Using an Information Theoretic Metric for Compressive Recovery under Poisson Noise

Sukanya Patil[a], Karthik S. Gurumoorthy[b], Ajit Rajwade[c,*]

[a]*Department of Electrical Engineering, IIT Bombay*
[b]*International Center for Theoretical Sciences, TIFR (ICTS-TIFR), Bangalore*
[c]*Department of Computer Science and Engineering, IIT Bombay*

Abstract

Recovery error bounds in compressed sensing under Gaussian or uniform bounded noise do not translate easily to the case of Poisson noise. Reasons for this include the signal dependent nature of Poisson noise, and also the fact that the negative log likelihood in case of a Poisson distribution (which is directly related to the generalized Kullback-Leibler divergence) is not a metric and does not obey the triangle inequality. There exist prior theoretical results in the form of provable error bounds for computationally tractable estimators for compressed sensing problems under Poisson noise. However, these results do not apply to *realistic* compressive systems, which must obey some crucial constraints such as non-negativity and flux preservation. On the other hand, there exist provable error bounds for such realistic systems in the published literature, but they are for estimators that are computationally intractable. In this paper, we develop error bounds for a computationally tractable estimator which also applies to realistic compressive systems obeying the required constraints. The focus of our technique is on the replacement of the generalized Kullback-Leibler divergence, with an information theoretic metric - namely the square root of the Jensen-Shannon divergence, which is related to an approximate, symmetrized version of the Poisson log likelihood function. We show that our method allows for very simple proofs of the error bounds. We also propose and prove several interesting statistical properties of the square root of Jensen-Shannon divergence, a well-known information-theoretic metric, and exploit other known ones. Numerical experiments are performed showing the practical use of the technique in signal and image reconstruction from compressed measurements under Poisson noise. Our technique has the following features: (i) It is applicable to signals that are sparse or compressible in any orthonormal basis. (ii) It works with high probability for any randomly generated sensing matrix that obeys the non-negativity and flux preservation constraints, and is derived from a 'base matrix' that obeys

*Corresponding author
Email addresses: sukanyapatil1993@gmail.com (Sukanya Patil), karthik.gurumoorthy@icst.res.in (Karthik S. Gurumoorthy), ajitvr@cse.iitb.ac.in (Ajit Rajwade)

[1]Karthik S. Gurumoorthy thanks the AIRBUS Group Corporate Foundation Chair in Mathematics of Complex Systems established in ICTS-TIFR.
[2]Ajit Rajwade gratefully acknowledges support from IIT Bombay Seed Grant number 14IRCCSG012.

the restricted isometry property. (iii) Most importantly, our proposed estimator uses parameters that are purely statistically motivated and signal independent, as opposed to techniques (such as those based on the Poisson negative log-likelihood or ℓ_2 data-fidelity) that require the choice of a regularization or signal sparsity parameter which are unknown in practice.

Keywords: Compressed sensing, Poisson noise, reconstruction error bounds, information theoretic metric, Jensen-Shannon divergence, triangle inequality

1. Introduction

Compressed sensing is today a very mature field of research in signal processing, with several advances on the theoretical, algorithmic as well as application fronts. The theory essentially considers measurements of the form $\boldsymbol{y} = \boldsymbol{\Phi x} = \boldsymbol{\Phi\Psi\theta} = \boldsymbol{A\theta}$ where $\boldsymbol{y} \in \mathbb{R}^N$ is a measurement vector, $\boldsymbol{A} \in \mathbb{R}^{N\times m} \triangleq \boldsymbol{\Phi\Psi}$, $\boldsymbol{\Psi} \in \mathbb{R}^{m\times m}$ is a signal representation orthonormal basis, and $\boldsymbol{\theta} \in \mathbb{R}^m$ is a vector that is sparse or compressible such that $\boldsymbol{x} = \boldsymbol{\Psi\theta}$. Usually $N \ll m$. Under suitable conditions on the sensing matrix such as the restricted isometry property (RIP) and sparsity-dependent lower bounds on N, it is proved that $\boldsymbol{x}$ can be recovered near-accurately given $\boldsymbol{y}$ and $\boldsymbol{\Phi}$, even if the measurement $\boldsymbol{y}$ is corrupted by signal-independent, additive noise $\boldsymbol{\eta}$ of the form $\boldsymbol{y} = \boldsymbol{\Phi x} + \boldsymbol{\eta}$ where $\boldsymbol{\eta} \sim \mathcal{N}(0, \sigma^2)$ or $\|\boldsymbol{\eta}\|_2 \leq \varepsilon$ (bounded noise). The specific error bound [1] on $\boldsymbol{\theta}$ in the case of $\|\boldsymbol{\eta}\|_2 \leq \varepsilon$ is given as:

$$\|\boldsymbol{\theta} - \boldsymbol{\theta^\star}\|_2 \leq C_1\varepsilon + \frac{C_2}{\sqrt{s}}\|\boldsymbol{\theta} - \boldsymbol{\theta_s}\|_1 \tag{1}$$

where $\boldsymbol{\theta_s}$ is a vector created by setting all entries of $\boldsymbol{\theta}$ to 0 except for those containing the s largest absolute values, $\boldsymbol{\theta^\star}$ is the minimum of the following optimization problem denoted as (P1),

$$\text{(P1): minimize}\|\boldsymbol{z}\|_1 \text{ such that } \|\boldsymbol{y} - \boldsymbol{Az}\|_2 \leq \varepsilon, \tag{2}$$

and C_1 and C_2 are increasing functions of δ_{2s}, the so-called restricted isometry constant (RIC) of $\boldsymbol{A}$. These bounds implicity require that $N \sim \Omega(s \log m)$, and $\boldsymbol{\Phi}$ (and hence $\boldsymbol{\Phi\Psi}$) is said to obey the RIP if $\delta_{2s} < 1$. Recently, such bounds have been extended to high SNR systems as in [2]. Other innovations, such as recovery with partially known support [3], or with a recent two-level weighted optimization have also been proposed [4], with promising results.

The noise affecting several different types of imaging systems is, however, known to follow the Poisson distribution. Examples include photon-limited imaging systems deployed in night-time photography [5], astronomy [6], low-dosage CT or X-ray imaging [7] or fluorescence microscopy [8, 9]. The Poisson noise model is given as follows:

$$\boldsymbol{y} \sim \text{Poisson}(\boldsymbol{\Phi x}) \tag{3}$$

where $\boldsymbol{x} \in \mathbb{R}^m_{\geq 0}$ is the *non-negative* signal or image of interest. The likelihood of observing a given measurement vector $\boldsymbol{y}$ is given as

$$p(\boldsymbol{y}|\boldsymbol{\Phi x}) = \prod_{i=1}^{N} \frac{[(\boldsymbol{\Phi x})_i]^{y_i} e^{-(\boldsymbol{\Phi x})_i}}{y_i!} \tag{4}$$

where y_i and $(\boldsymbol{\Phi x})_i$ are the i^{th} component of the vectors $\boldsymbol{y}$ and $\boldsymbol{\Phi x}$ respectively.

Unfortunately, the mathematical guarantees for compressive reconstruction from bounded or Gaussian noise [10, 1, 11] are no longer *directly* applicable to the case where the measurement noise follows a Poisson distribution, which is the case considered in this paper. One important reason for this is a feature of the Poisson distribution - that the mean and the variance are equal to the underlying intensity, thus deviating from the signal independent or bounded nature of other noise models.

Furthermore, the aforementioned practical imaging systems essentially act as photon-counting systems. Not only does this require non-negative signals of interest, but it also imposes crucial constraints on the nature of the sensing matrix $\boldsymbol{\Phi}$:

1. Non-negativity: $\forall i, \forall j, \Phi_{ij} \geq 0$
2. Flux-preservation: The total photon-count of the observed signal $\boldsymbol{\Phi x}$ can never exceed the photon count of the original signal $\boldsymbol{x}$, *i.e.*, $\sum_{i=1}^{N} (\boldsymbol{\Phi x})_i \leq \sum_{i=1}^{m} x_i$. This in turn imposes the constraint that every column of $\boldsymbol{\Phi}$ must sum up to a value no more than 1, i.e. $\forall j, \sum_{i=1}^{N} \Phi_{ij} \leq 1$.

A randomly generated non-negative and flux-preserving $\boldsymbol{\Phi}$ matrix does *not* (in general) obey the RIP. This situation is in contrast to randomly generated Gaussian or Bernoulli (± 1) random matrices which obey the RIP with high probability [12], and poses several challenges. However following prior work, we construct a related matrix $\tilde{\boldsymbol{\Phi}}$ from $\boldsymbol{\Phi}$ which obeys the RIP (see Section 2.1).

1.1. Main Contributions

The derivation of the theoretical performance bounds in Eqn. 1 based on the optimization problem in Eqn. 2 cannot be used in the Poisson noise model case, as it is well known that the use of the ℓ_2 norm between $\boldsymbol{y}$ and $\boldsymbol{\Phi x}$ leads to oversmoothing in the lower intensity regions and undersmoothing in the higher intensity regions. To estimate an unknown parameter set $\boldsymbol{x}$ given a set of Poisson-corrupted measurements $\boldsymbol{y}$, one proceeds by the maximum likelihood method. Dropping terms involving only $\boldsymbol{y}$, this reduces to maximization of the quantity $\sum_{i=1}^{N} y_i \log \frac{y_i}{(\boldsymbol{\Phi x})_i} - \sum_{i=1}^{N} y_i + \sum_{i=1}^{N} (\boldsymbol{\Phi x})_i$ which is called the generalized Kullback-Leibler divergence [13] between $\boldsymbol{y}$ and $\boldsymbol{\Phi x}$ - denoted as $G(\boldsymbol{y}, \boldsymbol{\Phi x})$. This divergence measure, however, does not obey the triangle inequality, quite unlike the ℓ_2 norm term in Eqn. 2 which is a metric. This 'metric-ness' of the ℓ_2 norm constraint is an important requirement for the error bounds in Eqn. 1 proved in [1]. For instance, the triangle inequality of the ℓ_2 norm is used to prove that $\|\boldsymbol{A}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\|_2 \leq 2\varepsilon$ where $\boldsymbol{\theta}^\star$ is the minimizer of

Problem (P1) in Eqn. 2. This is done in the following manner:

$$\|\boldsymbol{A}(\boldsymbol{\theta}-\boldsymbol{\theta}^{\star})\|_2 \leq \|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}\|_2 + \|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\theta}^{\star}\|_2 \leq 2\varepsilon. \tag{5}$$

This upper bound on $\|\boldsymbol{A}(\boldsymbol{\theta}-\boldsymbol{\theta}^{\star})\|_2$ is a crucial step in [1], for deriving the error bounds of the form in Eqn. 1.

The ℓ_2 norm is however not appropriate for the Poisson noise model for the aforementioned reasons. The first major contribution of this paper is to replace the ℓ_2 norm error term by a term which is more appropriate for the Poisson noise model and which, at the same time, is a metric. The specific error term that we choose here is the square root of the Jensen-Shannon divergence (defined in Section 2.2), which is a well-known information theoretic metric [14]. Hereafter we abbreviate the Jensen-Shannon divergence as JSD, its square-root as SQJSD, and denote them as J and $\sqrt{J}$ respectively within equations. Let $\boldsymbol{\theta}^{\star}$ be the minimizer of the following optimization problem which we denote as (P2):

$$\text{(P2): minimize}\|\boldsymbol{z}\|_1 \text{ such that } \sqrt{J(\boldsymbol{y},\boldsymbol{A}\boldsymbol{z})} \leq \varepsilon, \boldsymbol{\Psi}\boldsymbol{z} \succeq \boldsymbol{0}, \|\boldsymbol{\Psi}\boldsymbol{z}\|_1 = I, \tag{6}$$

where $I \triangleq \sum_{i=1}^{m} x_i$ is the total intensity of the signal of interest and ε is an upper bound on $\sqrt{J(\boldsymbol{y},\boldsymbol{A}\boldsymbol{z})}$ that we set to $\sqrt{N}(\frac{1}{2}+\sqrt{\frac{11}{8}+\frac{21}{16c}})$ where c is a constant (for reasons that will be clear in Section 2 and 7). We then prove that with high probability

$$\frac{\|\boldsymbol{\theta}-\boldsymbol{\theta}^{\star}\|_2}{I} \leq C_1\mathcal{O}\left(\frac{N}{\sqrt{I}}\right) + \frac{C_2}{I\sqrt{s}}\|\boldsymbol{\theta}-\boldsymbol{\theta}_{\boldsymbol{s}}\|_1 \tag{7}$$

where C_1 and C_2 are increasing functions of the RIC δ_{2s} of the sensing matrix $\tilde{\boldsymbol{\Phi}}$ derived from $\boldsymbol{\Phi}$. This result is proved in Section 2, followed by an extensive discussion. Note that for orthonormal $\boldsymbol{\Psi}$, we also have $\|\boldsymbol{x}-\boldsymbol{x}^{\star}\|_2 = \|\boldsymbol{\theta}-\boldsymbol{\theta}^{\star}\|_2$ where $\boldsymbol{x}^{\star} = \boldsymbol{\Psi}\boldsymbol{\theta}^{\star}$. In particular, we explain the reason behind the apparently counter-intuitive first term which is increasing in N: namely, that a Poisson imaging system distributes the total incident photon flux across the N measurements, reducing the SNR per measurement and hence affecting the performance. This phenomenon has been earlier observed in [15]. Our performance bounds derived independently and via a completely different method confirm the same phenomenon.

While there exists a body of earlier work on reconstruction error bounds for Poisson regression, the approach taken in this paper is different, and has the following features:

1. *Statistically motivated parameters:* Our proposed estimator does not require tweaking of a regularization or signal sparsity parameter, but uses a constrained optimization procedure with a *signal-independent* parameter dictated by the statistical properties of the SQJSD as shown in Section 2.2. This is in contrast with estimators based on the Poisson NLL or the ℓ_2 error between $\boldsymbol{y}$ and $\boldsymbol{A}\boldsymbol{\theta}$, which require regularization or constraint parameters which are *dependent on the unknown signal*. Hence, our estimator has significant advantages in terms of practical implementation.

2. *Confluence of computational tractability and realizability:* Existing techniques such as [15] work with intractable estimators for Poisson compressed sensing although they are designed to deal with physically realizable compressive systems. On the other hand, there are several techniques such as [16, 17, 18, 19] which are applicable to computationally efficient estimators (convex programs) for sparse Poisson regression and produce provable guarantees, but they do not impose important constraints required for physical implementability. Our approach, however, works with a computationally tractable estimator involving regularization with the ℓ_1 norm of the sparse coefficients representing the signal, while at the same time being applicable to physically realizable compressive systems. See Section 4 for a detailed comparison.
3. *Novel estimator:* Our technique demonstrates successfully (for the first time, to the best of our knowledge) the use of the JSD and the SQJSD for Poisson compressed sensing problems, at a theoretical as well as experimental level. Our work exploits several interesting properties of the JSD, some of which we derive in this paper.
4. *Simplicity:* Our technique affords (arguably) much simpler proofs than existing methods.

1.2. Organization of the Paper

The main theoretical result is derived in detail in Section 2, especially Section 2.2. Numerical simulations are presented in Section 3. Relation to prior work on Poisson compressed sensing is examined in detail in Section 4, followed by a discussion in Section 6. The proofs of some key theorems are presented in Section 7. The relation between the JSD and a symmetrized version of the Poisson likelihood is examined in Section 5.

2. Main Result

2.1. Construction of Sensing Matrices

We construct a sensing matrix $\mathbf{\Phi}$ ensuring that it corresponds to the forward model of a real optical system, based on the approach in [15]. Therefore it has to satisfy certain properties imposed by constraints of a physically realizable optical system - namely non-negativity and flux preservation. One major difference between Poisson compressed sensing and conventional compressed sensing emerges from the fact that conventional randomly generated sensing matrices which obey RIP do not follow the aforementioned physical constraints (although sensing matrices can be *designed* to obey the RIP, non-negativity and flux preservation simultaneously as in [20], and we comment upon this aspect in the remarks following the proof of our key theorem, later on in this section). In the following, we construct a sensing matrix $\mathbf{\Phi}$ which has only zeroes or (scaled) ones as entries. Let $\boldsymbol{Z}$ be a $N \times m$ matrix whose entries $Z_{i,j}$ are i.i.d random variables defined

as follows,

$$Z_{i,j} = \begin{cases} -\sqrt{\dfrac{1-p}{p}} & \text{with probability } p, \quad (8a) \\ \sqrt{\dfrac{p}{1-p}} & \text{with probability } 1-p. \quad (8b) \end{cases}$$

Let us define $\tilde{\boldsymbol{\Phi}} \triangleq \dfrac{\boldsymbol{Z}}{\sqrt{N}}$. For $p = 1/2$, the matrix $\tilde{\boldsymbol{\Phi}}$ now follows RIP of order $2s$ with a very high probability given by $1 - 2e^{-Nc(1+\delta_{2s})}$ where δ_{2s} is its RIC of order $2s$ and function $c(h) \triangleq \dfrac{h^2}{4} - \dfrac{h^3}{6}$ [12]. In other words, for any $2s$-sparse signal $\boldsymbol{\rho}$, the following holds with high probability

$$(1 - \delta_{2s})\|\boldsymbol{\rho}\|_2^2 \leq \|\tilde{\boldsymbol{\Phi}}\boldsymbol{\rho}\|_2^2 \leq (1 + \delta_{2s})\|\boldsymbol{\rho}\|_2^2.$$

Given any orthonormal matrix $\boldsymbol{\Psi}$, arguments in [12] show that $\tilde{\boldsymbol{\Phi}}\boldsymbol{\Psi}$ also obeys the RIP of the same order as $\tilde{\boldsymbol{\Phi}}$.

However $\tilde{\boldsymbol{\Phi}}$ will clearly contain negative entries with very high probability, which violates the constraints of a physically realizable system. To deal with this, we construct the flux-preserving and positivity preserving sensing matrix $\boldsymbol{\Phi}$ from $\tilde{\boldsymbol{\Phi}}$ as follows:

$$\boldsymbol{\Phi} \triangleq \sqrt{\frac{p(1-p)}{N}}\tilde{\boldsymbol{\Phi}} + \frac{(1-p)}{N}\mathbf{1}_{N\times m}, \quad (9)$$

which ensures that each entry of $\boldsymbol{\Phi}$ is either 0 or $\dfrac{1}{N}$. In addition, one can easily check that $\boldsymbol{\Phi}$ satisfies both the non-negativity as well as flux-preservation properties. We refer to $\tilde{\boldsymbol{\Phi}}$ as the 'base matrix' for $\boldsymbol{\Phi}$.

