Estimation Error Guarantees for Poisson Denoising with Sparse and Structured Dictionary Models

Akshay Soni and Jarvis Haupt
Department of Electrical and Computer Engineering
University of Minnesota – Twin Cities
Minneapolis, MN 55455
Email: {sonix022,jdhaupt}@umn.edu

***Abstract*—Poisson processes are commonly used models for describing discrete arrival phenomena arising, for example, in photon-limited scenarios in low-light and infrared imaging, astronomy, and nuclear medicine applications. In this context, several recent efforts have evaluated Poisson denoising methods that utilize contemporary sparse modeling and dictionary learning techniques designed to exploit and leverage (local) shared structure in the images being estimated. This paper establishes a theoretical foundation for such procedures. Specifically, we formulate sparse and structured dictionary-based Poisson denoising methods as constrained maximum likelihood estimation strategies, and establish performance bounds for their mean-square estimation error using the framework of complexity penalized maximum likelihood analyses.**

I. Introduction

Across a broad range of engineering application domains, Poisson processes have been utilized to describe discrete event or arrival phenomena. For example, in a host of imaging applications (including infrared and thermal imaging, night vision, astronomical imaging, and nuclear medicine, to name a few) the random arrival of photons at each detector in an array may be modeled using Poisson-distributed random variables, with unknown rates or intensities. A fundamental problem in these applications is that of estimating the unknown rates associated with each of the sources, a task typically referred to as *Poisson denoising*.

We consider here a denoising task along these lines. Suppose that we are equipped with a collection of detectors, and that the arrival of photons at each individual detector may be accurately described by a Poisson process with some unknown (non-negative) rate. At each detector we acquire a single integer-valued observation, corresponding to the number of photons arriving at the detector over some fixed (but not necessarily specified) time interval that we assume to be the same across all detectors. It follows that the observation at each detector is a Poisson-distributed random variable whose parameter is the product of the underlying rate parameter of the process and the length of the time interval (see, e.g., [1]). We assume the Poisson processes giving rise to the observations at each detector are mutually independent.

Suppose that there are a total of d detectors. For each $\ell \in [d]$ where $[d]$ is shorthand for the set $\{1, 2, \ldots, d\}$, we denote the Poisson-distributed observation at the ℓ-th detector as y_ℓ and denote by x_ℓ^* its unknown parameter. Letting $\mathrm{Poi}(y_\ell | x_\ell^*) = (x_\ell^*)^{y_\ell} \exp(-x_\ell^*)/(y_\ell)!$ denote the univariate Poisson probability mass function (pmf) defined on nonnegative integers $y_\ell \in \mathbb{N}_0$, we may write the joint pmf of the d observations, defined on $\mathbb{N}_0^d$, as

$$p(\{y_\ell\}_{\ell\in[d]} | \{x_\ell^*\}_{\ell\in[d]}) = \prod_{\ell\in[d]} \mathrm{Poi}(y_\ell | x_\ell^*), \tag{1}$$

where the product form on the right-hand side follows from our independence assumption on the individual Poisson processes.

A. Exploiting Data Structure in Poisson Denoising Tasks

In the absence of any structural dependencies among the collection of rates $\{x_\ell^*\}_{\ell\in[d]}$, the Poisson denoising task is somewhat trivial – in this case, classical estimation theoretic analyses establish that each observation is itself the minimum variance unbiased estimator of its underlying parameter (see, e.g., [2]). More interesting approaches to the denoising task, then, seek to exploit some form of underlying structure among the individual rates. Efforts along these lines include [3], [4], which proposed and analyzed estimation strategies applicable in scenarios where the collection of rates (appropriately arranged) admits a simple representation in terms of a wavelet representation, and [5], which also examined multiresolution representations of the collection of rates. Along similar lines, the work [6] analyzed estimation procedures tailored to signals that are sparse (or nearly so) in any orthonormal basis, within the context of a compressed sensing approach to the Poisson denoising problem.

A number of related efforts have examined Poisson denoising tasks using data representations or bases that are learned from the data themselves, in contrast to the efforts described above that utilize fixed bases or representations. Such "data-driven" estimation strategies include Poisson-specific extensions of classical methods like principal component analysis and other matrix factorization methods [7], [8], as well as application of contemporary ideas from sparse dictionary learning [9]–[11] to Poisson-structured data [12]. We note, in particular, the recent works [13] and [14], which describe estimation tasks employing models that may be described as sparse or structured dictionary-based models; our effort here is motivated by a desire to provide theoretical justification for these dictionary-based techniques.

B. Our Approach

The sparse and structured dictionary-based models upon which our analyses are based describe underlying data structure in terms of matrix factorization models. To that end, we will find it useful here to formulate our model so that the collection of d observations are interpreted as elements of an $m \times n$ matrix (with $d = mn$) denoted by $\mathbf{Y}$, and having elements $Y_{i,j}$, where for $i \in [m]$ and $j \in [n]$, $Y_{i,j}$ is a Poisson random variable with rate $X_{i,j}^*$. Letting $\mathbf{X}^*$ be the $m \times n$ matrix with entries $X_{i,j}^*$, we overload (slightly) the notation in (1), and write the joint pmf of the observations in this case as

$$p(\mathbf{Y}|\mathbf{X}^*) = \prod_{i\in[m], j\in[n]} \mathrm{Poi}(Y_{i,j} | X_{i,j}^*) \triangleq \mathrm{Poi}(\mathbf{Y}|\mathbf{X}^*). \tag{2}$$

Our interest here is primarily on settings where the matrix $\mathbf{X}^*$ admits a dictionary-based factorization, so that $\mathbf{X}^* = \mathbf{D}^*\mathbf{A}^*$, where

This work was supported by DARPA/ONR Award No. N66001-11-1-4090, and by NSF Award CCF-1217751.

