Two Penalized Estimators based on Variance Stabilization Transforms for Sparse Compressive Recovery with Poisson Measurement Noise

Ajit Rajwade[a], Karthik S. Gurumoorthy[b]

[a]*Department of Computer Science and Engineering, IIT Bombay*
[b]*India Machine Learning, Amazon, Bangalore, India*

Abstract

In this paper, we consider compressive inversion of a signal/image that is sparse in typical orthonormal bases used in image processing, given its measurements that have been corrupted by Poisson noise. The square-root operation is known to convert a Poisson random variable into one that is approximately Gaussian distributed with a constant variance. We present two different computationally tractable, penalized estimators with a data-fidelity term based on the aforementioned square-root based 'variance stabilization transform'. The first estimator has been proposed earlier in the literature, but this is the first paper to analyze its theoretical performance in compressed sensing. Our second estimator is completely novel, and also has the interesting statistical property of being an approximately pivotal estimator. For both estimators, we specifically consider the case of a physically realistic sensing matrix in our analysis. We present detailed performance bounds on

Email addresses: ajitvr@cse.iitb.ac.in (Ajit Rajwade), gurumoor@amazon.com (Karthik S. Gurumoorthy)

[1]AR acknowledges support from SERB Matrics grant MTR/2019/000691.

the ℓ_2 recovery error for purely sparse signals for both estimators, making use of many different Poisson concentration inequalities. Several numerical results are presented, showing the practicality of the proposed estimators.

Keywords: Compressed sensing, Sparse regression, Poisson noise, Variance stabilization transform, Poisson concentration inequalities

1. Introduction

Compressed sensing (CS) is today a popular branch of signal processing. The main aim of CS is to reconstruct a signal/image $\boldsymbol{x} \in \mathbb{R}^m$ from its 'compressive measurements' of the form $\boldsymbol{y} = \boldsymbol{\Phi x} + \boldsymbol{\eta}$ where $\boldsymbol{\Phi} \in \mathbb{R}^{N\times m}, N \ll m$, is a sensing matrix that represents the forward model of the sensing device, and $\boldsymbol{y} \in \mathbb{R}^N$ is a usually noisy measurement vector. Here, $\boldsymbol{\eta} \in \mathbb{R}^N$ is the noise vector, whose elements are usually assumed to be i.i.d. Gaussian distributed. CS theory states that this recovery task is well-posed under certain conditions. In particular, the signal $\boldsymbol{x}$ can be recovered either uniquely or with high accuracy [1] if two sufficient conditions hold: (1) $\boldsymbol{x} \triangleq \boldsymbol{\Psi\theta}$ produces a sparse (or compressible) coefficient vector $\boldsymbol{\theta}$ in some orthonormal basis $\boldsymbol{\Psi} \in \mathbb{R}^{m\times m}$, and (2) $\boldsymbol{A} \triangleq \boldsymbol{\Phi\Psi}$ obeys the so-called restricted isometry property (RIP). The matrix $\boldsymbol{A}$ is said to obey the RIP of order s, if for any s-sparse vector $\boldsymbol{\theta}$[2], we have $(1-\delta_s)\|\boldsymbol{\theta}\|_2^2 \leq \|\boldsymbol{A\theta}\|_2^2 \leq (1+\delta_s)\|\boldsymbol{\theta}\|_2^2$. Here, δ_s is the so-called s-order restricted isometry constant (RIC) of $\boldsymbol{A}$. There exist precise error bounds for the recovery of $\boldsymbol{x}$ [1]. Moreover, most of the algorithms for CS recovery are also efficient and have proven performance bounds. A well-known example is Basis Pursuit Denoising (BPDN), which

[2]This is a vector with at the most s non-zero elements.

optimizes the following cost function:

$$\min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_1 \text{ s.t. } \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{\theta}\|_2 \leq \epsilon, \tag{1}$$

where ϵ is an upper bound on $\|\boldsymbol{\eta}\|_2$. A second example is the LASSO [2], which seeks to minimize the objective function

$$J_{lasso}(\boldsymbol{\theta}) \triangleq \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}\|_2^2 + \rho\|\boldsymbol{\theta}\|_1, \tag{2}$$

given a regularization parameter ρ chosen based on noise statistics. Although BPDN and LASSO are equivalent from an optimization perspective [3], the latter has shown significantly tighter (in fact, minimax optimal) performance bounds than the former. This can be understood by comparing corollary 1.1 of [4], which deals with performance bounds for BPDN to Theorem 11.1 and Example 11.1 of [2], which deals with performance bounds for LASSO.

Most of the theoretical treatment for CS assumes Gaussian noise in $\boldsymbol{y}$. However, most optical or X-Ray systems (compressive or otherwise) exhibit noise that is dominantly Poisson in nature [5, 6, 7, 8, 9]. Such Poisson noise statistics innately dictate that the sensing matrix $\boldsymbol{\Phi}$ as well as the signal $\boldsymbol{x}$, must necessarily be non-negative. There exist a series of estimators with proven bounds for CS recovery under Poisson noise [10, 11, 12, 13, 14, 15, 16]. Almost all of them assume the following form for $\boldsymbol{\Phi}$, first proposed in [10]:

$$\boldsymbol{\Phi} = \sqrt{\frac{1}{4N}}\widetilde{\boldsymbol{\Phi}} + \frac{1}{2N}\mathbf{1}_{N\times m}, \tag{3}$$

where $\mathbf{1}$ is a matrix of ones. Every entry of $\widetilde{\boldsymbol{\Phi}}$ is either $-1/\sqrt{N}$ or $1/\sqrt{N}$ with equal probability, and $\widetilde{\boldsymbol{\Phi}}$ obeys the RIP [17]. This construction ensures that each entry of $\boldsymbol{\Phi}$ is either 0 or $1/N$ (thus non-negative). Moreover $\boldsymbol{\Phi}$

satisfies the important flux-preserving property for Poisson systems, which states that the total measurement flux $\sum_{i=1}^{N} \boldsymbol{\Phi^i x}$ cannot exceed the incident signal flux $\|\boldsymbol{x}\|_1$ [10].

Problem statement: In this paper, we consider the following forward model for the measurements:

$$\forall i \in 1, ..., N, y_i \sim \text{Poisson}(\boldsymbol{\Phi^i \Psi \theta}), \tag{4}$$

where $\boldsymbol{\Phi}$ follows the model in Eqn. 3 and $\boldsymbol{\Phi^i}$ denotes its i^{th} row. Such a forward model is widely used in compressive architectures such as the Rice Single Pixel Camera [6]. We seek to recover $\boldsymbol{\theta}$, the vector of coefficients for signal $\boldsymbol{x}$ in a commonly used orthonormal basis $\boldsymbol{\Psi}$, such as the Haar wavelet transform (HWT) or the discrete cosine transform (DCT).

Contributions: The contributions of our paper are listed here below:

1. We present two computationally tractable penalized estimators for $\boldsymbol{\theta}$ given $\boldsymbol{y}, \boldsymbol{A} \triangleq \boldsymbol{\Phi \Psi}$. For both, we prove bounds on the estimation error $\|\boldsymbol{\theta} - \boldsymbol{\theta^*}\|_2$, where $\boldsymbol{\theta^*}$ is the estimated signal coefficient vector and $\boldsymbol{\theta}$ is the true signal coefficient vector which is assumed to be purely sparse. Both estimators make use of the square-root based variance stabilization transform which converts a Poisson random variable into one which has close to constant (signal-independent) variance.
2. The first estimator has been used earlier in the context of image deblurring under Poisson noise [18], but its theoretical analysis in Poisson compressed sensing is novel (see Sec. 2.4 for advantages over our previous work in [15]). The second estimator that we present is completely novel, and is an approximately pivotal estimator, i.e. the distribution of

Symbol	Dimensions	Meaning
m	Integer	Signal dimension
N	Integer	Number of measurements
$\boldsymbol{x}$	$m \times 1$	Signal vector
$\boldsymbol{y}$	$N \times 1$	Measurement vector
$\boldsymbol{\Phi}$	$N \times m$	Measurement matrix
$\widetilde{\boldsymbol{\Phi}}$	$N \times m$	RIP-obeying component of measurement matrix Φ
$\boldsymbol{\Psi}$	$m \times m$	Orthonormal sparsifying matrix
$\boldsymbol{\Psi_k}$	$m \times 1$	kth column of matrix $\boldsymbol{\Psi}$
$\boldsymbol{\Phi^k}$	$1 \times N$	kth row of matrix $\boldsymbol{\Phi}$
$\boldsymbol{\theta}$	$m \times 1$	Coefficient vector with $\boldsymbol{\theta} = \boldsymbol{\Psi}^T \boldsymbol{x}$
$\boldsymbol{A}$	$N \times m$	$\boldsymbol{A} = \boldsymbol{\Phi\Psi}$
a^*	Real-valued	$a^* \triangleq \max_{i,j} \lvert A_{ij} \rvert$
I	real scalar	signal intensity, i.e. $I = \lVert \boldsymbol{x} \rVert_1$

Table 1: Glossary of symbols used in this paper

the regularization parameter in this estimator is approximately signal independent (Sec. 9.2.2 of [19]). (The approximation becomes increasingly more accurate with increase in signal intensity.) Commonly used estimators in Poisson compressed sensing such as the LASSO [14] or those based on the Poisson negative log likelihood (PNLL) [20, 11] do not possess this property.

