Variance Stabilization Based Compressive Inversion under Poisson or Poisson-Gaussian Noise with Analytical Bounds

Pakshal Bohra[1], Deepak Garg[2], Karthik S. Gurumoorthy[3] and Ajit Rajwade[2]‡

[1]Department of Electrical Engineering, IIT Bombay
[2]Department of Computer Science and Engineering, IIT Bombay
[3]International Center for Theoretical Sciences, Bengaluru

E-mail: pakshalbohra@gmail.com, 19deepak94@gmail.com,karthik.gurumoorthy@icts.res.in, ajitvr@cse.iitb.ac.in

Abstract. Most existing bounds for signal reconstruction from compressive measurements make the assumption of additive signal-independent noise. However in many compressive imaging systems, the noise statistics are more accurately represented by Poisson or Poisson-Gaussian noise models. In this paper, we derive upper bounds for signal reconstruction error from compressive measurements which are corrupted by Poisson or Poisson-Gaussian noise. The features of our bounds are as follows: (1) The bounds are derived for a computationally tractable convex estimator with statistically motivated parameter selection. The estimator penalizes signal sparsity subject to a constraint that imposes a novel statistically motivated upper bound on a term based on variance stabilization transforms to approximate the Poisson or Poisson-Gaussian distributions by distributions with (nearly) constant variance. (2) The bounds are applicable to signals that are sparse as well as compressible in any orthonormal basis, and are derived for compressive systems obeying realistic constraints such as non-negativity and flux-preservation. Our bounds are motivated by several properties of the variance stabilization transforms that we develop and analyze. We present extensive numerical results for signal reconstruction under varying number of measurements and varying signal intensity levels. Ours is the first piece of work to derive bounds on compressive inversion for the Poisson-Gaussian noise model. We also use the properties of the variance stabilizer to develop a principle for selection of the regularization parameter in penalized estimators for Poisson and Poisson-Gaussian inverse problems.

‡ PB and DG are first authors with equal contribution. Corresponding author is AR.
AR acknowledges support from IITB Seed Grant 14IRCCSG012 and thanks NVIDIA corporation for the generous donation of a TitanX GPU. KSG acknowledges the support of the AIRBUS Group Corporate Foundation Chair in Mathematics of Complex Systems established in ICTS-TIFR.

1. Introduction

Compressed sensing (CS) is a flourishing branch of signal processing with many theoretical and algorithmic advances, along with emerging applications in the form of actual systems in medicine (especially MRI acquisition), astronomy, photography and various other fields. The basic philosophy is to efficiently acquire signals by reducing the number of measurements, and reconstruct the signal from this reduced set later. Theoretical bounds for performance of compressive reconstruction algorithms have shown great promise [1]. 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 or RIP (i.e. the sensing matrix approximately preserves the magnitude of sparse vectors) 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 (G1),

$$(\mathsf{G1})\text{: minimize}\|\boldsymbol{z}\|_1 \text{ such that } \|\boldsymbol{y} - \boldsymbol{Az}\|_2 \leq \varepsilon, \tag{2}$$

and C_1 and C_2 are constants dependent only on δ_{2s}, the so-called restricted isometry constant (RIC) of $\boldsymbol{A}$ of order s. 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 $\delta_{2s} < 1$. Note that the RIC of order s of matrix $\boldsymbol{A}$ is defined as the smallest value δ_s for which the following is true for all s-sparse signals $\boldsymbol{\theta}$: $(1 - \delta_s)\|\boldsymbol{\theta}\|_2^2 \leq \|\boldsymbol{A\theta}\|_2^2 \leq (1 + \delta_s)\|\boldsymbol{\theta}\|^2$. In other words, the matrix $\boldsymbol{A}$ obeys RIP if it approximates preserves the squared magnitude of all s-sparse signals $\boldsymbol{\theta}$. An intuition behind this property is as follows: If $\boldsymbol{A}$ were to obey RIP of order $2s$, it implies that no $2s$-sparse signal would lie in its null-space. Hence for two different signals $\boldsymbol{\theta^{(1)}}$ and $\boldsymbol{\theta^{(2)}}$, we would necessarily have $\boldsymbol{A\theta^{(1)}} \neq \boldsymbol{A\theta^{(2)}}$, implying that unique recovery of $\boldsymbol{\theta}$ from $\boldsymbol{y}$ and $\boldsymbol{A}$ is possible. Ideally, one desires that δ_{2s} be as close to 0 as possible, within the limits imposed on $\boldsymbol{\Phi}$ by the imaging system.

The aforementioned bounds are based on the assumption of additive signal independent noise. However the noise in many compressive imaging systems can be more accurately described as Poisson-Gaussian. The Poisson component, which is signal dependent, is typically known to emerge from photon-counting principles in the acquisition of signals. The Gaussian component is signal-independent and is due to

fluctuations in the electronic parts of the imaging system. The Poisson component is quite dominant particularly at lower signal intensities [2], and is a non-additive form of noise. Given a non-negative signal $\boldsymbol{x} \in \mathbb{R}^m$ and a compressive measuring device with a non-negative sensing matrix $\boldsymbol{\Phi} \in \mathbb{R}^{N \times m}$, $N \ll m$, the measurement vector $\boldsymbol{y} \in \mathbb{R}^N$ can be described as follows:

$$\boldsymbol{y} \sim \alpha \text{Poisson}(\boldsymbol{\Phi x}) + \boldsymbol{\eta}, \boldsymbol{\eta} \sim \mathcal{N}(g, \sigma^2), \tag{3}$$

where α represents a gain factor, and g, σ represent the mean and standard deviation of the Gaussian component respectively. The Gaussian component of the noise cannot be ignored, and such a mixed Poisson-Gaussian noise model is ubiquitous in imaging systems in astronomy [3], microscopy [4] and compressive imagers such as the Rice Single Pixel camera [5, 6], to name a few.

There exists a large amount of literature on denoising of signals or images under Poisson-Gaussian noise. For instance, recent work in [7] denoises and deblurs images using an exact Poisson-Gaussian likelihood, which is approximated in a very principled way during an iterative optimization. Earlier work on image denoising using this model includes approximations based on variance stabilization transforms [3] or PURElet-based approaches [8], among others. However, the Poisson-Gaussian noise model has not been presented heretofore in the context of CS, and in particular with a derivation of performance bounds. There does exist fairly recent literature on performance bounds for CS under purely Poisson noise using either the penalized Poisson negative log-likelihood (hereafter referred to as PNLL) or the LASSO (see Section 6 for a detailed discussion), or using least squares estimation for Poisson inverse problems with $N > m$ [9]. Efficient algorithms have also been proposed for Poisson CS [10, 11, 12, 13] or Poisson deconvolution [14, 15]. A comprehensive survey of algorithms and applications of Poisson inverse problems has been presented in [16].

In this paper, we derive performance bounds for CS under Poisson noise using a variance stabilization transform (VST) approach. As has been shown in [17], if $y \sim \text{Poisson}(\lambda)$, then $\sqrt{y + \frac{3}{8}}$ has variance approximately $\frac{1}{4}$ and mean $\sqrt{\lambda + \frac{3}{8}}$ when $\lambda \to \infty$. This motivates the following objective function for compressive inference:

$$\min \|\boldsymbol{\theta}\|_1 \text{ subject to } \|\sqrt{\boldsymbol{y} + c} - \sqrt{\boldsymbol{\Phi\Psi\theta} + c}\|_2 \leq \varepsilon, \boldsymbol{\Psi\theta} \succeq \boldsymbol{0} \tag{4}$$

where $\boldsymbol{\Psi}$ is a $m \times m$ orthonormal basis in which the signal $\boldsymbol{x}$ yields a sparse set of coefficients $\boldsymbol{\theta} = \boldsymbol{\Psi}^T \boldsymbol{x}$, c is a coefficient that defines the VST (*e.g.*, $c = \frac{3}{8}$ for the Anscombe transform) and the symbol $\succeq$ in $\boldsymbol{a} \succeq \boldsymbol{b}$ means that $a_i \geq b_i$ for every index i in vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. Here ε is a statistically motivated upper bound on $\|\sqrt{\boldsymbol{y} + c} - \sqrt{\boldsymbol{\Phi\Psi\theta} + c}\|_2$ that we derive later in this paper. We also extend these bounds to the case of Poisson-Gaussian noise. Moreover, we use variance stabilizers for a particular aspect of all Poisson or Poisson-Gaussian inverse problems (i.e. not restricted to just CS but other problems such as deblurring) - that of choice of the regularization parameter in penalized estimators. We develop a statistically motivated principle for this purpose, which works well in practice.

The contribution of our work is summarized as follows:

(i) To the best of our knowledge, this is the first piece of work to provide performance bounds for CS under Poisson-Gaussian noise. In fact, we have a unified approach to handle Poisson as well as Poisson-Gaussian noise, and it can be easily extended to Poisson-Gaussian-uniform quantization noise.

(ii) Our bounds apply to a computationally tractable and probabilistically motivated estimator, under realistic CS matrices, and for sparse or compressible signals in any orthonormal basis. A detailed comparison with earlier work is presented in Section 6.

(iii) Due to the VST, our estimator allows for very principled, statistically motivated parameter tuning, since the term $\|\sqrt{\boldsymbol{y}+c}-\sqrt{\boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{\theta}+c}\|_2^2$ is a metric and since (as we show later in the paper) the magnitude of the difference term, *i.e.* $\|\sqrt{\boldsymbol{y}+c}-\sqrt{\boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{\theta}+c}\|_2$, has a bounded variance which does not depend on the original signal or the number of measurements. This bounded variance property of $\|\sqrt{\boldsymbol{y}+c}-\sqrt{\boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{\theta}+c}\|_2$, which is a *major contribution* of this work, leads to a neat concentration inequality. The statistically motivated parameter tuning in our work is different from the case of the PNLL which is not a metric, which does not have a signal-independent value, and where choosing the regularization parameter for signal sparsity is not easy in practice. Again, see Section 5.1 and 6.

(iv) We develop a statistically motivated principle, based on variance stabilization transforms, to choose the regularization parameter for penalized estimators for any Poisson or Poisson-Gaussian inverse problem.

A part of this work earlier appeared in our conference paper [18], but this work contains an extension to the Poisson-Gaussian case, as well as many refinements to the theory and experiments for the Poisson noise case.

This paper is organized as follows. Some preliminaries are presented in Sec. 2, the main theoretical results are derived in Sec. 3 and Sec. 4 along with a discussion, numerical results are presented in Section 5, followed by a summary of the contributions, a more detailed comparison with existing work and directions for future work in Section 6.

2. Preliminaries

In this section, we go over some preliminary concepts briefly, so as to make the paper self-contained.

2.1. Construction of Sensing Matrices

We construct a sensing matrix $\boldsymbol{\Phi}$ that corresponds to the forward model of a real optical system, based on the approach in [19]. Clearly $\boldsymbol{\Phi}$ has to satisfy certain constraints natural to a realizable imaging system - non-negativity and flux preservation. The

latter is due to the fact that the total photon-count of the noise-free measurement $\boldsymbol{\Phi x}$ can never exceed that of the original signal $\boldsymbol{x}$, *i.e.*, $\sum_{i=1}^{N}(\boldsymbol{\Phi x})_i \leq \sum_{k=1}^{m} x_k$. This in turn imposes the constraint that every column of $\boldsymbol{\Phi}$ must sum up to a value no more than 1, i.e. $\forall j, \sum_{i=1}^{N} \Phi_{ij} \leq 1$.

One major difference between Poisson CS and conventional CS emerges from the fact that conventional randomly generated sensing matrices which obey restricted isometry (RIP) do not follow the aforementioned physical constraints. This is a drawback as the RIP is a well-known sufficient condition which guarantees bounds on compressive recovery. We now construct a sensing matrix $\boldsymbol{\Phi}$ which has only zero or (scaled) ones as entries. Let us define p to be the probability that a matrix entry is 0, then $1-p$ is the probability that the matrix entry is a scaled 1. Let $\boldsymbol{Z}$ be a $N \times m$ matrix whose entries $Z_{i,j}$ are i.i.d random variables taking only these two different values, *i.e.*,

$$Z_{i,j} = \begin{cases} -\sqrt{\dfrac{1-p}{p}} & \text{with probability } p, \quad (5a) \\ \sqrt{\dfrac{p}{1-p}} & \text{with probability } 1-p. \quad (5b) \end{cases}$$

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 as $1 - 2e^{-Nc(1+\delta_{2s})}$ where δ_{2s} is its RIC of order $2s$ and function $c(h) \triangleq \dfrac{h^2}{4} - \dfrac{h^3}{6}$ [20]. In other words, for any $2s$-sparse signal $\boldsymbol{\rho}$, the following holds with high probability

$$(1-\delta_{2s})\|\boldsymbol{\rho}\|_2^2 \leq \|\tilde{\boldsymbol{\Phi}}\boldsymbol{\rho}\|_2^2 \leq (1+\delta_{2s})\|\boldsymbol{\rho}\|_2^2. \quad (6)$$

Given any orthonormal matrix $\boldsymbol{\Psi}$, arguments in [20] 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 non-negative sensing matrix $\boldsymbol{\Phi}$ from $\tilde{\boldsymbol{\Phi}}$ as follows [19]:

$$\boldsymbol{\Phi} = \sqrt{\frac{p(1-p)}{N}}\tilde{\boldsymbol{\Phi}} + \frac{(1-p)}{N}\mathbf{1}_{N\times m}, \quad (7)$$

which ensures that each entry of $\boldsymbol{\Phi}$ is either 0 or $\dfrac{1}{N}$. One can easily check that $\boldsymbol{\Phi}$ satisfies both the non-negativity as well as flux-preservation properties.

We note that the model for sensing matrices presented here, has been used in a variety of compressed sensing applications. For example, it has been used for compressive acquisition of astronomical images in [21, 22], for flourescence microscopy in [23], for depth mapping in [24] and for remote sensing in [25]. The core architecture for all these applications follows the model of the Rice Single Pixel camera [6] (i.e. the model for $\boldsymbol{Z}$ here, or equivalently the model for $\boldsymbol{\Phi}$ here, with a flux-preserving constraint), where the point-wise multiplication with binary sensing patterns are modelled by an

array of mirrors that are turned on (= +1) or off (= 0). The final addition to obtain the dot product between the rows of $\boldsymbol{\Phi}$ and the signal $\boldsymbol{x}$ are implemented by means of a diode.

2.2. Variance Stabilization Transforms

VSTs are a popular method of converting Poisson data into data that are approximately Gaussian. In particular, [17] proves that if $y \sim \text{Poisson}(\lambda)$, then we have the following:

$$E(\sqrt{y+c}) = \sqrt{\lambda+c} - \frac{1}{8\sqrt{\lambda}} + \mathcal{O}(\lambda^{-1.5}) \tag{8}$$

$$\text{Var}(\sqrt{y+c}) = \frac{1}{4} + \frac{3-8c}{32\lambda} + \mathcal{O}(\lambda^{-2}). \tag{9}$$

Setting $c = \frac{3}{8}$ yields the so-called Anscombe Transform (AT) and produces data with a 'stable' noise variance of approximately $\frac{1}{4}$ and a mean of approximately $\sqrt{\lambda+c}$. The higher order moments are approximately zero for a reasonably large λ. The approximation to the mean is further expressed as $\sqrt{\lambda}$ in some papers [16]. All these approximations improve as λ grows beyond 4, and the noise distribution becomes closer and closer to $\mathcal{N}(0, \frac{1}{4})$ as shown rigorously in [26]. In the case of Poisson-Gaussian noise, i.e. when $y \sim \alpha\text{Poisson}(\lambda)+\eta$ where $\eta \sim \mathcal{N}(g, \sigma^2)$, the AT is replaced by the Generalized AT (GAT) which is given as $t = \frac{1}{\alpha}\sqrt{\alpha y + \frac{3}{8}\alpha^2 + \sigma^2 - \alpha g}$. As λ grows in value, it can be shown [3] that t has a mean of $\sqrt{\lambda + \frac{3}{8}\alpha + \frac{\sigma^2-\alpha g}{\alpha}}$ and variance of approximately $\frac{1}{4}$. In this paper, we keep $\alpha = 1, g = 0$ for simplicity, although our framework is general enough to handle deviations from this assumption.

3. Theory

The main theoretical development is presented in this section. First, for noisy measurements $\boldsymbol{y} \sim \text{Poisson}(\boldsymbol{\Phi x})$, we prove that the quantity $R(\boldsymbol{y}, \boldsymbol{\Phi x}) \triangleq \|\sqrt{\boldsymbol{y}+c} - \sqrt{\boldsymbol{\Phi x}+c}\|_2$ (henceforth called the 'residual magnitude') has a mean which is $\mathcal{O}(\sqrt{N})$ and a variance which is constant (independent of the signal $\boldsymbol{x}$ and also suprisingly independent of the number of measurements N) as long as $\boldsymbol{\Phi x} \succeq \beta\mathbf{1}$ where $\beta > 0.125$. This result is extended to the case of Poisson-Gaussian noise. Using these results, we then state and prove two theorems for upper error bounds for the reconstruction of a signal from Poisson corrupted CS measurements in a realistic system as per Eqn. 7. For the case of Poisson-Gaussian CS, we present and prove two more theorems. An extensive discussion on the theorem statements is presented. The proofs of the theorems on error bounds follow the broad techniques from [1].