$\mathbf{D}^* \in \mathbb{R}^{m\times p}$ and $\mathbf{A}^* \in \mathbb{R}^{p\times n}$. Since such factorization models are themselves fairly general, we restrict our attention here to two specific settings – the first being when the matrix $\mathbf{A}^*$ is sparse so that only a small fraction of its elements are nonzero (along the lines of models employed in dictionary learning efforts), and the second when p, the number of columns of $\mathbf{D}^*$ and rows of $\mathbf{A}^*$, is small relative to m and n (in which case $\mathbf{X}^*$ admits a *low-rank* decomposition). That said, the analytical approach we develop here is fairly general, and thus may readily be extended to other factorization models (e.g., non-negative matrix factorization, structured dictionary models, etc.).

The estimation approaches we analyze here amount to constrained maximum likelihood estimation procedures. Abstractly, we consider a set $\mathcal{X}$ of candidate estimates $\mathbf{X}$ for $\mathbf{X}^*$, each of which admits a factorization of the form $\mathbf{X} = \mathbf{DA}$. The elements of the factors $\mathbf{D}$ and $\mathbf{A}$ may themselves be constrained to enforce the type of structure that we assume present in $\mathbf{X}^*$. Formally, we construct sets $\mathcal{D}$ and $\mathcal{A}$ and a set

$$\mathcal{X} \triangleq \left\{ \mathbf{X} = \mathbf{DA} \ : \mathbf{D} \in \mathcal{D},\ \mathbf{A} \in \mathcal{A},\ \max_{i,j} |X_{i,j}| \le \mathrm{X}_{\max} \right\}$$

where $0 < \mathrm{X}_{\max} < \infty$ is a constant that describes the maximum rate of the underlying processes (and whose specific role will become evident in our analysis), and we consider estimates $\widehat{\mathbf{X}}$ of $\mathbf{X}^*$ constructed according to

$$\widehat{\mathbf{X}} = \arg\min_{\mathbf{X}\in\mathcal{X}} -\log p(\mathbf{Y}|\mathbf{X}) + \lambda\ \mathrm{pen}(\mathbf{X}), \tag{3}$$

where $\mathrm{pen}(\mathbf{X})$ is a non-negative penalty that quantifies the inherent "complexity" of each estimate $\mathbf{X} \in \mathcal{X}$, and $\lambda > 0$ is a user-specified regularization parameter. For both the low-rank and the sparse dictionary based models we consider here, we describe the construction of suitable sets $\mathcal{D}$ and $\mathcal{A}$, cast each corresponding estimation procedure in terms of an optimization of the form (3) (with appropriately constructed penalties), and derive mean-square estimation error rates using analysis techniques motivated by those employed in [5], [6], [15]–[21].

C. Related Efforts in Poisson Restoration

While our focus here is on Poisson denoising, we briefly note several related efforts that examine restoration and deblurring methods for Poisson-distributed data [22]–[26]. These works employ regularized maximum likelihood estimation strategies similar in form to those we analyze in this effort. More recently, [27] proposed a dictionary-based approach to the Poisson deblurring task.

D. Organization and Notation

The remainder of this paper is organized as follows. We present our main theoretical results, stated in terms of the estimation procedures proposed in [13], [14], in Section II, and provide proofs in Section III. In Section IV we briefly discuss how our analytical approach overcomes somewhat limiting minimum rate assumptions inherent in several prior works that use penalized maximum likelihood methods for Poisson denoising. In Section V, we conclude with a discussion of potential extensions of our analysis.

A brief note on notation employed in the sequel – for a matrix $\mathbf{A}$, we denote its number of nonzero elements by $\|\mathbf{A}\|_0$, the sum of absolute values of its elements by $\|\mathbf{A}\|_1$, and its dimension (the product of its row and column dimensions) by $\dim(\mathbf{A})$. For an integer $m \in \mathbb{N}$, the notation $\mathbf{1}_m$ denotes an all-ones length m column vector.

II. Main Results

As noted above, our analyses here are motivated by recent efforts ([13], [14]) that examine Poisson denoising tasks arising in imaging problems and provide empirical evaluations of procedures that exploit local shared structure in the rates being estimated. These prior works each utilize "patch-level" structural models for the underlying image, in which the shared structure arises in terms of factorizations of matrices comprised of vectorized versions of small image patches.

The first procedure proposed in [13] is a non-local variant of a principal component analysis (PCA) method. That approach uses an initial clustering step designed to identify collections of similar patches, then obtains estimates of the underlying rate functions of the image by performing low-rank factorizations of patch-level matrix representations of each data cluster. In terms of our model here, the approximation step inherent to this approach may be described by assuming the true matrix of rates $\mathbf{X}^* \in \mathbb{R}^{m\times n}$ giving rise to independent Poisson-distributed observations $\mathbf{Y}$ in each data cluster admits a decomposition of the form $\mathbf{X}^* = \mathbf{D}^*\mathbf{A}^*$, where $\mathbf{D}^* \in \mathbb{R}^{m\times p}$ and $\mathbf{A}^* \in \mathbb{R}^{p\times n}$ for some $p \le \min(m, n)$.

Both [13] and [14] also examine sparse dictionary-based denoising methods along the lines of recent efforts in the dictionary learning literature (see, e.g., [11]), which seek to model the image patches as sparse linear combinations of columns of a learned dictionary matrix. Here, this model assumes that the true rate matrix $\mathbf{X}^*$ admits a decomposition of the form $\mathbf{D}^*\mathbf{A}^*$ where $\mathbf{A}^*$ is sparse (e.g., having fewer than some $k_{\max}$ non zeros per column). Sparse dictionary-based models may be interpreted as a natural extension of low-rank models; the latter essentially fits the data to a single low-dimensional linear subspace, while the former utilizes a union of linear subspaces.

