Robust Dequantized Compressive Sensing

Ji Liu Stephen J. Wright
Department of Computer Sciences, University of Wisconsin−Madison
{*ji-liu, swright*}*@cs.wisc.edu*

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.

Keywords: compressive sensing, signal reconstruction, quantization, optimization.

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 $\Delta > 0$, and setting $G := 2^{B-1}\Delta$, we obtain the following values for representable measurements:

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

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, $[-G + \frac{\Delta}{2}, G - \frac{\Delta}{2}]$. This setup is seen in some compressive sensing hardware architectures (see, for example, Laska et al., 2007; Tropp et al., 2009; Romberg, 2009; Tropp et al., 2006; Duarte et al., 2008).

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

- 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 (Candès et al., 2006; Jacques et al., 2011), 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 variable with zero expectation.

The state-of-the-art formulation to this problem (see Laska et al., 2011) is to combine the basis pursuit model with saturation constraints, as follows:

$$\min_x \ \|x\|_1 \tag{2a}$$
$$\text{s.t.} \quad \|\tilde{\Phi}x - \tilde{y}\|^2 \le \epsilon^2\Delta^2 \qquad (\ell_2) \tag{2b}$$
$$\bar{\Phi}_+ x \ge (G-\Delta)\mathbf{1} \qquad (+\text{ saturation}) \tag{2c}$$
$$\bar{\Phi}_- x \le (\Delta-G)\mathbf{1}, \qquad (-\text{ saturation}) \tag{2d}$$

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. 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 Laska et al., 2011; Candès, 2008; Jacques et al., 2011).

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

$$\min_x \ \frac{1}{2}\|\tilde{\Phi}x - \tilde{y}\|^2 + \lambda\Delta\|x\|_1 \tag{3a}$$
$$\text{s.t.} \quad \|\tilde{\Phi}x - \tilde{y}\|_\infty \le \Delta/2 \qquad (\ell_\infty) \tag{3b}$$
$$\bar{\Phi}_+ x \ge (G-\Delta)\mathbf{1} \qquad (+\text{ saturation}) \tag{3c}$$
$$\bar{\Phi}_- x \le (\Delta-G)\mathbf{1}. \qquad (-\text{ saturation}) \tag{3d}$$

The ℓ_∞ constraint arises immediately from the observation that the quantization error is 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 the least-squares loss can lead a tighter bound than using (2b).

The analysis in this paper shows that solutions obtained from this formulation are, in the worst case, better than the state-of-the-art model (2), and also better than the model in which only the ℓ_∞ constraint (3b) are applied (in place of the ℓ_2 constraint (2b)), as suggested by Jacques et al. (2011). More important, when the number $\tilde{M}$ of unsaturated measurements goes to infinity faster than $S\log(N)$ (where S is the sparsity number—the number of nonzero components in x^*), the estimation error for the solution of (3) vanishes with high probability. By comparison, the state-of-the-art model (2) does not guarantee improvement given more measurements. Although Jacques et al. (2011) show that the estimation error can be eliminated only using the ℓ_∞ constraint, their "oversampling" condition is much stronger than our requirement.

We apply the alternating direction method of multipliers (ADMM) (see Eckstein and Bertsekas, 1992; Boyd et al., 2011) to solve (3). The computational results reported in Section 4 compare the solution properties for (3) to those for (2) and other formulations in which some of the constraints are omitted.

1.1 Related Work

There have been several recent works on CS with quantization and saturation. Laska et al. (2011) propose the formulation (2). Jacques et al. (2011) replace the ℓ_2 constraint (2b) by an ℓ_p constraint ($2 \le p \le \infty$) to handle the oversampling case, and show that values p greater than 2 are advantageous. The model of Zymnis et al. (2010) allows Gaussian noise in the measurements before quantization, and solves the resulting formulation with an ℓ_1-regularized maximum likelihood formulation. The average distortion introduced by scalar, vector, and entropy coded quantization of CS is studied by Dai et al. (2011).

The extreme case of 1-bit CS (in which only the sign of the observation is recorded) has been studied by Gupta et al. (2010) and Boufounos and Baraniuk (2008). In the latter paper, the ℓ_1 norm objective is minimized on the unit ball, with a sign consistency constraint. The former paper proposes two algorithms that need no more than $O(S\log N)$ measurements to recover the unknown support of the true signal, but they cannot recover the magnitudes of the nonzeros reliably.

1.2 Notation

We use $\|\cdot\|_p$ to denote the ℓ_p norm, where $1 \leq p \leq \infty$, with $\|\cdot\|$ denoting the ℓ_2 norm. We use x^* for the true signal, $\hat{x}$ as the estimated signal (the solution of (3), and $h = \hat{x} - x^*$ as the difference between them. As mentioned above, S denotes the number of nonzero elements of x^*.

For any $z \in \mathbb{R}^N$, we use z_i to denote the ith component and z_T to denote the subvector corresponding to index set $T \subset \{1, 2, ..., N\}$. Similarly, we use $\tilde{\Phi}_T$ to denote the column submatrix of $\tilde{\Phi}$ consisting of the columns indexed by T. The cardinality of T is denoted by $|T|$.

In discussing the dimensions of the problem and how they are related to each other in the limit (as N and $\tilde{M}$ both approach ∞), we make use of order notation. If α and β are both positive quantities that depend on the dimensions, we write $\alpha = O(\beta)$ if α can be bounded by a fixed multiple of β for all sufficiently large dimensions. We write $\alpha = o(\beta)$ if for *any* positive constant $\phi > 0$, we have $\alpha \leq \phi\beta$ for all sufficiently large dimensions. We write $\alpha = \Omega(\beta)$ if both $\alpha = O(\beta)$ and $\beta = O(\alpha)$.

The projection onto the ℓ_∞ norm ball with the radius λ is

$$\mathcal{P}_\infty(x, \lambda) := \text{sign}(x) \odot \min(|x|, \lambda)$$

where $\odot$ denotes componentwise multiplication and sign(x) is the sign vector of x. (The ith entry of sign(x) is 1, -1, or 0 depending on whether x_i is positive, negative, or zero, respectively.)

The indicator function $\mathbb{I}_\Pi(\cdot)$ for a set Π is defined to be 0 on Π and ∞ otherwise.

Let

$$\bar{\Phi} = \begin{bmatrix} -\bar{\Phi}_- \\ \bar{\Phi}_+ \end{bmatrix} \text{ and } \Phi = \begin{bmatrix} \tilde{\Phi} \\ \bar{\Phi} \end{bmatrix}. \tag{4}$$

The maximum column norm in $\tilde{\Phi}$ is denoted by $f_{\max}$, that is,

$$f_{\max} := \max_{i \in \{1,2,...,N\}} \|\tilde{\Phi}_{\cdot i}\|. \tag{5}$$

We define the following quantities associated with a matrix Ψ with N columns:

$$\rho_p^-(k, \Psi) := \min_{|T| \leq k, h \in \mathbb{R}^N} \frac{\|\Psi_T h_T\|_p^2}{\|h_T\|^2} \tag{6}$$

and

$$\rho_p^+(k, \Psi) := \max_{|T| \leq k, h \in \mathbb{R}^N} \frac{\|\Psi_T h_T\|_p^2}{\|h_T\|^2}. \tag{7}$$

Sometimes we use following notations for short:

$$\begin{aligned} \rho_p^-(k) &:= \rho_p^-(k, \Phi), \quad \rho_p^+(k) := \rho_p^+(k, \Phi), \\ \tilde{\rho}_p^-(k) &:= \rho_p^-(k, \tilde{\Phi}), \quad \tilde{\rho}_p^+(k) := \rho_p^+(k, \tilde{\Phi}), \\ \bar{\rho}_p^-(k) &:= \rho_p^-(k, \bar{\Phi}), \quad \bar{\rho}_p^+(k) := \rho_p^+(k, \bar{\Phi}). \end{aligned}$$

Note that the ℓ_2 norm is used in each denominator, regardless of p.

Finally, we denote $\max\{z, 0\}$ as $(z)_+$ for short.

1.3 Organization

The ADMM optimization framework for solving (3) is discussed in Section 2. Section 3 analyzes the properties of the solution of (3) in the worst case and compares with existing results. The numerical simulation is reported in Section 4 and some conclusions are offered in Section 5. Proofs appear in the appendix.

2 Algorithm

This section describes the ADMM algorithm for solving (3). For simpler notation, we combine the saturation constraints as follows:

$$\begin{bmatrix} -\bar{\Phi}_- \\ \bar{\Phi}_+ \end{bmatrix} x \geq \begin{bmatrix} (G-\Delta)\mathbf{1} \\ (G-\Delta)\mathbf{1} \end{bmatrix} \quad \Leftrightarrow \quad \bar{\Phi}x \geq \bar{y},$$

where $\bar{\Phi}$ is defined in (4) and $\bar{y}$ is defined in an obvious way. To specify ADMM, we introduce some auxiliary variables and write (3) as follows.

$$\begin{aligned} \min_x \ & \frac{1}{2}\|\tilde{\Phi}x - \tilde{y}\|^2 + \lambda\|x\|_1 \\ \text{s.t.} \quad & u = \tilde{\Phi}x - \tilde{y} \\ & v = \bar{\Phi}x - \bar{y} \\ & \|u\|_\infty \leq \Delta/2 \\ & v \geq \mathbf{0}. \end{aligned} \tag{8}$$

Introducing Lagrange multipliers α and β for the two equality constraints in (8), we write the augmented Lagrangian for this formulation, with prox parameter $\theta > 0$ as follows:

$$\begin{aligned} & L_A(x,u,v,z,w,\alpha,\beta,\eta,\gamma) \\ =& \frac{1}{2}\|\tilde{\Phi}x - \tilde{y}\|^2 + \lambda\|x\|_1 + \langle \alpha, u - \tilde{\Phi}x + \tilde{y}\rangle + \langle \beta, v - \bar{\Phi}x + \bar{y}\rangle + \\ & \frac{\theta}{2}\|u - \tilde{\Phi}x + \tilde{y}\|^2 + \frac{\theta}{2}\|v - \bar{\Phi}x + \bar{y}\|^2 + \mathbb{I}_{\|u\|_\infty \leq \Delta/2}(u) + \mathbb{I}_{v\geq 0}(v) \end{aligned} \tag{9}$$

At each iteration of ADMM, we optimize this function with respect to the primal variables u and v in turn, then update the dual variables α and β in a manner similar to gradient descent. The parameter θ may then be increased before proceeding to the next iteration.