3.1. Theorem for Properties of the Residual Magnitude

The theorem we present in this section was inspired by our simulations with the quantity $R(\boldsymbol{y}, \boldsymbol{\Phi x})$ defined above. We simulated Poisson-corrupted CS measurements

$\boldsymbol{y} \sim \text{Poisson}(\boldsymbol{\Phi x})$ for sensing matrix $\boldsymbol{\Phi} \in \mathbb{R}^{N\times m}$ as per Eqn. 7 and for a non-negative signal $\boldsymbol{x}$ of $m = 1000$ dimensions. Each element of $\boldsymbol{x}$ was generated independently from $\text{Unif}(0, 1)$, followed by rescaling to ensure that the signal intensity $I \triangleq \|\boldsymbol{x}\|_1$ was 1000 (i.e. we divided $\boldsymbol{x}$ by $\|\boldsymbol{x}\|_1$ and multiplied the result by $I = 1000$). The chosen values of N were from 20 to 6000. For each N, we executed 2000 trials keeping $\boldsymbol{\Phi}, \boldsymbol{x}$ fixed. We empirically observed that $E[R(\boldsymbol{y}, \boldsymbol{\Phi x})]$ was $\mathcal{O}(\sqrt{N})$, i.e. independent of I. We also observed that $\text{Var}[R(\boldsymbol{y}, \boldsymbol{\Phi x})]$ was upper bounded by a small constant value around 0.14 independent of both I and N. We repeated this experiment for a fixed $N = 500$, a fixed $\boldsymbol{x}/\|\boldsymbol{x}\|_1$, but varying I from 10^2 to 10^9 in powers of 10 (i.e. each time we divided the signal $\boldsymbol{x}$ by $\|\boldsymbol{x}\|_1$ and multiplied the result by I). Again, we observed the same properties of $E[R(\boldsymbol{y}, \boldsymbol{\Phi x})]$ and $\text{Var}[R(\boldsymbol{y}, \boldsymbol{\Phi x})]$. Moreover, we observed that the empirical CDF of the values of $R(\boldsymbol{y}, \boldsymbol{\Phi x})$ was similar to a Gaussian. These results are shown in Fig. 1. These results were independent of the specific instances of $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\Phi}$.

Theorem 1: Let $\boldsymbol{y} \in \mathbb{Z}_+^N$ be a vector of independent CS 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. 7 and $\boldsymbol{x} \in \mathbb{R}^m$ is a non-negative signal. Define $\gamma_i \triangleq (\boldsymbol{\Phi x})_i$. Then we have:

(i) $E[R(\boldsymbol{y}, \boldsymbol{\Phi x})] \leq \sqrt{N/2}$

(ii) Define $v \triangleq \text{Var}[R(\boldsymbol{y}, \boldsymbol{\Phi x})]$. Then if $\forall i, \gamma_i > \beta \triangleq 1/2 - c$ and $N \geq \dfrac{3/8}{(\frac{\beta}{4(\beta+c)} - \frac{\beta}{8(\beta+c)^2})^2}$, we have

(a) $v \leq \dfrac{\sum_{i=1}^N \frac{\gamma_i(1+3\gamma_i)}{4(\gamma_i+c)^2}}{\sum_{i=1}^N \frac{\gamma_i}{4(\gamma_i+c)} - \frac{\gamma_i}{8(\gamma_i+c)^2}} \leq \bar{v} \triangleq \dfrac{3/4}{(\frac{\beta}{4(\beta+c)} - \frac{\beta}{8(\beta+c)^2})}$

(b) $P\Big(R(\boldsymbol{y}, \boldsymbol{\Phi x}) \leq \sqrt{N}(\frac{1}{\sqrt{2}} + \sqrt{\bar{v}})\Big) \geq 1 - 1/N.$

(iii) Specifically, if $\forall i, \gamma_i \geq 1$ and $N \geq 29$, then the results in (ii) become:

(a) $v \leq \frac{3N/4}{N(2c+1)/(8(1+c)^2)} \lessapprox 6.48$

(b) $P\Big(R(\boldsymbol{y}, \boldsymbol{\Phi x}) \leq \sqrt{N}(\frac{1}{\sqrt{2}} + 2.545)\Big) \geq 1 - 1/N.$ ♢

All statements of this theorem are proved in Section 7.1. We make a few comments below:

(i) $E[R(\boldsymbol{y}, \boldsymbol{\Phi x})]$ has a signal-independent upper bound.

(ii) The upper bound $\bar{v}$ on the variance is signal independent (also see statement 3(a)) and independent of N. This property is *not* shared by the PNLL. In Lemma 2 of [28], it is shown that an approximate form of the PNLL (APNLL) expressible in our context as $APNLL(\boldsymbol{y}, \boldsymbol{\Phi x}) \triangleq \sum_{i=1}^N y_i \log(y_i/[\boldsymbol{\Phi x}]_i) + [\boldsymbol{\Phi x}]_i - y_i$ obeys the property that $E[APNLL(\boldsymbol{y}, \boldsymbol{\Phi x})] \approx N/2$. This is because the expected value of each term in the summation is approximately 0.5. However, the variance of $APNLL(\boldsymbol{y}, \boldsymbol{\Phi x})$ cannot be bounded by a constant independent of N, unlike the case with $R(\boldsymbol{y}, \boldsymbol{\Phi x})$. This can be seen in the last four sub-figures of Fig. 1. Hence in our framework, $R(\boldsymbol{y}, \boldsymbol{\Phi x})$ enjoys stronger theoretical properties compared to $APNLL(\boldsymbol{y}, \boldsymbol{\Phi x})$. Later in Sec. 4 and 5.3, we develop a similar principle for choosing

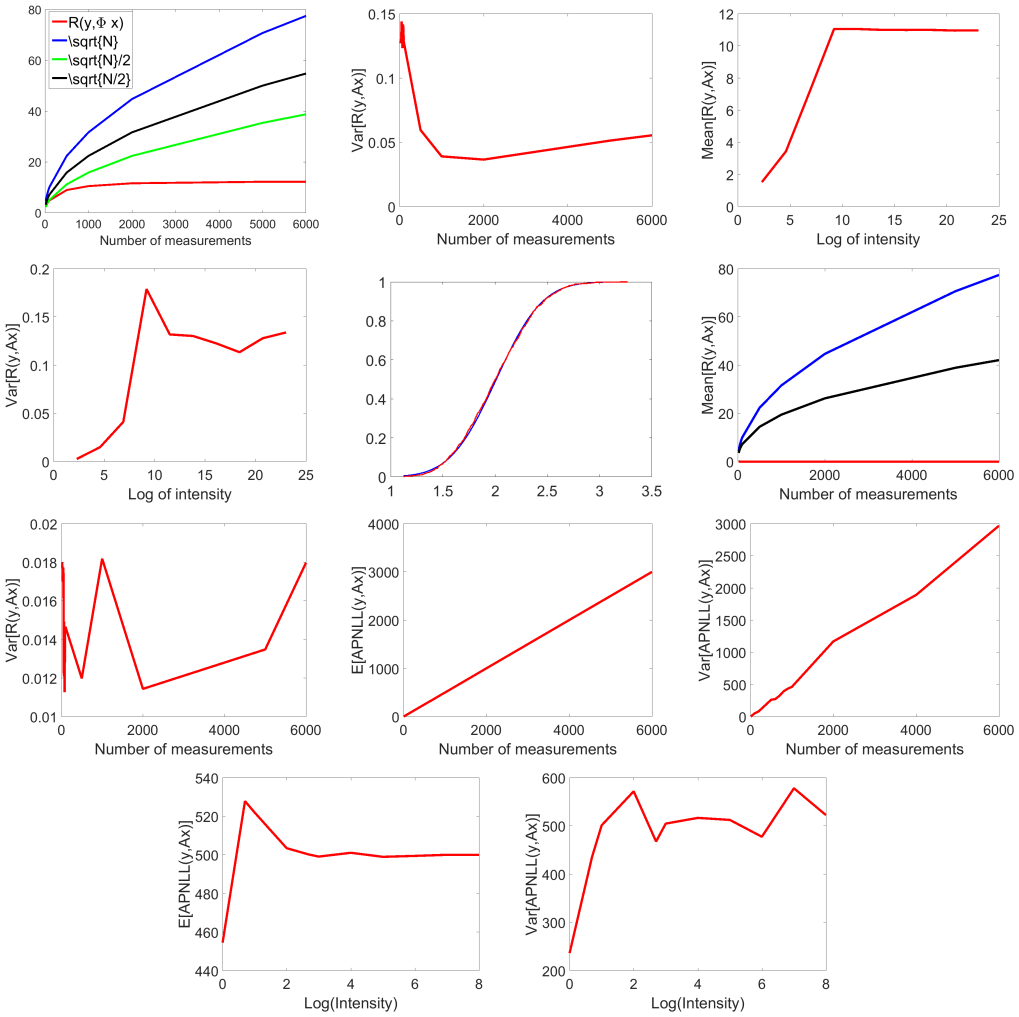

Figure 1: In the left to right, top to bottom order. First two sub-figures: Plot of mean and variance of the values of $R(\boldsymbol{y}, \boldsymbol{\Phi x})$ versus N for a fixed $I = 10^3$ for a signal of dimension $m = 1000$. Third and fourth sub-figures: Plot of mean and variance of the values of $R(\boldsymbol{y}, \boldsymbol{\Phi x})$ versus $\log(I)$ for a fixed $N = 500$ for a signal of dimension $m = 1000$. Fifth sub-figure: Empirical CDF of $R(\boldsymbol{y}, \boldsymbol{\Phi x})$ (red curve) for $N = 20, I = 10^3, m = 1000$ compared to a Gaussian CDF (blue curve) with mean and variance equal to that of the values of $R(\boldsymbol{y}, \boldsymbol{\Phi x})$. The curves overlap significantly as the empirical CDFs are very close. Sixth and seventh sub-figures: Same as first two sub-figures but with a signal of $I = 0.1$ and $m = 1000$. Eight and ninth: Mean and variance of $APNLL(\boldsymbol{y}, \boldsymbol{\Phi x})$ w.r.t. N for $I = 10^3, m = 10^3$. Tenth and eleventh: Mean and variance of $APNLL(\boldsymbol{y}, \boldsymbol{\Phi x})$ w.r.t. I for $N = 500, m = 10^3$. Scripts for reproducing these results are available at [27].

regularization parameters for penalized estimators in Poisson as well as Poisson-Gaussian inverse problems.

(iii) We would like to emphasize that the difference in the variance of $R(\boldsymbol{y}, \boldsymbol{\Phi x})$ and

$APNLL(\boldsymbol{y}, \boldsymbol{\Phi x})$ is not just an artifact due to scaling. That is, even though $APNLL(\boldsymbol{y}, \boldsymbol{\Phi x})/\sqrt{N}$ has a constant variance, its usage as a data fidelity term (via an estimator of the form $\min\|\boldsymbol{\theta}\|_1$ such that $APNLL(\boldsymbol{y}, \boldsymbol{\Phi\Psi\theta})/\sqrt{N} \leq \varepsilon$) will lead to unnecessarily loose performance bounds. This is because when deriving the performance bounds for compressed sensing (Theorem 2), the actual data fidelity term (related to the RIP) is the $APNLL$ and not $APNLL/\sqrt{N}$. Hence in step (ii)-a of the Proof of Theorem 2, the term ε on the RHS of Eqn. 36 will get scaled by $\sqrt{N}$, and so a mere scaling will not help§.

(iv) In practice, we have observed a smaller value of the upper bound on the variance than what is predicted in statement 2(a) or 3(a). The value is close to 0.14 *even when* the condition that $\forall i, \gamma_i > 1/2 - c$ is disobeyed. The assumption that $\gamma_i > 1/2 - c$, is not restrictive in most signal or image processing applications, except those that work with extremely low intensity levels. But in such cases the performance of Poisson CS is itself very poor due to the very low SNR [29]. However even for very low intensity signals for which the condition $\gamma_i > 1/2 - c$ is disobeyed for *every* measurement, we have empirically observed that this predicted upper bound on the variance is not violated. This can be observed from the last sub-figure of Fig. 1. In other words, we believe the imposed condition on γ_i is only sufficient for the variance bound, and *not* a necessary condition.

(v) The last statement of this theorem can be further tightened to yield a probability of $1 - 2e^{-N/2}$ by using the central limit theorem (CLT). Of course, the latter is an asymptotic result and hence for a finite value of N, it is an approximation. However, the approximation is empirically observed to be tight even for small $N \sim 20$ as confirmed by a Kolmogorov-Smirnov test even at 1% significance (see [27]). Further details can be found at the end of the proof in Section 7.1.

(vi) The bounds in this theorem do *not* assume (or require) that $\sqrt{\boldsymbol{y} + c} - \sqrt{\boldsymbol{\Phi x} + c}$ is Gaussian distributed. Indeed such an assumption would not be rigorous enough. This is because as shown in [26], the Gaussianity is obeyed only asymptotically when the mean of $\boldsymbol{y}$ tends to infinity.

3.2. Key Theorem for Poisson CS

Theorem 2 : Consider a non-negative signal $\boldsymbol{x}$ with total intensity $I \triangleq \|\boldsymbol{x}\|_1$ expressed using the orthornormal basis $\boldsymbol{\Psi}$ in the form $\boldsymbol{x} = \boldsymbol{\Psi\theta}$. Consider Poisson corrupted CS measurements of the form $\boldsymbol{y} \sim \text{Poisson}(\boldsymbol{\Phi x})$ where $\boldsymbol{\Phi}$ is constructed as per Eqn. 7. Define $\boldsymbol{A} \triangleq \boldsymbol{\Phi\Psi}$ so that $\boldsymbol{\Phi x} = \boldsymbol{A\theta}$. Let $\boldsymbol{\theta^\star}$ be the result of the following optimization problem:

$$\begin{aligned} &\mathsf{P1} : \min\|\boldsymbol{\theta}\|_1 \text{ such that } \|\sqrt{\boldsymbol{y} + c} - \sqrt{\boldsymbol{A\theta} + c}\|_2 \leq \varepsilon, \\ &\|\boldsymbol{\Psi\theta}\|_1 = I, \boldsymbol{\Psi\theta} \succeq \boldsymbol{0}, \end{aligned} \tag{10}$$

§ Another reason why we do not use $APNLL$ is that it does not obey the triangle inequality and cannot be used for Poisson-Gaussian noise.

where $\varepsilon \triangleq \sqrt{N}(\sqrt{\bar{v}} + 1/\sqrt{2})$ with $\bar{v} \triangleq \dfrac{3/4}{(\frac{\beta}{4(\beta+c)} - \frac{\beta}{8(\beta+c)^2})}$ as defined in Theorem 1 is a statistical upper bound (that holds with a high probability $1 - 1/N$) on the magnitude of the noise in the measurements *after* application of the AT. Let $\boldsymbol{\theta_s}$ denote a vector containing the s largest magnitude elements of $\boldsymbol{\theta}$ with the rest being 0. If $\widetilde{\boldsymbol{\Phi}}$ obeys RIP of order $2s$ with RIC $\delta_{2s} < \sqrt{2} - 1$, if the condition $\boldsymbol{\Phi x} \succeq \beta \mathbf{1}$ holds where $\beta > 1/2 - c$ and if $N \geq \dfrac{3/8}{(\frac{\beta}{4(\beta+c)} - \frac{\beta}{8(\beta+c)^2})^2}$, then we have for any $\kappa > 0$:

$$P\left(\frac{\|\boldsymbol{\theta} - \boldsymbol{\theta}^{\star}\|_2}{I} \leq C_1\sqrt{N}\tau\sqrt{\frac{1}{I} + \frac{cN}{I^2}} + \frac{C_2 s^{-\frac{1}{2}}\|\boldsymbol{\theta} - \boldsymbol{\theta_s}\|_1}{I}\right) \geq 1 - \kappa^2/N \text{ where } \tau \triangleq (\sqrt{\bar{v}}/\kappa + 1/\sqrt{2}).\diamondsuit \tag{11}$$

This theorem is proved in Section 7.2. Comments on this theorem follow.

Remarks on the Theorem and its Proof:

(i) If $\beta > 1$ and $N \geq 29$, then we have $v \leq 6.48$ as per statement 3 of Theorem 1. Further note that the condition $\boldsymbol{\Phi x} \succeq \beta \mathbf{1}$ is sufficient, but not necessary (as per our simulations).

(ii) The tighest upper bounds we have are for the case when $c = 0$, i.e. the original square-root VST developed by Bartlett [30], since the term $\sqrt{cN/I^2}$ would then disappear. We still chose $c = 3/8$ in the experiments as it gives better variance stabilization [17] and yields a cost function with a Lipschitz continuous derivative.

(iii) Our proof architecture is inspired from [1], but the points of departure are steps 2(a), 2(b), 2(c) as well as step 4(a) which gives a relationship between $\|\boldsymbol{Ah}\|_2$ and $\|\boldsymbol{Bh}\|_2$. These steps exploit the non-negativity and flux-preserving property of $\boldsymbol{\Phi}$, and the constraint $\|\boldsymbol{\Psi\theta}\|_1 = I$. See Section 7.2.

(iv) Given that we are dealing with a *Poisson* inverse problem, it is more intuitive to analyze the *relative* reconstruction error (RRE) rather than the (absolute) reconstruction error. This is because as the mean of the Poisson distribution increases, so does its variance, causing an increase in the mean squared error but a decrease in the relative mean squared error.

(v) Notice that our derived RRE bound is inversely proportional to the signal intensity I, which is typical in Poisson problems.

(vi) For a fixed I, if N is increased, the incident photon flux I is distributed across the N measurements, causing a decrease in SNR per measurement and possibly degrading performance. In fact, this affects the bounds in the $c \neq 0$ case, giving a scaling of $\mathcal{O}(N)$. This phenomenon differs from CS under Gaussian noise, and has earlier been noted in [19]. For $c = 0$, however, the flux-preserving nature of the matrix does not affect the bounds. Rather the $\sqrt{N}$ term is due to the fact that the variance of the noise after VST is a constant independent of N although there are N measurements. This is similar to Equation (17) of [31], where pure Gaussian noise

is considered. We conjecture that this behaviour is common to estimators such as (P1) with a constrained formulation, and is also exhibited by the quadratically constrained basis pursuit estimator from [31], [1]. Our method meticulously adapts the bounds from [1] for Poisson noise. Indeed, 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 [32] while comparing the quadratically constrained formulation with other estimators. There is currently no consensus in the literature as to whether this $\mathcal{O}(\sqrt{N})$ behaviour is a fundamental limit on the error bounds of such constrained estimators, 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 C_1 and C_2) 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.

