \) problem, it is given by
\[ f_t(v) = \operatorname{sign}(v) \max(|v| - 1, 0)/\xi^t, \]  

\[ (15) \]
\[ (16) \]
\[ (17) \]
\[ (18) \]
\[ (19) \]
whereas the $\ell_2$ problem is solved by means of
\[ f_t(v) = \frac{v}{1 + \xi^t}. \] (20)

The output step of the algorithm evaluates the error vector $z$ (i.e., the uncertainty) between the estimated noiseless output and the observation. We notice that the term $\theta^T \cdot z^t$ ensures the convergence behavior and thus the reconstruction performance. Based on this error vector, an update is carried out for the estimated vector using the input function $f_t(v)$ obtained according to the special norm regularization.

A. Factor Graph Representation

To derive the algorithm, and in analogy to [7], a factor graph representation of the quantized measurement system is introduced in Fig. 2. Each unknown coefficient $x_i$ is represented by a circle, referred to symbol node, and each received quantized signal $r_j$ corresponds to a square, called the signal node. Each edge connecting $i$ and $j$ represents the corresponding gain factor $a_{j,i}$. If $a_{j,i} \neq 0$. Ignoring the cycles in the graph, let us derive the so-called “loopy” message passing algorithm (or MAX-SUM algorithm) from the factor graph representation.¹

Fig. 2. Factor-graph representation of the quantized MIMO channel.

¹We note that the MAX-SUM algorithm is optimal for cycle-free graphs and performs nearly optimal in sparse graphs. In the case of dense, large enough, channel matrices, it may provide good approximate solution as we will see later.

B. Derivation of the approximate MAX-SUM algorithm

For the derivation of the presented algorithm, a message passing algorithm based on the MAX-SUM rules while ignoring cycles in the underlying factor graph is considered. This approach is fairly similar to the derivation in [3]. It is made possible due to the structure of the cost function (8) which consists of a sum of input (input vector norm) and output (output probabilities) terms. Each iteration $t$ of this algorithm consists in, first sending messages from each symbol node $j$ to each symbol node $i$ (output step), and then vice versa (input step). The output step message is given by
\[ \lambda^t_{j,i}(x_i) = \max_{x_i \in x_i} \frac{\ln p_r(r_j | a^T_j x)}{\lambda^t_{j,i}} + \sum_{i' \neq i} \pi^t_{j,i'}(x_{i'}) \] (21)
The input step message is then
\[ \pi^t_{j,i}(x_i) = -\frac{1}{p} |x_i|^p + \sum_{j' \neq j} \lambda^t_{j',i}(x_i). \] (22)

The messages are initialized at $t = 0$ with $\lambda^0_{j,i}(x_i) = 0$, and the final estimate can be calculated as the maximizer of
\[ \pi^t_{j,i}(x_i) = \frac{1}{p} |x_i|^p + \sum_j \lambda^t_{j,i}(x_i). \] (23)

We now provide an approximation for the MAX-SUM algorithm, which becomes exact in the large system limit. It is based on the fact that the entries $a_{j,i}$ scales as $1/\sqrt{N}$. To this end, we introduce the following values
\[ \hat{x}^t_{i,j} = \arg \max_{x_i} \pi^t_{j,i}(x_i) \]
\[ \hat{x}^t_{j,i} = \arg \max_{x_i} \pi^t_{j,i}(x_i) \]
\[ 1 \mu^t_{j,i} = -\frac{\partial}{\partial x_i} \pi^t_{j,i}(\hat{x}^t_{i,j}) \] (24)
\[ 1 \mu^t_{j,i} = -\frac{\partial}{\partial x_i} \pi^t_{j,i}(\hat{x}^t_{i,j}). \]

Then, we can use the following second order approximation for the input step messages around its maximizer,
\[ \pi^t_{j,i}(x_i) \approx \pi^t_{j,i}(\hat{x}^t_{i,j}) - \frac{1}{2\mu^t_{j,i}} (x_i - \hat{x}^t_{j,i})^2, \] (25)
where we assume $\mu^t_{j,i} \approx \mu^t_{j,i}$ in the large system limit. Thus, the cost function in (21) becomes
\[ \frac{\ln p_r(r_j | a^T_j x)}{\lambda^t_{j,i}} + \sum_{i' \neq i} \pi^t_{j,i'}(\hat{x}^t_{i,j}) - \frac{1}{2\mu^t_{j,i}} (x_i - \hat{x}^t_{j,i})^2. \] (26)