We summarize the ADMM algorithm in **Algorithm 1**.

Algorithm 1 ADMM for (8)

Require: $\tilde{\Phi}$, $\tilde{y}$, $\bar{\Phi}$, $\bar{y}$, Δ, K, and x;
1: Initialize $\theta > 0$, $\alpha = 0$, $\beta = 0$, $u = \tilde{\Phi}x - \tilde{y}$, and $v = \bar{\Phi}x - \bar{y}$;
2: **for** $k = 0 : K$ **do**
3: $u \leftarrow \arg\min_v : L_A(x,u,v,\alpha,\beta,) = \mathcal{P}_\infty(\tilde{\Phi}x - \tilde{y} - \alpha/\theta, \Delta/2)$;
4: $w \leftarrow \arg\min_w : L_A(x,u,v,\alpha,\beta) = \max(\bar{\Phi}x - \bar{y} - \beta/\theta, 0)$;
5: $x \leftarrow \arg\min_x : L_A(x,u,v,\alpha,\beta)$;
6: $\alpha \leftarrow \alpha + \theta(u - \tilde{\Phi}x + \tilde{y})$;
7: $\beta \leftarrow \beta + \theta(v - \bar{\Phi}x + \bar{y})$;
8: Possibly increase θ;
9: **if** stopping criteria is satisfied **then**
10: break
11: **end if**
12: **end for**

The updates in Steps 3 and 4 have closed-form solutions, as shown. The function to be minimized in Step 5 consists of an $\|x\|_1$ term in conjunction with a quadratic term in x. Many algorithms can be applied to solve this problem, e.g., the SpaRSA algorithm proposed by Wright et al. (2009), the accelerated first order method (Nesterov, 2007), and the FISTA algorithm (Beck and Teboulle, 2009). The update strategy for θ in Step 7 is flexible. We use the following simple and useful scheme from He et al. (2000) and Boyd et al. (2011):

$$\theta := \begin{cases} \theta\tau & \text{if } \|r\| > \mu\|d\| \\ \theta/\tau & \text{if } \|r\| < \mu\|d\| \\ \theta & \text{otherwise,} \end{cases} \tag{10}$$

where r and d denote the primal and dual residual errors respectively, specifically,

$$r = \begin{bmatrix} u - \tilde{\Phi}x + \tilde{y} \\ v - \bar{\Phi}v + \bar{y} \end{bmatrix} \text{ and } d = \theta \begin{bmatrix} \tilde{\Phi}(x - x_{\text{last}}) \\ \bar{\Phi}(x - x_{\text{last}}) \end{bmatrix},$$

where x_{last} denotes the previous value of x. The parameters μ and τ should be greater than 1; we used $\mu = 10$ and $\tau = 2$.

3 Analysis

The section analyzes the properties of the solution obtained from our formulation (3). In Subsection 3.1, we obtain bounds on the norm of the difference h between the estimator $\hat{x}$ given by (3) and the true signal x^*. Our bounds require the true solution x^* to be feasible for the formulation (3); we derive conditions that guarantee that this condition holds, with a specified probability. In Subsection 3.2, we estimate the constants that appear in our bounds under certain assumptions, including an assumption that the full sensing matrix Φ is Gaussian.

We formalize our assumption about quantization errors as follows.

Assumption 1. *The quantization errors $(\tilde{\Phi}x^* - \tilde{y})_i$, $i = 1, 2, \ldots, \tilde{M}$ are indepently distributed with expectation 0.*

3.1 Estimation Error Bounds

The following error estimate is our main theorem, proved in the appendix.

Theorem 1. *Assume that the true signal x^* satisfies*

$$\|\tilde{\Phi}^T(\tilde{\Phi}x^* - \tilde{y})\|_\infty \leq \lambda\Delta/2. \tag{11}$$

Let s and l be positive integers in the range $1, 2, \ldots, N$, and define

$$B_L := 6\sqrt{s}\lambda\Delta, \tag{12a}$$

$$B_\infty := \sqrt{\tilde{M}}\Delta, \tag{12b}$$

$$A_0(\Psi) := \rho_2^-(s + l, \Psi) - 3\sqrt{s/l}[\rho_2^+(s + 2l, \Psi) - \rho_2^-(s + 2l, \Psi)] \tag{12c}$$

$$A_1(\Psi) := 4[\rho_2^+(s + 2l, \Psi) - \rho_2^-(s + 2l, \Psi)], \tag{12d}$$

$$C_0(\Psi) := \frac{\sqrt{1 + 9s/l}}{A_0(\Psi)}, \tag{12e}$$

$$C_1(\Psi) := 4 + \frac{\sqrt{1 + 9s/l}A_1(\Psi)}{A_0(\Psi)}, \tag{12f}$$

$$C_2(\Psi) := \sqrt{\frac{1 + 9s/l}{A_0(\Psi)}}. \tag{12g}$$

We have that for any $T_0 \subset \{1, 2, ..., N\}$ with $s = |T_0|$, if $A_0(\tilde{\Phi}) > 0$, then

$$\|h\| \leq C_0(\tilde{\Phi})B_L + C_1(\tilde{\Phi})/\sqrt{l}\|x^*_{T_0^c}\|_1 + C_2(\tilde{\Phi})B_L^{1/2}/s^{1/4}\|x^*_{T_0^c}\|_1^{1/2}, \tag{13a}$$

$$\|h\| \leq C_2(\tilde{\Phi})B_\infty + C_1(\tilde{\Phi})/\sqrt{l}\|x^*_{T_0^c}\|_1. \tag{13b}$$

Suppose that Assumption 1 holds, and let $\pi \in (0, 1)$ be given. If we define $\lambda = \sqrt{2\log 2N/\pi}f_{\max}$ in (3)*, then with probability at least $P = 1 - \pi$, the inequalities* (13a) *and* (13b) *hold.*

From the proof in the appendix, one can see that the estimation error bound (13a) is mainly determined by the least square term in the objective (3a), whereas the estimation error bound (13b) arises from the L_∞ constraint (3b).

If we take T_0 as the support set of x^*, only the first term in (13a) and (13b) remains. It follows that $\|h\| \leq \min\{C_0(\tilde{\Phi})B_L, C_2(\tilde{\Phi})B_\infty\}$. We will compare their values in next subsection.

3.2 Estimating the Constants

Here we discuss the behaviors of the least square term and the ℓ_∞ constraints by comparing the values of $C_0(\tilde{\Phi})B_L$ and $C_2(\tilde{\Phi})B_\infty$. In order to simplify the comparison, we make the following assumptions.

(i) Φ is a Gaussian random matrix, that is, each entry is i.i.d., drawn from a standard Gaussian distribution $\mathcal{N}(0,1)$.

(ii) the confidence level $P = 1 - \pi$ is fixed.

(iii) s and l are both equal to the sparsity number S.

(iv) $S\log(N) = o(M)$.

(v) the saturation ratio $\chi := \bar{M}/M$ is smaller than a small positive threshold that is defined in Theorem 2.

(vi) T_0 is taken as the support set of x^*, so that $x^*_{T_0^c} = 0$.

Note that (iii) and (iv) together imply that $s = l = S \ll M$, while (iv) implies that $\tilde{M} = \Omega(M)$.

The discussion following Theorem 2 in Appendix indicates that under these assumptions, the quantities defined in (12e), (12f), (12f), and (5) satisfy the following estimates:

$$C_0(\tilde{\Phi}) = \Omega(1/M), \quad C_1(\tilde{\Phi}) = \Omega(1), \quad C_2(\tilde{\Phi}) = \Omega(1/\sqrt{M}), \quad f_{\max} = \Omega(\sqrt{M}),$$

with high probability, for sufficiently high dimensions. Using the estimates in Theorem 2, we have from the definitions of B_L and B_∞ in Theorem 1, with the setting of λ from Theorem 1, that

$$C_0(\tilde{\Phi})B_L = O\left(\frac{\sqrt{S\log(N)}f_{\max}\Delta}{M}\right) = O\left(\sqrt{\frac{S\log(N)}{M}}\Delta\right) \to 0, \tag{14a}$$

$$C_2(\tilde{\Phi})B_\infty = O\left(\frac{\sqrt{\tilde{M}}\Delta}{\sqrt{M}}\right) = O\left(\Delta\right). \tag{14b}$$

Combining the estimation error bounds (13a) and (13b), we have

$$\|h\| \le \min\{C_0(\tilde{\Phi})B_L, C_2(\tilde{\Phi})B_\infty\} = \min\{O(\sqrt{S\log(N)/M}), O(1)\}\Delta. \tag{15}$$

In the regime described by assumption (iv), $C_0(\tilde{\Phi})B_L$ will be asymptotically smaller than $C_2(\tilde{\Phi})B_\infty$. The bound in (15) has size $O(\sqrt{S\log(N)/M}\Delta)$, consistent with the upper bound of the Dantzig selector (Candès and Tao, 2007) and the LASSO (Zhang, 2009)[1]. Recall that the estimation error of the state-of-the-art algorithm (2) for quantized compressed sensing is $O(\|\tilde{\Phi}x^* - \tilde{y}\|/\sqrt{\tilde{M}})$ (Jacques et al., 2011; Laska et al., 2011). Since $\|\tilde{\Phi}x^* - \tilde{y}\| = O(\sqrt{\tilde{M}}\Delta)$ (Jacques et al., 2011), this estimate is consistent with the error that would be obtained if we imposed only the ℓ_∞ constraint (3b) in our formulation. Under the assumption (iv), the estimation error for (3) will vanish as the dimensions grow, with probability at least $1 - \pi$. By contrast, Jacques et al. (2011) do not account for saturation in their formulation and show that the estimation error converges to 0 using the ℓ_∞ constraint only when oversampling happens (in particular, $M \ge Ne^S/S$) and $M \to \infty$. Our formulation yields a much weaker condition on M than oversampling. For example, $M = S\log^2(N)$ would produce consistency in our formulation, but not in (2).