(vii) As s increases, the restricted isometry constant (RIC) δ_{2s} of the sensing matrix will increase. Hence the factors C_1 and C_2, which are monotonically increasing functions of δ_{2s} for the domain $[0, 1]$, are also monotonically increasing in s. A precise mathematical formula for δ_{2s} (and hence for C_1, C_2) w.r.t. s is not straightforward and is an open problem in compressed sensing, to the best of our knowledge. Our bound is a function of C_1, C_2, N and s. Moreover the number of measurements N must satisfy $N \geq \mathcal{O}(s \log m)$ for RIP to hold. So, as s increases, the minimum number of measurements N required for the theoretical guarantees to hold, also increases. Therefore, it does not appear to be a straightforward task to theoretically establish the precise formula for our bound w.r.t. s. Furthermore, we would like to underscore that we are building upon the fundamental work in [1]. This issue regarding the variation of δ_{2s}, C_1, C_2 with respect to s is not exclusive to our work alone, and is also observed in [1]. As part of our future work, we would like to further understand this behavior and attempt to theoretically establish the same. However, we have experimentally observed that the reconstruction error does *increase* w.r.t. s.

(viii) Our experimental results in the next section show that the constraint $\|\boldsymbol{x}\|_1 = I$ is not necessary, although we required it *only* for our theoretical analysis.

(ix) The RRE bounds are also applicable to the Freeman-Tukey transform [33] given as $\sqrt{\boldsymbol{y}} + \sqrt{\boldsymbol{y}+1}$ with minor changes to the constant C_1.

(x) As has been mentioned earlier, the VST approximation is not so accurate for measurements with low mean, however at such low intensity levels Poisson CS is considered to be undesirable in itself [29].

(xi) It is tempting to treat $\sqrt{\boldsymbol{y}+c} - \sqrt{\boldsymbol{\Phi x}+c}$ as a Gaussian random variable, and

hence $R(\boldsymbol{y}, \boldsymbol{\Phi x})$ as a chi random variable. This would ignore the fact that the Gaussianity of the former has been established only asymptotically if all the values in $\boldsymbol{\Phi x}$ tend to ∞ [26]. However we have in practice seen that even for moderate values of $\boldsymbol{\Phi x}$, its distribution can be approximated very closely by a Gaussian as affirmed by Kolmogorov Smirnov hypothesis tests [27], even though we are unable to prove this theoretically. In fact, we have found no literature that establishes even the sub-Gaussianity or sub-exponentiality of $R(\boldsymbol{y}, \boldsymbol{\Phi x})$. Nonetheless, treating this approximation as exact allows us to improve the probability in the second part of the theorem from $1 - 1/N$ (for $\kappa = 1$) to $1 - 2e^{-N\tau}$ for an appropriately defined constant τ. If we treat ε as equal to the magnitude of a vector with elements drawn from $\mathcal{N}(0, \frac{1}{4})$, then ε^2 follows a chi distribution with N degrees of freedom. Hence, we can use tail bounds on the chi-square random variable [34] (Lemma 1) to arrive at the following bound:

$$P\Big(\frac{\|\boldsymbol{\theta} - \boldsymbol{\theta}^{\star}\|_2}{I} \leq C_{\mathrm{I}}\sqrt{N\tilde{\tau}}\sqrt{\frac{1}{I} + \frac{cN}{I^2}} + \frac{C_2 s^{-\frac{1}{2}}\|\boldsymbol{\theta} - \boldsymbol{\theta_s}\|_1}{I}\Big) \geq 1 - \exp(-N\tau)$$

for some $\tau > 0$ where $\tilde{\tau} \triangleq (1 + 2\tau + \sqrt{2\tau})$.

(xii) *Advantage of our estimator P1 over G1*: The major advantage of P1 over G1 is that G1, the tail bound on $\|\boldsymbol{y} - \boldsymbol{\Phi x}\|_2$ is signal-dependent because $\mathrm{Var}(y_i) = E(y_i) = (\boldsymbol{\Phi x})_i$. This unlike in our case where we have signal-independent tail bounds. One could however consider the normalized term $NL2(\boldsymbol{y}, \boldsymbol{\Phi x}) \triangleq \|(\boldsymbol{y} - \boldsymbol{\Phi x})./\sqrt{\boldsymbol{\Phi x}}\|_2$ where './' stands for pointwise division. We have observed in our experiments that tail-bounds based on $NL2(\boldsymbol{y}, \boldsymbol{\Phi x})$ are signal-independent. In fact, it can be easily shown that $\forall i, E[(y_i - (\boldsymbol{\Phi x})_i)^2/(\boldsymbol{\Phi x})_i] = 1$ and $\mathrm{Var}[(y_i - (\boldsymbol{\Phi x})_i)^2/(\boldsymbol{\Phi x})_i] = 2 + 1/(\boldsymbol{\Phi x})_i$. However, from the proof of Theorem 1 in Section 7, we see that $\forall i, E(\sqrt{y_i + c} - \sqrt{(\boldsymbol{\Phi x})_i + c}) \leq 0.5, \mathrm{Var}(\sqrt{y_i + c} - \sqrt{(\boldsymbol{\Phi x})_i + c}) \leq 0.75$. Hence we believe that the error bounds with P1 will be tighter than those with such a normalized ℓ_2-constrained estimator.

3.3. Theorem for Residual Magnitude in the Poisson-Gaussian case

Here, we state a theorem for the case of Poisson-Gaussian noise in the compressed measurements (with a known standard deviation for the Gaussian part of the noise), equivalent to Theorem 1 for Poisson noise. The proof can be found in Section 7.3. This theorem is inspired by experimentally observed behaviour of $R_d(\boldsymbol{y}, \boldsymbol{\Phi x}) \triangleq \|\sqrt{\boldsymbol{y} + d} - \sqrt{\boldsymbol{\Phi x} + d}\|_2$ where $d \triangleq c + \sigma^2$, which was quite similar to the Poisson case. That, is the mean of $R_d(\boldsymbol{y}, \boldsymbol{\Phi x})$ appeared to be $\mathcal{O}(\sqrt{N})$ and the variance appeared to be a constant independent of N, I, σ. This can be seen in Fig. 2.

Theorem 3 : Let $\boldsymbol{y}$ be a vector of N independent CS measurements such that $y_i \sim \mathrm{Poisson}[(\boldsymbol{\Phi x})_i] + \eta_i$ where $\boldsymbol{\Phi} \in \mathbb{R}^{N \times m}$ is a non-negative flux-preserving matrix as per Eqn. 7, $\boldsymbol{x} \in \mathbb{R}^m$ is a non-negative signal and $\eta_i \sim \mathcal{N}(0, \sigma^2)$. Define $\gamma_i \triangleq (\boldsymbol{\Phi x})_i$, $d \triangleq c + \sigma^2$ and $R_d(\boldsymbol{y}, \boldsymbol{\Phi x}) \triangleq \|\sqrt{\boldsymbol{y} + d} - \sqrt{\boldsymbol{\Phi x} + d}\|_2$. Then we have:

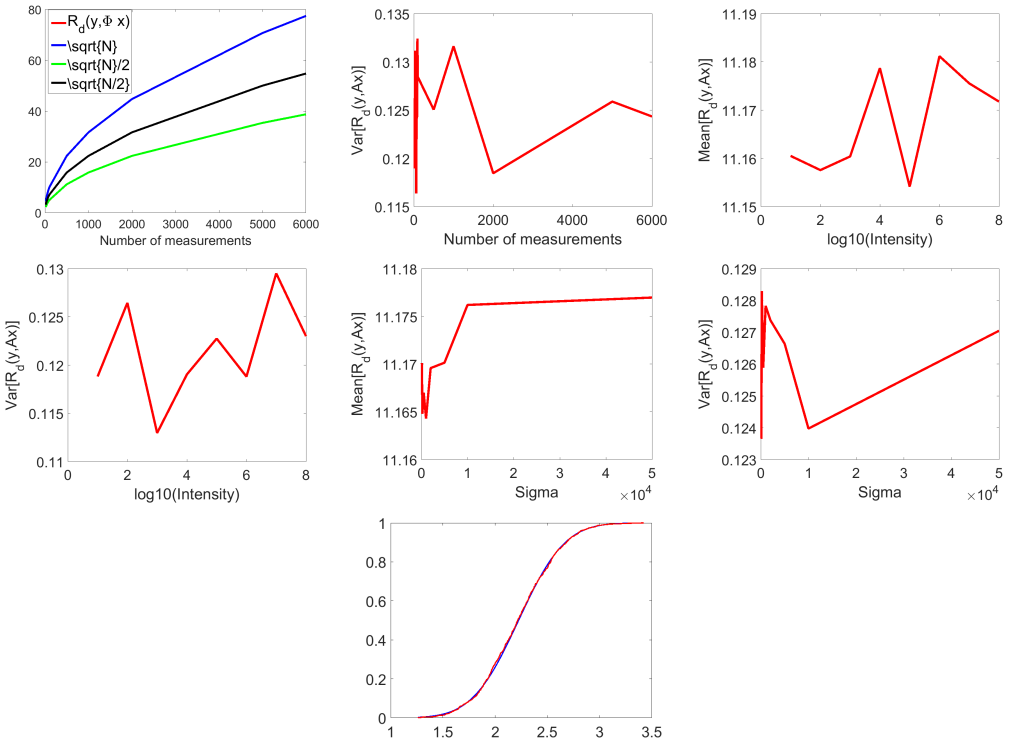

Figure 2: In the left to right, top to bottom order. First two sub-figures: Plot of mean and variance of the values of $R_d(\boldsymbol{y}, \boldsymbol{\Phi x})$ versus N for a fixed $I = 10^3, \sigma = 200$ for a signal of dimension $m = 1000$. (For the left subfigure in the first row, the blue line represents the plot of $\sqrt{N}$. The green line represents $\sqrt{N}/2$ which *coincides with* the red plot for $R_d(\boldsymbol{y}, \boldsymbol{\Phi x})$.) Third and fourth sub-figures: Plot of mean and variance of the values of $R_d(\boldsymbol{y}, \boldsymbol{\Phi x})$ versus $\log_{10}(I)$ for a fixed $N = 500, \sigma = 200$ for a signal of dimension $m = 1000$. Fifth and sixth sub-figures: Plot of mean and variance of the values of $R_d(\boldsymbol{y}, \boldsymbol{\Phi x})$ versus σ for a fixed $I = 10^3, N = 50$ for a signal of dimension $m = 1000$. Last sub-figure: Empirical CDF of $R_d(\boldsymbol{y}, \boldsymbol{\Phi x})$ (red curve) for $N = 20, I = 10^3, m = 1000$ compared to a Gaussian CDF (blue curve) with mean and variance equal to that of the values of $R_d(\boldsymbol{y}, \boldsymbol{\Phi x})$. The curves overlap significantly as the empirical CDFs are very close. Scripts for reproducing these results are available at [27].

(i) $E[R_d(\boldsymbol{y}, \boldsymbol{\Phi x})] \leq \sqrt{N/2}$

(ii) Define $v \triangleq \mathrm{Var}[R_d(\boldsymbol{y}, \boldsymbol{\Phi x})]$ and $w \triangleq \frac{8(1+d)^2}{2(d+1)(\sigma^2+1)-1}$. Then if $\forall i, \gamma_i > \beta_d \triangleq 0.5 - d$ and $N \geq \dfrac{3/8}{\left(\frac{\beta_d+\sigma^2}{4(\beta_d+d)} - \frac{\beta_d}{8(\beta_d+d)^2}\right)^2}$, we have

(a) $v \leq \dfrac{\sum_{i=1}^{N} \frac{\gamma_i+3(\gamma_i+\sigma^2)^2}{4(\gamma_i+d)^2}}{\sum_{i=1}^{N} \frac{\gamma_i+\sigma^2}{4(\gamma_i+d)} - \frac{\gamma_i}{8(\gamma_i+d)^2}} \leq \bar{v_d} \triangleq \dfrac{3/4}{\frac{\beta_d+\sigma^2}{4(\beta_d+d)} - \frac{\beta_d}{8(\beta_d+d)^2}}.$

(b) $P\Big(R_d(\boldsymbol{y}, \boldsymbol{\Phi x}) \leq \sqrt{N}(\frac{1}{\sqrt{2}} + \sqrt{\bar{v_d}})\Big) \geq 1 - 1/N.$

(iii) Specifically, if $\forall i, \gamma_i \geq 1$ and $N > 0.375w^2$, then statement (ii) becomes:
   (a) $v \leq 0.75w$.
   (b) $P\Big(R_d(\boldsymbol{y}, \boldsymbol{\Phi x}) \leq \sqrt{N}(\frac{1}{\sqrt{2}} + \sqrt{0.75w})\Big) \geq 1 - 1/N$. $\diamondsuit$

We make a few comments below:

(i) Yet again, $E[R_d(\boldsymbol{y}, \boldsymbol{\Phi x})]$ has an upper bound which is signal-independent. This property is *not* shared by the negative log-likelihood of the Poisson-Gaussian distribution. Also when $\boldsymbol{\Phi x} \succeq \beta_d \mathbf{1}$, we see that $\mathrm{Var}[R_d(\boldsymbol{y}, \boldsymbol{\Phi x})]$ is upper bounded by a constant that is effectively independent of σ (since the factors involving σ nearly cancel out from the numerator and denominator of w or $\bar{v_d}$), N as well as the signal values. This is further confirmed by Fig. 2.

(ii) Setting $\sigma = 0$ produces the statement of Theorem 1.

(iii) The bounds in this theorem can be easily modified for the case of uniform quantization noise from $\mathrm{Unif}[-\delta, +\delta]$, or Gaussian noise coupled with uniform quantization noise.

(iv) Similar to the case of Poisson noise in Theorem 1, the lower bound on γ_i is not strictly required in practice and is only a sufficient condition for the variance bound.

3.4. Key Theorem for Poisson-Gaussian CS

For the Poisson-Gaussian case, a theorem similar to Theorem 2 follows.

Theorem 4 : Consider a non-negative signal $\boldsymbol{x}$ with total intensity $I \triangleq \|\boldsymbol{x}\|_1$ expressed using the orthornormal basis $\boldsymbol{\Psi}$ in the form $\boldsymbol{x} = \boldsymbol{\Psi\theta}$. Consider Poisson-Gaussian corrupted CS measurements of the form $\boldsymbol{y} \sim \mathrm{Poisson}(\boldsymbol{\Phi x}) + \boldsymbol{\eta}$ where $\boldsymbol{\eta} \sim \mathcal{N}(0, \sigma^2)$ is signal-independent noise, and $\boldsymbol{\Phi}$ is constructed as per Eqn. 7. Let $\boldsymbol{\theta^\star}$ be the result of the following optimization problem:

$$\begin{aligned} &\mathsf{PG2} : \min \|\boldsymbol{\theta}\|_1 \text{ such that } \|\sqrt{\boldsymbol{y} + d} - \sqrt{\boldsymbol{A\theta} + d}\|_2 \leq \varepsilon, \\ &\|\boldsymbol{\Psi\theta}\|_1 = I, \boldsymbol{\Psi\theta} \succeq \mathbf{0}, \end{aligned} \tag{12}$$

where $d \triangleq c + \sigma^2$, $\boldsymbol{A} \triangleq \boldsymbol{\Phi\Psi}$ so that $\boldsymbol{\Phi x} = \boldsymbol{A\theta}$ and $\varepsilon \triangleq \sqrt{N}(\sqrt{\bar{v_d}} + \frac{1}{\sqrt{2}})$ is an upper bound on the magnitude of the noise in the measurements *after* application of the GAT, with $\bar{v_d}$ as defined in Theorem 3. Let $\boldsymbol{\theta_s}$ denote a vector containing the s largest magnitude elements of $\boldsymbol{\theta}$ with the rest being 0. If $\widetilde{\boldsymbol{\Phi}}$ obeys RIP of order $2s$ with RIC $\delta_{2s} < \sqrt{2} - 1$, if $\boldsymbol{\Phi x} \succeq \beta_d \mathbf{1}$ and if $N \geq \dfrac{3/8}{\left(\frac{\beta_d + \sigma^2}{4(\beta_d + d)} - \frac{\beta_d}{8(\beta_d + d)^2}\right)^2}$, where $\beta_d > 1/2 - d$, then we have for any $\kappa > 0$:

$$P\left(\frac{\|\boldsymbol{\theta} - \boldsymbol{\theta^\star}\|_2}{I} \leq C_1 \sqrt{N} \tau_d \sqrt{\frac{1}{I} + \frac{dN}{I^2}} + \frac{C_2 s^{-\frac{1}{2}} \|\boldsymbol{\theta} - \boldsymbol{\theta_s}\|_1}{I}\right) \geq 1 - \kappa^2/N \text{ where } \tau_d \triangleq (\sqrt{\bar{v_d}}/\kappa + \frac{1}{\sqrt{2}}).$$

We note several comments on this theorem here below.

Remarks on Theorem and its Proof:

(i) If $\beta_d > 1$ and $N \geq 0.375w^2$, then we have $v \leq 0.75w$ as per statement 3 of Theorem 3. Further note that the condition $\boldsymbol{\Phi x} \succeq \beta_d \mathbf{1}$ is sufficient, but not necessary (as per our simulations).