2.2. The Jensen-Shannon Divergence and its Square Root

The well-known Kullback-Leibler Divergence between vectors $\boldsymbol{p} \in {\mathbb{R}_{\geq 0}}^{N\times 1}$ and $\boldsymbol{q} \in {\mathbb{R}_{\geq 0}}^{N\times 1}$ denoted by $D(\boldsymbol{p}, \boldsymbol{q})$ is defined as[3]

$$D(\boldsymbol{p}, \boldsymbol{q}) \triangleq \sum_{i=1}^{N} p_i \log \frac{p_i}{q_i}. \quad (10)$$

Whenever $p_i = 0$ or $p_i = q_i = 0$, the i^{th} term is considered 0, since $\lim_{h\to 0^+} h \log h = 0$ as per Section 2.3 of [21]. The Jensen-Shannon Divergence between $\boldsymbol{p}$ and $\boldsymbol{q}$ denoted by $J(\boldsymbol{p}, \boldsymbol{q})$ is defined as

$$\begin{aligned} J(\boldsymbol{p}, \boldsymbol{q}) \triangleq \frac{D(\boldsymbol{p}, \boldsymbol{m}) + D(\boldsymbol{q}, \boldsymbol{m})}{2} &= \frac{1}{2}\sum_{i=1}^{N}(p_i \log p_i + q_i \log q_i) - \sum_{i=1}^{N} m_i \log m_i \quad (11) \\ &= \frac{1}{2}\sum_{i=1;p_i\neq 0}^{N} p_i \log p_i + \frac{1}{2}\sum_{i=1;q_i\neq 0}^{N} q_i \log q_i - \sum_{i=1;m_i\neq 0}^{N} m_i \log m_i, \end{aligned}$$

[3]Note that the Kullback-Leibler and other divergences are usually defined for probability mass functions, but they have also been used in the context of general non-negative vectors in the same manner as we do in this paper.

where $\boldsymbol{m} \triangleq \frac{1}{2}(\boldsymbol{p}+\boldsymbol{q}), m_i \triangleq (p_i+q_i)/2$. We note that each term in the summation has the form $h \log h$ for some real-valued non-negative h.

The performance bounds derived in this paper for reconstruction from Poisson-corrupted measurements deal with the estimate obtained by solving the constrained optimization problem (P2) in Eqn. 6, where we consider a statistically motivated upper bound of ε on the quantity $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}$. Now, while $\boldsymbol{\Phi x}$ is real-valued and non-negative, $\boldsymbol{y}$ consists of non-negative integers, possibly including zeros. The quantity $h \log h$ is not defined for $h = 0$ although $\lim_{h \to 0^+} h \log h = 0$. In our formulation, we *set* $h \log h = 0$ in the definition of J, to maintain continuity of J and mathematical cogency. This is completely in tune with the afore-mentioned definition of the Kullback-Leibler divergence as per Section 2.3 of [21].

The motivation for using the JSD will be evident from the following features of the JSD considered in this section: (1) the metric nature of (including the triangle inequality observed by) its square-root, (2) its relation with the total variation distance $V(\boldsymbol{p}, \boldsymbol{q}) \triangleq \sum_i |p_i - q_i|$, and (3) interesting statistical properties of $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}$.

We now enumerate the properties of the JSD, the last of which is *discovered and proved* in this paper. These properties are very useful in deriving the performance bounds in the following sub-section.

[Lemma 1 [Theorem 1 of [14]]: The square root of the Jensen-Shannon Divergence is a metric.

[Lemma 2 [Theorem 2 of [22]]: Let us define

$$V(\boldsymbol{p}, \boldsymbol{q}) \triangleq \sum_{i=1}^{N} |p_i - q_i|, \Delta(\boldsymbol{p}, \boldsymbol{q}) \triangleq \sum_{i=1}^{N} \frac{|p_i - q_i|^2}{p_i + q_i}.$$

If $\boldsymbol{p}, \boldsymbol{q} \succeq \boldsymbol{0}$ and $\|\boldsymbol{p}\|_1 \leq 1$, $\|\boldsymbol{q}\|_1 \leq 1$. Then,

$$\frac{1}{2} V(\boldsymbol{p}, \boldsymbol{q})^2 \leq \Delta(\boldsymbol{p}, \boldsymbol{q}) \leq 4 J(\boldsymbol{p}, \boldsymbol{q}). \tag{12}$$

Additionally, *we have experimentally observed some interesting properties of the distribution of the SQJSD values*, across different Poisson realizations of compressive measurements of a signal $\boldsymbol{x}$ with values generated from Unif$[0, 1]$ (with appropriate scaling). We considered a fixed and realistic sensing matrix $\boldsymbol{\Phi}$ as described in Section 2.1 (but with $p = 0.5$) . In other words, if $\boldsymbol{y} \sim$ Poisson$(\boldsymbol{\Phi x})$, then we consider the distribution of $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}$ across different realizations of $\boldsymbol{y}$. Our observations, shown in Fig. 1 are as follows. We make these observations rigorous via Theorem 1.

1. Beyond a threshold τ on the intensity I, the expected value of $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}$ is nearly constant (say some κ), and independent of I, given a fixed number of measurements N. For $I \leq \tau$, we have $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})} \leq \kappa$.
2. The variance of $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}$ is small, irrespective of the value of I and N.
3. For any I, the mean (and any chosen percentile, such as the 99 percentile) of $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}$ scales at a rate lower than $O(N^{0.5})$ w.r.t. N.

4. Irrespective of I, N or m, the distribution of $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}$ is Gaussian with mean and standard deviation equal to the empirical mean and empirical standard deviation of the values of $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}$. This is confirmed by a Kolmogorov-Smirnov (KS) test even at 1% significance (see [23]).

We emphasize that as per our extensive simulations, these properties are independent of specific realizations of $\boldsymbol{\Phi}, \boldsymbol{x}$ or the dimensionality or sparsity of $\boldsymbol{x}$. Our scripts to reproduce these results are included at [23]. Our attempt to formalize these observations lead to the following theorem which we prove in Section 7.

Theorem 1: Let $\boldsymbol{y} \in \mathbb{Z}_+^N$ be a vector of compressive measurements such that $y_i \sim \text{Poisson}[(\boldsymbol{\Phi x})_i]$ where $\boldsymbol{\Phi} \in \mathbb{R}^{N \times m}$ is a non-negative flux-preserving matrix as per Eqn. 9 and $\boldsymbol{x} \in \mathbb{R}^m$ is a non-negative signal. Define $\gamma_i \triangleq (\boldsymbol{\Phi x})_i$. Then we have:

1. $E[\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}] \leq \sqrt{N/4}$
2. If $N \geq \frac{(10.5+11c)(1+2c)}{4c^2}$ and $\forall i, \gamma_i \geq 0.5 + c$ where $c > 0$,
   (a) $\text{Var}[\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}] \leq \dfrac{11N + 5\sum_{i=1}^N 1/\gamma_i}{\sum_{i=1}^N \max(0, 4(2 - 1/\gamma_i))} \leq \dfrac{11}{8} + \dfrac{21}{16c}$
   (b) $P\Big(\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})} \leq \sqrt{N}(\frac{1}{2} + \sqrt{\frac{11}{8} + \frac{21}{16c}})\Big) \geq 1 - 1/N$. This probability can be refined to approximately $1 - 2e^{-N/2}$ when $N \to \infty$, using the Central limit theorem.

We make a few comments below:

1. $E[\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}]$ does not increase with I. This property is *not* shared by the negative log-likelihood of the Poisson distribution. This forms one major reason for using SQJSD as opposed to the latter, for deriving the bounds in this paper.
2. If each γ_i is sufficiently large in value (*i.e.* $\gg 0.5$), the upper bound $\frac{11}{8} + \frac{21}{16c}$ (which is independent of N as well as the measurement or signal values) moves closer towards $\frac{11}{8}$. See also the simulation in Fig. 1. In practice, we have observed this constant upper bound on $\text{Var}[\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}]$ even when the condition $\gamma_i \gg 0.5$ is violated for *all* measurements. Therefore, we consider this condition to be *sufficient* but not *necessary* in practice.
3. If $c = 0.5$ in the above Theorem, then the lower bound on N is $N \geq 32$.
4. The assumption that $\gamma_i \gg 0.5$ is not restrictive in most signal or image processing applications, except those that work with extremely low intensity levels. In the latter case, it should be noted that the performance of Poisson compressed sensing is itself very poor due to the very low SNR [24].
5. The refinement to the probability in the last statement of this theorem is based on the central limit theorem, and hence for a finite value of N, it is an approximation. However, the approximation is empirically observed to be tight even for small $N \sim 10$ as confirmed by a Kolmogorov-Smirnov test even at 1% significance (see [23]).

2.3. Theorem on Reconstruction Error Bounds

Theorem 2: Consider a non-negative signal of interest $\boldsymbol{x} = \boldsymbol{\Psi\theta}$ for orthonormal basis $\boldsymbol{\Psi}$ with sparse vector $\boldsymbol{\theta}$. Define $\boldsymbol{A} \triangleq \boldsymbol{\Phi\Psi}$ for sensing matrix $\boldsymbol{\Phi}$ defined in Eqn. 9. Suppose $\boldsymbol{y} \sim \text{Poisson}(\boldsymbol{\Phi\Psi\theta})$, *i.e.* $\boldsymbol{y} \sim \text{Poisson}(\boldsymbol{A\theta})$, represents a vector of $N \ll m$ independent Poisson-corrupted compressive measurements of $\boldsymbol{x}$, *i.e.*, $\forall i, 1 \leq i \leq N, y_i \sim \text{Poisson}((\boldsymbol{A\theta})_i)$. Let $\boldsymbol{\theta^\star}$ be the solution to the problem (P2) defined earlier, with the upper bound ε in (P2) set to $\sqrt{N}\Big(\frac{1}{2} + \sqrt{\frac{11}{8} + \frac{21}{16c}}\Big)$. If $\tilde{\boldsymbol{\Phi}}$ constructed from $\boldsymbol{\Phi}$ obeys the RIP of order $2s$ with RIC $\delta_{2s} < \sqrt{2} - 1$, if $N \geq \frac{(10.5+11c)(1+2c)}{4c^2}$ with $c > 0$, and if $\boldsymbol{\Phi x} \succeq (1/2 + c)\boldsymbol{1}$, then we have

$$\Pr\Big(\frac{\|\boldsymbol{\theta} - \boldsymbol{\theta^\star}\|_2}{I} \leq \tilde{C}\frac{N}{\sqrt{I}} + \frac{C'' s^{-1/2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_s\|_1}{I}\Big) \geq 1 - 1/N, \tag{13}$$

where $\tilde{C} \triangleq C'(1/2 + \sigma)$, $C' \triangleq \frac{4\sqrt{8(1+\delta_{2s})}}{\sqrt{p(1-p)}(1-(1+\sqrt{2})\delta_{2s})}$, $C'' \triangleq (\frac{2 - 2\delta_{2s} + 2\sqrt{2\delta_{2s}}}{1-(1+\sqrt{2})\delta_{2s}})$, $\boldsymbol{\theta}_s$ is a vector containing the s largest absolute value elements from $\boldsymbol{\theta}$, and σ is the standard deviation of $\sqrt{J(y_i, (\boldsymbol{\Phi x})_i)}$, which is upper bounded by $\sqrt{\frac{11}{8} + \frac{21}{16c}}$.

Theorem 2 is proved in Section 7.2. We make several comments on these bounds below.

1. **Behaviour of $\tilde{C}$ and C''**: Both $\tilde{C}$ and C'' are increasing functions of δ_{2s} over the domain $[0, 1]$.
2. **Value of ε:** Practical implementation of the estimator (P2) would require supplying a value for ε, which is the upper bound on $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}$. This can be provided based on the theoretical analysis of $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}$ from Theorem 1, which motivates the choice $\varepsilon = \sqrt{N}\Big(\frac{1}{2} + \sqrt{\frac{11}{8} + \frac{21}{16c}}\Big)$. In our experiments, we provided a 99 percentile value (see Section 3) which also turns out to be $\mathcal{O}(\sqrt{N})$ and is independent of $\boldsymbol{x}$.
3. **Dependence on I:** We have derived upper bounds on the *relative* reconstruction error, i.e. on $\frac{\|\boldsymbol{\theta} - \boldsymbol{\theta^\star}\|_2}{I}$ and not on $\|\boldsymbol{\theta} - \boldsymbol{\theta^\star}\|_2$. This is because as the mean of the Poisson distribution increases, so does its variance, which would cause an increase in the root mean squared error. But this error would be small in comparison to the average signal intensity. Hence the *relative* reconstruction error is the correct metric to choose in this context. Indeed, $\frac{\|\boldsymbol{\theta} - \boldsymbol{\theta^\star}\|_2}{I}$ is upper bounded by a term that is inversely proportional to I, reflecting the common knowledge that recontruction under Poisson noise is more challenging if the original signal intensity is lower. This is a common feature of Poisson bounds including those in [18]. The second term (due to the compressibility and not full sparsity of the signal) is independent of I. There is no such term in [18] because they have not considered compressible signals.
4. The usage of SQJSD plays a critical role in this proof. First, the term J is related to the Poisson likelihood as will be discussed in Section 5. Second, $\sqrt{J}$ is a metric and hence obeys the triangle inequality. Furthermore, J also upper-bounds the total variation norm, as shown in Lemma 2. Both these properties are essential for the derivation of the critical Step 1 of the proof of Theorem 2 (see Sec. 7.2).

5. **Dependence on N:** It may seem counter-intuitive that the first error term increases with N. There are two reasons for this. First, if the original signal intensity remains fixed at I, an increase in N simply distributes the photon flux across multiple measurements thereby decreasing the SNR at each measurement and degrading the performance. Similar arguments have been made previously in [15], where the error bounds also increase w.r.t. N. This behaviour is a feature of Poisson imaging systems with flux-preserving sensing matrices. The second reason is that the problem P2 has a constrained formulation, quite similar to the the quadratically constrained formulation in [1] (a very fundamental paper in the field of compressed sensing), though modified for Poisson noise. If the error bounds in [1] are applied for the case of $\mathcal{N}(0, \sigma^2)$ noise, they can be proved to scale as $\mathcal{O}(\sigma\sqrt{N})$, i.e. they increase w.r.t. N. Similar arguments have been put forth in Sec. 5.2 of [25] while comparing the quadratically constrained formulation with other estimators. There is currently no consensus in the literature as to whether this $\mathcal{O}(\sigma\sqrt{N})$ behaviour is a fundamental limit on the error bounds of such constrained problems, or whether it is a consequence of the specific proof technique. Nevertheless, it should be borne in mind that like most literature in CS, these are *worst-case* bounds and consider worst-case combinations of signal, sensing matrix and noise values. In practice, the results are much better in comparison to the predicted bounds. Moreover, like most of the literature in CS, the decrease in RIC δ_{2s} (and hence the decrease of $\tilde{C}$ and C") w.r.t. N has been ignored. A precise relationship for the variation of δ_{2s} w.r.t. N has not been derived in the literature and is an open problem, to the best of our knowledge.
6. **Dependence on s:** It may appear that the second term in the error bound decreases with increase in s. However, this is not true, because δ_{2s} and hence both $\tilde{C}$ and C" will increase with s, and hence the upper bound also decreases. This is exactly in tune with earlier work such as [1]. The exact dependence of δ_{2s} on s has not been mathematically established in the literature, to the best of our knowledge.
7. The above bound holds for a signal sparse/compressible in some orthonormal basis $\boldsymbol{\Psi}$. However, for reconstruction bounds for a non-negative signal sparse/compressible in the *canonical* basis, *i.e.* $\boldsymbol{\Psi} = \boldsymbol{I}$ and hence $\boldsymbol{x} = \boldsymbol{\theta}$, one can solve the following optimization problem which penalizes the ℓ_q $(0 < q < 1)$ norm instead of the ℓ_1 norm:

$$\min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_q \text{ subject to} \sqrt{J(\boldsymbol{y}, \boldsymbol{A\theta})} \leq \varepsilon, \|\boldsymbol{\theta}\|_1 = I, \boldsymbol{\theta} \succeq \boldsymbol{0}$$

   Performance guarantees for this case can be developed along the lines of the work in [26]. Other sparsity-promoting terms such as those based on a logarithmic penalty function (which approximates the original ℓ_0 norm penalty more closely than the ℓ_1 norm) may also be employed [27, 28].
8. While imposition of the constraint that $\|\boldsymbol{z}\|_1 = I$ with I being known may appear as a strong assumption, it must be noted that in some compressive camera architectures, it is easy to obtain an estimate

of I during acquisition. One example is the Rice Single Pixel Camera [29], where I can be obtained by turning on all the micro-mirrors, thereby allowing the photo-diode to measure the sum total of all values in the signal. The imposition of this constraint has been considered in earlier works on Poisson compressed sensing. Furthermore, we note that in our experiments in Section 3, we have obtained excellent reconstructions even without using this constraint.

9. Measurement matrices in compressed sensing can be specifically designed to have very low coherence, as opposed to the choice of random matrices. Such approaches have been proposed for a Poisson setting in [20]. Since the coherence value can be used to put an upper bound on the RIC, one can conclude that such matrices will obey RIP even while obeying non-negativity and flux preservation. In case of such matrices which already obey the RIP, the upper bound on the reconstruction error would potentially tighten by a factor of at least $\sqrt{N}$. However, such matrices are obtained as the output of non-convex optimization problems, and there is no guarantee on how low their coherence, and hence their RIC, will be. Indeed, they may not respect the sufficient condition in our proof that $\delta_{2s} < \sqrt{2}-1$. Matrices can also be designed based on an MSE optimization criterion [30, 31] for excellent performance. If the noise is Gaussian and the signal is assumed to be a sample from the Gaussian mixture model, the estimate of the signal from the compressive measurements can be obtained via a modified Wiener filter in closed form (which is also the MAP or MMSE estimate). Moreover the expected MSE between the estimate and the true signal also has a closed form expression. One can perform a descent on this expression in order to design better sensing matrices for compressed sensing, as has been accomplished in [30, 31]. This closed form is no longer applicable if the noise is Poisson and hence extending this work to the Poisson setting is challenging.