3. We present extensive analysis of the theoretical bounds for both estimators, assuming purely sparse $\boldsymbol{\theta}$.
4. We also present a series of numerical experiments showing the practical usage of our estimators.

The rest of this paper is organized as follows. The two estimators and their properties are presented in Sec. 2. In Sec. 2.4, we present a discussion of the relative merits/demerits of our work *in relation to previous literature* in Poisson CS, including our group's previous work on this topic [15, 21]. Several numerical results are presented and discussed in Sec. 3. We conclude in Sec. 4 with a summary of our contributions and directions for future work. For the reader's easy reference, we have prepared a glossary of symbols commonly used in this paper along with their meaning and dimensions, in Table 1.

2. Theory

Given a random variable (r.v.) $y \sim \text{Poisson}(\zeta)$, the r.v. $\sqrt{y}$ is known to be approximately Gaussian distributed [22, 23] with an approximate variance of 0.25, and these approximations become tighter as $\zeta \to \infty$. Inspired by this, we propose two penalized estimators for Poisson CS below. We then comment on the estimators and outline their theoretical properties.

2.1. The Two Estimators

Our first estimator, which we henceforth refer to as VST-1, is presented below:

$$\begin{aligned}\text{VST-1} : \boldsymbol{\theta^*} = \text{argmin}_{\boldsymbol{\theta}} J_1(\boldsymbol{\theta}) &\triangleq \|\sqrt{\boldsymbol{y}} - \sqrt{\boldsymbol{\Phi\Psi\theta}}\|_2^2 + \rho\|\boldsymbol{\theta}\|_1 \\ &= \|\sqrt{\boldsymbol{y}} - \sqrt{\boldsymbol{A\theta}}\|_2^2 + \rho\|\boldsymbol{\theta}\|_1, \\ &\text{s. t. } \boldsymbol{\Psi\theta} \succeq \mathbf{0}_{m\times 1}, \|\boldsymbol{\Psi\theta}\|_1 = I, \end{aligned} \tag{5}$$

where the square-root is performed element-wise; $\succeq$ represents the element-wise $\geq$ inequality; $\mathbf{0}$ stands for a vector with all zeroes; I is the ℓ_1 norm of the signal $\boldsymbol{x}$, also called the signal intensity; and as defined before in Sec. 1, $\boldsymbol{A} \triangleq \boldsymbol{\Phi\Psi}$ where the sensing matrix $\boldsymbol{\Phi}$ obeys the constraints from Eqn. 3. The imposition of the constraint $\|\boldsymbol{x}\|_1 = I$ is necessary only for the theoretical analysis of various estimators in Poisson compressed sensing, but is seen to be not required in numerical experiments in the literature [14, 12, 13, 16, 15]. Note that the data fidelity term $L_1(\boldsymbol{y}; \boldsymbol{A\theta}) \triangleq \|\sqrt{\boldsymbol{y}} - \sqrt{\boldsymbol{A\theta}}\|_2^2$ in the cost function J_1 is convex and differentiable [15], just as the fidelity terms for LASSO and PNLL.

Our second estimator VST-2 is as follows:

$$\begin{aligned}\text{VST-2} : \boldsymbol{\theta^*} = \text{argmin}_{\boldsymbol{\theta}} J_2(\boldsymbol{\theta}) &\triangleq \sum_{i=1}^{N} (\sqrt{y_i} - \sqrt{\boldsymbol{\Phi^i\Psi\theta}})^2 (2\sqrt{\boldsymbol{\Phi^i\Psi\theta}} + \sqrt{y_i}) + \rho\|\boldsymbol{\theta}\|_1 \\ &= \sum_{i=1}^{N} (\sqrt{y_i} - \sqrt{\boldsymbol{A^i\theta}})^2 (2\sqrt{\boldsymbol{A^i\theta}} + \sqrt{y_i}) + \rho\|\boldsymbol{\theta}\|_1 \\ &\text{s. t. } \boldsymbol{\Psi\theta} \succeq \mathbf{0}_{m\times 1}, \|\boldsymbol{\Psi\theta}\|_1 = I. \end{aligned} \tag{6}$$

Just as used in Eqn. 5, we have used $\boldsymbol{A} \triangleq \boldsymbol{\Phi\Psi}$ where $\boldsymbol{\Phi}$ follows the model

defined in Eqn. 3. The advantage of VST-2 is that the choice of the regularization parameter ρ that ensures theoretical performance guarantees, is independent of the intensity I. We will elaborate on this and other advantages in Sec. 2.4. This property is not observed for VST-1. Another advantage of VST-2 over VST-1 is that the derivative of the data-fidelity term (w.r.t. $\boldsymbol{\theta}$) in VST-2 is always bounded for all non-negative signals, including for $\boldsymbol{\theta} = \mathbf{0}$. This advantage is *not* shared by the negative log-likelihood of the Poisson distribution which given by $\sum_{i=1}^{N} \boldsymbol{A^i\theta} - y_i \log \boldsymbol{A^i\theta} +$ constants . Clearly, the Poisson negative log-likelihood as well as its gradient both have an unbounded value if $\boldsymbol{A^i\theta} = 0$ for any $i \in \{1, ..., N\}$. Additionally, besides being differentiable, the data-fidelity term in VST-2 is also convex, as stated in Lemma 4A in Sec. 2.3.3 (and proved in the supplemental material [24]).

2.2. Theoretical Guarantees: Preliminaries

Our proof technique for the performance of VST-1 and VST-2 uses a key lemma and theorem from the seminal work in [25], which we state below using the notation from this paper for an estimator of the following form, using a prototypical data fidelity term L:

$$\begin{aligned} \mathsf{(PP)} : \boldsymbol{\theta^*} &= \text{argmin}_{\boldsymbol{\theta}} L(\boldsymbol{y}, \boldsymbol{A\theta}) + \rho \|\boldsymbol{\theta}\|_1 \\ &\text{s. t. } \boldsymbol{\Psi\theta} \succeq \mathbf{0}_{m\times 1}, \|\boldsymbol{\Psi\theta}\|_1 = I. \end{aligned} \tag{7}$$

Nonetheless, we emphasize that our work contains many novel, non-trivial innovations, using different properties of the Poisson distribution, for deriving the theoretical results for both estimators (which will follow soon).

Lemma L1: (Lemma 1 of [25]): Let $\boldsymbol{\theta^*}$ be the optimum of the cost function in Eqn. 7 with a regularization parameter $\rho \geq 2\|\nabla L(\boldsymbol{y}; \boldsymbol{A\theta})\|_\infty$. Then

the error vector $\boldsymbol{\Delta} \triangleq \boldsymbol{\theta}^* - \boldsymbol{\theta}$ belongs to the set $\mathbb{C}(S;\boldsymbol{\theta}) \triangleq \{\boldsymbol{\Delta} | \|\boldsymbol{\Delta}_{S^c}\|_1 \leq 3\|\boldsymbol{\Delta}_S\|_1 + 4\|\boldsymbol{\theta} - \boldsymbol{\theta}_S\|_1\}$, where S is the set of indices of the s largest absolute value elements of $\boldsymbol{\theta}$, and $\forall i \in S, \theta_S(i) = \theta_i; \forall i \notin S, \theta_S(i) = 0$. ■

Lemma L1 is interesting because it states that for the optimization problem in Eqn. 7 with an appropriate regularization parameter, the error vector $\boldsymbol{\Delta}$ is guaranteed to be restricted to lie in a very specific set $\mathbb{C}(S;\boldsymbol{\theta})$ which depends on the unknown vector $\boldsymbol{\theta}$. If $\boldsymbol{\theta}$ is s-sparse (i.e. it has at the most s non-zero elements), then the constraint set for $\mathbb{C}(S;\boldsymbol{\theta})$ reduces to $\{\boldsymbol{\Delta} | \|\boldsymbol{\Delta}_{S^c}\|_1 \leq 3\|\boldsymbol{\Delta}_S\|_1\}$ which is a cone.

Definition D1: A loss function L is said to obey the restricted strong convexity (RSC) property with curvature $\kappa_L > 0$ and tolerance function $\tau_L(\boldsymbol{\theta})$ if the Bregman divergence

$$\begin{aligned} \delta L(\boldsymbol{\Delta}, \boldsymbol{\theta}) &\triangleq L(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{\theta}^*) - L(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{\theta}) - \nabla L(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{\theta})^t(\boldsymbol{\Delta}) \\ &\geq \kappa_L \|\boldsymbol{\Delta}\|_2^2 - \tau_L^2(\boldsymbol{\theta}), \end{aligned}$$

for every vector $\boldsymbol{\Delta} \in \mathbb{C}(S;\boldsymbol{\theta})$. ■

The Bregman divergence $\delta L(\boldsymbol{\Delta}, \boldsymbol{\theta})$ is the error between the loss function value at $\boldsymbol{\theta}^*$ and its first order Taylor series expansion about $\boldsymbol{\theta}$. Intuitively, a loss function that obeys RSC is sharply curved around $\boldsymbol{\theta}$, so that any difference in the loss function $|L(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{\theta}) - L(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{\theta}^*)|$ will imply a *proportional* estimation error $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_1$ for all error vectors $\boldsymbol{\theta}^* - \boldsymbol{\theta} \in \mathbb{C}(S;\boldsymbol{\theta})$. We refer the reader to [25] for more details.