(ii) The proof of this theorem follows Theorem 2 very closely with a replacement of c by d. Hence we omit its proof.

(iii) Theorem 2 and Theorem 4 show that using the VST, a unified treatment of Poisson CS as well as Poisson-Gaussian CS is possible. Methods based on purely the PNLL do not have this feature. Theorem 4 can be easily extended to include uniform quantization noise (with or without Gaussian noise).

(iv) For the same probability, the upper bounds increase with σ due to the d term in the square root. Also setting $\sigma = 0$ gives us Theorem 2.

(v) Similar to the case of Theorem 2, the probability with which the bound holds can be approximated to $1 - e^{-\mathcal{O}(N)}$ using the CLT for large N.

3.5. Properties of $R(\boldsymbol{y}, \boldsymbol{\Phi x})$ and $R_d(\boldsymbol{y}, \boldsymbol{\Phi x})$

First, we note $R^2(\boldsymbol{y}, \boldsymbol{\Phi x})$ is convex in $\boldsymbol{x}$, which can be seen by a simple algebraic expansion and due to the concavity of $\sqrt{\boldsymbol{x}}$. Also, it is convex in $\boldsymbol{\theta}$ due to the affine mapping property of convex functions (see Section 3.2.2. of [35]). Second, for finite $\boldsymbol{y}$ and $c \neq 0$, $R^2(\boldsymbol{y}, \boldsymbol{\Phi x})$ is Lipschitz continuous as it has a bounded first derivative. Both these properties are also true for $R_d^2(\boldsymbol{y}, \boldsymbol{\Phi x})$. These properties allow for efficient optimization and have been pointed out earlier in [14].

4. VSTs for Regularization Parameter Selection in Poisson Inverse Problems

In this section, we explore a special property of our data-fidelity term $R_d(\boldsymbol{y}, \boldsymbol{\Phi x})$ for Poisson-Gaussian (and thereby also of the term $R(\boldsymbol{y}, \boldsymbol{\Phi x})$ for Poisson noise). Consider penalized estimators (similar to P5 or PG5 for CS) which seek to minimize

$$J(\boldsymbol{\theta}) = z_f(\boldsymbol{y}, \boldsymbol{A\theta}) + \rho z_r(\boldsymbol{\theta}), \tag{13}$$

where z_f is a data-fidelity term and z_r is a regularizer term. Such estimators are very popular in the literature on Poisson-Gaussian (or Poisson) inverse problems such as deblurring. However in most of the literature, a principled way to choose the regularization parameter $\rho > 0$ has not been specified. We note while choice of ρ is well-principled in case of Gaussian problems, the case with Poisson-Gaussian noise is much more involved due to noise heteroscedasticity. In the following, let $\boldsymbol{\theta_\rho}$ be the minimizer of $J(\boldsymbol{\theta})$ for a chosen value of the regularization parameter ρ. We now refer to two lemmas obtained from Lemma 3.4 of [36], which enable efficient and principled choice of ρ. We have included a proof of both lemmas in the supplemental material.

Lemma 3: $J(\boldsymbol{\theta_\rho})$ is a strictly increasing function of ρ. $\diamondsuit$

Lemma 4: $z_f(\boldsymbol{y}, \boldsymbol{A\theta_\rho})$ is a non-decreasing function of ρ. $\diamondsuit$

Now, consider the case when $z_f(\boldsymbol{y}, \boldsymbol{A\theta}) = R_d^2(\boldsymbol{y}, \boldsymbol{A\theta})$ as considered in Theorem 3. From Theorem 3, we see that $E[R_d(\boldsymbol{y}, \boldsymbol{A\theta})]$ is upper-bounded by $\sqrt{N/2}$, and that it has a constant variance. In practical simulations, we have seen that $E[R_d(\boldsymbol{y}, \boldsymbol{A\theta})] \approx \sqrt{N}/2$, as seen in Fig. 2 (top-left subfigure where the plots coincide). This behaviour can be explained as follows: $E[R_d(\boldsymbol{y}, \boldsymbol{A\theta})] = \sqrt{\sum_{i=1}^{N} E[(\sqrt{y_i + d} - \sqrt{[\boldsymbol{A\theta}]_i + d})^2]} \approx \sqrt{\sum_{i=1}^{N} \mathrm{Var}[\sqrt{y_i + d}]} \approx \sqrt{N}/2$. The first approximation is because $E(\sqrt{y_i + d}) \approx \sqrt{[\boldsymbol{A\theta}]_i + d}$ as per [17] (eqn. 2.10) and [12] (Appendix A.1). Hence $E[(\sqrt{y_i + d} - \sqrt{[\boldsymbol{A\theta}]_i + d})^2]$ is only an *approximation* to $\mathrm{Var}[\sqrt{y_i + d}]$. The second approximation is because the variance of each such term is shown to be roughly 1/4 in [17] (eqn. 2.11) and [12] (Appendix A.1). We emphasize that $E[R_d(\boldsymbol{y}, \boldsymbol{A\theta})] \approx \sqrt{N}/2$ is only an approximation and this derivation is not subjected to the rigorous treatment as in Theorem 3 (or Theorem 1). Note that the upper bounds of $\sqrt{N/2}$ predicted by these theorems are still valid. We only use this approximation as a guide to choose ρ by means of a statistical selection principle. As per this, we recommend a choice of ρ such that

$$\sqrt{N}/2 - \delta_\rho \leq R_d(\boldsymbol{y}, \boldsymbol{A\theta_\rho}) \leq \sqrt{N}/2 + \delta_\rho. \tag{14}$$

Here $\delta_\rho > 0$ is a tolerance parameter which we set to $3\sqrt{\mathrm{Var}[R_d(\boldsymbol{y}, \boldsymbol{A\theta_\rho})]}$. A non-zero δ_ρ is *required* due to noise stochasticity. In fact, exact equality of $R_d(\boldsymbol{y}, \boldsymbol{A\theta_\rho})$ to a particular value cannot be guaranteed even with a large number of measurements. Note that given a value of ρ, the minimum $\boldsymbol{\theta_\rho}$ in Eqn. 13 is unique if z_f is strictly convex.

Computation Time: A linear search for ρ in an interval $[\rho_l, \rho_h]$ is time consuming, as the optimization has to be carried out for $\mathcal{O}(\rho_h - \rho_l)$ values of ρ. However, due to the monotonicity of z_f w.r.t. ρ, we can choose ρ more efficiently than just a linear search. We can instead use a recursive procedure similar to the bisection method in root-finding to determine a ρ which satisfies the condition in Eqn. 14. This can reduce the number of optimizations to $\mathcal{O}(\log((\rho_h - \rho_l)/\delta_\rho))$ - see Sec. 5.5.1 of [37]. The secant method can also be used for faster convergence, but it requires second order differentiability of z_f unlike the bisection method which requires only continuity.

Existence of Solution: In practice, we simply chose a ρ such that $\sqrt{N}/2 - \delta_\rho \leq R_d(\boldsymbol{y}, \boldsymbol{A\theta_\rho}) \leq \sqrt{N}/2$. As $R_d(\boldsymbol{y}, \boldsymbol{A\theta_\rho})$ is a non-increasing function, such a rule for choosing ρ produces the smallest value of ρ that satisfies the constraint. This is desirable and optimal as $J(\boldsymbol{\theta_\rho})$ is an increasing function of ρ from Lemma 3.

Relation to Prior Work on Parameter Selection: There exists literature for statistical choice of ρ in Gaussian problems, for example [36]. However in a large body of the literature on Poisson-Gaussian inverse problems [7, 14], the parameter ρ is chosen either by cross-validation or else omnisciently (i.e. choosing the value of ρ that yielded the result closest to the ground truth). Here, we have instead provided a statistical principle for the choice of ρ in Poisson-Gaussian inverse problems. The earlier work in [28, 38] provides a selection principle for Poisson problems via the term $APNLL$ as

defined in the comments after Theorem 1 in Sec. 3.1. However it applies only for purely Poisson and not for Poisson-Gaussian problems. Moreover the variance of that term is proportional to N unlike our case where it is a constant.

Experimental results for the choice of ρ are presented in Sec. 5.3.

5. Results

In this section, we show signal reconstruction results from CS measurements with Poisson and Poisson-Gaussian noise. Box-plots for the results of all these experiments are presented in the supplemental material accompanying this paper. Our scripts for reproducing the results in this section are available at [27].

5.1. Experiments on Poisson CS

Signal and Measurement Generation: We ran experiments for the reconstruction of $Q = 100$ non-negative signals in 1D with 100 elements each, from their Poisson corrupted CS measurements. The sensing matrix $\boldsymbol{\Phi}$ followed Eqn. 7. The signals were synthetically constructed using sparse linear combinations of DCT basis vectors. The non-zero indices of the coefficient vector $\boldsymbol{\theta}$ for the Q different signals were chosen randomly (i.e. allowing different supports for each signal), and the values of those entries were drawn randomly from Unif$[0, 1]$. The signals $\boldsymbol{x} = \boldsymbol{\Psi\theta}$ thus generated were forced to be non-negative by adjusting the DC component, followed by a scaling to ensure that they had a desired value of I (see description of experiments later in this section).
Methods Compared: For the Poisson noise case, we ran our simulations on the following problem which is a variant of P1 *without* the constraint $\|\boldsymbol{\Psi\theta}\|_1 = I$ as its exclusion had a negligible impact on the results (see later in this section):

$$\mathsf{P3}: \min \|\boldsymbol{\theta}\|_1 \text{ such that } \|\sqrt{\boldsymbol{y}+c} - \sqrt{\boldsymbol{A\theta}+c}\|_2 \leq \varepsilon, \boldsymbol{\Psi\theta} \succeq \mathbf{0}.$$

Here we set $c = 3/8$, and the bound ε was set to $2\sqrt{N}$ based on the tail bound from Theorem 1 (note that $2\sqrt{N} = \sqrt{N}/\sqrt{2} + \sqrt{N}(\sqrt{6.48}/2.5)$, and that this bound holds with probability $1 - (2.5)^2/N$, i.e. $\kappa = 2.5$). Note that the same value of ε was used in all experiments, and that this is a *very conservative* upper bound. Problem P3, being convex, was implemented using the well-known CVX package [39] with the SDPT3 solver. We compared the performance of P3 to the following problem based on the negative log-likelihood of the Poisson distribution (again without the constraint $\|\boldsymbol{\Psi\theta}\|_1 = I$ for the same reason as for P3):

$$\mathsf{P4}: \min \rho\|\boldsymbol{\theta}\|_1 + \sum_{i=1}^{N}((\boldsymbol{A\theta})_i - y_i \log(\boldsymbol{A\theta})_i), \boldsymbol{\Psi\theta} \succeq \mathbf{0}.$$

For P4, the regularization parameter ρ was chosen omnisciently from the set $\mathcal{S} \triangleq \{10^{-10}, 10^{-9}, ..., 0.1, 1, 10\}$, i.e. choosing the particular value of $\rho \in \mathcal{S}$ that yielded the least squared difference between the true $\boldsymbol{\theta}$ (assuming it were known) and its estimate.

P4 was implemented using the well-known SPIRAL-TAP algorithm [10] with a penalty for the ℓ_1 norm of DCT coefficients, for a maximum of 500 iterations (in many cases, the algorithm converged and exited in just 300-400 iterations). For the default choice of a maximum of 100 iterations set in the SPIRAL-TAP code, the performance was significantly worse. We used default choices for all other parameters except ρ. Recent work in [40] analyzed the following estimator based on the negative likelihood, instead of P4:

$$\mathsf{P6}: \min_{\|\boldsymbol{\theta}\|_1 \leq I_\theta, \boldsymbol{\Psi\theta} \succeq \mathbf{0}} \sum_{i=1}^{N} (\boldsymbol{A\theta})_i - y_i \log(\boldsymbol{A\theta})_i, \tag{15}$$

where I_θ is an upper bound on $\|\boldsymbol{\theta}\|_1$. This method, i.e. P6, requires prior knowledge of I_θ for the analysis as well as the implementation even for matrices that obey RIP. P6 was implemented using CVX with the value of $\|\boldsymbol{\theta}\|_1$ supplied to it omnisciently. Recently, LASSO-based techniques for Poisson compressed sensing have emerged, as in [41]. Hence we also compared with a LASSO estimator of the following form:

$$\mathsf{LASSO}: \min \rho \|\boldsymbol{\theta}\|_1 + \|\boldsymbol{A\theta} - \boldsymbol{y}\|^2, \boldsymbol{\Psi\theta} \succeq \mathbf{0}.$$

We implemented LASSO using CVX, and we set the ρ parameter either omnisciently (referred to henceforth as just LASSO), or we set it to $2\sqrt{8 \log m/I}$ as per Theorem 2 of [41] (referred to henceforth as LASSO-fixed). Additionally, we also compared the results to a version of P3 which we had used in [18], given by the following:

$$\mathsf{P5}: \min \rho \|\boldsymbol{\theta}\|_1 + \|\sqrt{\boldsymbol{y} + c} - \sqrt{\boldsymbol{A\theta} + c}\|_2^2, \boldsymbol{\Psi\theta} \succeq \mathbf{0},$$

where ρ was chosen omnisciently from $\mathcal{S}$. P5, being convex, was again implemented using CVX and SDPT3. An estimator similar to P5 has been used earlier in [14] and [42], however only for Poisson denoising and deblurring (without performance bounds), and not for Poisson CS.

Study of variation of signal/measurement parameters: We show comparisons between P3, P4 with SPIRAL-TAP, P6, LASSO, and P5 for three types of experiments for the following RRMSE (relative root mean-squared error) metric: RRMSE $= \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2 / \|\boldsymbol{x}\|_2$, where $\boldsymbol{x}$ and $\boldsymbol{x}^\star$ denote the true/original and reconstructed signal respectively. In the *first experiment*, we studied the effect of change in signal intensity I on the reconstruction results. For this, we generated Poisson corrupted measurements of the Q different signals in $\mathbb{R}^{100}$, each with a fixed number of measurements $N = 50$. The sparsity of each signal in the DCT basis was fixed to $s = 10$ (but with different supports), and the signal intensity was varied from $I = 10$ to $I = 10^8$ in powers of 10. For each value of I, the median RRMSE value over the Q signals was computed. This is shown in the top sub-figure in Fig. 3. The performance of all methods improves with increase in I as expected. In the *second experiment*, for the Q different signals, the number of Poisson corrupted CS measurements was fixed to $N = 50$, the signal intensity was fixed to $I = 10^8$, and the signal sparsity was varied from $s = 5$ to $s = 50$ in steps of 5. For each value of s, median RRMSE values were recorded over the Q signals,

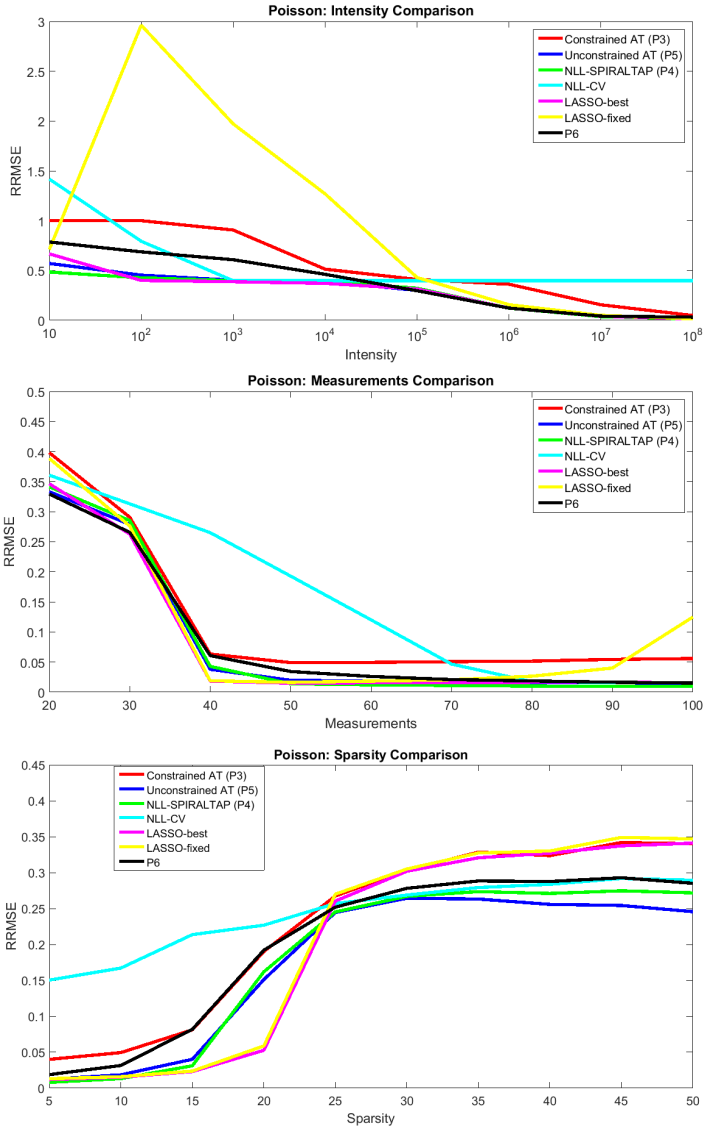

Figure 3: Median RRMSE comparisons between P3 using CVX with $\varepsilon = 2\sqrt{N}$ (termed 'Constrained Anscombe'), P4 using SPIRAL-TAP (termed 'NLL SPIRAL-TAP) with omniscient ρ, P4 with cross-validation for ρ (termed 'NLL-CV'), P5 using CVX (termed 'Unconstrained Anscombe'), LASSO using omniscient ρ, LASSO using $\rho = \mathcal{O}(1/I)$ (termed 'LASSO-fixed' - see main text), and P6 - a norm-constrained version of P4 - with omniscient choice of $\|\boldsymbol{\theta}\|_1$. Top row: fixed $N = 50$ and $s = 10$ but varying I, middle row: fixed $I = 10^8$ and $s = 10$ but varying N, bottom row: fixed $I = 10^8$ and $N = 50$ but varying s. See supplemental material for box-plots and [27] for code. Note that in many cases, the curve for P5 overlaps with that of other methods.

as shown in the middle sub-figure in Fig. 3. The performance of all methods worsens with increase in s as expected. In the *third experiment*, for the Q different signals, the sparsity of the signals was fixed to $s = 10$, and their intensity was fixed to $I = 10^8$. The number of measurements was varied from $N = 20$ to $N = 100$ in steps of 10. For each value of N, median RRMSE values were recorded over the Q signals, as shown in the bottom sub-figure in Fig. 3. We do see an improvement in the reconstruction results with increase in N, but this is not guaranteed in the worst case similar to [19].