2.4. Advantages of SQJSD over Poisson NLL or ℓ_2 difference

Here, we summarize the essential advantage of the SQJSD over the Poisson NLL derived from Eqn. 4 or the ℓ_2 difference, i.e. $\|\boldsymbol{y} - \boldsymbol{A\theta}\|_2$. Estimators based on the Poisson NLL require regularization parameters [32] or constraint parameters [18] that are signal dependent and hence very difficult to tune in practice as the underlying signal is unknown. In contrast, our SQJSD-based estimator (P2) uses a value of ε based on the signal-independent tail bounds of $\sqrt{J(\boldsymbol{y}, \boldsymbol{A\theta})}$. A more detailed comparison with previous methods based on NLL is presented in Section 4.

It is also natural to question how (P2) would compare to an estimator of the following form, which we name P2-L2: $\min \|\boldsymbol{\theta}\|_1$ s. t. $\|\boldsymbol{y} - \boldsymbol{A\theta}\|_2 \leq \tilde{\varepsilon}, \|\boldsymbol{\Psi\theta}\|_1 = I, \boldsymbol{\Psi\theta} \succeq \boldsymbol{0}$. In problem (P2-L2), the tail bound $\tilde{\varepsilon}$ would be clearly signal-dependent as $\mathrm{Var}(y_i) = E(y_i) = (\boldsymbol{\Phi x})_i = (\boldsymbol{A\theta})_i$, unlike in problem (P2). This is a major disadvantage of (P2-L2) as compared to (P2). One could counter-argue in the following manner: (a) For the forward model used in this paper, we have $(\boldsymbol{A\theta})_i \leq I/N$, which imposes an upper bound on the measurement variance. This can be used to put a tail bound on $\tilde{\varepsilon}$ either using a Gaussian approximation

for the elements of $\boldsymbol{y} - \boldsymbol{A\theta}$, or else via Chebyshev's inequality. (b) Moreover, both (P2) as well as (P2-L2) impose the constraint $\|\boldsymbol{\Psi\theta}\|_1 = I$ which is necessary for the theoretical proofs.

This counter-argument however misses two important points. (1) First, in practice while implementing (P2), this constraint is not required as stated before and in Section 3. (2) Second, the tail bound for $\tilde{\varepsilon}$ used in this manner in a practical implementation of P2-L2 will be loose since the values of $(\boldsymbol{A\theta})_i$ (which are of course, unknown) could be significantly less than I/N.

Instead of the term $\|\boldsymbol{y}-\boldsymbol{A\theta}\|_2$ in (P2-L2), one could however consider the term $L(\boldsymbol{y}, \boldsymbol{A\theta}) \triangleq \|(\boldsymbol{y}-\boldsymbol{A\theta})./\sqrt{\boldsymbol{A\theta}}\|_2$ where './' indicates element-wise division. We conjecture and have experimentally observed that tail-bounds based on $L(\boldsymbol{y}, \boldsymbol{A\theta})$ are signal-independent. However we can easily prove that for any i, $E[(y_i - (\boldsymbol{A\theta})_i)^2/(\boldsymbol{A\theta})_i] = 1$ and $\mathrm{Var}[(y_i - (\boldsymbol{A\theta})_i)^2/(\boldsymbol{A\theta})_i] = E[(y_i - (\boldsymbol{A\theta}_i))^4/((\boldsymbol{A\theta})_i)^2] - E^2[(y_i - (\boldsymbol{A\theta})_i)^2/(\boldsymbol{A\theta})_i] = 2 + 1/(\boldsymbol{A\theta})_i$. These are greater than the corresponding values for the JSD, as can be seen from the proof of Theorem 1 in the Appendix (see Eqns. 26 and 33). This leads us to conjecture that the bounds with the SQJSD will be tighter. An estimator using $L(\boldsymbol{y}, \boldsymbol{A\theta})$ is essentially a normalized form of the LASSO. Experimental results with it have been shown in [18] and its sign-consistency has been analyzed in [33]. But there is no work which presents signal estimation bounds with it, to the best of our knowledge. At this point, we consider the complete development of an estimator using $L(\boldsymbol{y}, \boldsymbol{A\theta})$ to be beyond the scope of this paper, as it is non-trivially different from estimators based on SQJSD, Poisson NLL or ℓ_2 difference.

3. Numerical Experiments

Generation of Test Measurements: Experiments were run on Poisson-corrupted compressed measurements obtained from each signal taken from an ensemble of 1D signals with 100 elements. Each signal $\boldsymbol{x} = \boldsymbol{\Psi\theta}$ in the ensemble was generated using sparse linear combinations of DCT basis vectors, and was forced to be non-negative by adjusting the DC component. The support sets of the sparse coefficients were randomly selected, and different signals had different supports. In some experiments (see later in this section), all the signals were normalized so that they had a fixed intensity I. The sensing matrix followed the architecture discussed in Section 2.

Comparisons: We show results on numerical experiments for problem (P2). We omitted the explicit constraint that $\|\boldsymbol{\Psi\theta}\|_1 = I$, as its inclusion did not affect the results significantly (see Fig. 8), and refer to it simply as (P2) in this section. We compared our results with those obtained using the following estimators, all without the $\|\boldsymbol{\Psi\theta}\|_1 = I$ constraint:

1. A regularized version of P2-L2 (from the previous section) referred to here as (P2-L4):
   argmin $\rho\|\boldsymbol{\theta}\|_1 + \|\boldsymbol{y} - \boldsymbol{\Phi\Psi\theta}\|_2^2$ s.t. $\boldsymbol{\Psi\theta} \succeq \boldsymbol{0}$.
2. A regularized estimator using the Poisson NLL, referred to here as (P-NLL):
   argmin $\rho\|\boldsymbol{\theta}\|_1 + \sum_{i=1}^{N}[\boldsymbol{\Phi\Psi\theta}]_i - y_i \log[\boldsymbol{\Phi\Psi\theta}]_i$ s.t. $\boldsymbol{\Psi\theta} \succeq \boldsymbol{0}$.

3. A regularized version of (P2), referred to here as (P4):
   argmin $\rho\|\boldsymbol{\theta}\|_1 + J(\boldsymbol{y}, \boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{\theta})$ s.t. $\boldsymbol{\Psi}\boldsymbol{\theta} \succeq \mathbf{0}$.
4. An estimator P-VST of the following form based on our work on variance stabilization transforms for Poisson noise [34] (see also Sec. 4.3): argmin $\rho\|\boldsymbol{\theta}\|_1 + \|\sqrt{\boldsymbol{y} + 3/8} - \sqrt{\boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{\theta} + 3/8}\|_2^2$ s.t. $\boldsymbol{\Psi}\boldsymbol{\theta} \succeq \mathbf{0}$. Note: In [34], we have analyzed the theoretical properties of a constrained version of P-VST.

In all of these, ρ is a regularization parameter. Before describing our actual experimental results, we state a lemma which shows that solving (P4) is equivalent to solving (P2) for some pair of (ρ, ε) values, but again without the constraint $\|\boldsymbol{\Psi}\boldsymbol{\theta}\|_1 = I$. The proof of this lemma follows [35] and can be found in the supplemental material.

Lemma 3: Given $\boldsymbol{\theta}$ which is the minimizer of problem (P4) for some $\rho > 0$, there exists some value of $\varepsilon = \varepsilon_\theta$ for which $\boldsymbol{\theta}$ is the minimizer of problem (P2), but without the constraint $\|\boldsymbol{\Psi}\boldsymbol{\theta}\|_1 = I$.

Note that despite this equivalence, in practice we preferred (P2) over (P4) as selection of ρ poses practical difficulties, as opposed to the statistically motivated choice for ε.

Implementation Packages: As JSD is a convex function and $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi}\boldsymbol{x})} \leq \varepsilon$ implies $J(\boldsymbol{y}, \boldsymbol{\Phi}\boldsymbol{x}) \leq \varepsilon^2$, we solved both (P2) and (P4) using the well-known CVX package [36] with the SCS solver for native implementation of logarithmic functions[4]. Likewise, (P2-L4) and P-VST were also implemented using CVX. For (P-NLL), we used the SPIRAL-TAP algorithm from [32].

Parameter Choices and Description of Experiments: In all experiments, the value of ε for (P2) was chosen as per the tail bounds on SQJSD, which are independent of $\boldsymbol{x}$ as noted in Section 2.2. To be specific, we set $\varepsilon = \sqrt{N/2}$ for *all experiments* with no further tweaking whatsoever. This value is lower than what is predicted by our theoretical analysis which deals with *worst case bounds*, and gave us better results. We also report results for (P2) with ε set to the 99-percentile of the SQJSD values, computed empirically on an arbitrary set of signals and their compressive measurements. This is perfectly principled, because we know that the statistics of SQJSD depend only on N and not on the (unknown) signals. For (P-NLL), (P2-L4), (P4) and P-VST, the value of ρ was chosen by cross-validation (CV). For (P-NLL), the optimization was run for a maximum of 500 iterations, which was more than the default parameter of 100 specified in the associated package [32]. We ran a total of three experiments on each of the competing methods. The comparison metric was the relative reconstruction error given as $RRMSE(\boldsymbol{x}, \boldsymbol{x}^\star) \triangleq \dfrac{\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2}{\|\boldsymbol{x}\|_2}$ where $\boldsymbol{x}^\star$ is the estimate of $\boldsymbol{x}$. In the *intensity experiment*, we studied the effect of change in signal intensity I on the RRMSE, keeping the signal sparsity s fixed to 10 (out of 100 elements) and $N = 50$. For CV, the parameter ρ was chosen to be the parameter from the set $\mathcal{PV} \triangleq \{10^{-7}, 10^{-6}, ..., 10^{-2}, 0.1, 1\}$ which yielded the best RRMSE reconstruction of an ensemble of synthetic signals with sparsity $s = 10$ and $I = 1000$, from

[4] `http://web.cvxr.com/cvx/beta/doc/solver.html`

$N = 50$ compressive measurements. We also separately report results when ρ was chosen **omnisciently** (i.e. we used the value of ρ from a chosen range, that yielded the best signal reconstruction results in terms of RRMSE, assuming the ground truth was known). In the *experiment on number of measurements*, we studied the effect of change in N on the RRMSE, keeping $I = 10^6$ and $s = 10$ fixed. For CV, the parameter ρ was chosen to be the parameter from the set $\mathcal{PV}$ which yielded the best RRMSE reconstruction of an ensemble of synthetic signals with sparsity $s = 10$, $I = 10^6$ from $N = 20$ compressive measurements. We also separately report results when ρ was chosen omnisciently. In the *signal sparsity experiment*, we studied the effect of change in s on the RRMSE, keeping $I = 10^6$ and $N = 50$ fixed. For CV, the parameter ρ was chosen to be the parameter from the set $\mathcal{PV}$ which yielded the best RRMSE reconstruction of an ensemble of synthetic signals with sparsity $s = 40$, $I = 10^6$ from $N = 50$ compressive measurements. Again, we also separately report results when ρ was chosen omnisciently.

Observations and Comments: The results (i.e. average RRMSE values computed over $Q = 100$ signals) for the intensity experiment, the experiment on N and the sparsity experiment are respectively presented in Figs. 2, 3 and 4. Note that the best tuning parameters ρ for (P2-L4) and (P-NLL) are *signal-dependent*. As can be seen from the plots, an omniscient choice of ρ (defined as the value of ρ from a chosen range, that yields the best signal reconstruction results in terms of RRMSE, assuming the ground truth is known) for (P4), (P-NLL), (P2-L4) and P-VST no doubt improves their performance (as it would also for (P2)). However an omniscient choice is not practical, and improper choice of ρ indeed adversely affects the performance of (P4), (P-NLL), (P2-L4), and P-VST[5] CV-based methods can help, but here again they require some prior knowledge of signal properties in order to be effective. *Moreover, a very important point to be noted here is that for (P2), we have a statistically consistent and signal independent parameter ε. The methods (P4), (P-NLL), (P2-L4) do not have this benefit.* From the plots for (P2) in Fig. 2, we observe that the RRMSE decreases on an average with increase in I. We would have observed such a trend even with (P-NLL) and (P2-L4) with omnisciently picked parameters or CV procedures that require a priori knowledge of signal properties such as intensity or sparsity, but that is not practical. From the plots for (P2) in Fig. 3, we observe that the RRMSE is not always guaranteed to decreases on an average with increase in N, owing to the flux-preserving nature of $\mathbf{\Phi}$ which causes poorer SNR with increase in N. The results for the sparsity experiment in Fig. 4, we see that the RRMSE can increase with increase in s. All these trends are in line with our worst case bounds.

Low signal intensity: We ran experiments using P2 with $\varepsilon = \sqrt{N/2}$ and ε set to the empirically computed 99-percentile of the SQJSD values. The experiment were run for compressive measurements of

[5] At cursory glance, the results of P-VST in our work in [34] may appear to different from those reported in this paper. However several settings are different in the two papers. For example, in [34], many experiments have been run at higher intensity levels, and that paper also contains experiments on a contrained version of P-VST.

a DCT-sparse signal with $s = 10, I = 10^3, m = 100$ with flux-preserving sensing matrices (with $p = 0.5$) and with $N \in \{10, 20, ..., 100\}$. Thus in expectation, the maximum value of the noiseless measurements was upper bounded by 5 when $N = 100$. In such low-intensity regimes, we indeed observe from the results plotted in Fig 5, that the RRMSE increases with increase in N as predicted by the bounds. As seen, the RRMSE increase is much sharper for $\varepsilon = \sqrt{N/2}$ than for the 99 percentile of SQJSD, because the latter values were lower than $\sqrt{N/2}$. This is because our theoretical bounds are *worst case*, and the empirical results can thus be better than what is predicted by the bounds.

Image Reconstruction Experiments: We also tested the performance of all competing methods on an image reconstruction task from compressed measurements under Poisson noise. Each patch of size 7×7 from a gray-scale image was vectorized and 25 Poisson-corrupted measurements were generated for this patch using the sensing matrix discussed in Section 2. This model is based on the architecture of the compressive camera designed in [37, 38] except that we considered overlapping patches here. Each patch was reconstructed from its compressed measurements independently by solving (P2) with sparsity in a 2D-DCT basis, with $\varepsilon = \sqrt{N/2}$. The final image was reconstructed by averaging the reconstructions of overlapping patches (which is similar to running a deblocking algorithm on reconstructions from non-overlapping patches). This experiment was repeated for different I values by suitably rescaling the intensities of the original image before simulation of the compressive measurements. In Fig. 6, we show reconstruction results with (P2) with $\varepsilon = \sqrt{N/2}$ under different values of I. There is a sharp decrease in relative reconstruction error with increase in I. For (P4), (P-NLL) and (P2-L4), the ρ parameter was picked omnisciently on a small set of patches at a fixed intensity level of $I = 10^5$ and used for all other intensities. For these experiments, we observed nearly identical numerical results with (P4), (P-NLL) and (P2-L4), as with (P2) with a fixed $\varepsilon = \sqrt{N/2}$. However, for the lowest intensity level of $I = 10^5$, we observed that (P2) produced a lower RRMSE than (P-NLL) (0.13 as against 0.18).

The constraint $\|\boldsymbol{x}^\star\|_1 = I$: Note that in our experiments, we have not made use of the hard constraint $\|\boldsymbol{x}^\star\|_1 = I$ in problems (P2) or in any of the competing methods (P4), (P-NLL), (P2-L4),P-VST . In practice, we however observed that the estimated $\|\boldsymbol{x}^\star\|_1$ was close to the true I, especially for higher values of $I \geq 10^6$, and moreover even imposition of the constraint did not significantly alter the results as can be seen in Fig. 7 for a 100-dimensional signal with 50 measurements and sparsity 5.

Computational Complexity: Also, to get an idea of the computational complexity of various estimators, we plot a graph (Fig. 8) of the average reconstruction time (across noise instances) till convergence w.r.t. m with $N = m/2$ each time, and also w.r.t. N keeping m fixed. The experiments were run for signals with $s = 10, I = 10^6$. For the former plot, we chose m ranging from 100 to 3500, with $N = m/2$ measurements in each case. For the latter plot, we chose signals with $m = 100$. From the plots, it appears that (P2) is more time-intensive than other estimators, including (P4). However there are two reasons for this apparent

trend. First, we ran (P4), (P-NLL),(P-VST) and (P2-L4) with a fixed value of ρ and hence the time for cross-validation is ignored. Second, it is well-known that constrained formulations such as (P2) are often implemented using their corresponding unconstrained formulations (i.e. (P4) here), and are hence less efficient. Such arguments have been made in [39] in the context of support vector machines.

Handling zero-valued measurements: Note that zero-valued measurements pose no problem, given the definition of $J(\boldsymbol{y}, \boldsymbol{A\theta})$ as in Section 2.2.

Summary: All the numerical experiments in this section confirm the efficacy of using the JSD/SQJSD in Poisson compressed sensing problems. In particular, the statistical properties of the SQJSD allow for compressive reconstruction with statistically motivated parameter selection, unlike methods based on the Poisson negative log-likelihood which require tweaking of the regularization/signal sparsity parameter.

Reproducible Research: Our supplemental material at `https://www.cse.iitb.ac.in/~ajitvr/SQJSD/` contains scripts for execution of these results in CVX.

4. Relation to Prior Work

There exist excellent algorithms for Poisson reconstruction such as [28, 6, 40, 41], but these methods do not provide performance bounds. In this section, we put our work in the context of existing work on Poisson compressed sensing with theoretical performance bounds. These techniques are based on one of the following categories: (a) optimizing either the Poisson negative log-likelihood (NLL) along with a regularization term, or (b) the LASSO, or (c) using constraints motivated by variance stabilization transforms (VST).