Theorem T1: (Theorem 1 of [25]) If L is convex, differentiable and satisfies the restricted strong convexity (RSC) property for perturbations $\boldsymbol{\Delta}$ in $\mathbb{C}(S;\boldsymbol{\theta})$ with curvature κ_L and tolerance $\tau_L^2(\boldsymbol{\theta})$, and $\boldsymbol{\theta}^*$ is the solution to the

problem in Eqn. 7 with $\rho \geq 2\|\nabla L(\boldsymbol{y}; \boldsymbol{A\theta})\|_\infty$, then we have:

$$\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2^2 \leq \frac{9\rho^2 s}{\kappa_L^2} + \frac{\rho}{\kappa_L}(2\tau_L^2(\boldsymbol{\theta}) + 4\|\boldsymbol{\theta} - \boldsymbol{\theta}_S\|_1). \blacksquare$$

In this paper, we specifically concentrate on sparse $\boldsymbol{\theta}$ for easier exposition. Hence we would have $\boldsymbol{\theta} - \boldsymbol{\theta_S} = \mathbf{0}$. For both our estimators VST-1 and VST-2, we also actually prove that $\tau_L^2(\boldsymbol{\theta}) = 0$. (As per [25, Defn. D2, Eqn. 19], it is expected that $\tau_L^2(\boldsymbol{\theta}) = 0$ for most statistical models.) We comment later on possible extensions of our work to compressible (or weakly sparse) signals.

2.3. Theoretical Guarantees for our Estimators

With the mathematical background from Sec. 2.2, we now first present several theoretical properties of VST-1 in Sec. 2.3.1, and prove performance bounds for the minimizer of $J_1(.)$. We emphasize that the proofs of our theoretical bounds do not directly assume that $\sqrt{\boldsymbol{y}}$ is Gaussian-distributed. The Gaussianity is only asymptotically true for $\sqrt{\boldsymbol{y}}$ (i.e. when $E(\boldsymbol{y})$ is very large), and hence the assumption would reduce the rigour of the proofs. We then present theoretical analysis of VST-2 in Sec. 2.3.2. For both VST-1 and VST-2, we present analysis on recovery of *purely* sparse signals. Currently our theoretical analysis does not encompass weakly sparse signals, though numerically both estimators show good performance on them.

2.3.1. Analysis of VST-1

We state two important lemmas for VST-1, both proved in the supplemental material in [24]. Before that, we state the definition of a useful property of a sensing matrix.

Definition D2: Restricted Eigenvalue Condition of a Sensing Matrix: A sensing matrix $\boldsymbol{\Phi}$ is said to obey the s-order restricted eigenvalue

condition (REC) if for every vector $\boldsymbol{\Delta} \in \mathbb{C}(S;\boldsymbol{\theta})$ (see definition in Lemma L1), we have $\frac{\|\boldsymbol{\Phi\Delta}\|^2}{m} \geq \gamma_s \|\boldsymbol{\Delta}\|^2$, where $\gamma_s > 0$ is the restricted eigenvalue constant.

As argued in [25, Sec. 4.1, Eqns. 28 and 29] and [2, Sec. 11.2.2], the RSC property of the squared loss function implies the REC because in such a case the first-order Taylor series expansion for the Bregman divergence (definition D1) is exact and dependent only on $\boldsymbol{\Delta}$ and not on $\boldsymbol{\theta}$. The REC is obeyed with high probability by a wide variety of randomly generated sub-Gaussian matrices, including by $\widetilde{\boldsymbol{\Phi}}$ from Eqn. 3 [25].

Lemma 1: As defined before, let I be the signal intensity and N be the number of measurements. If the matrix $\tilde{\boldsymbol{\Phi}}\boldsymbol{\Psi}$, where $\tilde{\boldsymbol{\Phi}}$ is as per Eqn. 3, satisfies the s-order restricted eigenvalue condition with constant γ_s, then the function $L(\boldsymbol{y};\boldsymbol{A\theta}) \triangleq \|\sqrt{\boldsymbol{y}} - \sqrt{\boldsymbol{A\theta}}\|_2^2$ (the data fidelity term for VST-1) satisfies the restricted strong convexity property with curvature term $\kappa_L \geq \frac{\gamma_s}{16I\sqrt{2}}$ and tolerance function $\tau_L^2(\boldsymbol{\theta}) = 0$, with probability greater than or equal to $(1 - N^{-C_1/12})^N$, if $\forall i \in \{1, 2, ..., N\}$, $\boldsymbol{A^i\theta} \geq CI$ and $I \geq C_1 N \log N$, for some constants $C > 0, C_1 > 12$. ■

The assumed lower bounds on I and $\boldsymbol{A^i\theta}$ merely reflect that the measurement or signal intensity should be sufficiently large for the bounds to hold. This is typical of Poisson inverse problems (eg: [14, 10, 16, 15]).

Lemma 2: Define $\lambda_i \triangleq \boldsymbol{A^i\theta}; w_{ij} \triangleq A_{ij}/\sqrt{\lambda_i}; w^* \triangleq \max_{ij}|w_{ij}|; v_i \triangleq E(y_i)E(1/(4y_i + 1)); \bar{v} \triangleq \sum_{i=1}^N v_i$.

Also consider for all i that $\lambda_i \geq CI$ for $C > 0$, and that $I \geq C_1 N \log N$ for $C_1 > 12$. Then for any $\tau > 0$, $P(\|\nabla L(\boldsymbol{y};\boldsymbol{A\theta})\|_\infty \geq 2\tau w^* \log m) \leq$

$4\exp\Big(-\frac{\log m}{2}\Big[\tau\log\Big(\frac{\tau\log m}{2\bar{v}}+1\Big)-2\Big]\Big)$. ■

The optimal regularization parameter ρ must satisfy $\rho \geq 2\|\nabla L\|_\infty$ for the bounds stated in Theorem T1 to hold. Hence, we have $\rho \geq 4\tau w^* \log m$. But $w^* \leq \dfrac{a^*}{\sqrt{CI}}$ since $\boldsymbol{A^i\theta} \geq CI$ as used in Lemma 1 (where a^* is defined in Table 1). This leads us to the relation $\rho \geq 4\tau a^* \log m/\sqrt{CI}$. Given these lemmas, we now state the following theorem (proved in the supplemental material [24]).

Theorem 3: Consider N compressive measurements $y_i \sim \text{Poisson}(\boldsymbol{\Phi^i\Psi\theta})$, where $\boldsymbol{\Phi}$ obeys the structure in Eqn. 3 and $\widetilde{\boldsymbol{\Phi}}\boldsymbol{\Psi}$ obeys the restricted eigenvalue condition (REC) of some order s with constant γ_s[3]. Consider that $\forall i, \boldsymbol{\Phi^i\Psi\theta} \geq CI$ for some constant C. Let $\boldsymbol{\theta^*}$ be the minimum of the cost function in Eqn. 5 using $\rho = 4\tau a^* \log m/\sqrt{CI}$ where $\tau > 0$ and a^* is defined in Table 1, and let us assume that the cost function obeys the RSC (which, as per Lemma 1 holds with high probability). Consider $\bar{v}$ as defined in Lemma 2. Then, the following upper bound holds with probability $1-4\exp\Big(-\frac{\log m}{2}\Big[\tau\log\Big(\frac{\tau\log m}{2\bar{v}}+1\Big)-2\Big]\Big)$:

$$\frac{\|\boldsymbol{\theta^*}-\boldsymbol{\theta}\|_2^2}{I^2} \leq \frac{C_2(\log m)^2 s}{I}$$

for a suitably defined constant C_2. ■

We present a careful discussion of Lemmas 1 and 2, and Theorem 3 in Sec. 2.4.

[3]By construction, $\widetilde{\boldsymbol{\Phi}}\boldsymbol{\Psi}$ satisfies the RIP [17], hence also the weaker REC condition[26, 27]. The REC is tied to the cone constraint -Lemma 1 of [25].

2.3.2. Motivation for VST-2

Consider a random variable $X \sim F(x;\chi)$ where $F(.)$ is a distribution parametrized by χ. The random variable $q(X;\chi)$ is said to be a **pivotal quantity** (or a pivot) w.r.t. χ if $q(X;\chi)$ has the same distribution for all values of χ [19, Sec 9.2.2]. Pivotal quantities have many applications in determining confidence intervals for various distribution, as shown in [19, Sec 9.2.2].