Our main results establish mean square error guarantees for estimates for these tasks that are obtained via penalized maximum likelihood estimation strategies. In order to state our results, we need to first construct a set $\mathcal{X}$ of candidate reconstructions, with appropriate penalties. To that end, we fix parameters $\mathrm{A}_{\max} > 0$, and $\mathrm{X}_{\max} > 0$, and $\lambda' > 1$, let q be a positive integer satisfying

$$q \ge \max\left\{4, 3 + \log\left(\frac{18\mathrm{A}_{\max}}{\lambda' \log(2)}\right), 1 + \log\left(\frac{36\mathrm{A}_{\max}}{\mathrm{X}_{\max}}\right)\right\}, \tag{4}$$

and let L be the smallest integer exceeding $(\max(m,n))^q$. Now, for any positive integer $p \le \min(m,n)$ we let $\mathcal{X}$ be the set of candidate reconstructions of the form $\mathbf{X} = \mathbf{DA}$ satisfying $\max_{i,j}|X_{i,j}| \le X_{\max}$, where $\mathbf{D} \in \mathcal{D}$ are in $\mathbb{R}^{m\times(p+1)}$ and $\mathbf{A} \in \mathcal{A}$ are in $\mathbb{R}^{(p+1)\times n}$, so that each entry of $\mathbf{D}$ takes values on one of L uniformly-spaced quantization levels in the range $[-1, 1]$ and each element of $\mathbf{A}$ takes on one of L possible uniformly spaced quantization levels in the range $[-\mathrm{A}_{\max},\ \mathrm{A}_{\max}]$.

Our first result, stated here as a theorem, pertains to sparse dictionary-based models.

Theorem II.1. *Let the true rate matrix* $\mathbf{X}^*$ *be* $m \times n$, *where* $\max(m,n) \ge 3$. *Suppose* $\mathbf{X}^*$ *satisfies the constraint* $\max_{i,j} X^*_{i,j} < \mathrm{X}_{\max}/2$, *and admits a dictionary-based decomposition of the form* $\mathbf{D}^*\mathbf{A}^*$, *where the dictionary* $\mathbf{D}^*$ *is* $m \times p$ *for* $p < n$ *with entries bounded in magnitude by* 1, *and the coefficient matrix* $\mathbf{A}^*$ *is* $p \times n$ *whose elements are bounded in magnitude by* $\mathrm{A}_{\max}$. *Let observations* $\mathbf{Y}$ *of* $\mathbf{X}^*$ *be acquired according to the model* (2).

Form the set $\mathcal{X}$ *as above, and let* $\mathrm{pen}(\mathbf{X}) = [q \cdot \dim(\mathbf{D}) + (q+2) \cdot \|\mathbf{A}\|_0] \cdot \log(\max(m,n))$. *The estimate* $\widehat{\mathbf{X}} = \widehat{\mathbf{X}}(\mathbf{Y}) = \widehat{\mathbf{D}}\widehat{\mathbf{A}}$ *formed using the solution of the penalized maximum likelihood problem*

$$\{\widehat{\mathbf{D}}, \widehat{\mathbf{A}}\} = \arg\min_{\mathbf{D}\in\mathcal{D},\mathbf{A}\in\mathcal{A}:\mathbf{DA}\in\mathcal{X}} -\log p(\mathbf{Y}|\mathbf{DA}) + \lambda\|\mathbf{A}\|_0, \tag{5}$$

with $\lambda = \lambda' \cdot (q+2) \cdot \log(\max(m,n)) \log(2)$ *(and where* λ' *is as specified in the construction of* $\mathcal{X}$*) satisfies*

$$\frac{\mathbb{E}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{mn} \preceq \lambda' \mathrm{X}_{\max} \left(\frac{m(p+1)}{mn} + \frac{\|\mathbf{A}^*\|_0 + n}{mn} \right) \log(\max(m,n)).$$

Here, the expectation is with respect to the distribution of $\mathbf{Y}$ *parameterized by the true rate matrix* $\mathbf{X}^*$*, and the notation* $\preceq$ *suppresses leading (finite, positive) constants.*

The salient take-away point here is that the average per-element estimation error is upper bounded by a term that decays essentially in proportion to the number of "degrees of freedom" in the model divided by the number of observations. In other words, our result here establishes that the estimation error exhibits characteristics of the well-known parametric rate.

The result of Thm. II.1 also provides guidance on when dictionary-based estimation procedures are viable. Consider, for example, a setting where the true matrix $\mathbf{A}^*$ in the dictionary-based decomposition of $\mathbf{X}^*$ has some $k_{\max}$ nonzero elements per column. Here, Theorem II.1 establishes that the mean-square estimation error for estimating $\mathbf{X}^*$ decays in proportion to $(p+1)/n + (k_{\max}+1)/m$, ignoring leading constants and logarithmic factors. This result implies natural conditions on the estimation task – that accurate estimation is possible when the number of columns of $\mathbf{X}^*$ exceeds (by a multiplicative constant times a factor logarithmic in the dimension) the number of true dictionary elements p, and the number of rows of $\mathbf{X}^*$ exceeds (by a multiplicative constant times a factor logarithmic in the dimension) the number of non zeros in the sparse representation of each column. This latter condition is reminiscent of conditions arising in compressive sensing (see, e.g., [21], [28], [29]).

