Sparse Signal Recovery under Poisson Statistics

Delaram Motamedvaziri, Mohammad H. Rohban, and Venkatesh Saligrama
ECE Department Boston University
Email: {deli, mhrohban, srv}@bu.edu

Abstract—We are motivated by problems that arise in a number of applications such as explosives detection and online Marketing, where the observations are governed by Poisson statistics. Here each observation is a Poisson random variable whose mean is a sparse linear superposition of known patterns. Unlike many conventional problems observations here are not identically distributed since they are associated with different sensing modalities. We analyse the performance of a Maximum Likelihood (ML) decoder, which for our Poisson setting is computationally tractable. We derive fundamental sample complexity bounds for sparse recovery in the high-dimensional setting. We show that when the sensing matrix satisfies the so-called Restricted Eigenvalue (RE) condition the $\ell_1$ regularized ML decoder is consistent. Moreover, it converges exponentially fast in terms of number of observations. Our results apply to both deterministic and random sensing matrices and we present several results for both cases.

Index Terms—Poisson Model Selection, Sparse Recovery, Regularized Maximum Likelihood

I. INTRODUCTION

In this paper, we study the problem of high dimensional sparse model estimation under Poisson model for observations. This problem is motivated by many practical applications where the observations are the counts of an event. The mean count in these applications depends linearly on a sparse subset of parameters. Our goal is to extract the sparse subset from a potentially large number of parameters. Some of these practical applications include explosive identification based on photon counts received in fluorescence based methods [1], and eMarketing based on website traffic [2].

In fluorescence based explosive identification, observations are the number of photon counts observed when an unknown mixture of explosives is exposed to different fluorophores. In general the mixture is sparse, namely, only a few basic explosives constitute the mixture. The goal is to identify how much of each explosive is contained in the mixture.

In the eMarketing application, the goal is to find the best advertising website that brings traffic to a business website. Due to the variety and the high cost of link purchases, businesses are interested in discovering a small number of dominant advertisement websites that direct online traffic to their websites. In this application, observations are the weekly website traffic (for different entities within a similar market) and is modeled using Poisson statistics. Specifically, this traffic is modeled as a superposition of independent traffic emanating from various advertisement websites. The goal is to find dominant advertisement websites. This requires recovering the portion of business traffic that each advertisement website provides.

We propose a general model that is applicable to a broad class of problems involving Poisson statistics. We consider the case where observations are obtained from heterogeneous sensors or different measurement settings and therefore not identically distributed. To simplify the model, we assume that the rates of the underlying Poisson model for observations are affine functions of some positive signal we want to estimate. In other words, if the signal of interest is $w^* \in \mathbb{R}_+^n$, the $i$-th observation, $y_i$, is distributed as follows:

$$\forall i \in \{1, \ldots, n\} : y_i \sim \text{Poisson}(\lambda_{0,i} + a_i^T w^*)$$

where $\lambda_{0,i}$ is the rate of the background Poisson noise and each $a_i = [a_{i,1}, \ldots, a_{i,p}]^T$ is a distinct vector corresponding to the $i$-th sensor. The collection of these vectors form the sensing matrix, $A = [a_1, \ldots, a_n]^T$. Our goal is to recover the sparse vector, $w^*$, from $\{y_1, \ldots, y_n\}$.

In explosive identification example, $y_i$ and $\lambda_{0,i}$ are the photon counts and background emission for fluorophore $i$. $a_{ij}$ is the quenching effect of explosive $j$ on fluorophore $i$, and $w_j$ is the weight of explosive $j$ in the mixture.

In eMarketing example, $y_i$ is the weekly online traffic for business website $i$. $\lambda_{0,i}$ is the average traffic that visits website $i$ directly (not through intermediate advertisement website), $a_{i,j}$ is the number of backward links that business website $i$ has bought from advertisement website $j$, and $w_j$ measures popularity/dominance of advertisement website $j$.

Fig. 1 illustrates this eMarketing model:

In this paper we analyse the performance of a Maximum
Likelihood (ML) decoder. The ML decoder in our Poisson setting is a convex optimization problem and is computationally tractable. We derive fundamental sample complexity bounds for sparse recovery in the high-dimensional setting. We show that when the sensing matrix, $A$ satisfies the so-called Restricted Eigenvalue (RE) condition the $\ell_1$ regularized ML decoder is consistent. Moreover, our estimates converge exponentially fast in terms of number of observations. Our results apply to both deterministic and random sensing matrices and we present several results for both cases. We also conduct several synthetic and real-world experiments and demonstrate the efficacy of our method. Specifically, it has been suggested in the literature \cite{10} that LASSO can handle exponential family noise such as that arises in our application. It turns out that our solution uniformly performs better than LASSO and has significantly superior performance in many interesting regimes.

The paper is organized as follows: In section \[4\] we introduce our notations and state our sparse estimation problem. Section \[3\] is dedicated to our theoretical results on the convergence of the regularized ML decoder. The numeric results under specific settings are demonstrated in sections \[5\]. Finally, the detailed proof of the main theorems and lemmas are provided in section \[6\] and \[7\].

\textbf{A. Related Work}

The proof of consistency of maximum-likelihood estimators for most conventional estimation methods under Poisson Statistics \cite{4, 5} hinges on identically distributed observations and it is unclear whether their results apply to our setting. As we argued our problem involves non-identical observations. For instance different observations correspond to different fluorophores etc. This requirement distinguishes our work from most of the previous Poisson parameter estimation studies such as \cite{6, 7, 8, 9}.

Parameter estimation for non-identical Poisson distributions, as a member of exponential family, has been studied in the context of Generalized Linear Models (GLMs). However, our model is inherently different from the exponential family of GLM models that has been studied in \cite{6, 7, 8}. In particular the GLM model has observations distributed as follows:

\begin{equation}
\text{Model I : } \Pr(y_i = k) \propto (a_i^\top w) \exp(k a_i^\top w)
\end{equation}

In contrast our observations are distributed as follows:

\begin{equation}
\text{Model II : } \Pr(y_i = k) \propto (a_i^\top w)^k \exp(a_i^\top w)
\end{equation}

Due to this difference in the models, the techniques developed in the context of GLM models do not apply to our setting. We will summarize some of the work in the context of GLM for the purpose of reference here to draw out some of the technical aspects of our work. \cite{6} studies the robustness of regularized ML under Model I in non asymptotic settings. On the other hand, \cite{8} provides asymptotic results for slow (sub exponential) growth of the number of dimensions with no sparsity assumption on the latent parameter. Finally, \cite{7} studies the asymptotic behavior of regularized ML with sparsity assumption for Model I. In addition \cite{7} requires the data distribution to be sub-gaussian and that the data be identically distributed. A unified framework for analysis of regularized $M-$ estimators in high dimensions is provided in \cite{9}. Generalization of this unified framework to GLMs is briefly mentioned. They described “Strong Convexity” as a natural setting to study estimator instances under Model I. The statistical aspects in that work requires identically distributed samples and assumes that the components of the sensing matrix is characterized by sub-Gaussian distributions.