The motivation for such approximation is that the impact of $x_i$ on the function $ln p_r(r_j | a^T_j x)$ is asymptotically very small and that the maximizing value of $x_i$ for the cost function (21) will not deviate much from $\hat{x}_{j,i}$. We solve the maximization in (21) in two steps
\[ \max_{q_j} p_r(r_j | q_j) - \frac{1}{2\mu^t_{j,i}} (x_i - \hat{x}^t_{j,i})^2, \] (27)
The solution for the inner maximization leads to
\[ \max_{q_j} p_r(r_j | q_j) - \frac{1}{2\mu^t_{j,i}} (q_j - \hat{a}_{j,i} - a_{j,i} x_i)^2 \] (28),
where, we introduced the variables
\[ q_j = a_{j,i}x_i + \sum_{i' \neq i} a_{j,i'}x_{i'}, \]
\[ \hat{a}^j_{j,i} = \sum_{i' \neq i} a_{j,i'}\hat{x}_{j,i'}, \]
\[ \theta^j_{j,i} = \sum_{i' \neq i} |a_{j,i'}|^2 \mu^j_{i'} \approx \sum_i |a_{j,i}|^2 \mu_i = \theta_j \ \forall i, \]
and
where the approximation holds for the large system limit. Furthermore, by defining
\[ \hat{a}^j_i = \sum_i a_{j,i}\hat{x}_{j,i}, \]
we get
\[ \max_{q_j} \frac{\ln(\rho(r_j|q_j))}{\lambda'} - \frac{1}{2\theta_j}(q_j - \hat{a}^j_i - a_j,\hat{x}_{j,i} - a_{j,i}x_i)^2. \]

These obtained output step messages can be then approximated as
\[ \lambda^j_{j,i}(x_i) = \max_{q_j} \frac{\ln(\rho(r_j|q_j))}{\lambda'} - \frac{1}{2\theta_j}(q_j - \hat{a}^j_i - a_j,\hat{x}_{j,i} - a_{j,i}x_i)^2 \]
\[ = \max_{q_j} \frac{\ln(\rho(r_j|q_j))}{\lambda'} - \frac{1}{2\theta_j}(q_j - \hat{a}^j_i - a_j,\hat{x}_{j,i} - \hat{x}^j_i)^2 \]
\[ \approx \max_{q_j} \frac{\ln(\rho(r_j|q_j))}{\lambda'} - \frac{1}{2\theta_j}(q_j - \hat{a}^j_i - a_j,\hat{x}_{j,i} - \hat{x}^j_i)^2, \]
where the approximation is obtained by neglecting the terms of order \( a^2_{j,i} \). Let us now define
\[ \hat{a}^j_i = \arg\max_{q_j} \frac{\ln(\rho(r_j|q_j))}{\lambda'} - \frac{1}{2\theta_j}(q_j - \hat{a}^j_i)^2, \]
and
\[ z^j_i = \frac{1}{\theta_j}(\hat{a}^j_i - \hat{a}^j_i)\hat{g}_r(-\hat{a}^j_i, r_j), \]
As done before for the input step message, we derive now a second order expansion of the output step message around \( \hat{x}^j_i \). Evaluating the first derivative
\[ \frac{\partial \lambda^j_{j,i}(x_i)}{\partial x_i} \bigg|_{x_i = \hat{x}^j_i} = a_{j,i}\hat{x}_{j,i}, \]
and the second derivative
\[ \frac{\partial^2 \lambda^j_{j,i}(x_i)}{\partial x_i^2} \bigg|_{x_i = \hat{x}^j_i} = a_{j,i}\hat{g}_r(-\hat{a}^j_i, r_j)\hat{a}_{j,i}^2, \]
leads to
\[ \lambda^j_{j,i}(x_i) \approx \text{const} - \frac{1}{2\theta_j^2}(\mu^j_{i'}\hat{x}_{j,i'} - a_{j,i}(x_i - \hat{x}^j_i))^2. \]

