Received: 18 May 2020 | Accepted: 23 November 2020

DOI: 10.1002/pamm.202000015

# Quantized Compressed Sensing by Rectified Linear Units

**Hans Christian Jung**[1], **Johannes Maly**[1,*], **Lars Palzer**[2], and **Alexander Stollenwerk**[3]

[1] Lehrstuhl für Mathematik der Informationsverarbeitung, RWTH Aachen, Pontdriesch 10, DE-52062 Aachen
[2] Lehrstuhl für Nachrichtentechnik, TU München, Theresienstraße 90, DE-80333 München
[3] Institut für Mathematik, TU Berlin, Straße des 17. Juni 136, DE-10623 Berlin

This work is concerned with the problem of recovering high-dimensional signals, which belong to a convex set of low-complexity, from a small number of quantized measurements. We propose to estimate the signals via a convex program based on rectified linear units (ReLUs) for two different quantization schemes, namely one-bit and uniform multi-bit quantization. Assuming that the linear measurement process can be modelled by a sensing matrix with i.i.d. subgaussian rows, we obtain for both schemes near-optimal uniform reconstruction guarantees by adding well-designed noise to the linear measurements prior to the quantization step. In the one-bit case, we show that the program is robust against adversarial bit corruptions as well as additive noise on the linear measurements. Further, our analysis quantifies precisely how the rate-distortion relationship of the program changes depending on whether we seek reconstruction accuracies above or below the noise floor. The proofs rely on recent results by Dirksen and Mendelson on non-Gaussian hyperplane tessellations.



## 1 Introduction

We consider the problem of estimating high-dimensional signals $\mathbf{x} \in \mathbb{R}^n$ from quantized measurements $\mathbf{q} = Q(\mathbf{A}\mathbf{x}) \in \mathcal{A}^m$, where $\mathbf{A} \in \mathbb{R}^{m\times n}$ models the linear sensing process, $\mathcal{A} \subset \mathbb{R}$ is a discrete quantization alphabet and $Q : \mathbb{R}^m \to \mathcal{A}^m$ quantizes the sampled signal values. The quantization step is a crucial component of any real-world sensing scheme, where the measurements have to be quantized to a finite number of bits before the signal can be reconstructed.

In many scenarios of interest, the number of observations $m$ is much smaller than the ambient dimension $n$, which implies that the measurement system is highly underdetermined. Such types of recovery tasks have recently caught increasing attention in various research areas, for example in the field of 1-bit compressed sensing [1–3]. Here, only one bit per measurement is stored, that is, $\mathcal{A} = \{-1, 1\}$. Although appealing due to the fact that these quantization schemes are especially easy to implement in hardware and have low energy consumption, the coarse quantization drastically distorts the linear measurements and exact recovery of the signal clearly becomes impossible. Nevertheless, for suitable measurement matrices $\mathbf{A}$ and quantizers $Q$ it was shown that signals $\mathbf{x}$ can be accurately estimated via various tractable recovery programs from 1-bit observations of the above form. Moreover, the number of sufficient measurements can be drastically smaller than $n$ provided that $\mathbf{x}$ has some type of low-complexity structure that can be exploited during reconstruction [4–7]. These works show that the dependence of the number of observations on the complexity of $\mathbf{x}$, which is sufficient to achieve a good reconstruction, is comparable to the unquantized case. However, the developed theory suffers from several drawbacks: the guarantees are often non-uniform, the proven reconstruction error decays suboptimally in $m$, and most importantly, the results only permit Gaussian measurement matrices (or demand additional structural assumptions on $\mathbf{x}$) and are only partially or not at all robust against noise that corrupts the measurement process.

By adding well-designed noise to the linear measurements prior to the quantization step, the work [8] improved on all of these points showing reconstruction from 1-bit subgaussian measurements via a tractable recovery program [8, Theorem 1.7]. The recovery result is uniform in nature and the program is provably robust against adversarial bit corruptions as well as additive noise on the linear measurements. However, the result is not sensitive to the magnitude of the additive noise. This leads to suboptimal recovery guarantees in a low-noise setting.