In summary at a high-level our models are different. At a technical level our observations are not identically distributed and the components of our sensing matrix do not satisfy sub-Gaussian properties required in much of the literature \cite{6, 7, 8}. These reasons require new techniques of analysis which is the subject of this paper.

Compressive sensing problem under Poisson statistics has been studied in \cite{10}, where a related problem setting is introduced

$$y_i \sim \text{Poisson}(a_i^\top w^*)$$

The main focus of \cite{10} is to provide a lower bound on error for Maximum Likelihood estimation of sparse signal based on non-identical Poisson distributed data. However, no results were provided for consistency, achieveability or convergence. Another distinguishing feature of that work is that an additional constraint, namely, a so called Flux-preserving property, arises in their application. However, no such constraint arises in our setting and consequently requires new methods for analysis.

\textbf{II. PROBLEM SETUP}

\textbf{A. Notations}

We will use the following notation in this paper. Recall that $a_i$’s are the rows of the sensing matrix $A$ and $\lambda_{0,i} + a_i^\top w$ is the Poisson rate of the $i$th sensor.

- $a_{max} = \max_{i,j} |a_{i,j}|$
- $a_{min} = \min_{i,j} |a_{i,j}|$
- $\|a_{max}\|_2 = \max_j |a_{i,j}|$
- $\lambda_{w,i} = \lambda_{0,i} + a_i^\top w$
- $\lambda_{max} = \max_i \lambda_{w,i}$
- $\lambda_{min} = \min_i \lambda_{w,i}$
- $\lambda_0 = \max_i \lambda_{0,i}$

To simplify the analysis of the behavior of ML estimator, we also assume that $\lambda_{w,i}$ is both lower and upper bounded.

$$1 < \lambda_{min} \leq \lambda_{w,i} \leq \lambda_{max}$$

\textbf{B. Problem Formulation}

Consider $n$ independent Poisson distributed observations generated as:

$$\forall i \in \{1, \ldots, n\} : y_i \sim \text{Poisson}(\lambda_{w^*,i})$$

where $\lambda_{w,i}$ is a sparse linear superposition of patterns with weights $w^* \in \mathbb{R}_+^p$:

$$\lambda_{w^*,i} = \lambda_{0,i} + a_i^\top w^* = \lambda_{0,i} + \sum_{j=1}^p a_{ij} w_j^*$$

$$y_i \sim \text{Poisson}(\lambda_{w^*,i})$$
for some known patterns, \( a_i = [a_{i,1}, \ldots, a_{i,p}]^T \).

This model arises in applications where the measurements are superposition of independent arrival processes of interest contaminated by some independent background arrival. Our goal is to recover \( k \)-sparse weight vector, \( w^* \), from \( y_i \)'s. Although \( y_i \)'s are non-identically distributed, their distributions are related through \( w^* \). Hence, estimating the weight vector \( w^* \) can be interpreted as a parameter estimation problem using \( n \) independent non-identical Poisson distributed samples, which are related through \( k \) non-zero elements of \( w^* \). We study the high dimensional problem where the number of parameters \( p \) can grow rapidly with \( n \), and \( k \) can scale with \( p \). Our goal is to prove that under appropriate conditions on \( a_i \)'s, \( \hat{w} \), the \( \ell^1 \) regularized ML estimate of \( w^* \) from \( y_i \)'s, is consistent with the ground truth:

\[
\lim_{n \to \infty} \Pr \{ \| \hat{w} - w^* \|_2 \geq \epsilon \} = 0
\]

Moreover, we want to show exponential rate of convergence with respect to the number of observations:

\[
\Pr \{ \| \hat{w} - w^* \|_2 \geq \epsilon \} \leq C' \exp(-nC)
\]

where \( C \) and \( C' \) are some positive constants.

C. Regularized Maximum Likelihood

The regularized ML estimate of \( w \) from \( y_1, \ldots, y_n \) is defined by:

\[
\hat{w} = \arg \max_{\Theta_k} \log p(y_1, \ldots, y_n | w)
\]

where \( \Theta_k \) is the set of feasible solutions dictated by sparsity and physics of the problem.

\[ \Theta_k = \{ w \mid w \geq 0, \| w \|_0 \leq k, \forall i: \lambda_{min} \leq \lambda_{w,i} \leq \lambda_{max} \} \]

Since \( \Theta_k \) is not a convex set, we define the set

\[ \Theta_s = \{ w \mid w \geq 0, \sum_{j=1}^{p} w \leq s, \forall i: \lambda_{min} \leq \lambda_{w,i} \leq \lambda_{max} \} \]

as a convex relaxation of the set \( \Theta_k \).

In our problem, the independence of observations along with the Poisson distribution of the observations, implies that the regularized ML estimation will have the form:

\[
\hat{w} = \arg \max_{w \in \Theta_s} \sum_{i=1}^{n} y_i \log (\lambda_{0,i} + a_i^T \hat{w}) - a_i^T \hat{w} \quad (1)
\]

The constrained maximization problem defined in Eqn. (1) is equivalent to the following unconstrained minimization problem for a suitably chosen \( \eta_s \):

\[
\hat{w} = \arg \min_{w} -\frac{1}{n} \sum_{i=1}^{p} y_i \log (\lambda_{0,i} + a_i^T w) - a_i^T w + \eta_s \sum_{j=1}^{p} w_j \quad (2)
\]

The latter problem is a convex optimization and can be solved efficiently by conventional optimization algorithms to find the global optimum. It needs to be mentioned that \( y_i \)'s are not identically distributed so the consistency of the regularized ML does not trivially follow from the consistency of ordinary maximum likelihood. In the next section, we will describe sufficient conditions on consistency of regularized ML estimation.

III. MAIN RESULTS

Under some mild condition on the response vectors, \( a_i \)'s, we can prove consistency of the estimation:

**Assumption 1.** Restricted Eigenvalue (RE) condition: Suppose \( S = \text{Supp}(w^*) \) with \( |S| \leq k \). There exists a constant \( \gamma_k > 0 \), such that for any vector \( u \in C(S) \triangleq \{ u \neq 0 : \| u_S \|_1 \geq \| u_S \|_2 \} \), we have:

\[
\frac{1}{n} \| Au \|_2^2 \geq \gamma_k \| u \|_2^2 \quad (3)
\]

where \( u_S \) is the restriction of the vector \( u \) to the indices in \( S \), and \( S^c = \{ 1, \ldots, p \} \setminus S \).