Observations and Comments: Observing Fig. 3, we see that the reconstruction results with P5, LASSO and P4 are comparable in most cases. P5 and P4 showed better results than P3 due to the omnisicent selection of ρ, as against the fixed, statistically motivated ε in P3. The performance of P3 would improve and become equivalent to that of P5 with omniscient selection of ε. However, note that omniscient choices are difficult to implement in practice, and have significant computational costs. Improper choice of ρ led to arbitrary increase in reconstruction error. We have found that the optimal ρ depended on the unknown signal (see also [43] and Table 1). While model-selection approaches for Poisson problems exist [44], no performance bounds with such methods have been proven. For the sake of comparison, we collected results on P4 via cross-validation. For this, we omnisciently chose ρ which yielded the best RRMSE for $I = 10^4$ and used the same ρ for all other intensity levels in the first experiment. For the second experiment, ρ was chosen omnisciently for $s = 30$ and used for all other values of s. For the third experiment, ρ was chosen omnisciently for $N = 20$ and used for all other N. The results for this variant of P4 (termed 'P4 with cross-validation') are shown in Fig. 3. Similarly, we see that the results of LASSO-fixed (defined earlier in this section) suffer when compared to LASSO with omniscient ρ. The estimator P6 in Eqn. 15 from [40] requires prior knowledge of I for the analysis as well as the implementation even for matrices that obey RIP. In our case, as also in [29, 19, 41], the constraint $\|\boldsymbol{x}\|_1 = I$ is required in the theoretical analysis for the specific type of matrices from Eqn. 7. The constraint would not be required for RIP-obeying matrices, and was not deemed necessary even in the numerical experiments for matrices from Eqn. 7. For example, RRMSE of a typical signal of 100 dimensions with $s = 10, I = 10^8$ with $N = 50$ CS measurements using P3 was greater than that using P1 by only $\mathcal{O}(10^{-4})$. Also, we observed that P5 with omniscient choice of ρ outperformed P6 with omniscient choice of I.

Execution Times: We also saw that P4 for a single fixed ρ (that is, not counting execution times for different $\rho \in \mathcal{S}$) was 3-4 times more computationally expensive than P3 with a fixed ε. On a 2GHz CPU with 8 GB RAM, typical execution times were 58 seconds and 18.6 seconds for P4 and P3 respectively, for $N = 50, m = 100, s = 10$.

Image Reconstruction: Lastly, we ran an experiment to simulate image-patch and image reconstruction from Poisson-corrupted CS measurements, for a camera following the architecture of [45],[46]. The architecture of these cameras is similar to the Rice SPC [6], but the measurements are acquired patch-wise. That is, for each patch $\boldsymbol{x}_i \in \mathbb{R}_+^m$ extracted from an image, the measurement vector is given by $\boldsymbol{y}_i \sim \text{Poisson}(\boldsymbol{\Phi}_i \boldsymbol{x}_i)$ where

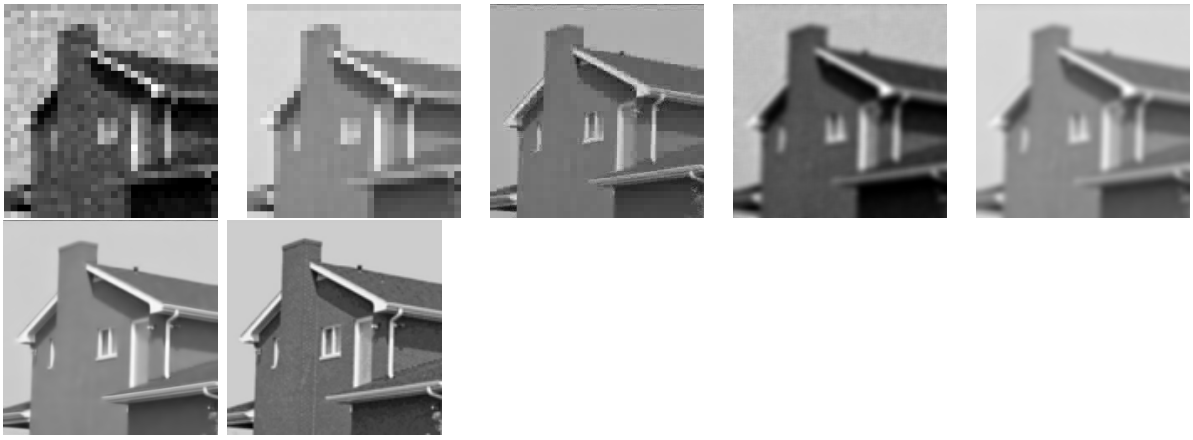

Figure 4: First row left to right: Image reconstruction results for non-overlapping 8×8 patches from 32 CS measurements per patch, using P3 for $I = 10^6$ (left, RRMSE = 0.743), $I = 10^8$ (RRMSE = 0.16), $I = 10^{10}$ (right, RRMSE = 0.068). Fourth, fifth and sixth: Same as ealrier two but with overlapping patches and averaging in sliding window fashion: for $I = 10^6$ (RRMSE = 0.7408), $I = 10^8$ (RRMSE = 0.148) and $I = 10^{10}$ (second row leftmost, RRMSE = 0.054). Second row (right): original image for reference.

$\boldsymbol{y}_i \in \mathbb{Z}_+^N, \boldsymbol{\Phi}_i \in \mathbb{R}_+^{N \times m}, N \ll m$ and i is a spatial location index. The model for each $\boldsymbol{\Phi}_i$ follows Eqn. 7. In our experiments, we set $m = 64$ (from 8×8 patches) and $N = 32$. Each (non-overlapping) patch $\boldsymbol{x}_i$ was independently reconstructed by solving P3 using $\boldsymbol{\Psi}$ as the 2D-DCT basis and $\varepsilon = 2\sqrt{N}$, as per the tail bound on $R(\boldsymbol{y}, \boldsymbol{\Phi x})$. Since there are inevitable patch-seam artifacts, we also ran these experiments for overlapping patches followed by sliding-window averaging. Though in [45],[46], CS measurements are not acquired on overlapping blocks, this simulates the use of a deblocking algorithm to get rid of patch-seam artifacts. The reconstruction results for this experiment are presented in Fig. 4 on the popular 'house' image (size 256×256) for values of total image-intensity $I \in \{10^6, 10^8, 10^{10}\}$. The results show clear improvement with increase in I and are evidence that our method works for compressible signals as well, since image patches are compressible (not sparse) in 2D-DCT bases.

5.2. Experiments on Poisson-Gaussian CS

The signal generation model for experiments on Poisson-Gaussian CS was the same as that used for Poisson CS. Throughout, we assumed known values of σ. Experiments were performed for the problem PG3 defined below, which is identical to PG2 except that we did not impose the $\|\boldsymbol{x}\|_1 = I$ constraint as its exclusion had negligible impact on the results:

$$\mathsf{PG3} : \min \|\boldsymbol{\theta}\|_1 \text{ s.t. } \|\sqrt{\boldsymbol{y} + d} - \sqrt{\boldsymbol{A\theta} + d}\|_2 \leq \varepsilon, \boldsymbol{\Psi\theta} \succeq \mathbf{0}. \tag{16}$$

Here as defined before $d \triangleq c + \sigma^2, c = 3/8$. For *all* experiments using PG3, the bound ε was set to $2\sqrt{N}$ based on Theorem 3 (see also Fig. 2). Note that this is

still a very conservative upper bound. We removed all measurements y_i for which $y_i + d < 0$. This happened very rarely (see supplemental material), and is akin to the so-called 'saturation rejection' for CS with saturation and quantization [47]. PG3 was implemented using CVX and the SDPT3 solver. We compared the results for PG3 with those produced by problem P4. P4 was implemented using SPIRAL-TAP for a maximum of 500 iterations (ensuring convergence in each case) under default parameters except ρ which was chosen omnisciently from $\mathcal{S}$. For P4, all negative measurements were removed. We also compared the results with problem PG5 defined below:

$$\mathsf{PG5}: \min \rho\|\boldsymbol{\theta}\|_1 + \|\sqrt{\boldsymbol{y}+d} - \sqrt{\boldsymbol{A}\boldsymbol{\theta}+d}\|_2^2, \boldsymbol{\Psi}\boldsymbol{\theta} \succeq \mathbf{0}.$$

PG5 was implemented using CVX-SDPT3, using an omniscient choice of $\rho \in \mathcal{S}$ and with removal of measurements for which $y_i + d < 0$. We did not compare with the Poisson-Gaussian technique in [7] because it is a *deconvolution* algorithm with a total variation prior, whereas we are dealing with CS and sparsity of transform coefficients. We also observed that empirical results with AT (i.e. P3) were similar to those with GAT (i.e. PG3) for small to moderate values of σ. For larger σ, GAT outperformed AT, besides being statistically more principled. Moreover for AT, measurements for which $y_i + c < 0$ need to be removed. This occurs more often than $y_i + d < 0$ since $d \triangleq c + \sigma^2$.

Study of variation of signal/measurement parameters: We ran three sets of experiments here. In the first experiment, we fixed $N = 50, s = 10, \sigma = 200$ and varied only I from 10^3 to 10^8 in multiples of 10. In the second experiment, we fixed $I = 10^8, \sigma = 200, s = 10$ and varied only N from 10 to 100 in steps of 10. In the third experiment, we fixed $I = 10^8, N = 50, s = 10$ and varied σ in $\{10, 50, 100, 250, 500, 1000, 2000, 10^4\}$. Comparative median RRMSE plots (across Q signals) are presented in Fig. 5.

Observations and Comments: The performance of our methods improved with increase in I and N, and worsened gradually with increase in σ (gradually because of the term dN/I^2 in the bounds for Theorem 4 which increases very slowly with σ for large values of I, such as $I = 10^8$ as chosen in Fig. 5). The presented results establish the usefulness of our proposed method for Poisson-Gaussian CS. We observed that P4 and PG5 with omnisicent ρ outperformed PG3 with fixed ε. The performance of PG3 would no doubt improve with omniscient choice of ε and would be quite similar to that of PG5. Quite surprisingly, P4 with omniscient ρ performed very well, even though it is not designed for Poisson-Gaussian noise. However we emphasize that no theoretical performance bounds for P4 have been established for this noise model. Moreover, with improperly chosen ρ, the performance of PG5 and P4 was worse than PG3, and even for a single fixed ρ, P4 was computationally more expensive than PG3. In Fig. 5, we also show results for P4 with cross-validation. In the first experiment, the value ρ was omnisciently chosen for $I = 10^4$ and used for other intensities. In the second experiment, the value ρ was chosen omnisciently for $N = 30$ and used for other values of N. For the third experiment, we chose the best ρ omnisciently for $\sigma = 20$ and used it for other

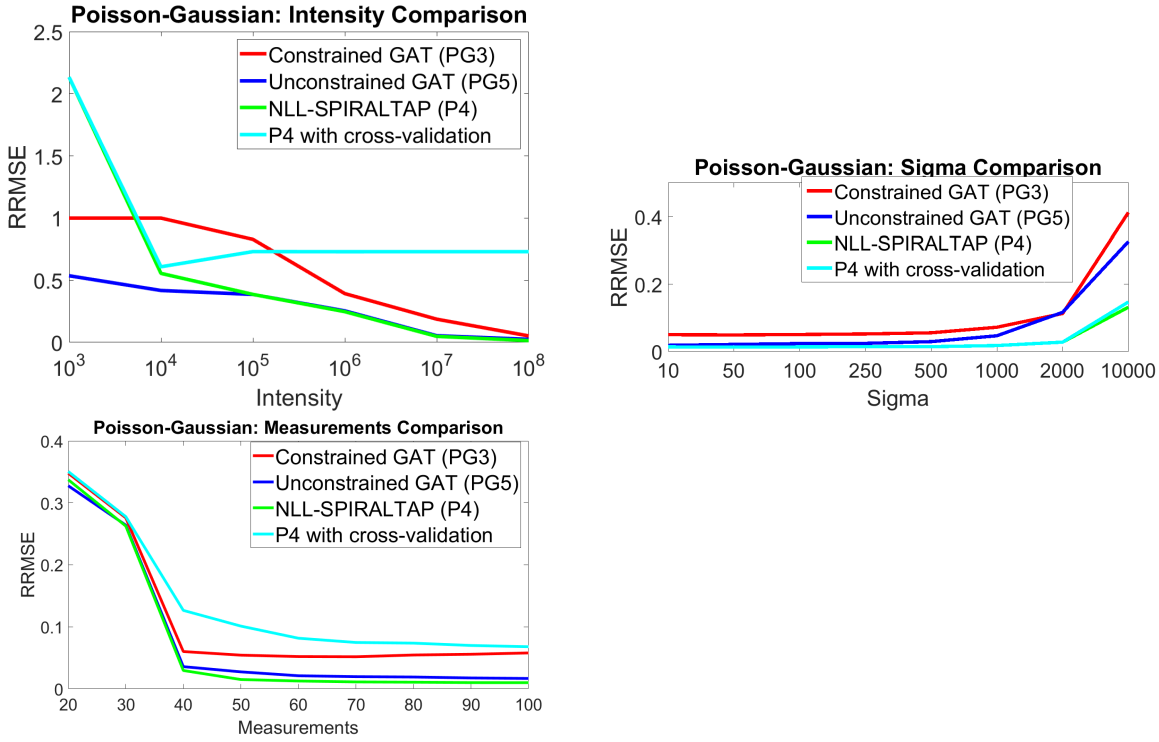

Figure 5: Median RRMSE comparisons between PG3 using CVX with $\varepsilon = 2\sqrt{N}$ (termed 'Constrained GAT'), P4 using SPIRAL-TAP (termed 'NLL SPIRAL-TAP'), P4 with cross-validation for ρ, and PG5 using CVX (termed 'Unconstrained GAT'). Top row (left): fixed $N = 50$, $\sigma = 200$ and $s = 10$ but varying I, top row (right): fixed $I = 10^8$, $s = 10$ and $N = 50$ but varying σ, bottom row: fixed $I = 10^8$, $\sigma = 200$ and $s = 10$ but varying N. See supplemental material for box-plots and [27] for code.

values of σ. Surprisingly, the best ρ did not depend on σ for a wide range.

Image Reconstruction: Lastly, we ran an image-patch and image reconstruction experiment similar to the one described for Poisson noise. We simulated $N = 32$ measurements of the form $\boldsymbol{y}_i \sim \text{Poisson}(\boldsymbol{\Phi}_i \boldsymbol{x}_i) + \boldsymbol{\eta}_i$, for patch $\boldsymbol{x}_i$ of $m = 64$ pixels. The σ for $\boldsymbol{\eta}_i$ was 200. The reconstruction was done independently patch-wise by solving PG3 using $\boldsymbol{\Psi}$ as the 2D-DCT basis and $\varepsilon = 2\sqrt{N}$. Results are presented on the 256×256 house image, for image-intensity $I \in \{10^6, 10^8, 10^{10}\}$ in Fig. 6. Due to the high σ relative to the measurement values, the reconstruction failed at $I = 10^6$ and is not reported here, but improved for higher intensities. Compared to Fig. 4, the results in Fig. 6 show higher RRMSE on non-overlapping blocks due to the presence of Gaussian noise. (The errors in both cases reduce upon sliding window averaging.) These experiments are evidence that our method works for compressible signals.

Figure 6: First row from left to right: Image reconstruction results for non-overlapping 8×8 patches from 32 Poisson-Gaussian CS measurements per patch with $\sigma = 200$, using PG3 for $I = 10^8$ (left, RRMSE = 0.5) and $I = 10^{10}$ (right, RRMSE = 0.07). Third and fourth images: Same as earlier two, but with overlapping patches and averaging in sliding window fashion: for $I = 10^8$ (left, RRMSE = 0.116), $I = 10^{10}$ (right, RRMSE = 0.0326). Rightmost: original image for reference.

5.3. Experiments with Regularization Parameter Selection in Deblurring

We now report results on choice of ρ for image deblurring under Poisson-Gaussian noise based on minimization of the following objective function:

$$J(\boldsymbol{x}) = R_d^2(\boldsymbol{y}, h * \boldsymbol{x}) + \rho \mathrm{TV}(\boldsymbol{x}) = \| \sqrt{\boldsymbol{y} + 3/8 + \sigma^2} - \sqrt{h * \boldsymbol{x} + 3/8 + \sigma^2} \|^2 + \rho \mathrm{TV}(\boldsymbol{x}), \tag{17}$$

where h is a known blur kernel, '*' is the convolution operator, $\boldsymbol{x}$ is the underlying image of size $n_1 \times n_2$, $\mathrm{TV}(\boldsymbol{x}) \triangleq \sum_{i=0}^{n_1-1} \sum_{j=0}^{n_2-1} \sqrt{(x(i+1,j) - x(i,j))^2 + (x(i,j+1) - x(i,j))^2}$ is the total variation of $\boldsymbol{x}$, and the forward model for the noisy image $\boldsymbol{y}$ is $\boldsymbol{y} = \mathrm{Poisson}[h * \boldsymbol{x}] + \boldsymbol{\eta}$ where each element of $\boldsymbol{\eta}$ is drawn iid from $\mathcal{N}(0, \sigma^2)$ with known σ. We chose the TV regularizer due to its popularity in the deblurring literature as it tends to enhance image edges, though we could have chosen any other regularizer. Anscombe-based deblurring methods have been earlier proposed in [14], but there ρ was chosen based on cross-validation which has its limitations and is also expensive.