## 2 Contribution

Our contribution is twofold: **(i)** We propose to estimate vectors $\mathbf{x} \in \mathcal{T} \subset \mathbb{R}^n$ from (corrupted) one-bit measurements $\mathbf{q}_{\text{corr}} \in \{-1, 1\}^m$ by solving

$$\min_{\mathbf{z}\in\mathbb{R}^n} \mathcal{L}_{\mathbf{q}_{\text{corr}}}(\mathbf{z}) \quad \text{subject to} \quad \mathbf{z} \in \mathcal{T}, \quad \text{where } \mathcal{L}_{\mathbf{q}_{\text{corr}}}(\mathbf{z}) := \frac{1}{m}\sum_{i=1}^{m} \left[-(q_{\text{corr}})_i(\langle \mathbf{a}_i, \mathbf{z}\rangle + \tau_i)\right]_+ . \tag{1}$$

Here, $[x]_+ = \max\{x, 0\}$ denotes the ReLU. The program is convex whenever the set $\mathcal{T}$ is a convex subset of $\mathbb{R}^n$. In Theorem 3.1, we provide robust reconstruction guarantees for (1). For reconstruction accuracies below the noise floor (high-noise regime), the guarantees for (1) match those given in [8, Theorem 1.7]. However, when the noise level is low and we

* Corresponding author: e-mail maly@mathc.rwth-aachen.de

seek reconstruction guarantees above the noise floor (low-noise regime), the guarantees for (1) are superior to those given in [8, Theorem 1.7] and in fact match the near-optimal guarantees established in [8, Theorem 1.5] for the non-tractable Hamming distance minimization program. Our results build upon tools developed in [8].

**(ii)** We extend the estimation scheme proposed in (1) to a memoryless uniform multi-bit quantization model with refinement level $\Delta > 0$ leading to a modified version of $\mathcal{L}$ and (1). The recovery guarantees we derive characterize both the high quantizer resolution and the low quantizer resolution regime and can be found in the full article [9].

Both in the one-bit and multi-bit setting our reconstruction guarantees are near-optimal (see the discussion on optimality below [9, Theorem 3] and [10, Section 5]).

## 3 One-Bit Result

We assume that the rows $\mathbf{a}_i^T$ of the sensing matrix $\mathbf{A}$ are independent copies of an isotropic, symmetric and subgaussian random vector $\mathbf{a} \in \mathbb{R}^n$. To formulate our results, define the Gaussian width of a set $\mathcal{T} \subset \mathbb{R}^n$ by $w_*(\mathcal{T}) := \mathbb{E}\sup_{\mathbf{x}\in\mathcal{T}} |\langle \mathbf{g}, \mathbf{x}\rangle|$, where $\mathbf{g}$ is an $n$-dimensional standard Gaussian random vector. Further, for $\varepsilon > 0$, we denote the $\varepsilon$-covering number of $\mathcal{T}$ by $\mathcal{N}(\mathcal{T}, \varepsilon)$. It is the smallest number of Euclidean balls with radius $\varepsilon$ needed to cover $\mathcal{T}$. Let us remark that for many interesting structured sets, there are tight upper bounds for both complexity measures, see for example [5, 6].

We are interested in recovering signals $\mathbf{x} \in \mathbb{R}^n$ from possibly (adversarially) corrupted one-bit measurements $\mathbf{q}_{\text{corr}} \in \{-1, 1\}^m$ which satisfy $d_H(\mathbf{q}_{\text{corr}}, q(\mathbf{x})) \le \beta m$, for a parameter $\beta \in (0, 1)$. Here, $d_H(\mathbf{q}_{\text{corr}}, q(\mathbf{x})) := \sum_{i=1}^m 1_{(\mathbf{q}_{\text{corr}})_i \neq (q(\mathbf{x}))_i}$ denotes the Hamming distance between $\mathbf{q}_{\text{corr}}$ and the vector of 1-bit measurements of $\mathbf{x}$ given by $q(\mathbf{x}) = \operatorname{sign}(\mathbf{A}\mathbf{x} + \boldsymbol{\tau} + \boldsymbol{\nu}) \in \{-1, 1\}^m$. Importantly, we add a vector of random thresholds $\boldsymbol{\tau}$ to the linear measurements $\mathbf{A}\mathbf{x}$ prior to the entry-wise quantization. The origin of this technique, called dithering, goes back to the work [11] where it was used to remove artefacts from quantized pictures. The vector $\boldsymbol{\nu}$ accounts for additive noise, which may disturb the linear sensing process. We assume that $\mathbf{A}, \boldsymbol{\tau}, \boldsymbol{\nu}$ are independent. In contrast to $\mathbf{A}$ and $\boldsymbol{\tau}$, the noise vector $\boldsymbol{\nu}$ is not known to us.

The following theorem is our main result in the one-bit case.