The input step messages can be obtained now as (c.f. 22)
\[ \pi^j_{j,i}(x_i) = -\frac{1}{p}|x_i|^p - \frac{1}{2\xi_{j,i}}(\hat{a}^j_i - \xi_{j,i}x_i)^2, \]
\[ \xi_{j,i} = \max_{i' \neq j} (a_{j,i'}z_{j,i'} + \sum_{i,i'} a_{j,i}z_{j,i'}^2) \]
\[ \approx \xi_{j,i},\hat{x}_{j,i}^2 \sum_{i} a_{j,i}z_{j,i}^2 - a_{j,i}z_{j,i}^2, \]

and
\[ \xi_{j,i} = \sum_{i} a_{j,i}^2 \mu_i \approx \sum_i a_{j,i}^2 \mu_i = \xi_{j,i} \ \forall j \]
\[ \rightarrow \frac{1}{N}\sum_j \hat{g}_r(-\hat{a}^j_i, r_j)\xi_{j,i}. \]

We can proceed to compute \( \hat{x}^j_i \)
\[ \hat{x}^j_i = \arg\max_{x_i} \frac{-1}{p}|x_i|^p - \frac{1}{2\xi_{j,i}}(\hat{a}^j_i - \xi_{j,i}x_i)^2 \]
\[ \approx \hat{x}^j_i - \Gamma_{j}^i a_{j,i}\hat{z}_{j,i}, \]
where we define
\[ \hat{z}^j_i = \arg\max_{x_i} \rho_{j}(x_i) \]
\[ = \arg\max_{x_i} \frac{-1}{p}|x_i|^p - \frac{1}{2\xi_{j,i}}(\hat{a}^j_i - \xi_{j,i}x_i)^2 \]
\[ \approx \hat{z}^j_i, \]
\[ \hat{f}_i(\hat{v}_i), \]
and
\[ \Gamma_{j}^i = \frac{\partial \hat{x}^j_i}{\partial \hat{v}_i} = \hat{f}_i(\hat{v}_i). \]

Next, it can be shown due to the fact that the first derivative of \( \pi_{j}^{j+1}(\hat{x}_{j}^{j+1}) = 0 \) that \( \Gamma_{j}^i = \mu_{i}^j \) (c.f. 24). Therefore, 31 becomes
\[ \theta_{j} = \sum_{i} a_{j,i}^2 \hat{f}_i(\hat{v}_i) \rightarrow \frac{\beta}{M} \sum_i \hat{f}_i(\hat{v}_i) = \theta \ \forall j. \]

Finally, 32 becomes using 43 and the fact that \( \mu_{i}^j = \Gamma_{j}^i = \hat{f}_i(\hat{v}_i) \) as follows
\[ \hat{u}_j = \sum_i a_{j,i}\hat{x}_i - \theta_{j}\hat{x}_j. \]
It can be observed that the steps presented in the algorithm corresponds to Eqs 47, 36, 42, 41, 44 and 46.

IV. STATE EVOLUTION ANALYSIS

State evolution analysis (also known as density evolution for the case of sparse matrices) is a powerful tool to study the behavior of message passing algorithms in the large-system limit [8, 3]. The large-system limit means that we consider the limit when \( M \) and \( N \) go to infinity, while the ratio \( \beta = M/N \) is kept fixed. Under the conjecture that the presented algorithm converges to a certain fixed point, this analysis would deliver useful theoretical results about its reconstruction performance under quantization. For the analysis, we assume
a random matrix \( A \), where the entries \( \{ a_{j,i} \} \) are i.i.d. with zero mean and variance \( 1/N \).

Such analysis has been used in coding theory for analyzing the behavior of iterative algorithms and was justified on the fact that the underlying graph is sparse. In general, the state evolution analysis is not mathematically rigorous and it is based on some heuristic assumptions. However, it was shown in [6] that the state evolution analysis provide mathematically exact results for the case of dense graphs, where the entries of the matrix is drawn from a Gaussian distribution but when the observation vector is not quantized. For the case of quantized observations, a mathematical rigorous proof of its correctness is still missing. Nevertheless, simulations demonstrate the usefulness of this analysis, especially as an analytical way for the offline optimization of the system parameters, such as the quantizer characteristic, as done in [3].