We obtain an analogous result for the case where the true rate matrix $\mathbf{X}^*$ admits a low-rank decomposition. We state the result here as a corollary of Theorem II.1.

Corollary II.1. *Suppose that* $\max(m,n) \geq 3$*, and that the true rate matrix* $\mathbf{X}^* \in \mathbb{R}^{m \times n}$ *admits a low-rank decomposition, so that it may be written as* $\mathbf{X}^* = \mathbf{D}^*\mathbf{A}^*$*, where* $\mathbf{D}^*$ *is* $m \times p$ *and* $\mathbf{A}^*$ *is* $p \times n$ *with* $p \leq \min(m,n)$*, and such that* $X^*_{i,j} \leq \mathrm{X}_{\max}/2$*,* $\forall i,j$*. Let observations* $\mathbf{Y}$ *be acquired via the model* (2)*. Form the set* $\mathcal{X}$ *as above, and let* $\mathrm{pen}(\mathbf{X}) = [q \cdot \dim(\mathbf{D}) + (q+2) \cdot \dim(\mathbf{A})] \cdot \log(\max(m,n))$*.*

The estimate $\widehat{\mathbf{X}} = \widehat{\mathbf{X}}(\mathbf{Y}) = \widehat{\mathbf{D}}\widehat{\mathbf{A}}$ *formed using the solution of the following penalized maximum likelihood problem*

$$\{\widehat{\mathbf{D}}, \widehat{\mathbf{A}}\} = \arg \min_{\mathbf{D} \in \mathcal{D}, \mathbf{A} \in \mathcal{A} : \mathbf{DA} \in \mathcal{X}} -\log p(\mathbf{Y}|\mathbf{DA}), \tag{6}$$

satisfies

$$\frac{\mathbb{E}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{mn} \preceq \lambda' \mathrm{X}_{\max} \left(\frac{(p+1)(m+n)}{mn} \right) \log(\max(m,n)),$$

where as above the expectation is with respect to the distribution of $\mathbf{Y}$ *parameterized by the true rate matrix* $\mathbf{X}^*$*, and the notation* $\preceq$ *suppresses leading (finite) constants.*

Note that in this case the penalty $\mathrm{pen}(\mathbf{X})$ is actually the same for all $\mathbf{X} \in \mathcal{X}$, as it depends only on the dimensions of the two factors, which are the same for all candidates by construction of $\mathcal{X}$. Thus, the estimation approach here reduces to just a maximum likelihood estimation over constrained sets. As above, the estimation error rate exhibits characteristics of the parametric rate, as the low-rank model here has $\mathcal{O}(p(m+n))$ degrees of freedom.

III. Proofs of Main Results

We write $p_{X_{i,j}}(\cdot)$ as shorthand for the scalar Poisson pmf with rate $X_{i,j}$, and we denote the multivariate Poisson pmf $p(\cdot|\mathbf{X})$ defined in (2) (parameterized by the collection of rates $\{X_{i,j}\}_{i,j}$) by $p_{\mathbf{X}}(\cdot)$.

Central to our analysis will be the aforementioned countable sets $\mathcal{X}$ of candidate reconstructions of the unknown (non-negative) rate matrix $\mathbf{X}^*$. We consider sets $\mathcal{X}$ constructed as above, and assign to each $\mathbf{X} \in \mathcal{X}$ a non-negative "penalty" quantity denoted by $\mathrm{pen}(\mathbf{X})$ (which here will quantify the "complexity" of the corresponding estimate), so that the collection of penalties satisfies the summability condition $\sum_{\mathbf{X} \in \mathcal{X}} 2^{-\mathrm{pen}(\mathbf{X})} \leq 1$. Note that this condition is just the Kraft-McMillan inequality; in constructing penalties for elements of $\mathcal{X}$ we will employ the well-known fact that the Kraft-McMillan inequality is satisfied provided we may construct a *uniquely decodable code* for the elements $\mathbf{X} \in \mathcal{X}$; see [30]. With this, we begin by establishing a fundamental result, from which our results follow.

Lemma III.1. *Suppose that the elements of the unknown non-negative rate matrix* $\mathbf{X}^*$ *are bounded in amplitude, so that for some fixed* $\mathrm{X}_{\max} > 0$*, we have* $0 \leq X^*_{i,j} \leq \mathrm{X}_{\max}/2$ *for all* $i \in [m]$ *and* $j \in [n]$*. Let* $\mathcal{X}$ *be a countable set of candidate solutions* $\mathbf{X}$ *satisfying the uniform bound* $\max_{i \in [m], j \in [n]} |X_{i,j}| \leq X_{\max}$*, with associated non-negative penalties* $\{\mathrm{pen}(\mathbf{X})\}_{\mathbf{X} \in \mathcal{X}}$ *satisfying the Kraft-McMillan inequality as stated above. Collect a total of* mn *independent Poisson measurements* $\mathbf{Y} = \{Y_{i,j}\}_{i \in [m], j \in [n]}$*, parameterized by* $\mathbf{X}^*$*, according to the model* (2)*. If there exists* $\mathbf{X}^+ \in \mathcal{X}$ *such that* $X^+_{i,j} - X^*_{i,j} \geq 0$ *for all* $i \in [m]$ *and* $j \in [n]$*, then for any choice of* $\lambda' > 1$*, the complexity penalized maximum likelihood estimate*

$$\widehat{\mathbf{X}} = \arg \min_{\mathbf{X} \in \mathcal{X}} \{-\log p(\mathbf{Y}|\mathbf{X}) + \lambda' \log(2) \cdot \mathrm{pen}(\mathbf{X})\}, \tag{7}$$

satisfies,

$$\frac{\mathbb{E}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{mn} \leq \frac{4\mathrm{X}_{\max}}{mn} \left[\|\mathbf{X}^* - \mathbf{X}^+\|_1 + \lambda' \log(2) \cdot \mathrm{pen}(\mathbf{X}^+)\right], \tag{8}$$

where the expectation is taken with respect to the distribution of $\mathbf{Y} \sim p_{\mathbf{X}^*}$*.*

