SIAM J. Imaging Sciences
Vol. xx, pp. x

Signal Recovery and System Calibration from Multiple Compressive Poisson Measurements

Liming Wang[*][†], Jiaji Huang[*][†], Xin Yuan[*][†], Kalyani Krishnamurthy[†], Joel Greenberg[†], Volkan Cevher[‡], Miguel R.D. Rodrigues[§], David Brady[†], Robert Calderbank[†], and Lawrence Carin[†][¶]

Abstract. The measurement matrix employed in compressive sensing typically cannot be known precisely *a priori*, and must be estimated via calibration. One may take *multiple* compressive measurements, from which the measurement matrix and underlying signals may be estimated *jointly*. This is of interest as well when the measurement matrix may change as a function of the details of what is measured. This problem has been considered recently for Gaussian measurement noise, and here we develop this idea with application to Poisson systems. A collaborative maximum likelihood algorithm and alternating proximal gradient algorithm are proposed, and associated theoretical performance guarantees are established based on newly derived concentration-of-measure results. A Bayesian model is then introduced, to improve flexibility and generality. Connections between the maximum likelihood methods and the Bayesian model are developed, and example results are presented for a real compressive X-ray imaging system.

Key words. Compressive sensing, Poisson compressive sensing, system calibration, concentration-of-measure, Bayesian compressive sensing, X-ray imaging.

AMS subject classifications.

1. Introduction and Related Work.

There is increasing interest in realizing the potential of compressive sensing (CS) in actual physical systems, with the goal of efficiently (compressively) measuring the information characteristic of an entity under test. Examples include a single-pixel camera [6], hyperspectral imaging [14], and compressive video [12, 19, 34]. The Gaussian measurement model is assumed in each of these examples, and is widely employed in existing theory and applications.

A Gaussian measurement model is not appropriate for many important applications, including X-ray [9, 20] and chemical imaging [29, 30, 31]. The observed data in these applications are characterized by counts, typically under Poisson statistics. The properties of the Poisson measurement model have been studied from various perspectives. Algorithms for recovering the (sparse) Poisson rate function (*i.e.,* the associated parameter of the Poisson distribution) have been studied in [10], and performance bounds for inversion algorithms have been developed in [22, 24]. However, these algorithms assume perfect knowledge of the sensing matrix. In real physical systems that motivate this paper, it is usually impossible to build a device that perfectly matches the desired sensing matrix, and a calibration step is required to learn the

[*] Equal contribution.

[†] Department of Electrical and Computer Engineering Duke University, Durham, NC 27708, USA.

[‡] Laboratory for Information and Inference Systems (LIONS), Ecole Polytechnique Federale de Lausanne (EPFL), CH1015 - Lausanne, Switzerland.

[§] Department of Electronic and Electrical Engineering, University College London, London WC1E 7JE, U.K.

[¶] E-mails: liming.w@duke.edu, jiaji.huang@duke.edu, eiex.yuan@connect.polyu.hk, kk63@duke.edu, joel.greenberg@duke.edu, volkan.cevher@epfl.ch, dbrady@duke.edu, m.rodrigues@ee.ucl.ac.uk, robert.calderbank@duke.edu, lcarin@duke.edu.

sensing matrix. There are other situations for which the target to be sensed may perturb the sensing matrix [2], and hence perturbations to this matrix must be inferred when performing CS inversion. The problem of fluctuating sensing matrices has been studied for the Gaussian measurement model in [11, 33, 37].

To the authors' knowledge, almost all previous work focuses on establishing various theoretical properties of both Gaussian and Poisson measurement models with *a single* compressive measurement. CS with multiple measurements has only been addressed in [8, 21], and there for the Gaussian measurement model. Furthermore, randomly constituted sensing matrices employed in the aforementioned work violate the nonnegativity constraint of the Poisson measurement model. Our focus is on the use of *multiple* measurements to estimate the sensing matrix and recover the signal, which is often practical as a calibration step or in multi-view compressive measurements. Further, we focus on the Poisson measurement model, for which there are no previous results on estimating the sensing matrix. The new theory developed here is motivated by the practical problem of compressive X-ray *scatter* (*not* transmission) measurements, and therefore we begin with a brief introduction to this motivating system.

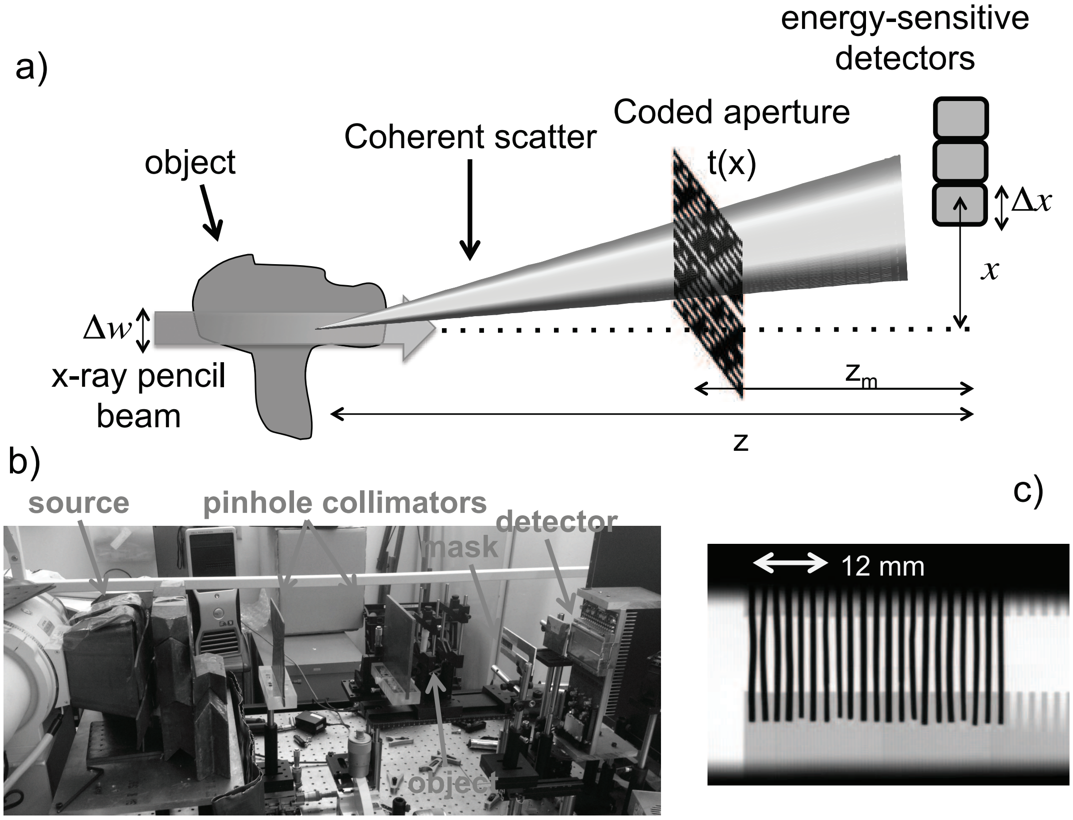


Figure 1: Illustration of the motivating compressive X-ray system. a) Schematic, b)Photograph of the system, and c) X-ray transmission image of the coded aperture [9].

1.1. Hardware Review. The basic structure of the coded aperture coherent scatter spectral imaging (CACSSI) [9] system considered in this paper is illustrated in Figure 1. The object is illuminated with an X-ray pencil beam (transmitted beam along a single linear direction, or "pencil"), and a *coded aperture* is placed in front of a linear array of *energy-sensitive* detectors; the system measures coherently scattered (energy-dependent) X-rays. The angle with which a scattered X-ray impinges on a particular detector element is a function of the detector distance from the scattering source. Multiple angle (depth) dependent X-rays are measured by each detector element, and each ray is distinguished by interacting with (being encoded by) a different portion of the coded aperture. Hence, the coded aperture manifests angle- and hence depth-dependent coding, and we model this with the sensing matrix. The measurements are compressive in that one may recover a depth-energy image of the entity under test (energy-position two-dimensional image along the length of the pencil beam), based on a linear array of energy-dependent measurements (one-dimensional measurements). The overall entity under test is characterized by sequentially scanning the position of the pencil beam across the full volume of the object, yielding ultimately an energy-dependent characterization over the volume of the object under test. However, one typically cannot manufacture or situate the coded aperture precisely, and these imperfections manifest mismatches between the designed and actual measurement matrix. The detailed measurement model is introduced in Section 2.

1.2. Contributions. The specified contributions of this paper are:

a) We propose a regularized maximum-likelihood (ML) algorithm to jointly estimate the signal and the imperfectly-known sensing matrix.
b) A new concentration-of-measure result is derived for a Rademacher-type random sensing matrix, suitable for both Gaussian and Poisson multiple-measurement models. Several theoretical applications based on the derived concentration-of-measure result are presented.
c) Performance bounds are derived for the regularized ML method, assuming Rademacher-type randomly constituted sensing matrices.
d) A new Bayesian method is proposed to improve the inversion performance, eliminating the need for cross-validation, and dropping some assumptions in the ML algorithm.
e) The proposed algorithms and theoretical results are demonstrated on both synthetic data and real data captured by the X-ray imaging system summarized in Figure 1.

The remainder of the paper is organized as follows. Section 2 introduces the Poisson measurement model. In Section 3 the reconstruction algorithm is developed, considering a regularized maximum likelihood estimator. We derive a new concentration-of-measure result, under the assumption of multiple measurements, and several theoretical applications are developed. Using these results, performance bounds are presented. In Section 4 we develop the Bayesian inversion algorithm, to improve inversion performance and robustness. Experimental results are summarized in Section 5, and Section 6 concludes the paper.

1.3. Notation. Bold upper and lower case letters are used to denote matrices and vectors, respectively, *e.g.,* $\mathbf{\Phi}$ is a matrix and $\mathbf{f}$ is a vector. $\mathbf{I}_m$ denotes the $m \times m$ identity matrix, and $\mathbf{1}_m$ denotes an m-dimensional vector with all entries being 1. An element in a matrix/vector is represented by non-bold lower case letter with its indices as the subscript, *e.g.,* $\Phi_{i,j}$ is the element at the i-th row and j-th column of matrix $\mathbf{\Phi}$, and f_i is the i-th element in vector $\mathbf{f}$. A variable with no superscript, *e.g.,* $\mathbf{f}$, $\mathbf{\Phi}$, is used for general purposes, such as describing the model/system or used as a dummy variable in an optimization problem. The symbol $\mathbf{f}^*$ reflects the true value of variable $\mathbf{f}$, and $\hat{\mathbf{f}}$ denotes an estimate of $\mathbf{f}^*$. $|\cdot|$ denotes the cardinality of the argument set. $\|\cdot\|_p$ denotes the l_p norm of the argument vector. $\|\cdot\|_2$ may respectively denote l_2 norm or operator norm (spectral norm) for vector or matrix, and its specific meaning should be clear from the argument variable. $\|\cdot\|_F$ denotes the Frobenius norm for argument matrix. The big-theta Θ notation is used to denote two-sided boundedness for functions of real numbers. For example, $f(K) \sim \Theta(g(K))$ means $c_1 g(K) \leq f(K) \leq c_2 g(K)$ with constants $c_1, c_2 > 0$ for all K large enough.

2. Model and Problem Statement.

Assume K compressive measurements are performed, with the k-th measurement $\mathbf{y}_k \in \mathbb{Z}_+^m$ a Poisson-distributed vector of counts:

$$\mathbf{y}_k \sim \mathrm{Pois}(\mathbf{\Phi}\mathbf{f}_k + \mathbf{u}) \overset{\text{def}}{=} \prod_{i=1}^{m} \mathrm{Pois}((\mathbf{\Phi}\mathbf{f}_k + \mathbf{u})_i), \quad \forall k = 1, \dots, K, \tag{2.1}$$

where $\mathbb{Z}_+$ denotes the collection of nonnegative integers, $\mathbf{f}_k \in \mathbb{R}_+^n$ is the k-th input signal to be estimated (Poisson rate vector) and $\mathbf{\Phi} \in \mathbb{R}_+^{m\times n}$, with $m \ll n$, is the sensing matrix (imperfectly known *a priori*). The first Pois$(\cdot)$ in (2.1) has a vector argument for the rate, and corresponds to a Poisson distribution implemented independently on each component of the rate vector; the second Pois$(\cdot)$ in (2.1) denotes the common scalar Poisson distribution with a rate parameter. The $\mathbf{u} \in \mathbb{R}_+^m$ accounts for background noise, which may exist even when $\mathbf{f}_k = \mathbf{0}$, and $\mathbf{u}$ is often termed as the "dark-current". We further model

$$\mathbf{\Phi} = \mathbf{\Phi}_0 + \mathbf{\Phi}_E, \tag{2.2}$$

where $\mathbf{\Phi}_0$ is known *a priori* as the designed sensing matrix, and $\mathbf{\Phi}_E$ is the *unknown* perturbation. We wish to *jointly* recover $\{\mathbf{f}_k\}_{k=1}^K$, $\mathbf{\Phi}_E$ and $\mathbf{u}$ from the multiple measurements $\{\mathbf{y}_k\}_{k=1}^K$.

It is instructive to place the model in the context of the system discussed in Section 1.1. For a material with effective lattice spacing d, excited by X-rays at energy E, the angle θ at which the fields are scattered satisfies $\frac{1}{2d} = \frac{E}{hc}\sin(\theta/2)$, where h is Plank's constant and c is the speed of light in vacuum. If the X-ray beam is characterized by a range of energies E over a finite bandwidth, then each energy within that bandwidth scatters at an angle $\theta(E,d)$ defined by material properties d and energy E. The gray-scale gradient of scattered waves in Figure 1(a) reflects variation in $\theta(E,d)$ as a function of E over the bandwidth, for fixed d; the d is fixed by the properties of the local material, and E represents the energy, which varies across the bandwidth of the source. The reader is referred to [9] for further details.

Each gray scale in Figure 1(a) denotes one energy over the bandwidth, with each energy scattering at a specific angle. In Figure 1(a), note that the gray-scale gradient of energies is emitted from a point along the pencil beam, with the properties of that gray-scale gradient defined by the material properties around that emission point. Each emission point along the pencil beam is characterized in general by a unique gray-scale gradient, with energies scattered at angles defined by the properties of the local material. The signal $\mathbf{f}_k$ may be viewed as a 3D entity, with position along the pencil beam representing one axis (ζ_1), energy level representing the second axis (ζ_2), and angle of scatter representing the third axis (ζ_3); $\mathbf{f}_k$ represents the scatter strength from each position along the beam, as a function of energy and angle of scatter. For any fixed point along ζ_1, the two-dimensional image in the (ζ_2, ζ_3) plane characterizes the strength of scatter, as a function of angle and energy; by integrating across the ζ_3 dimension, we obtain an energy-dependent signature for the material at the selected position in ζ_1, used for material characterization.

The detectors are energy-dependent. The measured signal $\mathbf{y}_k$ may be viewed as two-dimensional, defined by an energy-dependent photon count as a function of detector location. The $\mathbf{f}_k$ and $\mathbf{y}_k$ are "unwrapped" to constitute vectors. Matrix $\mathbf{\Phi}$ characterizes the *linear* process through which $\mathbf{f}_k$ is mapped to on overall energy-dependent Poisson rate as observed at each detector, and it accounts for modulation induced by the coded aperture. The coded aperture modulates the scattered photons in a manner that depends on the position of emission and on the angle of scatter, helping to disambiguate the source of observed photons when seeking to recover $\mathbf{f}_k$ from $\mathbf{y}_k$. Further details on the X-ray system may be found in [9, 20].

One may take multiple "off-line" measurements with all inputs $\mathbf{f}_k$ set to $\mathbf{0}$ (*i.e.*, in the absence of any material in the measurement system), from which one may yield an estimate of $\mathbf{u}$. Hence, for simplicity, we assume that $\mathbf{u}$ is known *a priori* when developing theory, and the theory focuses on the maximum-likelihood-based estimation of $\{\mathbf{f}_k\}$ and $\mathbf{\Phi}_E$. However, for the Bayesian inversion we develop in Section 4, $\mathbf{u}$ is estimated along with $\{\mathbf{f}_k\}_{k=1}^K$ and $\mathbf{\Phi}_E$. This latter flexibility is important, as the dark current is characteristic of spurious scatter within the measurement system, which may change with the item under test.

3. Collaborative Reconstruction via Regularized MLE.

3.1. Collaborative MLE. The model for multiple Poisson measurements (2.1) can be rewritten concisely as

$$\mathbf{y} \sim \mathrm{Pois}(\mathbf{A}\mathbf{f} + \boldsymbol{\lambda}), \tag{3.1}$$

where

$$\mathbf{y} \overset{\text{def}}{=} \left[\mathbf{y}_1^\top, \ldots, \mathbf{y}_K^\top\right]^\top, \quad \mathbf{f} \overset{\text{def}}{=} \left[\mathbf{f}_1^\top, \ldots, \mathbf{f}_K^\top\right]^\top, \quad \boldsymbol{\lambda} \overset{\text{def}}{=} \mathbf{1}_K \otimes \mathbf{u},$$

where $\otimes$ represents the Kronecker (tensor) product, $\mathbf{A}$ is constituted via block-diagonalization, $\mathbf{A} \overset{\text{def}}{=} \mathbf{I}_K \otimes \mathbf{\Phi}$. We define $\mathbf{A}_0 \overset{\text{def}}{=} \mathbf{I}_K \otimes \mathbf{\Phi}_0$ and $\mathbf{A}_E \overset{\text{def}}{=} \mathbf{I}_K \otimes \mathbf{\Phi}_E$, and $\mathbf{A} = \mathbf{A}_0 + \mathbf{A}_E$.

When developing the theory, we make the following assumptions:

A1) The intensity of each signal $\mathbf{f}_k$ is known and fixed, *i.e.*, $\|\mathbf{f}_k\|_1 = I$ for $k = 1, \ldots, K$; a similar assumption was made in [24], and is necessary to make $\{\mathbf{f}_k\}$ and $\mathbf{\Phi}$ identifiable.

A2) $\mathbf{A}\mathbf{f} \succeq cI\mathbf{1}_{Km}$, where constant $c > 0$ and $\succeq$ denotes the entry-wise inequality. This is used to exclude the singular case, where some Poisson rates asymptotically approach zero.

A3) The system perturbation is bounded as $\frac{\|\mathbf{\Phi}_E\|_2}{\|\mathbf{\Phi}_0\|_2} \leq \epsilon_1$, where $\|\cdot\|_2$ denotes the operator norm (spectral norm) of the argument matrix.

A4) The energy of the dark-current is assumed to be bounded as $\|\mathbf{u}\|_1 \leq U$.