4.1. Comparison with Poisson NLL based methods

These methods include [15, 24, 42, 18, 19, 43]. One primary advantage of the SQJSD-based approach over the Poisson NLL is that the former (unlike the latter) is a metric, *and* can be bounded by values independent of I as demonstrated in Section 2.2. In principle, this allows for an estimator that in practice does not require tweaking a regularization or signal sparsity parameter, and instead requires a statistically motivated bound ε to be specified, which is more intuitive. Moreover, the methods in [15, 24] (and their extensions to the matrix completion problem in [44, 45, 46]) employ ℓ_0-regularizers for the signal, due to which the derived bounds are applicable only to computationally intractable estimators. The results in both papers have been presented using estimators with ℓ_1 regularizers with the regularization parameters (as in [15]) or signal sparsity parameter (as in [24]) chosen omnisciently, but the derived bounds are not applicable for the implemented estimator. In contrast, our approach proves error bounds with the ℓ_1 sparsity regularizer for which efficient and tractable algorithms exist. Moreover, the analysis in [24] is applicable to exactly sparse signals, whereas our work is applicable to signals that are sparse or compressible in any orthonormal basis. Recently, NLL-based tractable minimax estimators have been presented in [18], but knowledge of an upper

Method	Objective Function
This paper	Problem (P2) from Section 1.1, with ε chosen using properties of the SQJSD
[15]	$\text{NLL}(\boldsymbol{y}, \boldsymbol{\Phi x}) + \rho \text{pen}(\boldsymbol{\Psi}^T \boldsymbol{x})$ such that $\boldsymbol{x} \succeq \boldsymbol{0}, \|\boldsymbol{x}\|_1 = I$ where $\text{pen}(\boldsymbol{\Psi}^T \boldsymbol{x}) = \|\boldsymbol{\Psi}^T \boldsymbol{x}\|_0$
[24]	$\text{NLL}(\boldsymbol{y}, \boldsymbol{\Phi x})$ such that $\boldsymbol{x} \succeq \boldsymbol{0}, \|\boldsymbol{x}\|_1 = I, \|\boldsymbol{\Psi}^T \boldsymbol{x}\|_0 \leq s$
[18]	$\text{NLL}(\boldsymbol{y}, \boldsymbol{\Phi x})$ such that $\boldsymbol{x} \succeq \boldsymbol{0}, \|\boldsymbol{\Psi}^T \boldsymbol{x}\|_1 \leq s$
[47]	$\|\boldsymbol{y} - \boldsymbol{\Phi x}\|^2 + \rho \|\boldsymbol{\Psi}^T \boldsymbol{x}\|_1$ such that $\boldsymbol{x} \succeq \boldsymbol{0}, \|\boldsymbol{x}\|_1 = I$; ρ is chosen as $\mathcal{O}(1/I)$
[34]	$\|\boldsymbol{\Psi}^T \boldsymbol{x}\|_1$ such that $\|\sqrt{\boldsymbol{y}} - \sqrt{\boldsymbol{\Phi x}}\|_2 \leq \varepsilon, \boldsymbol{x} \succeq \boldsymbol{0}, \|\boldsymbol{x}\|_1 = I$ with ε picked based on chi-square tail bounds
[17]	$\|\boldsymbol{y} - \boldsymbol{\Phi x}\|^2 + \rho \sum_k d_k \lvert(\boldsymbol{\Psi}^T \boldsymbol{x})_k\rvert$, with weights d_k picked statistically
[48]	$\|\boldsymbol{\Psi}^T \boldsymbol{x}\|_1$ such that $\text{NLL}(\boldsymbol{y}, \boldsymbol{\Phi x}) \leq \varepsilon$ where no criterion to choose ε is analyzed

Table 1: Objective functions optimized by various Poisson compressed sensing methods. Note that $\boldsymbol{\Psi}$ refers to an orthonormal signal basis.

bound on the signal sparsity parameter (ℓ_q norm of the signal, $0 < q \leq 1$) is required for the analysis, even if the sensing matrix were to obey the RIP. A technique for deriving a regularization parameter to ensure statistical consistency of the ℓ_1-penalized NLL estimator has been proposed in [19], but that again requires knowledge of the signal sparsity parameter. In our work, the constraint $\|\boldsymbol{x}\|_1 = I$ was required only due to the specific structure of the sensing matrix, and even there, it was not found to be necessary in practical implementation. For clarity the specific objective functions used in these techniques is summarized in Table 4.1. The work in [42] deals with a *specific* type of sensing matrices called the expander-based matrices, unlike the work in this paper which deals with any randomly generated matrices of the form Eqn. 9, and the bounds derived in [42] are only for signals that are sparse in the *canonical* basis. In [43], performance bounds are derived *in situ* with system calibration error estimates for *multiple* measurements, which is essentially a different computational problem, which again requires knowledge of regularization parameters.

4.2. Comparison with LASSO-based methods

These methods include [16, 17, 49, 33, 50, 48] and are based on optimization of a convex function of the form $\sum_{i=1}^{N} (y_i - [\boldsymbol{\Phi x}]_i)^2 + \rho \|\boldsymbol{\Psi}^T \boldsymbol{x}\|_1$. The performance of the LASSO (designed initially for homoscedastic noise) under heterscedasticity associated with the Poisson noise model is examined in [33] and necessary and sufficient conditions are derived for the sign consistency of the LASSO. Weighted/adaptive LASSO and group LASSO schemes with provable guarantees based on Poisson concentration inequalities have been proposed in [16, 17]. Group LASSO based bounds have also been derived in [49] and applied to Poisson regression. Bounds on recovery error using an ℓ_1 penalty are derived in [48] and [50] based on the RIP and maximum eigenvalue condition respectively. These techniques do not provide bounds for realistic physical constraints

in the form of flux-preserving sensing matrices. The quantity ε is not analyzed theoretically in [48] unlike in our method - see Table 4.1. Moreover the LASSO is not a probabilistically motivated (i.e. penalized likelihood based) estimator for the case of Poisson noise. Even considering an approximation of Poisson(λ) by $\mathcal{N}(\lambda, \lambda)$, the approximated likelihood function would be $K(\boldsymbol{y}, \boldsymbol{\Phi x}) \triangleq \sum_{i=1}^{N} \frac{(y_i - [\boldsymbol{\Phi x}]_i)^2}{[\boldsymbol{\Phi x}]_i} + \log[\boldsymbol{\Phi x}]_i$ and not $\sum_{i=1}^{N}(y_i - [\boldsymbol{\Phi x}]_i)^2$ as considered in the LASSO. While $K(\boldsymbol{y}, \boldsymbol{\Phi x})$ is nonconvex, $J(\boldsymbol{y}, \boldsymbol{\Phi x})$ is a convex function. Moreover $J(\boldsymbol{y}, \boldsymbol{\Phi x})$ is a lower bound on $K(\boldsymbol{y}, \boldsymbol{\Phi x})$ if $[\boldsymbol{\Phi x}]_i \geq 1$. This is shown in Eqn. 25 while proving Theorem 1. Therefore our SQJSD method provides a tractable way to implement an estimator using $K(\boldsymbol{y}, \boldsymbol{\Phi x})$ if the parameter ε is chosen based on the statistics of $\sqrt{K(\boldsymbol{y}, \boldsymbol{\Phi x})}$.

4.3. Comparison with VST-based methods

VST-based methods, especially those based on variants of the square-root transformations, have been used extensively in denoising [51] and deblurring [52] under Poisson noise, but without performance bounds. In the context of Poisson CS, the VST converts a linear problem into a non-linear one via a square-root transform. The basic motivation is that if $y \sim$ Poisson(λ), then $\sqrt{y + 3/8} \sim \mathcal{N}(\sqrt{\lambda + 3/8}, 1/4)$ approximately, with improvement in the quality of approximation to the Gaussian distribution as well as the fixed variance when λ is large. Our group has recently shown [53, 34] that despite the conversion to non-linear measurements, this non-linear regression via a data fidelity of the form $\|\sqrt{\boldsymbol{y} + 3/8} - \sqrt{\boldsymbol{\Phi x} + 3/8}\|_2$ has various advantages for Poisson CS reconstructions, with a similar statistically motivated parameter (ε) selection, as for the SQJSD. The specific estimator developed there is as follows:

$$\min \|\boldsymbol{\theta}\|_1 \text{ such that } \|\sqrt{\boldsymbol{y} + 3/8} - \sqrt{\boldsymbol{\Phi\Psi\theta} + 3/8}\|_2 \leq \epsilon, \boldsymbol{\Psi\theta} \succeq \boldsymbol{\theta}. \tag{14}$$

There are many similarities as well as difference between the properties of the SQJSD estimator in this paper and the aforementioned constrained formulation based on the VST from [34]. Let us denote the data fidelity term as $g(\boldsymbol{y})$ (see proof of Theorem 1), which is given by SQJSD in this paper and by $\|\sqrt{\boldsymbol{y} + 3/8} - \sqrt{\boldsymbol{\Phi x} + 3/8}\|_2$ in [34]. In both cases, we have shown that $E[g(\boldsymbol{y})]$ is $\mathcal{O}(\sqrt{N})$ and that Var$[g(\boldsymbol{y})]$ is independent of the signal and N. In both cases, we require a small lower bound on the magnitude of the underlying measurement, for the theory to hold, but the experiments reveal that the lower bound is not strictly required. In terms of numerical simulations shown in this paper, we also see cases when both (P2) and omniscient (P-VST) outperform each other. However there are important differences as well. We conjecture based on empirical simulation that the expected value of SQJSD is *smaller than* $\mathcal{O}(\sqrt{N})$ (see also Fig. 1), whereas the same does not hold for our VST-based work. This may help tighten our SQJSD bounds further. On the other hand, our VST-based work has been extended to handle Poisson-Gaussian noise. A full exploration of CS with Poisson-Gaussian noise using the SQJSD (or a modified form of SQJSD) is beyond the scope of this paper. However most importantly, in this paper, we have presented the interesting result that *the SQJSD also possesses variance stabilizing properties for the Poisson distribution. To the best of our knowledge, there*

is no prior literature in statistics or signal processing reporting such a result. Apart from this, the data fidelity term in the VST-based work is essentially a negative log *quasi*-likelihood, whereas the SQJSD is related to a *symmetrized version* of the Poisson negative log-likelihood (as shown in Sec. 5).

5. Relation between the JSD and a Symmetrized Poisson Negative Log Likelihood

In this section, we demonstrate the relationship between the JSD and an approximate symmetrized version of the Poisson negative log likelihood function. As such, this relationship does not affect the performance bounds, but is interesting in its own right. Consider an underlying noise-free signal $\boldsymbol{x} \in {\mathbb{R}_+}^{m\times 1}$. Consider that a compressive sensing device acquires $N \ll m$ measurements of the original signal $\boldsymbol{x}$ to produce a measurement vector $\boldsymbol{y} \in {\mathbb{Z}_+}^{N\times 1}$. Assuming independent Poisson noise in each entry of $\boldsymbol{y}$, we have $\forall i, 1 \leq i \leq N, y_i \sim \text{Poisson}(\boldsymbol{\Phi x})_i$, where as considered before, $\boldsymbol{\Phi}$ is a non-negative flux-preserving sensing matrix. The main task is to estimate the original signal $\boldsymbol{x}$ from $\boldsymbol{y}$. A common method is to maximize the following likelihood in order to infer $\boldsymbol{x}$:

$$\begin{aligned}\mathcal{L}(\boldsymbol{y}|\boldsymbol{\Phi x}) &= \prod_{i=1}^{N} p(y_i|(\boldsymbol{\Phi x})_i) \\ &= \prod_{i=1}^{N} \frac{(\boldsymbol{\Phi x})_i{}^{y_i}}{y_i!} e^{-(\boldsymbol{\Phi x})_i}.\end{aligned} \tag{15}$$

The negative log-likelihood $\mathcal{NLL}$ can be approximated as:

$$\mathcal{NLL}(\boldsymbol{y}, \boldsymbol{\Phi x}) \approx \sum_{i=1}^{N} y_i \log \frac{y_i}{(\boldsymbol{\Phi x})_i} - y_i + (\boldsymbol{\Phi x})_i + \frac{\log y_i}{2} + \frac{\log 2\pi}{2}. \tag{16}$$

This expression stems from the Stirling's approximation [54] for $\log y_i!$ given by

$$\log y_i! \approx y_i \log y_i - y_i + \frac{\log y_i}{2} + \frac{\log 2\pi}{2}. \tag{17}$$

This is derived from Stirling's series given below as follows for some integer $n \geq 1$:

$$n! \approx \sqrt{2\pi n}\big(\frac{n}{e}\big)^n \big(1 + \frac{1}{12n} + \frac{1}{288n^2}\big) \approx \sqrt{2\pi n}\big(\frac{n}{e}\big)^n. \tag{18}$$

Consider the generalized Kullback-Leibler divergence between $\boldsymbol{y}$ and $\boldsymbol{\Phi x}$, denoted as $G(\boldsymbol{y}, \boldsymbol{\Phi x})$ and defined as

$$G(\boldsymbol{y}, \boldsymbol{\Phi x}) \triangleq \sum_{i=1}^{N} y_i \log \frac{y_i}{(\boldsymbol{\Phi x})_i} - y_i + (\boldsymbol{\Phi x})_i. \tag{19}$$

The generalized Kullback-Leibler divergence turns out to be the Bregman divergence for the Poisson noise model [55] and is used in maximum likelihood fitting and non-negative matrix factorization under the Poisson

noise model [13]. The negative log-likelihood can be expressed in terms of the generalized Kullback-Leibler divergence in the following manner:

$$\mathcal{NLL}(\boldsymbol{y}, \boldsymbol{\Phi x}) \approx G(\boldsymbol{y}, \boldsymbol{\Phi x}) + \sum_{i=1}^{N} \Big(\frac{\log y_i}{2} + \frac{\log 2\pi}{2} \Big). \tag{20}$$

Let us consider the following symmetrized version of the $\mathcal{NLL}$:

$$\begin{aligned} \mathcal{SNLL}(\boldsymbol{y}, \boldsymbol{\Phi x}) = \mathcal{NLL}(\boldsymbol{y}, \boldsymbol{\Phi x}) + \mathcal{NLL}(\boldsymbol{\Phi x}, \boldsymbol{y}) &\approx G(\boldsymbol{y}, \boldsymbol{\Phi x}) + G(\boldsymbol{\Phi x}, \boldsymbol{y}) + \sum_{i=1}^{N} \Big(\frac{\log y_i}{2} + \frac{\log (\boldsymbol{\Phi x})_i}{2} + \log 2\pi \Big) \\ &\geq G(\boldsymbol{y}, \boldsymbol{\Phi x}) + G(\boldsymbol{\Phi x}, \boldsymbol{y}) = D(\boldsymbol{y}, \boldsymbol{\Phi x}) + D(\boldsymbol{\Phi x}, \boldsymbol{y}), \end{aligned} \tag{21}$$

where D is the Kullback-Leibler divergence from Eqn. 10. The inequality above is true when the term in parantheses is non-negative, which is true when either (1) for each i, we must have $y_i \geq \dfrac{1}{4\pi^2 (\boldsymbol{\Phi x})_i}$, or (2) the *minimum* value for $y_i \geq d \triangleq \dfrac{1}{4\pi^2 \big(\prod_{i=1}^{N} (\boldsymbol{\Phi x})_i \big)^{(1/N)}}$. We collectively denote these conditions as 'Condition 1' henceforth. Note that, given the manner in which $\boldsymbol{\Phi}$ is constructed, we have the guarantee that $(\boldsymbol{\Phi x})_i \geq \dfrac{x_{min}}{N}$ with a probability of $1 - Np^m$ where x_{min} is the minimum value in $\boldsymbol{x}$. The quantity on the right hand side of the last equality above follows from Eqns. 10 and 19, and yields a symmetrized form of the Kullback-Leibler divergence $D_s(\boldsymbol{y}, \boldsymbol{\Phi x}) \triangleq D(\boldsymbol{y}, \boldsymbol{\Phi x}) + D(\boldsymbol{\Phi x}, \boldsymbol{y})$. Now, we have the following useful lemma giving an inequality relationship between D_s and J, the proof of which follows [56] and can be found in the supplemental material.

Lemma 4: Given non-negative vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, we have $\dfrac{1}{4} D_s(\boldsymbol{u}, \boldsymbol{v}) \geq J(\boldsymbol{u}, \boldsymbol{v})$.

Combining Eqns. 22 and Lemma 4, we arrive at the following conclusion if 'Condition 1' holds true:

$$\mathcal{SNLL}(\boldsymbol{y}, \boldsymbol{\Phi x}) \leq \varepsilon \implies J(\boldsymbol{y}, \boldsymbol{\Phi x}) \leq \varepsilon/4 \implies \sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})} \leq \varepsilon' \triangleq \sqrt{\varepsilon}/2. \tag{22}$$

Let us consider the following optimization problem:

$$\text{(P3): minimize} \|\boldsymbol{z}\|_1 \text{ such that } \mathcal{SNLL}(\boldsymbol{y}, \boldsymbol{Az}) \leq \varepsilon, \boldsymbol{\Psi z} \succeq \boldsymbol{0}, \|\boldsymbol{\Psi z}\|_1 = I. \tag{23}$$

Following Eqn. 22, we observe that a solution to (P2) is also a solution to (P3) if the parameter ε is chosen based on the statistics of $\sqrt{\mathcal{SNLL}(\boldsymbol{y}, \boldsymbol{Az})}$. Note that Condition 1 can fail with higher probability if $(\boldsymbol{\Phi x})_i$ is small, due to which the $J \leq \mathcal{SNLL}$ bound may no longer hold. However, this does not affect the validity of Theorems 1 or 2, or the properties of the estimator proposed in this paper. Note that we choose to solve (P2) instead of (P3) in this paper, as the SQJSD is a metric unlike $\mathcal{SNLL}$, which makes it easier to establish theoretical bounds using SQJSD. Also, there is no literature on the statistical properties of $\sqrt{\mathcal{SNLL}(\boldsymbol{y}, \boldsymbol{Az})}$, established so far.