Proof: Our proof utilizes a straight-forward extension of a result stated and utilized in [5], [6] (based on the essential ideas of [18], [19]), which we provide here without proof: for any $\lambda' > 1$ the complexity regularized maximum likelihood solution $\widehat{\mathbf{X}}$ of the form (7), obtained by optimizing over any countable set $\mathcal{X}$ of candidates having penalties $\{\mathrm{pen}(\mathbf{X})\}_{\mathbf{X} \in \mathcal{X}}$ satisfying the Kraft-McMillan inequality, satisfies

$$-2\mathbb{E} \log \mathrm{A}(p_{\mathbf{X}^*}, p_{\widehat{\mathbf{X}}}) \leq \min_{\mathbf{X} \in \mathcal{X}} \left[\mathrm{K}(p_{\mathbf{X}^*}, p_{\mathbf{X}}) + \lambda' \log(2) \cdot \mathrm{pen}(\mathbf{X})\right], \tag{9}$$

where the expectation is with respect to the distribution of $\mathbf{Y} \sim p_{\mathbf{X}^*}$. Here,

$$\mathrm{K}(p_{\mathbf{X}^*}, p_{\mathbf{X}}) \triangleq \sum_{\mathbf{Y} \in \mathbb{N}_0^{m \times n}} \log \left(\frac{p(\mathbf{Y}|\mathbf{X}^*)}{p(\mathbf{Y}|\mathbf{X})} \right) p(\mathbf{Y}|\mathbf{X}^*)$$

denotes the *Kullback-Leibler divergence* (or KL divergence)[1] of $p_{\mathbf{X}}$

[1] Note that the KL divergence is only well-defined here for non-negative $\mathbf{X}$, when the corresponding Poisson pmf $p(\mathbf{Y}|\mathbf{X})$ is well-defined. We make no specific constraint here that each $\mathbf{X} \in \mathcal{X}$ be non-negative, but without loss of generality we may take $\mathrm{K}(p_{\mathbf{X}^*}, p_{\mathbf{X}})$ to be infinite also when $\mathbf{X}$ has any non-negative entries. Further, note the KL divergence is infinite here if for any i, j, $X_{i,j} = 0$ but $X^*_{i,j} \neq 0$ (i.e., when the distribution $p_{\mathbf{X}^*}$ is not absolutely continuous with respect to $p_{\mathbf{X}}$).

from $p_{\mathbf{X}^*}$, and the quantity

$$\mathrm{A}(p_{\mathbf{X}^*}, p_{\widehat{\mathbf{X}}}) \triangleq \sum_{\mathbf{Y} \in \mathbb{N}_0^{m \times n}} \sqrt{p(\mathbf{Y}|\mathbf{X}^*) \cdot p(\mathbf{Y}|\widehat{\mathbf{X}})}$$

is the *Hellinger Affinity* between $p_{\mathbf{X}^*}$ and $p_{\widehat{\mathbf{X}}}$. Now, since the upper bound in (9) holds for $\mathbf{X} \in \mathcal{X}$ which achieves the minimum, it holds for all $\mathbf{X} \in \mathcal{X}$. Considering, specifically, the estimator $\mathbf{X}^+ \in \mathcal{X}$, we have

$$-2\mathbb{E}\log \mathrm{A}(p_{\mathbf{X}^*}, p_{\widehat{\mathbf{X}}}) \leq \mathrm{K}(p_{\mathbf{X}^*}, p_{\mathbf{X}^+}) + \lambda' \log(2) \cdot \mathrm{pen}(\mathbf{X}^+). \quad (10)$$

Specializing to the Poisson case, we use the results of Lemmas III.2 and III.3 (in Section III-C) to obtain, respectively, a lower bound for the left-hand side and an upper bound for the right-hand side of (10). The result follows. ■

Our main results of Section II follow from specializing this result to each of the two structural models. We establish first a proof of the sparse dictionary-based inference estimation procedure; the analogous result for estimation in low-rank models follows as a simple corollary.

A. *Proof of Theorem II.1*

The proof of our first main result follows directly from Lemma III.1 above. First, note that each candidate estimate $\mathbf{X} = \mathbf{DA} \in \mathcal{X}$ may be described via a code, in which each element of $\mathbf{D}$ is encoded using $\log(L) = q\log(\max(m,n))$ bits and each nonzero element of $\mathbf{A}$ is encoded using $\log(\dim(\mathbf{A}))$ bits to denote its location, and $\log(L)$ bits for its amplitude. Thus, a total of $q \cdot \dim(\mathbf{D}) \cdot \log(\max(m,n))$ bits suffice to encode $\mathbf{D}$, and since $\log(\dim(\mathbf{A})) < \log(\max(m,n)^2)$, matrices $\mathbf{A}$ having $\|\mathbf{A}\|_0$ nonzero entries can be described using no more than $\|\mathbf{A}\|_0 \cdot (q+2) \cdot \log(\max(m,n))$ bits. Overall, this implies we may choose $\mathrm{pen}(\mathbf{X}) = q \cdot \dim(\mathbf{D}) \cdot \log(\max(m,n)) + \|\mathbf{A}\|_0 \cdot (q+2) \cdot \log(\max(m,n))$. Note that while constructing the codes we did not care about the uniform bounded condition (i.e., that each entry should be bounded by $\mathrm{X}_{\max}$); in effect, we formed uniquely decodable codes for a bigger set $\mathcal{X}'$ such that $\mathcal{X} \subseteq \mathcal{X}'$, so we always have $\sum_{\mathbf{X} \in \mathcal{X}} 2^{-\mathrm{pen}(\mathbf{X})} \leq \sum_{\mathbf{X} \in \mathcal{X}'} 2^{-\mathrm{pen}(\mathbf{X})} \leq 1$.