Recall that as per Lemma 2, we require that the regularization parameter $\rho \geq 2\|\nabla L(\boldsymbol{y};\boldsymbol{A\theta})\|_\infty$. When $L(\boldsymbol{y};\boldsymbol{A\theta}) \triangleq \|\sqrt{\boldsymbol{y}} - \sqrt{\boldsymbol{A\theta}}\|^2$ as in VST-1, then $\nabla L(\boldsymbol{y};\boldsymbol{A\theta}) = -\boldsymbol{A}^T(\sqrt{\boldsymbol{y}} - \sqrt{\boldsymbol{A\theta}}) \oslash \sqrt{\boldsymbol{A\theta}}$ where $\oslash$ denotes element-wise division. Noting that $\sqrt{\boldsymbol{y}}$ is a random variable with a distribution approximately $\mathcal{N}(\sqrt{\boldsymbol{A\theta}}, 1/4)$, we see that the distribution of $\|\nabla L(\boldsymbol{y};\boldsymbol{A\theta})\|_\infty$ will be dependent on intensity I as well as the measurement $\boldsymbol{A\theta}$. This is due to the term $\boldsymbol{A\theta}$ in the denominator of the expression for $\nabla L(\boldsymbol{y};\boldsymbol{A\theta})$.

On the other hand, when the data fidelity term is $L(\boldsymbol{y};\boldsymbol{A\theta}) = \sum_{i=1}^{N}(\sqrt{y_i} - \sqrt{\boldsymbol{A^i\theta}})^2(2\sqrt{\boldsymbol{A^i\theta}} + \sqrt{y_i})$ as in VST-2, then we see that $\nabla L(\boldsymbol{y};\boldsymbol{A\theta}) = 3\boldsymbol{A}^T(\sqrt{\boldsymbol{A\theta}} - \sqrt{\boldsymbol{y}})$. Since $\sqrt{\boldsymbol{y}} \sim^{approx} \mathcal{N}(\sqrt{\boldsymbol{A\theta}}, 1/4)$, in this case we observe that the distribution of $\|\nabla L(\boldsymbol{y};\boldsymbol{A\theta})\|_\infty$ is approximately independent of the signal or its intensity, due to which it is **approximately pivotal** ([19, Sec. 9.2.2]) w.r.t. signal intensity I and individual signal values. Hence the VST-2 estimator is also (approximately) pivotal, forming a primary motivation for its development. We provide a more rigorous explanation of this property in Lemma 5. For more details, please refer to Lemmas 4 and 5 and their proofs, as well as Fig. 1.

We note that a pivotal estimator in the context of sparse Poisson regression

has been presented recently in [28], inspired from the square-root LASSO estimator for white noise in [29]. However their estimator is applicable for a non-linear measurement model of the form $y_i \sim \text{Poisson}(e^{-\boldsymbol{\Phi^i\Psi\theta}})$, which is natural in computed tomography due to the well-known Beer's law [30, 8], but not in compressed sensing. Consequently, their analysis cannot be used in our setting, and vice versa. The square-root LASSO estimator in [29] replaces the $\|\boldsymbol{y} - \boldsymbol{A\theta}\|_2^2$ term in the LASSO by $\|\boldsymbol{y} - \boldsymbol{A\theta}\|_2$ to arrive at a regularizer term whose distribution is independent of the noise variance. Note however that in our work, merely replacing the $\|\sqrt{\boldsymbol{y}} - \sqrt{\boldsymbol{A\theta}}\|_2^2$ term in VST-1 by $\|\sqrt{\boldsymbol{y}} - \sqrt{\boldsymbol{A\theta}}\|_2$ does not suffice. This is because the distribution of the gradient of the latter term is *not* signal independent.

VST-2 has two more advantages over VST-1. The derivative of the data-fidelity term in VST-1 will be unbounded if $\exists i$ for which $\boldsymbol{A^i\theta} = 0$, potentially causing problems in implementing gradient-based optimization methods. However the corresponding derivative in VST-2 will always remain bounded. Secondly, we are able to establish the RSC property of the data-fidelity term in VST-1 only if certain lower bounds on the values of intensity I as well as on $\boldsymbol{A^i\theta}$ are satisfied, as is seen in the statement and proof of Lemma 1. At this point, we are unaware whether such a lower bound is absolutely necessary. Nonetheless, we are able to establish the RSC of the data fidelity term in VST-2 without any such restrictions, as will be evident from the statement of Lemma 4 and its proof.

2.3.3. Analysis of VST-2

We now state important lemmas and a theorem for the performance of VST-2, all of which are proved in the supplemental material [24].

Lemma 4: If the matrix $\widetilde{\boldsymbol{\Phi}}\boldsymbol{\Psi}$, where $\widetilde{\boldsymbol{\Phi}}$ is as per Eqn. 3, satisfies the s-order restricted eigenvalue condition with constant γ_s, then the function $L(\boldsymbol{y}; \boldsymbol{A\theta}) = \sum_{i=1}^{N}(\sqrt{y_i} - \sqrt{\boldsymbol{A^i\theta}})^2(2\sqrt{\boldsymbol{A^i\theta}} + \sqrt{y_i})$ (the data fidelity term in VST-2) satisfies the restricted strong convexity property with curvature $\kappa_L \geq \frac{\gamma_s}{8\sqrt{NI}}$ and tolerance function $\tau_L^2(\boldsymbol{\theta}) = 0$, where I is the signal intensity and N is the number of measurements. ■

Lemma 4A: The data fidelity function $L(\boldsymbol{y}; \boldsymbol{A\theta}) = \sum_{i=1}^{N}(\sqrt{y_i} - \boldsymbol{A^i\theta})^2(2\sqrt{\boldsymbol{A^i\theta}} + \sqrt{y_i})$ is convex in $\boldsymbol{\theta}$. ■

Lemma 5: Given the matrix $\boldsymbol{A}$, define $v_i \triangleq E(y_i)E(1/(4y_i+1)); \bar{v} \triangleq \sum_{i=1}^{N} v_i$. Let $L(\boldsymbol{y}; \boldsymbol{A\theta}) = \sum_{i=1}^{N}(\sqrt{y_i} - \sqrt{\boldsymbol{A^i\theta}})^2(2\sqrt{\boldsymbol{A^i\theta}} + \sqrt{y_i})$. Then for any $\tau > 0$, we have
$P(\|\nabla L(\boldsymbol{y}; \boldsymbol{A\theta})\|_\infty \geq 2\tau a^* \log m) \leq 4\exp\Big(-\frac{\log m}{2}\Big[\tau \log\Big(\frac{\tau \log m}{2\bar{v}} + 1\Big) - 2\Big]\Big)$
where a^* is defined in Table 1. ■

Theorem 6: Consider N compressive measurements $y_i \sim \text{Poisson}(I\boldsymbol{\Phi^i\Psi\theta})$, where $\boldsymbol{\Phi}$ obeys the structure in Eqn. 3 and $\widetilde{\boldsymbol{\Phi}}\boldsymbol{\Psi}$ obeys the restricted eigenvalue condition (REC) of some order s with constant γ_s. Let $\boldsymbol{\theta^*}$ be the minimum of the cost function in Eqn. 6 using $\rho = 4\tau a^* \log m$ where $\tau > 0$ and a^* is defined in Table 1. Consider $\bar{v}$ as defined in Lemma 5. Then, the following upper bound holds with probability $1 - 4\exp\Big(-\frac{\log m}{2}\Big[\tau \log\Big(\frac{\tau \log m}{2\bar{v}} + 1\Big) - 2\Big]\Big)$:
$\frac{\|\boldsymbol{\theta^*} - \boldsymbol{\theta}\|_2^2}{I^2} \leq \frac{C_2(\log m)^2 s}{I}$, for a suitably defined constant C_2. ■

2.4. Discussion and Comparisons with Previous Work

1. *Constraint on signal L1-norm:* For facilitating theoretical analysis, our estimators impose the constraint that $\|\boldsymbol{x}\|_1 = I$, just as in other estimators [14, 12, 10] for Poisson compressive recovery, including our own previous work in [16, 15]. Simulations in prior work reveal that this constraint is not required in practice (i.e. the quality of signal recovery in numerical experiments is not affected if this constraint is dropped), and it is required only for the theoretical analysis.
2. *Behaviour of performance bounds w.r.t. N, I, s:* The performance bounds for estimators VST-1 as well as VST-2 are measured in terms of squared relative errors of the form $\|\boldsymbol{\theta} - \boldsymbol{\theta^*}\|^2/I^2$, which is typical in Poisson problems since the noise variance is proportional to I. The relative error decreases w.r.t. I, remains constant with N (as long as N is large enough to satisfy the RIP of $\widetilde{\boldsymbol{\Phi}}$) and increases with s. These trends are in line with existing literature on Poisson compressed sensing, such as [14, 12]. The flux-preserving construction of Φ from Eqn. 3 is obtained by dividing elements of a Bernoulli matrix by N. Division by $\sqrt{N}$ instead of by N would have given rise to an RIP-obeying matrix. The division by an extra factor of $\sqrt{N}$ (which is required for flux-preservation - see Eqn. 3) leads to more rapid reduction in measurement SNR with increase in N. Due to this, the upper bounds on the error do not decrease with increase in N (beyond what is required for ensuring RIP). On the other hand, typical bounds with estimators such as the LASSO with matrices that are not flux-preserving show decrease in upper bounds w.r.t. N [2, Theorem 11.1].