We performed experiments on two different images, for a 25×25 Gaussian blur kernel with standard deviation 1.5 and Gaussian noise of variance 9. The average intensity of the two images was 69 and 4.5 respectively. The optimization was implemented by a primal-dual method based on the code released by the authors of [7]. We ran the algorithm for different values of ρ and recorded the value which satisfied the criterion in Eqn. 14. In Fig. 7, we plot the value of $|R_d(\boldsymbol{y}, h*\boldsymbol{x}) - \sqrt{n_1 n_2}/2|$ (referred to henceforth as the fidelity offset FO) against $\log_{10} \rho$. We also plot the mean absolute error (MAE) between $\boldsymbol{x}$ and the minimizer of $J(\boldsymbol{x})$ for a given value of ρ (denoted $\boldsymbol{x}_{\boldsymbol{\rho}}$) on the same graph. From the plots, we see that there is a good agreement between the value of ρ as predicted by our selection principle and the one that yields least MAE. For the first image, the least MAE was 0.74 for $\rho = 0.1$ and FO of 0.63. For an FO of 0.02, the MAE was 0.84 (slightly higher) with $\rho = 0.05$. For the second image, the least MAE was 0.46 for $\rho = 0.15$ and FO of 1.32. For an FO of 0.07, the MAE was 0.59 (slightly higher) with $\rho = 0.06$. From this, we empirically demonstrate the efficacy of using this principle for parameter selection in deblurring problems.

Comments:

(i) Though our principle to choose ρ is based on the Anscombe transform, it can be used for selection of ρ even if the z_f term is different from $R_f(\boldsymbol{y}, \boldsymbol{A\theta})$, i.e. for example a term based on PNLL. However the monotonicity of $z_f(\boldsymbol{y}, h * \boldsymbol{x_\rho})$ suggests that the same z_f (i.e., in our case R_d) be used for parameter selection for the sake of efficient selection via the bisection method.

(ii) Our principle (which is based on Theorems 1 and 3) could have been used to guide parameter selection for various estimators such as P4, P5, LASSO, PG5 for CS reconstruction in Secs. 5.1 and 5.2. However we did not do so, as the performance was comparable to that obtained with omniscient choice of ρ. Moreover, the aim there was to show the dependence of these estimators (based on unconstrained formulations) on an *external mechanism* of parameter selection, which is expensive in practice and for which rigorous theoretical bounds do not exist. In contrast, our estimators P3 and PG3 have no such dependencies, and we have established analytical bounds for them.

6. Conclusion, Comparisons to Prior Art and Future Work

Contributions: We have presented a convex implementable estimator for sparse/compressible signal reconstruction from CS measurements acquired by realistic sensing models, but corrupted by Poisson or Poisson-Gaussian noise. The estimator allows for statistically motivated and principled parameter tuning. To the best of our knowledge, there is no earlier work on analyzing Poisson CS using VSTs since the VSTs convert a problem with linear measurements to non-linear measurements [10]. We have demonstrated here, both theoretically as well as experimentally, that the non-linearity is actually not a problem, and that it does in fact have some advantages over the PNLL - namely more intuitive parameter tuning, besides Lipschitz continuity of the objective function and its derivative for $c \neq 0$. This is our first major contribution. Our second major contribution is the unification of analysis of Poisson CS and Poisson-Gaussian CS that our VST-based framework so readily allows for. Also ours is the first work to develop bounds for Poisson-Gaussian CS to the best of our knowledge. The extension of our method to Poisson-Gaussian noise also retains all the advantages of the method for Poisson noise. We emphasize that square-root transformations can be used stabilize variance *whenever the variance of the random variable is proportional to the mean* (chapter 14.6 of [48]), of which Poisson is just a special case. Thus, our approach is also applicable to other such noise models, including average of Poisson random variables, which appears in color image demosaicing [49]. Our third contribution is the development of a principle based on Theorems 1 and 3, to select the regularization parameter in penalized estimators in Poisson or Poisson-Gaussian inverse problems.

Note: In Table 1, we show succinct comparisons of our work in this paper to six recent techniques for Poisson compressed sensing. Here below, we present an elaborate

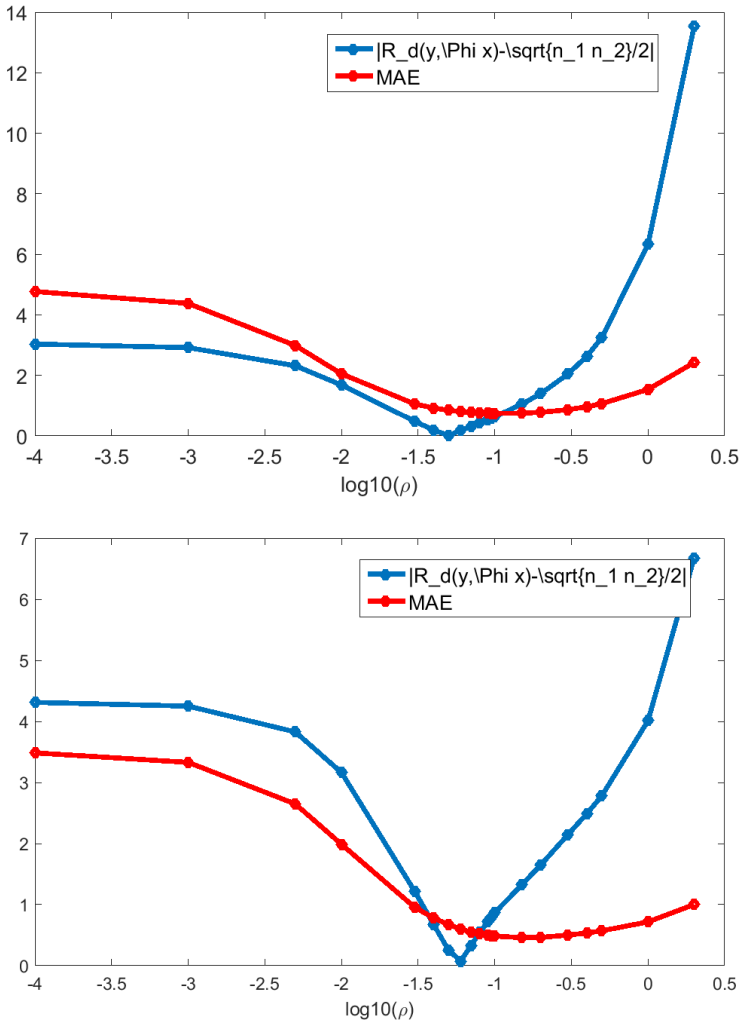

Figure 7: Plots of $|R_d(\boldsymbol{y}, h * \boldsymbol{x}) - \sqrt{n_1 n_2}/2|$ and MAE between $\boldsymbol{x}$ and $\boldsymbol{x}_{\boldsymbol{\rho}}$ versus ρ, for two images. See images displayed in Fig. 8. We see a good agreement between the value of ρ that produces least MAE and that predicted by our selection principle.

discussion.

Comparisons with Negative Log Likelihood Approaches: There exists some previous work on Poisson CS in [19, 29] using the penalized PNLL that applies to physically realizable sensing matrices, but the theory there is developed only for computationally intractable estimators with ℓ_0 regularizers. Moreover the latter work applies only to sparse (and not compressible) signals. The work in [40] applies to computationally tractable estimators using the PNLL, but does not explicitly address the important case of flux-preserving matrices and uses the estimator P6 defined in Section 5. However P6 cannot be implemented without knowledge of a signal-dependent parameter I. The consistency of an ℓ_1 regularized maximum likelihood (ML) estimator for compressive inversion is examined in [43] under the model $\lambda = \exp(-\boldsymbol{a}^t\boldsymbol{\theta})$ where $\boldsymbol{a}$ is a known vector, $\boldsymbol{\theta}$ is an unknown vector of sparse coefficients and λ is the mean of the Poisson

Table 1: Comparison of various methods analyzing performance bounds in Poisson and Poisson-Gaussian CS (Y = Yes, N = No) based on various criteria: TE =tractability of estimator; FP = whether the method handles flux-preserving sensing matrices explicitly; Comp = whether the method's bounds are applicable to only purely sparse or also to compressible signals; PE = whether the estimator has free or signal-dependent parameters; ML = whether the estimator uses a log-likelihood based data fidelity term; LCF,LCD = whether the objective function and its derivative are Lispschitz continuous; LB = whether the method presents lower bounds; PG = whether the method's analysis extends to Poisson-Gaussian noise apart from pure Poisson noise; NLCS = whether the estimator solves a non-linear inversion problem for CS; AvP = whether the method extends to handle noise modelled as average of Poisson random variables

Feature	Our Method	[19]	[29]	[40]	[50]	[41]	[51]
TE	Y	N	N	Y	Y	Y	Y
FP	Y	Y	Y	N	N	Y	Y
Comp	Y	Y	Sparse only	Sparse only	Y	Y	Y
PE	None (or statistically motivated ε)	Y (regularization parameter)	Y (signal ℓ_0 norm)	Y (signal ℓ_1 norm)	None (statistically motivated regularization parameter)	Y (signal intensity)	None (or statistically motivated ε)
LCF,LCD	Y,Y (for $c \neq 0$)	N,N	N,N	N,N	Y,Y	Y,Y	N,N
ML	N	Y	Y	Y	N	N	N
LB	N	N	Y	Y	N	Y	N
PG	Y	N	N	N	N	N	N
NLCS	Y (due to VST)	N	N	N	N	N	N
AvP	Y	N	N	N	N	N	N

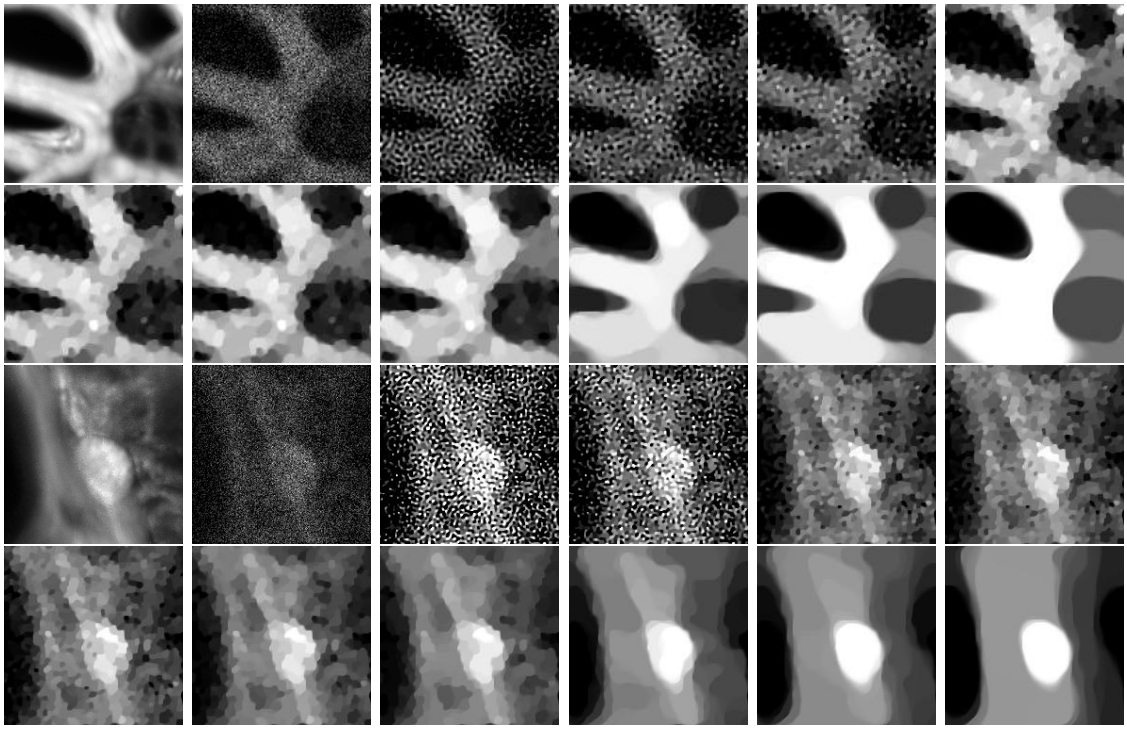

Figure 8: In each group from left to right, top to bottom: ground truth; blurry and noisy image; results for $\rho \in \{0.001, 0.005, 0.01, 0.05, 0.06, 0.07, 0.1, 0.5, 1, 2\}$. For the first image, the least MAE was 0.74 for $\rho = 0.1$ and fidelity offset (FO) of 0.63. For an FO of 0.02, the MAE was 0.84 (slightly higher) with $\rho = 0.05$. For the second image, the least MAE was 0.46 for $\rho = 0.15$ and FO of 1.32. For an FO of 0.07, the MAE was 0.59 (slightly higher) with $\rho = 0.06$. Refer Fig. 7.

distribution. This work in fact shows that the regularization parameter to guarantee consistency is dependent on the signal sparsity, which is unknown in practice. Again, none of these approaches have been extended to handle Poisson-Gaussian noise, which has a complicated likelihood function with infinite summation.

Comparison with LASSO-based Approaches: Recent work from [41] uses the LASSO, which is a computationally tractable estimator, and applies to physical constraints and for sparse/compressible signals. However their estimator requires the choice of a regularization parameter, which is dependent on the signal intensity I for statistical consistency (see Theorem 2 of [41]), or else a cross-validation type approach is needed. This is unlike our technique which has an easier choice of parameter during implementation. (In particular, the constraint $\|\boldsymbol{x}\|_1 = I$ was required only for the theoretical analysis and was not deemed necessary in the actual results.) There exist other papers which provide performance guarantees for some variant of the LASSO for Poisson-related problems. For example, [52] and [53] provide bounds using the RIP and maximum eigenvalue condition respectively. Necessary and sufficient conditions are derived for the sign consistency of the LASSO with the Poisson noise model in [54]. These techniques however do not explicitly deal with flux-preserving matrices.

Weighted/adaptive LASSO and group LASSO schemes with provable guarantees based on Poisson concentration inequalities have been proposed in [55, 50], and the technique in [50] can be extended for flux-preserving matrices. However again, none of these techniques handle the case of Poisson-Gaussian noise.

Comparison of Error Bounds: The work in [41] presents minimax estimators for compressible signals, whereas that in [29] presents minimax estimators for purely sparse signals. The bounds in [29, 40] are in fact applicable exclusively for purely sparse signals and extensions to approximately sparse signals are not presented. We now present a comparison between the error bounds derived in our paper with those in [41, 50] for the case of only Poisson noise (because they do not present analysis for Poisson-Gaussian noise unlike our work). As per Theorem 2 of [41], their relative error has a bound of the form $\mathcal{O}(\sqrt{R_1}(\log m/I)^{0.25})$ for the ℓ_1 sparsity ball, where $R_q \triangleq \sum_{i=2}^{m} |\theta_i/I|^q, 0 < q \leq 1$ and hence $R_1 \triangleq \sum_{i=2}^{m} |\theta_i/I|$‖. Their bound is valid with a probability of $1 - \mathcal{O}(1/p)$ whereas our bounds are valid with a probability of $1 - \mathcal{O}(1/N)$ which can increased to $1 - \exp(-N)$ via a central limit theorem argument as explained in the comments after Theorem 1. Compared to their bounds, we have a $\mathcal{O}(\sqrt{N})$ term in the numerator of the first term of our upper bounds for $c = 0$ (see Theorem 2 of our paper). This is because P3 is based on the formulation from [1] which puts a constraint on the noise residual. For such an estimator, the worst case error even in Gaussian compressed sensing scales as $\mathcal{O}(\sqrt{N})$. This is unlike the LASSO-based formulation for which the error scales as $\mathcal{O}(\sqrt{s \log m})$ for s-sparse signals. Of course, if we set $N = \mathcal{O}(s \log m)$ (which are the minimum number of measurements for compressed sensing bounds to hold), we get similar bounds as [41], but if N grows further, our bounds become looser. This issue has been discussed in [56] (problem 9.11), but we have observed that the apparent worst case behaviour of the constrained formulation rarely (if ever) shows up in practice. Our extensive simulations have shown no difference between the constrained estimator and the LASSO for Gaussian noise (or for Poisson noise). Moreover, implementation of the estimator in [41] requires the user to know a signal-dependent parameter I for computation of the regularization parameter in the LASSO formulation, which is not required in our case.

The work in [50] presents a purely data driven estimator for Poisson noise. However as mentioned in [41], their bounds are best suited for purely sparse signals. For approximately sparse signals, their bounds scale in the form $\mathcal{O}(\sqrt{R_q(\log m/I)^{1-q/2}}) + \mathcal{O}(R_q(\log m/I)^{(1-q)/2})$. For the ℓ_1 ball, we see that their error has the form $\mathcal{O}(\sqrt{R_1(\log m/I)^{0.5}}) + \mathcal{O}(R_1)$. Keeping other parameters fixed but varying only intensity, our bounds show a faster decay, and the bias term in our bounds $s^{-0.5}\|\boldsymbol{\theta} - \boldsymbol{\theta_s}\|_1/I$ is smaller than R_1.