6. Conclusion

In this paper, we have presented new upper bounds on the reconstruction error from compressed measurements under Poisson noise in a realistic imaging system obeying the non-negativity and flux-preservation constraints, for a *computationally tractable* estimator using the ℓ_1 norm sparsity regularizer. Our bounds are easy to derive and follow the skeleton of the technique laid out in [1]. The bounds are based on the properties of the SQJSD from Section 2.2, of which some, such as signal-independent mean and variance, are derived in this paper. Our bounds are applicable to sparse as well as compressible signals in any chosen orthonormal basis. We have presented numerical simulations with parameters chosen based on noise statistics (unlike the choice of regularization or signal sparsity parameters in other techniques), demonstrating the efficacy of the method in reconstruction from compressed measurements under Poisson noise. We observe that the derived upper bounds decrease with an increase in the original signal flux, i.e. I. However the bounds do not decrease with an increase in the number of measurements N, unlike conventional compressed sensing. This observation, though derived independently and using different techniques, agrees with existing literature on Poisson compressed sensing or Poisson matrix completion [15, 45, 44, 46]. The reason for this observation is the division of the signal flux across the N measurements, thereby leading to poorer signal to noise ratio per measurement.

There exist several avenues for future work, as follows. A major issue is to derive lower-bounds on the reconstruction error, and to derive bounds for (P4) along with a statistical criterion for the selection of ρ. Another avenue to use more general models for signal and noise correlation as in [57], to consider effect of quantization [58] over and above Poisson noise, and to consider simultaneous dictionary learning and sensing matrix inference in the context of Poisson compressed sensing [59].

7. Appendix

7.1. Proof of Theorem 1

To prove this theorem, we first begin by considering $y \sim \text{Poisson}(\gamma)$ where $\gamma \in \mathbb{R}_+$ and derive bounds for the mean and variance of $J(y, \gamma)$. Thereafter, we generalize to the case with multiple measurements.
Let $f(y) \triangleq J(y, \gamma)$ for non-negative and real-valued y for the purpose of deriving bounds. Hence we have

$$\begin{aligned} f(y) &= \frac{1}{2}(\gamma \log \gamma + y \log y) - \frac{\gamma + y}{2} \log\Big(\frac{\gamma + y}{2}\Big). \\ \therefore f^{(1)}(y) &= \frac{1}{2}[\log y - \log\Big(\frac{\gamma + y}{2}\Big)]. \\ \therefore f(y) &= \int_{\gamma}^{y} f^{(1)}(t)dt \text{ as } f(\gamma) = 0. \end{aligned}$$

where $f^{(k)}(y)$ stands for the k^{th} derivative of $f(y)$. As $f^{(1)}(y)$ is a non-decreasing function (since $f^{(2)}(y)$ is non-negative for all y), we have

$$f(y) \leq (y-\gamma) f^{(1)}(y). \tag{24}$$

Likewise, noting that $f^{(1)}(\gamma) = 0$ we get $f^{(1)}(y) = \int_{\gamma}^{y} f^{(2)}(t)dt$. We know that $f^{(2)}(y) = \frac{1}{2}\Big[\frac{1}{y} - \frac{1}{y+\gamma}\Big]$ is a decreasing function as $f^{(3)}(y)$ is negative for all y.

If $y \geq \gamma$ then $f^{(2)}(y) \leq f^{(2)}(\gamma)$. Therefore, $f^{(1)}(y) \leq (y-\gamma) f^{(2)}(\gamma)$. If $y \leq \gamma$ then $f^{(2)}(y) \geq f^{(2)}(\gamma)$. Therefore, $-f^{(1)}(y) \geq (\gamma - y) f^{(2)}(\gamma)$. Combining Eqn. 24 with the above inequality, we get

$$f(y) \leq (y-\gamma)^2 f^{(2)}(\gamma) = \frac{1}{4\gamma}(y-\gamma)^2. \tag{25}$$

Therefore, using $E[(y-\gamma)^2] = \gamma$ for a Poisson random variable, we have

$$E[f(y)] \leq \frac{1}{4\gamma} E[(y-\gamma)^2] = \frac{1}{4}. \tag{26}$$

Thus, we have found an upper bound on $E[f(y)]$ which is independent of γ.

We will now derive a lower bound on $E[f(y)]$, as it will be useful in deriving an upper bound for $\text{Var}(f(y))$. We can expand $f(y)$ using a second order Taylor series about γ along with a (third order) Lagrange remainder term as follows:

$$\begin{aligned} f(y) &= f(\gamma) + f^{(1)}(\gamma)(y-\gamma) + \frac{f^{(2)}(\gamma)}{2!}(y-\gamma)^2 + \frac{f^{(3)}(z(y))}{3!}(y-\gamma)^3 \\ &= \frac{1}{8\gamma}(y-\gamma)^2 - \frac{1}{12}(y-\gamma)^3\Big[\frac{1}{z^2(y)} - \frac{1}{(\gamma + z(y))^2}\Big] \end{aligned}$$

for some $z(y)$ that lies in the interval (y, γ) or (γ, y). Therefore,

$$\begin{aligned} E[f(y)] &= \frac{1}{8\gamma} E[(y-\gamma)^2] - \frac{1}{12}\Big[\sum_{y=0}^{\infty} \frac{e^{-\gamma}\gamma^y}{y!}(y-\gamma)^3\Big[\frac{1}{z^2(y)} - \frac{1}{(\gamma+z(y))^2}\Big]\Big] \\ &= \frac{1}{8} - \frac{1}{12}\Big[\sum_{y=0}^{\infty} \frac{e^{-\gamma}\gamma^y}{y!}(y-\gamma)^3\Big(\frac{1}{z^2(y)} - \frac{1}{(\gamma+z(y))^2}\Big)\Big]. \end{aligned}$$

Let α be the largest integer less than or equal to γ. We can split the second term in the RHS of the above expression into the sum of two terms I_1 and $-I_2$, depending upon whether y is greater than α or not. I_1 and I_2 are defined as follows:

$$\begin{aligned} I_1 &= \frac{1}{12}\Big[\sum_{y=0}^{\alpha} \frac{e^{-\gamma}\gamma^y}{y!}(\gamma-y)^3\Big(\frac{1}{z^2(y)} - \frac{1}{(\gamma+z(y))^2}\Big)\Big] \\ I_2 &= \frac{1}{12}\Big[\sum_{y=\alpha+1}^{\infty} \frac{e^{-\gamma}\gamma^y}{y!}(y-\gamma)^3\Big(\frac{1}{z^2(y)} - \frac{1}{(\gamma+z(y))^2}\Big)\Big]. \end{aligned}$$

In order to lower bound $E[f(y)]$, we want to minimize I_1 and maximize I_2 w.r.t. $z(y)$. Since $\frac{1}{z^2(y)} - \frac{1}{(\gamma+z(y))^2}$ is a decreasing function of $z(y)$, it can be proved that I_1 is minimized when $z(y) = \gamma$ and that I_2

attains a maximum when $z(y) = \gamma$. Therefore, we obtain

$$E[f(y)] \geq \frac{1}{8} - \frac{1}{16\gamma^2} E[(y-\gamma)^3] = \frac{1}{8} - \frac{1}{16\gamma}. \tag{27}$$

This lower bound is loose if $\gamma < 0.5$ since we know that $E[f(y)]$ must clearly be non-negative. Hence it is more apt to express the lower bound as follows:

$$E[f(y)] \geq \max(0, \frac{1}{8} - \frac{1}{16\gamma}). \tag{28}$$

In summary, we have

$$\max(0, \frac{1}{8} - \frac{1}{16\gamma}) \leq E[f(y)] \leq \frac{1}{4}. \tag{29}$$

We now proceed to derive an upper bound on the variance of $f(y)$.

Using Eqn. 25 we get,

$$E[(f(y))^2] \leq \frac{1}{16\gamma^2} E[(y-\gamma)^4] = \frac{\gamma(1+3\gamma)}{16\gamma^2} \leq \frac{3}{16} + \frac{1}{16\gamma}. \tag{30}$$

Recall that $\mathrm{Var}[f(y)] = E[(f(y))^2] - (E[f(y)])^2$. Using Eqns. 30 and 28, we get

$$\begin{aligned} \mathrm{Var}(f(y)) &\leq \tfrac{3}{16} + \tfrac{1}{16\gamma} - \left(\max(0, \left[\tfrac{1}{8} - \tfrac{1}{16\gamma}\right])\right)^2 && (31) \\ &\leq \max(0, \tfrac{11}{64} + \tfrac{5}{64\gamma} - \tfrac{1}{256\gamma^2}) && (32) \\ &\leq \tfrac{11}{64} + \tfrac{5}{64\gamma}. && (33) \end{aligned}$$

Now consider that $\boldsymbol{y}$ is a vector of N measurements such that $\forall i \in \{1, 2, ..., N\}, y_i \sim \text{Poisson}(\gamma_i)$ and all measurements are independent. We will later replace γ_i by $(\boldsymbol{\Phi x})_i$ where $\boldsymbol{\Phi}$ is a non-negative flux-preserving matrix and $\boldsymbol{x}$ is the unknown signal to be estimated. Let us define some terminology as follows:

$$f_i(y_i) \triangleq \frac{(\gamma_i \log \gamma_i + y_i \log y_i)}{2} - \frac{\gamma_i + y_i}{2} \log\left(\frac{\gamma_i + y_i}{2}\right), f(\boldsymbol{y}) \triangleq \sum_{i=1}^{N} f_i(y_i), g(\boldsymbol{y}) \triangleq \sqrt{f(\boldsymbol{y})}.$$

Jensen's inequality gives the following upper bound on the expected value of $g(\boldsymbol{y})$:

$$E[g(\boldsymbol{y})] = E[\sqrt{f(\boldsymbol{y})}] \leq \sqrt{\sum_{i=1}^{N} E[f_i(y_i)]} \leq \sqrt{\frac{N}{4}}. \tag{34}$$

In order to lower bound $E[g(\boldsymbol{y})]$ we use the following inequality for the non-negative variable f:

$$\sqrt{f} \geq 1 + \frac{f-1}{2} - \frac{(f-1)^2}{2}.$$

This inequality follows since it is equivalent to $3f - f^2 \leq 2\sqrt{f}$ which implies $3b - b^3 \leq 2$ which is true for

any $b \geq 0$. Define $\tilde{f} \triangleq \dfrac{f}{E[f]}$ such that $E[\tilde{f}] = 1$. Therefore, we have the following inequalities:

$$\begin{aligned}
\sqrt{\tilde{f}} &\geq 1 + \frac{\tilde{f}-1}{2} - \frac{(\tilde{f}-1)^2}{2} \\
\therefore E[\sqrt{\tilde{f}}] &\geq 1 - \frac{\mathrm{Var}(\tilde{f})}{2} \\
\therefore E[\sqrt{f}] &\geq \sqrt{E[f]}\Big(1 - \frac{\mathrm{Var}(f)}{2E[f]^2}\Big) \\
\therefore E[g] &\geq \sqrt{E[f]}\Big(1 - \frac{\mathrm{Var}(f)}{2E[f]^2}\Big).
\end{aligned}$$

Now, we can find an upper bound on $\mathrm{Var}[g(\boldsymbol{y})]$

$$\begin{aligned}
\mathrm{Var}(g) &= E[g^2] - E[g]^2 \\
&\leq E[f] - E[f]\Big(1 - \frac{\mathrm{Var}(f)}{2E[f]^2}\Big)^2 \\
&\leq \frac{\mathrm{Var}(f)}{E[f]} - \frac{1}{4}\frac{\mathrm{Var}(f)^2}{E[f]^3}.
\end{aligned}$$

Note that the first inequality in the chain above requires that $(E[g])^2 \geq E[f]\Big(1 - \frac{\mathrm{Var}(f)}{2E[f]^2}\Big)^2$. This follows from the earlier relationship $E[g] \geq \sqrt{E[f]}\Big(1 - \frac{\mathrm{Var}(f)}{2E[f]^2}\Big)$, only if its RHS is non-negative. Since $E[f] \geq 0$, this is equivalent to the condition that $\Big(1 - \frac{\mathrm{Var}(f)}{2E[f]^2}\Big) \geq 0$. It can be shown that this is guaranteed if $N \geq 32$ for the case when $\forall i, \gamma_i \geq 1$, by invoking the lower bound on $E[f]$ from Eqn. 28 and the upper bound on $\mathrm{Var}(f)$ from Eqn. 33. More generally, if $\forall i, \gamma_i \geq 0.5 + c$ where $c > 0$, then we must have $N \geq \frac{(10.5+11c)(1+2c)}{4c^2}$ to guarantee the afore-mentioned result.

As for different i, the variables $f_i(y_i)$ are independent of each other, we get $\mathrm{Var}(f) = \sum_{i=1}^{N} \mathrm{Var}(f_i)$, due to which we have:

$$\mathrm{Var}(g) \leq \frac{\sum_{i=1}^{N} \mathrm{Var}(f_i)}{\sum_{i=1}^{N} E(f_i)} - \frac{1}{4}\frac{(\sum_{i=1}^{N} \mathrm{Var}(f_i))^2}{(\sum_{i=1}^{N} E(f_i))^3} \leq \frac{11N + 5\sum_{i=1}^{N} 1/\gamma_i}{\sum_{i=1}^{N} \max(0, 4(2 - 1/\gamma_i))}. \tag{35}$$

The last step follows from Eqn. 33 and 28 and gives us the final upper bound. The expression on the RHS is a decreasing function of γ_i, and the upper bound is reached when $\forall i, 1 \leq i \leq N, \gamma_i = 1/2 + c$ where $c > 0$. This upper bound is $\frac{11}{8} + \frac{21}{16c}$.

In order to obtain a tail bound on $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}$, we can use Chebyshev's inequality to prove that $P(\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})} \leq \sqrt{N/4} + \sqrt{\frac{11}{8} + \frac{21}{16c}}\sqrt{N}) \geq 1 - \frac{1}{N}$, since the variance of $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}$ is upper bounded by $\frac{11}{8} + \frac{21}{16c}$. However, we show here that $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}$ is approximately Gaussian distributed which leads to tighter bounds and with an even higher probability: $P(\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})} \leq \sqrt{N/4} + \sqrt{\frac{11}{8} + \frac{21}{16c}}\sqrt{N}) \geq 1 - 2e^{-N/2}$ using upper bounds on the mean and variance of $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}$ from Eqns. 34 and 35 respectively. However while proving the Gaussianity, we further get a constant factor improvement as shown in the following paragraph.

By the central limit theorem, we know that $P(\frac{f(\boldsymbol{y}) - N\mu}{\sigma\sqrt{N}} \leq \alpha) \to \Phi_g(\alpha)$ as $N \to \infty$, where Φ_g is the CDF for $\mathcal{N}(0, 1)$, and μ, σ are respectively the expected value and standard deviation of f_i. All the f_i values

will have near-identical variances ($\leq \frac{11}{64} + \frac{5}{64(0.5+c)}$ from Eqn. 33) if the intensity of the measurements is sufficiently high. Due to the continuity of Φ_g[6], we have $P(\frac{f(\boldsymbol{y})-N\mu}{\sigma\sqrt{N}} \leq \alpha + \frac{\alpha^2\sigma^2}{4\mu\sigma\sqrt{N}}) \rightarrow \Phi_g(\alpha)$ as $N \rightarrow \infty$. Hence we have $P(f(\boldsymbol{y}) \leq (\sqrt{N\mu} + \frac{\alpha\sigma}{2\sqrt{\mu}})^2) \rightarrow \Phi_g(\alpha)$ as $N \rightarrow \infty$, and taking square roots we get $P(\sqrt{f(\boldsymbol{y})} \leq (\sqrt{N\mu} + \frac{\alpha\sigma}{2\sqrt{\mu}})) \rightarrow \Phi_g(\alpha)$ as $N \rightarrow \infty$. By rearrangement, we obtain $P(\frac{\sqrt{f(\boldsymbol{y})}-\sqrt{N\mu}}{\sigma/(2\sqrt{\mu})} \leq \alpha) \rightarrow \Phi_g(\alpha)$ as $N \rightarrow \infty$. With this development and since $\mu \leq 1/4, \sigma \leq \sqrt{\frac{11}{64} + \frac{5}{64(0.5+c)}}$ from Eqns. 26 and 33, we can now invoke a Gaussian tail bound to establish that $P(\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})} \leq \sqrt{N/4} + \sqrt{\frac{11}{64} + \frac{5}{64(0.5+c)}}\sqrt{N}) \geq 1 - 2e^{-N/2}$. Note that the Gaussian nature of $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}$ emerges from the central limit theorem and is only an asymptotic result. However we consistently observe it to be true even for small values of $N \sim 10$ as confirmed by a Kolmogorov-Smirnov test (see [23]). □

7.2. Proof of Theorem 2

Our proof follows the approach for the proof of the key results in [1, 11] for the case of bounded, signal-independent noise, but meticulously adapted here for the case of Poisson noise.

1. Consider an upper bound ε on $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}$, *i.e.*, $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})} \leq \varepsilon$. We will later set ε using tail bounds on the distribution of the random variable $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}$ from Theorem 1 of the main paper. For now, we prove the following result:

$$\|\boldsymbol{\Phi\Psi}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\|_2 \leq 2\sqrt{8I}\varepsilon. \tag{36}$$

We have

$$\begin{aligned}
\|\boldsymbol{\Phi\Psi\theta} - \boldsymbol{\Phi\Psi\theta}^\star\|_2 \leq \|\boldsymbol{\Phi\Psi}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\|_1 &= I\|\boldsymbol{\Phi\Psi}(\frac{\boldsymbol{\theta}}{I} - \frac{\boldsymbol{\theta}^\star}{I})\|_1 \\
&\leq I\sqrt{8J(\frac{\boldsymbol{\Phi\Psi\theta}}{I}, \frac{\boldsymbol{\Phi\Psi\theta}^\star}{I})} \quad \text{by Lemma 2} \\
&\leq I\sqrt{8J(\frac{\boldsymbol{y}}{I}, \frac{\boldsymbol{\Phi\Psi\theta}}{I})} + I\sqrt{8J(\frac{\boldsymbol{y}}{I}, \frac{\boldsymbol{\Phi\Psi\theta}^\star}{I})} \text{ by Lemma 1} \\
&= \frac{I}{\sqrt{I}}\sqrt{8J(\boldsymbol{y}, \boldsymbol{\Phi\Psi\theta})} + \frac{I}{\sqrt{I}}\sqrt{8J(\boldsymbol{y}, \boldsymbol{\Phi\Psi\theta}^\star)} \leq 2\sqrt{8I}\varepsilon.
\end{aligned}$$

Note that Lemma 2 can be used in the third step above because we have imposed the constraint that $\|\boldsymbol{\Psi\theta}^\star\|_1 = \|\boldsymbol{\Psi\theta}\|_1 = I$ and because by the flux-preserving property of $\boldsymbol{\Phi}$, we have $\|\boldsymbol{\Phi\Psi\theta}\|_1 \leq I$ and $\|\boldsymbol{\Phi\Psi\theta}^\star\|_1 \leq I$.