3. *Comments regarding a^*:* Since we assume $\|\boldsymbol{x}\|_1 = I$, this implies $\theta_1 = I/\sqrt{m}$ for any orthonormal basis $\boldsymbol{\Psi}$ whose first column is a vector with all values equal to $1/\sqrt{m}$ (eg: DCT or HWT bases). Therefore, the unknowns in the objective function for both VST-1 and VST-2 are only the coefficients in $\{\theta_j\}_{j=2}^m$. Hence in the computation of a^*, we can ignore the first column of $\boldsymbol{\Psi}$, and hence the first column of $\boldsymbol{A}$. In such cases, we have observed experimentally that a^* is independent of m - see the last section of the supplemental material [24]. In general, $a^* \propto 1/N$ due to the construction of $\boldsymbol{\Phi}$ in Eqn. 3.
4. *Comments regarding $\bar{v}$*: For VST-1, the value of $\bar{v}$ is $O(N)$, as seen in the proof of Lemma 2 in the supplemental material [24]. This is primarily because of the lower bounds $\boldsymbol{A^i\theta} \geq CI$ and $I \geq C_1 N \log N$ in Lemma 1. In the case of VST-2, no such lower bounds on I are required for the RSC of the cost function (compared Lemmas 1 and 4). Hence $\bar{v}$ is asymptotically a constant independent of N. However for small values of N (which are meaningful in compressed sensing), we empirically observe that that $\bar{v}$ is $o(\sqrt{N})$ as seen in the last section of the supplemental material [24].
5. *Comments regarding τ:* In the case of VST-1, we have seen earlier that $\bar{v}$ is $O(N)$ but $a^* \propto O(1/N)$. For the probability with which the bounds of Theorem 3 hold to be meaningful, it is necessary to consider $\tau \log\left(\frac{\tau \log m}{2\bar{v}} + 1\right) > 2$. Hence we consider $\tau = O(N)$ which ensures that $a^*\tau$ is $O(1)$ (since a^* is $O(N)$), and makes ρ independent of N just like the curvature κ_L (see Lemma 1). Considering τ and $\bar{v}$ to be $O(N)$, the probability with which the upper bound in Theorem

3 holds is at least $1 - 4\exp\Big(-\frac{\log m}{2}\Big[N\log\Big(\log m + 1\Big) - 2\Big]\Big)$. See the comments at the end of the proof of Lemma 2 in [24] for more details. In the case of VST-2, we choose τ to be $O(\sqrt{N})$ so that $a^*\tau$ is $O(1/\sqrt{N})$. This also makes $\rho = O(1/\sqrt{N})$, i.e. of the same order as κ_L which is also $O(1/\sqrt{N})$ (see Lemma 4). For the probability with which the bounds of Theorem 6 hold to be meaningful, it is necessary to consider $\tau\log\Big(\frac{\tau\log m}{2\bar{v}} + 1\Big) > 2$. With τ and $\bar{v}$ both being $O(\sqrt{N})$, this probability is at least $1 - 4\exp\Big(-\frac{\log m}{2}\Big[\sqrt{N}\log\Big(\log m + 1\Big) - 2\Big]\Big)$. See the comments at the end of the proof of Lemma 5 in [24] for more details.

6. *Tractability:* Our estimators VST-1 and VST-2 are both computationally tractable just like those in [14, 15, 16, 13, 11] as opposed to intractable estimators in [10, 12].
7. *Flux-preserving matrices:* Our estimator uses flux preserving sensing matrices like those in [14, 15, 16] but unlike the work in [11].
8. *Constrained/unconstrained VST-based estimator:* Our group has previously performed extensive analysis [15, 21] for the following constrained version of the estimator in Eqn. 5:

$$\min\|\boldsymbol{\theta}\|_1, \text{ s. t. } \|\sqrt{\boldsymbol{y} + 3/8} - \sqrt{\boldsymbol{A}\boldsymbol{\theta} + 3/8}\|_2 \leq \varepsilon,$$
$$\|\boldsymbol{\Psi}\boldsymbol{\theta}\|_1 = 1, \text{ and } \boldsymbol{\Psi}\boldsymbol{\theta} \succeq \boldsymbol{0},$$

where ε is carefully chosen based on the first few moments of the Poisson distribution. The bounds from Theorems 3 and 6 for the *unconstrained* estimators VST-1 and VST-2 considered in this paper are significantly tighter than the $O(\sqrt{N})$ type bounds for the *constrained* estimator in [15].

9. *Main innovations:* One key theoretical innovation in this paper is the proof of the RSC property of the two data-fidelity terms for VST-1 and VST-2 respectively. The other key innovation is the use of a concentration inequality for the square-root of a Poisson r.v. from [31], and the use of the one-sided Bernstein bound [32], both in the proofs of Lemmas 2 and 5.

10. *Comparisons with* LASSO*:* The LASSO has been used for Poisson CS in [14], and also in [13] but with a carefully designed weighted regularizer. The work in [14] imposes a lower bound of $I \geq O(N \log m)$ on the intensity, whereas we require $I > C_1 N \log N$ (where C_1 is constant) in VST-1, in addition to a lower bound on each $\boldsymbol{A^i\theta}$. Furthermore, we do not require any such lower bound in VST-2. Moreover, our estimator VST-2 is approximately pivotal with respect to the intensity, unlike the LASSO estimator whose optimal regularization parameter is inversely proportional to the intensity I [14]. Moreover the probability with which the performance bounds for VST-1 and VST-2 hold, increase with the number of measurements, as opposed to the $1 - 2/m$ probability for the LASSO bounds reported in [14].

11. *VST versus negative log likelihood:* The work in [11] establishes the RSC for the Poisson negative log-likelihood (PNLL) for strictly sparse signals and not for flux-preserving matrices. Unlike the PNLL, the square-root based VST is also applicable to Poisson-Gaussian noise or to any noise model whether the variance is proportional to the mean [33] (Sec. 14.6 and 14.7), though we haven't explored this aspect in this paper. Compared to PNLL, our VST-2 has the advantage of pivotal

estimation. Also, our data fidelity function (as well as its gradient) has a bounded value even if $\boldsymbol{A^i\theta} = 0$ for the i^{th} measurement, as mentioned in Sec. 2.3.2.

12. *Extension to weakly sparse signals:* The bounds for estimators VST-1 and VST-2 have been derived only for sparse signals. However, the analysis can be extended to handle weakly sparse signals following the technique in [14]. Moreover, in Sec. 3 we show numerical results with weakly sparse signals as well.

13. *Non-linear transformation of measurements:* Among all prevailing estimators except for our work in [15], our technique is the only one to apply a non-linear transformation on the Poisson-corrupted linear measurements in $\boldsymbol{y}$. The non-linearity is not necessarily an advantage but only a peculiar feature of our technique. It is interesting to see that such non-linear transformations do not adversely affect the performance of the estimator in any aspect.

3. Experimental Results

In this section, we present empirical results to show the behavior of our estimators on toy 1D signals in Sec. 3.1 and for image reconstruction in Sec. 3.2.

3.1. Experiments on 1D signals

We present results on reconstruction of non-negative 1D signals with $m = 256$ elements. The signals were constructed to be sparse/weakly sparse in the 1D-DCT basis. The active signal coefficients were chosen randomly

from Uniform$(0, 1)$, and every signal had a different support. The signals were normalized to be non-negative and have a particular intensity I. Compressive measurements were generated using the model in Eqn. 3.

Comparisons: We compared the performance of our estimators VST-1 and VST-2 with that of the LASSO and the Poisson negative log-likelihood estimator (PNLL) defined as $\text{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{N} (\boldsymbol{A^i\theta} - y_i \log \boldsymbol{A^i\theta}) + \rho\|\boldsymbol{\theta}\|_1)$. In all these estimators, the non-negativity constraint $\boldsymbol{\Psi\theta} \succeq \mathbf{0}$ was always imposed. However, we did not explicitly impose the the constraint $\|\boldsymbol{\Psi\theta}\|_1 = I$ as it had negligible impact on the performance of Poisson CS as also previously reported in [10, 15, 16]. The comparison measure was the median RRMSE $\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2/I$ between the true $\boldsymbol{\theta}$ and its reconstruction $\hat{\boldsymbol{\theta}}$ from compressive measurements corrupted by 10 different Poisson noise realizations. The optimal ρ in all cases was chosen omnisciently to yield the best RRMSE assuming knowledge of the ground truth signal. However, a major advantage of VST-2 is its approximately pivotal nature and the consequent signal-independent nature of the parameter ρ that guarantees statistical consistency. To highlight this, we **auto-tune** ρ by training on a set of 20 non-negative signals with elements chosen from a uniform distribution, and with sparsity and intensity values that were *different* from the signal to be reconstructed (i.e. the signals on which RRMSE values are reported). Only the number of measurements N was kept the same for auto-tuning ρ and for signal reconstruction. We set ρ in VST-2 to be equal to twice the 95 percentile of the absolute values of the data fidelity gradient $\nabla L = 3\boldsymbol{A}^T(\sqrt{\boldsymbol{y}} - \sqrt{\boldsymbol{A\theta}})$, as computed empirically from the training set. We henceforth refer to this version of VST-2 as VST2-AUTO. Despite having a somewhat conservative ρ, we shall see

that VST2-AUTO yields performance that is in many cases comparable to the omniscient VST-2 and omniscient versions of the other estimators. We demonstrate the relative invariance of the statistics of the gradient of the data fidelity for estimator VST-2 term w.r.t. the signal or signal intensity in Fig. 1. The statistics are collected from 40 randomly generated signals with $m = 256$ and for $N = 200$ Poisson-corrupted compressive measurements following the model in Eqn. 3. Note that the statistics are shown on the actual values of the gradient and not the absolute values. However, we consider the absolute values in our estimate of the 95-percentile used in determining ρ. Further, we also note that such auto-tuning is not possible for VST-1, LASSO or PNLL as they are not pivotal estimators, and L in these estimators will be signal-dependent.