Now, consider a candidate reconstruction of the form $\mathbf{X}_Q = \mathbf{D}_Q\mathbf{A}_Q + \mathbf{1}_m(\alpha\mathbf{1}_n^T) \triangleq \tilde{\mathbf{D}}_Q\tilde{\mathbf{A}}_Q$, where $\mathbf{D}_Q$ and $\mathbf{A}_Q$ are the closest quantized surrogates of the true parameters $\mathbf{D}^*$ and $\mathbf{A}^*$, and $0 \leq \alpha \leq A_{\max}$ is a quantity to be specified. Denote $\mathbf{D}_Q = \mathbf{D}^* + \triangle_{\mathbf{D}}$ and $\mathbf{A}_Q = \mathbf{A}^* + \triangle_{\mathbf{A}}$, where $\triangle_{\mathbf{D}}$ and $\triangle_{\mathbf{A}}$ are the quantization error matrices. Then, it is easy to see that

$$\tilde{\mathbf{D}}_Q\tilde{\mathbf{A}}_Q - \mathbf{D}^*\mathbf{A}^* = \mathbf{1}_m(\alpha\mathbf{1}_n^T) + \mathbf{D}^*\triangle_{\mathbf{A}} + \triangle_{\mathbf{D}}\mathbf{A}^* + \triangle_{\mathbf{D}}\triangle_{\mathbf{A}}. \quad (11)$$

To satisfy the conditions of Lemma III.1, we must have that $\mathbf{X}_Q$ overestimates (element-wise) the true rate matrix, and that the right-hand side of (11) be no larger than $\mathrm{X}_{\max}/2$. To that end, our aim is to choose α so that the right side of (11) becomes element-wise nonnegative, but no larger than $\mathrm{X}_{\max}/2$. It is straightforward to see that each entry of the matrices $\mathbf{D}^*\triangle_{\mathbf{A}}$ and $\triangle_{\mathbf{D}}\mathbf{A}^*$ is bounded in magnitude by $2p\mathrm{A}_{\max}/L$. Also, the elements of the matrix $\triangle_{\mathbf{D}}\triangle_{\mathbf{A}}$ are bounded in magnitude by $4p\mathrm{A}_{\max}/L^2 \leq 4p\mathrm{A}_{\max}/L$. Thus, it suffices to choose α as the smallest quantization level exceeding $8p\mathrm{A}_{\max}/L$ to ensure the each element of the matrix on the right-hand side of (11) is nonnegative. Since we choose α to be the higher quantization level of $8p\mathrm{A}_{\max}/L$, and the quantization levels for elements of $\mathbf{A}$ are of size $2\mathrm{A}_{\max}/L$, we have that $\alpha \leq (8p+2)\mathrm{A}_{\max}/L$. In order for α to be a valid entry of $\mathbf{A}$, it must be bounded by $\mathrm{A}_{\max}$, which is true whenever $L \geq (8p+2)$.

We can now bound each entry of $\tilde{\mathbf{D}}_Q\tilde{\mathbf{A}}_Q - \mathbf{D}^*\mathbf{A}^*$ as follows

$$\begin{aligned}
&(\tilde{\mathbf{D}}_Q\tilde{\mathbf{A}}_Q - \mathbf{D}^*\mathbf{A}^*)_{i,j} \\
&\quad = (\mathbf{1}_m(\alpha\mathbf{1}_n)^T + \mathbf{D}^*\triangle_{\mathbf{A}} + \triangle_{\mathbf{D}}\mathbf{A}^* + \triangle_{\mathbf{D}}\triangle_{\mathbf{A}})_{i,j} \\
&\quad \leq \frac{(8p+2)\mathrm{A}_{\max}}{L} + \frac{2p\mathrm{A}_{\max}}{L} + \frac{2p\mathrm{A}_{\max}}{L} + \frac{4p\mathrm{A}_{\max}}{L} \\
&\quad = \frac{16p\mathrm{A}_{\max}}{L} + \frac{2\mathrm{A}_{\max}}{L} \leq \frac{18p\mathrm{A}_{\max}}{L},
\end{aligned}$$

where the second inequality follows from bounds on the entries of each matrix mentioned above and the last inequality is valid for $p \geq 1$. This quantity is no larger than $\mathrm{X}_{\max}/2$ whenever $L \geq 36p\mathrm{A}_{\max}/\mathrm{X}_{\max}$, and in this case, we ensure that $\mathbf{X}_Q \in \mathcal{X}$.