We propose to estimate $\mathbf{f}$ and $\mathbf{A}_E$ simultaneously via the following collaborative maximum-likelihood estimator (CMLE):

$$(3.2)\qquad (\hat{\mathbf{f}}, \hat{\mathbf{A}}_E) = \mathop{\arg\min}_{(\mathbf{f},\mathbf{A}_E)\in\Gamma} \left\{ -\log \mathrm{Pois}[\mathbf{y}; (\mathbf{A}_0 + \mathbf{A}_E)\mathbf{f} + \boldsymbol{\lambda}] + 2\,\mathrm{pen}(\mathbf{f}) + 2\frac{\|\mathbf{A}_E\|_F}{K} \right\},$$

where Γ is a collection of all candidate estimators and should satisfy the following constraint

$$(3.3)\qquad \Gamma \subset T \overset{\text{def}}{=} \left\{ (\mathbf{f}, \mathbf{A}_E) \;\middle|\; \begin{array}{l} \mathbf{f} \succeq 0,\ \|\mathbf{f}_k\|_1 = I, \forall k = 1, \ldots, K; \\ \frac{\|\boldsymbol{\Phi}_E\|_2}{\|\boldsymbol{\Phi}_0\|_2} \leq \epsilon_1; \\ \mathbf{A}_0 + \mathbf{A}_E \succeq \mathbf{0},\ (\mathbf{A}_0 + \mathbf{A}_E)\mathbf{f} \succeq cI\mathbf{1}_{Km}. \end{array} \right\}.$$

We further assume the technical condition that Γ is a countable or finite set. In order to remedy the mismatch between this technical assumption and the fact that T is a continuous domain, we require Γ to be selected as a quantized version of T. Note that we may always utilize uniform quantization on T to obtain such a Γ. In particular, we define the quantization step $\mathcal{QS}(\Gamma)$ as

$$(3.4)\qquad \mathcal{QS}(\Gamma) = \inf_{\substack{(\mathbf{f}',\mathbf{A}'_E),(\mathbf{f}'',\mathbf{A}''_E)\in\Gamma \\ (\mathbf{f}',\mathbf{A}'_E)\neq(\mathbf{f}'',\mathbf{A}''_E)}} \left\{ \min(\|\mathbf{f}'_i - \mathbf{f}''_i\|_1, \ldots, \|\mathbf{f}'_i - \mathbf{f}''_i\|_K, \|\boldsymbol{\Phi}'_E - \boldsymbol{\Phi}''_E\|_F) \right\}.$$

Therefore, $\mathcal{QS}(\Gamma)$ characterizes the minimal quantization level occurring within Γ. Throughout the paper, we assume that $\mathcal{QS}(\Gamma) \geq \frac{d}{\sqrt[4]{K}}$, where d is a constant independent of K. In other words, the discrete set Γ is assumed to be asymptotically dense with the increase of the number of measurements K.

The penalty term $\mathrm{pen}(\mathbf{f})$ is required to satisfy the Kraft inequality $\sum_{\mathbf{f}\in\Gamma_1} e^{-\mathrm{pen}(\mathbf{f})} \leq 1$, where $\Gamma_1 = \{\mathbf{f} | (\mathbf{f}, \mathbf{A}_E) \in \Gamma\}$, and $\|\cdot\|_F$ denotes the Frobenius norm. We also denote $\Gamma_2 = \{\mathbf{A}_E | (\mathbf{f}, \mathbf{A}_E) \in \Gamma\}$. The $\mathrm{pen}(\cdot)$ acts as a logarithmic prior on the signal and can be designated as many popular penalty functions, when a proper scaling is applied in order to satisfy the Kraft inequality. Typical choices for the $\mathrm{pen}(\cdot)$ include the l_1 norm, of interest for sparse signals [4], and the total-variation norm for smooth signals [26]. In fact, the Kraft-compliant penalty is related to the prefix codes for estimators, and more concrete examples of this penalty functions are presented in [32]. As elaborated in the proof of Theorem 3.1, we utilize a main result from [15], which requires the countability assumption as reflected in Γ.

In practice, the signal energy level I and the perturbation bound ϵ_1 may not be known precisely. In order to increase the flexibility of regularizers in (3.2), one may relax the CMLE in (3.2) to the following form

$$(3.5)\qquad (\hat{\mathbf{f}}, \hat{\mathbf{A}}_E) = \mathop{\arg\min}_{(\mathbf{f},\mathbf{A}_E)\in\Gamma} \left\{ -\log \mathrm{Pois}[\mathbf{y}; (\mathbf{A}_0 + \mathbf{A}_E)\mathbf{f} + \boldsymbol{\lambda}] + \tau_1\,\mathrm{pen}(\mathbf{f}) + \tau_2\|\mathbf{A}_E\|_F \right\},$$

where $\tau_1 > 0$ and $\tau_2 > 0$ are preset constants. We refer to the above estimator as the relaxed-CMLE.

The CMLE and the above associated assumptions make theoretical analysis tractable. As we will see subsequently, some of these assumptions can be relaxed further within our practical Bayesian approach (developed in Section 4). For example, rather than the strong assumption

$\|\mathbf{f}_k\|_1 = I$ for $k = 1, \ldots, K$, that may not always be true in practice due to variation of the scatter strength of the K targets, we assume that the prior *distributions* placed on $\{\mathbf{f}_k\}$ are all the same. Further, rather than assuming $\mathbf{u}$ is known, we assume the mean of $\mathbf{u}$ is known.

3.2. Performance Analysis for the CMLE. The work presented here is inspired by previous research that assumed Gaussian sensor noise [4] and a randomized sensing matrix. We consider Poisson noise and also consider randomized design of the sensing matrix. However, the nonnegativity constraint on $\mathbf{A}$ and its block-diagonal structure prevent direct application of many results for the Gaussian model [4, 21, 6]. We develop a new concentration-of-measure inequality for a randomized sensing matrix $\mathbf{A}$ satisfying the positivity and block-diagonal constraints. Several theoretical applications of the concentration inequality are presented as well, and in particular we establish performance bounds for the proposed CMLE estimator.

3.2.1. Construction of the Sensing Matrix. To constitute the matrix $\mathbf{\Phi}_0$, we first generate $\mathbf{Z} \in \{1, -1\}^{m \times n}$, with each entry drawn i.i.d. from the Rademacher distribution (*i.e.*, random variables take values 1 or -1 with equal probability). Let $\mathbf{\Psi} \overset{\text{def}}{=} \frac{\mathbf{Z}}{\sqrt{m}}$ and $\mathbf{\Phi}_0 \overset{\text{def}}{=} \mathbf{\Psi} + \frac{1}{\sqrt{m}}\mathbf{1}_{m \times n}$; the sensing matrix is $\mathbf{A}_0 = \mathbf{I}_K \otimes \mathbf{\Phi}_0$ and, for use below, we define $\tilde{\mathbf{A}}_0 = \mathbf{I}_K \otimes \mathbf{\Psi}$. Note that $\mathbf{\Phi}_0$ is a matrix with entries being either 0 or $\frac{2}{\sqrt{m}}$. In other words, the sensing matrix $\mathbf{A}_0$ consists of a scaled-Rademacher matrix $\mathbf{\Psi}$ and a DC offset $\frac{1}{\sqrt{m}}\mathbf{1}_{m \times n}$ keeping the sensing matrix nonnegative.

3.2.2. Concentration-of-measure Inequalities. Concentration of measure is a phenomenon describing the tendency of certain functions of a high-dimensional random process to concentrate sharply around their means [16]. Our first result is a concentration-of-measure inequality for the block-diagonal Rademacher-distributed matrix $\tilde{\mathbf{A}}_0$, which serves as a key ingredient in the proof of the performance bounds.

Theorem 3.1. *Let $\tilde{\mathbf{A}}_0$ be generated as described in Section 3.2.1, and let $\Delta \subset \{\mathbf{f} | \mathbf{f} \succeq 0\}$ be a countable or finite set. Then the matrix $\tilde{\mathbf{A}}_0$ satisfies the following concentration-of-measure inequality*

$$\mathbb{P}\left(\left|\|\tilde{\mathbf{A}}_0\mathbf{f}\|_2^2 - \|\mathbf{f}\|_2^2\right| \geq \epsilon\|\mathbf{f}\|_2^2\right) \leq e \cdot \exp\left(-\frac{c_1\epsilon^2\|\mathbf{f}\|_2^4}{mn^2}\right) \; \forall \mathbf{f} \in \Delta, \; \epsilon \in (0, 1), \tag{3.6}$$

where $c_1 > 0$ is a constant.

Proof. The proof is presented in Appendix A. ■

In contrast to many previous concentration-of-measure results for matrices populated with i.i.d. sub-Gaussian entries [23], the decay rate indicated by Theorem 3.1 depends on the signal being measured. In particular, for the estimator candidates set Γ introduced in Section 3.1, we have the following corollary which will be used later in the proof, serving an analogy of the restricted isometry property (RIP) for sparse signals [4].

Corollary 3.2. *Let $\tilde{\mathbf{A}}_0$ be generated as described in Section 3.2.1. We have*

$$(1 - \epsilon)\|\mathbf{f} - \mathbf{g}\|_2^2 \leq \|\tilde{\mathbf{A}}_0(\mathbf{f} - \mathbf{g})\|_2^2 \leq (1 + \epsilon)\|\mathbf{f} - \mathbf{g}\|_2^2, : \; \forall \mathbf{f}, \mathbf{g} \in \Gamma_1, \; \epsilon \in (0, 1) \tag{3.7}$$

with probability at least $1 - e \cdot \exp\left(-\frac{c_1\epsilon^2 K}{mn^4}\right)$, where $c_1 > 0$ is a constant.

Proof. The proof is presented in Appendix B. ■

In previous work, for both Gaussian and Poisson measurement models, a sparseness assumption has been placed on the source signal $\mathbf{f}_k$ [4] and RIP conditions suitable for the sparsity assumption have been utilized accordingly. Our RIP results as in Corollary 3.2 are valid even when the signal is not sparse, but rather a general Kraft relationship $\sum_{\mathbf{f}\in\Gamma_1} e^{-\operatorname{pen}(\mathbf{f})} \leq 1$ is satisfied (a special case of which includes a sparsity constraint). We note that it is possible to derive tighter bounds via a more accurate RIP condition by leveraging further assumptions, such as sparsity of the source signal (or sparsity in a particular basis).

Compressive sensing with multiple measurements has only been addressed in [8, 21], and there for the Gaussian measurement model. The case of a Poisson single-measurement model (without perturbation on the sensing matrix) has been considered in [22, 24]. However, it is worth noting that even though the multiple-measurement case can be formulated concisely as $\mathbf{y} \sim \operatorname{Pois}(\mathbf{A}\mathbf{f} + \boldsymbol{\lambda})$, akin to the single-measurement situation [22, 24], a fundamental difference is that the sensing matrix $\mathbf{A}$ in this case is limited to a *block-diagonal* structure, rather than being arbitrary, as in the single-measurement case. The block-diagonal measurement matrix poses a more challenging (and practical) problem. Hence, the proof techniques from [22, 24] cannot be applied to the multiple-measurements case, and the nonnegativity constraint on the sensing matrix also invalidates adaptation of the results in [8, 21].

Our concentration-of-measure results provide a new strategy, by constituting $\boldsymbol{\Phi}_0$ via the Rademacher distribution. More importantly, as we elaborate on later, such a Rademacher configuration facilitates an easy construction of a nonnegative sensing matrix necessary for Poisson sensing, by simply adding a DC offset. The Gaussian configuration proposed in [8, 21] cannot be easily adapted to guarantee such a nonnegativity constraint. Nevertheless, in addition to their value for Poisson sensing, our concentration-of-measure results also shed light on CS for multiple linear measurements; in the next section, we present some applications of our concentration-of-measure results in that case.

3.2.3. Applications of the concentration-of-measure inequalities. The concentration-of-measure inequality is a powerful characterization for the behavior of a random operator, which possesses a number of implications in various areas [23]. In the previous section, a new concentration-of-measure inequality for a block-diagonal Rademacher-distributed random matrix has been derived, and we now present several theoretical applications of this result. Specifically, we formulate a modified version of the Johnson-Lindenstrauss (JL) Lemma [13] for block-diagonal matrices. Recall the definition of the stable embedding [5]

Definition 3.3. *For $U, V \subset \mathbb{R}^n$, a map $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ is called an ϵ-stable embedding of (U, V) if*

$$(1-\epsilon)\|\mathbf{x}-\mathbf{y}\|_2^2 \leq \|\Phi(\mathbf{x}-\mathbf{y})\|_2^2 \leq (1+\epsilon)\|\mathbf{x}-\mathbf{y}\|_2^2,\ \forall \mathbf{x} \in U,\ \mathbf{y} \in V. \tag{3.8}$$

In other words, a map is a stable embedding of (U, V) if it almost preserves all pairwise distances between U and V. The classical JL Lemma [13] assures the existence of such an ϵ-stable embedding of (U, U) if $m \sim \Theta(\frac{\log|U|}{\epsilon^2})$.

Via Theorem 3.1, a modified version of the JL Lemma can be stated as follows.

Theorem 3.4. *Let $U, V \subset \mathbb{R}^{Kn}$ be two finite sets and $\tilde{\mathbf{A}}_0 \in \mathbb{R}^{Km\times Kn}$ be generated as described in Section 3.2.1. For $0 < \rho < 1$ being fixed, $\tilde{\mathbf{A}}_0$ is an ϵ-stable embedding of (U, V)*

with

$$\epsilon = \sqrt{\frac{mn^2(\log\rho + 1 + \log|U| + \log|V|)}{c_1 \min_{\substack{\mathbf{x}\in U,\mathbf{y}\in V\\ \mathbf{x}\neq\mathbf{y}}} \|\mathbf{x}-\mathbf{y}\|_2^4}}, \tag{3.9}$$

which holds with probability at least $1-\rho$*, where* $c_1 > 0$ *is a constant.*

Proof. The proof is presented in Appendix C. ■

In the above theorem, we note that the performance of the stable-embedding depends on the pairwise distance between two classes U and V and the total number of measurement K is not directly revealed there. Theorem 3.4 is of particular interests when the minimal energy of signal difference between two classes U and V scales to the number of measurements K, *i.e.*, $\min_{\substack{\mathbf{x}\in U,\mathbf{y}\in V\\ \mathbf{x}\neq\mathbf{y}}} \|\mathbf{x}-\mathbf{y}\|_2^2 \sim \Theta(K)$. Whenever such an assumption is valid, it is straightforward to see that embedding performance will keep improving with increase of K.

As a matter of fact, this assumption can be satisfied for many common classes of signals, akin to the notion of "favorable signal class" proposed in [21]. Specifically, for video signals and frequency-sparse signals satisfying additional assumptions, it has been justified numerically and theoretically in [21] that the energy of the signal difference $\|\mathbf{x}_i - \mathbf{y}_i\|_2^2, i = 1,\ldots,K$ tends to be uniformly distributed. In other words, we have $\|\mathbf{x}-\mathbf{y}\|_2^2 = \sum_{i=1}^K \|\mathbf{x}_i - \mathbf{y}_i\|_2^2 \sim \Theta(K)$. We refer the readers to [21] for details and more examples. The following corollary summarizes the previous discussion, which explicitly reveals the effect of K towards the embedding performance.

Corollary 3.5. *Let* $U, V \subset \mathbb{R}^{Kn}$ *be two finite sets and* $\tilde{\mathbf{A}}_0 \in \mathbb{R}^{Km\times Kn}$ *be generated as described in Section 3.2.1. Assume that* $\min_{\substack{\mathbf{x}\in U,\mathbf{y}\in V\\ \mathbf{x}\neq\mathbf{y}}} \|\mathbf{x}-\mathbf{y}\|_2^2 \sim \Theta(K)$*. For* $0 < \rho < 1$ *being fixed,* $\tilde{\mathbf{A}}_0$ *is an* ϵ*-stable embedding of* (U, V) *with*

$$\epsilon = \sqrt{\frac{mn^2(\log\rho + 1 + \log|U| + \log|V|)}{c_2 K^2}}, \tag{3.10}$$

which holds with probability at least $1-\rho$*, where* $c_2 > 0$ *is a constant.*

Indeed, an analogy of the classical JL Lemma can be derived from Corollary 3.5 by considering a finite set U satisfying the assumptions. Setting $m = 1$, Corollary 3.5 essentially claims the existence of an ϵ-stable embedding for (U, U) which maps U to $\mathbb{R}^K$, provided $K \sim \Theta(\frac{\sqrt{\log|U|}}{\epsilon})$. Note that this result can be regarded as a JL-Lemma-type result for sequential embedding of a sequence of signals. Furthermore, rather than being a pure existence result, as in the classical JL Lemma, this result provides a randomized method to realize such a stable embedding.

It is possible to apply them to various applications for signal processing in the compressed domain, where the stable embedding result plays a pivotal role [5]. For example, one may use the above theorems to generalize results in [5], for performing classification directly based on compressive measurements.

3.2.4. Performance Bounds for CMLE.

We consider a performance analysis for the proposed CMLE in (3.2), with estimate $(\hat{\mathbf{A}}_E, \hat{\mathbf{f}})$ for true $(\mathbf{A}_E^*, \mathbf{f}^*)$ evaluated via the risk function

$$R(\{\hat{\mathbf{A}}_E, \hat{\mathbf{f}}\}, \{\mathbf{A}_E^*, \mathbf{f}^*\}) = \frac{1}{K}\left(\frac{\|\hat{\mathbf{f}} - \mathbf{f}^*\|_2}{\|\mathbf{f}^*\|_2} + \frac{\|\hat{\mathbf{A}}_E - \mathbf{A}_E^*\|_F}{\|\mathbf{A}_E^*\|_F}\right). \tag{3.11}$$

Note that the above risk function calculates the average total-estimation error per measurement, where the error terms for both the signal and the sensing matrix have been normalized. The adopted risk function measures the average recovery error per measurement. Although small average error does not necessarily lead to small total recovery error, the employed risk function reflects the general recovery performance of the sensing system and serves as a meaningful evaluation criterion. We assume that $\{\mathbf{A}_E^*, \mathbf{f}^*\}$ are drawn from a distribution whose support satisfies assumptions A1–A3, and we present a performance guarantee for the CMLE which quantifies the *expected* risk bounds with respect to that distribution.