2. Let us define vector $\boldsymbol{h} \triangleq \boldsymbol{\theta}^\star - \boldsymbol{\theta}$ which is the difference between the estimated and true coefficient vectors. Let us denote vector $\boldsymbol{h}_T$ as the vector equal to $\boldsymbol{h}$ only on an index set T and zero at all other indices. Let T^c denote the complement of the index set T. Let T_0 be the set of indices containing the

[6] inspired from `https://tinyurl.com/ybmc7rgs`

s largest entries of $\boldsymbol{\theta}$ (in terms of absolute value), T_1 be the set of indices of the next s largest entries of $\boldsymbol{h}_{T_0^c}$, and so on. We will now decompose $\boldsymbol{h}$ as the sum of $\boldsymbol{h}_{\boldsymbol{T_0}}, \boldsymbol{h}_{T_1}, \boldsymbol{h}_{T_2}, \ldots$. Our aim is to prove a logical and intuitive bound for both $\|\boldsymbol{h}_{T_0 \cup T_1}\|_2$ and $\|\boldsymbol{h}_{(T_0 \cup T_1)^c}\|_2$.

3. We will first prove the bound on $\|\boldsymbol{h}_{(T_0 \cup T_1)^c}\|_2$, in the following way:
   (a) We have

$$\|\boldsymbol{h}_{T_j}\|_2 = \sqrt{\sum_k \boldsymbol{h}_{T_{j_k}}^2} \leq s^{1/2} \|\boldsymbol{h}_{T_j}\|_\infty,$$
$$s\|\boldsymbol{h}_{T_j}\|_\infty \leq \sum_i |\boldsymbol{h}_{T_{j-1_i}}| = \|\boldsymbol{h}_{T_{j-1}}\|_1.$$

   Therefore,

$$\|\boldsymbol{h}_{T_j}\|_2 \leq s^{1/2} \|\boldsymbol{h}_{T_j}\|_\infty \leq s^{-1/2} \|\boldsymbol{h}_{T_{j-1}}\|_1.$$

   (b) Using Step 3(a), we get

$$\begin{aligned} \|\boldsymbol{h}_{(T_0 \cup T_1)^c}\|_2 = \|\sum_{j \geq 2} \boldsymbol{h}_{T_j}\|_2 &\leq \sum_{j \geq 2} \|\boldsymbol{h}_{T_j}\|_2 \\ &\leq s^{-1/2} \sum_{i \geq 1} \|\boldsymbol{h}_{T_i}\|_1 \\ &\leq s^{-1/2} \|\boldsymbol{h}_{(T_0)^c}\|_1. \end{aligned}$$

   (c) Using the reverse triangle inequality and the fact that $\boldsymbol{\theta}^\star$ is the solution of (P2), we have

$$\begin{aligned} \|\boldsymbol{\theta}\|_1 &\geq \|\boldsymbol{\theta} + \boldsymbol{h}\|_1 \\ &= \sum_{i \in T_0} |\theta_i + h_i| + \sum_{i \in (T_0)^c} |\theta_i + h_i| \\ &\geq \|\boldsymbol{\theta}_{T_0}\|_1 - \|\boldsymbol{h}_{T_0}\|_1 + \|\boldsymbol{h}_{(T_0)^c}\|_1 - \|\boldsymbol{\theta}_{(T_0)^c}\|_1. \end{aligned}$$

   Rearranging the above equation gives us

$$\|\boldsymbol{h}_{(T_0)^c}\|_1 \leq \|\boldsymbol{h}_{(T_0)}\|_1 + 2\|\boldsymbol{\theta} - \boldsymbol{\theta_s}\|_1$$

   (d) We have

$$\begin{aligned} \|\boldsymbol{h}_{(T_0 \cup T_1)^c}\|_2 &\leq s^{-1/2} \|\boldsymbol{h}_{(T_0)^c}\|_1 \\ &\leq s^{-1/2} (\|\boldsymbol{h}_{(T_0)}\|_1 + 2\|\boldsymbol{\theta} - \boldsymbol{\theta_s}\|_1) \\ &\leq \|\boldsymbol{h}_{(T_0)}\|_2 + 2s^{-1/2} \|\boldsymbol{\theta} - \boldsymbol{\theta_s}\|_1 \end{aligned}$$

   Using $\|\boldsymbol{h}_{(T_0)}\|_2 \leq \|\boldsymbol{h}_{T_0 \cup T_1}\|_2$, we get

$$\|\boldsymbol{h}_{(T_0 \cup T_1)^c}\|_2 \leq \|\boldsymbol{h}_{T_0 \cup T_1}\|_2 + 2s^{-1/2} \|\boldsymbol{\theta} - \boldsymbol{\theta_s}\|_1. \tag{37}$$

4. We will now prove the bound on $\|\boldsymbol{h}_{(T_0 \cup T_1)}\|_2$, in the following way:

(a) We have

$$\begin{aligned}
\boldsymbol{\Phi} &= \sqrt{\frac{p(1-p)}{N}}\tilde{\boldsymbol{\Phi}} + \frac{(1-p)}{N}\mathbf{1}_{N\times m} \\
\boldsymbol{\Phi\Psi}(\boldsymbol{\theta}-\boldsymbol{\theta}^\star) &= \sqrt{\frac{p(1-p)}{N}}\tilde{\boldsymbol{\Phi}}\boldsymbol{\Psi}(\boldsymbol{\theta}-\boldsymbol{\theta}^\star) + \\
&\quad \frac{(1-p)}{N}\mathbf{1}_{N\times m}\boldsymbol{\Psi}(\boldsymbol{\theta}-\boldsymbol{\theta}^\star) \\
&= \sqrt{\frac{p(1-p)}{N}}\tilde{\boldsymbol{\Phi}}\boldsymbol{\Psi}(\boldsymbol{\theta}-\boldsymbol{\theta}^\star) + \\
&\quad \frac{(1-p)}{N}(\|\boldsymbol{\Psi\theta}\|_1 - \|\boldsymbol{\Psi\theta}^\star\|_1)
\end{aligned}$$

As $\|\boldsymbol{\Psi\theta}^\star\|_1 = \|\boldsymbol{\Psi\theta}\|_1 = I$, we get

$$\boldsymbol{\Phi\Psi}(\boldsymbol{\theta}-\boldsymbol{\theta}^\star) = \sqrt{\frac{p(1-p)}{N}}\tilde{\boldsymbol{\Phi}}\boldsymbol{\Psi}(\boldsymbol{\theta}-\boldsymbol{\theta}^\star). \tag{38}$$

Let us define $\boldsymbol{B} \triangleq \tilde{\boldsymbol{\Phi}}\boldsymbol{\Psi}$. If $N \geq O(s\log m)$, then $\tilde{\boldsymbol{\Phi}}$ obeys RIP of order $2s$ with very high probability, and so does the product $\boldsymbol{B}$ since $\boldsymbol{\Psi}$ is an orthonormal matrix [12].

From Eqn. 38 above we have,

$$\begin{aligned}
\|\boldsymbol{B}(\boldsymbol{\theta}-\boldsymbol{\theta}^\star)\|_2 &= \sqrt{\frac{N}{p(1-p)}}\|\boldsymbol{\Phi\Psi}(\boldsymbol{\theta}-\boldsymbol{\theta}^\star)\|_2 \\
&\leq 2\sqrt{\frac{8NI}{p(1-p)}}\varepsilon \text{ using Eqn. 36} \\
\therefore \|\boldsymbol{Bh}\|_2 &\leq 2\sqrt{\frac{8NI}{p(1-p)}}\varepsilon
\end{aligned}$$

Defining $C_1 \triangleq 2\sqrt{\frac{8}{p(1-p)}}$, we have

$$\|\boldsymbol{Bh}\|_2 \leq C_1\sqrt{NI}\varepsilon \tag{39}$$

(b) The RIP of $\boldsymbol{B}$ with RIC δ_{2s} gives us,

$$\|\boldsymbol{Bh}_{T_0\cup T_1}\|_2 \leq \sqrt{1+\delta_{2s}}\|\boldsymbol{h}_{T_0\cup T_1}\|_2$$

Using Eqn. 39 and the Cauchy-Schwartz inequality,

$$\begin{aligned}
|\langle \boldsymbol{Bh}_{T_0\cup T_1}, \boldsymbol{Bh}\rangle| &\leq \|\boldsymbol{Bh}_{T_0\cup T_1}\|_2\|\boldsymbol{Bh}\|_2 \\
&\leq C_1\varepsilon\sqrt{NI(1+\delta_{2s})}\|\boldsymbol{h}_{T_0\cup T_1}\|_2.
\end{aligned} \tag{40}$$

(c) Note that the vectors $\boldsymbol{h}_{T_0}$ and $\boldsymbol{h}_{T_j}$, $j \neq 0$ have disjoint support. Consider

$$|\langle \boldsymbol{B}\boldsymbol{h}_{T_0}, \boldsymbol{B}\boldsymbol{h}_{T_j}\rangle| = \|\boldsymbol{h}_{T_0}\|_2 \|\boldsymbol{h}_{T_j}\|_2 |\langle \boldsymbol{B}\hat{\boldsymbol{h}}_{T_0}, \boldsymbol{B}\hat{\boldsymbol{h}}_{T_j}\rangle|$$

where $\hat{\boldsymbol{h}}_{T_0}$ and $\hat{\boldsymbol{h}}_{T_j}$ are unit-normalized vectors. This further yields,

$$\begin{aligned}
&|\langle \boldsymbol{B}\boldsymbol{h}_{T_0}, \boldsymbol{B}\boldsymbol{h}_{T_j}\rangle| \\
&= \|\boldsymbol{h}_{T_0}\|_2 \|\boldsymbol{h}_{T_j}\|_2 \frac{\|\boldsymbol{B}(\hat{\boldsymbol{h}}_{T_0} + \hat{\boldsymbol{h}}_{T_j})\|^2 - \|\boldsymbol{B}(\hat{\boldsymbol{h}}_{T_0} - \hat{\boldsymbol{h}}_{T_j})\|^2}{4} \\
&\leq \|\boldsymbol{h}_{T_0}\|_2 \|\boldsymbol{h}_{T_j}\|_2 \frac{(1+\delta_{2s})(\|\hat{\boldsymbol{h}}_{T_0}\|^2 + \|\hat{\boldsymbol{h}}_{T_j}\|^2) - (1-\delta_{2s})(\|\hat{\boldsymbol{h}}_{T_0}\|^2 + \|\hat{\boldsymbol{h}}_{T_j}\|^2)}{4} \\
&\leq \delta_{2s} \|\boldsymbol{h}_{T_0}\|_2 \|\boldsymbol{h}_{T_j}\|_2. \qquad (41)
\end{aligned}$$

Analogously,

$$|\langle \boldsymbol{B}\boldsymbol{h}_{T_1}, \boldsymbol{B}\boldsymbol{h}_{T_j}\rangle| \leq \delta_{2s} \|\boldsymbol{h}_{T_1}\|_2 \|\boldsymbol{h}_{T_j}\|_2. \qquad (42)$$

(d) We observe that

$$\begin{aligned}
\boldsymbol{B}\boldsymbol{h}_{T_0 \cup T_1} &= \boldsymbol{B}\boldsymbol{h} - \sum_{j \geq 2} \boldsymbol{B}\boldsymbol{h}_{T_j} \\
\|\boldsymbol{B}\boldsymbol{h}_{T_0 \cup T_1}\|_2^2 &= \langle \boldsymbol{B}\boldsymbol{h}_{T_0 \cup T_1}, \boldsymbol{B}\boldsymbol{h}\rangle - \langle \boldsymbol{B}\boldsymbol{h}_{T_0 \cup T_1}, \sum_{j \geq 2} \boldsymbol{B}\boldsymbol{h}_{T_j}\rangle. \qquad (43)
\end{aligned}$$

(e) Using the RIP of $\boldsymbol{B}$ and Eqns. 40, 41, 42, 43, we obtain

$$(1-\delta_{2s})\|\boldsymbol{h}_{T_0 \cup T_1}\|_2^2 \leq \|\boldsymbol{B}\boldsymbol{h}_{T_0 \cup T_1}\|_2^2 \leq C_1 \varepsilon \sqrt{NI(1+\delta_{2s})} \|\boldsymbol{h}_{T_0 \cup T_1}\|_2 + \delta_{2s}(\|\boldsymbol{h}_{T_0}\|_2 + \|\boldsymbol{h}_{T_1}\|_2) \sum_{j \geq 2} \|\boldsymbol{h}_{T_j}\|_2.$$

As $\boldsymbol{h}_{T_0}$ and $\boldsymbol{h}_{T_1}$ are vectors with disjoint sets of non-zero indices, it follows that

$$\|\boldsymbol{h}_{T_0}\|_2 + \|\boldsymbol{h}_{T_1}\|_2 \leq \sqrt{2} \|\boldsymbol{h}_{T_0 \cup T_1}\|_2.$$

Therefore, we get

$$(1-\delta_{2s})\|\boldsymbol{h}_{T_0 \cup T_1}\|_2^2 \leq \|\boldsymbol{h}_{T_0 \cup T_1}\|_2 \left(C_1 \varepsilon \sqrt{NI(1+\delta_{2s})} + \sqrt{2}\delta_{2s} \sum_{j \geq 2} \|\boldsymbol{h}_{T_j}\|_2 \right). \qquad (44)$$

(f) We have

$$\begin{aligned}
\sum_{j \geq 2} \|\boldsymbol{h}_{T_j}\|_2 &\leq s^{-1/2} \|\boldsymbol{h}_{(T_0)^c}\|_1 \\
&\leq s^{-1/2} \|\boldsymbol{h}_{(T_0)}\|_1 + 2s^{-1/2} \|\boldsymbol{\theta} - \boldsymbol{\theta_s}\|_1 \\
&\leq \|\boldsymbol{h}_{(T_0)}\|_2 + 2s^{-1/2} \|\boldsymbol{\theta} - \boldsymbol{\theta_s}\|_1 \\
&\leq \|\boldsymbol{h}_{T_0 \cup T_1}\|_2 + 2s^{-1/2} \|\boldsymbol{\theta} - \boldsymbol{\theta_s}\|_1. \qquad (45)
\end{aligned}$$

Combining Eqns. 44 and 45,

$$\|\boldsymbol{h}_{T_0\cup T_1}\|_2 \leq C_1\varepsilon\frac{\sqrt{NI(1+\delta_{2s})}}{1-(1+\sqrt{2})\delta_{2s}} + \frac{2\sqrt{2}\delta_{2s}}{1-(1+\sqrt{2})\delta_{2s}}s^{-1/2}\|\boldsymbol{\theta}-\boldsymbol{\theta_s}\|_1. \tag{46}$$

5. Combining the upper bounds on $\|\boldsymbol{h}_{(T_0\cup T_1)}\|_2$ and $\|\boldsymbol{h}_{(T_0\cup T_1)^c}\|_2$ yields the final result as follows:

$$\|\boldsymbol{h}\|_2 = \|\boldsymbol{h}_{T_0\cup T_1} + \boldsymbol{h}_{(T_0\cup T_1)^c}\|_2 \leq \|\boldsymbol{h}_{T_0\cup T_1}\|_2 + \|\boldsymbol{h}_{(T_0\cup T_1)^c}\|_2 \leq 2\|\boldsymbol{h}_{T_0\cup T_1}\|_2 + 2s^{-1/2}\|\boldsymbol{\theta}-\boldsymbol{\theta_s}\|_1.$$

Using Eqn. 46, we get

$$\|\boldsymbol{h}\|_2 \leq 2C_1\varepsilon\frac{\sqrt{NI(1+\delta_{2s})}}{1-(1+\sqrt{2})\delta_{2s}} + \Big(\frac{2-2\delta_{2s}+2\sqrt{2\delta_{2s}}}{1-(1+\sqrt{2})\delta_{2s}}\Big)s^{-1/2}\|\boldsymbol{\theta}-\boldsymbol{\theta_s}\|_1.$$

Let us define $C' \triangleq \dfrac{4\sqrt{8(1+\delta_{2s})}}{\sqrt{p(1-p)}(1-(1+\sqrt{2})\delta_{2s})}$ and $C'' \triangleq \Big(\dfrac{2-2\delta_{2s}+2\sqrt{2\delta_{2s}}}{1-(1+\sqrt{2})\delta_{2s}}\Big)$. This yields

$$\|\boldsymbol{h}\|_2 \leq C'\sqrt{NI}\varepsilon + C''s^{-1/2}\|\boldsymbol{\theta}-\boldsymbol{\theta_s}\|_1. \tag{47}$$

The positivity requirements for C' and C'' are met by $\delta_{2s} < \sqrt{2}-1$. Dividing both sides by I we obtain the first part of the theorem,

$$\frac{\|\boldsymbol{\theta}-\boldsymbol{\theta^\star}\|_2}{I} \leq C'\sqrt{\frac{N}{I}}\varepsilon + \frac{C''s^{-1/2}\|\boldsymbol{\theta}-\boldsymbol{\theta_s}\|_1}{I}.$$

However using tail bounds on $\sqrt{J(\boldsymbol{y},\boldsymbol{\Phi x})}$ from Theorem 1 from the main paper, we can set $\varepsilon = \sqrt{N}(\frac{1}{2}+\frac{\sqrt{11}}{\sqrt{8}})$. This yields the following:

$$\Pr\Big(\frac{\|\boldsymbol{\theta}-\boldsymbol{\theta^\star}\|_2}{I} \leq \tilde{C}\frac{N}{\sqrt{I}} + \frac{C''s^{-1/2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_s\|_1}{I}\Big) \geq 1-1/N, \tag{48}$$

where $\tilde{C} \triangleq C'(1/2+\sigma)$ where σ is the upper bound of $\sqrt{\frac{11}{8}+\frac{21}{64c}}$ on the standard deviation of the SQJSD as stated in Theorem 1. For sufficiently high intensity signals, the previous analysis shows that σ is independent of both I and N. Also, the probability of $1-1/N$ can be refined to $1-2e^{-N/2}$ as argued in the comments after Theorems 1 and 2. □

References

[1] E. Candes, The restricted isometry property and its implications for compressed sensing, Comptes Rendus Mathematique 346 (910) (2008) 589 – 592.