4 Simulations

This section compares results for five variant formulations. The first one is our formulation (3), which we refer to in the plots as **LASSO**∞. We also tried a variant in which the ℓ_∞ constraint (3b) was omitted from (3). The recovery

[1] Their bound is $O(\sqrt{S\log(N)/M}\sigma)$ where σ is the standard deviation of the observation noise which, in the classical setting for the Dantzig selector and the LASSO, is assumed to follow a Gaussian distribution.

performance for this variant was uniformly worse than for **LASSO**∞ , so we do not show it in our figures. (It is, however, sometimes better than the formulations described below, and uniformly better than **Dantzig** .)

The remaining four alternatives are based on the following model, in which the ℓ_2 norm of the residual appears in a constraint (rather than the objective) and a constraint of Dantzig type also appears:

$$\min_x \ \|x\|_1 \tag{16a}$$

$$\text{s.t.} \quad \|\tilde{\Phi}x - \tilde{y}\|^2 \le \epsilon^2\Delta^2 \qquad (\ell_2) \tag{16b}$$

$$\|\tilde{\Phi}x - \tilde{y}\|_\infty \le \Delta/2 \qquad (\ell_\infty) \tag{16c}$$

$$\|\tilde{\Phi}^T(\tilde{\Phi}x - \tilde{y})\|_\infty \le \lambda\Delta/2 \qquad \text{(Dantzig)} \tag{16d}$$

$$\bar{\Phi}_+ x \ge (G - \Delta)\mathbf{1} \quad (+ \text{ saturation}) \tag{16e}$$

$$\bar{\Phi}_- x \le (\Delta - G)\mathbf{1}. \ (- \text{ saturation}) \tag{16f}$$

The four formulations are obtained by omitting certain constraints from this model.

- **L**∞ : an ℓ_∞ constraint model that enforces (16c), (16e), and (16f), but not (16b) or (16d) (that it, the model with $p = \infty$ in Jacques et al. (2011) but with the addition of saturation constraints)
- **L**2 : an ℓ_2 constraint model (that is, the state-of-the-art model in (2) (Laska et al., 2011)) that enforces (16b), (16e), and (16f) but not (16c) or (16d);
- **Dantzig** : the Dantzig constraint algorithm with saturation constraints, that is, it enforces (16d), (16e), and (16f) but not (16b) or (16c);
- **L**2**Dantzig**∞ : the full model defined by (16).

Note that we use the same value of λ in (16d) as in (3), since in both cases they lead to a constraint that the true signal x^* satisfies $\|\tilde{\Phi}^T(\tilde{\Phi}x^* - \tilde{y})\|_\infty \le \lambda\Delta/2$ with a certain probability; see (16d) and (11).

The synthetic data is generated as follows. The measurement matrix $\tilde{\Phi} \in \mathbb{R}^{M\times N}$ is a Gaussian matrix, each entry being independently generated from $\mathcal{N}(0, 1/R^2)$, for a given parameter R. The S nonzero elements of x^* are in random locations and are drawn from independently from $\mathcal{N}(0, 1)$. We use SNR $= -20\log_{10}(\|\hat{x} - x^*\|/\|x^*\|)$ as the error metric, where $\hat{x}$ is the signal recovered from each of the formulations under consideration. Given values of saturation parameter G and number of bits B, the interval Δ is defined accordingly as $\Delta = 2^{B-1}G$. All experiments are repeated for 30 times; we report the average performance.

We now describe how the bounds λ for (3a) and (16d) and ϵ for (16b) were chosen for these experiments. Essentially, ϵ and λ should be chosen so that the constraints (16b) and (16d) admit the true signal x^* with a a high (specified) probability. There is a tradeoff between tightness of the error estimate and confidence. Larger values of ϵ and λ can give a more confident estimate, since the defined feasible region includes x^* with a higher probability, while smaller values provide a tighter estimate. Although Lemma 2 suggests how to choose λ and Jacques et al. (2011) show how to determine ϵ, the analysis it not tight, especially when M and N are not particularly large. We use instead an approach based on simulation and on making the assumption (not used elsewhere in the analysis) that the non-saturated quantization errors $\xi_i = (\tilde{\Phi}x^* - \tilde{y})_i$ are i.i.d. uniform in $U_{[-\Delta/2,\Delta/2]}$. (As noted earlier, this stronger assumption makes sense in some settings, and has been used in previous analyses.) We proceed by generating numerous independent samples of $Z \sim U_{[-\Delta/2,\Delta/2]}$. Given a confidence level $1 - \pi$ (for $\pi > 0$), we set ϵ to the value for which $\mathbb{P}(Z \ge \epsilon\Delta) = \pi$ is satisfied empirically. A similar technique is used to determine λ. When we seek certainty ($\pi = 0$, or confidence $P = 100\%$), we set ϵ and λ according to the true solution x^*, that is, $\epsilon = \|\tilde{\Phi}x^* - \tilde{y}\|/\Delta$ and $\lambda = 2\|\tilde{\Phi}^T(\tilde{\Phi}x^* - \tilde{y})\|_\infty/\Delta$.

To summarize the various parameters that are varied in our experiments: M and N are dimensions of Φ, S is sparsity of solution x^*, G is saturation level, B is number of bits, R is the inverse standard deviation of the elements of Φ, and $P = 1 - \pi$ denotes the confidence levels, expressed as a percentage.

In Figure 1, we fix the values of M, S, G, R, and P, choose two values of B: 3 and 5. Plots show the average SNRs (over 30 trials) of the solutions $\hat{x}$ recovered from the five models against the dimension N. In this and all subsequent figures, the saturation ratio is defined to be $(M - \tilde{M})/M$, the fraction of extreme measurements. Our

LASSO∞ formulation and the full model **L**2**Dantzig**∞ give the best recovery performance for small N, while for larger N, **LASSO**∞ is roughly tied with the the **L**2 model. The **L**∞ and **Dantzig** models have poorer performance — a pattern that we continue to observe in subsequent tests.

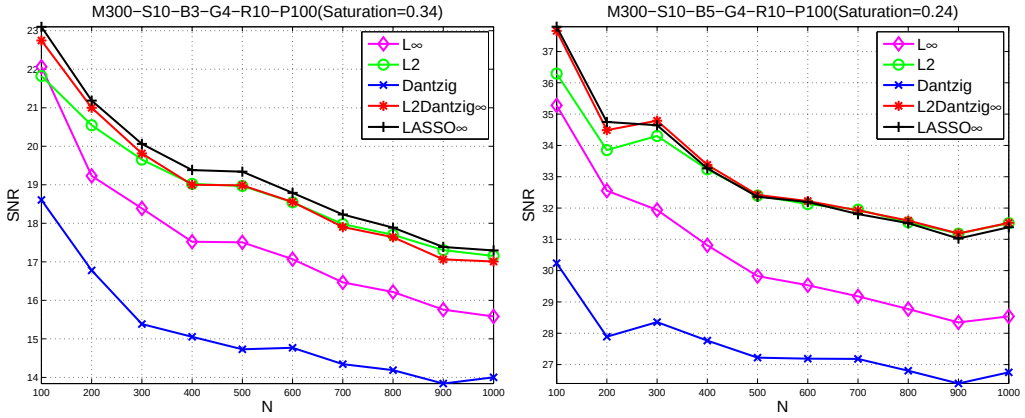

Figure 1: Comparison among various models for fixed values $M = 300$, $S = 10$, $G = 4$, $R = 10$, and $P = 100\%$, and two values of B (3 and 5, respectively). The graphs show dimension N (horizontal axis) against SNR (vertical axis) for values of N between 100 and 1000, averaged over 30 trials for each combination of parameters.

Figure 2 fixes N, M, B, G, R, and P, and plots SNR as a function of sparsity level S. For all models, the quality of reconstruction decreases rapidly with S. **LASSO**∞ and **L**2**Dantzig**∞ achieve the best results overall, but are roughly tied with the **L**2 model for all but the sparsest signals. The **L**∞ model is competitive for very sparse signals, while the **Dantzig** model lags in performance.

We now examine the effect of number of measurements M on SNR. Figure 3 fixes N, S, G, R, and P, and tries two values of B: 3 and 5, respectively. Figure 4 fixes $B = 4$, and allows N to increase with M in the fixed ratio $5/4$. These figures indicate that the **LASSO**∞ and **L**2**Dantzig**∞ models are again roughly tied with the **L**2 model when the number of measurements is limited. For larger M, our models have a slight advantage over the **L**2 and **L**∞ models, which is more evident when the quantization intervals are smaller (that is, $B = 4$). Another point to note from Figure 4 is that **L**∞ outperforms **L**2 when both M and N are much larger than the sparsity S.

In Figure 5 we examine the effect of the number of bits B on SNR, for fixed values of N, M, S, G, R, and P. The fidelity of the solution from all models increases linearly with B, with the **LASSO**∞ , **L**2**Dantzig**∞ , and **L**2 models being slightly better than the alternatives.

Next we examine the effect on SNR of the confidence level, for fixed values of N, M, B, G, and R. In Figure 6, we set $M = 300$ and plot results for two values of S: 5 and 15. In Figure 7, we use the same values of S, but set $M = 150$ instead. Note first that the confidence level does not affect the solution of the **L**∞ model, since this is a deterministic model, so the reconstruction errors are constant for this model. For the other models, we generally see degradation as confidence is higher, since the constraints (16b) and (16d) are looser, so the feasible point that minimizes the objective $\|\cdot\|_1$ is further from the optimum x^*. Again, we see a clear advantage for **LASSO**∞ when the sparsity is low, M is larger, and the confidence level P is high. For less sparse solutions, the **L**2 and **L**2**Dantzig**∞ models have similar or better performance.

In Figure 8 we examine the effect of saturation bound G on SNR. We fix N, M, B, R, and P, and try two values of S: 5 and 10. A tradeoff is evident — the reconstruction performances are not monotonic with G. As G increases, the proportion of saturated measurements drops sharply, but the quantization interval also increases, degrading the quality of the measured observations. We again note a slight advantage for the **LASSO**∞ and **L**2**Dantzig**∞ models, with very similar performance by **L**2 when the oversampling is lower.