Implementation Details: Our estimators VST-1 and VST-2, as well as PNLL and LASSO were implemented using the well-known CVX library[4] with the SDPT3 solver.

Experiments: In the **intensity experiment**, we set $N = 200$ and enforced signal sparsity to $s = \|\boldsymbol{\theta}\|_0 = 20$, and varied the intensity I from 10 to 10^7 in multiples of 10. We observed decrease in RRMSE w.r.t. I for all estimators in Fig. 2, as is typical of Poisson problems. At low intensities, PNLL tends to perform slightly better than other estimators.

In the **measurements experiment**, we set $I = 10^7, s = 20$ and varied the number of measurements N from 50 to 240 in steps of 10. With increase in N beyond a lower bound of about $N = 100$ (which is essential for properties

[4] `http://cvxr.com/cvx/`

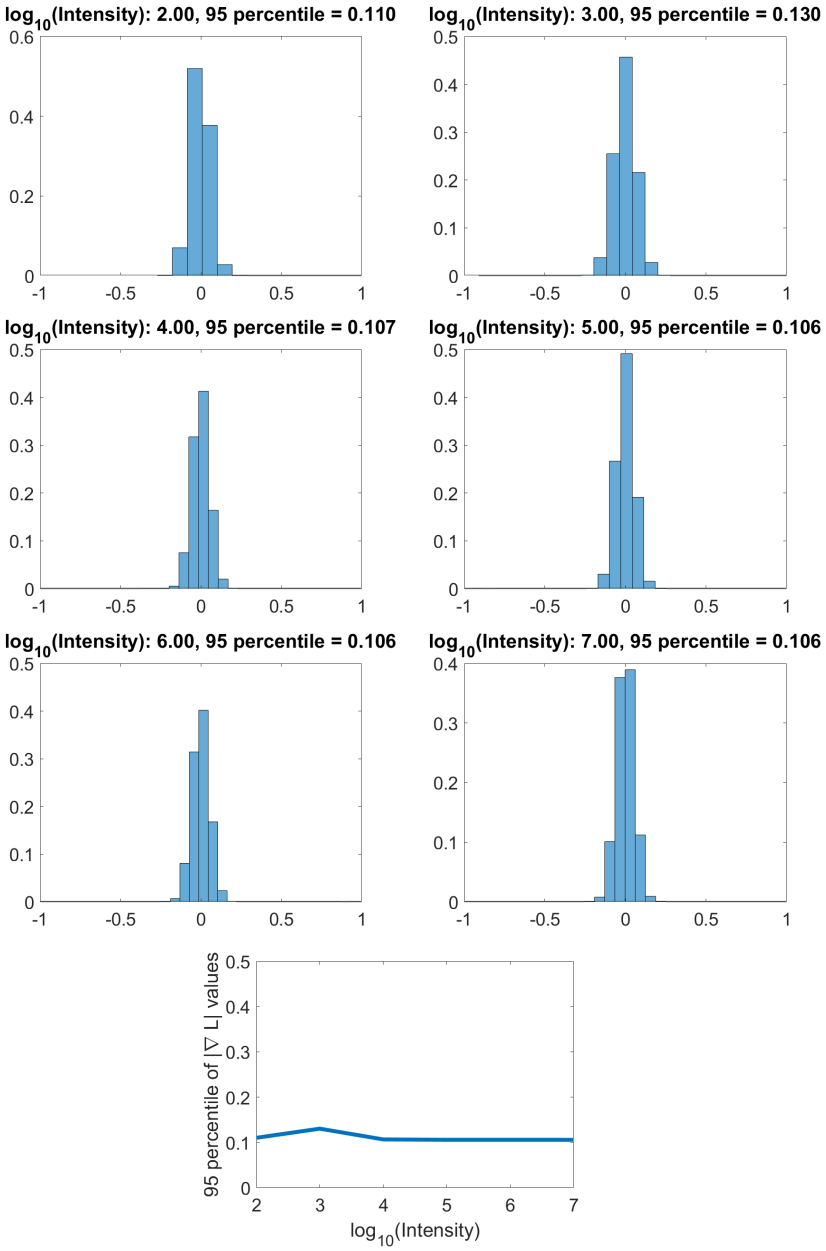

Figure 1: Left to right, top to bottom: histograms of the values of $\nabla L = 3\boldsymbol{A}^T(\sqrt{y} - \sqrt{\boldsymbol{A}\boldsymbol{\theta}})$ for intensity $I \in \{10^2, 10^3, ..., 10^7\}$. Last figure: Plot of the 95 percentile of the absolute gradient value versus $\log_{10}(I)$. The final ρ was chosen to be the average of these values. The statistics are collected from 40 randomly generated signals with $m = 256$ and for $N = 200$ Poisson-corrupted compressive measurements following the model in Eqn. 3.

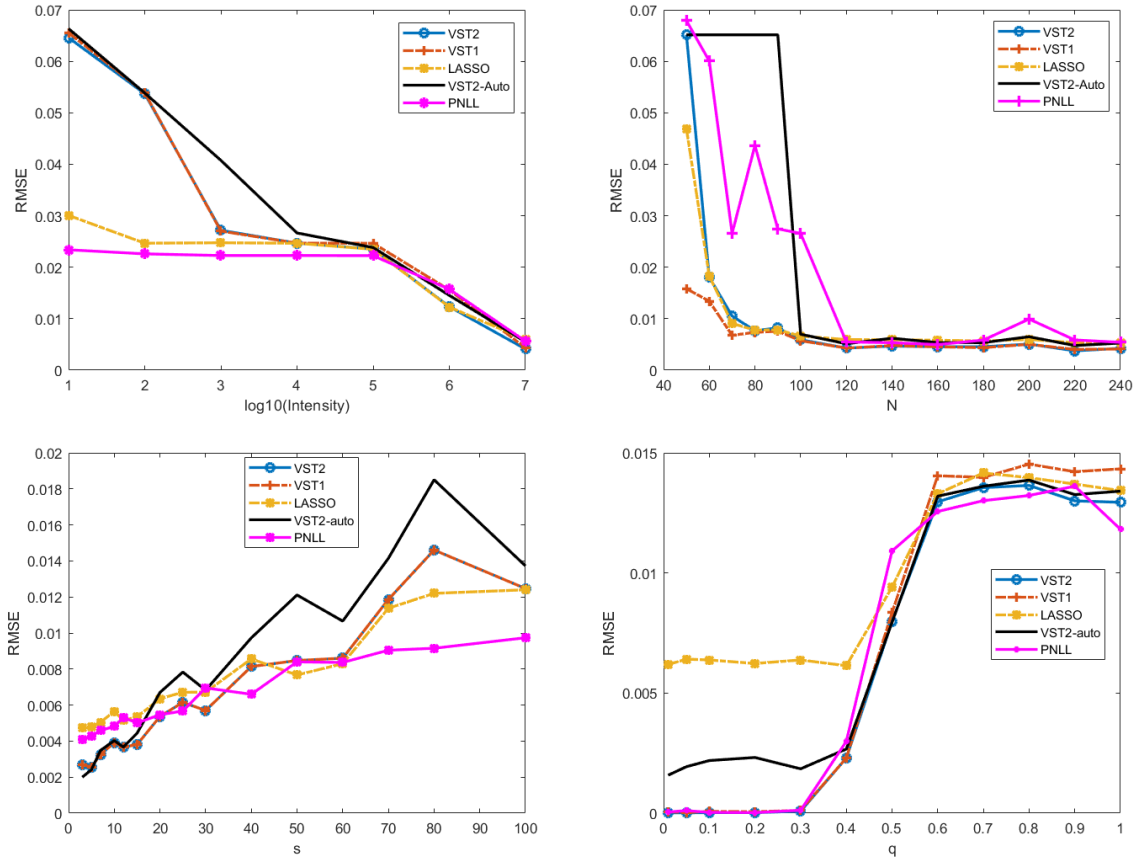

Figure 2: From left to right, top to bottom: RRMSE versus $\log_{10} I$, N, $\|\boldsymbol{\theta}\|_0$ and q for five estimators: VST-1, VST-2, VST2-AUTO, LASSO, PNLL. See Sec. 3.1 for details of experiment parameters.

such as REC/RIP to emerge), we observed that the RRMSE for the various estimators did not decrease much w.r.t. N, as seen in Fig. 2. This again tallies with our bounds from Theorems 3 and 6, which do not depend on N as long as N is large enough for $\widetilde{\mathbf{\Phi}}$ to satisfy REC/RIP. We also observed that for small N, the estimators VST-1, VST-2 outperformed PNLL.