[2] S. Kallummil, S. Kalyani, High snr consistent compressive sensing, Signal Processing 146 (2018) 1 – 14.

[3] L. Jacques, A short note on compressed sensing with partially known signal support, Signal Processing 90 (12) (2010) 3308 – 3312.

[4] X. Huang, Y. Liu, L. Shi, S. V. Huffel, J. A. K. Suykens, Two-level l1 minimization for compressed sensing, Signal Processing 108 (2015) 459 – 475.

[5] F. Alter, Y. Matsushita, X. Tang, An intensity similarity measure in low-light conditions, in: ECCV, 2006.

[6] J. L. Starck, J. Bobin, Astronomical data analysis and sparsity: From wavelets to compressed sensing, Proceedings of the IEEE 98 (6) (2010) 1021–1030.

[7] J. Boone, E. Geraghty, J. Seibert, S. Wootton-Gorges, Dose reduction in pediatric CT: A rational approach, Radiology 228 (2) (2003) 352–360.

[8] J. Boulanger, C. Kervrann, P. Bouthemy, P. Elbau, J.-B. Sibarita, J. Salamero, Patch-based nonlocal functional for denoising fluorescence microscopy image sequences, IEEE Trans. Med. Imag. 29 (2) (2010) 442454.

[9] S. Yang, et al, Estimation of multiexponential fluorescence decay parameters using compressive sensing, Journal of Biomedical Optics 20 (9).

[10] T. T. Cai, A. Zhang, Sharp rip bound for sparse signal and low-rank matrix recovery, Applied and Computational Harmonic Analysis 35 (1) (2013) 74 – 93.

[11] C. Studer, R. Baraniuk, Stable restoration and separation of approximately sparse signals, Applied and Computational Harmonic Analysis 37 (1) (2014) 12 – 35.

[12] R. Baraniuk, M. Davenport, R. DeVore, M. Wakin, A simple proof of the restricted isometry property for random matrices, Constructive Approximation 28 (3) (2008) 253–263.

[13] C. Fevotte, A. T. Cemgil, Nonnegative matrix factorizations as probabilistic inference in composite models, in: Signal Processing Conference, 2009 17th European, 2009, pp. 1913–1917.

[14] D. Endres, J. Schindelin, A new metric for probability distributions, IEEE Trans. Inf. Theory 49 (7) (2003) 18581860.

[15] M. Raginsky, R. Willett, Z. Harmany, R. Marcia, Compressed sensing performance bounds under Poisson noise, Signal Processing, IEEE Transactions on 58 (8) (2010) 3990–4002.

[16] S. Ivanoff, F. Picard, V. Rivoirard, Adaptive lasso and group-lasso for functional Poisson regression, Journal of Machine Learning Research 17 (55) (2016) 1–46.

[17] X. Jiang, P. Reynaud-Bouret, V. Rivoirard, L. Sansonnet, R. Willett, A data-dependent weighted LASSO under Poisson noise, `http://arxiv.org/abs/1509.08892`, online; accessed Feb. 2019.

[18] M.-H. Rohban, V. Saligrama, D.-M. Vaziri, Minimax optimal sparse signal recovery with Poisson statistics, IEEE Trans. Signal Processing 64 (13) (2016) 3495–3508.

[19] Y.-H. Li, V. Cevher, Consistency of l1-regularized maximum-likelihood for compressive Poisson regression, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, 2015, pp. 3606–3610.

[20] M. Mordechay, Y. Y. Schechner, Matrix optimization for Poisson compressed sensing, in: IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2014, pp. 684–688.

[21] T. Cover, J. Thomas, Elements of Information Theory 2nd Edition, Wiley-Interscience, 2006.

[22] F. Topsoe, Some inequalities for information divergence and related measures of discrimination, IEEE Transactions on Information Theory 46 (4) (2000) 1602–1609.

[23] Supplemental material and code for reproducing results in the paper, `https://www.cse.iitb.ac.in/~ajitvr/SQJSD/`.

[24] X. Jiang, G. Raskutti, R. Willett, Minimax optimal rates for Poisson inverse problems with physical constraints, IEEE Trans. Information Theory 61 (8) (2015) 4458–4474.

[25] E. Candes, T. Tao, The dantzig selector: Statistical estimation when p is much larger than n, The Annals of Statistics 35 (6) (2007) 2313–2351.

[26] R. Saab, O. Yilmaz, Sparse recovery by non-convex optimization instance optimality, Applied and Computational Harmonic Analysis 29 (1) (2010) 30 – 48.

[27] E. Candes, M. Wakin, S. Boyd, Enhancing sparsity by reweighted l1 minimization, Journal of Fourier Analysis and Applications 14 (5) (2008) 877–905.

[28] D. Lingenfelter, J. Fessler, Z. He, Sparsity regularization for image reconstruction with Poisson data, Vol. 7246, 2009.

[29] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly, R. Baraniuk, Single pixel imaging via compressive sampling, IEEE Signal Processing Magazine.

[30] B. Li, L. Zhang, T. Kirubarajan, S. Rajan, A projection matrix design method for MSE reduction in adaptive compressive sensing, Signal Process. 141 (C) (2017) 16–27.

[31] B. Li, L. Zhang, T. Kirubarajan, S. Rajan, Projection matrix design using prior information in compressive sensing, Signal Processing 135 (2017) 36–47.

[32] Z. T. Harmany, R. F. Marcia, R. M. Willett, This is SPIRAL-TAP: Sparse Poisson intensity reconstruction algorithms - theory and practice, IEEE Trans. Image Processing 21 (3) (2012) 1084–1096.

[33] J. Jinzhu, R. Karl, Y. Bin, The LASSO under Poisson-like heterscedasticity, Statistica Sinica 23 (1) (2013) 99–118.

[34] D. Garg, P. Bohra, K. S. Gurumoorthy, A. Rajwade, Variance stabilization based compressive inversion under poisson or poisson-gaussian noise with analytical bounds.
URL https://www.cse.iitb.ac.in/~ajitvr/Poisson_PoissonGaussian_VST.pdf

[35] S. Foucart, H. Rauhut, A Mathematical Introduction to Compressive Sensing, Birkhauser, 2013.

[36] M. Grant, S. Boyd, CVX: Matlab software for disciplined convex programming, version 2.1, http://cvxr.com/cvx (Mar. 2014).

[37] Y. Oike, A. El Gamal, CMOS image sensor with per-column sigma delta ADC and programmable compressed sensing, Solid-State Circuits, IEEE Journal of 48 (1) (2013) 318–328.

[38] K. Kulkarni, S. Lohit, P. Turaga, R. Kerviche, A. Ashok, Reconnet:non-iterative reconstruction of images from compressively sensed measurements, in: CVPR, 2016.

[39] L. Oneto, S. Ridella, D. Anguita, Tikhonov, ivanov and morozov regularization for support vector machine learning, Machine Learning 103 (1) (2016) 103–136.

[40] B. Zhang, M. Fadili, J. Starck, Wavelets, ridgelets, and curvelets for Poisson noise removal, IEEE Transactions on Image Processing 17 (7) (2008) 1093–1108.

[41] S. Sra, D. Kim, B. Schlkopf, Non-monotonic Poisson likelihood maximization, Tech. Rep. 170, Max Planck Institute (2008).

[42] M. Raginsky, S. Jafarpour, Z. T. Harmany, R. F. Marcia, R. M. Willett, R. Calderbank, Performance bounds for expander-based compressed sensing in Poisson noise, IEEE Transactions on Signal Processing 59 (9) (2011) 4139–4153.

[43] L. Wang, et al, Signal recovery and system calibration from multiple compressive Poisson measurements, SIAM J. Imaging Sciences 8 (3) (2015) 1923–1954.

[44] Y. Xie, Y. Chi, R. Calderbank, Low-rank matrix recovery with Poisson noise, in: IEEE GlobalSIP, 2013, pp. 622–622.

[45] Y. Cao, Y. Xie, Poisson matrix recovery and completion, IEEE Transactions on Signal Processing 64 (6) (2016) 1609–1620.

[46] A. Soni, S. Jain, J. Haupt, S. Gonella, Noisy matrix completion under sparse factor models, IEEE Trans. Information Theory 62 (6) (2016) 3636–3661.

[47] Y. Li, G. Raskutti, Minimax optimal convex methods for Poisson inverse problems under lq-ball sparsity, https://arxiv.org/abs/1604.08943, online; accessed Feb. 2019.

[48] I. Rish, G. Grabarnik, Sparse signal recovery with exponential-family noise, in: ACCS, 2009, pp. 60–66.

[49] M. Blazere, J. M. Loubes, F. Gamboa, Oracle inequalities for a group lasso procedure applied to generalized linear models in high dimension, IEEE Transactions on Information Theory 60 (4) (2014) 2303–2318.

[50] S. Kakade, O. Shamir, K. Sindharan, A. Tewari, Learning exponential families in high-dimensions: Strong convexity and sparsity, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS, 2010, pp. 381–388.

[51] M. Makitalo, A. Foi, Optimal inversion of the generalized anscombe transformation for Poisson-gaussian noise, IEEE Transactions on Image Processing 22 (1) (2013) 91–103.

[52] F.-X. Dupé, J. Fadili, J.-L. Starck, A proximal iteration for deconvolving poisson noisy images using sparse representations, IEEE Trans. Image Processing 18 (2) (2009) 310–321.

[53] D. Garg, A. Rajwade, Performance bounds for poisson compressed sensing using variance stabilization transforms, in: ICASSP, 2017, pp. 6080–6084.

[54] Stirling's approximation, https://en.wikipedia.org/wiki/Stirling%27s_approximation, online; accessed May 2016.

[55] M. Collins, S. Dasgupta, R. Schapire, A generalization of principal component analysis to the exponential family, in: Advances in Neural Information Processing Systems, 2001.

[56] J. Lin, Divergence measures based on the shannon entropy, IEEE Transactions on Information Theory 37 (1) (1991) 6958–6975.

[57] T. Arildsen, T. Larsen, Compressed sensing with linear correlation between signal and measurement noise, Signal Processing 98 (2014) 275 – 283.

[58] J. Fang, Y. Shen, L. Yang, H. Li, Adaptive one-bit quantization for compressed sensing, Signal Processing 125 (2016) 145 – 155.

[59] T. Hong, Z. Zhu, Online learning sensing matrix and sparsifying dictionary simultaneously for compressive sensing, Signal Processing 153 (2018) 188 – 196.

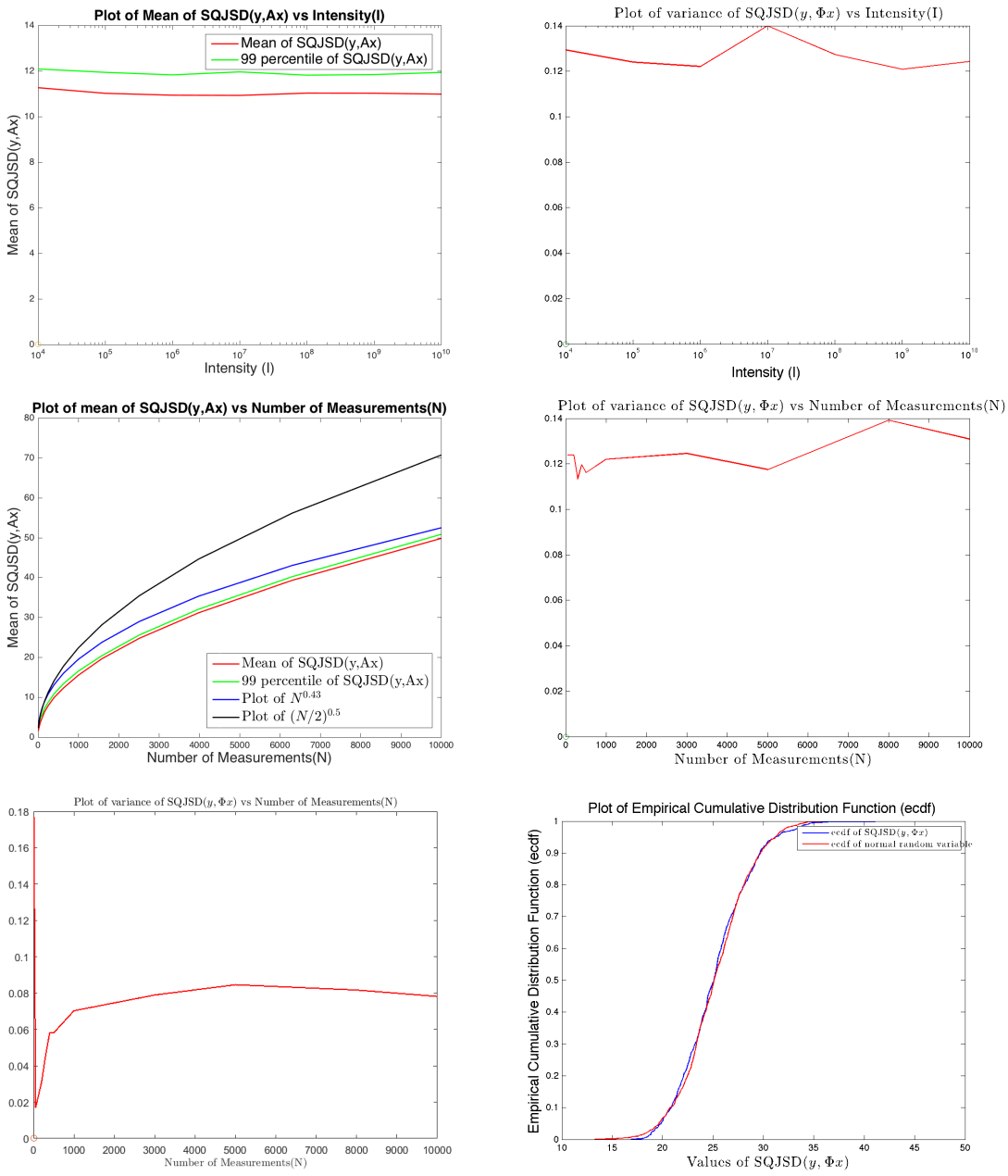

Figure 1: First row: Plot of mean and 99 percentile (left), and plot of variance (right) of the values of $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}$ versus I for a fixed $N = 500$ for a signal of dimension $m = 1000$. Second row: Plot of mean and 99 percentile (left), and plot of the variance (right) of the values of $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}$ versus N for a fixed $I = 10^6$ for a signal of dimension $m = 1000$. The left plot also contains a plot of $N^{0.43}$ and $(N/2)^{0.5}$ for comparison. Third row: Left - Plot of variance of $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}$ versus N for a signal with $I = 40, m = 500$, i.e. very low values of $\boldsymbol{\Phi x}$. Right - Empirical CDF of $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}$ for $N = 100, I = 10^4, m = 500$ compared to a Gaussian CDF with mean and variance equal to that of the values of $\sqrt{J(\boldsymbol{y}, \boldsymbol{\Phi x})}$. Scripts for reproducing these results are available at [23].

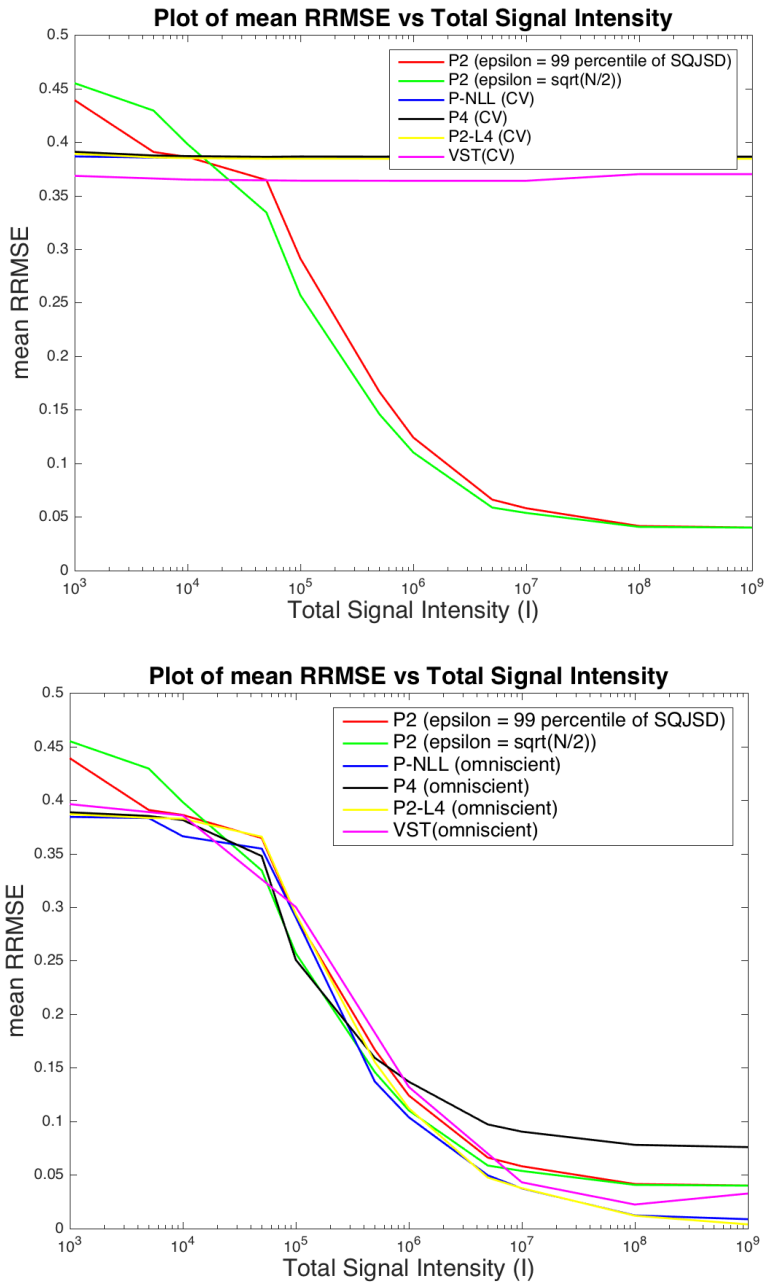

Figure 2: Results for *intensity experiment*. Top: Mean of RRMSE versus I for the following problems: (P2) with $\varepsilon = \sqrt{N/2}$, (P2) with $\varepsilon = 99$ percentile of SQJSD values, and following estimators with ρ by cross-validation (CV): (P4), (P-NLL), (P2-L4) and (P-VST). Bottom: Mean of RRMSE versus I for the following problems: (P2) with $\varepsilon = \sqrt{N/2}$, (P2) with $\varepsilon = 99$ percentile of SQJSD values, and following estimators with omniscient ρ: (P4), (P-NLL), (P2-L4) and (P-VST). All results are for an ensemble of $Q = 100$ 1D signals of 100 elements, for fixed $N = 50$ and fixed signal sparsity $s = 10$ in the 1D-DCT basis.

