Appl. Comput. Harmon. Anal. 37 (2014) 325–346

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Robust dequantized compressive sensing

Ji Liu *, Stephen J. Wright

Department of Computer Sciences, University of Wisconsin–Madison, 1210 W. Dayton St., Madison, WI 53706-1685, United States

ARTICLE INFO

Article history:
Received 23 May 2012
Received in revised form 22 July 2013
Accepted 29 December 2013
Available online 6 January 2014
Communicated by Richard Baraniuk

Keywords:
Compressive sensing
Signal reconstruction
Quantization
Optimization

ABSTRACT

We consider the reconstruction problem in compressed sensing in which the observations are recorded in a finite number of bits. They may thus contain quantization errors (from being rounded to the nearest representable value) and saturation errors (from being outside the range of representable values). Our formulation has an objective of weighted ℓ_2–ℓ_1 type, along with constraints that account explicitly for quantization and saturation errors, and is solved with an augmented Lagrangian method. We prove a consistency result for the recovered solution, stronger than those that have appeared to date in the literature, showing in particular that asymptotic consistency can be obtained without oversampling. We present extensive computational comparisons with formulations proposed previously, and variants thereof.

1. Introduction

This paper considers a compressive sensing (CS) system in which the measurements are represented by a finite number of bits, which we denote by B. By defining a quantization interval $\varDelta > 0$, and setting $G := 2^{B-1}\varDelta$, we obtain the following values for representable measurements:

$$-G + \frac{\varDelta}{2}, -G + \frac{3\varDelta}{2}, \ldots, -\frac{\varDelta}{2}, \frac{\varDelta}{2}, \ldots, G - \frac{\varDelta}{2}. \tag{1}$$

We assume in our model that actual measurements are recorded by rounding to the nearest value in this set. The recorded observations thus contain (a) quantization errors, resulting from rounding of the true observation to the nearest represented number, and (b) saturation errors, when the true observation lies beyond the range of represented values, namely, $[-G + \frac{\varDelta}{2}, G - \frac{\varDelta}{2}]$. This setup is seen in some compressive sensing hardware architectures (see, for example, [15,20,19,21,9]).

Given a sensing matrix $\varPhi \in \mathbb{R}^{M \times N}$ and the unknown vector x, the true observations (without noise) would be $\varPhi x$. We denote the recorded observations by the vector $y \in \mathbb{R}^M$, whose components take on the values in (1). We partition $\varPhi$ into the following three submatrices:

* Corresponding author.
E-mail addresses: ji-liu@cs.wisc.edu (J. Liu), swright@cs.wisc.edu (S.J. Wright).

1063-5203/$ – see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.acha.2013.12.006

- The saturation parts $\bar{\Phi}_-$ and $\bar{\Phi}_+$, which correspond to those recorded measurements that are represented by $-G+\Delta/2$ or $G-\Delta/2$, respectively — the two extreme values in (1). We denote the number of rows in these two matrices combined by $\bar{M}$.
- The unsaturated part $\tilde{\Phi} \in \mathbb{R}^{\tilde{M}\times N}$, which corresponds to the measurements that are rounded to non-extreme representable values.

In some existing analyses [5,13], the quantization errors are treated as a random variables following an i.i.d. uniform distribution in the range $[-\frac{\Delta}{2}, \frac{\Delta}{2}]$. This assumption makes sense in many situations (for example, image processing, audio/video processing), particularly when the quantization interval Δ is tiny. However, the assumption of a uniform distribution may not be appropriate when Δ is large, or when an inappropriate choice of saturation level G is made. In this paper, we assume a slightly weaker condition, namely, that the quantization errors for non-saturated measurements are independent random variables with zero expectation. (These random variables are of course bounded uniformly by $\Delta/2$.)

The state-of-the-art formulation to this problem (see [14]) is to combine the basis pursuit model with saturation constraints, as follows:

$$\min_x \|x\|_1 \tag{2a}$$

$$\text{s.t.} \quad \|\tilde{\Phi}x - \tilde{y}\|^2 \leqslant \epsilon^2\Delta^2 \qquad (\ell_2) \tag{2b}$$

$$\bar{\Phi}_+ x \geqslant (G-\Delta)\mathbf{1} \qquad (+\text{ saturation}) \tag{2c}$$

$$\bar{\Phi}_- x \leqslant (\Delta-G)\mathbf{1}, \qquad (-\text{ saturation}) \tag{2d}$$

where $\mathbf{1}$ is a column vector with all entries equal to 1 and $\tilde{y}$ is the quantized subvector of the observation vector y that corresponds to the unsaturated measurements. We refer to this model as "**L2**" in later discussions. It has been shown that the estimation error arising from the formulation (2) is bounded by $O(\epsilon\Delta)$ in the ℓ_2 norm sense (see [14,6,13]).

The paper proposes a robust model that replaces (2b) with a least-square loss term in the objective and adds an ℓ_∞ constraint:

$$\min_x \frac{1}{2}\|\tilde{\Phi}x - \tilde{y}\|^2 + \lambda\Delta\|x\|_1 \tag{3a}$$

$$\text{s.t.} \quad \|\tilde{\Phi}x - \tilde{y}\|_\infty \leqslant \Delta/2 \qquad (\ell_\infty) \tag{3b}$$

$$\bar{\Phi}_+ x \geqslant (G-\Delta)\mathbf{1} \qquad (+\text{ saturation}) \tag{3c}$$

$$\bar{\Phi}_- x \leqslant (\Delta-G)\mathbf{1}. \qquad (-\text{ saturation}) \tag{3d}$$

We refer to this model as **LASSO**∞ in later discussions. The ℓ_∞ constraint (3b) arises from the fact that (unsaturated) quantization errors are bounded by $\Delta/2$. This constraint may reduce the feasible region for the recovery problem while retaining feasibility of the true solution x^*, thus promoting more robust signal recovery. From the viewpoint of optimization, the constraint (2b) plays the same role as the least-square loss term in the objective (3a), when the values of ϵ and λ are related appropriately. However, it will become clear from our analysis that inclusion of this term in the objective rather than applying the constraint (2b) can lead a tighter bound on the reconstruction error.

The analysis in this paper shows that when Φ is a Gaussian ensemble, and provided that $S\log N = o(M)$ and several mild conditions hold, the estimation error of for the solution of (3) is bounded by

$$\min\{O(\sqrt{S(\log N)/M}), O(1)\}\Delta,$$