Theorem 3.6. *Assume the perturbation level* $\epsilon_1 < \sqrt{2} - 1$ *and let* $\epsilon' \overset{\text{def}}{=} (1+\epsilon)(1+\epsilon_1)^2 - 1$, *where* ϵ *is an arbitrary constant in* $(0, \frac{2}{(1+\epsilon_1)^2} - 1)$. *With assumptions A1–A3 and the designed sensing matrix* $\mathbf{A}_0$ *generated as described in Section 3.2.1, the expected risk function between the true signal* $\{\mathbf{A}_E^*, \mathbf{f}^*\}$ *and the estimate* $\{\hat{\mathbf{A}}_E, \hat{\mathbf{f}}\}$ *output by the CMLE in (3.2) is bounded by*

$$\begin{aligned} &\mathbb{E}[R(\{\hat{\mathbf{A}}_E, \hat{\mathbf{f}}\}, \{\mathbf{A}_E^*, \mathbf{f}^*\})] \\ &\leq \mathbb{E}\left\{\sqrt{C_1 \min_{\{\mathbf{A}_E,\mathbf{f}\}\in\Gamma}\left\{\left(\frac{2mK}{cI}\left((1+\epsilon')\|\mathbf{f}-\mathbf{f}^*\|_2^2 + K^2I^2\|\mathbf{A}_E - \mathbf{A}_E^*\|_F^2\right) + 2\,\mathrm{pen}(\mathbf{f}) + 2\frac{\|\mathbf{A}_E\|_F}{K}\right)\right\}} \right. \\ &\left. +\sqrt{C_1\left(2P + \min_{\mathbf{A}_E\in\Gamma_2}\left\{\frac{2K^3mI}{c}\|\mathbf{A}_E - \mathbf{A}_E^*\|_F^2 + 2\frac{\|\mathbf{A}_E\|_F}{K}\right\}\right)} + C_2\frac{\epsilon_1}{\|\mathbf{A}_E^*\|_2}\right\}, \end{aligned} \tag{3.12}$$

with probability at least $1 - \frac{m}{2^n} - e \cdot \exp\left(-\frac{c_1\epsilon^2K}{mn^4}\right)$, *where* $P = \max_{\mathbf{f}\in\Gamma_1}\mathrm{pen}(\mathbf{f})$, $C_1 = \left(\frac{n\sqrt{m}(1+\epsilon_1)}{(1-\epsilon')IK\sqrt{K}} + \frac{4Un}{K^2I^2(1-\epsilon')}\right)$, $C_2 = \frac{4\sqrt{mn}}{K}$ *and* $c_1 > 0$ *is a constant. The expectation is taken with respect to an arbitrary joint distribution of* $\{\mathbf{A}_E^*, \mathbf{f}^*\}$ *whose support satisfies assumptions A1–A3.*

Proof. The proof of the theorem is presented in Appendix D. ■

Similarly, we also establish a performance bound for the relaxed-CMLE in (3.5).

Theorem 3.7. *Assume the perturbation level* $\epsilon_1 < \sqrt{2} - 1$ *and let* $\epsilon' \overset{\text{def}}{=} (1+\epsilon)(1+\epsilon_1)^2 - 1$, *where* ϵ *is an arbitrary constant in* $(0, \frac{2}{(1+\epsilon_1)^2} - 1)$. *Fix* $\tau_1 \geq 2$ *and* $\tau_2 > 0$. *With assumptions A1–A4 and the designed sensing matrix* $\mathbf{A}_0$ *generated as described in Section 3.2.1 the expected risk function between the true signal* $\{\mathbf{A}_E^*, \mathbf{f}^*\}$ *and the estimate* $\{\hat{\mathbf{A}}_E, \hat{\mathbf{f}}\}$ *output by the relaxed-*

CMLE in (3.5) is bounded by

(3.13)

$$\mathbb{E}[R(\{\hat{\mathbf{A}}_E, \hat{\mathbf{f}}\}, \{\mathbf{A}_E^*, \mathbf{f}^*\})]$$
$$\leq \mathbb{E}\Bigg\{\sqrt{C_1 \min_{\{\mathbf{A}_E, \mathbf{f}\}\in\Gamma} \Bigg\{\Bigg(\frac{2mK}{cI}\left((1+\epsilon')\|\mathbf{f}-\mathbf{f}^*\|_2^2 + K^2I^2\|\mathbf{A}_E - \mathbf{A}_E^*\|_F^2\right) + \tau_1 \operatorname{pen}(\mathbf{f}) + \tau_2\|\mathbf{A}_E\|_F\Bigg)\Bigg\}}$$
$$+\sqrt{C_1\left(\tau_1 P + \min_{\mathbf{A}_E\in\Gamma_2}\left\{\frac{2K^3mI}{c}\|\mathbf{A}_E - \mathbf{A}_E^*\|_F^2 + \tau_2\|\mathbf{A}_E\|_F\right\}\right)} + C_2\frac{\epsilon_1}{\|\mathbf{A}_E^*\|_2}\Bigg\},$$

with probability at least $1 - \frac{m}{2^n} - e \cdot \exp\left(-\frac{c_1\epsilon^2 K}{mn^4}\right)$, *where* $P = \max_{\mathbf{f}\in\Gamma_1}\operatorname{pen}(\mathbf{f})$, $C_1 = \left(\frac{n\sqrt{m}(1+\epsilon_1)}{(1-\epsilon')IK\sqrt{K}} + \frac{4Un}{K^2I^2(1-\epsilon')}\right)$, $C_2 = \frac{4\sqrt{mn}}{K}$ *and* $c_1 > 0$ *is a constant. The expectation is taken with respect to an arbitrary joint distribution of* $\{\mathbf{A}_E^*, \mathbf{f}^*\}$ *whose support satisfies assumptions A1–A3.*

Proof. The proof of the theorem is presented in Appendix E. ■

Theorems 3.6 and 3.7 provide quantitative performance characterizations for the CMLE algorithm with respect to the number of measurements K. According to both theorems, when the assumptions are valid and m, n, K are fixed, the performance of the proposed CMLE is governed by two factors. The first is the perturbation level ϵ_1, and CMLE has been shown robust in expectation to a perturbation on the sensing matrix. Namely, a perturbation ϵ_1 on the sensing matrix can only result at most in a proportional perturbation $C_2\mathbb{E}[\frac{\epsilon_1}{\|\mathbf{A}_E^*\|_2}]$ on the accuracy of the estimates. The other factor is the two minimization terms in (3.12), which represent the minimal error one could ever achieve over all the candidate estimators in Γ. As we discussed in Section 3.1, the set Γ is designated as a quantized version of the continuous domain and the quantization step $\mathcal{QS}(\Gamma)$ asymptotically approaches to zero with the increase of K. Therefore, we may alway practically assume that the true signal $\{\mathbf{A}_E^*, \mathbf{f}^*\}$ is contained in the candidate set Γ, when large enough K is considered. In particular, we summarize this scenario for Theorem 3.6 as the following corollary (similar argument applies for Theorem 3.7), where the two minimization terms are replaced by bounded terms.

Corollary 3.8. *Assume the perturbation level* $\epsilon_1 < \sqrt{2} - 1$ *and let* $\epsilon' \overset{\text{def}}{=} (1+\epsilon)(1+\epsilon_1)^2 - 1$, *where* ϵ *is an arbitrary constant in* $(0, \frac{2}{(1+\epsilon_1)^2} - 1)$. *If the true signal* $\{\mathbf{A}_E^*, \mathbf{f}^*\}$ *is contained in* Γ, *together with assumptions A1–A3 and the designed sensing matrix* $\mathbf{A}_0$ *generated as described in Section 3.2.1, the expected risk function between the true signal* $\{\mathbf{A}_E^*, \mathbf{f}^*\}$ *and the estimate* $\{\hat{\mathbf{A}}_E, \hat{\mathbf{f}}\}$ *output by the CMLE in (3.2) is bounded by*

(3.14)
$$\mathbb{E}[R(\{\hat{\mathbf{A}}_E, \hat{\mathbf{f}}\}, \{\mathbf{A}_E^*, \mathbf{f}^*\})]$$
$$\leq \mathbb{E}\left\{\sqrt{C_1\left(2\operatorname{pen}(\mathbf{f}^*) + 2\frac{\|\mathbf{A}_E^*\|_F}{K}\right)} + \sqrt{C_1\left(2P + \left\{2\frac{\|\mathbf{A}_E^*\|_F}{K}\right\}\right)} + C_2\frac{\epsilon_1}{\|\mathbf{A}_E^*\|_2}\right\},$$

with probability at least $1 - \frac{m}{2^n} - e \cdot \exp\left(-\frac{c_1\epsilon^2 K}{mn^4}\right)$, *where* $P = \max_{\mathbf{f}\in\Gamma_1}\operatorname{pen}(\mathbf{f})$, $C_1 = \left(\frac{n\sqrt{m}(1+\epsilon_1)}{(1-\epsilon')IK\sqrt{K}} + \frac{4Un}{K^2I^2(1-\epsilon')}\right)$, $C_2 = \frac{4\sqrt{mn}}{K}$ *and* $c_1 > 0$ *is a constant. The expectation is taken* *with respect to an arbitrary joint distribution of* $\{\mathbf{A}_E^*, \mathbf{f}^*\}$ *whose support satisfies assumptions A1–A3.*

When large enough K is considered, the set Γ essentially represents a dense quantization of the underlying continuous domain, and the performance bounds derived are asymptotically valid for the case that CMLE is performed over some continuous domain Γ.

The coefficients C_1 and C_2 are clearly decreasing functions of K, via which it has been suggested that the performance of CMLE can be potentially improved by increasing the number of measurements K. This result has rigorously justified the intuition that reconstruction quality enhances when more measurements are available; a similar phenomenon under the Gaussian measurement model has been observed and justified in [21]. By unifying the estimation of the true sensing matrix and signals, the accuracy of the estimated sensing matrix improves with increasing measurements, thus enhancing the signal-recovery quality. Furthermore, more-accurately estimated signals simultaneously lead to a better learning of the true underlying sensing matrix, and these complementary relationships advance to promising overall performance.

Via Corollary 3.8, it also suggest a convergence rate when one is seeking the unique ground truth $\{\mathbf{A}_E^*, \mathbf{f}^*\}$ via CMLE or relaxed-CMLE. Since that C_1 and C_2 are multiplied with bounded terms, it is straightforward to observe that the convergence rate is of $\Theta(\frac{1}{K^{3/4}})$, provided that m, n, ϵ are fixed. This convergence rate also applies to the case when the support of $\{\mathbf{A}_E^*, \mathbf{f}^*\}$ is discrete and is contained in Γ.

Moreover, we note that it is undesirable to derive a performance bound by repeatedly applying the single measurement result in [24] for each individual measurement, which claims a performance bound valid with probability p for recovering a single measurement. This simple strategy would yield a bound for recovering multiple measurements valid with probability p^K, and this probability decays to 0 with increasing number of measurements K, thereby eventually invalidating the derived performance bound.

3.3. CMLE Algorithm.

In practice, it is easier to conduct a continuous-domain optimization than searching over a discrete set. Hence, we extend the proposed CMLE with a continuous version

$$(\hat{\mathbf{A}}_E, \hat{\mathbf{f}}) = \underset{\mathbf{A}_E, \mathbf{f}}{\arg\min} \{-\log \mathrm{Pois}[\mathbf{y}; (\mathbf{A}_0 + \mathbf{A}_E)\mathbf{f} + \boldsymbol{\lambda}] + \tau_1 \,\mathrm{pen}(\mathbf{f}) + \tau_2 \|\mathbf{A}_E\|_F\} \tag{3.15}$$
$$\text{s.t.} \qquad \mathbf{A}_E = \mathbf{I}_K \otimes \boldsymbol{\Phi}_E, \quad \boldsymbol{\Phi}_0 + \boldsymbol{\Phi}_E \succeq \mathbf{0}, \quad \mathbf{f} \succeq \mathbf{0}$$

where the two regularization parameters τ_1 and τ_2 on $\mathrm{pen}(\mathbf{f})$ and $\|\mathbf{A}_E\|_F$ are incorporated to accommodate any inaccuracies on the signal energy level I and the perturbation bound ϵ_1. Practical $\mathrm{pen}(\cdot)$ choices include l_p norm $\|\cdot\|_p$ for $p \geq 1$ and the total-variation norm [26] $\|\cdot\|_{\mathrm{TV}}$. In these cases, τ_1 can also act as a scaling factor, to ensure Kraft-inequality compliance, as assumed in the previous theorems. Suitable parameters τ_1 and τ_2 can be empirically determined via cross-validation. It is straightforward to see that (3.15) is a continuous version of CMLE and previous performance bounds in Theorem 3.6 and 3.7 may still apply when the candidate set Γ is chosen as a countable set of the continuous searching domain.

We express the objective of (3.15) in terms of $\boldsymbol{\Phi}$ in the following equivalent manner. We define $\mathbf{F} \overset{\text{def}}{=} [\mathbf{f}_1, \ldots, \mathbf{f}_K]$, $\hat{\mathbf{F}} \overset{\text{def}}{=} [\hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_K]$, $\mathbf{Y} \overset{\text{def}}{=} [\mathbf{y}_1, \ldots, \mathbf{y}_K]$, $\boldsymbol{\Lambda} \overset{\text{def}}{=} \mathbf{1}_K^\top \otimes \mathbf{u}$. Then (3.15) is equivalent to

$$(\hat{\mathbf{\Phi}}, \hat{\mathbf{F}}) = \underset{\mathbf{\Phi}\succeq\mathbf{0}, \mathbf{f}\succeq\mathbf{0}}{\arg\min} \left\{ -\log \mathrm{Pois}[\mathbf{Y}; \mathbf{\Phi}\mathbf{F} + \mathbf{\Lambda}] + \tau_1 \, \mathrm{pen}(\mathbf{F}) + \tau_2 \|\mathbf{\Phi} - \mathbf{\Phi}_0\|_F \right\}, \tag{3.16}$$

which can be solved efficiently via an alternating proximal-gradient method. Specifically, we define the data-fitting term as $\ell(\mathbf{F}, \mathbf{\Phi}) \overset{\text{def}}{=} -\log \mathrm{Pois}[\mathbf{Y}; \mathbf{\Phi}\mathbf{F} + \mathbf{\Lambda}]$. In the $(t+1)$-th iteration, we solve the following two subproblems sequentially:

1) updating $\hat{\mathbf{F}}^t$:

$$\hat{\mathbf{F}}^{t+1} = \underset{\mathbf{F}\succeq\mathbf{0}}{\arg\min} \left\langle \nabla_{\mathbf{F}} \ell(\hat{\mathbf{F}}^t, \hat{\mathbf{\Phi}}^t), \mathbf{F} - \hat{\mathbf{F}}^t \right\rangle + \tau_1 \, \mathrm{pen}(\mathbf{F}) + \frac{L_f^t}{2} \|\mathbf{F} - \hat{\mathbf{F}}^t\|_F^2, \tag{3.17}$$

2) updating $\hat{\mathbf{\Phi}}^t$:

$$\hat{\mathbf{\Phi}}^{t+1} = \underset{\mathbf{\Phi}\succeq\mathbf{0}}{\arg\min} \left\langle \nabla_{\mathbf{\Phi}} \ell(\hat{\mathbf{F}}^{t+1}, \hat{\mathbf{\Phi}}^t), \mathbf{\Phi} - \hat{\mathbf{\Phi}}^t \right\rangle + \tau_2 \|\mathbf{\Phi} - \mathbf{\Phi}_0\|_F + \frac{L_\phi^t}{2} \|\mathbf{\Phi} - \hat{\mathbf{\Phi}}^t\|_F^2, \tag{3.18}$$

where L_f^t and L_ϕ^t are local Lipschitz constants. Notice that by fixing either $\mathbf{F}$ or $\mathbf{\Phi}$, $\ell(\mathbf{F}, \mathbf{\Phi})$ is a self-concordant function [27] of the other variable, for which an optimal step-size [27] is available. Details of the alternating minimization steps are summarized in Algorithm 1.

4. Bayesian Model for Real Systems: Poisson-Gamma Model.

The performance of the proposed CMLE is guaranteed by the previously derived performance bounds. However, the proposed CMLE algorithm relies significantly on tuning the parameters τ_1 and τ_2, and an inconvenient cross-validation procedure is required to adjust these parameters. Further, extra effort is demanded for a separate learning of the *dark-current*, $\mathbf{u}$, whose estimation (in)accuracy may also affect final results. Towards this end, we also propose a Bayesian model, in which $\mathbf{u}$ is *inferred* jointly with $\{\mathbf{f}_k\}$ and $\mathbf{\Phi}$, and the procedure of tuning parameters becomes unnecessary (we do have to set model hyperparameters, but the results are stable to a wide range of "reasonable" settings). If an estimate of the mean of $\mathbf{u}$ is available, it may be used in the prior, but such prior knowledge of $\mathbf{u}$ has been found unnecessary in practice.

We propose a Poisson-Gamma (PG) [36] Bayesian model:

$$\begin{array}{lcl} \mathbf{y}_k & \sim & \mathrm{Pois}\left(\mathbf{\Phi}\mathbf{f}_k + \mathbf{u}\right), \\ \mathbf{f}_k & \sim & \prod_{j=1}^n \mathrm{Gamma}(f_{k,j}; \alpha_f, \beta_f), \\ \Phi_{i,j} & \sim & \mathrm{Gamma}(\Phi_{i,j}; \beta_\Phi \Phi_{0,i,j},\ \beta_\Phi), \\ \mathbf{u} & \sim & \prod_{i=1}^m \mathrm{Gamma}(u_i; \alpha_{u,i}, \beta_{u,i}), \end{array} \tag{4.1}$$

where $\{\alpha_f, \beta_f, \beta_\Phi, \alpha_{u,i}, \beta_{u,i}\}$ are hyper-parameters and $f_{k,j}$ denotes the j-th entry of vector $\mathbf{f}_k$. Note that the gamma priors are now playing triple roles:

(a) Shrinkage priors are imparted on $\mathbf{f}_k$ to impose sparsity (the signals are indeed sparse in our problem; refer to the top row of Figure 4) by setting $\{\alpha_f, \beta_f\}$ in the way that the gamma distribution concentrates at zero, with heavy tails. Specifically, we set $\alpha_f = 1$ and $\beta_f = 10^{-6}$.

(b) The mean of the gamma prior on the sensing matrix is set equal to our initial estimation $\mathbf{\Phi}_0$ (used in its original design).