Remark: The \( 1/n \) factor on the left hand side of Eqn. (3) can be considered as a column normalization of \( A \), i.e. each column is divided by \( \sqrt{n} \).

RE condition is a well known sufficient condition for consistency of several sparse recovery algorithms. Specifically, various forms of it was used to derive the oracle inequalities for LASSO and Dantzig selector [11], [12]. In this paper, we are going to use this condition to establish the consistency of regularized ML for our highly non-linear Poisson model.

There are a number of well known results for random designs \( A \), for which Assumption 1 holds with high probability in terms of \( n \) [13]. For example, consider the case that elements of \( A \) are i.i.d. samples from a subgaussian distribution. Then, Assumption 1 is satisfied for all \( n \geq ck \log(p) \), with probability at least \( 1 - c_1 \exp(-c_2 n) \), where \( c_1 \), and \( c_2 \) are universal constants [14]. Moreover, in these cases \( \gamma_k \) is invariant to sparsity level \( k \), so long as \( n \geq ck \log(p) \).

In our experimental data, however, matrix \( A \) is deterministic and given to us. In general, testing RE condition is an NP-hard problem. Nevertheless, our numerical results still show fast rate of convergence for regularized ML.

Our results for deterministic and random designs are provided in the following sections. To simplify the theorem statements, we will use the following definition:

\[
\beta \triangleq \frac{\lambda_{max} (\lambda_{min} + 2s a_{max})}{\lambda_{min}} \sqrt{\log(\lambda_{max})} \quad (4)
\]

A. Deterministic Design

Our first result is on consistency of regularized ML estimation in Eqn. (2).

**Theorem 1.** Under Assumption 1, if \( \gamma_k = \omega \left( \frac{1}{\sqrt{n}} \right) \) and the elements of \( A \) are bounded, then for a suitable choice of \( \eta_s \), we have:

\[
\lim_{n \to \infty} \Pr \{ \| \hat{w} - w^* \|_2 \geq \epsilon \} = 0
\]
Our second result is on sample complexity of regularized ML in Eqn. (4):

**Theorem 2.** If Assumption 1 holds for \( \gamma_k \) and elements of \( A \) are bounded, and

\[
n \geq \frac{c \lambda_{\text{max}} \beta^4 \log \left( \frac{2}{\delta} \right)}{\gamma_k^2 \lambda_{\text{min}}^2 \epsilon^4}
\]

then for a suitable choice of \( \eta_k \) and any \( 0 < \delta < 1 \), we have:

\[
\Pr \{ \| \hat{w} - w^* \|_2 \geq \epsilon \} \leq \delta
\]

where

\[
\epsilon' \triangleq \min \left( \epsilon, \sqrt{\frac{1}{2} \frac{L}{\log(n)}} \right),
\]

\( L \) and \( c \) are constants, and \( \beta \) is as defined in Eqn. (4).

**B. Random Design**

**Theorem 3.** If elements of \( A \) are i.i.d. samples from a distribution with a bounded support on \( \mathbb{R}_+ \) or \( \mathbb{R}_- \), and the dimensionality \( p \) and sparsity levels \( k \) scale such a way that \( k^2 \log(p) \log^3(k \log(p)) = o(n) \), then for a suitable choice of \( \eta_k \), \( \hat{w} \) converges to \( w^* \) in probability. Furthermore, if \( k = o \left( \sqrt{n / \log(n)} \right) \), \( \hat{w} \) converges to \( w^* \) almost surely.

In the next theorem, we will give the exact convergence rate for the same design matrix as in Theorem 3.

**Theorem 4.** If elements of \( A \) are i.i.d. samples from a distribution with a bounded support on \( \mathbb{R}_+ \) or \( \mathbb{R}_- \), then for a suitable choice of \( \eta_k \), any \( n \geq c k^2 \log(p) \log^3(c' k \log(p)) \), we have:

\[
\Pr \{ \| \hat{w} - w^* \|_2 \geq \epsilon \} \leq 2 \exp \left( - \frac{c_1 \lambda_{\text{min}}^2 \epsilon^{\delta}}{k^2 \lambda_{\text{max}}^3 \epsilon^4} \right) + \exp(-c_2 n/k)
\]

where

\[
\epsilon' \triangleq \min \left( \epsilon, \sqrt{\frac{k}{\log(n)}} \right),
\]

\( L, c, c', c_1, \) and \( c_2 \) are constants, and \( \beta \) is as defined in Eqn. (4).

If both the positive and negative entries are allowed for the design matrix \( A \), then the well known sample complexity \( n = \Omega(k \log(p)) \) is possible by drawing elements of \( A \) from a sub-gaussian distribution:

**Theorem 5.** If rows of \( A \) are i.i.d. samples from an isotropic subgaussian distribution with a bounded support and the dimensionality \( p \) and sparsity levels \( k \) scale such a way that \( k \log(p) = o(n) \), then for a suitable choice of \( \eta_k \), \( \hat{w} \) converges to \( w^* \) almost surely.

The corresponding convergence of Theorem 5 is given in the next theorem:

**Theorem 6.** If rows of \( A \) are i.i.d. samples from an isotropic sub-gaussian distribution with a bounded support, then for a suitable choice of \( \eta_k \), any \( n \geq c k \log(p) \), we have:

\[
\Pr \{ \| \hat{w} - w^* \|_2 \geq \epsilon \} \leq 2 \exp \left( - \frac{c_1 \lambda_{\text{min}}^2 \epsilon^{\delta}}{k^2 \lambda_{\text{max}}^3 \epsilon^4} \right) + 2 \exp(-c_2 n) \quad (7)
\]

where

\[
\epsilon' \triangleq \min \left( \epsilon, L' \beta \right),
\]

\( L', c, c', c_1, \) and \( c_2 \) are constants, and \( \beta \) is as defined in Eqn. (4).

**IV. PROOF SKETCH**

**A. Deterministic Design**

Since Theorem 1 is a direct consequence of Theorem 2, we will prove the latter first. We use the idea of Extremum Estimators, which are a broad class of estimators for parametric models calculated through maximization (or minimization) of an objective function \( Q_n(w) \), which depends on the data [15].

**Lemma 1.** If \( \hat{w} \) and \( w^* \) are the minimizers of \( Q_n(w) \) and \( \overline{Q}_n(w) \) subject to \( w \in \Theta_s \), respectively, we have an upper bound of the form:

\[
\Pr \{ \| w^* - \hat{w} \|_2 \geq \epsilon \} \leq \Pr \left\{ \sup_{w \in \Theta_s} | Q_n(w) - \overline{Q}_n(w) | \geq \frac{\delta}{2} \right\}
\]

\[
(8)
\]

where \( \delta = \min_{\| w^* - \hat{w} \|_2 \geq \epsilon} \min_{\hat{w} \in \Theta_s} (\overline{Q}_n(\hat{w}) - \overline{Q}_n(w^*)) \).