The state evolution analysis is based on the assumption that the components of the estimated vector \( u^t = Ax^t - \theta^t z^t \) (c.f. (9)) at each iteration and the noiseless vector \( v = A \cdot x \) as well as the quantized observation vector \( r \) converge empirically to three random variables \( u, w \) and \( r \) given by the joint distribution

\[
 f(u, w, r) = \rho(r|w) \cdot \rho_z(u, w),
\]

where \( \rho(r|w) \) is given in (6) and \( \rho_z(u, w) \) is the bivariate Gaussian probability density function with the covariance matrix

\[
 E \left[ \begin{bmatrix} w \\ u \end{bmatrix} \right]^T \cdot \left[ \begin{bmatrix} w \\ u \end{bmatrix} \right] = C_{w,u}^t,
\]

in other words,

\[
 (w, u) \sim \mathcal{N}(0, C_{w,u}^t).
\]

On the other hand, the components of the input vector \( x, v^t = A^T z^t + \xi^t x^{t-1} \) (c.f. (11)), as well as \( x^t \) at each iteration \( t \) converges empirically to a joint distribution described by the linear one-dimensional model

\[
 v^t = x + n \text{ with } n \sim \mathcal{N}(0, \nu^{t,2}),
\]

and the nonlinear scalar estimation rule

\[
 x^t = f_t(v^t).
\]

This one-dimensional equivalent model for the input step of the algorithm is illustrated in Fig. 3. In fact, the state evolution analysis implies that the performance can be fully described by the scalar one dimensional estimation problem, where the underlying observation channel is purely Gaussian and surprisingly linear and does not present any nonlinear effects except the estimation function \( f_t(v) \). The state evolution equations of the input step (11) and (12) are then given by the expectations

\[
 c_x = E_x[x],
\]
\[
 c_{x,\hat{x}} = E_{x,u}[x f_t(v)],
\]
\[
 c_{\hat{x},\hat{x}} = E_{x,u}[f_t(v)^2],
\]
\[
 \theta^t = \beta \cdot E_{x,u}[f(v)].
\]

On the other hand, the state evolution equations of the output step are given by [3], [5]

\[
 \nu^{t,2} = E_{w,u,r}[g_t(-u, r)],
\]
\[
 \xi^t = E_{w,u,r}[g_t(-u, r)],
\]
\[
 \eta^t = E_{w,u,r}[\rho(r|w)g_t(-u, r)],
\]

where \( \rho(r|w) \) is the derivative of \( \rho(r|w) \) with respect to \( w \). It can be shown that the covariance matrix \( C_{w,u} \) from (49) is given by [5]

\[
 (w, u^t) \sim \mathcal{N}(0, C_{w,u}^t) \text{, with } C_{w,u}^t = \beta \left[ \begin{bmatrix} c_x & c_{x,\hat{x}} \\ c_{x,\hat{x}} & c_{\hat{x},\hat{x}} \end{bmatrix} \right],
\]

and thus we have the following joint density function

\[
 \rho_z(u, w) = \frac{e^{-\frac{(w - c_x u)^2}{2 c_{u}^2}}}{\sqrt{2 \pi c_{u}^2}} \cdot \frac{e^{-\frac{\xi^t}{2 c_{\hat{x}}^2}}}{\sqrt{2 \pi c_{\hat{x}}^2}}.
\]

With these definitions we get the recursive formulas (54) characterizing the dynamics of the algorithm and initialized as \( c_{\hat{x}} = c_{x,\hat{x}} = 0, \theta^t = \text{const}, \)

\[
 \nu^{t,2} = \int_{-\infty}^{\infty} g_t(-u, r)^2 du \quad \xi^t = \int_{-\infty}^{\infty} g_t(-u, r) \rho(r|w) dwdw
\]

\[
 \eta^t = \int_{-\infty}^{\infty} \rho(r|w) g_t(-u, r) dwdw
\]

and

\[
 \xi^t = \int_{-\infty}^{\infty} \rho(r|w) g_t(-u, r) dwdw
\]

while

\[
 \xi^t = \int_{-\infty}^{\infty} \rho(r|w) g_t(-u, r) dwdw
\]

\[
 \theta^t = \beta \cdot E_{x,u}[f(v)].
\]
where
\[
\beta_t(v) \frac{e^{v \beta_t \ell}}{\mathcal{E}_t} = 
\Phi \left( \frac{\beta_{t^*} v - \beta_{t^*} \ell}{\sqrt{\beta_{t^*} v + \beta_{t^*} \ell}} \right) - \Phi \left( \frac{\beta_{t^*} v - \beta_{t^*} \ell}{\sqrt{\beta_{t^*} v + \beta_{t^*} \ell}} \right)
\]

We note that for the special case of unquantized measurements, above state evolution parameters takes the values [4]
\[
\xi^t |_{\text{unquantized}} = \eta^t |_{\text{unquantized}} = 1,
\]
and
\[
\nu^{t,2} |_{\text{unquantized}} = \sigma_0^2 + \beta (c_x + c_{x,\\hat{x}} - 2n_{x,\\hat{x}}).\]