In Figure 9, we fix N, M, S, B, R, and tune the value of G to achieve specified saturation ratios of 2% and 10%. We plot SNR against the confidence level P, varied from 0% to 100%. Again, we see generally good performance from the **LASSO**∞ and **L**2**Dantzig**∞ models, with **L**2 being competitive for less sparse solutions.

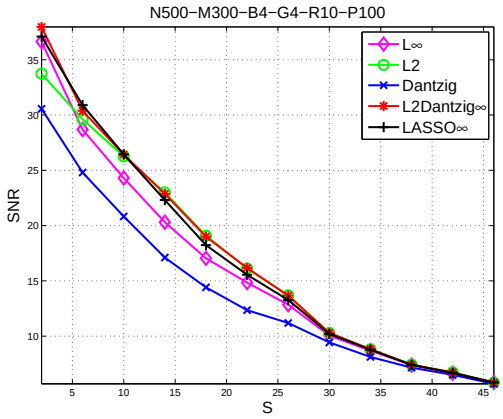

Figure 2: Comparison among various models for $N = 500$, $M = 300$, $B = 4$, $G = 0.4$, $R = 10$, and $P = 100\%$. The graph shows sparsity level S (horizontal axis) plotted against SNR (vertical axis), averaged over 30 trials.

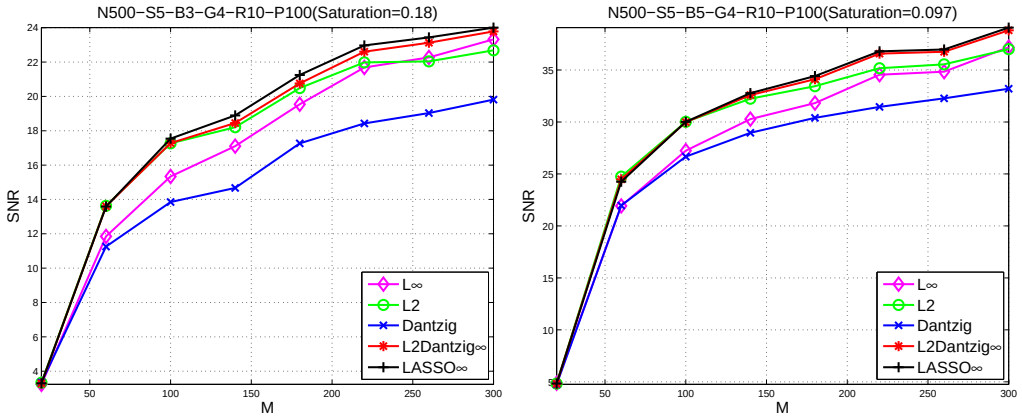

Figure 3: Comparison among various models for fixed values $N = 500$, $S = 5$, $G = 0.4$, $R = 15$, and $P = 100\%$, and two values of B (3 and 5). The graphs show the number of measurements M (horizontal axis) against SNR (vertical axis) for values of M between 20 and 300, averaged over 30 trials for each combination of parameters.

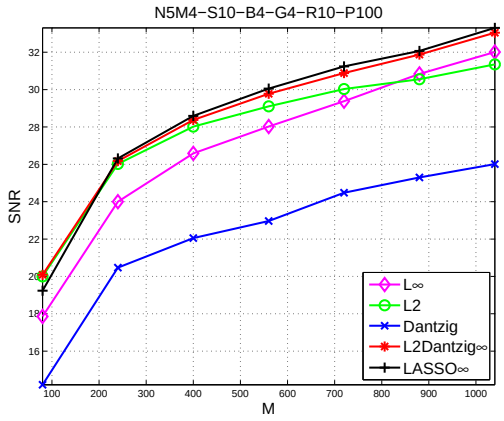

Figure 4: Comparison among various models for fixed ratio $N/M = 5/4$, and fixed values $S = 10$, $B = 4$, $G = 0.4$, $R = 15$, and $P = 100\%$. The graph shows the number of measurements M (horizontal axis) against SNR (vertical axis) for values of M between 100 and 1680, averaged over 30 trials for each combination of parameters.

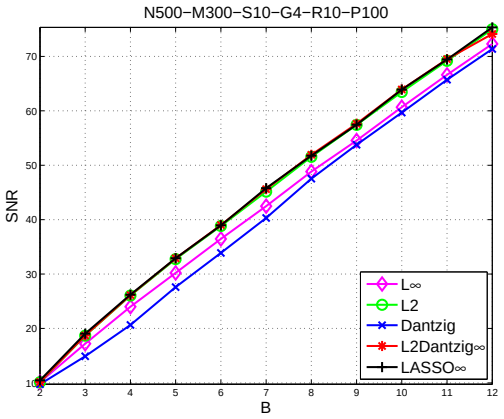

Figure 5: Comparison among various models for fixed values $N = 500$, $M = 300$, $S = 10$, $G = 0.4$, $R = 10$, and $P = 100\%$. This graph shows the bit number B (horizontal axis) against SNR (vertical axis), averaged over 30 trials.

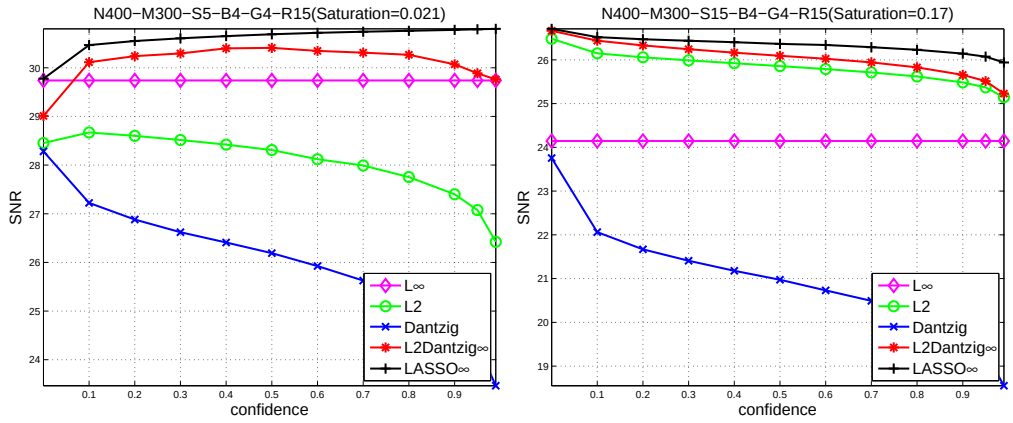

Figure 6: Comparison among various models for fixed values $N = 400$, $M = 300$, $B = 4$, $G = 0.4$, and $R = 15$, and sparsity levels $S = 5$ and $S = 15$. The graphs show saturation bound G (horizontal axis) against SNR (vertical axis) for values of P between 0.0001 and 0.99, averaged over 30 trials for each combination of parameters.

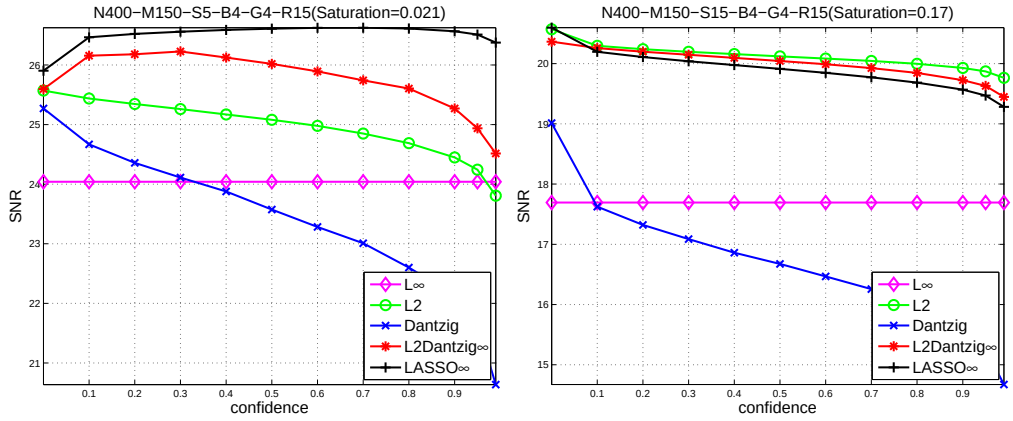

Figure 7: Comparison among various models for fixed values $N = 400$, $M = 150$, $B = 4$, $G = 0.4$, and $R = 15$, and sparsity levels $S = 5$ and $S = 15$. The graphs show confidence P (horizontal axis) against SNR (vertical axis) for values of P between 0.0001 and 0.99, averaged over 30 trials for each combination of parameters.

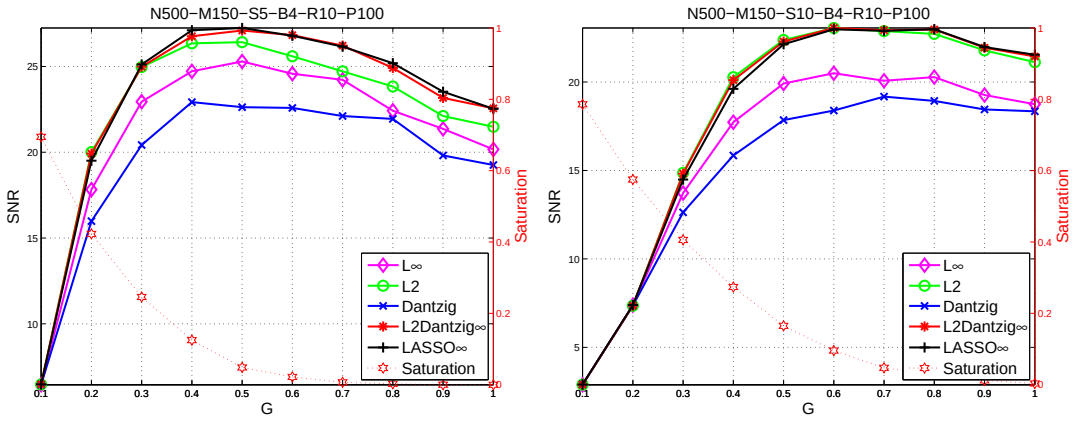

Figure 8: Comparison among various models for fixed values of $N = 500$, $M = 150$, $B = 4$, $R = 15$, $P = 100\%$, and two values of S: 5 and 10. The graphs show confidence P (horizontal axis) against SNR (left vertical axis) and saturation ratio (right vertical axis), averaged over 30 trials for each combination of parameters.