**Proof.** The detailed proof is provided in Appendix A. \( \square \)

Intuitively, \( \delta \) in Lemma 1 represents the minimum increase in the function \( \overline{Q}_n(\hat{w}) \), when \( \hat{w} \in \Theta_s \) is at least \( \epsilon \) far away from the function minimizer \( w^* \). When the function \( \overline{Q}_n \) is strongly convex, \( \delta \) would be strictly positive, if \( \epsilon > 0 \). Moreover, if \( Q_n \) is uniformly convergent to \( \overline{Q}_n \) on \( \Theta_s \), the right hand side of (8) will converge to zero. Then, this will imply the consistency of \( \hat{w} \). Based on this argument, the proof to Theorem 2 consists of these parts:

- Assuming \( Q_n(w) = -\log p(y_1, \ldots, y_n|w^*) \) and \( \overline{Q}_n(w) = \mathbb{E}(Q_n(w)) \).
- Characterizing \( \delta \) in terms of \( \epsilon \) and the function \( \overline{Q}_n \).
- Obtaining some upper bound on right hand side of (8) through the appropriate concentration of measure inequalities.

We will briefly explain each part in what follows. We start with defining a random sequence \( Q_n \), and its expectation over data \( \overline{Q}_n \):

\[
Q_n(w) = -\frac{1}{n} \sum_{i=1}^{n} y_i \log(\lambda_{0,i} + a_i^T w) - a_i^T w \quad (9)
\]

\[
\overline{Q}_n(w) = -\frac{1}{n} \sum_{i=1}^{n} (\lambda_{0,i} + a_i^T w^*) \log(\lambda_{0,i} + a_i^T w) - a_i^T w \quad (10)
\]

It can be seen that \( w^* \) is the minimizer of \( \overline{Q}_n \) over \( \Theta_s \). Hence, preconditions of Lemma 1 are satisfied and we may use the
upper bound in Eqn. (8). Next, by exploiting Assumption 1, we may characterize δ in terms of ε:

**Lemma 2.** Suppose that there exists a number \( \gamma_k \) such that Assumption 1 holds for the sensing matrix \( A \) with bounded elements. Furthermore, consider \( Q_n(w) \) and \( \overline{Q}_n(w) \) as described in Eqn. (9) and (10). Then,

\[
\delta = \frac{\min_{\hat{w} \in \Theta_s} |Q_n(\hat{w}) - \overline{Q}_n(w)|}{\|\hat{w} - w\|} \geq \frac{\gamma_k \lambda_{\min}^2}{\lambda_{\max}(\lambda_{\min} + 2\alpha_{\max})^2}
\]

(11)

**Proof.** The detailed proof is provided in Appendix A. \( \Box \)

Finally, we are going to upper bound the right hand side of Eqn. (8) using the following lemma:

**Lemma 3.** For \( 0 < \delta \leq \frac{2\lambda_{\min} \log(\lambda_{\max})}{L} \) and a constant \( L \), we have the following inequality

\[
Pr \left( \sup_{w \in \Theta_s} |Q_n(w) - \overline{Q}_n(w)| \geq \frac{\delta}{2} \right) \leq \exp \left( -\frac{n\delta^2}{16\lambda_{\max} \log^2(\lambda_{\max})} \right)
\]

(12)

**Proof.** The detailed proof is provided in Appendix A. \( \Box \)

By plugging the δ into Eqn. (11) (Lemma 2) into Eqn. (8) and Eqn. (12) (Lemma 1 and 3) the convergence rate in Theorem 2 will be obtained.

**B. Random Design**

As Theorems 1 and 2 require RE condition as well as elements of \( A \) to be bounded, one may extend these results to case that \( A \) is chosen randomly and the preconditions are satisfied with high probability. RE condition can be guaranteed with high probability for various classes of random designs. Specifically, if rows of \( A \) is an ensemble of isotropic subgaussian random variables (or a linear transformation of them), it is well known that RE condition will be satisfied with overwhelming probability \( (1 - 2 \exp(-cn)) \) when \( n = \Omega(k \log(p)) \) with \( \gamma_k = \Omega(1) \). Then, Theorem 6 can be obtained from Theorem 2 by conditioning on the event that \( A \) satisfies RE. Moreover, since the convergence rate in Theorem 6 is summable, by Borel Cantelli Lemma \( \hat{w} \) converges to \( w^* \) almost surely (Theorem 5).

However, our setting needs all entries of \( A \) to be positive (or negative), which is not satisfied for a sub-gaussian ensemble. Therefore, we use the following Lemma from [16], which guarantees RE condition for the case that elements of \( A \) are i.i.d. samples from a bounded random variable (a variation of Theorem 1.8 to adapt to our definition of RE condition and notations).

**Lemma 4.** If elements of \( A \) are i.i.d. samples from a distribution with the bounded support, and \( n = \Omega(k \log(p)) \log^2(k \log(p)) \), then \( A \) satisfies RE condition with \( \gamma_k = \Omega(1/k) \) with probability at least \( 1 - \exp(-cn/k) \), where \( c \) is a constant.

Based on Lemma A, Theorem A can be obtained from Theorem 2 by conditioning the probability of error on the fact that \( A \) satisfies RE with \( \gamma_k = \Omega(1/k) \). Then, applying a union bound on the probability of the previous event and the event that \( A \) satisfies RE, the convergence rate of error in Theorem 4 will be yielded.

According to Lemma 4 if \( k \log(p) \log^2(k \log(p)) = o(n) \), one gets \( \gamma_k = \Omega(1/k) \). However, based on Theorem 1 \( \gamma_k = \Omega(1/\sqrt{n}) \) leads to the convergence of \( \ell^2 \) error to zero. Therefore, if both \( k^2 = o(n) \) and \( k \log(p) \log^2(k \log(p)) = o(n) \), the error will converge to zero in probability. This condition can be summarized to \( k^2 \log(p) \log^3(k \log(p)) = o(n) \), and Theorem 3 will be concluded. In addition, if \( k = o \left( \sqrt{\frac{n}{\log(n)}} \right) \), the convergence rate in Theorem 4 would become summable. Then, by Borel Cantelli Lemma, the error converges almost surely.

**V. Numerical Results**

**A. Rescaled LASSO vs. Regularized ML**

Parameter estimation based on LASSO for the Poisson setting has been studied in [3]. The idea is to view the problem as an additive noise problem where noise belongs to an exponential family of distributions. Alternatively, in [3] the problem is viewed as an additive Gaussian noise problem with noise variance equal to its mean to mimic “Poisson like” behavior. This results in a rescaled version of LASSO, which is then used to estimate model parameters. This amounts to scaling the loss function associated with each observation by the mean(or equivalently the variance).