Now, note that $\|\mathbf{X}^* - \mathbf{X}_Q\|_1 = \sum_{i\in[m],j\in[n]}(\tilde{\mathbf{D}}_Q\tilde{\mathbf{A}}_Q - \mathbf{D}^*\mathbf{A}^*)_{i,j} \leq 18p \cdot (mn) \cdot \mathrm{A}_{\max}/L$, and if we now evaluate the oracle bound (8) from Lemma III.1 at the candidate $\mathbf{X}_Q$ which overestimates $\mathbf{X}^*$ (entry-wise), we have

$$\begin{aligned}
&\frac{\mathbb{E}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{mn} \\
&\qquad \leq \frac{4\mathrm{X}_{\max}}{mn}\left[\|\mathbf{X}^* - \mathbf{X}_Q\|_1 + \lambda'\log(2) \cdot \mathrm{pen}(\mathbf{X}_Q)\right] \\
&\qquad \leq \frac{72pX_{\max}\mathrm{A}_{\max}}{L} + \lambda' \cdot 4\log(2)\mathrm{X}_{\max} \cdot \frac{\mathrm{pen}(\mathbf{X}_Q)}{mn} \\
&\qquad \leq \lambda' \cdot 8\log(2)\mathrm{X}_{\max} \cdot \frac{\mathrm{pen}(\mathbf{X}_Q)}{mn},
\end{aligned}$$

where the last line follows whenever $L \geq \frac{18\mathrm{A}_{\max}mnp}{\lambda'\log(2)}$ (since $\mathrm{pen}(\mathbf{X}_Q)$ corresponds to a binary code having length greater than 0, we have $\mathrm{pen}(\mathbf{X}_Q) \geq 1$).

Overall, then, the result follows since by construction, we have $\dim(\widetilde{\mathbf{D}}_Q) \leq mp+m$, and $\|\widetilde{\mathbf{A}}_Q\|_0 \leq \|\mathbf{A}^*\|_0 + n$, and the assumption (4) implies

$$L \geq \max\left\{8p+2, \frac{18\mathrm{A}_{\max}mnp}{\lambda'\log(2)}, \frac{36p\mathrm{A}_{\max}}{\mathrm{X}_{\max}}\right\}.$$

B. *Proof of Corollary II.1*

The proof of Corollary II.1 follows directly from the proof of Theorem II.1 – in particular, by substituting $\|\mathbf{A}^*\|_0 = pn$.

C. *Useful Lemmata*

The following lemmata are used in the proof of Lemma III.1.

Lemma III.2 (From [6]). *For any two (non-negative) Poisson rate matrices $\mathbf{X}^a$ and $\mathbf{X}^b$, having entries uniformly bounded above by $\mathrm{X}_{\max}$, we have*

$$\frac{1}{4\mathrm{X}_{\max}}\|\mathbf{X}^a - \mathbf{X}^b\|_F^2 \leq -2 \cdot \log \mathrm{A}(p_{\mathbf{X}^a}, p_{\mathbf{X}^b}).$$

Lemma III.3. *For non-negative Poisson rate matrices $\mathbf{X}^a$ and $\mathbf{X}^b$ such that $\mathbf{X}^b$ over-estimates $\mathbf{X}^a$ element-wise i.e., $X^b_{i,j} - X^a_{i,j} \geq 0$ for all $i \in [m]$ and $j \in [n]$, we have $\mathrm{K}(p_{\mathbf{X}^a}, p_{\mathbf{X}^b}) \leq \|\mathbf{X}^b - \mathbf{X}^a\|_1$.*

Proof: By independence and the definition of the KL divergence,

$$\begin{aligned}
\mathrm{K}(p_{\mathbf{X}^a}, p_{\mathbf{X}^b}) &= \sum_{i\in[m],j\in[n]}\left[X^a_{i,j}\log\frac{X^a_{i,j}}{X^b_{i,j}} + X^b_{i,j} - X^a_{i,j}\right] \\
&\leq \sum_{i\in[m],j\in[n]}\left[X^b_{i,j} - X^a_{i,j}\right] = \|\mathbf{X}^a - \mathbf{X}^b\|_1,
\end{aligned}$$

where the inequality follows from the fact that $X^a_{i,j}\log\frac{X^a_{i,j}}{X^b_{i,j}} \leq 0$ since $X^b_{i,j} \geq X^a_{i,j}$ (and following standard convention that $a\log(a/0) = \infty, 0\log(0/a) = 0$ for $a > 0$). ■

IV. Discussion

It is worthwhile to explicitly point out a unique point in our analysis – introducing the additional dimension in the model to ensure that our class of candidate solutions contains an element that always overestimates, element-wise, the rates in the true parameter matrix $\mathbf{X}^*$ – enables us to obtain estimation error rates without making any assumptions on the *minimum* rate of the underlying Poisson processes. This is a significant contrast with prior efforts employing penalized maximum likelihood analyses (but with different structural models) on Poisson-distributed data [5], [6], each of which prescribe adopting an assumption that the rates associated with each Poisson-distributed observation be strictly bounded away from 0.

Our extension here is an important advance, especially in the context of extremely photon-limited scenarios. Indeed, in these settings it is somewhat counter-intuitive (or at least, restrictive) to assume that the rates be bounded away from zero, as it is precisely in these scenarios when one might be most interested in estimating rates that are very near zero. Further, classical analyses suggest that there may be no *fundamental* reason why zero or nearly-zero rates become more difficult to estimate. For instance, in the scalar Poisson rate estimation problem, the Cramer-Rao lower bound for estimating a Poisson rate parameter from n iid $\text{Poi}(\cdot|\theta)$ observations (achievable with the sample average estimator) is θ/n, suggesting that the estimation problem actually becomes easier as the rate decreases. The analytical framework we develop here facilitates analysis of these important low-rate cases under sparse and structured data model assumptions.

Finally, we note that Poisson models also find utility other application domains beyond imaging. In networking tasks, for example, Poisson processes are a natural choice to model arrival events, such as packets arriving at each of a number of network routers our flows across network links (see, e.g., [31]). Our techniques and analysis here would extend directly to other application domains, as well.