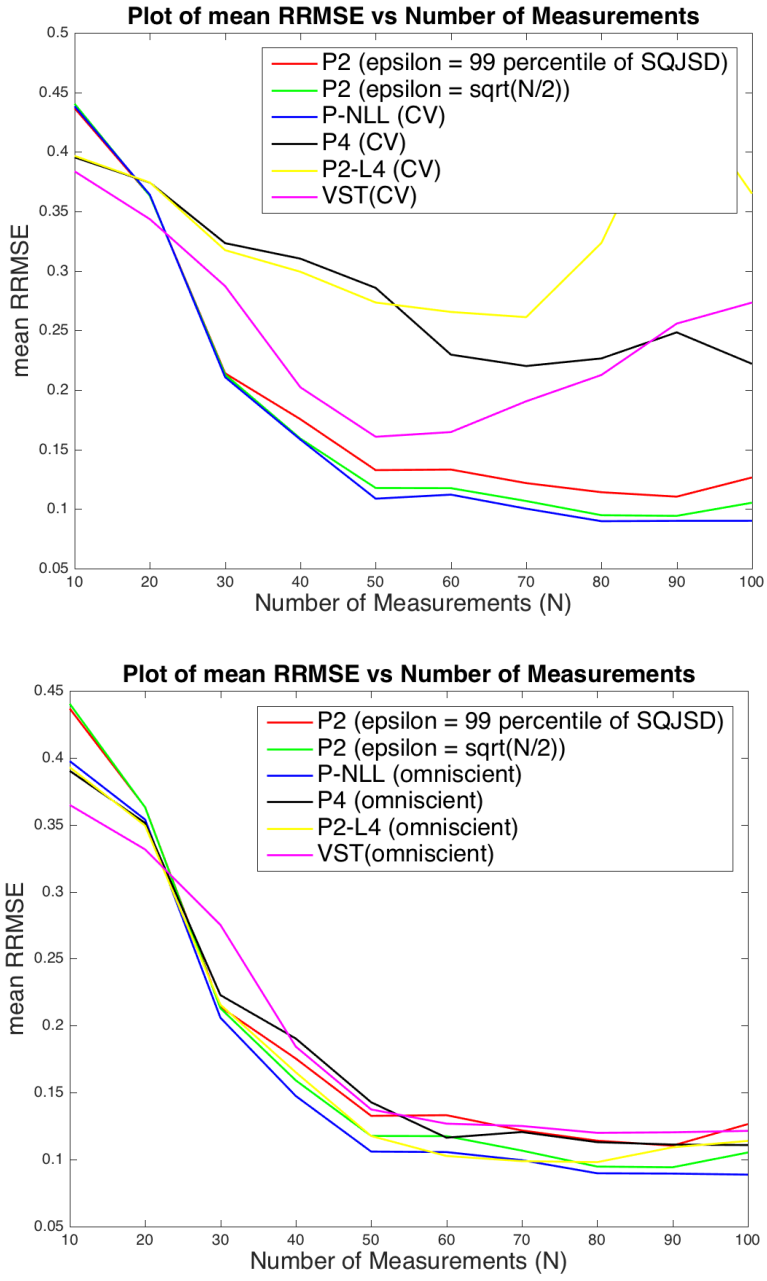

Figure 3: Results for *experiment on number of measurements*. Top: Mean of RRMSE versus N for the following problems: (P2) with $\varepsilon = \sqrt{N/2}$, (P2) with $\varepsilon = 99$ percentile of SQJSD values, and following estimators with ρ by cross-validation (CV): (P4), (P-NLL), (P2-L4) and (P-VST). Bottom: Mean of RRMSE versus N for the following problems: (P2) with $\varepsilon = \sqrt{N/2}$, (P2) with $\varepsilon = 99$ percentile of SQJSD values, and following estimators with omniscient ρ: (P4), (P-NLL), (P2-L4) and (P-VST). All results are for an ensemble of $Q = 100$ 1D signals of 100 elements, for fixed $I = 10^6$ and fixed signal sparsity $s = 10$ in the 1D-DCT basis.

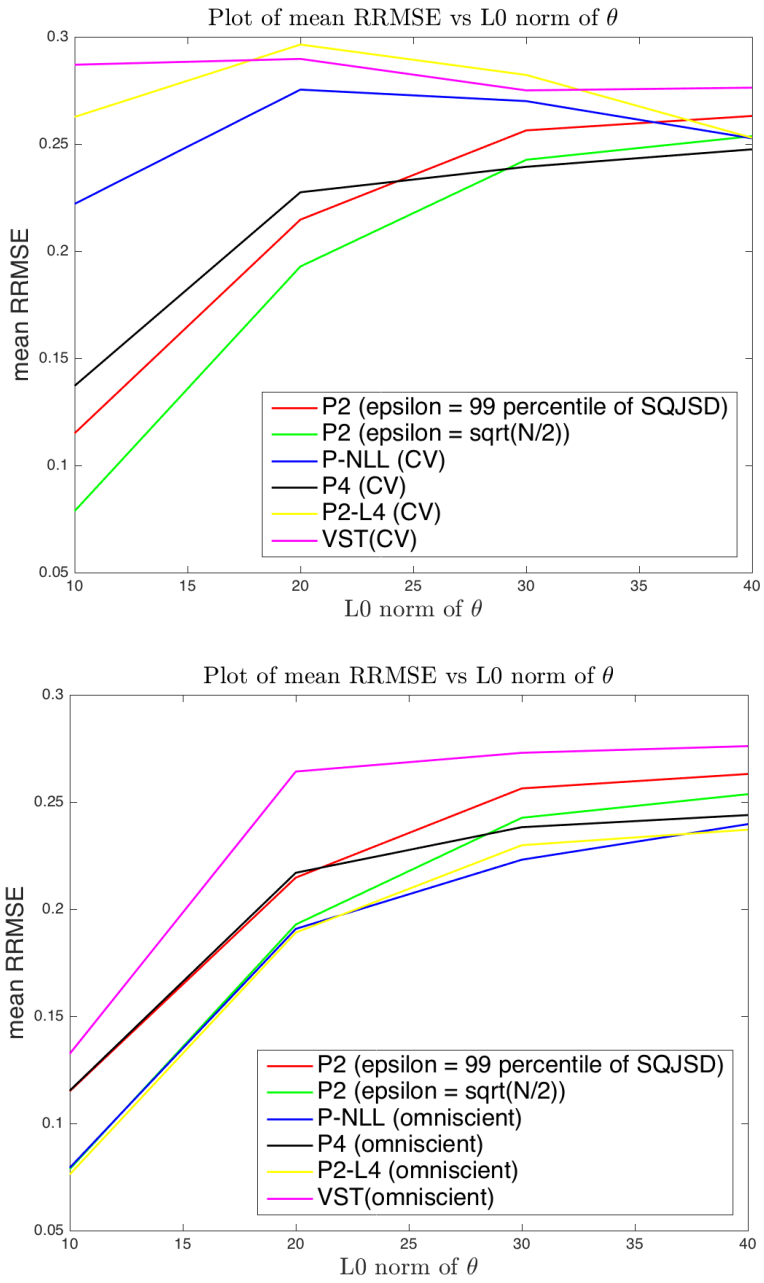

Figure 4: Results for *sparsity experiment*. Top: Mean of RRMSE versus s for the following problems: (P2) with $\varepsilon = \sqrt{N/2}$, (P2) with $\varepsilon = 99$ percentile of SQJSD values, and following estimators with ρ by cross-validation (CV): (P4) , (P-NLL), (P2-L4) and (P-VST). Bottom: Mean of RRMSE versus s for the following problems: (P2) with $\varepsilon = \sqrt{N/2}$, (P2) with $\varepsilon = 99$ percentile of SQJSD values, and following estimators with omniscient ρ: (P4), (P-NLL), (P2-L4) and (P-VST). All results are for an ensemble of $Q = 100$ 1D signals of 100 elements, for fixed $I = 10^6$ and fixed $N = 50$. Signal sparsity is in the 1D-DCT basis.

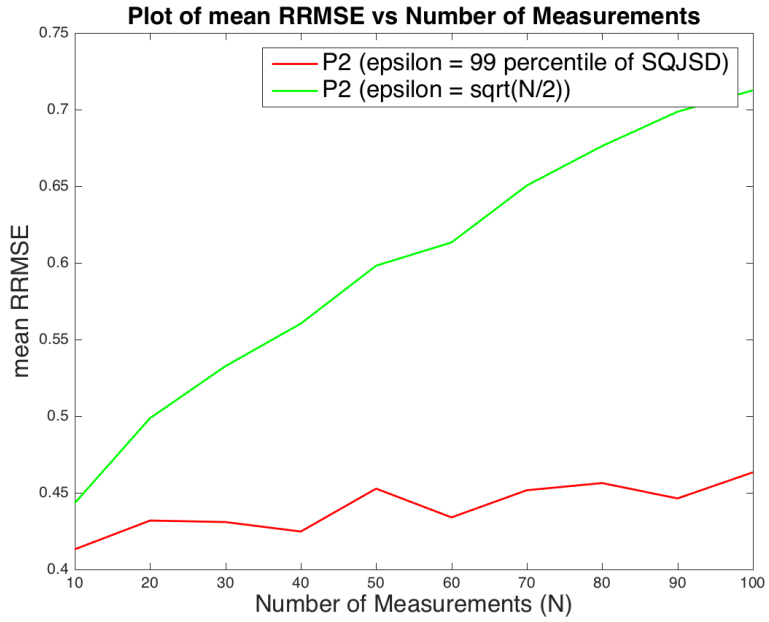

Figure 5: Results for low-intensity experiment with increase in N: The RRMSE for P2 can increase with N. In this experiment, the signal was DCT-sparse with $s = 10, m = 10^3, I = 10^3$.

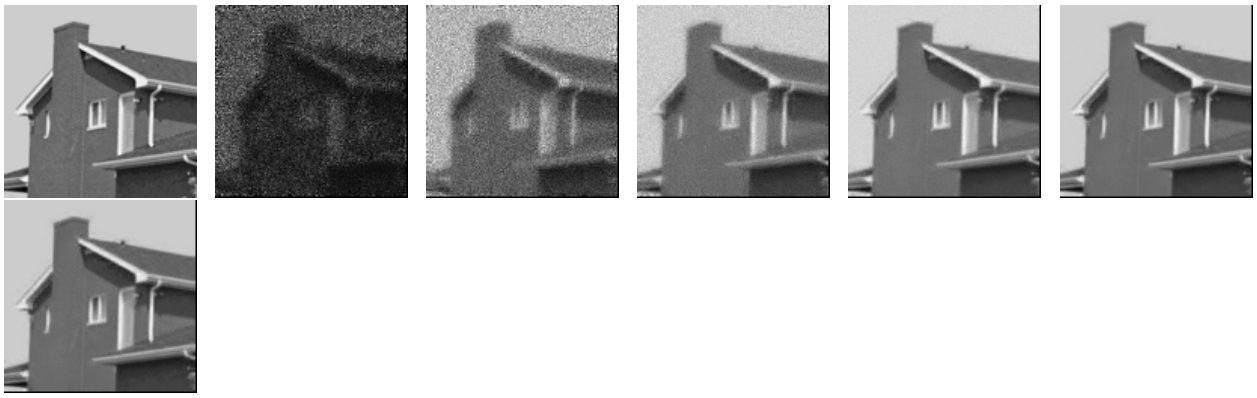

Figure 6: Sample reconstruction results for Poisson-corrupted compressed measurements of an image using (P2) with $\varepsilon = \sqrt{N/2}$ and a 2D-DCT basis. Left to right, top to bottom: original image, reconstructions for $I = 10^5$, $I = 10^6$, $I = 10^7$, $I = 10^8$, $I = 10^9$, $I = 10^{10}$. The respective relative reconstruction errors (RRMSE) are 0.17, 0.11, 0.092, 0.089, 0.0885, 0.0884. Refer to Section 3 for more details.

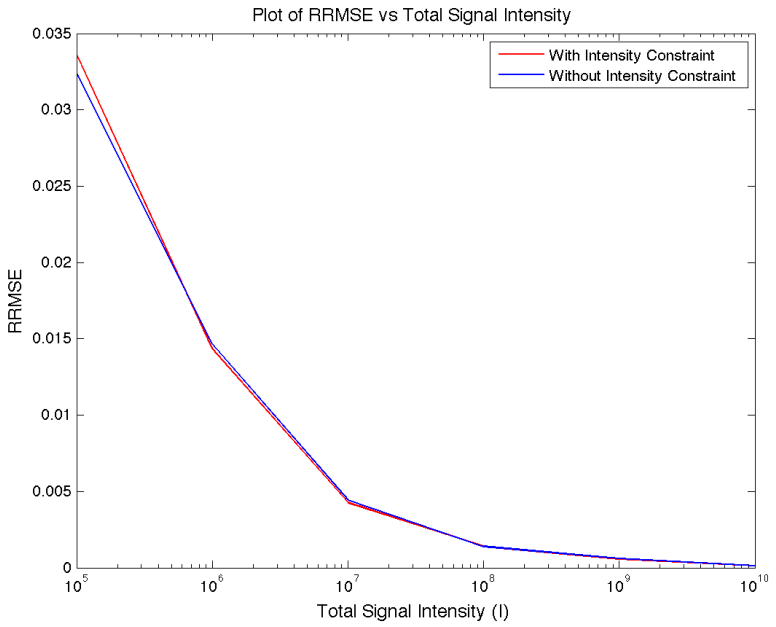

Figure 7: RRMSE comparison for (P2) with and without imposition of the $\|\boldsymbol{x}^\star\|_1 = I$ constraint.

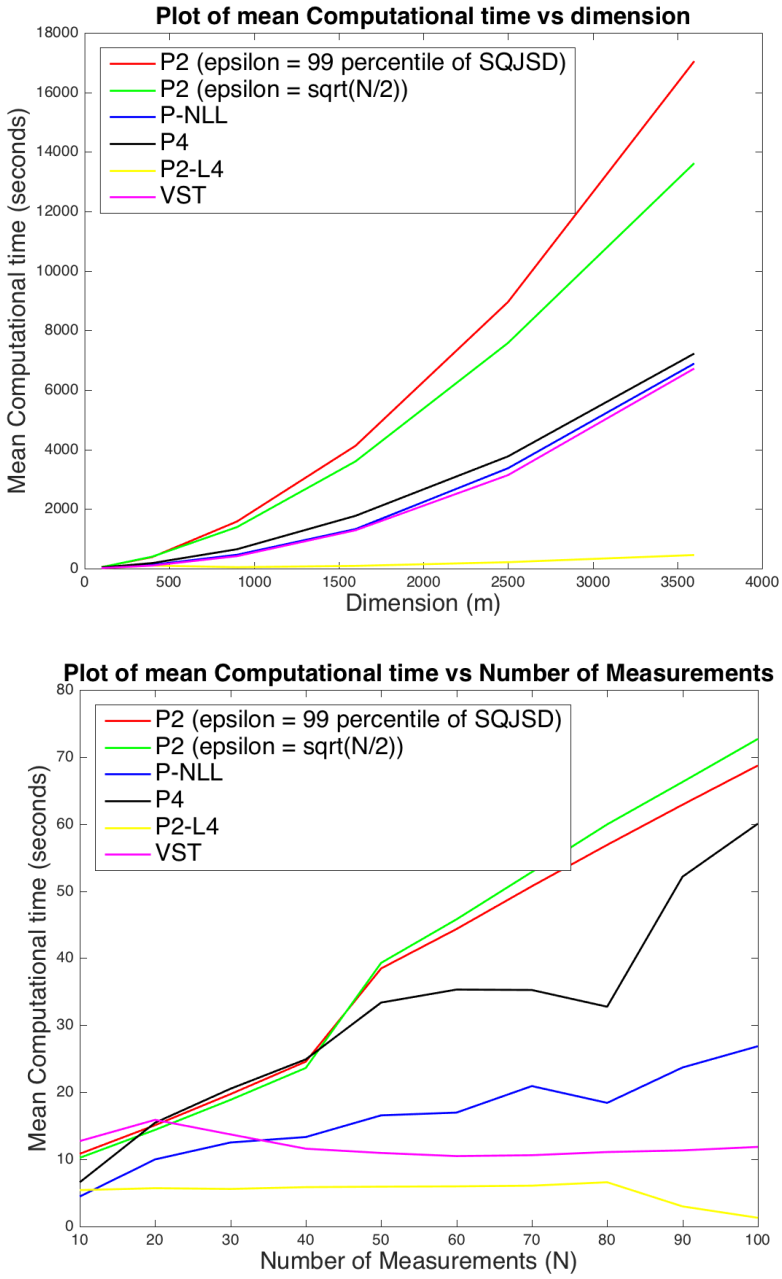

Figure 8: Comparison of the running times of various estimators w.r.t. m with $N = m/2$ (top), and w.r.t. N for fixed $m = 100$ (bottom). For all competing estimators including (P4), the time is recorded only for a single value of ρ without any cross-validation.