This approach motivates us to compare our regularized ML method against re-scaled LASSO for poisson distributed data. In this section we will demonstrate that our regularized ML outperforms re-scaled LASSO in several regimes including low SNR, high dimensions, and moderate to low sparsity levels.

To compare the performance of regularized ML and rescaled LASSO, we first generate a random sensing matrix \( A \in \mathbb{R}^{n \times p} \) where each element \( a_{i,j} \) is an independent truncated Gaussian random variable. Due to the fact that entries in \( A \) are i.i.d. instantiation of a Gaussian random variable, this matrix satisfies RE condition with high probability and regardless of how \( p \) scales with \( n \). We also generate a random vector of the base rates, \( \lambda_0 \in \mathbb{R}^n \), and some sparse vector \( w^* \in \mathbb{R}^p \), with \( \|w^*\|_1 = s \). To recover \( w \), we generate \( n \) Poisson distributed data with coefficients specified in \( A \) as:

\[
y_i = \text{Poisson}(\lambda_{0,i} + a_i^T w^*)
\]

We first solve the non linear optimization where \( w \) is constrained to be in \( \Theta_s \).

\[
\hat{w}_{ML} = \arg \min_{w \in \Theta_s} -\frac{1}{n} \sum_{i=1}^{n} y_i \log(\lambda_{0,i} + a_i^T w) - a_i^T w
\]

Then, we apply the rescaled LASSO:

\[
\hat{w}_{LASSO} = \arg \min_{w \in \Theta_s} -\frac{1}{n} \sum_{i=1}^{n} (y_i - \lambda_{0,i} - a_i^T w)^2 \frac{1}{\lambda_{0,i} + a_i^T w} + \gamma_k \|w\|_1
\]
For comparison purposes we then threshold the solution by zeroing out components of \(\hat{w}_{ML}\) and \(\hat{w}_{LASSO}\) below a pre-defined small threshold \(t\). We average the estimation performance over 100 Monte Carlo loops. The performance of the two methods are compared in Fig. 1, Fig. 2, and Fig. 3. The results are compared in terms of different sparsity levels, \(k\), base rates \(\lambda_0\) and number of observations, \(n\), respectively.

In Fig. 2, we compare the result of regularized ML estimation to that of re-scaled LASSO for different regimes of the base rate, \(\lambda_0\). Here, we fix \(n = 100\), \(p = 200\), \(t = 10^{-3}\) and \(k = 20\). We compare the performance of the two approaches based on the Hamming distance error between the support set of the thresholded estimation and that of the Ground truth. This error indicates the number of mismatches between the two support sets. We average this error over 100 random samples of \(w\) for each value of \(\lambda_0\). We see that for small values of \(\lambda_0\) we notice a significant difference. To understand this effect consider the following sequence of expressions:

\[
\log(\lambda_{0,i} + a_i^Tw^*) = \log(\lambda_{0,i}) + \log(1 + \frac{a_i^Tw^*}{\lambda_{0,i}})
\]

Notice that this is first term in regularized ML in Eq. 2. For large values of \(\lambda_0\) we see that \(\log()\) behaves linearly and this accounts for the negligible difference between rescaled LASSO and ML. On the other when \(\lambda_0\) is small this is no longer true and re-scaled LASSO cannot be justified and we notice a significant difference between the two approaches.

Alternatively, we can view the problem in terms of validity of Gaussian approximations. For large values of \(\lambda_0\), e.g. \(10^5\) or \(10^6\), Poisson distribution behavior is very similar to a narrow Gaussian distribution. Therefore, the performance of both methods are similar. However, normal distribution is not a good approximation for the behavior of a Poisson model for smaller \(\lambda_0\). This deviation from the model degrades the performance of rescaled LASSO in terms of Hamming distance error in the support recovery.

In Fig. 3, we compare the result of regularized ML estimation with rescaled LASSO as a function of \(n\). This time, we fix \(\lambda_0 = 100\), \(p = 400\), \(t = 10^{-4}\) and \(k = 40\). At each iteration, we estimate \(w\) based on \(n\) observations where \(n\) varies from 2 to 400. We compare the performance of the two approaches based on probability of successful recovery of the support set.

This error is 0 if the thresholded support set of the estimation is equal to that of the ground truth and I otherwise. We average this error over 100 samples of \(w\) per each \(n\).

In Fig. 4, we compare the result of regularized ML estimation with rescaled LASSO for different sparsity levels, \(k\). This time, we fix \(\lambda_0 = 100\), \(p = 200\), and \(n = 100\). For each \(k\), we generate 100 samples of \(k\)-sparse \(w\)’s and recover them from \(n\) observations. Since \(||w^*||_1 = 1\) for all values of \(k\), we threshold \(w\)’s element according to their corresponding \(k\), \(t < \frac{0.01}{k}\), to obtain their sparse support set. We measure the performances of the two estimations based on average probability of successful recovery of the thresholded support set for each value of \(k\). Again the intuition here is similar to the one in Fig. 2. For small sparsity levels the log expression tends to behave linearly and so we do not notice a significant difference between the two approaches.

Notice that the error bars in Fig. 3 and Fig. 4 indicate that our result is indeed statistically significant.

In Fig. 5, we compare the result of regularized ML estimation with rescaled LASSO in terms of the ROC curves. In an ROC curve, the average number of true detections is plotted against the average number of false alarms. True detections are indices that are common in the thresholded estimated support set and that of the Ground Truth, whereas, false alarms are the indices in the thresholded estimated support set that are not included in the support set of the Ground Truth. This time, we fix \(\lambda_0 = 100\), \(p = 200\), \(n = 100\), and \(k = 20\). By applying different thresholds \(t = \frac{1}{k}\) to \(t = \frac{0.01}{k}\) we obtain the different points in the ROC plot. We average Probability of Detection
Weights recovered from Rescaled LASSO

Weights recovered from Regularized ML

ROC curve for ML and Lasso

Fig. 5. ROC curve for Regularized ML and LASSO for \( n = 100, p = 200, k = 20, \lambda = 100, \) and \( m = 100 \) (monte carlo loops).

(PD) and Probability of False alarm (PF) over 100 Monte Carlo loops.

The error bars show that despite the randomness of our data, this result is statistically significant.

B. Explosive Identification

In this experiment, we first measure the light intensities of different fluorophores before and after separate exposures to a unit weight of different explosives. The intensities are measured by counting the number of photons received at each photo-sensor. Each explosive \( j \) has a unique quenching effect in the fluorescence property of each fluorophore \( i \), which we denote by \( a_{i,j} \). In the experimental setting with \( \lambda_i \) being the before exposure intensity for fluorophore \( i \), we model the after exposure intensity \( y_i \) as:

\[
y_i \sim \text{Poisson}(\lambda_i(1-a_{i,j}))
\]

In the next step, fluorophores are exposed to an unknown mixture of these explosives. The goal is to recover which and how much of each explosive is contained in that mixture.