In the **sparsity experiment**, we set $I = 10^7, N = 180$ and varied s in $\{3, 5, 7, 10, 12, 15, 20, 25, 30, 40, 50, 60, 70, 80, 100\}$. We observed a predictably steady increase in RRMSE w.r.t. s for all estimators, as seen in Fig. 2. This is also predicted by our bounds for VST-1 and VST-2, as well as those from [11] and [14] for PNLL and LASSO respectively.

In the **weak sparsity experiment**, we experimented with weakly sparse signals in the 1D-DCT basis, as opposed to strictly sparse signals in all earlier experiments. Signals were created with DCT coefficients sampled from Uniform$(0, 1)$ and renormalized to make them non-negative as well as to ensure $\|\boldsymbol{x}\|_1 = I$ for some chosen I. All DCT coefficients except the DC coefficient were adjusted such that $\sum_{i=1}^{n} |\theta_i|^q = R_q$ (for a fixed value of R_q), and the signal was renormalized. Here $0 < q \leq 1$ is a sparsity factor reflecting how fast the coefficients decay, which is the fastest when q is close to 0 and slowest when q is close to 1. In our experiments, we set set a fixed $I = 10^6, m = 256, N = 180$. We adjusted the coefficients by varying $q \in \{0.01, 0.1, 0.2, ..., 0.9, 1\}$, but keeping I fixed. The RRMSE for VST-1, VST-2, PNLL and LASSO increased with q, as seen in Fig. 2. In general, from Fig. 2, we see that VST-1, VST-2 and LASSO have comparable performance in most regimes. However, as we have highlighted, VST-2 has many theoretical advantages over LASSO due to (1) the pivotal nature of VST-2,

(2) no lower bound restrictions on I for the VST-2 bounds to hold, and (3) higher probability for a similar performance bound in case of VST-2. In our experiments, the performance of our estimators was broadly comparable to that of other estimators such as LASSO and PNLL.

3.2. Image Reconstruction Experiments

Next, we show image reconstruction results given compressive measurements corrupted with Poisson noise. The forward model for the compressive measurements follows the architecture of a variant of the Rice Single Pixel Camera (SPC) [6], which operates on non-overlapping image patches as opposed to the entire image. Such a patch-based architecture has been implemented in hardware in [34, 35, 36]. The compressive measurements for the i^{th} image patch $\boldsymbol{f_i} = \boldsymbol{\Psi\theta_i}$ acquire the form

$$\boldsymbol{y_i} \sim \text{Poisson}(\boldsymbol{\Phi\Psi\theta_i}), \tag{8}$$

where $\boldsymbol{y_i} \in \mathbb{Z}_{\geq 0}^{N\times 1}$, $\boldsymbol{f_i} \in \mathbb{R}_{\geq 0}^{m\times 1}$, $\boldsymbol{\Psi}$ is an $m \times m$ orthonormal basis and $\boldsymbol{\theta_i}$ is the vector of m coefficients of the patch $\boldsymbol{f_i}$ given the basis $\boldsymbol{\Psi}$. The sensing matrix $\boldsymbol{\Phi}$ still follows the model from Eqn. 3, but it operates patch-wise. Image reconstruction using various estimators like VST-1, VST-2, VST2-AUTO, PNLL and LASSO now proceeds independently on image patches. For our experiments, we used a 2D-DCT as the sparsifying basis $\boldsymbol{\Psi}$ with the ℓ_1 penalty on the 2D-DCT coefficients. We used a patch-size of 8×8 with $N = 45$ compressive measurements per patch (70%), and tested all algorithms on images of size 256×256. Each image was scaled to different intensity levels $I = \|\boldsymbol{f}\|_1$ where $I \in \{10^9, 10^8, 5\times 10^7, 10^7, 5\times 10^6, 10^6\}$. The average photon flux per compressive measurement (i.e. the expected value of each

compressive measurement) is given as $\frac{I}{N} \times \frac{\#\text{pixels per patch}}{\#\text{image pixels}} \times 0.5$. The factor of 0.5 arises because the matrix $\mathbf{\Phi}$ in our model consisted of equal number of zeros and ones in expectation. For our specific settings of image size, patch-size and N, the average photon flux values for the aforementioned values of I are $\sim \{10^4, 10^3, 500, 100, 54, 11\}$ respectively, respectively corresponding to noise-to-signal ratios of $\{0.96\%, 3\%, 4\%, 9.6\%, 13.5\%, 30\%\}$. The regularization parameters ρ for VST2-AUTO was selected using the **auto-tuning procedure** based on the 95 percentile of the absolute gradient of the data fidelity function, as explained in Sec. 3.1. Note that auto-tuning is not possible for estimators such as LASSO, VST-1 or PNLL as their optimal regularization parameters depend upon signal properties such as I. The regularization parameter ρ for LASSO, PNLL, VST-1 and VST-2 were obtained using **cross-validation** (CV), a popular procedure for choice of regularization parameter selection in compressed sensing [37, 38]. CV was implemented as follows. The N compressive measurements in $\boldsymbol{y}$ were divided into two disjoint sets: a reconstruction set $\mathcal{M}_{\mathcal{R}}$ and a validation set $\mathcal{M}_{\mathcal{V}}$. For each of the different values of ρ from a candidate set $\mathcal{C}_\rho$, the estimators were run independently. Denoting the estimate of patch $\boldsymbol{g}$ using regularization parameter $\rho \in \mathcal{C}_\rho$ as $\hat{\boldsymbol{g}}_\rho$, we evaluated the corresponding validation error as $VE(\rho) \triangleq \sum_{l \in \mathcal{M}_{\mathcal{V}}} L(y_l, \mathbf{\Phi}^{\boldsymbol{l}} \hat{\boldsymbol{g}}_\rho)$ where L stands for the data-fidelity term used for the particular estimator. The value of ρ which yielded the least value of $VE(\rho)$ is then chosen. Following this, a final estimate of $\boldsymbol{g}$ is computed using all N measurements using this chosen the regularization parameter value. In our experiments, the set $\mathcal{C}_\rho$ was chosen to be $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.5, 1, 1.5, 25, 8, 10\}$ whereas the auto-tuned parameter for VST2-AUTO turned out to be around 0.2448.

CV has nice theoretical properties for CS with Gaussian [37] as well as Poisson noise [38]. However, the CV procedure needs to be executed separately for each patch and for each value from $\mathcal{C}_\rho$, making it computationally expensive. On the other hand, VST2-AUTO does not require any such external procedure, leading to two advantages:

1. VST-AUTO is effectively more computationally efficient (by a factor of nearly 10-fold in our experiments compared to other estimators).
2. It also more accurate because the values in $\mathcal{C}_\rho$ can have only a limited resolution for cross-validation purposes. Moreover, in an application such as patch-wise compressive reconstruction, only a small number of measurements will be realistically available for cross-validation, which limits its accuracy [38]. For example, here for each 8×8 patch, we have $N = 45$ measurements, which would allow for a validation set with only a handful of measurements.

This reflects in superior RRMSE performance (defined as $\|\boldsymbol{f} - \hat{\boldsymbol{f}}\|_2 / \|\boldsymbol{f}\|_2$) for VST2-AUTO compared to all other estimators, as can be seen in Table 3.2. Visual reconstruction results for VST2-AUTO across different values of I are presented in Fig. 3. It is clear that superior results are obtained for larger intensity values. The results do show some blocking artifacts at lower intensities because the reconstruction is being performed separately on each non-overlapping patch. We did not perform any deblocking procedure as detailed in [34] nor did we use more sophisticated signal priors apart from the ℓ_1 sparsity, as it is not central to the main aim of this paper.

Image	I	VST1	VST2	VST2-AUTO	LASSO	PNLL
Peppers	10^9	0.0994	0.0859	0.0842	0.0838	0.11
Peppers	10^8	0.1418	0.1649	0.1335	0.1643	0.178
Peppers	5×10^7	0.172	0.2040	0.1516	0.2105	0.22
Peppers	10^7	0.2851	0.3546	0.1923	0.389	0.378
Peppers	5×10^6	0.3712	0.4678	0.2118	0.4933	0.48
Peppers	10^6	0.6742	0.7441	0.2860	0.752	0.68
Barbara	10^9	0.1	0.093	0.0872	0.09	0.085
Barbara	10^8	0.139	0.1631	0.1268	0.1633	0.176
Barbara	5×10^7	0.1662	0.202	0.14	0.213	0.22
Barbara	10^7	0.2678	0.3505	0.1668	0.3878	0.4
Barbara	5×10^6	0.36	0.4739	0.1825	0.5084	0.477
Barbara	10^6	0.7359	0.7916	0.2624	0.7817	0.83

Table 2: RRMSE values for image reconstruction with estimators VST-1, VST-2, VST2-AUTO, LASSO and PNLL for different intensity values, with $N = 45$ measurements per 8×8 patch. Refer to Sec. 3.2 for more details.