Algorithm 1 Collaborative MLE (CMLE)

```
Input: $\mathbf{\Phi}_0$, $\mathbf{u}$, $\mathbf{Y}$, initial signal estimators $\hat{\mathbf{F}}^0$, $\tau_1$, $\tau_2$
Output: Sensing matrix estimator $\hat{\mathbf{\Phi}}$, signal estimator $\hat{\mathbf{F}}$
Initialize sensing matrix estimator $\hat{\mathbf{\Phi}} \leftarrow \mathbf{\Phi}_0$
$t \leftarrow 0$
while stopping criteria not met do
    //Update $\hat{\mathbf{F}}^t$
    Compute search direction $\mathbf{d} \leftarrow \mathcal{P}_{\tau_1/L_f^t}(\hat{\mathbf{F}}^t - \nabla_{\mathbf{F}}\ell(\hat{\mathbf{F}}^t, \hat{\mathbf{\Phi}}^t)/L_f^t) - \hat{\mathbf{F}}^t$, where $\mathcal{P}_{\tau_1/L_f}$ is proximal projection operator depending on the specific pen$(\cdot)$.
    Compute optimal step-size $\alpha = \frac{L_f^t \|\mathbf{d}\|_F^2}{\|\mathbf{d}\|_{\mathbf{f}}(\|\mathbf{d}\|_{\mathbf{f}} + L_f^t \|\mathbf{d}\|_F^2)}$, where $\|\mathbf{d}\|_{\mathbf{f}}^2 = \mathbf{d}^\top \nabla_{\mathbf{F}}^2 \ell(\hat{\mathbf{F}}^t, \hat{\mathbf{\Phi}})\mathbf{d}$.
    $\hat{\mathbf{F}}^{t+1} \leftarrow \hat{\mathbf{F}}^t + \alpha\mathbf{d}$.
    //Update $\hat{\mathbf{\Phi}}$
    Compute search direction $\mathbf{g} \leftarrow \mathcal{Q}_{\tau_2/L_\phi^t}(\hat{\mathbf{\Phi}}^t - \nabla_{\mathbf{\Phi}}\ell(\hat{\mathbf{F}}^{t+1}, \hat{\mathbf{\Phi}}^t)/L_\phi^t) - \hat{\mathbf{\Phi}}^t$, where
        $$\mathcal{Q}_{\tau_2/L_\phi^t}(\mathbf{\Phi}) = \left[\mathbf{\Phi}_0 + \frac{\hat{\mathbf{\Phi}} - \mathbf{\Phi}_0}{\|\hat{\mathbf{\Phi}} - \mathbf{\Phi}_0\|_F} \cdot S_\eta(\|\hat{\mathbf{\Phi}} - \mathbf{\Phi}_0\|_F)\right]^+ .$$
        And $S_a(x) = \begin{cases} x - a & x > a \\ 0 & \text{otherwise} \end{cases}$
    Compute optimal step-size $\beta$ in the same way as computing $\alpha$.
    $\hat{\mathbf{\Phi}}^{t+1} \leftarrow \hat{\mathbf{\Phi}}^t + \beta\mathbf{g}$.
    $t \leftarrow t + 1$.
end while
```

(c) A diffuse prior is imposed on $\mathbf{u}$, i.e., $\{\alpha_{u,i} = \beta_{u,i} = 10^{-6}\}$. As discussed earlier, it is possible to estimate the dark-current by taking measurements for which it is known $\mathbf{f}_k = \mathbf{0}$, *i.e.*, for these "off line" measurements $\mathbf{y}_k \sim \text{Pois}(\mathbf{u})$. From such measurements one may estimate $\mathbf{u}$, with the estimate denoted $\mathbf{u}_0$. In this case it is appropriate to set $\alpha_{u,i} = 1$ and $\beta_{u,i} = 1/u_{0,i}$, where $u_{0,i}$ is the i-th element of the dark current $\mathbf{u}_0$ estimated (*i.e.*, $\mathbb{E}(\mathbf{u}) = \mathbf{u}_0$).

Inference for the PG model involves an augmentation scheme [7], introducing latent variable $\boldsymbol{\xi} \in \mathbb{R}_+^{K \times m \times (n+1)}$, with

$$\begin{aligned} \xi_{k,i,j} &= \frac{\hat{\Phi}_{i,j}\hat{f}_{k,j}y_{k,i}}{\sum_{j=1}^n \hat{\Phi}_{i,j}\hat{f}_{k,j} + \hat{u}_i}, \ j = 1, \ldots, n; \\ \xi_{k,i,n+1} &= \frac{\hat{u}_i y_{k,i}}{\sum_{j=1}^n \hat{\Phi}_{i,j}\hat{f}_{k,j} + \hat{u}_i}. \end{aligned}$$

We have developed both Markov Chain Monte Carlo (MCMC) sampling and Expectation Maximization (EM) inference methods, with details provided in Appendix F. The update equations of $\{\hat{\mathbf{f}}_k\}_{k=1}^K$, $\hat{\boldsymbol{\Phi}}$ and $\hat{\mathbf{u}}$ in the EM inference are

$$\begin{aligned} \hat{f}_{k,j} &= \left[\frac{\alpha_f+\sum_{i=1}^m \xi_{k,i,j}-1}{\beta_f+\sum_{i=1}^m \hat{\Phi}_{i,j}}\right]^+, \quad \forall k=1,\ldots,K;\ i=1,\ldots,m;\ j=1,\ldots,n; \\ \hat{\Phi}_{i,j} &= \left[\frac{\beta_\Phi \Phi_{0,i,j}+\sum_{k=1}^K \xi_{k,i,j}-1}{\beta_\Phi+\sum_{k=1}^K \hat{f}_{k,j}}\right]^+, \\ \hat{u}_i &= \left[\frac{\alpha_{u,i}+\sum_{k=1}^K \xi_{k,i,n+1}-1}{\beta_{u,i}+K}\right]^+, \end{aligned} \tag{4.2}$$

and we refer to the proposed Bayesian model with EM inference as the EMPG algorithm.

4.1. Connections to Other Methods. The Bayesian and optimization-based methods are closely related, especially when a point estimate (*e.g.*, EM) is used. Let $\boldsymbol{\Theta} = \{\boldsymbol{\Phi}, \{\mathbf{f}_k\}_{k=1}^K, \mathbf{u}\}$ denote all the variables to be inferred. The log-posterior of the PG model may be expressed as:

$$\begin{aligned} \log p(\boldsymbol{\Theta}|\{\mathbf{y}_k\}_{k=1}^K) = &\sum_\ell \log\left(\mathrm{Pois}(\boldsymbol{\Phi}\mathbf{f}_k + \mathbf{u})\right) - \beta_f \sum_k \|\mathbf{f}_k\|_1 \\ &+(\alpha_f - 1)\sum_{k,j} \log f_{k,j} - \beta_\Phi \sum_{i,j} \Phi_{i,j} + (\beta_\Phi \Phi_{0,i,j} - 1)\sum_{i,j} \log \Phi_{i,j} \\ &- \sum_i \beta_{u,i}\mathbf{u}_i + \sum_i (\alpha_{u,i} - 1)\log \mathbf{u}_i + const. \end{aligned} \tag{4.3}$$

Note in (4.3) that the ℓ_1 penalty is inherently imposed on $\mathbf{f}_k$, and the third term accounts for the over-dispersion [35] effect of count data in a variety of real applications. Interestingly, when $\alpha_f = 1$, (4.3) provides the same formulation as CMLE with the ℓ_1 penalty, and β_f plays the role of τ_1 in (3.15). However, in CMLE, a tuning of the parameter τ_1 is needed, while for the PG model we can readily infer β_f via a conjugate prior (gamma distribution) if desired.

Beyond the Poisson measurement model considered in this paper, the PG model also connects to other existing algorithms. Recall the update equation of $\hat{f}_{\ell,j}$ in (4.2), and notice that if we fix $\mathbf{u} = \mathbf{0}$, $\alpha_f = 1$, $\beta_f = 0$, it is equivalent to

$$\hat{f}_{k,j}^t \leftarrow \hat{f}_{k,j}^t \cdot \frac{\sum_{i=1}^m \frac{\hat{\Phi}_{i,j} y_{k,i}}{\sum_j \hat{\Phi}_{i,j}\hat{f}_{k,j}}}{\sum_{i=1}^m \hat{\Phi}_{i,j}}, \tag{4.4}$$

where the superscript t indexes the iteration of the inference. Equation (4.4) is the same multiplicative update as in the Richardson–Lucy algorithm [25], a classical image-deblurring method. It has also been noticed in [36] that (4.4) is closely connected to nonnegative matrix factorization guided by a KL-divergence-minimization criterion, *i.e.*, $\min_{\boldsymbol{\Phi},\mathbf{f}} \mathrm{KL}(\mathbf{y}||\boldsymbol{\Phi}\mathbf{f})$ [17]. In addition to these methods, the PG model is able to accommodate a sparsity assumption on $\{\hat{\mathbf{f}}_k\}_{k=1}^K$ as previously explained, thereby exhibiting more flexibility. Compared with a recently proposed gradient-descent based methods, Spiral-TAP [10], multiplicative updating is *exempt* from step-size searching, which is costly yet irreplaceable in balancing the algorithm convergence and speed. In the experiments, we have noticed that each iteration of EMPG is much faster than that of CMLE, due to the omitted step-size searching.

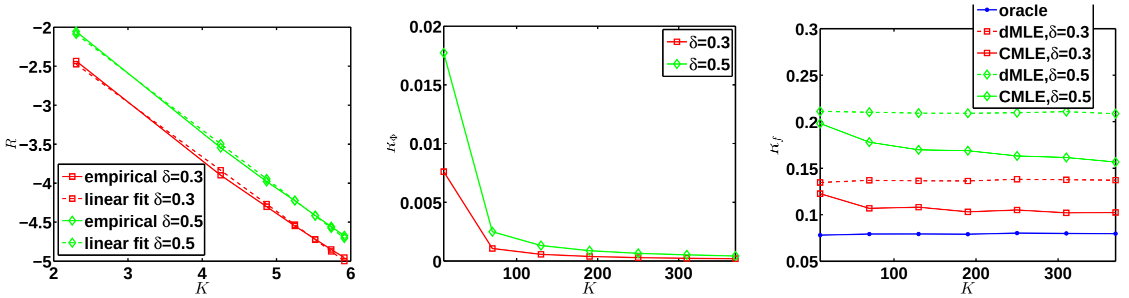
5. Experiments. We present results on both synthetic and real data. As a baseline, we compare the proposed algorithms with an estimator that assumes $\mathbf{\Phi}_0$ is the sensing matrix, *ignoring* perturbation $\mathbf{\Phi}_E$. We refer to this as a "degraded" estimator, allowing assessment of how uncertainty in the sensing matrix affects performance, if it is not accounted. For the "degraded" estimator, we consider an MLE solution, denoted dMLE, in which we skip the update $\hat{\Phi}$ step in Algorithm 1; the dMLE algorithm provides a fair comparison with CMLE, connecting theory and experiments. For the synthetic data, since ground truth is available, we also compare with the *oracle* estimator, which knows exactly the ground-truth sensing matrix $\mathbf{\Phi}^*$. The simulated perturbation matrix $\mathbf{\Phi}_E$ is generated by uniformly drawing entries from a small interval $(0, \delta]$ around zero, with $\delta = \frac{\|\mathbf{\Phi}_E\|_F}{\|\mathbf{\Phi}_0\|_F}$ set within a given experiment. The normalized Mean Square Error (R) of $\hat{\mathbf{\Phi}}$ and $\hat{\mathbf{f}}_k$ are performance metrics

$$(5.1) \qquad R_\Phi \stackrel{\text{def}}{=} \frac{1}{K}\frac{\|\hat{\mathbf{\Phi}} - \mathbf{\Phi}^*\|_F}{\|\mathbf{\Phi}^*\|_F}, \qquad R_f \stackrel{\text{def}}{=} \frac{1}{K}\sum_{k=1}^{K}\frac{\|\hat{\mathbf{f}}_k - \mathbf{f}_k^*\|_2}{\|\mathbf{f}_k^*\|_2}.$$

consistent with the risk metric in (3.11). The code was written in MATLAB and executed on a 2.6GHz CPU with 4GB RAM.

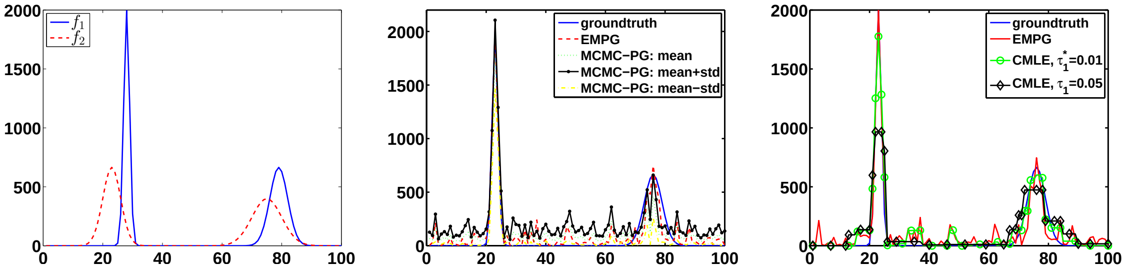
5.1. Synthetic Data: Consideration of the CMLE Theory. We first verify the theoretical properties of the CMLE algorithm, considering the special randomized sensing matrix $\mathbf{\Phi}_0$ developed in the theory of Section 3.2. The $\mathbf{\Phi}_0 \in \mathbb{R}_+^{200\times1000}$ is drawn from the shifted Rademacher distribution, as described in Section 3.1, and we consider here $\delta = 0.3$ and 0.5. Each signal $\mathbf{f}_k \in \mathbb{R}_+^{1000}$ is sparse (with 1% nonzero entries) and has a fixed intensity $\|\mathbf{f}_k\|_1 = 10^4$. For each $\mathbf{f}_k$, the corresponding measurement $\mathbf{y}_k$ is generated by $\mathbf{y}_k \sim \text{Pois}(\mathbf{\Phi}\mathbf{f}_k)$ (assuming no dark-current in this experiment). To verify the effect of multiple sets of measurements, we vary K from 10 to 370. The calculated risk function R, R_Φ and R_f are averaged over 100 trials. The ℓ_1 penalty is selected to capture the sparsity of $\mathbf{f}_k$. The regularizers τ_1 and τ_2 are cross-validated, *i.e.*, they are chosen such that the R_f is minimized on a smaller training set. We compare R_f for CMLE, dMLE and an oracle estimator. Both dMLE and the oracle estimator fix their knowledge of the sensing matrix, and recover each signal separately; for the oracle estimator the sensing matrix $\mathbf{\Phi} = \mathbf{\Phi}_0 + \mathbf{\Phi}_E$ is assumed known exactly, where for dMLE $\mathbf{\Phi}_E$ is ignored. By contrast, CMLE jointly estimates $\mathbf{\Phi}_E$ and the underlying signals $\{\mathbf{f}_k\}$. Figure 2 demonstrates that R, R_Φ and R_f associated with CMLE constantly decrease as K grows, consistent with the performance guarantees in Theorem 3.6 and 3.7. Moreover, it can be observed from Figure 2(a) that $\log R$ asymptotically exhibits a decay rate of $\Theta(\frac{1}{K^{3/4}})$ with respect to $\log K$, and this coincides with the $\Theta(\frac{1}{K^{3/4}})$ convergence rate suggested by Theorems 3.6 and 3.7. Figure 2(c) compares the R_f from all three estimators. CMLE consistently yields smaller R_f than dMLE, and approaches the oracle estimator for large K.

5.2. Synthetic Data: A Realistic Scenario. We next consider a more realistic scenario to verify the efficacy of the PG model. In real applications, the sensing matrix $\mathbf{\Phi}_0$ may not be designed as described in Section 3.1, and a significant amount of *dark-current* may be present. We generate $\mathbf{\Phi}_0 \in \mathbb{R}_+^{50\times100}$ by randomly setting half the entries to nonzero, with non-zero values distributed uniformly over $[0, 1]$, and generate $\mathbf{\Phi}_E$ via $\delta = 0.3$. Each $\{\mathbf{f}_k\}_{k=1}^{30}$ is generated by concatenating several bell-shaped curves of different bandwidths (examples



(a) (b) (c)

Figure 2: CMLE performance with varying number of measurements, K. (a) Logarithm of risk R decreases roughly linearly w.r.t. $\log K$, and the two dashed lines (labeled "linear fit") are of slope $-3/4$. (b) R_Φ for CMLE. (c) R_f, with comparison between CMLE and dMLE.



(a) (b) (c)

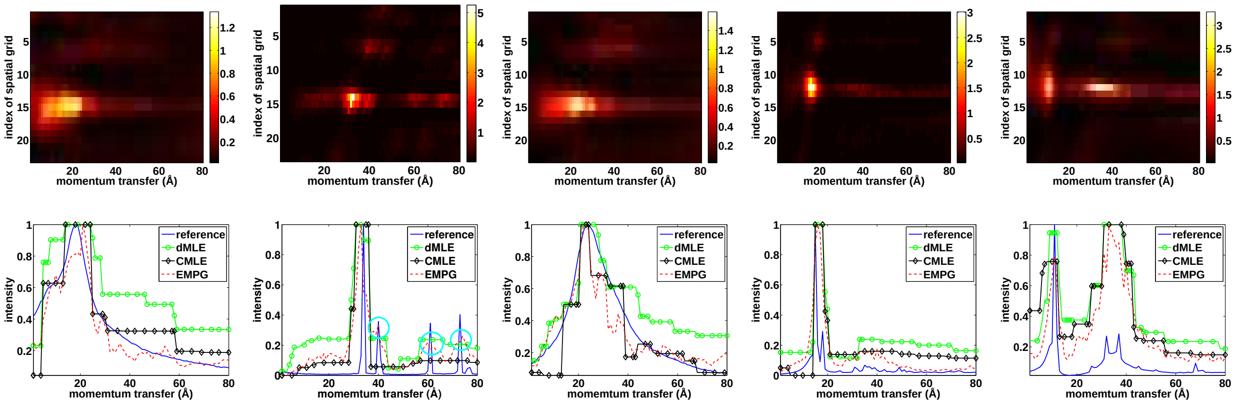
Figure 3: Comparison of EMPG, MCMC-PG and CMLE on synthetic data: (a) Example groundtruth signals. (b) Reconstruction of one example signal by MCMC-PG and EMPG. For MCMC-PG, we obtain the mean values of the collected samples as well as the confidence region defined by the standard deviation (std) of these samples. (c) Reconstruction of one example signal by EMPG and CMLE. Via CMLE, a small deviation of τ_1 (e.g., 5×10^{-2}) from the optimal $\tau_1^* = 10^{-2}$ gives inferior reconstruction.

shown in Figure 3(a)). This introduces variation in the signal smoothness and results in nontrivial tuning of τ_1 in CMLE. We simulate the noisy measurements by $\mathbf{y}_k \sim \text{Pois}(\mathbf{\Phi f}_k + \mathbf{u})$, where $\mathbf{u} = 100 \times \mathbf{1}_{50}$ is unknown to the algorithms. Both CMLE and the Bayesian inversion methods have access to an MLE of $\mathbf{u}$ from 30 measurements of the empty system (no input source), *i.e.*, $\hat{\mathbf{u}} = \frac{1}{30}\sum_{k=1}^{30} \mathbf{u}_{0,k}$ where $\mathbf{u}_{0,k} \sim \text{Pois}(\mathbf{u})$.

Both MCMC and EM inference of the PG model provide effective reconstruction results (Figure 3(b)). In MCMC, we collected 1000 samples after discarding the first 5000 samples as burn-in. Via these collected samples, we obtain not only the mean estimated signal but also the *confidence* of the estimation, where here the measure of confidence is quantified via the standard deviation shown in Figure 3(b). The EM-based solution only requires 500 iterations to achieve comparable results, but it does not provide a measure of confidence in the point estimate. Each iteration of the EM and MCMC inference methods takes about the same CPU time ($\sim 8.1 \times 10^{-3}$ seconds); we use EM-based inference for the PG model in the rest of the paper.