The physics of the problem suggests that when the fluorophore is exposed to a mixture of explosives, the quenching effects are additive in the regime where the mixture weights are small. Therefore, our observations are best modeled by a Poisson distribution with additive rate model for each fluorophore:

\[
y_i \sim \text{Poisson}\left(\lambda_i \left(1 - \sum_{j=1}^{p} a_{i,j} w_j \right)\right)
\]

where \( p \) is the total number of basic explosives and \( w_j \) is the amount of the explosive \( j \) in the mixture. We solve this problem through Regularized ML and Rescaled LASSO and compare the results.

In this problem, matrix \( A \), the responses of \( n = 8 \) fluorophores to \( p = 12 \) basic explosives is given. Based on this given data and our additive model for mixtures, we generated 10 mixtures by combining up to 3 random explosives. We used Regularized ML and Rescaled LASSO to identify these mixtures through their effect on fluorescence property of our fluorophores. The result is shown in the form of a \( 10 \times 12 \) grid in Fig. 5. In this grid, rows are different mixtures and columns are different explosives. Dark squares indicate the absence (or negligible contribution), where as lighter squares indicate higher amount of the corresponding explosive in the associated mixture.

Since photon count rates are of the order of \( 10^5 \), based on our discussion in the previous section, Regularized ML and Rescaled LASSO performances are very close.

C. Internet Marketing Application

In this application, we are going to extract strategies that result in higher website traffic for specific businesses. We specifically studied the Clothing Brand Market and based on the information provided at Alexa.com and the brands’ websites, we tried to extract the successful strategies for clothing market.

Our assumption is that the website traffic is generated as a superposition of the traffic generated from current costumers and the traffic from advertisement through backward links (links in advertisement websites that are linked to these business websites). In general, big business websites buy 1000-1500 backward links in advertisement websites. However, the hypothesis is that only a few of them are efficiently directing costumers. Our goal is to find those dominant advertisement websites.

We modeled the website traffic by:

\[
y_i = \text{Poisson}(\lambda_{i,0} + a_i^\top w)
\]

Fig. 6. Sparse recovery results for \( k \leq 3 \). Columns: Basic explosives. Rows: Synthesized mixtures with \( k \) basic explosives, non-black squares at each row show the explosives that the corresponding mixture is composed of. lighter colors show larger amounts.

Fig. 7. Average Hamming distance error in support set recovery for different sparsity levels and 100 (monte carlo loops).
where \( \lambda_{0,i} \) models the current costumers who visit the site directly, \( a_{i,j} \) is the number of backwards link for the website \( i \) in the advertisement website \( j \). Our model assumes that each of the backward links brings independent traffic to the website. Therefore, we used our additive model. Our analysis provided the top ten backward links for this market:

### TABLE I

| Backward link | ML estimated weight | Lasso estimated weight |
|---------------|---------------------|------------------------|
| Amazon        | 0.32                | 0.35                   |
| Twitter       | 0.21                | 0.17                   |
| Pinterest     | 0.17                | 0.16                   |
| Google        | 0.15                | 0.15                   |
| Blogger       | 0.06                | 0.09                   |
| Bing          | 0.05                | 0.09                   |
| douban        | 0.01                | 0.01                   |
| tumblr        | 0.01                | 0.05                   |

### VI. CONCLUSIONS

We provided convergence guarantees for the solution of ML decoder of a high dimensional sparse underlying parameter of heterogeneous Poisson distributed data. We assumed the design matrix to satisfy Restricted Eigenvalue (RE) condition, which has been originally used to prove the consistency of sparse linear models. Although our model is highly non-linear, we establish that RE is a sufficient condition to obtain various convergence results. We also provided the error bounds for the bounded random designs satisfying RE condition with high probability. In our experiments we compared rescaled LASSO against our regularized ML. We concluded that regularized ML can result in significantly superior performance in “low-SNR” regimes.

### VII. APPENDIX A

#### A. Useful Bounds:

To show an exponential rate of convergence for Poisson distributed data, we need a tool to bound the tail probability. We build this tool from Bernstein inequality for Poisson distribution:

**Lemma 5.** (Bernstein inequality)[17] Let \( y_1, \ldots, y_n \) be independent random variables with means \( \mu_1, \ldots, \mu_n \). Suppose that \( \exists L > 0 \) such that \( \forall k \in N \) and \( k > 1 \):

\[
E[(y_i - \mu_i)^k] \leq \frac{1}{2} E[(y_i - \mu_i)^2] L^{k-2} k!
\]

Then, we have:

\[
\Pr \left\{ \frac{1}{n} \sum_{i=1}^{N} |y_i - \mu_i| \geq \frac{2t}{n} \sqrt{\sum_{i=1}^{N} E[(y_i - \mu_i)^2]} \right\} \leq 2 \exp \left( -t^2 \right)
\]

where \( 0 < t \leq \sqrt{2 \sum_{i=1}^{N} E[(y_i - \mu_i)^2]} / 2L \).

**Lemma 6.** [7] For \( y_i \)'s distributed as:

\[
y_i \sim \text{Poisson}(\lambda_i)
\]

There exists a bounded constant \( L > 0 \), such that \( \forall k \in N \) and \( k > 1 \):

\[
E[(y_i - \lambda_i)^k] \leq \frac{1}{2} E[(y_i - \lambda_i)^2] L^{k-2} k!
\]

**Remark:** The proof of this Lemma is provided in [7] and is based on the fact that moment generating function for Poisson distribution with rate \( \lambda \),

\[
\exp(\lambda(\exp(t) - 1))
\]

is an analytic function. Therefore the \( k \)-th coefficient of Taylor series exists and is bounded.