**Theorem 3.1** *For a parameter $\lambda > 0$, let $\boldsymbol{\tau}$ be uniformly distributed in $[-\lambda, \lambda]^m$. Assume that the entries of $\boldsymbol{\nu} \in \mathbb{R}^m$ are independent copies of a mean-zero subgaussian random variable $\nu$. There are constants $c, c_0, c_1, c_2, c_3 > 0$ and $C \ge e$ (only depending on the subgaussian norms of $\mathbf{a}$ and $\nu$) such that the following holds. Let $\mathcal{T} \subset RB_2^n$ denote a convex set. Fix the reconstruction accuracy $\rho \in (0, R]$.*

(i) *Low-noise regime: if $\|\nu\|_{L^2} \le c_0\rho/\sqrt{\log(C\lambda/\rho)}$, choose $\lambda \gtrsim R$ and assume that the number of measurements satisfies*

$$m \gtrsim \frac{\lambda}{\rho}\left(\frac{w_*^2((\mathcal{T}-\mathcal{T})\cap \rho B_2^n)}{\rho^2} + \log \mathcal{N}(\mathcal{T}, \tfrac{c\rho}{\sqrt{\log(C\lambda/\rho)}})\right).$$

*With probability exceeding $1 - 2\exp(-c_3 m\rho/\lambda)$, the following holds: for all $\mathbf{x} \in \mathcal{T}$ and all $\mathbf{q}_{corr} \in \{-1, 1\}^m$ which satisfy $d_H(\mathbf{q}_{corr}, q(\mathbf{x})) \le \beta m$ for $\beta \log(e/\beta) \le c_2\rho/\lambda$, every minimizer $\mathbf{x}^\#$ of the program (1) satisfies $\left\|\mathbf{x} - \mathbf{x}^\#\right\|_2 \le \rho$.*

(ii) *High-noise regime: if $\|\nu\|_{L^2} \ge c_0\rho/\sqrt{\log(C\lambda/\rho)}$, choose $\lambda \gtrsim (R + \|\nu\|_{L^2})\sqrt{\log(\lambda/\rho)}$ and assume*

$$m \gtrsim \frac{\lambda^2}{\rho^2}\left(\frac{w_*^2((\mathcal{T}-\mathcal{T})\cap \rho B_2^n)}{\rho^2} + \log \mathcal{N}(\mathcal{T}, \tfrac{c\rho}{\log(C\lambda/\rho)}\right).$$

*With probability exceeding $1 - 2\exp(-c_3 m(\rho/\lambda)^2)$, the following holds: for all $\mathbf{x} \in \mathcal{T}$ and all $\mathbf{q}_{corr} \in \{-1, 1\}^m$ which satisfy $d_H(\mathbf{q}_{corr}, q(\mathbf{x})) \le \beta m$ for $\beta \log(e/\beta) \le c_2\rho/\lambda$, every minimizer $\mathbf{x}^\#$ of the program (1) satisfies $\left\|\mathbf{x} - \mathbf{x}^\#\right\|_2 \le \rho$.*

For the proof of Theorem 3.1, corresponding guarantees in the multi-bit setting, numerical comparison with state-of-the-art methods, and extensive discussion of the results, see [9].

**Acknowledgements** Open access funding enabled and organized by Projekt DEAL.

## References

[1] P. T. Boufounos, Greedy sparse signal reconstruction from sign measurements, in: Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers, (IEEE, 2009), pp. 1305–1309.
[2] P. T. Boufounos and R. G. Baraniuk, 1-bit compressive sensing, in: 42nd Annual Conference on Information Sciences and Systems, (IEEE, 2008), pp. 16–21.
[3] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk, IEEE Transactions on Information Theory **59**(4), 2082–2102 (2013).
[4] A. Ai, A. Lapanowski, Y. Plan, and R. Vershynin, Linear Algebra and its Applications **441**, 222–239 (2014).
[5] Y. Plan and R. Vershynin, Communications on Pure and Applied Mathematics **66**(8), 1275–1297 (2013).

 

[6] Y. Plan and R. Vershynin, IEEE Transactions on Information Theory **59**(1), 482–494 (2013).
[7] Y. Plan and R. Vershynin, IEEE Transactions on Information Theory **62**(3), 1528–1537 (2016).
[8] S. Dirksen and S. Mendelson, arXiv preprint arXiv:1805.09409 (2018).
[9] H. C. Jung, J. Maly, L. Palzer, and A. Stollenwerk, arXiv preprint arXiv:1911.07816 (2019).
[10] S. Dirksen and S. Mendelson, arXiv preprint arXiv:1812.06719 (2018).
[11] L. Roberts, IRE Transactions on Information Theory **8**(2), 145–154 (1962).

 