On the other hand, the state evolution equations of the input step can be obtained by evaluating the expectations given in (53) for the special choice of the function \( f_t(v) \) defined in (13) and the statistics of \( x \). Therefore, we consider the special case where we have that the elements of the vector \( x \) undergo the following statistics
\[
x_i \sim \begin{cases} 
N(0,1) & \text{w.p. } s \\
0 & \text{w.p. } 1-s,
\end{cases}
\]
with \( s \) being the fraction of non-zero elements, i.e., it represents the sparsity of the vector \( x \). Then, for the \( \ell_1 \) regularization function \( f_t(v) \) defined in (19), we obtain the following results at each iteration \( t \)
\[
c_x = E[x^2] = s
\]
\[
c_{x,\hat{x}} = E[x^{t,2}]
\]
\[
= 2s \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left( \frac{v^t v + 1}{\xi^t} \right)^2 \frac{1}{2\pi} e^{-\frac{x^2}{2}} dxdv + 2(1-s) \int_{-\infty}^{\infty} \left( \frac{v^t v + 1}{\xi^t} \right)^2 \frac{1}{2\pi} e^{-\frac{v^2}{2}} dv
\]
\[
= 2s \left[ \Phi \left( \frac{-1/\xi^t}{\sqrt{(b' + 1)/\xi^t}} \right) \right] + 2(1-s) \left[ \Phi \left( \frac{-1/\xi^t}{\sqrt{(b' + 1)/\xi^t}} \right) \right]
\]
\[
e^{-\frac{x^2}{2\pi \xi^t}} \sqrt{(b' + 1)/\xi^t}
\]
\[
= 2s \left[ \Phi \left( \frac{-1/\xi^t}{\sqrt{(b' + 1)/\xi^t}} \right) \right] + 2(1-s) \left[ \Phi \left( \frac{-1/\xi^t}{\sqrt{(b' + 1)/\xi^t}} \right) \right]
\]
\[
e^{-\frac{x^2}{2\pi \xi^t}} \sqrt{(b' + 1)/\xi^t}
\]
\[
\nu^{t,2} = \frac{1}{\xi^2}
\]
\[
\nu^{t,2} = \frac{1}{\xi^2}
\]
\[
E[\hat{f}_t(v)] = \frac{1}{\xi^2}
\]
\[
E[\hat{f}_t(v)] = \frac{1}{\xi^2}
\]

V. NUMERICAL RESULTS

Let us consider a simulation setup with a vector \( x \) of length \( M = 400 \) with a sparsity ratio of \( s \) which is the average fraction of non-zero entries, which are i.i.d. and generated according to the distribution
\[
x_i \sim \begin{cases} 
N(0,1) & \text{w.p. } s \\
0 & \text{w.p. } 1-s.
\end{cases}
\]

We apply the same scalar quantizers to the measurement vector \( y = A \cdot x + \eta \), where the measurement matrix \( A \) has i.i.d. Gaussian distributed elements with variance \( 1/N \). In the single bit case, we use an asymmetric 1-bit quantizer since taking just sign-measurements (symmetric quantizer), extensively treated in many previous works, would not deliver information about the norm of the vector, whereas the asymmetric 1-bit quantizer exhibits this feature. This aspect of optimizing the quantizer thresholds for general resolution can be done deterministically based on the state evolution analysis. For instance, Fig. 4 shows for different values of the sparsity \( s \) and different value of \( \beta \) the analytical MSE (70) normalized by \( c_x = s \) as function of the 1-bit threshold, assuming noiseless observations \( \sigma_0 = 0 \). We can observe from the Fig. 4 that choosing this threshold to correspond to the standard deviation of the noiseless output,
that is $\sqrt{3}s$, provides good estimation performance independently of the sparsity of the vector $x$. That is

$$\text{Threshold}_{1\text{-bit}} = \sqrt{\frac{\mathbb{E}[\|Ax\|^2]}{N}} = \sqrt{\beta \cdot s}. \quad (72)$$

To evaluate now the effect of quantization on the reconstruction performance by simulations, we first consider the noiseless case, i.e., $\sigma_0 = 0$. For resolutions higher than 1-bit, uniform mid-riser quantizers are considered, the step-size of which can be optimized based on the state evolution results. To evaluate the effect of quantization on the reconstruction performance, we first consider the noiseless case, i.e., $\sigma_0 = 0$. In Fig. 5, we compare the performance in terms of relative mean squared error (i.e. $\mathbb{E}[\|x - \tilde{x}\|^2/\|x\|^2]$) of the $\ell_1$ and $\ell_2$ regularized optimizations with noiseless ($\sigma_0 = 0$) 1-bit measurements and non-sparse Gaussian vector $x$ ($s = 1$) as function of the oversampling ratio $N/M$, while maintaining $M = 400$ constant. 100 Monte Carlo trials have been carried out to generate the curves. Interestingly, there is hardly no difference between both formulations in that case.