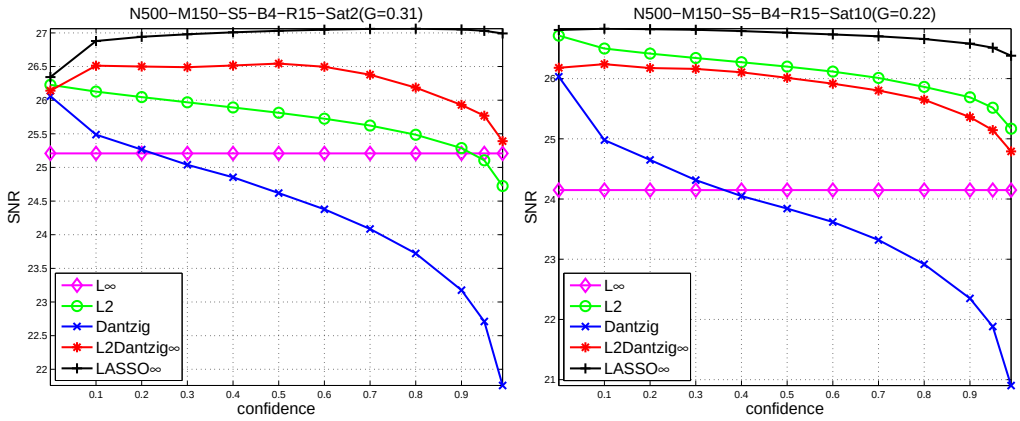

Figure 9: Comparison among various models for fixed values of $N = 500$, $M = 150$, $S = 5$, $B = 4$, $R = 15$, and two values of saturation ratio: 2% and 10%, which are achieved by tuning the value of G. The graphs show confidence P (horizontal axis) against SNR (vertical axis), averaged over 30 trials for each combination of parameters.

Summarizing, we note the following points.

(a) Our proposed **LASSO**∞ formulation gives either best or equal-best reconstruction performance in most regimes, with a more marked advantage when the signal is highly sparse and the number of samples is higher.

(b) The **L**2 model has similar performance to the full model, and is even slightly better than our model for less sparse signals with fewer measurements, since it is not sensitive to the measurement number as the upper bound suggested by Laska et al. (2011). Although the inequality in (15) also indicates the estimate error by our model is bounded by a constant due to the ℓ_∞ constraint, the error bound determined by the ℓ_∞ constraint is not so tight as the ℓ_2 constraint in general. This fact is evident when we compare the the **L**∞ model with the **L**2 model.

(c) The **L**∞ model performs well (and is competitive with the others) when the number of unsaturated measurements is relatively large.

(d) The **L**2**Dantzig**∞ model is competitive with **LASSO**∞ if ϵ and λ can be determined from the true signal x^*. Otherwise, **LASSO**∞ is more robust to choices of these parameters that do not require knowledge of the true signals, especially if one desires a high confidence level.

5 Conclusion

We have analyzed a formulation of the reconstruction problem from compressed sensing in which the measurements are quantized to a finite number of possible values. Our formulation uses an objective of ℓ_2-ℓ_1 type, along with explicit constraints that restrict the individual quantization errors to known intervals. We obtain bounds on the estimation error, and estimate these bounds for the case in which the sensing matrix is Gaussian. Finally, we prove the practical utility of our formulation by comparing with an approach that has been proposed previously, along with some variations on this approach that attempt to distil the relative importance of different constraints in the formulation.

Acknowledgments

The authors acknowledge support of National Science Foundation Grant DMS-0914524 and a Wisconsin Alumni Research Foundation 2011-12 Fall Competition Award.

Appendix

This section contains the proof to Theorem 1, developed via a number of technical lemmas. At the end, we state and prove a result (Theorem 2) concerning high-probability estimates of the bounds under additional assumptions on the sensing matrix $\tilde{\Phi}$.

The proof of Theorem 1 essentially follows the standard analysis procedure in compressive sensing. Some similar lemmas and proofs can be found in Bickel et al. (2009); Candès and Tao (2007); Candès (2008); Zhang (2009); Liu et al. (2010, 2012). For completeness, we include all proofs in the following discussion.

Let T_0 be any subset of $\{1, 2, ..., N\}$, with $s := |T_0|$. Given the error vector $h = \hat{x} - x^*$, divide the complementary index set $T_0^c := \{1, 2, ..., N\} \backslash T_0$ into a group of subsets T_j's ($j = 1, 2, \ldots, J$), without intersection, such that T_1 indicates the index set of the largest l entries of $h_{T_0^c}$, T_2 contains the next-largest l entries of $h_{T_0^c}$, and so forth.[2]

Lemma 1. *We have*

$$\|\tilde{\Phi} h\|_\infty \leq \Delta. \tag{17}$$

[2]The last subset may contain fewer than l elements.

Proof. From (3b), and invoking feasibility of $\hat{x}$ and x^*, we obtain

$$\|\tilde{\Phi}h\|_\infty = \|\tilde{\Phi}(\hat{x}-x^*)\|_\infty \le \|\tilde{\Phi}\hat{x}-\tilde{y}\|_\infty + \|\tilde{\Phi}x^*-\tilde{y}\|_\infty \le \Delta.$$

□

Lemma 2. *Suppose that Assumption 1 holds. Given $\pi \in (0,1)$, the choice $\lambda = \sqrt{2\log(2N/\pi)}f_{\max}$ ensures that the true signal x^* satisfies* (11)*, that is*

$$\|\tilde{\Phi}^T(\tilde{\Phi}x^*-\tilde{y})\|_\infty \le \lambda\Delta/2$$

with probability at least $1-\pi$.

Proof. Define the random variable $Z_j = \tilde{\Phi}_{\cdot j}^T(\tilde{\Phi}x^*-\tilde{y}) = \tilde{\Phi}_{\cdot j}^T\xi$, where $\xi = [\xi_1,...,\xi_{\tilde{M}}]$ is defined in an obvious way. (Note that $\|Z\|_\infty = \|\tilde{\Phi}^T(\tilde{\Phi}x^*-\tilde{y})\|_\infty$.) Since $\mathbb{E}(Z_j)=0$ (from Assumption 1) and all $\tilde{\Phi}_{ij}\xi_i$'s are in the range $[-\tilde{\Phi}_{ij}\Delta/2, \tilde{\Phi}_{ij}\Delta/2]$, we use the Hoeffding inequality to obtain

$$\begin{aligned}
\mathbb{P}(Z_j > \lambda\Delta/2) =& \mathbb{P}(Z_j - \mathbb{E}(Z_j) > \lambda\Delta/2)\\
=& \mathbb{P}(\sum_{i=1}^{\tilde{M}} \tilde{\Phi}_{ij}\xi_i - \mathbb{E}(Z_j) > \lambda\Delta/2)\\
\le& \exp\frac{-2(\lambda\Delta/2)^2}{\sum_{i=1}^{\tilde{M}}(\tilde{\Phi}_{ij}\Delta)^2}\\
=& \exp\frac{-\lambda^2}{2\sum_i \tilde{\Phi}_{ij}^2}\\
\le& \exp\frac{-\lambda^2}{2f_{\max}^2},
\end{aligned}$$

which implies (using the union bound) that

$$\begin{aligned}
&\mathbb{P}(|Z_j| > \lambda\Delta/2) \le 2\exp\frac{-\lambda^2}{2f_{\max}^2}\\
\Rightarrow &\mathbb{P}(\|Z\|_\infty = \max_j |Z_j| > \lambda\Delta/2) \le 2N\exp\frac{-\lambda^2}{2f_{\max}^2}\\
\Rightarrow &\mathbb{P}\left(\|Z\|_\infty > \sqrt{\frac{1}{2}\log\frac{2N}{\pi}}f_{\max}\Delta\right) \le \pi,
\end{aligned}$$

where the last line follows by setting λ to the prescribed value. This completes the proof. □

Lemma 3. *We have*

$$\|h_{T_{01}^c}\| \le \sum_{j=2}^{J}\|h_{T_j}\| \le \|h_{T_0^c}\|_1/\sqrt{l},$$

where $T_{01} = T_0 \cup T_1$.

Proof. First, we have for any $j \ge 1$ that

$$\|h_{T_{j+1}}\|^2 \le l\|h_{T_{j+1}}\|_\infty^2 \le l(\|h_{T_j}\|_1/l)^2 = \|h_{T_j}\|_1^2/l,$$

because the largest value in $|h_{T_{j+1}}|$ cannot exceed the average value of the components of $|h_{T_j}|$. It follows that

$$\|h_{T_{01}^c}\| \le \sum_{j=2}^{J}\|h_{T_j}\| \le \sum_{j=1}^{J-1}\|h_{T_j}\|_1/\sqrt{l} \le \|h_{T_0^c}\|_1/\sqrt{l}.$$

□

Similar claims or inequalities to Lemma 3 can be found in Zhang (2009); Candès and Tao (2007); Liu et al. (2010).