Figure 3: Left to right, top to bottom in each group: Original image, followed by image reconstruction results using VST2-AUTO for 8×8 image patches with $N = 45$ measurements per patch, for six different image intensity levels $I \in \{10^9, 10^8, 5 \times 10^7, 10^7, 5 \times 10^6, 10^6\}$. Refer to Sec. 3.2 for more details, and to Table 3.2 for RRMSE values.

3.3. Image Denoising Experiments

The principal goal of this paper is theoretical bounds for compressive reconstruction. But we cursorily explore the application of VST2-AUTO for image denoising. Given a noisy image $Y \sim \text{Poisson}(F)$, the aim is to estimate the underlying clean image F from Y. For this, we performed the denoising on overlapping 8×8 patches in sliding window fashion (considering $\boldsymbol{y_i} \sim \text{Poisson}(\boldsymbol{f_i})$ for the i^{th} noisy patch $\boldsymbol{y_i}$), and averaging of multiple hypotheses at any pixel. We used ℓ_1 sparsity of 2D-DCT coefficients. The denoising experiments were run on noisy versions of the 256×256 Barbara image for different values of $I = \|F\|_1$ in $\{10^5, 10^6, 10^7, 10^8\}$. This corresponds to an average per-pixel photon flux of $\{1526, 152.6, 15.26, 1.526\}$ respectively, which implies a noise to signal ratio (NSR) of $\{0.025, 0.081, 0.256, 0.81\}$ respectively. The visual results in Fig. 4 show proof of concept that this estimator works well for denoising. We consider detailed application of VST2-AUTO to image restoration with comparisons to other Poisson denoisers [39, 40, 41] to be out of scope of the present work, and more suitable for a separate investigation.

4. Conclusion

We have presented and theoretically analyzed two unconstrained, penalized estimators based on the square-root transform for Poisson random variables, in the context of compressive inversion of sparse signals. In Sec. 2.4, we have presented an extensive comparison to estimators from the literature in terms of the tightness of the bounds and the key assumptions made. As compared to the popular LASSO estimator, the presented estimators (espe-

Figure 4: Topmost row: original Barbara image; Second row: noisy images generated from underlying clean images with intensity $10^5, 10^6, 10^7, 10^8$ respectively; Third row: denoised images corresponding to their noisy versions from the row above (RMSE values $0.4, 0.128, 0.0597, 0.0264$ respectively). See Sec. 3.3 for more details.

cially VST-2) require less stringent lower bounds that the underlying signal intensity must meet. Moreover VST-2 is pivotal w.r.t. signal intensity unlike the LASSO and also allows for higher probability for similar performance bounds. Similarly, VST-2 also has advantages over PNLL which includes its pivotal property and stability for zero-valued measurements (unlike the PNLL). In this work, we have also shown tighter performance bounds as compared to our previous work on VSTs for Poisson CS [15].

There exist many avenues for future work: (1) deriving lower bounds for the estimator, (2) extending the estimator to handle Poisson-Gaussian noise or developing and analyzing variance stabilization transforms for other noise models such as correlated noise [42, 43], (3) extending the theoretical results to improve the bounds for blind CS in conjunction with variance stabilization transforms presented in [44], and (4) exploring applications of the proposed estimators in image restoration and tomographic reconstruction.

Acknowledgement: We acknowledge discussions with Prof. Venkatesh Saligrama (Boston University) which were useful for deriving Lemma 1. We also thank the anonymous reviewers for many insightful and useful comments.

References

[1] E. Candes, M. Wakin, An introduction to compressive sampling, IEEE Signal Processing Magazine.

[2] T. Hastie, R. Tibshirani, M. Wainwright, Statistical Learning with Sparsity: The LASSO and Generalizations, CRC Press, 2015.

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

[4] M. A. Davenport, M. F. Duarte, Y. C. Eldar, G. Kutyniok, Introduction to Compressed Sensing, 2012.

[5] H. J. Trussell, R. Zhang, The dominance of Poisson noise in color digital cameras, in: ICIP, IEEE, 2012, pp. 329–332.

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

[7] D. Lingenfelter, J. Fessler, Z. He, Sparsity regularization for image reconstruction with Poisson data, in: Proc. SPIE, Vol. 7246, 2009.

[8] P. Gopal, S. Chandran, I. D. Svalbe, A. Rajwade, Low radiation tomographic reconstruction with and without template information, Signal Processing 175.

[9] T. Pun, J. Ellis, Application of simulated poisson statistical processes to STEM imaging, Signal Processing 8 (1985) 51–62.

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

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

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

[13] X. Jiang, P. Reynaud-Bouret, V. Rivoirard, L. Sansonnet, R. Willett, A data-dependent weighted LASSO under Poisson noise, IEEE Trans. Information Theory 65 (3).

[14] Y. Li, G. Raskutti, Minimax optimal convex methods for Poisson inverse problems under lq-ball sparsity, IEEE Trans. Information Theory 64 (8).

[15] P. Bohra, D. Garg, K. S. Gurumoorthy, A. Rajwade, Variance-stabilization-based compressive inversion under Poisson or Poisson–Gaussian noise with analytical bounds, Inverse Problems 35 (10).

[16] S. Patil, K. S. Gurumoorthy, A. Rajwade, Using an information theoretic metric for compressive recovery under Poisson noise, Signal Processing 162 (2019) 35–53.

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

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

[19] G. Casella, R. Berger, Statistical Inference, 2001.

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

[21] D. Garg, A. Rajwade, Performance bounds for Poisson compressed sensing using variance-stabilization transforms, in: ICASSP, 2017, pp. 1–4.

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

[23] J. H. Curtiss, On transformations used in the analysis of variance, Ann. Math. Statist. 14 (2) (1943) 107–122. doi:10.1214/aoms/1177731452. URL https://doi.org/10.1214/aoms/1177731452

[24] Supplemental material for 'two penalized estimators based on variance stabilization transforms for sparse compressive recovery with poisson measurement noise', Uploaded on Journal Portal.

[25] S. Negahban, P. Ravikumar, M. Wainwright, B. Yu, A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers, Statistical Science 27 (4).

[26] P. Bickel, Y. Ritov, A. Tsybakov, Simultaneous analysis of lasso and dantzig selector, Annals of Statistics 37 (4).

[27] G. Raskutti, M. Wainwright, B. Yu, Restricted eigenvalue properties for correlated gaussian designs, J. Mach. Learn. Res. 11 (2010) 2241–2259.

[28] J. Jia, F. Xie, L. Xu, Sparse poisson regression with penalized weighted score function, Electron. J. Statist. 13 (2) (2019) 2898–2920.

[29] A. Belloni, V. Chernozhukov, L. Wang, Square-root lasso: pivotal recovery of sparse signals via conic programming, Biometrika 98 (4) (2011) 791–806.

[30] F. Natterer, The Mathematics of Computerized Tomography, SIAM, 1986.

[31] S. Boucheron, G. Lugosi, P. Massart, Concentration inequalities: A nonasymptotic theory of independence, Oxford University Press, 2013.

[32] M. Wainwright, Basic tail and concentration bounds, `https://www.stat.berkeley.edu/~mjwain/stat210b/Chap2_TailBounds_Jan22_2015.pdf`.

[33] J. H. Pollard, A Handbook of Numerical and Statistical Techniques, Cambridge University Press, 1979.

[34] R. Kerviche, N. Zhu, A. Ashok, Information-optimal scalable compressive imaging system, in: Classical Optics 2014, Optical Society of America, 2014.

[35] K. Kulkarni, S. Lohit, P. Turaga, R. Kerviche, A. Ashok, Reconnet: Non-iterative reconstruction of images from compressively sensed measurements, in: CVPR, 2016, pp. 449–458.

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

[37] J. Zhang, L. Chen, P. T. Boufounos, Y. Gu, On the theoretical analysis of cross validation in compressive sensing, in: ICASSP, 2014, pp. 3370–3374.

[38] R. Sudarsanan, A. Rajwade, Analyzing cross-validation in compressed sensing with poisson noise, Signal Processing 182.

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

[40] J. Salmon, Z. Harmany, C. Deledalle, R. Willett, Poisson noise reduction with non-local PCA, Journal of Mathematical Imaging Vision 48 (2014) 279–294.

[41] S. Patil, A. Rajwade, Poisson noise removal for image demosaicing, in: BMVC, 2016.

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

[43] B. G. Jeong, B. C. Kim, Y. H. Moon, I. K. Eom, Simplified noise model parameter estimation for signal-dependent noise, Signal Processing 96 (2014) 266 – 273.

[44] R. Das, A. Rajwade, Nonlinear blind compressed sensing under signal dependent noise, in: International Conference on Image Processing, 2019.