Next, Fig. 6 shows for the same scenario but for a sparsity ratio of $s = 0.1$ the estimation performance in terms of the relative mean square error (MSE) of the proposed algorithm for the $\ell_1$ and $\ell_2$ regularized optimization as function of the oversampling ratio $N/M$. As expected the $\ell_1$ outperforms in this case the $\ell_2$ regularization, but in both cases the MSE decrease with the order of $N^{-2}$ with respect to the number of measurements, which corresponds to the maximal possible decay predicted by the theory [9]. Furthermore, the analytical MSE curves found by the state evolution analysis of the iterative algorithm are also shown and they match closely to the experimental results.

We now concentrate on the $\ell_1$ formulation and perform again for the same setting ($s = 0.1$, $M = 400$, $\sigma_0 = 0$) 100 Monte Carlo trials for varying ratio $N/M$ and different bit resolutions (1-, 2-, 3- and 4-bit). The step sizes of the quantizers can be optimized, and the following rule-of-thumb is obtained\textsuperscript{2}

$$\Delta_{b=\text{bit}} = 4.8 \cdot 2^{-b} \cdot \sqrt{\frac{\mathbb{E}[\|Ax\|^2]}{N}}. \quad (73)$$

It can be observed from Fig. 7 that increasing the resolution by one bit improve the MSE by roughly factor 4 (i.e. 6dB) in terms of relative MSE. This leads to the conclusion that the MSE decreases as $2^{-2b}$, which is consistent to the fact that the step-size (in other words the uncertainty) decreases as $2^{-b}$.

Let us now consider noisy measurements, in combination with 1-bit quantization. Again, we fix $s = 0.1$ and $M = 400$. Fig. 8 and 9 present the performance of the $\ell_1$ based iterative algorithm in terms of the relative MSE for two different SNR= $\mathbb{E}[\|Ax\|^2]/N/\sigma_0^2$ values (i.e. the ratio of signal variance and noise variance at each output) are considered. For comparison the relative MSE in the ideal case with unquantized observations is also considered. The regularization parameter $\lambda$ in the function (17) was adapted to the noise level (see the legend of the figure) by trial to get the lowest MSE. First of all, we see that the MSE decreases now inverse proportionally to the number of measurements ($\propto N^{-1}$), which means that at high oversampling ratios the estimation performance becomes noise limited. Additionally, it turns out that the high SNR value (20dB) the performance loss due to quantization is significant, while for the low SNR value (0dB), the performance loss becomes much small, where roughly tripling the number of measurements leads to the same MSE. This suggests the use of low resolution quantizers in the medium or low SNR regimes, which is a result that has been also observed in the context multi-antenna communications [10]. For comparison the analytical curves found by the state evolution algorithm are also plotted, where again we see that the performance is predicted very well.

VI. CONCLUSION

We presented an low complexity algorithm for performing an $\ell_1$ or $\ell_2$ regularized estimation of a vector from quantized measurements assuming a random measurement matrix. As expected, the $\ell_1$ outperforms the $\ell_2$ formulation when the vector is sparse. Contrary to previous works no prior information, in terms of density function or amplitude of the unknown vector $x$ is need for the reconstruction. Simulations show that the iterative algorithm presents good reconstruction performance.

\textsuperscript{2}The prescalar 4.8 was obtained for $s = 0.1$ but for other sparsity levels, other values might provide better results.
while having significant low complexity. We note that the number of measurements needs to be higher than the vector dimension when the quantization is very coarse, e.g., 1-bit, which might be advantageous in some applications, since a 1-bit quantizer is a simple device and can operate at very high speed. Remarkably, when noise becomes significant, it turns out that the loss due to quantization compared to the ideal case become marginal, and the tradeoff between using finer quantization or increasing the number of observations becomes less effective. This suggests that the use of low resolution high speed sampling devices might reduce the number of measurement bits per estimated vector coefficient.

REFERENCES