#### B. Proof of Lemma 1

Now, we need to show that for any \( \epsilon > 0 \), and

\[
\delta = \min_{\|w - w^*\| \geq \epsilon} \left\{ |\mathcal{Q}_n(w) - \mathcal{Q}_n(w^*)| \right\}
\]

we have:

\[
\Pr \left\{ \|w^* - \hat{w}_n\|_2 \geq \epsilon \right\} \leq \Pr \left\{ 2 \sup_{w^* \in \Theta} |Q_n(w) - \mathcal{Q}_n(w)| \geq \delta \right\}
\]

The proof of this part, can be shown by combining two Lemmas:

**Lemma 7.** For any \( \epsilon > 0 \), and

\[
\delta = \min_{\|w - w^*\| \geq \epsilon} \left\{ |\mathcal{Q}_n(w) - \mathcal{Q}_n(w^*)| \right\}
\]

we have:

\[
\Pr \left\{ \|w^* - \hat{w}_n\|_2 \geq \epsilon \right\} \leq \Pr \left\{ 2 \sup_{w^* \in \Theta} |Q_n(w) - \mathcal{Q}_n(w)| \geq \delta \right\}
\]

**Lemma 8.** For any \( \delta > 0 \), we have:

\[
\Pr \left\{ |\mathcal{Q}_n(w) - \mathcal{Q}_n(w^*)| \geq \delta \right\} \leq \Pr \left\{ 2 \sup_{w^* \in \Theta} |Q_n(w) - \mathcal{Q}_n(w)| \geq \delta \right\}
\]

**Proof of Lemma 7** Consider

Event A: \( \|w^* - \hat{w}_n\|_2 \geq \epsilon \)

and

Event B: \( |\mathcal{Q}_n(w) - \mathcal{Q}_n(w^*)| \geq \delta \)

We have:

\( (A \Rightarrow B) \Rightarrow \Pr(A) \leq \Pr(B) \)

In other words, for any \( \epsilon > 0 \) and \( \delta \) as defined in Lemma 7 we have:

\[
\Pr \left\{ \|w^* - \hat{w}_n\|_2 \geq \epsilon \right\} \leq \Pr \left\{ |\mathcal{Q}_n(w) - \mathcal{Q}_n(w^*)| \geq \delta \right\}
\]

**Proof of Lemma 8** Based on the definition of \( \mathcal{Q}_n(w) \):

\[
w^* = \arg \min_{w} \mathcal{Q}_n(w)
\]

Therefore:

\[
|\mathcal{Q}_n(\hat{w}) - \mathcal{Q}_n(w^*)| = \mathcal{Q}_n(\hat{w}) - \mathcal{Q}_n(w^*)
\]
minimizer of $\delta$ we can choose corresponding

Assumption 1 is satisfied for some

The minimum of Lemma 9.

Lemma 10.

The proof is provided in the Appendix B.

Proof. [C. Proof of Lemma 2]

Let's assume $\hat{w} = w^* + u$. We have:

$$|Q_n(\hat{w}) - Q_n(w^*)| = \min_{\|u\|_2 \geq \epsilon} f(u)$$

where $\lambda_{w^*} = \lambda_0 + a^T w^*$.

Our goal is to find:

$$\delta = \min_{\|u\|_2 \geq \epsilon, \hat{w} \in \Theta} \|f(u)\|$$

According to definition, $\forall i, \lambda_{\min} \leq \lambda_{w^*, i}$, and we have:

$$f(u) \geq \frac{1}{n} \sum_{i=1}^{n} -\lambda_{w^*, i} \log \left( 1 + a_i^T u \right) + a_i^T u$$

and from inequality $-\log(1+x) \geq -x$, we can show that for all $u$, $f(u) \geq 0$. This confirms the fact that $w^*$ is the minimizer of $Q_n(w)$.

**Lemma 9.** The minimum of $f(u)$, as defined in Eqn. (14), over the set $\|u\|_2 \geq \epsilon$, is located on the boundary; $\|u\|_2 = \epsilon$:

$$\min_{\|u\|_2 \geq \epsilon} f(u) = \min_{\|u\|_2 = \epsilon} f(u)$$

**Proof.** The proof is provided in the Appendix B.

**Lemma 10.** For function $f(u)$, as defined in Eqn. (14), if Assumption 1 is satisfied for some $0 < \gamma_k < 1$, we have:

$$\min_{\|u\|_2 = \epsilon} f(u) \geq \frac{\gamma_k^2 \lambda_{w^*}^2\epsilon^2}{\lambda_{\max}^2 (\lambda_{\min} + 2s_{\max})^2}$$

**Proof.** The proof is provided in the Appendix B.

Therefore, to guarantee that Eqn. (13) holds for any $\epsilon > 0$, we can choose corresponding $\delta$ to be:

$$\delta = \frac{\gamma_k^2 \lambda_{w^*}^2\epsilon^2}{\lambda_{\max}^2 (\lambda_{\min} + 2s_{\max})^2}$$

**D. proof of Lemma 3**

We start with:

$$\Pr \left\{ \sup_{w \in \Theta} |Q_n(w) - \bar{Q}_n(w)| \geq \frac{\delta}{2} \right\}$$

$$\leq \Pr \left\{ \sup_{w \in \Theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \lambda_{w^*, i}) \log(\lambda_{w, i}) \geq \frac{\delta}{2} \right\}$$

$$\leq \Pr \left\{ \sum_{i=1}^{n} (y_i - \lambda_{w^*, i}) \geq \frac{n\delta}{2} \log(\lambda_{\max}) \right\}$$

where for the last inequality, we exploit the assumption $\lambda_{\min} > 1$. To use the result of Lemma 5, we have to set:

$$\frac{n\delta}{2} \log(\lambda_{\max}) = 2t \sqrt{\sum_{i=1}^{n} \lambda_{w^*, i}}$$

Combining Eqn. (19) and Eqn. (20), we have:

$$\frac{n\delta}{4 \log(\lambda_{\max}) \sqrt{\sum_{i=1}^{n} \lambda_{w^*, i}}} \leq \sqrt{\sum_{i=1}^{n} \lambda_{w^*, i}}$$

Therefore, $\delta$ must be upper bounded by:

$$\delta \leq \frac{2\lambda_{\min} \log(\lambda_{\max})}{L}$$

to guarantee:

$$\Pr \left\{ \sup_{w \in \Theta} |Q_n(w) - \bar{Q}_n(w)| \geq \frac{\delta}{2} \right\}$$

$$\leq 2 \exp \left( -\frac{n\delta^2}{16 \log^2(\lambda_{\max}) \lambda_{\max}} \right)$$

**VIII. APPENDIX B**

**A. Proof of Lemma 9**

To prove this, let's assume the unique minimum is obtained somewhere else, $\|u^*\|_2 > \epsilon$. Also consider $u'$ to be defined as $u' = \frac{\epsilon}{\|u^*\|_2} u^*$. We can also write $u'$ as

$$u' = \frac{\epsilon}{\|u^*\|_2} u^* + \left( 1 - \frac{\epsilon}{\|u^*\|_2} \right) 0$$

From convexity of $f(u)$, we have:

$$f(u') \leq \frac{\epsilon}{\|u^*\|_2} f(u^*) + \left( 1 - \frac{\epsilon}{\|u^*\|_2} \right) f(\bar{w})$$

From definition of $f(u)$, we know that $f(\bar{w}) = 0$, hence:

$$f(u') \leq \frac{\epsilon}{\|u^*\|_2} f(u^*) < f(u^*)$$

This is contrary to assuming that $u^*$ is the minimizer of $f$. 