V. Conclusions

In this paper we described a framework for quantifying the mean-square error of constrained maximum likelihood Poisson denoising strategies, in settings where the collection of underlying rates (appropriately arranged) admits a low-rank or sparse dictionary-based decomposition. We established that, in these cases, the mean-square estimation error exhibits characteristics of the familiar parametric rate, in that the error essentially takes the form of "degrees of freedom" divided by "number of observations." In analogy to related analyses in [6], [21], our analysis can also be used to obtain error rates for data adhering to models that are not exactly sparse, but instead are characterized by coefficients whose ordered amplitudes decay (e.g., at a polynomial rate). Finally, while our analysis here was formulated in terms of matrix-structured data and factorization models, these methods may be extended straightforwardly to encompass also sparse and low-rank models for higher-order tensor structure data. We defer in-depth investigations of these extensions to a future effort.

References

[1] J. A. Gubner, *Probability and random processes for electrical and computer engineers*, Cambridge University Press, 2006.

[2] S. M. Kay, *Fundamentals of Statistical signal processing, Volume 1: Estimation Theory*, Prentice Hall PTR, 1993.

[3] R. D. Nowak and R. G. Baraniuk, "Wavelet-domain filtering for photon imaging systems," *IEEE Trans. Image Processing*, vol. 8, no. 5, pp. 666–678, 1999.

[4] K. E. Timmermann and R. D. Nowak, "Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging," *IEEE Trans. Information Theory*, vol. 45, no. 3, pp. 846–862, 1999.

[5] E. D. Kolaczyk and R. D. Nowak, "Multiscale likelihood analysis and complexity penalized estimation," *Ann. Statist.*, pp. 500–527, 2004.

[6] M. Raginsky, R. M. Willett, Z. T. Harmany, and R. F. Marcia, "Compressed sensing performance bounds under Poisson noise," *IEEE Trans. Signal Processing*, vol. 58, no. 8, pp. 3990–4002, 2010.

[7] M. Collins, S. Dasgupta, and R. E. Schapire, "A generalization of principal components analysis to the exponential family," in *Advances in neural information processing systems*, 2001, pp. 617–624.

[8] A. P. Singh and G. J. Gordon, "A unified view of matrix factorization models," in *Machine Learning and Knowledge Discovery in Databases*, pp. 358–373. 2008.

[9] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vision Research*, vol. 37, pp. 3311–3325, 1997.

[10] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Proc.*, vol. 54, no. 11, pp. 4311–4322, 2006.

[11] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proc. ICML*, 2009.

[12] P. Chainais, "Towards dictionary learning from images with non Gaussian noise," in *IEEE Intl. Workshop on Machine Learning for Signal Processing*, 2012, pp. 1–6.

[13] J. Salmon, Z. Harmany, C.-A. Deledalle, and R. Willett, "Poisson noise reduction with non-local PCA," *Journal of Mathematical Imaging and Vision*, April 2013.

[14] R. Giryes and M. Elad, "Sparsity based Poisson denoising with dictionary learning," *arXiv preprint arXiv:1309.4306*, 2013.

[15] A. R. Barron, "Complexity regularization with application to artificial neural networks," in *Nonparametric functional estimation and related topics*, pp. 561–576. Springer, 1991.

[16] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Trans. Information Theory*, vol. 37, no. 4, pp. 1034–1054, 1991.

[17] A. Barron, L. Birgé, and P. Massart, "Risk bounds for model selection via penalization," *Probability theory and related fields*, vol. 113, no. 3, pp. 301–413, 1999.

[18] Q. J. Li, *Estimation of mixture models*, Ph.D. thesis, Yale University, Dept. of Statistics, 1999.

[19] Q. J. Li and A. R. Barron, "Mixture density estimation," in *Advances in Neural Information Processing Systems*, 2000, vol. 12, pp. 279–285.

[20] T. Zhang, "On the convergence of MDL density estimation," in *Learning Theory*, pp. 315–330. Springer, 2004.

[21] J. Haupt and R. Nowak, "Signal reconstruction from noisy random projections," *IEEE Trans. Information Theory*, vol. 52, no. 9, pp. 4036–4048, 2006.

[22] F.-X. Dupé, J. M. Fadili, and J.-L. Starck, "A proximal iteration for deconvolving Poisson noisy images using sparse representations," *IEEE Trans. Image Processing*, vol. 18, no. 2, pp. 310–321, 2009.

[23] M. A. T. Figueiredo and J. M. Bioucas-Dias, "Restoration of Poissonian images using alternating direction optimization," *IEEE Trans. Image Processing*, vol. 19, no. 12, pp. 3133–3145, 2010.

[24] S. Setzer, G. Steidl, and T. Teuber, "Deblurring Poissonian images by split Bregman techniques," *Journal of Visual Communication and Image Representation*, vol. 21, no. 3, pp. 193–199, 2010.

[25] M. Carlavan and L. Blanc-Féraud, "Sparse Poisson noisy image deblurring," *IEEE Trans. Image Processing*, vol. 21, no. 4, pp. 1834–1846, 2012.

[26] Z. T. Harmany, R. F. Marcia, and R. M. Willett, "This is SPIRAL-TAP: Sparse Poisson intensity reconstruction algorithms – theory and practice," *IEEE Trans. Image Processing*, vol. 21, no. 3, pp. 1084–1096, 2012.

[27] L. Ma, L. Moisan, J. Yu, and T. Zeng, "A dictionary learning approach for Poisson image deblurring," *IEEE Trans. Medical Imaging*, vol. 32, no. 7, pp. 1277–1289, 2013.

[28] D. Donoho, "Compressed sensing," *IEEE Trans. Information Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[29] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[30] T. M. Cover and J. A. Thomas, *Elements of information theory*, John Wiley & Sons, 2012.

[31] Y. Vardi, "Network tomography: Estimating source-destination traffic intensities from link data," *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 365–377, 1996.