We have found that the PG results are insensitive to initialization of the dark current and do not require its offline estimate (via the measurements of the empty system), while this estimate was found to be crucial for CMLE. A TV (total variation) [26] regularization is adopted in CMLE, imposing smoothness of the signal and $\{\tau_1 = 10^{-2}, \tau_2 = 10^4\}$ are chosen such that the smallest R_f (defined in (5.1)) is achieved. Parameters $\beta_\Phi = 1000$, $\alpha_{u,i} = 1$ and $\beta_{u,i} = 1/u_{0,i}$ are utilized in the PG model. We observed that the values of τ_1 and τ_2 significantly impacted the CMLE performance (Figure 3(c)). Tuning these two parameters is challenging, particularly when the signal smoothness varies. Therefore, the Bayesian model is a better choice than CMLE in these realistic cases. We emphasize again that the PG model does not need to tune parameters (with appropriate hyperparameter settings, to which the results were found relatively insensitive), and thus PG saves much computation time. Each iteration of EMPG and CMLE takes 8.1×10^{-3} and 2.3×10^{-1} seconds, respectively. The R_f achieved by CMLE and EMPG are 0.3007 and 0.2343, respectively. One example signal and its estimates by these two methods are shown in Figure 3(c), providing a representative demonstration that EMPG outperforms CMLE in this realistic CS scenario. Based on these simulated experiments, for the real data considered next we perform inversion with the PG model, based on EM inference.

5.3. Real Data: X-ray Scatter Imaging. Recalling Figure 1, the X-ray system considered recovers both spatial (depth along z-axis) and spectral (manifested by the momentum-transfer spectrum) information of the material. Therefore, the input signal has two degrees of freedom, and can be considered as a two-dimensional image $\mathbf{G}_k \in \mathbb{R}_+^{d_l \times d_s}$, where $d_l = 27$ and $d_s = 101$ are the number of quantized grids in space and spectrum ($n = 2727$), respectively. The measurements are expressed mathematically as in (2.1), where $\mathbf{f}_k$ is defined as $\mathbf{G}_k$.



(a) Methanol (b) Al (c) H_2O_2 (d) HDPE (e) Teflon

Figure 4: Top: Reconstructed two-dimensional input signal (a cropped region of interest) by EMPG. Bottom: extracted spectrums (corresponding to the strongest row in the two-dimensional image). The dMLE is CMLE-based but ignores the perturbation of the sensing matrix and performs K inversions independently.

In each compressive measurement, a small piece of material (*i.e.*, a point object) is placed in the system, occupying approximately a single spatial grid point. This results (ideally) in a single nonzero row in the two-dimensional image of $\mathbf{G}_k$, for which example reconstructions are presented in the top row of Figure 4. In each of these images, we see that the signal intensity is indeed concentrated near one point.

These most intense rows are extracted and considered as the reconstructed spectrums for the materials. The spectrum obtained in this manner is denoted as $\hat{\mathbf{s}}_k \in \mathbb{R}_+^{d_s}$, and it ideally resembles the reference spectrum $\mathbf{s}_k$ from a spectrum library (this reference library is measured separately for each material, using a conventional X-ray diffraction sensor). The correlation $\text{corr}_k = \frac{\hat{\mathbf{s}}_k^\top \mathbf{s}_k}{\|\hat{\mathbf{s}}_k\| \cdot \|\mathbf{s}_k\|}$ between $\hat{\mathbf{s}}_k$ and $\mathbf{s}_k$ is employed as a performance metric.

We collect $K = 25$ measurements, and for each we place different materials in the system, or the same material at different depths; for each the measurement dimension is $m = 2679$. The mean of the dark-current $\mathbf{u}$ is obtained from measurements with an empty system (no material placed in the machine). Note that while here m is only slightly smaller than n, and hence the system is only slightly compressive from that standpoint, the significant advantage is that the sensor does not have to sequentially scan the spatial dimension, markedly speeding sensing. We make comparisons to dMLE, CMLE and EMPG. The dMLE assumes perfect knowledge of the designed sensing matrix, the $\mathbf{\Phi}_0$ for which the system was designed, *ignoring* the perturbation $\mathbf{\Phi}_E$. Since the sensing matrix is not estimated in this case, each of the K inversions are performed independently. The poor quality of these results were the motivation for the work reported in this paper.

A two-dimensional-TV regularizer [26] is used within CMLE to impose smoothness in both spectrum and depth; setting $\tau_1 = 0.02$ and $\tau_2 = 10^4$ yields the best results. For the PG model, *shrinkage* priors ($\alpha_f = 1, \beta_f = 10^{-6}$) are imposed on $\mathbf{f}$, and $\beta_\Phi = 1$, $\alpha_{u,i} = 1$, $\beta_{u,i} = 1/u_{0,i}$. In fact, we observed that PG is not sensitive to these parameters. Figure 4 shows the reconstructed $\hat{\mathbf{G}}_k$ (5 representative examples selected out of 25) by EMPG and compares the estimated spectrums $\hat{\mathbf{s}}_k$ with other methods. Since the material only occupies one spatial grid, we can locate it by finding the strongest row in the two-dimensional image $\hat{\mathbf{G}}_k$ (thus estimating the depth); normalizing this row gives us $\hat{\mathbf{s}}_k$ (the bottom row of Figure 4). Table 1 summarizes the correlation between the reference spectrum and the estimates by each approach. From the spectrum plots and correlations, we see that both CMLE and EMPG achieve marked improvements over the degraded estimator, and EMPG consistently performs the best. Furthermore, from the second row of Figure 4, we see that EMPG is more capable of capturing the *peaks* of the spectrum.

Table 1: Correlation between $\hat{\mathbf{s}}_k$ (estimates) and $\mathbf{s}_k$ (reference)

	Methanol	Al	H_2O_2	HDPE	Teflon
dMLE	0.9436	0.4190	0.9549	0.6906	0.7256
CMLE	0.9572	0.4299	0.9661	0.7100	0.6991
EMPG	**0.9796**	**0.4356**	**0.9713**	**0.7321**	**0.7312**

Considering Figure 4, results are plotted as a function of momentum transfer, which we wish to relate back to the experimental system and model, as discussed in Sec. 2. Recall that the relationship between the lattice spacing d, excitation energy E and scatter angle θ is given by Bragg's Law: $q = \frac{1}{2d} = \frac{E}{hc}\sin(\theta/2)$. The expression $q = 1/2d$ is called the *momentum* *transfer*. For each value of q, the variation in E across the source bandwidth yields a curve in the (E,θ) space. The intensity of each of these curves is inferred as a function of q, and these q-dependent intensity curves are plotted in the second row of Figure 4. The intensity of the signal as a function of momentum transfer q is used as a signature of the material (a given material is characterized by the intensity with which each momentum transfer is manifested).

6. Conclusion.

We have developed collaborative reconstruction algorithms for multiple compressive Poisson measurements, to address the physical perturbations of sensing matrix in real systems. The signals and measurement matrix are jointly estimated. The CMLE algorithm has first been proposed. A new concentration-of-measure result has been established and its theoretical applications on the linear multiple-measurement model have been presented. Theoretical performance guarantees for the proposed algorithm have been established. In order to improve the flexibility and robustness, we have also developed a Bayesian model to jointly estimate signals, dark-current and the sensing matrix, where tuning of parameters via cross-validation is unnecessary. We have demonstrated in our experiments that both the CMLE and EMPG algorithms achieve promising results, while a superior performance of the Bayesian model is suggested in realistic scenarios, such as the actual X-ray scatter imaging system we considered.

Acknowledgment. The research performed at Duke was supported by DARPA under the KECOM program and by ONR under a Research Challenge Program focused on compressive sensing. VC's work was supported in part by the European Commission under grants MIRG-268398 and ERC Future Proof and by the Swiss Science Foundation under grants SNF 200021-146750 and SNF CRSII2-147633.

Appendix. Proof of Theorem 3.1.

Proof. Let $\mathbf{y} \overset{\text{def}}{=} \tilde{\mathbf{A}}_0\mathbf{f}$. Define $\mathbf{L} \overset{\text{def}}{=} \mathbf{X}\mathbf{X}^\top$ where $\mathbf{X} \overset{\text{def}}{=} [\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_K]$. We have $\|\mathbf{y}\|_2^2 = \sum_{i=1}^m \boldsymbol{\psi}_i^\top \mathbf{L}\boldsymbol{\psi}_i$, where $\boldsymbol{\psi}_i^\top$ is the i-th row of $\boldsymbol{\Psi}$. Consider the eigen-decompositon of $\mathbf{L} = \mathbf{V}^\top\mathbf{D}\mathbf{V}$ with $\mathbf{V}$ being orthonormal, where $\mathbf{D} = \mathrm{diag}\{\lambda_1, \dots, \lambda_n\}$ and let $\mathbf{w}_i = \mathbf{V}\boldsymbol{\psi}_i$, we have

$$\|\mathbf{y}\|_2^2 = \sum_{i=1}^m \boldsymbol{\psi}_i^\top \mathbf{L}\boldsymbol{\psi}_i = \sum_{i=1}^m \boldsymbol{\psi}_i^\top \mathbf{V}^\top\mathbf{D}\mathbf{V}\boldsymbol{\psi}_i = \sum_{i=1}^m \mathbf{w}_i^\top \mathbf{D}\mathbf{w}_i = \sum_{j=1}^n\sum_{i=1}^m \lambda_j w_{ij}^2, \tag{A.1}$$

where w_{ij} denotes the j-th component of $\mathbf{w}_i$. The expectation of $\|\mathbf{y}\|_2^2$ can be calculated as

$$\mathbb{E}[\|\mathbf{y}\|_2^2] = \sum_{j=1}^n\sum_{i=1}^m \lambda_j \mathbb{E}[w_{ij}^2] = \sum_{j=1}^n \lambda_j = \mathrm{tr}(\mathbf{X}\mathbf{X}^\top) = \|\mathbf{f}\|_2^2, \tag{A.2}$$

where we use the fact that $\mathbb{E}[w_{ij}^2] = \mathbb{E}[(\sum_{k=1}^n V_{jk}\psi_{ik})^2] = \frac{1}{m}\sum_{k=1}^n V_{jk}^2 = \frac{1}{m}$. Hence, we can derive the following expressions

$$\mathbb{P}\left(\left|\|\mathbf{y}\|_2^2 - \|\mathbf{f}\|_2^2\right| > \epsilon\|\mathbf{f}\|_2^2\right) = \mathbb{P}\left(\left|\|\mathbf{y}\|_2^2 - \mathbb{E}[\|\mathbf{y}\|_2^2]\right| > \epsilon\|\mathbf{f}\|_2^2\right) \tag{A.3}$$

$$= \mathbb{P}\left(\left|\sum_{i=1}^m\left(\sum_{j=1}^n \lambda_j w_{ij}^2 - \mathbb{E}\left[\sum_{j=1}^n \lambda_j w_{ij}^2\right]\right)\right| > \epsilon\|\mathbf{f}\|_2^2\right) \tag{A.4}$$

$$= \mathbb{P}\left(\left|\sum_{i=1}^m (s_i - \mathbb{E}[s_i])\right| > \epsilon\|\mathbf{f}\|_2^2\right), \tag{A.5}$$

where $s_i \overset{\text{def}}{=} \sum_{j=1}^n \lambda_j w_{ij}^2$ and the probability measure is associated with $\mathbf{\Psi}$. Note that $\mathbf{L}$, λ_i and $\mathbf{V}$ are deterministic and $\boldsymbol{\psi}_i$, $i = 1, \dots, m$ are random vectors.

Consider

$$\mathcal{W} = \{\mathbf{w}|\mathbf{w} = \mathbf{V}\psi_i,\ i = 1, \dots, m \text{ and } \mathbf{V} \text{ is associated with all } \mathbf{f} \in \Delta\}$$

and define a map $\phi : \mathcal{W} \to \mathbb{R}$ by $\phi(w_i) \mapsto \sum_{j=1}^n \lambda_j w_{ij}^2$ where λ_i, $i = 1, \dots, n$ are the eigenvalues corresponding to the eigen-matrix $\mathbf{V}$ that is associated with $\mathbf{w}_i$. Equip respectively $\mathcal{W}$ and $\mathbb{R}$ with the Hamming distance and Euclidean distance, where the Hamming distance between $\mathbf{u} \in \mathcal{W}$ and $\mathbf{z} \in \mathcal{W}$ is defined as $|\mathbf{u} - \mathbf{z}|_H = \sum_{i=1}^n (1 - \delta_{u_i - z_i})$, where δ is the Kronecker delta function. By our assumption that Δ is at most countable, we can generate at most countably many $\mathbf{V}$ from Δ. Thus $\mathcal{W}$ is countable at most. We first show that the map ϕ is a Lipschitz map. For all $\mathbf{u},\ \mathbf{z} \in \mathcal{W}$, we have

$$\begin{aligned} |\phi(\mathbf{u}) - \phi(\mathbf{z})| &= \left|\sum_{j=1}^n \lambda_j(u_j - z_j)\right| \\ &\le \sum_{j=1}^n \lambda_j\,|(u_j - z_j)| \\ &\le L_n \sum_{j=1}^n (1 - \delta_{u_j - z_j}) \\ &= L_n|\mathbf{u} - \mathbf{z}|_H, \end{aligned} \tag{A.6}$$

where L_n is a constant depending only on n and (A.6) follows from the fact that $\mathcal{W}$ is a bounded space as $\mathbf{\Psi}$ and $\mathbf{V}$ are bounded. Hence ϕ is an L_n-Lipschitz map. Together with the fact that $\mathcal{W}$ is at most countable, by the main theorem in [15], we have the following concentration inequality on ϕ,

$$\mathbb{P}(|\phi(w_i) - \mathbb{E}[\phi(w_i)]| \ge t) \le 2e^{\frac{-t^2}{nC}},\ \forall t > 0, \tag{A.7}$$

where C is a constant. The result is remarkable since that one does not require $w_i \in \mathbb{R}^n$ to have independent entries and this result may serve as a generalization of the well-known McDiarmid's inequality.

By the concentration inequality (A.7), we may conclude that $s_i = \phi(w_i)$ is a sub-Gaussian random variable by its definition. Moreover, s_i, $i = 1, \dots, m$ are independent sub-Gaussian random variables. Via the Hoeffding inequality for sub-Gaussian random variables [28], we have

$$\mathbb{P}\left(\left|\sum_{i=1}^{m}(s_i - \mathbb{E}[s_i])\right| > \epsilon \|\mathbf{f}\|_2^2\right) \le e \cdot \exp\left(-\frac{c_3 \epsilon^2 \|\mathbf{f}\|_2^4}{B^2 m}\right), \tag{A.8}$$

where $c_3 > 0$ is a constant. $B \overset{\text{def}}{=} \max_i \|s_i\|_{\mathcal{O}}$ and $\|\cdot\|_{\mathcal{O}}$ is the *Orlicz* norm [28].

By the property of *Orlicz* norm [28], we have the following bound for all $\|s_i\|_{\mathcal{O}}$, $i = 1, \dots, m$

$$\|s_i\|_{\mathcal{O}} \le c_2 n, \tag{A.9}$$

where $c_2 > 0$ is a constant.

Combining this bound with (A.8), we have

$$\mathbb{P}\left(\left|\sum_{i=1}^{m}(s_i - \mathbb{E}[s_i])\right| > \epsilon \|\mathbf{f}\|_2^2\right) \le e \cdot \exp\left(-\frac{c_1 \epsilon^2 \|\mathbf{f}\|_2^4}{m n^2}\right). \tag{A.10}$$

Note that above result is equivalent to the statement of the theorem. ■

Appendix. Proof of Corollary 3.2.

Proof. Apply Theorem 3.1 to the vector $\mathbf{f} - \mathbf{g}$, we obtain

$$\begin{aligned}
\mathbb{P}(\left|\|\tilde{\mathbf{A}}_0(\mathbf{f}-\mathbf{g})\|_2^2 - \|\mathbf{f}-\mathbf{g}\|_2^2\right| \ge \epsilon \|\mathbf{f}-\mathbf{g}\|_2^2) &\le e \cdot \exp\left(-\frac{c_1' \epsilon^2 \|\mathbf{f}-\mathbf{g}\|_2^4}{m n^2}\right) \\
&\le e \cdot \exp\left(-\frac{c_1' \epsilon^2 \|\mathbf{f}-\mathbf{g}\|_1^4}{m K^2 n^4}\right) = e \cdot \exp\left(-\frac{c_1 \epsilon^2 K}{m n^4}\right), \qquad \text{(B.1)}
\end{aligned}$$

where (B.1) follows from the fact $\|\mathbf{f}-\mathbf{g}\|_2 \ge \frac{\|\mathbf{f}-\mathbf{g}\|_1}{\sqrt{Kn}}$ and the assumption $\mathcal{QS}(\Gamma) \ge \frac{d_1}{\sqrt[4]{K}}$. Notice that we have $\|f - g\|_1 \ge d_1 K^{\frac{3}{4}}$. Above inequality is equivalent to the statement that

$$(1-\epsilon)\|\mathbf{f}-\mathbf{g}\|_2^2 \le \|\tilde{\mathbf{A}}_0(\mathbf{f}-\mathbf{g})\|_2^2 \le (1+\epsilon)\|\mathbf{f}-\mathbf{g}\|_2^2,\ \forall \mathbf{f} \in \Gamma_1, \tag{B.2}$$

with probability at least

$$1 - e \cdot \exp\left(-\frac{c_1 \epsilon^2 K}{m n^4}\right). \tag{B.3}$$

■

Appendix. Proof of Theorem 3.4.

Proof. For $\mathbf{x} \in U$ and $\mathbf{y} \in V$, apply Theorem 3.1 to vector $\mathbf{f} = \mathbf{x} - \mathbf{y}$. By the product rule, we have that the demanded ϵ-stable embedding holds for all $\mathbf{x} - \mathbf{y}$ with probability at least

(C.1)

$$\prod_{\substack{\mathbf{x}\in U,\mathbf{y}\in V\\ x\neq\mathbf{y}}}\left[1-e\cdot\exp\left(-\frac{c_1\epsilon^2\|\mathbf{x}-\mathbf{y}\|_2^4}{mn^2}\right)\right]\geq 1-\sum_{\substack{\mathbf{x}\in U,\mathbf{y}\in V\\ x\neq\mathbf{y}}}\left[e\cdot\exp\left(-\frac{c_1\epsilon^2\|\mathbf{x}-\mathbf{y}\|_2^4}{mn^2}\right)\right]$$

$$\geq 1-|U||V|\left[e\cdot\exp\left(-\frac{c_1\epsilon^2\min_{\substack{\mathbf{x}\in U,\mathbf{y}\in V\\ x\neq\mathbf{y}}}\|\mathbf{x}-\mathbf{y}\|_2^4}{mn^2}\right)\right]. \tag{C.2}$$

Equate the right hand side of above inequality to $1-\rho$ and solve for ϵ. We have

$$\epsilon=\sqrt{\frac{mn^2(\log\rho+1+\log|U|+\log|V|)}{c_1\min_{\substack{\mathbf{x}\in U,\mathbf{y}\in V\\ \mathbf{x}\neq\mathbf{y}}}\|\mathbf{x}-\mathbf{y}\|_2^4}}. \tag{C.3}$$

Hence, we have proved the claim. ■

Appendix. Proof of Theorem 3.6.

In order to prove Theorem 3.6, we also need to establish the following lemmas which will be useful later.