B. Proof of Lemma [10]

Now that we know the minimizer lies on the boundary, from Eqn. (15) we have:

\[
\min_{w^* + u \in \Theta_s} f(u) \geq \min_{w^* + u \in \Theta_s} \frac{1}{n} \lambda_m \sum_{i=1}^{n} \log \left( 1 + \frac{a_i^T u}{\lambda_{w^{*},i}} \right) + \frac{a_i^T u}{\lambda_{w^{*},i}} \\
\geq \frac{1}{n} \lambda_m \
\sum_{i=1}^{n} \left( \frac{1}{(1 + \frac{2a_{\max}}{\lambda_{\min}}) \|X\|^2} \right) = \frac{1}{\lambda_{\min}} \sum_{i=1}^{n} X_i^2 \\
\geq \frac{\lambda_{\min}}{\lambda_{\max} (1 + \frac{2a_{\max}}{\lambda_{\min}})^2} \|X\|^2 \\
= \frac{\lambda_{\min}^2}{\lambda_{\max}^2 (\lambda_{\min} + 2a_{\max})^2}
\]

By change of variables:

\[X_i = \frac{a_i^T u}{\lambda_{w^{*},i}}\]

Before we proceed to apply the result of Assumption 1, we need to check that \[\|u_S\|_1 \geq \|u_{S^c}\|_1\]. We know:

\[\|u_S\|_1 = \|\tilde{w}_S - w^*\|_1 \geq \|w^*\|_1 - \|\tilde{w}_S\|_1\]

Moreover, from \[\|\tilde{w}\|_1 \in \Theta_s\], we have:

\[\|\tilde{w}_S\|_1 + \|\tilde{w}_{S^c}\|_1 = \|\tilde{w}\|_1 \leq s = \|w^*\|_1\]

Therefore,

\[\|u_{S^c}\|_1 = \|\tilde{w}_{S^c}\|_1 \leq \|w^*\|_1 - \|\tilde{w}_S\|_1 \leq \|u_S\|_1\]

Now, from Assumption 1, we have:

\[\frac{1}{n} \|X\|^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{(a_i^T u)^2}{\lambda_{w^{*},i}} \geq \gamma k \frac{\epsilon^2}{\lambda_{\max}}\]

Now, by applying Taylor series expansions around \(X_i = 0\) to each term in the sum in Eqn. (23), we have:

\[-\log (1 + X_i) + X_i = -X_i + \frac{1}{1 + X_i} X_i^2 + X_i\]

where \(|\tilde{X}_i|\) lies between 0 and \(|X_i|\):

\[|\tilde{X}_i| = |c \times 0 + (1 - c) \times X_i| \leq |X_i| \leq \frac{2sa_{\max}}{\lambda_{\min}}\]

where the last inequality follows from the fact that \(\|u\|_1 \leq 2s\). Therefore, we can rewrite Eqn. (23) as:

\[\min_{w^* + u \in \Theta_s} f(u) \geq \min_{\|X\|^2 \geq \gamma k \frac{\epsilon^2}{\lambda_{\max}} \sum_{i=1}^{n} \left( \frac{1}{(1 + \frac{2a_{\max}}{\lambda_{\min}}) \|X\|^2} \right) = \frac{\lambda_{\min}}{\lambda_{\max} (1 + \frac{2a_{\max}}{\lambda_{\min}})^2} \|X\|^2 \]

REFERENCES

[1] R. C. Stringer, S. Gangopadhyay, and S. A. Grant, “Detection of nitroaromatic explosives using a fluorescent-labeled imprinted polymer,” in *Analytical Chemistry*, 2010.

[2] J. Beel, B. Gipp, and E. Wilde, “Academic search engine optimization (aseo): Optimizing scholarly literature for google scholar and co.,” *Journal of Scholarly Publishing*, 2010.

[3] J. Jia, K. Rohe, and B. Yu, “The lasso under poisson-like heteroscedasticity,” *arXiv*, 2010.

[4] D. L. Snyder, *Random Point Processes*, Wiley-Interscience Publication, 1975.

[5] Z. T. Harmany, R. F. Marcia, and R. M. Willett, “Sparse poisson intensity reconstruction algorithms,” in *SSP*, 2009.

[6] I. Rish and G. Grabarnik, “Sparse signal recovery with exponential-family noise,” *Allerton*, 2009.

[7] S. Kakade, O. Shamir, K. Sridharan, and A. Tewari, “Learning exponential families in high-dimensions: Strong convexity and sparsity,” *arXiv.org*, 2009.

[8] S. Portnoy, “Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity,” *Annals of Statistics*, 1988.

[9] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, “A unified framework for high-dimensional analysis of m-estimators with descomposable regularizers,” *arXiv*, 2012.

[10] M. Raginsky, R. M. Willett, Z. T. Harmany, and R. F. Marcia, “Compressed sensing performance bounds under poisson noise,” *arXiv.org*, 2012.

[11] P. J. Bickel, Ya, acow Ritov, and Alexandre B. Tybakov, “Simultaneous analysis of lasso and dantzig selector,” *Annals of Statistics*, Volume 37, Number 4 (2009), 1705-1722.

[12] S. van de Geer and P. Buhlmann, “On the conditions used to prove oracle results for the lasso,” *Electron. J. Stat.*, vol. 3, pp. 1360–1392, 2009.

[13] G. Raskutti, M. J. Wainwright, and B. Yu, “Restricted eigenvalue properties for correlated gaussian designs,” *Journal of Machine Learning Research*, vol. 11, pp. 2241–2259, 2010.

[14] Shuheng Zhou, “Restricted eigenvalue conditions on subgaussian random matrices,” *arXiv*:0912.4045v2 [math.ST], 2009.

[15] A. Takeda, “Advanced econometrics,” *Harvard University Press*, 1985.

[16] M. Rudelson and S. Zhou, “Reconstruction from anisotropic random measurements,” *arXiv*:1106.1151 [math.ST], 2010.

[17] S. Bernstein, “The theory of probabilities,” *Gastehizdat Publishing House*, 1946.

[18] W. K. Newey, “Uniform convergence in probability and stochastic equicontinuity,” *Econometrica*, 1991.

[19] B. Hoadley, “Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case,” *The Annals of Mathematical Statistics*, 1971.

[20] F. Hayashi, *Econometrics*, Princeton University Press, 2000.

[21] F. X. Dupa, J. Fadili, and J. L. Starck, “Deconvolution under poisson noise using exact data delity and synthesis or analysis sparsity priors,” *International Conference on Image Processing (ICIP)*, 2011.

[22] F. X. Dupa, J. Fadili, and J. L. Starck, “A proximal iteration for deconvolving poisson noisy images using sparse representations,” *IEEE Transaction on Image Processing*, 2011.