Lemma 4. *Assume that* (11) *holds. We have*

$$\|h_{T_0^c}\|_1 \le 3\|h_{T_0}\|_1 + 4\|x^*_{T_0^c}\|_1, \tag{18a}$$

$$\|h\| \le \sqrt{1+9s/l}\|h_{T_{01}}\| + 4\|x^*_{T_0^c}\|_1/\sqrt{l}. \tag{18b}$$

Proof. Since $\hat{x}$ is the solution of (3), we have

$$\begin{aligned}
0 &\ge \frac{1}{2}\|\tilde{\Phi}\hat{x} - \tilde{y}\|^2 - \frac{1}{2}\|\tilde{\Phi}x^* - \tilde{y}\|^2 + \lambda\Delta(\|\hat{x}\|_1 - \|x^*\|_1) \\
&\ge h^T\tilde{\Phi}^T(\tilde{\Phi}x^* - \tilde{y}) + \lambda\Delta(\|\hat{x}\|_1 - \|x^*\|_1) \qquad \text{(by convexity of } (1/2)\|\tilde{\Phi}x - \tilde{y}\|^2) \\
&= h^T\tilde{\Phi}^T(\tilde{\Phi}x^* - \tilde{y}) + \lambda\Delta(\|\hat{x}_{T_0}\|_1 - \|x^*_{T_0}\|_1 + \|\hat{x}_{T_0^c}\|_1 - \|x^*_{T_0^c}\|_1) \\
&\ge -\|h\|_1\|\tilde{\Phi}^T(\tilde{\Phi}x^* - \tilde{y})\|_\infty + \lambda\Delta(\|\hat{x}_{T_0}\|_1 - \|x^*_{T_0}\|_1 + \|\hat{x}_{T_0^c}\|_1 - \|x^*_{T_0^c}\|_1) \\
&\ge -\|h\|_1\lambda\Delta/2 + \lambda\Delta(\|\hat{x}_{T_0}\|_1 - \|x^*_{T_0}\|_1 + \|\hat{x}_{T_0^c}\|_1 + \|x^*_{T_0^c}\|_1 - 2\|x^*_{T_0^c}\|_1) \quad \text{(from (11))} \\
&\ge -(\|h_{T_0}\|_1 + \|h_{T_0^c}\|_1)\lambda\Delta/2 + \lambda\Delta(-\|h_{T_0}\|_1 + \|h_{T_0^c}\|_1 - 2\|x^*_{T_0^c}\|_1) \\
&\ge \frac{1}{2}\lambda\Delta\|h_{T_0^c}\|_1 - \frac{3}{2}\lambda\Delta\|h_{T_0}\|_1 - 2\lambda\Delta\|x^*_{T_0^c}\|_1.
\end{aligned}$$

It follows that $3\|h_{T_0}\|_1 + 4\|x^*_{T_0^c}\|_1 \ge \|h_{T_0^c}\|_1$, proving (18a).

The second inequality (18b) is from

$$\begin{aligned}
\|h\|^2 &= \|h_{T_{01}}\|^2 + \|h_{T_{01}^c}\|^2 \\
&\le \|h_{T_{01}}\|^2 + \|h_{T_0^c}\|_1^2/l \quad \text{(from Lemma 3)} \\
&\le \|h_{T_{01}}\|^2 + (3\|h_{T_0}\|_1 + 4\|x^*_{T_0^c}\|_1)^2/l \quad \text{(from (18a))} \\
&\le \|h_{T_{01}}\|^2 + (3\sqrt{s}\|h_{T_{01}}\| + 4\|x^*_{T_0^c}\|_1)^2/l \\
&= (1+9s/l)\|h_{T_{01}}\|^2 + 24\sqrt{s}/l\|h_{T_{01}}\|\|x^*_{T_0^c}\|_1 + 16\|x^*_{T_0^c}\|_1^2/l \\
&\le \left[\sqrt{1+9s/l}\|h_{T_{01}}\| + 4\|x^*_{T_0^c}\|_1/\sqrt{l}\right]^2.
\end{aligned}$$

□

Lemma 5. *Assume that* (11) *holds. For any matrix Ψ with N columns, and $s, l \le N$, we have*

$$\|\Psi h\|^2 \ge A_0(\Psi)\|h_{T_{01}}\|^2 - A_1(\Psi)\|h_{T_{01}}\|\|x^*_{T_0^c}\|_1/\sqrt{l},$$

where $A_0(\Psi)$ and $A(\Psi)$ are defined in (12c) *and* (12d) *respectively.*

Proof. For any $j \ge 2$, we have

$$\begin{aligned}
&\frac{|h^T_{T_{01}}\Psi^T_{T_{01}}\Psi_{T_j}h_{T_j}|}{\|h_{T_{01}}\|\|h_{T_j}\|} \\
&= \frac{1}{4}\left|\left\|\Psi_{T_{01}}h_{T_{01}}/\|h_{T_{01}}\| + \Psi_{T_j}h_{T_j}/\|h_{T_j}\|\right\|^2 - \left\|\Psi_{T_{01}}h_{T_{01}}/\|h_{T_{01}}\| - \Psi_{T_j}h_{T_j}/\|h_{T_j}\|\right\|^2\right| \\
&= \frac{1}{4}\left|\left\|\begin{bmatrix}\Psi_{T_{01}} & : & \Psi_{T_j}\end{bmatrix}\begin{bmatrix}h_{T_{01}}/\|h_{T_{01}}\| \\ h_{T_j}/\|h_{T_j}\|\end{bmatrix}\right\|^2 - \left\|\begin{bmatrix}\Psi_{T_{01}} & : & \Psi_{T_j}\end{bmatrix}\begin{bmatrix}h_{T_{01}}/\|h_{T_{01}}\| \\ -h_{T_j}/\|h_{T_j}\|\end{bmatrix}\right\|^2\right| \\
&\le \frac{1}{4}(2\rho_2^+(s+2l) - 2\rho_2^-(s+2l)) \\
&= \frac{1}{2}(\rho_2^+(s+2l) - \rho_2^-(s+2l)).
\end{aligned} \tag{19}$$

The inequality above follows from the definitions (6) and (7), and the fact that fact that $h_{T_{01}}/\|h_{T_{01}}\|$ and $h_{T_j}/\|h_{T_j}\|$ are ℓ_2-unit vectors, so that

$$\left\| \begin{bmatrix} h_{T_{01}}/\|h_{T_{01}}\| \\ h_{T_j}/\|h_{T_j}\| \end{bmatrix} \right\|^2 = \left\| \begin{bmatrix} h_{T_{01}}/\|h_{T_{01}}\| \\ -h_{T_j}/\|h_{T_j}\| \end{bmatrix} \right\|^2 = 2.$$

Considering the left side of the claimed inequality, we have

$$\begin{aligned}
&\|\Psi h\|^2 \\
&= \|\Psi_{T_{01}} h_{T_{01}}\|^2 + 2h_{T_{01}}^T \Psi_{T_{01}}^T \Psi_{T_{01}^c} h_{T_{01}^c} + \|\Psi_{T_{01}^c} h_{T_{01}^c}\|^2 \\
&\geq \|\Psi_{T_{01}} h_{T_{01}}\|^2 - 2\sum_{j\geq 2} |h_{T_{01}}^T \Psi_{T_{01}}^T \Psi_{T_j} h_{T_j}| \\
&\geq \rho_2^-(s+l)\|h_{T_{01}}\|^2 - (\rho_2^+(s+2l) - \rho_2^-(s+2l))\|h_{T_{01}}\| \sum_{j\geq 2} \|h_{T_j}\| \quad \text{(from (19))} \\
&\geq \rho_2^-(s+l)\|h_{T_{01}}\|^2 - (\rho_2^+(s+2l) - \rho_2^-(s+2l))\|h_{T_{01}}\| \|h_{T_0^c}\|_1/\sqrt{l} \quad \text{(from Lemma 3)} \\
&\geq \rho_2^-(s+l)\|h_{T_{01}}\|^2 - (\rho_2^+(s+2l) - \rho_2^-(s+2l))\|h_{T_{01}}\|(3\|h_{T_0}\|_1/\sqrt{l} + 4\|x_{T_0^c}^*\|_1/\sqrt{l}) \quad \text{(from (18a))} \\
&\geq \left(\rho_2^-(s+l) - 3\sqrt{s/l}(\rho_2^+(s+2l) - \rho_2^-(s+2l))\right)\|h_{T_{01}}\|^2 - \\
&\quad 4(\rho_2^+(s+2l) - \rho_2^-(s+2l))\|x_{T_0^c}^*\|_1 \|h_{T_{01}}\|/\sqrt{l} \quad (\text{using } \|h_{T_0}\|_1 \leq \sqrt{s}\|h_{T_0}\| \leq \sqrt{s}\|h_{T_{01}}\|) \\
&\geq A_0(\Psi)\|h_{T_{01}}\|^2 - A_1(\Psi)\|h_{T_{01}}\| \|x_{T_0^c}^*\|_1/\sqrt{l},
\end{aligned}$$

which completes the proof. □

Lemma 6. *Assume that* (11) *holds. We have*

$$\|\tilde{\Phi}h\|^2 \leq \frac{3}{2}\lambda\Delta\|h\|_1 \leq 6\sqrt{s}\lambda\Delta\|h_{T_{01}}\| + 6\lambda\Delta\|x_{T_0^c}^*\|_1, \tag{20a}$$

$$\|\tilde{\Phi}h\|^2 \leq \tilde{M}\Delta^2, \tag{20b}$$

Proof. Denote the feasible region of (11) as

$$F := \{x \mid \bar{\Phi}x - \bar{y} \geq 0,\ \|\tilde{\Phi}x - \tilde{y}\|_\infty \leq \Delta/2\}.$$