Lemma D.1. *Consider* $\mathbf{G}\in\mathbb{R}^{m\times n}$ *and* $\tilde{\mathbf{G}}\overset{\text{def}}{=}\mathbf{I}_K\otimes\mathbf{G}$. *Then* $\|\mathbf{G}\|_2=\|\tilde{\mathbf{G}}\|_2$, *where* $\|\cdot\|_2$ *denotes the operator norm.*

Proof. [Proof of Lemma D.1] By definition, we have

$$\|\tilde{\mathbf{G}}\|_2^2=\sup_{\|\mathbf{x}\|_2=1}\|\tilde{\mathbf{G}}\mathbf{x}\|_2^2 \tag{D.1}$$

$$=\sup_{\|\mathbf{x}\|_2=1}\sum_{i=1}^{k}\|\mathbf{G}\mathbf{x}_i\|_2^2 \tag{D.2}$$

where $\mathbf{x}\overset{\text{def}}{=}(\mathbf{x}_1^T,\ldots,\mathbf{x}_K^T)^T\in\mathbb{R}^{Kn}$. Since $\|\mathbf{G}\mathbf{x}_i\|_2^2\leq\|\mathbf{G}\|_2^2\|\mathbf{x}_i\|_2^2$, we have

$$\|\tilde{\mathbf{G}}\|_2^2\leq\|\mathbf{G}\|_2^2\sup_{\|\mathbf{x}\|_2=1}\sum_{i=1}^{K}\|\mathbf{x}_i\|_2^2=\|\mathbf{G}\|_2^2\sup_{\|\mathbf{x}\|_2=1}\|\mathbf{x}\|_2^2=\|\mathbf{G}\|_2^2. \tag{D.3}$$

Conversely, consider $\mathbf{y}^*$ with $\|\mathbf{y}^*\|_2=1$ being a vector such that $\|\mathbf{G}\|_2=\|\mathbf{G}\mathbf{y}^*\|_2$. By letting $\mathbf{x}_1=\mathbf{y}^*$ and $\mathbf{x}_i=\mathbf{0}$ for $i=2,\ldots,k$, we have $\|\tilde{\mathbf{G}}\|_2\geq\|\mathbf{G}\|_2$. Hence $\|\mathbf{G}\|_2=\|\tilde{\mathbf{G}}\|_2$ ■

Lemma D.2.

$$2\mathbb{E}_{\mathbf{A}^*,\mathbf{f}^*}\left[\log\frac{1}{\int\sqrt{p(y|\mathbf{A}^*\mathbf{f}^*+\boldsymbol{\lambda})p(y|\hat{\mathbf{A}}\hat{\mathbf{f}}+\boldsymbol{\lambda})d\nu(y)}}\right]$$

$$\leq\mathbb{E}_{\mathbf{A}^*,\mathbf{f}^*}\left[\min_{\{\mathbf{A}_E,\mathbf{f}\}\in\Gamma}\left\{\mathrm{KL}(p(\cdot|(\mathbf{A}_0+\mathbf{A}_E^*)\mathbf{f}^*+\boldsymbol{\lambda})||p(\cdot|(\mathbf{A}_0+\mathbf{A}_E)\mathbf{f}+\boldsymbol{\lambda}))+2\,\mathrm{pen}(\mathbf{f})+2\frac{\|\mathbf{A}_E\|_F}{K}\right\}\right],$$

where ν is the counting measure on $\mathbb{Z}_+^{Km}$, $\boldsymbol{\lambda} = \mathbf{1}_K \otimes \mathbf{u}$ is the known dark-current and $\mathrm{KL}(\cdot||\cdot)$ *denotes the Kullback-Leibler distance. The expectation is taken with respect to an arbitrary joint distribution on $\{\mathbf{A}, \mathbf{f}\}$.*

Proof. [Proof of Lemma D.2] The proof is based on the techniques as in [18]. Denote $p_{\mathbf{A}^*\mathbf{f}^*} \overset{\text{def}}{=} p(y|(\mathbf{A}_0 + \mathbf{A}_E^*)\mathbf{f}^* + \boldsymbol{\lambda})$ and $p_{\mathbf{Af}} \overset{\text{def}}{=} p(y|(\mathbf{A}_0 + \mathbf{A}_E)\mathbf{f} + \boldsymbol{\lambda})$. Define $\mathcal{H}(\mathbf{A}^*\mathbf{f}^*, \mathbf{Af}) \overset{\text{def}}{=} \int \sqrt{p_{\mathbf{A}^*\mathbf{f}^*} p_{\mathbf{Af}}} d\nu$ as the Hellinger affinity. We have

(D.4)

$$2\log\frac{1}{\mathcal{H}(\mathbf{A}^*\mathbf{f}^*, \hat{\mathbf{A}}\hat{\mathbf{f}})} = 2\log\left[\frac{\sqrt{p_{\hat{\mathbf{A}}\hat{\mathbf{f}}}/p_{\mathbf{A}^*\mathbf{f}^*}}\, e^{-\operatorname{pen}(\hat{\mathbf{f}}) - \frac{\|\hat{\mathbf{A}}_E\|_F}{K}}}{\mathcal{H}(\mathbf{A}^*\mathbf{f}^*, \hat{\mathbf{A}}\hat{\mathbf{f}})}\right] + \log\frac{p_{\mathbf{A}^*\mathbf{f}^*}}{p_{\hat{\mathbf{A}}\hat{\mathbf{f}}}} + 2\operatorname{pen}(\hat{\mathbf{f}}) + 2\frac{\|\hat{\mathbf{A}}_E\|_F}{K}$$

The right hand side of above equation can be bounded above as

$$\begin{aligned} &2\log\left[\frac{\sqrt{p_{\hat{\mathbf{A}}\hat{\mathbf{f}}}/p_{\mathbf{A}^*\mathbf{f}^*}}\, e^{-\operatorname{pen}(\hat{\mathbf{f}}) - \frac{\|\hat{\mathbf{A}}_E\|_F}{K}}}{\mathcal{H}(\mathbf{A}^*\mathbf{f}^*, \hat{\mathbf{A}}\hat{\mathbf{f}})}\right] + \log\frac{p_{\mathbf{A}^*\mathbf{f}^*}}{p_{\hat{\mathbf{A}}\hat{\mathbf{f}}}} + 2\operatorname{pen}(\hat{\mathbf{f}}) + 2\frac{\|\hat{\mathbf{A}}_E\|_F}{K} \\ \text{(D.5)} \quad &\leq 2\log\sum_{\{\mathbf{A}_E, \mathbf{f}\}\in\Gamma}\left[\frac{\sqrt{p_{\mathbf{Af}}/p_{\mathbf{A}^*\mathbf{f}^*}}\, e^{-\operatorname{pen}(\mathbf{f}) - \frac{\|\mathbf{A}_E\|_F}{K}}}{\mathcal{H}(\mathbf{A}^*\mathbf{f}^*, \mathbf{Af})}\right] + \log\frac{p_{\mathbf{A}^*\mathbf{f}^*}}{p_{\hat{\mathbf{A}}\hat{\mathbf{f}}}} + 2\operatorname{pen}(\hat{\mathbf{f}}) + 2\frac{\|\hat{\mathbf{A}}_E\|_F}{K} \end{aligned}$$

Notice that the argument of the expectation in the left hand side of the claimed inequality is only a function of $\mathbf{A}_E^*$ and $\mathbf{f}^*$. Hence, we have $\mathbb{E}_{\mathbf{A}^*,\mathbf{f}^*}[\cdot] = \mathbb{E}_{\mathbf{A}^*,\mathbf{f}^*}[\mathbb{E}_{Y|\mathbf{A}^*,\mathbf{f}^*}[\cdot]]$, where $p_{Y=\mathbf{y}|\mathbf{A}^*,\mathbf{f}^*} = p(\mathbf{y}|\mathbf{A}^*\mathbf{f}^* + \boldsymbol{\lambda}) \overset{\text{def}}{=} p_{\mathbf{A}^*\mathbf{f}^*}$ and $p_{Y=\mathbf{y}|\mathbf{A},\mathbf{f}} = p(\mathbf{y}|\mathbf{Af} + \boldsymbol{\lambda}) = p_{\mathbf{Af}}$. Via the Jensen's inequality, we have the following bound for the right side of above inequality

$$\begin{aligned} &\mathbb{E}_{Y|\mathbf{A}^*,\mathbf{f}^*}\left[\log\sum_{\{\mathbf{A}_E, \mathbf{f}\}\in\Gamma}\left[\frac{\sqrt{p_{\mathbf{Af}}/p_{\mathbf{A}^*\mathbf{f}^*}}\, e^{-\operatorname{pen}(\mathbf{f}) - \frac{\|\mathbf{A}_E\|_F}{K}}}{\mathcal{H}(\mathbf{A}^*\mathbf{f}^*, \mathbf{Af})}\right]\right] \\ &\leq \log\sum_{\{\mathbf{A}_E, \mathbf{f}\}\in\Gamma}\left[\frac{e^{-\operatorname{pen}(\mathbf{f}) - \frac{\|\mathbf{A}_E\|_F}{K}}}{\mathcal{H}(\mathbf{A}^*\mathbf{f}^*, \mathbf{Af})}\mathbb{E}_{Y|\mathbf{A}^*,\mathbf{f}^*}\left[\sqrt{p_{\mathbf{Af}}/p_{\mathbf{A}^*\mathbf{f}^*}}\right]\right] \\ &\leq \log\sum_{\{\mathbf{A}_E, \mathbf{f}\}\in\Gamma}\left[\frac{e^{-\operatorname{pen}(\mathbf{f}) - \frac{\|\mathbf{A}_E\|_F}{K}}}{\mathcal{H}(\mathbf{A}^*\mathbf{f}^*, \mathbf{Af})}\mathbb{E}_{Y|\mathbf{A}^*,\mathbf{f}^*}\left[\sqrt{p_{\mathbf{Af}}/p_{\mathbf{A}^*\mathbf{f}^*}}\right]\right] \\ &= \log\sum_{\{\mathbf{A}_E, \mathbf{f}\}\in\Gamma}\left[e^{-\operatorname{pen}(\mathbf{f}) - \frac{\|\mathbf{A}_E\|_F}{K}}\right] \\ &\leq \log\sum_{\mathbf{f}\in\Gamma_1}\left[e^{-\operatorname{pen}(\mathbf{f})}\right] \leq 0. \end{aligned}$$

By the definition of $\{\hat{\mathbf{A}}_E, \hat{\mathbf{f}}\}$, we have

$$\begin{aligned}
&\mathbb{E}_{Y|\mathbf{A}^*,\mathbf{f}^*}\left[\log\frac{p_{\mathbf{A}^*\mathbf{f}^*}}{p_{\hat{\mathbf{A}}\hat{\mathbf{f}}}} + 2\mathrm{pen}(\hat{\mathbf{f}}) + \frac{2\|\hat{\mathbf{A}}_E\|_F}{K}\right] \\
&\leq \mathbb{E}_{Y|\mathbf{A}^*,\mathbf{f}^*}\left\{\min_{\{\mathbf{A}_E,\mathbf{f}\}\in\Gamma}\left[\log\frac{p_{\mathbf{A}^*\mathbf{f}^*}}{p_{\mathbf{A}\mathbf{f}}} + 2\,\mathrm{pen}(\mathbf{f}) + \frac{2\|\mathbf{A}_E\|_F}{K}\right]\right\} \\
&\leq \min_{\{\mathbf{A}_E,\mathbf{f}\}\in\Gamma}\left[\mathrm{KL}(p(\cdot|\mathbf{A}^*\mathbf{f}^*)||p(\cdot\,|\mathbf{A}\mathbf{f})) + 2\,\mathrm{pen}(\mathbf{f}) + 2\frac{\|\mathbf{A}_E\|_F}{K}\right]. \qquad (D.6)
\end{aligned}$$

By taking expectation on both sides of (D.4) and using the fact that $\mathbb{E}_{\mathbf{A}^*,\mathbf{f}^*}[\cdot\,] = \mathbb{E}_{\mathbf{A}^*,\mathbf{f}^*}[\mathbb{E}_{Y|\mathbf{A}^*,\mathbf{f}^*}[\cdot\,]]$ and (D.6), we have proved the claim. ■

Proof. [**Proof of Theorem 3.6**] Recall the definition $\mathbf{A}^* = \mathbf{A}_0 + \mathbf{A}_E^*$ and $\hat{\mathbf{A}} = \mathbf{A}_0 + \hat{\mathbf{A}}_E$. In the proof, we use the following well-known relationships among matrix norms. For $\mathbf{D} \in \mathbb{R}^{m\times n}$, we have $\|\mathbf{D}\|_2 \leq \|\mathbf{D}\|_F \leq \sqrt{m}\|\mathbf{D}\|_2$ and $\|\mathbf{D}\|_2 \leq \sqrt{mn}\|\mathbf{D}\|_{\max}$, where $\|\mathbf{D}\|_{\max} \overset{\text{def}}{=} \max_{1\leq i\leq m,\,1\leq j\leq n}|\mathbf{D}_{ij}|$. We first establish the following inequality via the triangle inequalities.

$$\begin{aligned}
\|\mathbf{A}^*\mathbf{f}^* - \hat{\mathbf{A}}\hat{\mathbf{f}}\|_2 &= \|(\mathbf{A}_0 + \mathbf{A}_E^*)(\mathbf{f}^* - \hat{\mathbf{f}}) + (\mathbf{A}_E^* - \hat{\mathbf{A}}_E)\hat{\mathbf{f}}\|_2 \\
&\geq \|\mathbf{A}(\mathbf{f}^* - \hat{\mathbf{f}})\|_2 - \|(\mathbf{A}_E^* - \hat{\mathbf{A}}_E)\hat{\mathbf{f}}\|_2. \qquad (D.7)
\end{aligned}$$

Hence,

$$\begin{aligned}
\|\mathbf{A}^*(\mathbf{f}^* - \hat{\mathbf{f}})\|_2 &\leq \|\mathbf{A}^*\mathbf{f}^* - \hat{\mathbf{A}}\hat{\mathbf{f}}\|_2 + \|(\mathbf{A}_E^* - \hat{\mathbf{A}}_E)\hat{\mathbf{f}}\|_2 \\
&\leq \|\mathbf{A}^*\mathbf{f}^* - \hat{\mathbf{A}}\hat{\mathbf{f}}\|_1 + \|(\mathbf{A}^* - \hat{\mathbf{A}})\hat{\mathbf{f}}\|_1, \qquad (D.8)
\end{aligned}$$

where (D.8) follows from properties of L_p norm. On the other hand, by Corollary 3.2, $\tilde{\mathbf{A}}_0$ satisfies the RIP condition

$$(1-\epsilon)\|\mathbf{f} - \hat{\mathbf{f}}\|_2^2 \leq \|\tilde{\mathbf{A}}_0(\mathbf{f} - \hat{\mathbf{f}})\|_2^2 \leq (1+\epsilon)\|\mathbf{f} - \hat{\mathbf{f}}\|_2^2, \qquad (D.9)$$

with probability at least $1 - e\cdot\exp\left(-\frac{c_1\epsilon^2 K}{mn^4}\right)$. As $\|\mathbf{f}_i\|_1 = \|\hat{\mathbf{f}}_i\|_1 = I$, $i = 1,\ldots,k$, we have that $\mathbf{A}_0 = \tilde{\mathbf{A}}_0 + \mathrm{diag}\{\frac{1}{\sqrt{m}}\mathbf{1}_{m\times n},\ldots,\frac{1}{\sqrt{m}}\mathbf{1}_{m\times n}\}$ also satisfies

$$(1-\epsilon)\|\mathbf{f} - \hat{\mathbf{f}}\|_2^2 \leq \|\mathbf{A}_0(\mathbf{f} - \hat{\mathbf{f}})\|_2^2 \leq (1+\epsilon)\|\mathbf{f} - \hat{\mathbf{f}}\|_2^2, \qquad (D.10)$$

with probability at least $1 - e\cdot\exp\left(-\frac{c_1\epsilon^2 K}{mn^4}\right)$. Via the result for perturbed sensing matrix in [11], we can derive the following expression

$$(1-\epsilon')\|\mathbf{f} - \hat{\mathbf{f}}\|_2^2 \leq \|\mathbf{A}(\mathbf{f} - \hat{\mathbf{f}})\|_2^2 \leq (1+\epsilon')\|\mathbf{f} - \hat{\mathbf{f}}\|_2^2, \qquad (D.11)$$

with probability at least $1 - e\cdot\exp\left(-\frac{c_1\epsilon^2 K}{mn^4}\right)$ and $\epsilon' \overset{\text{def}}{=} (1+\epsilon)(1+\epsilon_1)^2 - 1$. Note that by our assumption on the choice of ϵ in the statement of the theorem, we always have $1 - \epsilon' > 0$. Combining (D.8) and (D.11), we have the following inequality on the risk between the true underlying signal $\{\mathbf{A}_E^*, \mathbf{f}^*\}$ and the estimate $\{\hat{\mathbf{A}}_E, \hat{\mathbf{f}}\}$ output by CMLE, with probability at least $1 - e \cdot \exp\left(-\frac{c_1 \epsilon^2 K}{mn^4}\right)$

$$\begin{aligned}
R(\{\hat{\mathbf{A}}_E, \hat{\mathbf{f}}\}, \{\mathbf{A}_E^*, \mathbf{f}^*\})) &= \frac{\|\hat{\mathbf{f}} - \mathbf{f}^*\|_2}{K\|\mathbf{f}^*\|_2} + \frac{\|\hat{\mathbf{A}}_E - \mathbf{A}_E^*\|_F}{K\|\mathbf{A}_E^*\|_F} \\
&\leq \frac{\sqrt{n}\|\hat{\mathbf{f}} - \mathbf{f}^*\|_2}{\sqrt{K}\|\mathbf{f}^*\|_1} + \frac{\|\hat{\mathbf{A}}_E - \mathbf{A}_E^*\|_F}{K\|\mathbf{A}_E^*\|_2} \\
&= \frac{\sqrt{n}\|\hat{\mathbf{f}} - \mathbf{f}^*\|_2}{K\sqrt{KI}} + \frac{\|\hat{\mathbf{A}}_E - \mathbf{A}_E^*\|_F}{K\|\mathbf{A}_E^*\|_2} \\
&\leq \frac{\sqrt{n}\|\mathbf{A}^*\mathbf{f}^* - \hat{\mathbf{A}}\hat{\mathbf{f}}\|_1}{K\sqrt{KI}\sqrt{1-\epsilon'}} + \frac{\sqrt{n}\|(\mathbf{A}^* - \hat{\mathbf{A}})\hat{\mathbf{f}}\|_1}{K\sqrt{KI}\sqrt{1-\epsilon'}} + \frac{\|\hat{\mathbf{A}}_E\|_F + \|\mathbf{A}_E^*\|_F}{K\|\mathbf{A}_E^*\|_2} \\
&\leq \frac{\sqrt{n}\|\mathbf{A}^*\mathbf{f}^* - \hat{\mathbf{A}}\hat{\mathbf{f}}\|_1}{K\sqrt{KI}\sqrt{1-\epsilon'}} + \frac{\sqrt{n}\|(\mathbf{A}^* - \hat{\mathbf{A}})\hat{\mathbf{f}}\|_1}{K\sqrt{KI}\sqrt{1-\epsilon'}} + \frac{\sqrt{m}(\|\hat{\mathbf{A}}_E\|_2 + \|\mathbf{A}_E^*\|_2)}{K\|\mathbf{A}_E^*\|_2} \\
&\leq \frac{\sqrt{n}\|\mathbf{A}^*\mathbf{f}^* - \hat{\mathbf{A}}\hat{\mathbf{f}}\|_1}{K\sqrt{KI}\sqrt{1-\epsilon'}} + \frac{\sqrt{n}\|(\mathbf{A}^* - \hat{\mathbf{A}})\hat{\mathbf{f}}\|_1}{K\sqrt{KI}\sqrt{1-\epsilon'}} + \frac{2\sqrt{m}\epsilon_1}{K\|\mathbf{A}_E^*\|_2}\|\mathbf{\Phi}_0\|_2, \qquad \text{(D.12)}
\end{aligned}$$

where the inequalities directly follow from triangle inequality, the properties of matrix norm and Lemma D.1, respectively.