In general, we note that the proof techniques in both these papers use concentration

‖ Theorem 2 of [41] presents a squared error, and so we have taken the square root of their bound to match our bound on $\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2$. We also set $q = 1$ in their bounds for the ℓ_1 ball. We also use different notation for signal dimension (m in our work, and p in theirs) and signal intensity (I in our work and T in theirs).

properties of linear combinations of Poisson random variables. It is not trivial to extend these for the Poisson-Gaussian case. Moreover, the negative log likelihood for the Poisson-Gaussian case requires computation of an infinite sum, which poses practical difficulties in implementation as well as theoretical analysis. Our estimator PG3 does not have these problems.

Comparisons with Other Approaches: Besides our conference paper [18], our group has performed some other earlier work on Poisson CS for realistic matrices using a tractable estimator based on the Jensen-Shannon divergence (JSD) between $\boldsymbol{y}$ and $\boldsymbol{\Phi x}$ [51]. The work essentially makes use of the fact that the square-root of the JSD (SQJSD) is a metric, and that the SQJSD has values that scale as $o(\sqrt{N})$ but independent of I. The LASSO has been extended to deal with non-linear problems in [57, 58], of which our technique in this paper is a special case (albeit with an additional non-negativity constraint). The technique in [58] derives error bounds on any stationary point of the objective function $\|\boldsymbol{y}-f(\boldsymbol{\Phi x})\|^2+\rho\|\boldsymbol{x}\|_1$ for any differentiable monotonic function f with bounded derivatives. At this point, we have not succeeded in adapting the technique from [58] to Poisson CS via the VST, because such an adaptation requires imposition of the additional necessary constraint $\boldsymbol{x} \succeq \boldsymbol{0}$ and hence the associated Karush Kuhn Tucker (KKT) conditions, while obtaining the stationary point of the objective function.

Future Work: There are many directions for future work: (1) a derivation of lower bounds, (2) analysis of support recovery and prediction bounds $\|\boldsymbol{\Phi x} - \boldsymbol{\Phi x}^\star\|_2$, (3) analysis of the effect of clipping on Poisson-Gaussian CS measurements due to the limited dynamic range of sensors, (4) analysis using the original Poisson-Gaussian likelihood, (5) removal of the sufficient but not necessary conditions on γ_i for our theoretical results to hold, and (6) seeking an explanation for the good reconstruction results obtained even after ignoring the $\|\boldsymbol{x}\|_1 = I$ constraint.

7. Proofs

7.1. Proof of Theorem 1

To prove theorem 1, we first begin by considering the case of a scalar $y \sim \text{Poisson}(\gamma)$ and generalize later to the case of measurement vectors. Define $f(y) \triangleq (\sqrt{y+c} - \sqrt{\gamma+c})^2$. Hence $f^{(1)}(y) = 1 - \sqrt{\dfrac{\gamma+c}{y+c}}$, $f^{(2)}(y) = \dfrac{\sqrt{\gamma+c}(y+c)^{-1.5}}{2}$, and $f^{(3)}(y) = \dfrac{-3\sqrt{\gamma+c}(y+c)^{-2.5}}{4}$ where $f^{(k)}(y)$ denotes the k^{th} derivative of $f(y)$ at y. Now, observe that $f(\gamma) = 0, f^{(1)}(\gamma) = 0$. Now $f(y) = f(\gamma) + \int_\gamma^y f^{(1)}(t)dt = \int_\gamma^y f^{(1)}(t)dt \leq (y-\gamma)f^{(1)}(y)$ since $f^{(1)}(y)$ is an increasing function of y. Similarly, we have $f^{(1)}(y) = f^{(1)}(\gamma)+\int_\gamma^y f^{(2)}(t)dt = \int_\gamma^y f^{(2)}(t)dt \leq (y-\gamma)f^{(2)}(\gamma)$ since $f^{(2)}(y)$ is a decreasing function. Combining this, we have

$$f(y) \leq (y-\gamma)f^{(2)}(\gamma) = \frac{(y-\gamma)^2}{2(\gamma+c)}. \tag{18}$$

Recall that $f(y)$ is a random variable. Taking expectation on both sides, we obtain

$$E[f(y)] \leq \frac{E[(y-\gamma)^2]}{2(\gamma+c)} \leq 0.5 \text{ as } E[(y-\gamma)^2] = \gamma. \tag{19}$$

To obtain an upper bound on the variance of $f(y)$, we need a lower bound on $E[f(y)]$ since $\text{Var}(f(y)) = E[(f(y))^2] - (E[f(y)])^2$. For this, consider the following second order Taylor series expansion of $f(y)$ around γ with a third-order Lagrange remainder term:

$$f(y) = f(\gamma) + (y-\gamma)f^{(1)}(\gamma) + \frac{(y-\gamma)^2}{2!}f^{(2)}(\gamma) + \frac{(y-\gamma)^2}{3!}f^{(3)}(z(y)), \tag{20}$$

where $z(y) \in (\gamma, y)$ or $z(y) \in (y, \gamma)$. Using previous results for the derivatives, we have:

$$f(y) = \frac{(y-\gamma)^2}{4(\gamma+c)} - \frac{\sqrt{\gamma+c}(y-\gamma)^3}{8(z(y)+c)^{2.5}}. \tag{21}$$

Taking expectation on both sides, we have

$$E[f(y)] = \frac{\gamma}{4(\gamma+c)} - \frac{\sqrt{\gamma+c}}{8}\sum_{y=0}^{\infty}(y-\gamma)^3(z(y)+c)^{-2.5}e^{-\gamma}\gamma^y/y!. \tag{22}$$

Considering χ to be the largest integer less than or equal to γ, we can split the infinite summation in the equation above into two parts: one is a summation K_1 from $y=0$ to $y=\chi$, and the other is a summation K_2 from $y=\chi+1$ to $y=\infty$. In other words, we have

$$K_1 = -\frac{\sqrt{\gamma+c}}{8}\sum_{y=0}^{\chi}(y-\gamma)^3(z(y)+c)^{-2.5}e^{-\gamma}\gamma^y/y! \tag{23}$$

$$K_2 = -\frac{\sqrt{\gamma+c}}{8}\sum_{y=\chi+1}^{\infty}(y-\gamma)^3(z(y)+c)^{-2.5}e^{-\gamma}\gamma^y/y!.$$

For the lower bound on $E[f(y)]$, we seek a value of $z(y)$ which will minimize K_1 and a value of $z(y)$ which will maximize K_2. This is because K_1 is non-negative since $y \leq \gamma$ for terms in K_1, and K_2 is negative since $y > \gamma$ for terms in K_2. As $(z(y)+c)^{-2.5}$ is a decreasing function, we get $z(y) = \gamma$ in both cases. This yields

$$E[f(y)] \geq \frac{\gamma}{4(\gamma+c)} - \frac{\sqrt{\gamma+c}}{8}(\gamma+c)^{-2.5}E[(y-\gamma)^3] = \frac{\gamma}{4(\gamma+c)} - \frac{\gamma}{8(\gamma+c)^2}. \tag{24}$$

Here we have made use of the fact that $E[(y-\gamma)^3] = \gamma$ for a Poisson random variable y with mean γ. As $f(y)$ is non-negative, we can write instead

$$E[f(y)] \geq \max(0, \frac{\gamma}{4(\gamma+c)} - \frac{\gamma}{8(\gamma+c)^2}). \tag{25}$$

Note that if $\gamma > 1/2 - c$, we have $E[f(y)] > 0$. Also $E[f(y)]$ can be shown to be an increasing function of γ. Squaring both sides of Eqn. 18 and taking expectation, we have

$$E[(f(y))^2] \leq \frac{E[(y-\gamma)^4]}{4(\gamma+c)^2} = \frac{\gamma(1+3\gamma)}{4(\gamma+c)^2}, \tag{26}$$

since $E[(y-\gamma)^4] = \gamma(1+3\gamma)$ for a Poisson random variable y with mean γ. So we have

$$\mathrm{Var}[f(y)] = E[(f(y))^2] - (E[f(y)])^2 \tag{27}$$

$$\leq \frac{\gamma(1+3\gamma)}{4(\gamma+c)^2} - \max(0, \frac{\gamma}{4(\gamma+c)} - \frac{\gamma}{8(\gamma+c)^2})^2 \leq \frac{\gamma(1+3\gamma)}{4(\gamma+c)^2} \leq 3/4. \tag{28}$$

The last inequality follows using L'Hospital's rule and using the fact that $\frac{\gamma(1+3\gamma)}{4(\gamma+c)^2}$ is a strictly increasing function of γ. We have so far derived upper bounds on the mean and variance of $f(y)$. Now we move to the case of a vector, i.e. to the case where $\boldsymbol{y}$ is a vector of N measurements, where the i^{th} measurement is given as $y_i \sim \mathrm{Poisson}(\gamma_i)$ where $\gamma_i = (\boldsymbol{\Phi x})_i$. We also define $f_i(y_i) \triangleq (\sqrt{y_i + c} - \sqrt{\gamma_i + c})^2$, $f(\boldsymbol{y}) \triangleq \sum_{i=1}^N f_i(y_i)$, $g(\boldsymbol{y}) \triangleq \sqrt{f(\boldsymbol{y})}$. Hence we have $E[g(\boldsymbol{y})] = E[\sqrt{f(\boldsymbol{y})}] \leq \sqrt{E[f(\boldsymbol{y})]} \leq \sqrt{N/2}$ using Eqn. 19. This proves the first statement of Theorem 1.

To derive a bound for the variance of $g(\boldsymbol{y})$, we proceed as follows. Define $\tilde{f}(\boldsymbol{y}) = f(\boldsymbol{y})/E[f(\boldsymbol{y})]$. Using the non-negativity of $\tilde{f}(\boldsymbol{y})$, we have

$$\sqrt{\tilde{f}(\boldsymbol{y})} \geq 1 + (\tilde{f}(\boldsymbol{y}) - 1)/2 - (\tilde{f}(\boldsymbol{y}) - 1)^2/2. \tag{29}$$

To see why, consider that $l(h) \triangleq 3h - h^3 \leq 2$ for all $h \geq 0$ since $l(1) = 2$ and $l(h)$ is monotonically increasing in $[0, 1]$ and monotonically decreasing in $[1, \infty)$. Putting $h = \sqrt{\tilde{f}}$ yields $3\sqrt{\tilde{f}} - \tilde{f}^{1.5} \leq 2 \rightarrow 3\tilde{f} - \tilde{f}^2 \leq 2\sqrt{\tilde{f}}$ which after simple algebra yields Eqn. 29. Taking expectation on both sides of Eqn. 29, we have

$$E[\sqrt{\tilde{f}(\boldsymbol{y})}] \geq 1 - \mathrm{Var}(\tilde{f}(\boldsymbol{y}))/2. \tag{30}$$

Substituting the definition of $\tilde{f}(\boldsymbol{y})$, we have

$$E[g(\boldsymbol{y})] = E[\sqrt{f(\boldsymbol{y})}] \geq \sqrt{E[f(\boldsymbol{y})]}\left(1 - \frac{\mathrm{Var}[f(\boldsymbol{y})]}{2(E[f(\boldsymbol{y})])^2}\right). \tag{31}$$

We now wish to find an upper bound on $\mathrm{Var}(g(\boldsymbol{y})$. Since $\mathrm{Var}(g(\boldsymbol{y})) = E[f(\boldsymbol{y})] - (E[g(\boldsymbol{y})])^2$, we have

$$\mathrm{Var}(g(\boldsymbol{y})) \leq E[f(\boldsymbol{y})] - E[f(\boldsymbol{y})]\Big(1 - \frac{\mathrm{Var}[f(\boldsymbol{y})]}{2E^2[\boldsymbol{y}]}\Big)^2 \tag{32}$$

$$= \frac{\mathrm{Var}[f(\boldsymbol{y})]}{E[f(\boldsymbol{y})]} - \frac{(\mathrm{Var}[f(\boldsymbol{y})])^2}{4(E[f(\boldsymbol{y})])^3}$$

$$\leq \frac{\mathrm{Var}[f(\boldsymbol{y})]}{E[f(\boldsymbol{y})]} = \frac{\sum_{i=1}^N \mathrm{Var}[f_i(y_i)]}{\sum_{i=1}^N E[f_i(y_i)]}.$$

Note that the first inequality (and hence all the other inequalities in the chain) are true if and only if $E[g(\boldsymbol{y})]^2 \geq E[f(\boldsymbol{y})]\left(1 - \frac{\mathrm{Var}[f(\boldsymbol{y})]}{2(E[f(\boldsymbol{y})])^2}\right)^2$. As $E[f(\boldsymbol{y})]$ is non-negative, this in turn requires that $\left(1 - \frac{\mathrm{Var}[f(\boldsymbol{y})]}{2(E[f(\boldsymbol{y})])^2}\right) \geq 0$. Define $\beta \triangleq \min\{\gamma_i\}_{i=1}^N$. Substituting the lower

bound on $E[f_i(y_i)]$ from Eqn. 25 and the upper bound on $\mathrm{Var}[f_i(y_i)]$ from Eqn 28, we see that this condition is satisfied when $\beta > 1/2 - c$ and $N \geq \dfrac{3/8}{(\frac{\beta}{4(\beta+c)} - \frac{\beta}{8(\beta+c)^2})^2}$. In particular, if $\beta \geq 1$, then $N \geq 29$. Again using the upper bound on $\mathrm{Var}[f_i(y_i)]$ and the lower bound on $E[f_i(y_i)]$ and substituting in Eqn. 33, we have the following bound on the variance:

$$\mathrm{Var}(g(\boldsymbol{y})) \leq \frac{\sum_{i=1}^{N} \frac{3}{4}\frac{\gamma_i^2}{(\gamma_i+c)^2} + \frac{\gamma_i}{4(\gamma_i+c)^2}}{\sum_{i=1}^{N} \frac{\gamma_i}{4(\gamma_i+c)} - \frac{\gamma_i}{8(\gamma_i+c)^2}}. \tag{33}$$

This proves the statement 2(a) of Theorem 1. We see that each term in the summation in the numerator is upper bounded by $\frac{3}{4}$¶ as shown in Eqn. 28, leading to a numerator upper bound of $3N/4$. Moreover one can show that the term in the denominator is monotonically increasing as well as positive for $\forall i, \gamma_i > \beta \triangleq 1/2 - c$. This leads to the upper bound $\mathrm{Var}(g(\boldsymbol{y})) \leq \bar{v} \triangleq \dfrac{3/4}{(\frac{\beta}{4(\beta+c)} - \frac{\beta}{8(\beta+c)^2})}$ as mentioned in the latter half of statement 2(a). To prove statement 3(a), we see that this term is lower bounded by $\frac{2\beta(\beta+c)-\beta}{8(c+\beta)^2} \approx 0.1157$ for $\beta = 1$ and using $c = \frac{3}{8}$. This proves statement 3(a), and the approximate value of 6.48 can be obtained again by using $c = \frac{3}{8}$.

In order to obtain a tail bound on $R(\boldsymbol{y}, \boldsymbol{\Phi x})$ under the condition that $\boldsymbol{\Phi x} \succeq \beta \mathbf{1}$, we can use Chebyshev's inequality to prove that $P(R(\boldsymbol{y}, \boldsymbol{\Phi x}) \leq \sqrt{N/2} + \sqrt{\bar{v}}\sqrt{N}) \geq 1 - \frac{1}{N}$, since the variance of $R(\boldsymbol{y}, \boldsymbol{\Phi x})$ is upper bounded by $\bar{v}$ when $\boldsymbol{\Phi x} \succeq \beta \mathbf{1}$. This proves statement 2(b) of the theorem. In particular again note that when $\beta = 1$, we have $\bar{v} \approx 2.545$ which proves statement 3(b).

However, we show here that for large values of N, $R(\boldsymbol{y}, \boldsymbol{\Phi x})$ is approximately Gaussian distributed which leads to tighter bounds and with an even higher probability: $P(R(\boldsymbol{y}, \boldsymbol{\Phi x}) \leq \sqrt{N/2} + \sqrt{\bar{v}}\sqrt{N}) \geq 1 - 2e^{-N/2}$ using upper bounds on the mean and variance of $R(\boldsymbol{y}, \boldsymbol{\Phi x})$.

By the CLT, 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 have variances upper bounded by $\bar{v}$ if $\boldsymbol{\Phi x} \succeq \beta \mathbf{1}$. Due to the continuity of Φ_g[+], we have $P(\frac{f(\boldsymbol{y})-N\mu}{\sigma\sqrt{N}} \leq \alpha + \frac{\alpha^2\sigma^2}{4\mu\sigma\sqrt{N}}) \to \Phi_g(\alpha)$ as $N \to \infty$. Hence we have $P(f(\boldsymbol{y}) \leq (\sqrt{N\mu} + \frac{\alpha\sigma}{2\sqrt{\mu}})^2) \to \Phi_g(\alpha)$ as $N \to \infty$, and taking square roots we get $P(\sqrt{f(\boldsymbol{y})} \leq (\sqrt{N\mu} + \frac{\alpha\sigma}{2\sqrt{\mu}})) \to \Phi_g(\alpha)$ as $N \to \infty$. By rearrangement, we obtain $P(\frac{\sqrt{f(\boldsymbol{y})}-\sqrt{N\mu}}{\sigma/(2\sqrt{\mu})} \leq \alpha) \to \Phi_g(\alpha)$ as $N \to \infty$. With this development and since $\mu \leq 1/2, \sigma^2 \leq \bar{v}$ from Eqns. 19 and 28, we can now invoke a Gaussian tail bound to establish that $P(R(\boldsymbol{y}, \boldsymbol{\Phi x}) \leq \sqrt{N/2} + \sqrt{\bar{v}}\sqrt{N}) \geq 1 - 2e^{-N/2}$. Note that the Gaussian

¶ Strictly speaking, this upper bound is for $c = 3/8$. It turns out that if $c = 0$, then the constant changes to 1.125. However this does not change the fundamental nature of our proof and hence we do not dwell on this point further.