Since $\hat{x}$ is the optimal solution to (3), we have the optimality condition:

$$\tilde{\Phi}^T(\tilde{\Phi}\hat{x} - \tilde{y}) + \lambda\Delta\partial\|\hat{x}\|_1 \cap -N_F(\hat{x}) \neq \emptyset$$

where $N_F(\hat{x})$ denotes the normal cone of F at the point $\hat{x}$ and $\partial\|\hat{x}\|_1$ is the subgradient of the function $\|.\|_1$ at the point $\hat{x}$. It is equivalent to say that there exist $g \in \partial\|\hat{x}\|_1$ and $n \in N_F(\hat{x})$ such that

$$\tilde{\Phi}^T(\tilde{\Phi}\hat{x} - \tilde{y}) + \lambda\Delta g + n = 0.$$

It follows that

$$\begin{aligned}
&\tilde{\Phi}^T\tilde{\Phi}h + \tilde{\Phi}^T(\tilde{\Phi}x^* - \tilde{y}) + \lambda\Delta g + n = 0 \\
&\Rightarrow h^T\tilde{\Phi}^T\tilde{\Phi}h + h^T\tilde{\Phi}^T(\tilde{\Phi}x^* - \tilde{y}) + \lambda\Delta h^T g + h^T n = 0 \\
&\Rightarrow \|\tilde{\Phi}h\|^2 = -h^T\tilde{\Phi}^T(\tilde{\Phi}x^* - \tilde{y}) - \lambda\Delta h^T g - h^T n \\
&\Rightarrow \|\tilde{\Phi}h\|^2 \leq -h^T\tilde{\Phi}^T(\tilde{\Phi}x^* - \tilde{y}) - \lambda\Delta h^T g \quad (\text{using } x^* \in F \text{ and } h^T n \geq 0) \\
&\Rightarrow \|\tilde{\Phi}h\|^2 \leq \|h\|_1\|\tilde{\Phi}^T(\tilde{\Phi}x^* - \tilde{y})\|_\infty + \lambda\Delta\|h\|_1\|g\|_\infty.
\end{aligned}$$

From $\|g\|_\infty \le 1$ and (11), we obtain

$$\begin{aligned}\|\tilde{\Phi}h\|^2 &\le \lambda\Delta\|h\|_1/2 + \lambda\Delta\|h\|_1 \\ &= \frac{3}{2}\lambda\Delta\|h\|_1 \\ &= \frac{3}{2}\lambda\Delta(\|h_{T_0}\|_1 + \|h_{T_0^c}\|_1) \\ &\le \frac{3}{2}\lambda\Delta(4\|h_{T_0}\|_1 + 4\|x^*_{T_0^c}\|_1) \quad \text{(from (18a))} \\ &\le 6\sqrt{s}\lambda\Delta\|h_{T_0}\| + 6\lambda\Delta\|x^*_{T_0^c}\|_1,\end{aligned}$$

which proves the first inequality.

From (17), the second inequality is obtained by $\|\tilde{\Phi}h\|^2 \le \left(\sqrt{\tilde{M}\|\tilde{\Phi}h\|_\infty}\right)^2 \le \tilde{M}\Delta^2$. □

Proof of Theorem 1

Proof. First, assume that (11) holds. Take $\Psi = \tilde{\Phi}$ in Lemma 5 and apply (20a). We have

$$\begin{aligned}&A_0(\tilde{\Phi})\|h_{T_{01}}\|^2 - (A_1(\tilde{\Phi})/\sqrt{l})\|x^*_{T_{01}^c}\|_1\|h_{T_{01}}\| \\ \le&\|\tilde{\Phi}h\|^2 \\ \le&6\sqrt{s}\lambda\Delta\|h_{T_{01}}\| + 6\lambda\Delta\|x^*_{T_0^c}\|_1 \\ =&B_L\|h_{T_{01}}\| + (B_L/\sqrt{s})\|x^*_{T_0^c}\|_1.\end{aligned}$$

If follows that

$$A_0(\tilde{\Phi})\|h_{T_{01}}\|^2 - ((A_1(\tilde{\Phi})/\sqrt{l})\|x^*_{T_{01}^c}\|_1 + B_L)\|h_{T_{01}}\| \le B_L/\sqrt{s}\|x^*_{T_0^c}\|_1. \tag{21}$$

Using $A_0(\tilde{\Phi}) > 0$ (which is assumed in the statement of the theorem), we recall that for a quadratic inequality $ax^2 - bx \le c$ with $a, b, c > 0$, one has

$$x \le \frac{b + \sqrt{b^2 + 4ac}}{2a} \le \frac{2b + \sqrt{4ac}}{2a} = \frac{b}{a} + \sqrt{\frac{c}{a}}. \tag{22}$$

Hence (21) implies that

$$\begin{aligned}\|h_{T_{01}}\| &\le \frac{1}{A_0(\tilde{\Phi})}\left((A_1(\tilde{\Phi})/\sqrt{l})\|x^*_{T_{01}^c}\|_1 + B_L\right) + \sqrt{\frac{B_L\|x^*_{T_0^c}\|_1}{\sqrt{s}A_0(\tilde{\Phi})}} \\ &= \frac{B_L}{A_0(\tilde{\Phi})} + \frac{A_1(\tilde{\Phi})}{A_0(\tilde{\Phi})\sqrt{l}}\|x^*_{T_0^c}\|_1 + \sqrt{\frac{B_L}{\sqrt{s}A_0(\tilde{\Phi})}}\|x^*_{T_0^c}\|_1^{1/2}.\end{aligned}$$

By invoking (18b), we prove (13a) by

$$\begin{aligned}\|h\| &\le \sqrt{1 + 9s/l}\|h_{T_{01}}\| + (4/\sqrt{l})\|x^*_{T_0^c}\|_1 \\ &\le \frac{\sqrt{1 + 9s/l}B_L}{A_0(\tilde{\Phi})} + \left(4 + \frac{\sqrt{1 + 9s/l}A_1(\tilde{\Phi})}{A_0(\tilde{\Phi})}\right)(\|x^*_{T_0^c}\|_1/\sqrt{l}) + \sqrt{\frac{(1 + 9s/l)B_L}{\sqrt{s}A_0(\tilde{\Phi})}}\|x^*_{T_0^c}\|_1^{1/2} \\ &= C_0(\tilde{\Phi})B_L + C_1(\tilde{\Phi})\|x^*_{T_0^c}\|_1/\sqrt{l} + C_2(\tilde{\Phi})B_L^{1/2}(\|x^*_{T_0^c}\|_1^{1/2}/s^{1/4}).\end{aligned}$$

Next we prove (13b). Taking $\Psi = \tilde{\Phi}$ in Lemma 5 and applying (20b), we have

$$A_0(\tilde{\Phi})\|h_{T_{01}}\|^2 - (A_1(\tilde{\Phi})/\sqrt{l})\|x^*_{T_{01}^c}\|_1\|h_{T_{01}}\| \le \|\tilde{\Phi}h\|^2 \le \tilde{M}\Delta^2 = B_\infty^2.$$

Using (22) again, one has

$$\|h_{T_{01}}\| \leq \frac{A_1(\tilde{\Phi})}{A_0(\tilde{\Phi})}(\|x^*_{T_0^c}\|_1/\sqrt{l}) + \frac{B_\infty}{\sqrt{A_0(\tilde{\Phi})}}.$$

By invoking (18b), we have

$$\begin{aligned}\|h\| \leq & \sqrt{1+9s/l}\|h_{T_{01}}\| + (4/\sqrt{l})\|x^*_{T_0^c}\|_1 \\ \leq & \left(4 + \frac{\sqrt{1+9s/l}A_1(\tilde{\Phi})}{A_0(\tilde{\Phi})}\right)\|x^*_{T_0^c}\|_1/\sqrt{l} + \sqrt{\frac{1+9s/l}{A_0(\tilde{\Phi})}}B_\infty,\end{aligned}$$

proving (13b).

Note that all claims hold under the assumption that (11) is satisfied. Since Lemma 2 shows that (11) holds with probability at least $1-\pi$ with taking $\lambda = \sqrt{2\log(2N/\pi)}f_{\max}$, we conclude that all claims hold with the same probability. □

High-Probability Estimates of the Estimation Error

For use in these results, we define the quantity

$$\chi := \bar{M}/M = (M - \tilde{M})/M, \tag{23}$$

which is the fraction of saturated measurements.

Theorem 2. *Assume $\Phi \in \mathbb{R}^{M\times N}$ to be a Gaussian random matrix, that is, each entry is i.i.d. and drawn from a standard Gaussian distribution $\mathcal{N}(0,1)$. Let $\tilde{\Phi} \in \mathbb{R}^{\tilde{M}\times N}$ be the submatrix of Φ taking $\tilde{M}$ rows from Φ, with the remaining $\bar{M}$ rows being used to form the other submatrix $\bar{\Phi} \in \mathbb{R}^{\bar{M}\times N}$, as defined in* (4)*. Then by choosing a threshold τ sufficiently small, and assuming that χ satisfies the bound $\chi(1-\log\chi) \leq \tau$, we have for any $k \geq 1$ such that $k\log N = o(M)$ that, with probability larger than $1 - O(\exp(-\Omega(M)))$, the following estimates hold:*

$$\sqrt{\rho_2^+(k)} \leq \frac{17}{16}\sqrt{M} + o(\sqrt{M}), \tag{24a}$$

$$\sqrt{\rho_2^-(k)} \geq \frac{15}{16}\sqrt{M} - o(\sqrt{M}), \tag{24b}$$

$$\sqrt{\tilde{\rho}_2^+(k)} \leq \frac{17}{16}\sqrt{\tilde{M}} + o(\sqrt{M}), \tag{24c}$$

$$\sqrt{\tilde{\rho}_2^-(k)} \geq \frac{15}{16}\sqrt{\tilde{M}} - o(\sqrt{M}). \tag{24d}$$

Proof. From the definition of $\rho_2^+(k)$, we have

$$\sqrt{\rho_2^+(k)} = \max_{|T|\leq k, T\subset\{1,2,\ldots,N\}} \sigma_{\max}(\Phi_T),$$

where $\sigma_{\max}(\Phi_T)$ is the maximal singular value of Φ_T. From Vershynin (2011, Theorem 5.39), we have for any $t > 0$ that

$$\sigma_{\max}(\Phi_T) \leq \sqrt{M} + O(\sqrt{k}) + t$$

with probability larger than $1 - O(\exp(-\Omega(t^2))$. Since the number of possible choices for T is

$$\binom{N}{k} \leq \left(\frac{eN}{k}\right)^k,$$

we have with probability at least

$$1 - \binom{N}{k} O(\exp(-\Omega(t^2))) \geq 1 - O(\exp(k \log(eN/k) - \Omega(t^2))$$

that

$$\sqrt{\rho_2^+(k)} = \max_{|T| \leq k, T \subset \{1,2,\ldots,N\}} \sigma_{\max}(\Phi_T) \leq \sqrt{M} + O(\sqrt{k}) + t.$$

Taking $t = \sqrt{M}/16$, and noting that $k = o(M)$, we obtain the inequality (24a), with probability at least

$$\begin{aligned} & 1 - O(\exp(k \log(eN/k) - \Omega(t^2)) \\ &= 1 - O(\exp(k \log(eN/k) - \Omega(M)) \\ &= 1 - O(\exp(o(M) - \Omega(M)) \\ &\geq 1 - O(\exp(-\Omega(M))) \end{aligned}$$

The second inequality (24b) can be obtained similarly from

$$\min_{|T| \leq k, T \subset \{1,2,\ldots,N\}} \sigma_{\min}(\Phi_T) \leq \sqrt{M} - O(\sqrt{k}) - t,$$

where $\sigma_{\min}(\Phi_T)$ is the minimal singular value of Φ_T. (We set $t = \sqrt{M}/16$ as above.)