We first work on the bound for the term $\|\mathbf{A}^*\mathbf{f}^* - \hat{\mathbf{A}}\hat{\mathbf{f}}\|_1$. We establish the following inequalities

$$\begin{aligned}
\|\mathbf{A}^*\mathbf{f}^*\|_1 &\leq \sqrt{Km}\|\mathbf{A}^*\mathbf{f}^*\|_2 \\
&\leq \sqrt{Km}\|\mathbf{A}^*\|_2\|\mathbf{f}^*\|_2 \\
&\leq \sqrt{Km}\|\mathbf{A}^*\|_2\|\mathbf{f}^*\|_1 \\
&\leq KI\sqrt{Km}\|\mathbf{A}^*\|_2 \qquad \text{(D.13)} \\
&\leq KI\sqrt{Km}\|\mathbf{\Phi}^*\|_2 \\
&\leq KI\sqrt{Km}(\|\mathbf{\Phi}_0\|_2 + \|\mathbf{\Phi}_E^*\|_2) \\
&\leq KI\sqrt{Km}(1+\epsilon_1)\|\mathbf{\Phi}_0\|_2,
\end{aligned}$$

where various inequalities follow from the properties of matrix norm and Lemma D.1. Similarly, we can show

$$\|\hat{\mathbf{A}}\hat{\mathbf{f}}\|_1 \leq KI\sqrt{Km}(1+\epsilon_1)\|\mathbf{\Phi}_0\|_2. \qquad \text{(D.14)}$$

We have

$$
\begin{aligned}
\|\mathbf{A}^*\mathbf{f}^* - \hat{\mathbf{A}}\hat{\mathbf{f}}\|_1^2 &= \left(\sum_{i=1}^{Km}\left|\sqrt{(\mathbf{A}^*\mathbf{f}^*+\boldsymbol{\lambda})_i} - \sqrt{(\hat{\mathbf{A}}\hat{\mathbf{f}}+\boldsymbol{\lambda})_i}\right| \cdot \left|\sqrt{(\mathbf{A}^*\mathbf{f}^*+\boldsymbol{\lambda})_i} + \sqrt{(\hat{\mathbf{A}}\hat{\mathbf{f}}+\boldsymbol{\lambda})_i}\right|\right)^2 \\
&\le \sum_{i,j=1}^{Km}\left|\sqrt{(\mathbf{A}^*\mathbf{f}^*+\boldsymbol{\lambda})_i} - \sqrt{(\hat{\mathbf{A}}\hat{\mathbf{f}}+\boldsymbol{\lambda})_i}\right|^2 \cdot \left|\sqrt{(\mathbf{A}^*\mathbf{f}^*+\boldsymbol{\lambda})_j} + \sqrt{(\hat{\mathbf{A}}\hat{\mathbf{f}}+\boldsymbol{\lambda})_j}\right|^2 \qquad \text{(D.15)} \\
&\le 2\sum_{i,j=1}^{Km}\left|\sqrt{(\mathbf{A}^*\mathbf{f}^*+\boldsymbol{\lambda})_i} - \sqrt{(\hat{\mathbf{A}}\hat{\mathbf{f}}+\boldsymbol{\lambda})_i}\right|^2 \cdot \left|(\mathbf{A}^*\mathbf{f}^*+\boldsymbol{\lambda})_j + (\hat{\mathbf{A}}\hat{\mathbf{f}}+\boldsymbol{\lambda})_j\right| \qquad \text{(D.16)} \\
&\le (KI\sqrt{Km}(1+\epsilon_1)\|\boldsymbol{\Phi}_0\|_2 + 4KU)\sum_{i=1}^{km}\left|\sqrt{(\mathbf{A}^*\mathbf{f}^*+\boldsymbol{\lambda})_i} - \sqrt{(\hat{\mathbf{A}}\hat{\mathbf{f}}+\boldsymbol{\lambda})_i}\right|^2, \qquad \text{(D.17)}
\end{aligned}
$$

where (D.15) and (D.16) follow from the Cauchy-Schwartz inequality for the vectors and arithmetic-mean inequality [3]. (D.17) follows from (D.13) and (D.14).

Followed by the Bhattacharyya identity [1] and similar steps in [24], we can show that

$$
\begin{aligned}
\sum_{i=1}^{Km}\left|\sqrt{(\mathbf{A}^*\mathbf{f}^*+\boldsymbol{\lambda})_i} - \sqrt{(\hat{\mathbf{A}}\hat{\mathbf{f}}+\boldsymbol{\lambda})_i}\right|^2 &= -2\log\prod_{i=1}^{Km}\exp\left(-\frac{1}{2}\left[\sqrt{(\mathbf{A}^*\mathbf{f}^*+\boldsymbol{\lambda})_i} - \sqrt{(\hat{\mathbf{A}}\hat{\mathbf{f}}+\boldsymbol{\lambda})_i}\right]\right)^2 \\
&= 2\log\frac{1}{\int\sqrt{p(\mathbf{y}|\mathbf{A}^*\mathbf{f}^*+\boldsymbol{\lambda})p(\mathbf{y}|\hat{\mathbf{A}}\hat{\mathbf{f}}+\boldsymbol{\lambda})d\nu(\mathbf{y})}}, \qquad \text{(D.18)}
\end{aligned}
$$

where ν is the counting measure on $\mathbb{Z}_+^{Km}$. By Lemma D.2, we have

$$
\begin{aligned}
&2\mathbb{E}_{\mathbf{A}^*,\mathbf{f}^*}\left[\log\frac{1}{\int\sqrt{p(\mathbf{y}|\mathbf{A}^*\mathbf{f}^*+\boldsymbol{\lambda})p(\mathbf{y}|\hat{\mathbf{A}}\hat{\mathbf{f}}+\boldsymbol{\lambda})d\nu(\mathbf{y})}}\right] \\
&\le \mathbb{E}_{\mathbf{A}^*,\mathbf{f}^*}\left\{\min_{\{\mathbf{A}_E,\mathbf{f}\}\in\Gamma}\left[\mathrm{KL}(p(\cdot|\mathbf{A}^*\mathbf{f}^*+\boldsymbol{\lambda})||p(\cdot|\mathbf{A}\mathbf{f}+\boldsymbol{\lambda})) + 2\,\mathrm{pen}(\mathbf{f}) + 2\frac{\|\mathbf{A}_E\|_F}{K}\right]\right\}, \qquad \text{(D.19)}
\end{aligned}
$$

and the KL divergence can be bounded as

$$
\begin{aligned}
&\mathrm{KL}(p(\cdot|\mathbf{A}^*\mathbf{f}^*+\boldsymbol{\lambda})||p(\cdot|\mathbf{A}\mathbf{f}+\boldsymbol{\lambda}))\\
&= \sum_{i=1}^{Km}\left[(\mathbf{A}^*\mathbf{f}^*+\boldsymbol{\lambda})_i \log\frac{(\mathbf{A}^*\mathbf{f}^*+\boldsymbol{\lambda})_i}{(\mathbf{A}\mathbf{f}+\boldsymbol{\lambda})_i} - (\mathbf{A}^*\mathbf{f}^*+\boldsymbol{\lambda})_i + (\mathbf{A}\mathbf{f}+\boldsymbol{\lambda})_i\right] \quad \text{(D.20)}\\
&\le \sum_{i=1}^{Km}\left[(\mathbf{A}^*\mathbf{f}^*+\boldsymbol{\lambda})_i \left(\frac{(\mathbf{A}^*\mathbf{f}^*+\boldsymbol{\lambda})_i}{(\mathbf{A}\mathbf{f}+\boldsymbol{\lambda})_i}-1\right) - (\mathbf{A}^*\mathbf{f}^*+\boldsymbol{\lambda})_i + (\mathbf{A}\mathbf{f}+\boldsymbol{\lambda})_i\right] \quad \text{(D.21)}\\
&= \sum_{i=1}^{Km}\left[\frac{1}{(\mathbf{A}\mathbf{f}+\boldsymbol{\lambda})_i}[(\mathbf{A}\mathbf{f}+\boldsymbol{\lambda})_i^2 - 2(\mathbf{A}\mathbf{f}+\boldsymbol{\lambda})_i(\mathbf{A}^*\mathbf{f}^*+\boldsymbol{\lambda})_i + (\mathbf{A}^*\mathbf{f}^*+\boldsymbol{\lambda})_i^2]\right] \quad \text{(D.22)}\\
&\le \frac{Km}{cI}\|\mathbf{A}\mathbf{f}-\mathbf{A}^*\mathbf{f}^*\|_2^2 \quad \text{(D.23)}\\
&= \frac{Km}{cI}\|\mathbf{A}^*(\mathbf{f}^*-\mathbf{f}) - (\mathbf{A}_E^*-\mathbf{A}_E)\mathbf{f}^*\|_2^2 \quad \text{(D.24)}\\
&\le \frac{2Km}{cI}\|\mathbf{A}(\mathbf{f}^*-\mathbf{f})\|_2^2 + \|(\mathbf{A}_E^*-\mathbf{A}_E)\mathbf{f}^*\|_2^2 \quad \text{(D.25)}\\
&\le \frac{2Km}{cI}\left[(1+\epsilon')\|(\mathbf{f}^*-\mathbf{f})\|_2^2 + K^2I^2\|\mathbf{A}_E^*-\mathbf{A}_E\|_F^2\right] \quad \text{(D.26)}
\end{aligned}
$$

where (D.21) follows from the fact that $\log t \le t-1$, $\forall t>0$ and (D.26) follows the properties of matrix norm.

Now we try to bound $\|(\mathbf{A}^*-\hat{\mathbf{A}})\hat{\mathbf{f}}\|_1$. Followed by similar steps from (D.15) to (D.17), we have

$$\|(\mathbf{A}^*-\hat{\mathbf{A}})\hat{\mathbf{f}}\|_1^2 \le (KI\sqrt{Km}(1+\epsilon_1)\|\boldsymbol{\Phi}_0\|_2 + 4KU)\sum_{i,j=1}^{Km}\left|\sqrt{(\mathbf{A}^*\hat{\mathbf{f}}+\boldsymbol{\lambda})_i} - \sqrt{(\hat{\mathbf{A}}\hat{\mathbf{f}}+\boldsymbol{\lambda})_i}\right|^2.$$

Via the Bhattacharyya identity, we derive

$$
\begin{aligned}
\sum_{i,j=1}^{Km}\left|\sqrt{(\mathbf{A}^*\hat{\mathbf{f}}+\boldsymbol{\lambda})_i} - \sqrt{(\hat{\mathbf{A}}\hat{\mathbf{f}}+\boldsymbol{\lambda})_i}\right|^2 &= -2\log\prod_{i=1}^{Km}\exp\left(-\frac{1}{2}\left[\sqrt{(\mathbf{A}^*\hat{\mathbf{f}}+\boldsymbol{\lambda})_i} - \sqrt{(\hat{\mathbf{A}}\hat{\mathbf{f}}+\boldsymbol{\lambda})_i}\right]\right)^2\\
&= 2\log\frac{1}{\int\sqrt{p(\mathbf{y}|\mathbf{A}^*\hat{\mathbf{f}}+\boldsymbol{\lambda})p(\mathbf{y}|\hat{\mathbf{A}}\hat{\mathbf{f}}+\boldsymbol{\lambda})d\nu(\mathbf{y})}}.
\end{aligned}
$$

Followed by similar steps in the proof of Lemma D.2, we can establish

$$
\begin{aligned}
&2\mathbb{E}_{\mathbf{A}^*,\mathbf{f}^*}\left[\log\frac{1}{\int\sqrt{p(\mathbf{y}|\mathbf{A}^*\hat{\mathbf{f}}+\boldsymbol{\lambda})p(\mathbf{y}|\hat{\mathbf{A}}\hat{\mathbf{f}}+\boldsymbol{\lambda})d\nu(\mathbf{y})}}\right]\\
&\le \mathbb{E}_{\mathbf{A}^*,\mathbf{f}^*}\left\{\min_{\{\mathbf{A}_E,\mathbf{f}\}\in\Gamma}\left[\mathrm{KL}(p(\cdot|\mathbf{A}^*\hat{\mathbf{f}}+\boldsymbol{\lambda})||p(\cdot|\mathbf{A}\mathbf{f}+\boldsymbol{\lambda})) + 2\,\mathrm{pen}(\mathbf{f}) + 2\frac{\|\mathbf{A}_E\|_F}{k}\right]\right\},
\end{aligned}
$$

and the KL divergence can be bounded as

$$\begin{aligned}
&\mathrm{KL}(p(\cdot|\mathbf{A}^*\hat{\mathbf{f}}+\boldsymbol{\lambda})||p(\cdot|\mathbf{A}\mathbf{f}+\boldsymbol{\lambda}))\\
&=\sum_{i=1}^{Km}\left[(\mathbf{A}^*\hat{\mathbf{f}}+\boldsymbol{\lambda})_i\log\frac{(\mathbf{A}^*\hat{\mathbf{f}}+\boldsymbol{\lambda})_i}{(\mathbf{A}\mathbf{f}+\boldsymbol{\lambda})_i}-(\mathbf{A}^*\hat{\mathbf{f}}+\boldsymbol{\lambda})_i+(\mathbf{A}\mathbf{f}+\boldsymbol{\lambda})_i\right]\\
&\le\sum_{i=1}^{Km}\left[(\mathbf{A}^*\hat{\mathbf{f}}+\boldsymbol{\lambda})_i\left(\frac{(\mathbf{A}^*\hat{\mathbf{f}}+\boldsymbol{\lambda})_i}{(\mathbf{A}\mathbf{f}+\boldsymbol{\lambda})_i}-1\right)-(\mathbf{A}^*\hat{\mathbf{f}}+\boldsymbol{\lambda})_i+(\mathbf{A}\mathbf{f}+\boldsymbol{\lambda})_i\right] \quad \text{(D.27)}\\
&=\sum_{i=1}^{Km}\left[\frac{1}{(\mathbf{A}\mathbf{f}+\boldsymbol{\lambda})_i}[(\mathbf{A}\mathbf{f}+\boldsymbol{\lambda})_i^2-2(\mathbf{A}\mathbf{f}+\boldsymbol{\lambda})_i(\mathbf{A}^*\hat{\mathbf{f}}+\boldsymbol{\lambda})_i+(\mathbf{A}^*\hat{\mathbf{f}}+\boldsymbol{\lambda})_i^2]\right]\\
&\le\frac{Km}{cI}\|\mathbf{A}\mathbf{f}-\mathbf{A}^*\hat{\mathbf{f}}\|_2^2\\
&=\frac{Km}{cI}\|\mathbf{A}^*(\hat{\mathbf{f}}-\mathbf{f})+(\mathbf{A}^*-\mathbf{A})\mathbf{f}\|_2^2\\
&\le\frac{2Km}{cI}\|\mathbf{A}^*(\hat{\mathbf{f}}-\mathbf{f})\|_2^2+\|(\mathbf{A}_E^*-\mathbf{A}_E)f\|_2^2 \quad \text{(D.28)}\\
&\le\frac{2Km}{cI}\left[(1+\epsilon')\|(\hat{\mathbf{f}}-\mathbf{f})\|_2^2+K^2I^2\|\mathbf{A}_E^*-\mathbf{A}_E\|_F^2\right]. \quad \text{(D.29)}
\end{aligned}$$

Combining above derived inequalities, we have

$$\begin{aligned}
&\min_{\{\mathbf{A}_E,\mathbf{f}\}\in\Gamma}\left[\mathrm{KL}(p(\cdot|\mathbf{A}^*\hat{\mathbf{f}}+\boldsymbol{\lambda})||p(\cdot|\mathbf{A}\mathbf{f}+\boldsymbol{\lambda}))+2\,\mathrm{pen}(\mathbf{f})+2\frac{\|\mathbf{A}_E\|_F}{K}\right]\\
\text{(D.30)}\quad&\le\min_{\{\mathbf{A}_E,\mathbf{f}\}\in\Gamma}\left[\frac{2Km}{cI}[(1+\epsilon')\|(\hat{\mathbf{f}}-\mathbf{f})\|_2^2+K^2I^2\|\mathbf{A}_E^*-\mathbf{A}_E\|_F^2]+2\,\mathrm{pen}(\mathbf{f})+2\frac{\|\mathbf{A}_E\|_F}{K}\right]\\
\text{(D.31)}\quad&\le\min_{\mathbf{A}_E\in\Gamma_2}\left[\frac{2K^3mI}{c}\|\mathbf{A}_E^*-\mathbf{A}_E\|_F^2+2\,\mathrm{pen}(\hat{\mathbf{f}})+2\frac{\|\mathbf{A}_E\|_F}{K}\right]\\
\text{(D.32)}\quad&\le 2P+\min_{\mathbf{A}_E\in\Gamma_2}\left[\frac{2K^3mI}{c}\|\mathbf{A}_E^*-\mathbf{A}_E\|_F^2+2\frac{\|\mathbf{A}_E\|_F}{K}\right],
\end{aligned}$$

where (D.31) follows from the fact that the variables of the minimization argument are separable. Therefore, we derive