[+] inspired from `https://stats.stackexchange.com/questions/241504/central-limit-theorem-for-square-roots-of-\sums-of-i-i-d-random-variables`

nature of $R(\boldsymbol{y}, \boldsymbol{\Phi x})$ emerges from the CLT and is only an asymptotic result. However we consistently observe it to be approximately true even for small values of $N \sim 20$ as confirmed by a Kolmogorov-Smirnov test (see [27]). $\diamondsuit$

7.2. Proof of Theorem 2

We provide a sketch of the proof below, inspired from [1], but modified to suit our problem.

(i) Define a vector $\boldsymbol{h} \triangleq \boldsymbol{\theta} - \boldsymbol{\theta}^\star$. Denote vector $\boldsymbol{h}_T$ to be equal to $\boldsymbol{h}$ only for index set T and zero for other indices. Let T_0 be the set containing s largest absolute value indices of $\boldsymbol{h}$, T_1 be the set containing s largest absolute value indices of $h_{T_0^c}$ and so on, where T^c is the complement of the set T. Thus, vector $\boldsymbol{h}$ can be decomposed as the sum of $\boldsymbol{h}_{T0}$,$\boldsymbol{h}_{T_1}$,$\boldsymbol{h}_{T_2}$,...

(ii) Define $\boldsymbol{A} \triangleq \boldsymbol{\Phi\Psi}$. We have

$$\|\boldsymbol{Ah}\|_2^2 = \|\boldsymbol{A}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\|_2^2 \tag{34}$$

$$= \sum_{i=1}^{N} \left(\sqrt{(\boldsymbol{A\theta})_i + c} - \sqrt{(\boldsymbol{A\theta}^\star)_i + c} \right)^2 \left(\sqrt{(\boldsymbol{A\theta})_i + c} + \sqrt{(\boldsymbol{A\theta}^\star)_i + c} \right)^2 \tag{35}$$

(a) Consider an upper bound of ε on $\|\sqrt{\boldsymbol{y} + c} - \sqrt{\boldsymbol{\Phi x} + c}\|_2$. Later on, we shall assign a statistical meaning to ε based on Theorem 1. By triangle inequality and the nature of the constraint in P1, we have

$$\|\sqrt{\boldsymbol{A\theta} + c} - \sqrt{\boldsymbol{A\theta}^\star + c}\|_2 \leq \|\sqrt{\boldsymbol{y} + c} - \sqrt{\boldsymbol{A\theta} + c}\|_2 + \|\sqrt{\boldsymbol{y} + c} - \sqrt{\boldsymbol{A\theta}^\star + c}\|_2 \quad \leq 2\varepsilon. \tag{36}$$

(b) For scalars $v_1 \geq 0, v_2 \geq 0$, we have $(\sqrt{v_1} + \sqrt{v_2})^2 \leq 4\text{max}(v_1, v_2)$. We also have $(\boldsymbol{A\theta})_i = (\boldsymbol{\Phi x})_i = \Sigma_j \Phi_{ij} x_j \leq \dfrac{\|\boldsymbol{x}\|_1}{N} = \dfrac{I}{N}$. Likewise $(\boldsymbol{A\theta}^\star)_i \leq \dfrac{I}{N}$ as well, since $\|\boldsymbol{x}^\star\|_1 = I$. Hence $(\sqrt{(\boldsymbol{A\theta})_i + c} + \sqrt{(\boldsymbol{A\theta}^\star)_i + c})^2 \leq 4(\dfrac{I}{N} + c)$.

(c) Combining the earlier two results with Eqn. 36, we have $\|\boldsymbol{Ah}\|_2 \leq 4\varepsilon\sqrt{\dfrac{I}{N} + c}$.

(iii) To prove the bound on $\|\mathbf{h}_{(T_0 \cup T_1)^c}\|_2$, we follow steps similar to [1] to obtain

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

(iv) To prove error bounds on $\|\mathbf{h}_{(T_0 \cup T_1)}\|_2$, we adopt the following steps.

(a) Given the construction for $\boldsymbol{\Phi}$ in Eqn. 7, we have

$$\boldsymbol{\Phi\Psi}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star) = \frac{1}{2\sqrt{N}}\widetilde{\boldsymbol{\Phi}}\boldsymbol{\Psi}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star) + (\|\boldsymbol{\Psi\theta}\|_1 - \|\boldsymbol{\Psi\theta}^\star\|_1) = \frac{1}{2\sqrt{N}}\widetilde{\boldsymbol{\Phi}}\boldsymbol{\Psi}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star) \tag{38}$$

since we know that $\|\boldsymbol{\Psi\theta}\|_1 = \|\boldsymbol{\Psi\theta}^\star\|_1 = I$. Defining $\boldsymbol{B} \triangleq \widetilde{\boldsymbol{\Phi}}\boldsymbol{\Psi}$, we get

$$\|\boldsymbol{Bh}\|_2 = 2\sqrt{N}\|\boldsymbol{Ah}\|_2 \leq 8\varepsilon\sqrt{I + cN}. \tag{39}$$

(b) Following steps in [1] using the RIP and the Cauchy-Schwarz inequality, we can prove that

$$\|\boldsymbol{h}_{T_0\cup T_1}\|_2 \leq C'\varepsilon\sqrt{I+cN} + C''s^{-1/2}\|\boldsymbol{\theta}-\boldsymbol{\theta}^{\star}\|_1 \tag{40}$$

where $C' \triangleq \frac{2\sqrt{1+\delta_{2s}}}{1-\delta_{2s}(\sqrt{2}+1)}$ and $C'' \triangleq \frac{2\sqrt{2}\delta_{2s}}{1-\delta_{2s}(\sqrt{2}+1)}$.

(v) Combining the bounds on $\|\boldsymbol{h}_{T_0\cup T_1}\|_2$ and $\|\boldsymbol{h}_{T_0\cup T_1{}^c}\|_2$, we have

$$\|\boldsymbol{h}\|_2 \leq C_1\varepsilon\sqrt{I+cN} + C_2\sqrt{2}\|\boldsymbol{\theta}-\boldsymbol{\theta_s}\|_1 \tag{41}$$

where $C_1 \triangleq 2C'$ and $C_2 \triangleq 2+2C''$.

Finally, we divide by I to obtain upper RRE bounds:

$$\frac{\|\boldsymbol{\theta}-\boldsymbol{\theta}^{\star}\|_2}{I} \leq C_1\varepsilon\sqrt{\frac{1}{I}+\frac{cN}{I^2}} + \frac{C_2 s^{-\frac{1}{2}}\|\boldsymbol{\theta}-\boldsymbol{\theta_s}\|_1}{I}. \tag{42}$$

Using Theorem 1, we see that $\varepsilon \leq \sqrt{N}(\sqrt{\bar{v}}/\kappa + 1/\sqrt{2})$ with a probability of $1-\kappa^2/N$ for any $\kappa > 0$ under certain lower bounds on N and $(\boldsymbol{\Phi x})_i$. This proves Theorem 2. Note that this bound uses the fact that $\boldsymbol{y}$ is Poisson distributed. ♢

7.3. Proof of Theorem 3

The proof of this theorem is very similar to that of Theorem 1, so we mention only the points of difference. First, right through the proof, the constant c is replaced by $d \triangleq c+\sigma^2$. Moreover for Poisson-Gaussian noise where the Gaussian component is signal-independent, we have $E[(y-\gamma)^2] = \gamma+\sigma^2, E[(y-\gamma)^3] = \gamma, E[(y-\gamma)^4] = \gamma+3(\gamma+\sigma^2)^2$. Despite these changes, the upper bound for $E[f(y)]$ from Eqn. 19 remains unchanged (and so does the lower bound for $E[f(\boldsymbol{y})]$). The upper bound for the variance of $f(y)$ from Eqn. 28 becomes $\mathrm{Var}[f(y)] \leq \dfrac{\gamma+3(\gamma+\sigma^2)^2}{4(\gamma+d)^2} \leq 3/4$. Following similar steps, the final upper bound for the variance of $g(\boldsymbol{y})$ is given by:

$$\mathrm{Var}(g(\boldsymbol{y})) \leq \frac{\sum_{i=1}^{N}\frac{\gamma_i+3(\gamma_i+\sigma^2)^2}{4(\gamma_i+d)^2}}{\sum_{i=1}^{N}\frac{\gamma_i+\sigma^2}{4(\gamma_i+d)} - \frac{\gamma_i}{8(\gamma_i+d)^2}}. \tag{43}$$

This bound holds if $\forall i, \gamma_i > \beta_d \triangleq 0.5-d$ and $N \geq \dfrac{3/8}{(\frac{\beta_d+\sigma^2}{4(\beta_d+d)} - \frac{\beta_d}{8(\beta_d+d)^2})^2}$. This can be further shown to be upper bounded by $\bar{v_d} \triangleq \dfrac{3/4}{\frac{\beta_d+\sigma^2}{4(\beta_d+d)} - \frac{\beta_d}{8(\beta_d+d)^2}}$ as defined in the theorem statement. In particular, if $\beta_d = 1$, we get $N \geq 0.375w^2$ where $w = \frac{8(1+d)^2}{2(d+1)(\sigma^2+1)-1}$ and $v \leq 0.75w$. The statement 2(b) of the theorem can also be easily derived using similar arguments as in Theorem 1, and these bounds can be approximately refined via the CLT to yield $P(R_d(\boldsymbol{y},\boldsymbol{\Phi x}) \leq \sqrt{N}(\frac{1}{\sqrt{2}} + \sqrt{\bar{v_d}})) \geq 1-2e^{-N/2}$. ♢

References

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

[2] H. Joel Trussell and R. Zhang. The dominance of Poisson noise in color digital cameras. In *ICIP*, pages 329–332. IEEE, 2012.

[3] F. Murtagh, J.-L. Starck, and A. Bijaoui. Image restoration with noise suppression using a multiresolution support. *Astronomy and Astrophysics*, 112:179, July 1995.

[4] S. Delpretti, F. Luisier, S. Ramani, T. Blu, and M. Unser. Multiframe sure-let denoising of timelapse fluorescence microscopy images. In *ISBI*, page 149152, 2008.

[5] T. Du Bosq and B. Preece. Performance assessment of a singlepixel compressive sensing imaging system. In *Proc. SPIE 9820, Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XXVII, 98200F*.

[6] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly, and R. Baraniuk. Single pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, 2008.

[7] E. Chouzenoux, A. Jezierska, J.-C. Pesquet, and H. Talbot. A convex approach for image restoration with exact Poisson-Gaussian likelihood. *SIAM J. Imaging Sciences*, 8(4):2662–2682, 2015.

[8] Florian Luisier, Thierry Blu, and Michael Unser. Image denoising in mixed Poisson-Gaussian noise. *IEEE TIP*, 20(3):696–708, 2011.

[9] D. Shin, J.-H. Shapiro, and V. Goyal. Performance analysis of low-flux least-squares single-pixel imaging. *IEEE Signal Process. Lett.*, 23(12):1756–1760, 2016.

[10] Z. Harmany et al. This is SPIRAL-TAP: Sparse Poisson intensity reconstruction algorithms - theory and practice. *IEEE TIP*, 21(3):1084–1096, 2012.

[11] D. Lingenfelter, J. Fessler, and Z. He. Sparsity regularization for image reconstruction with Poisson data. In *Proc. SPIE*, volume 7246, 2009.

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

[13] Bo Zhang, Mohamed-Jalal Fadili, and Jean-Luc Starck. Wavelets, ridgelets, and curvelets for Poisson noise removal. *IEEE TIP*, 17(7):1093–1108, 2008.

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

[15] T. Jeong, H. Woo, and S. Yun. Frame-based poisson image restoration using a proximal linearized alternating direction method. *Inverse Problems*, 29, 2013.

[16] T. Hohage and F. Werner. Inverse problems with Poisson data: statistical regularization theory, applications and algorithms. *Inverse Problems*, 32, 2016.

[17] F. J. Anscombe. The transformation of Poisson, binomial and negative-binomial data. *Biometrika*, 35(3/4):246–254, 1948.

[18] D. Garg and A. Rajwade. Performance bounds for Poisson compressed sensing using variance-stabilization transforms. In *ICASSP*, pages 1–4, 2017.

[19] M. Raginsky, R. Willett, Z. Harmany, and R. Marcia. Compressed sensing performance bounds under Poisson noise. *IEEE TSP*, 58(8):3990–4002, Aug 2010.

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

[21] J. Bobin, J.-L. Starck, and R. Ottensamer. Compressed sensing in astronomy. *IEEE Journal of selected Topics in Signal Processing*, 2(5), 2008.

[22] A. Scaife, G. Puy, L. Jacques, P. Vandergheynst, and Y. Wiaux. Compressed sensing imaging techniques for radio interferometry. *Monthly Notices of the Royal Astronomical Society*, 395(3):1733–1742, 2009.

[23] V. Studer, J. Bobin, M. Chahid, H.-S. Mousavi, E. Candes, and M. Dahan. Compressive fluorescence microscopy for biological and hyperspectral imaging. *Proceedings of the National Academy of Sciences*, 109(26), 2012.

[24] G. Howland, D. Lum, M. Ware, and J. Howell. Photon counting compressive depth mapping. *Opt. Express*, 21(20):23822–23837, 2013.

[25] J. Ma. Single-pixel remote sensing. *IEEE Geoscience and Remote Sensing Letters*, 6(2), 2009.

[26] J. H. Curtiss. On transformations used in the analysis of variance. *Ann. Math. Statist.*, 14(2):107–122, 06 1943.

[27] Code for reproducing results in this paper. https://tinyurl.com/y8doso3n.

[28] M. Bertero, P. Boccacci, G. Talenti, R. Zanella, and L. Zanni. A discrepancy principle for Poisson data. *Inverse Problems*, 26, 2010.

[29] X. Jiang, G. Raskutti, and R. Willett. Minimax optimal rates for Poisson inverse problems with physical constraints. *IEEE TIT*, 61(8):4458–4474, 2015.

[30] M. S. Bartlett. The square root transformation in the analysis of variance. *Journal of the Royal Statistical Society*, 68, 1936.

[31] E. Candes, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

[32] E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 12 2007.

[33] M. Freeman and J.Tukey. Transformations related to the angular and the square root. *Annals of Mathematical Statistics*, 21:607–611, 1950.

[34] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

[35] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[36] S. Anzengruber and R. Ramlau. Morozovs discrepancy principle for tikhonov-type functionals with nonlinear operators. *Inverse Problems*, 26(2), 2009.

[37] M. Heath. *Scientific Computing: An Introductory Survey*. McGraw-Hill Higher Education, 2nd edition, 1996.

[38] L. Zanni, A. Benfenati, M. Bertero, and V. Ruggiero. Numerical methods for parameter estimation in poisson data inversion. *J Math Imaging Vis*, 52:397–413, 2015.

[39] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, March 2014.

[40] M.-H. Rohban, V. Saligrama, and D.-M. Vaziri. Minimax optimal sparse signal recovery with Poisson statistics. *IEEE TSP*, 64(13):3495–3508, 2016.

[41] Y. Li and G. Raskutti. Minimax optimal convex methods for Poisson inverse problems under lq-ball sparsity. online; accessed July 2016.

[42] K. Frick, P. Marnitz, and A. Munk. Statistical multiresolution estimation for variational imaging: With an application in poisson-biophotonics. *Journal of Mathematical Imaging and Vision*, 46:370–387, 2013.

[43] Y.-H. Li and V. Cevher. Consistency of l1-regularized maximum-likelihood for compressive Poisson regression. In *ICASSP*, pages 3606–3610, 2015.

[44] J. Bardsley and J. Goldes. Regularization parameter selection methods for ill-posed poisson maximum likelihood estimation. *Inverse Problems*, 25(9), 2009.

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

[46] K. Kulkarni, S. Lohit, P. Turaga, R. Kerviche, and A. Ashok. Reconnet: Non-iterative reconstruction of images from compressively sensed measurements. In *CVPR*, June 2016.

[47] J. Laska, P. Boufounos, M. Davenport, and R. Baraniuk. Democracy in action: Quantization, saturation, and compressive sensinng. *Applied and Computational Harmonic Analysis*, 31(10):429–443, 2011.

[48] J. H. Pollard. *A Handbook of Numerical and Statistical Techniques: With Examples Mainly from the Life Sciences*. Cambridge University Press, 1977.

[49] S. Patil and A. Rajwade. Poisson noise removal for image demosaicing. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

[50] X. Jiang, P. Reynaud-Bouret, V. Rivoirard, L. Sansonnet, and R. Willett. A data-dependent weighted LASSO under Poisson noise. online; accessed July 2016.

[51] S. Patil, K. Gurumoorthy, and A. Rajwade. Using an information theoretic metric for compressive recovery under poisson noise. *Signal Processing*, 2019.

[52] I. Rish and G. Grabarnik. Sparse signal recovery with exponential-family noise. In *ACCS*, pages 60–66, 2009.

[53] S. Kakade, O. Shamir, K. Sindharan, and A. Tewari. Learning exponential families in high-dimensions: Strong convexity and sparsity. In *AISTATS*, pages 381–388, 2010.

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

[55] S. Ivanoff, F. Picard, and V. Rivoirard. Adaptive LASSO and group-LASSO for functional Poisson regression. *JMLR*, 17(55):1–46, 2016.

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

[57] Y. Plan and R. Vershynin. The generalized LASSO with non-linear observations. *IEEE TIT*, 62(3):1528–1537, March 2016.

[58] Z. Yang, Z. Wang, H. Liu, Y. Eldar, and T. Zhang. Sparse nonlinear regression: Parameter estimation and asymptotic inference. `https://arxiv.org/abs/1511.04514`.