Next we prove (24c). We have

$$\sqrt{\tilde{\rho}_2^+(k)} = \max_{h, |T| \leq k} \frac{\|\tilde{\Phi}_T h_T\|}{\|h_T\|} \leq \max_{|T| \leq k, |R| \leq \tilde{M}} \sigma_{\max}(\Phi_{R,T}),$$

where $R \subset \{1, 2, \ldots, M\}$ and $T \subset \{1, 2, \ldots, N\}$ are subsets of the row and column indices of Φ, respectively, and $\Phi_{R,T}$ is the submatrix of Φ consisting of rows in R and columns in T. We now apply the result in Vershynin (2011, Theorem 5.39) again: For any $t > 0$, we have

$$\sigma_{\max}(\Phi_{R,T}) \leq \sqrt{\tilde{M}} + O(\sqrt{k}) + t$$

with probability larger than $1 - O(\exp(-\Omega(t^2)))$. The number of possible choices for R is

$$\binom{M}{\bar{M}} \leq \left(\frac{eM}{\bar{M}}\right)^{\bar{M}} = \left(\frac{e}{\chi}\right)^{\chi M} = \exp(M\chi \log(e/\chi)) \leq \exp(\tau M),$$

so that the number of possible combinations for (R, T) is bounded as follows:

$$\binom{M}{\bar{M}} \binom{N}{k} \leq \exp\left(\tau M + k \log(eN/k)\right).$$

We thus have

$$\begin{aligned} & \mathbb{P}\left(\sqrt{\tilde{\rho}_2^+(k)} \leq \sqrt{\tilde{M}} + O(\sqrt{k}) + t\right) \\ &\geq \mathbb{P}\left(\max_{|R| \leq \tilde{M}, |T| \leq k} \sigma(\Phi_{R,T}) \leq \sqrt{\tilde{M}} + O(\sqrt{k}) + t\right) \\ &\geq 1 - \binom{M}{\tilde{M}} \binom{N}{k} O(e^{-\Omega(t^2)}) \\ &= 1 - O\left[\exp\left(\tau M + k \log(eN/k) - \Omega(t^2)\right)\right]. \end{aligned}$$

Taking $t = \sqrt{\tilde{M}}/16$, and noting again that $k = o(M)$, we obtain the inequality in (24c). Working further on the probability bound, for this choice of t, we have

$$\begin{aligned} & 1 - O\left[\exp\left(\tau M + k\log(eN/k) - \Omega(\tilde{M})\right)\right] \\ &= 1 - O\left[\exp\left(\tau M + k\log(eN/k) - \Omega(M)\right)\right] \\ &= 1 - O(\exp(-\Omega(M))), \end{aligned}$$

where the first equality follows from $\tilde{M} = (1-\chi)M$ and for the second equality we assume that τ is chosen small enough to ensure that the $\Omega(M)$ term in the exponent dominates the τM term.

A similar procedure can be used to prove (24d). □

We finally derive estimates of $C_0(\tilde{\Phi})$, $C_1(\tilde{\Phi})$, $C_2(\tilde{\Phi})$, and $f_{\max}$, that are used in the discussion at the end of Section 3.

From Theorem 2, we have that under assumptions (iii), (iv), and (v) the quantity $A_1(\tilde{\Phi})$ defined in (12d) is bounded as follows:

$$\begin{aligned} A_1(\tilde{\Phi}) &= 4\left(\sqrt{\tilde{\rho}_2^+(s+2l)} + \sqrt{\tilde{\rho}_2^-(s+2l)}\right)\left(\sqrt{\tilde{\rho}_2^+(s+2l)} - \sqrt{\tilde{\rho}_2^-(s+2l)}\right) \\ &\leq 4(2\sqrt{\tilde{M}} + o(\sqrt{M}))(\frac{1}{8}\sqrt{\tilde{M}} + o(\sqrt{M})) \\ &= \tilde{M} + o(M) = \Omega(M). \end{aligned}$$

Using $s = l$, the quantity of $A_0(\tilde{\Phi})$ defined in (12c) is bounded as follows:

$$\begin{aligned} A_0(\tilde{\Phi}) &= \tilde{\rho}_2^-(s+l) - \frac{3\sqrt{s/l}}{4}A_1(\tilde{\Phi}) \\ &\geq \frac{15}{16}\tilde{M} - o(M) - \frac{3}{4}\tilde{M} - o(M) \\ &= \frac{3}{16}\tilde{M} - o(M) \\ &= \Omega(M), \end{aligned}$$

for all sufficiently large dimensions and small saturation ratio χ, since $\tilde{M} = (1-\chi)M$. Using the estimates above for $A_0(\tilde{\Phi})$ and $A_1(\tilde{\Phi})$, together with $s = l$, in the definitions (12e), (12f), and (12g), we obtain

$$\begin{aligned} C_0(\tilde{\Phi}) &= \frac{\sqrt{1+9s/l}}{A_0(\Psi)} = \Omega(1/M), \\ C_1(\tilde{\Phi}) &= 4 + \frac{\sqrt{1+9s/l}A_1(\Psi)}{A_0(\Psi)} = \Omega(1), \\ C_2(\tilde{\Phi}) &= \sqrt{\frac{1+9s/l}{A_0(\Psi)}} = \Omega(1/\sqrt{M}), \end{aligned}$$

as claimed. Finally, the estimate of $f_{\max}$ can be estimated by $f_{\max} = \sqrt{\tilde{\rho}_2^+(1)} \leq \frac{17}{16}\sqrt{\tilde{M}} + o(M) = O(\sqrt{M})$.

References

A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM Journal on Imaging Sciences 2 (2009) 183–202.

P.J. Bickel, Y. Ritov, A. Tsybakov, Simultaneous analysis of Lasso and Dantzig selector, Annals of Statistics 4 (2009) 1705–1732.

P.T. Boufounos, R.G. Baraniuk, 1-bit compressive sensing, CISS (2008) 19–21.

S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method and multipliers, Foundations and Trends in Machine Learning 3 (2011) 1–122.

E. Candès, J. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Comm. Pure Appl. Math. 59 (2006) 1207–1223.

E.J. Candès, The restricted isometry property and its implications for compressive sensing, C. R. Acad. Sci. Paris, Ser. I 346 (2008) 589–592.

E.J. Candès, T. Tao, The Dantzig selector: Statistical estimation when p is much larger than n, Annals of Statistics 35 (2007) 2392–2404.

W. Dai, H.V. Pham, O. Milenkovic, Quantized Compressive Sensing, Technical Report, Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, 2011.

M.F. Duarte, M.A. Davenport, D. Takhar, J.N. Laska, T. Sun, K.F. Kelly, R.G. Baraniuk, Single-pixel imaging via compressive sampling, IEEE Signal Processing Magazine 25 (2008) 83–91.

J. Eckstein, D.P. Bertsekas, On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators, Mathematical Programming 55 (1992) 293–318.

A. Gupta, R. Nowak, B. Recht, Sample complexity for 1-bit compressed sensing and sparse classification, ISIT (2010).

B.S. He, H. Yang, S.L. Wang, Alternating direction method with self- adaptive penalty parameters for monotone variational inequalities, Journal of Optimization Theory and Applications 106 (2000) 337–356.

L. Jacques, D.K. Hammond, M.J. Fadili, Dequantizing compressed sensing: When oversampling and non-gaussian constraints combine, IEEE Transactions on Information Theory 57 (2011) 559–571.

J.N. Laska, P.T. Boufounos, M.A. Davenport, R.G. Baraniuk, Democracy in action: Quantization, saturation, and compressive sensing, Applied and Computational Harmonic Analysis 39 (2011) 429–443.

J.N. Laska, S. Kirolos, M.F. Duarte, T. Ragheb, R.G. Baraniuk, Y. Massoud, Theory and implementation of an analog-to-information converter using random demodulation, ISCAS (2007) 1959–1962.

J. Liu, P. Wonka, J. Ye, Multi-stage Dantzig selector, NIPS (2010) 1450–1458.

J. Liu, P. Wonka, J. Ye, A multi-stage framework for dantzig selector and lasso, Journal of Machine Learning Research 13 (2012) 1189–1219.

Y. Nesterov, Gradient methods for minimizing composite objective function, CORE Discussion Papers 2007076, Universit catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2007.

J.K. Romberg, Compressive sensing by random convolution, SIAM J. Imaging Sciences 2 (2009) 1098–1128.

J.A. Tropp, J.N. Laska, M.F. Duarte, J.K. Romberg, R.G. Baraniuk, Beyond Nyquist: Efficient sampling of sparse bandlimited signals, CoRR abs/0902.0026 (2009).

J.A. Tropp, M.B. Wakin, M.F. Duarte, D. Baron, R.G. Baraniuk, Random filters for compressive sampling and reconstruction, ICASSP 3 (2006) 872–875.

R. Vershynin, Introduction to the non-asymptotic analysis of random matrices, arXiv:1011.3027 (2011).

S.J. Wright, R.D. Nowak, M.A.T. Figueiredo, Sparse reconstruction by separable approximation, IEEE Transactions on Signal Processing 57 (2009) 2479–2493.

T. Zhang, Some sharp performance bounds for least squares regression with l_1 regularization, Annals of Statistics 37 (2009) 2109–2114.

A. Zymnis, S. Boyd, E.J. Candès, Compressed sensing with quantized measurements, Signal Processing Letters (2010) 149–152.