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{A}_E^*,\mathbf{f}^*}[R(\{\hat{\mathbf{A}}_E,\hat{\mathbf{f}}\},\{\mathbf{A}_E^*,\mathbf{f}^*\})] \\
&\le \mathbb{E}_{\mathbf{A}_E^*,\mathbf{f}^*}\left[\frac{\sqrt{n}\|\mathbf{A}^*\mathbf{f}^*-\hat{\mathbf{A}}\hat{\mathbf{f}}\|_1}{K\sqrt{K}I\sqrt{1-\epsilon'}}+\frac{\sqrt{n}\|(\mathbf{A}^*-\hat{\mathbf{A}})\hat{\mathbf{f}}\|_1}{K\sqrt{K}I\sqrt{1-\epsilon'}}+\frac{2\sqrt{m}\epsilon_1}{K\|\mathbf{A}_E^*\|_2}\|\mathbf{\Phi}_0\|_2\right] \\
&= \mathbb{E}_{\mathbf{A}_E^*,\mathbf{f}^*}\{\sqrt{\frac{n\sqrt{m}(1+\epsilon_1)\|\mathbf{\Phi}_0\|_2}{(1-\epsilon')IK\sqrt{K}}+\frac{4Un}{K^2I^2(1-\epsilon')}} \\
&\cdot\sqrt{\min_{\{\mathbf{A}_E,\mathbf{f}\}\in\Gamma}\left\{\frac{2Km}{cI}\left((1+\epsilon')\|(\mathbf{f}^*-\mathbf{f})\|_2^2+K^2I^2\|\mathbf{A}_E^*-\mathbf{A}_E\|_F^2\right)+2\,\mathrm{pen}(\mathbf{f})+2\frac{\|\mathbf{A}_E\|_F}{K}\right\}} \\
&+\sqrt{\left(\frac{n\sqrt{m}(1+\epsilon_1)\|\mathbf{\Phi}_0\|_2}{(1-\epsilon')IK\sqrt{K}}+\frac{4Un}{K^2I^2(1-\epsilon')}\right)}\cdot\sqrt{2P+\min_{\{\mathbf{A}_E\}\in\Gamma}\left\{\frac{2K^3mI}{c}\|\mathbf{A}_E^*-\mathbf{A}_E\|_F^2+2\frac{\|\mathbf{A}_E\|_F}{K}\right\}} \\
&+\frac{2\sqrt{m}\epsilon_1}{K\|\mathbf{A}_E^*\|_2}\|\mathbf{\Phi}_0\|_2\}.
\end{aligned}
$$

We may further bound $\|\mathbf{\Phi}_0\|_2$ as $\|\mathbf{\Phi}_0\|_2\le\sqrt{mn}\|\mathbf{\Phi}_0\|_{\max}=2\sqrt{n}$. Finally, we have

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{A}_E^*,\mathbf{f}^*}\left[R(\{\hat{\mathbf{A}}_E,\hat{\mathbf{f}}\},\{\mathbf{A}_E^*,\mathbf{f}^*\})\right] \\
&\le \mathbb{E}_{\mathbf{A}_E^*,\mathbf{f}^*}\left\{\sqrt{C_1\min_{\{\mathbf{A}_E,\mathbf{f}\}\in\Gamma}\left\{\left(\frac{2Km}{cI}\left((1+\epsilon')\|(\mathbf{f}^*-\mathbf{f})\|_2^2+K^2I^2\|\mathbf{A}_E^*-\mathbf{A}_E\|_F^2\right)+2\,\mathrm{pen}(\mathbf{f})+2\frac{\|\mathbf{A}_E\|_F}{K}\right)\right\}}\right. \\
&\left.+\sqrt{C_1\left(2P+\min_{\{\mathbf{A}_E\}\in\Gamma}\left\{\frac{2K^3mI}{c}\|\mathbf{A}_E^*-\mathbf{A}_E\|_F^2+2\frac{\|\mathbf{A}_E\|_F}{K}\right\}\right)}+C_2\frac{\epsilon_1}{\|\mathbf{A}_E^*\|_2}\right\},
\end{aligned}
$$

where $C_1=\left(\frac{n\sqrt{m}(1+\epsilon_1)}{(1-\epsilon')IK\sqrt{K}}+\frac{4Un}{K^2I^2(1-\epsilon')}\right)$ and $C_2=\frac{4\sqrt{mn}}{K}$.

Further, we need to guarantee that each row of $\mathbf{A}_0$ has at least one non-zero entry and this is valid with probability at least $1-m(\frac{1}{2})^n$. Together with the probability guaranteeing the RIP condition as in Corollary 3.2, we have that above inequality holds with probability at least $1-m(\frac{1}{2})^n-e\cdot\exp\left(-\frac{c_1\epsilon^2K}{mn^4}\right)$. ■

Appendix. Proof of Theorem 3.7.

Proof. We first prove the following lemma.

Lemma E.1. *If $\tau_1\ge 2$ and $\tau_2>0$, then*

$$
\begin{aligned}
&2\mathbb{E}_{\mathbf{A}^*,\mathbf{f}^*}\left[\log\frac{1}{\int\sqrt{p(y|\mathbf{A}^*\mathbf{f}^*+\boldsymbol{\lambda})p(y|\hat{\mathbf{A}}\hat{\mathbf{f}}+\boldsymbol{\lambda})}d\nu(y)}\right] \\
&\le \mathbb{E}_{\mathbf{A}^*,\mathbf{f}^*}\left[\min_{\{\mathbf{A}_E,\mathbf{f}\}\in\Gamma}\{\mathrm{KL}(p(\cdot|(\mathbf{A}_0+\mathbf{A}_E^*)\mathbf{f}^*+\boldsymbol{\lambda})||p(\cdot|(\mathbf{A}_0+\mathbf{A}_E)\mathbf{f}+\boldsymbol{\lambda}))+\tau_1\,\mathrm{pen}(\mathbf{f})+\tau_2\|\mathbf{A}_E\|_F\}\right],
\end{aligned}
$$

where ν is the counting measure on $\mathbb{Z}_+^{Km}$, $\boldsymbol{\lambda}=\mathbf{1}_K\otimes\mathbf{u}$ is the known dark-current and $\mathrm{KL}(\cdot||\cdot)$ denotes the Kullback-Leibler distance. The expectation is taken with respect to an arbitrary joint distribution on $\{\mathbf{A},\mathbf{f}\}$.

Proof. [Proof of Lemma E.1] The proof of Lemma E.1 is very similar to Lemma D.2 and the following inequality is used here.

$$
\begin{aligned}
&\mathbb{E}_{Y|\mathbf{A}^*,\mathbf{f}^*}\left[\log \sum_{\{\mathbf{A}_E,\mathbf{f}\}\in\Gamma}\left[\frac{\sqrt{p_{\mathbf{A}\mathbf{f}}/p_{\mathbf{A}^*\mathbf{f}^*}}\, e^{-\frac{\tau_1 \operatorname{pen}(\mathbf{f})}{2}-\frac{\tau_2\|\mathbf{A}_E\|_F}{2}}}{\mathcal{H}(\mathbf{A}^*\mathbf{f}^*,\mathbf{A}\mathbf{f})}\right]\right]\\
&\le \log \sum_{\{\mathbf{A}_E,\mathbf{f}\}\in\Gamma}\left[\frac{e^{-\frac{\tau_1 \operatorname{pen}(\mathbf{f})}{2}-\frac{\tau_2\|\mathbf{A}_E\|_F}{2}}}{\mathcal{H}(\mathbf{A}^*\mathbf{f}^*,\mathbf{A}\mathbf{f})}\mathbb{E}_{Y|\mathbf{A}^*,\mathbf{f}^*}\left[\sqrt{p_{\mathbf{A}\mathbf{f}}/p_{\mathbf{A}^*\mathbf{f}^*}}\right]\right]\\
&\le \log \sum_{\{\mathbf{A}_E,\mathbf{f}\}\in\Gamma}\left[\frac{e^{-\frac{\tau_1 \operatorname{pen}(\mathbf{f})}{2}-\frac{\tau_2\|\mathbf{A}_E\|_F}{2}}}{\mathcal{H}(\mathbf{A}^*\mathbf{f}^*,\mathbf{A}\mathbf{f})}\mathbb{E}_{Y|\mathbf{A}^*,\mathbf{f}^*}\left[\sqrt{p_{\mathbf{A}\mathbf{f}}/p_{\mathbf{A}^*\mathbf{f}^*}}\right]\right]\\
&= \log \sum_{\{\mathbf{A}_E,\mathbf{f}\}\in\Gamma}\left[e^{-\frac{\tau_1 \operatorname{pen}(\mathbf{f})}{2}-\frac{\tau_2\|\mathbf{A}_E\|_F}{2}}\right]\\
&\le \log \sum_{\mathbf{f}\in\Gamma_1}\left[e^{-\operatorname{pen}(\mathbf{f})}\right] \le 0.
\end{aligned}
$$

■

The proof of the main theorem now follows similar steps in the proof of Theorem 3.6 and this lemma. ■

Appendix. MCMC and EM Inference for PG Model.

The model is expressed as

$$
\text{(F.1)}\qquad
\begin{array}{lcllcl}
\mathbf{y}_k & \sim & \operatorname{Pois}\left(\boldsymbol{\Phi}\mathbf{f}_k+\mathbf{u}\right), & \mathbf{f}_k & \sim & \prod_{j=1}^{n}\operatorname{Gamma}(f_{k,j};\ \alpha_f,\ \beta_f),\\
\Phi_{i,j} & \sim & \operatorname{Gamma}(\Phi_{i,j};\ \beta_\Phi\Phi_{0,i,j},\ \beta_\Phi), & \mathbf{u} & \sim & \prod_{i=1}^{m}\operatorname{Gamma}(u_i;\ \alpha_{u,i},\ \beta_{u,i}).
\end{array}
$$

1. MCMC inference

The augmented Poisson model [36] introduces auxiliary variables $g_{\ell,i,j}$ where

$$
\text{(F.2)}\quad g_{k,i,j}\sim\operatorname{Pois}(\Phi_{i,j}f_{k,j}), j=1,\dots,n \quad\text{and}\quad g_{k,i,n+1}\sim\operatorname{Pois}(u_i)\quad\text{with}\ \sum_{j=1}^{n+1}g_{k,i,j}=y_{k,i},
$$

and it states

$$
\begin{aligned}
&(g_{k,i,1},\dots,g_{k,i,n},g_{k,i,n+1})|y_{k,i}\\
\text{(F.3)}\qquad &\sim \operatorname{mult}\left(y_{k,i};\frac{\Phi_{i,1}f_{k,1}}{\sum_{j=1}^{n}\Phi_{i,j}f_{k,j}+u_i},\dots,\frac{\Phi_{i,n}f_{k,n}}{\sum_{j=1}^{n}\Phi_{i,j}f_{k,j}+u_i},\frac{u_i}{\sum_{j=1}^{n}\Phi_{i,j}f_{k,j}+u_i}\right)
\end{aligned}
$$

Thus we first sample the auxiliary variables from the above multinomial distribution. Then by the conjugacy between Gamma and Poisson distribution, we can show that the posteriors are

$$
\text{(F.4)}\qquad
\begin{array}{lcl}
f_{k,j}|- & \sim & \operatorname{Gamma}\left(\alpha_f+\sum_i g_{k,i,j},\ \beta_f+\sum_i\Phi_{i,j}\right)\\
u_i|- & \sim & \operatorname{Gamma}\left(\alpha_{u,i}+\sum_k g_{k,i,n+1},\ \beta_{u,i}+K\right)\\
\Phi_{i,j}|- & \sim & \operatorname{Gamma}\left(\beta_\Phi\Phi_{0,i,j}+\sum_k g_{k,i,j},\ \beta_\Phi+\sum_k f_{k,j}\right)
\end{array}
$$

2. EM inference

The EM inference is readily available once the posteriors are derived. Based on the multinomial distribution (F.3), the expectation of $g_{\ell,i,j}$ is

$$xi_{k,i,j} = \frac{\hat{\Phi}_{i,j}\hat{f}_{k,j}}{\sum_{j=1}^{n}\hat{\Phi}_{i,j}\hat{f}_{k,j}+\hat{u}_i}\cdot y_{k,i}, \quad j=1,\dots,n; \tag{F.5}$$

$$\xi_{k,i,n+1} = \frac{\hat{u}_i}{\sum_{j=1}^{n}\hat{\Phi}_{i,j}\hat{f}_{k,j}+\hat{u}_i}\cdot y_{k,i}, \tag{F.6}$$

where we add ˆ to all variables to denote their current estimates during iteration. The M-step then assigns to the variables the modes of their posteriors in equation (F.4).

$$\hat{f}_{k,j} = \left[\frac{\alpha_f+\sum_{i=1}^{m}\xi_{k,i,j}-1}{\beta_f+\sum_{i=1}^{m}\hat{\Phi}_{i,j}}\right]^{+}, \quad \forall k=1,\dots,K;\ i=1,\dots,m;\ j=1,\dots,n; \tag{F.7}$$

$$\hat{u}_i = \left[\frac{\alpha_{u,i}+\sum_{k=1}^{K}\xi_{k,i,n+1}-1}{\beta_{u,i}+k}\right]^{+}, \quad \hat{\Phi}_{i,j} = \left[\frac{\beta_\Phi\Phi_{0,i,j}+\sum_{k=1}^{K}\xi_{k,i,j}-1}{\beta_\Phi+\sum_{k=1}^{K}\hat{f}_{k,j}}\right]^{+}, \tag{F.8}$$

where $[a]^+$ is understood as $[a]^+ = \begin{cases} a, & a \geq 0 \\ 0, & a < 0 \end{cases}$.

REFERENCES

[1] A. Bhattacharyya, *On a measure of divergence between two statistical populations defined by their probability distributions*, Bulletin of Cal. Math. Soc., 35 (1943), pp. 99–109.

[2] D. J. Brady, *Optical Imaging and Spectroscopy*, John Wiley & Sons, 2009.

[3] I. Bronshtein, K. Semendyayev, G. Musiol, and H. Muehlig, *Handbook of Mathematics*, Springer Science & Business Media, 2007.

[4] E. J. Candès and M. B. Wakin, *An introduction to compressive sampling*, IEEE Signal Processing Magazine, 25 (2008), pp. 21–30.

[5] M. Davenport, P. Boufounos, M. Wakin, and R. Baraniuk, *Signal processing with compressive measurements*, IEEE Journal of Selected Topics in Signal Processing, 4 (2010), pp. 445–460.

[6] M. F. Duarte, M. A.Davenport, D. Takhar, J. N. Laska, S. Ting, K. F. Kelly, and R. G. Baraniuk, *Single-pixel imaging via compressive sampling*, IEEE Signal Processing Magazine, 25 (2008), pp. 83–91.

[7] D. B. Dunson and A. H. Herring, *Bayesian latent variable models for mixed discrete outcomes*, Biostatistics, 6 (2005), pp. 11–25.

[8] A. Eftekhari, H. L. Yap, C. J. Rozell, and M. B. Wakin, *The restricted isometry property for random block diagonal matrices*, Applied and Computational Harmonic Analysis, 35 (2015), pp. 1–31.

[9] J. A. Greenberg, K. Krishnamurthy, and D. J. Brady, *Snapshot molecular imaging using coded energy-sensitive detection*, Optics express, 21 (2013), pp. 25480–25491.

[10] Z. T. Harmany, R. F. Marcia, and R. M. Willett, *This is SPIRAL-TAP: Sparse poisson intensity reconstruction algorithms-theory and practice*, IEEE Transactions on Image Processing, 21 (2012), pp. 1084–1096.

[11] M. A. Herman and T. Strohmer, *General deviants: An analysis of perturbations in compressed sensing*, IEEE Journal of Selected Topics in Signal Processing, 4 (2010), pp. 342–349.

[12] Y. Hitomi, J. Gu, M. Gupta, T. Mitsunaga, and S. K. Nayar, *Video from a single coded exposure photograph using a learned over-complete dictionary*, in IEEE International Conference on In Computer Vision (ICCV), 2011.

[13] W. Johnson and J. Lindenstrauss, *Extensions of Lipschitz mappings into a Hilbert space*, Contemporary Mathematics, 26 (1984), pp. 189–206.

[14] D. Kittle, K. Choi, A. Wagadarikar, and D. J. Brady, *Multiframe image estimation for coded aperture snapshot spectral imagers*, Applied Optics, 49 (2010), pp. 6824–6833.

[15] L. A. Kontorovich and K. Ramanan, *Concentration inequalities for dependent random variables via the martingale method*, The Annals of Probability, 36 (2008), pp. 2126–2158.

[16] M. Ledoux, *The Concentration of Measure Phenomenon*, vol. 89, American Mathematical Soc., 2005.

[17] D. D. Lee and H. S. Seung, *Algorithms for non-negative matrix factorization*, in Advances in Neural Information Processing Systems, 2000.

[18] J. Li and A. Barron, *Mixture density estimation*, in Advances in Neural Information Processing System, 1999.

[19] P. Llull, X. Liao, X. Yuan, J. Yang, D. Kittle, L. Carin, G. Sapiro, and D. J. Brady, *Coded aperture compressive temporal imaging*, Optics Express, (2013).

[20] K. P. MacCabe, A. D. Holmgren, M. P. Tornai, and D. J. Brady, *Snapshot 2d tomography via coded aperture x-ray scatter imaging*, Applied Optics, 52 (2013), pp. 4582–4589.

[21] J. Y. Park, H. L. Yap, C. J. Rozell, and M. B. Wakin, *Concentration of measure for block diagonal matrices with applications to compressive signal processing*, IEEE Transactions on Signal Processing, 59 (2011), pp. 5859–5875.

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

[23] M. Raginsky and I. Sason, *Concentration of measure inequalities in information theory, communications, and coding*, Foundations and Trends in Communications and Information Theory, 10 (2013), pp. 1–246.

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

[25] W. H. Richardson, *Bayesian-based iterative method of image restoration*, Journal of Optical Society of America, 62 (1972), pp. 55–59.

[26] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D: Nonlinear Phenomena, 60 (1992), pp. 259–268.

[27] Q. Tran-Dinh, A. Kyrillidis, and V. Cevher, *Composite self-concordant minimization*, Journal of Machine Learning Research, 16 (2015), pp. 371–416.

[28] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, arXiv preprint arXiv:1011.3027, (2010).

[29] L. Wang, D. Carlson, M. Rodrigues, R. Calderbank, and L. Carin, *A Bregman matrix and the gradient of mutual information for vector Poisson and Gaussian channels*, IEEE Transactions on Information Theory, 60 (2014), pp. 2611–2629.

[30] L. Wang, D. Carlson, M. Rodrigues, D. Wilcox, R. Calderbank, and L. Carin, *Designed measurements for vector count data*, in Advances in neural information processing systems, 2013.

[31] D. S. Wilcox, G.T. Buzzard, B.J. Lucier, P. Wang, and D. Ben-Amotz, *Photon level chemical classification using digital compressive detection*, Analytica Chimica Acta, 755 (2012), pp. 17–27.

[32] R. Willett and R. Nowak, *Multiscale Poisson intensity and density estimation*, IEEE Transactions on Information Theory, 53 (2007), pp. 3171–3187.

[33] Z. Yang, C. Zhang, and L. Xie, *Robustly stable signal recovery in compressed sensing with structured matrix perturbation*, IEEE Transactions on Signal Processing, 60 (2012), pp. 4658–4671.

[34] X. Yuan, P. Llull, X. Liao, J. Yang, G. Sapiro, D. J. Brady, and L. Carin, *Low-cost compressive sensing for color video and depth*, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[35] M. Zhou and L. Carin, *Negative binomial process count and mixture modeling*, IEEE Transactions on Pattern Analysis and Machine Intelligence, (2013).

[36] M. Zhou, L. Hannah, D. Dunson, and L. Carin, *Beta-negative binomial process and poisson factor analysis*, in International Conference on Artificial Intelligence and Statistics, 2012.

[37] H. Zhu, G. Leus, and G.B. Giannakis, *Sparsity-cognizant total least-squares for perturbed compressive sampling*, IEEE Transactions on Signal Processing, 59 (2011), pp. 2002–2016.