SPARSE SIGNAL RECOVERY UNDER POISSON STATISTICS FOR ONLINE MARKETING APPLICATIONS

Delaram Motamedvaziri, Mohammad H. Rohban, and Venkatesh Saligrama

Boston University

ABSTRACT

We are motivated by many applications such as problems that arise in online marketing applications, where the observations are governed by non-homogeneous Poisson models. We analyze the performance of a Maximum Likelihood (ML) decoder. We prove consistency and show an exponential rate of converge for sparse recovery in the high-dimensional Poisson setting. After verifying the efficiency of ML estimator empirically, we apply the ML decoder to study the dynamics of online marketing methods over time.

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

1. INTRODUCTION

In this paper, we study the problem of high dimensional sparse model estimation under Poisson model for observations. This problem is motivated by online marketing applications where the observations are the counts of an event, e.g. number of online visitors for a website, number of clicks on an online ad, etc. The rate of this Poisson distribution depends linearly on a sparse subset of parameters. Our goal is to extract the sparse subset from a potentially large number of parameters.

Consider the problem of finding a small set of advertising websites that brings the most 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 advertisement websites that redirect the largest amount of online traffic to their business websites. In this problem, the online traffic for different business websites is modeled by non-identical Poisson distributions.

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 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 $w^* \in \mathbb{R}^p_+$, the i-th observation, y_i, is distributed as follows:

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

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

In the previous online marketing example, y_i is the daily 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 ad links that business website i has purchased from advertisement website j, and w^*_j is a relative score that measures the effectiveness of a single ad in advertisement website j. The advertisement websites j is more desirable if its relative score, w_j, is higher.

To estimate w^*, we use a Maximum Likelihood (ML) decoder. We prove consistency of ML recovery in the high-dimensional setting. Moreover, when the sensing matrix satisfies the so-called Restricted Eigenvalue (RE) condition, our estimates converge exponentially fast in terms of the number of observations. We also conduct several synthetic experiments to confirm the efficiency of ML estimator for our setting. After proving and verifying its efficiency for our settings, we apply the ML estimator to solve a practical problem in the realm of online marketing.

1.1. Related Work

Consistency of maximum-likelihood estimators for most conventional estimation methods under Poisson Statistics hinges on identically distributed observations [1, 2, 3, 4]. Therefore, their results do not apply to our setting.

Parameter estimation for non-identical Poisson distributions, as a member of exponential family, has been studied in the context of Generalized Linear Models (GLMs) [5, 6]. However, our model is inherently different from the the GLM models. In GLM models, the observations distributed as:

$$\text{Model I}: \Pr(y_i = k) \propto \exp\left(k a_i^\top w\right) \exp(\exp(a_i^\top w))$$

whereas in our problem, the observations are distributed as:

$$\text{Model II}: \Pr(y_i = k) \propto \left(a_i^\top w\right)^k \exp(a_i^\top w)$$

Due to this fundamental difference in the models, the

techniques developed in the context of GLM models do not apply to our setting.

Our linear additive model for the Poisson rate together with imposition of sparsity directly on w distinguishes our work from [7] and [8].

In another related work, compressive sensing problem under Poisson statistics has been studied in [9], where a similar problem setting is introduced

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

The main focus of [9] 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, achievability 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 which results in different type of analysis.

2. PROBLEM SETUP

2.1. 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 $w^* \in \mathbb{R}_+^p$:

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

for some known positive patterns, $a_i = [a_{i,1}, \ldots, a_{i,p}]^\top$.

This model arises in applications where the measurements are superposition of independent arrival processes contaminated by some independent background arrival. Our goal is to recover k-sparse weight vector, w^*, from y_i's. Estimating the weight vector w^* can be interpreted as a parameter estimation problem using n independent non-identical Poisson distributed samples. Despite non-identicality, these Poisson distributions 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, $\widehat{w}$, the ℓ^1 constrained ML estimate of w^* from y_i's, is consistent with w^*. Moreover, we want to show exponential rate of convergence with respect to the number of observations.

2.2. Constrained Maximum Likelihood

The constrained ML estimate of w from $y_1, \ldots, y_n$ is defined by:

$$\widehat{w} = \arg\max_{w \in \Theta_k} \log p(y_1, \ldots, y_n | w)$$

where Θ_k is the set of constrained feasible solutions dictated by sparsity and physics of the problem.

$$\Theta_k = \{w |\, w \geq 0, \|w\|_0 \leq k,\ \forall i\colon \lambda_{min} \leq \lambda_{w,i} \leq \lambda_{max}\}$$

Since Θ_k is not a convex set, we define the set

$$\Theta_s = \{w |\, w \geq 0, \|w\|_1 \leq s\ \forall i\colon \lambda_{min} \leq \lambda_{w,i} \leq \lambda_{max}\}$$

as a convex relaxation of the set Θ_k, where $s = \|w^*\|_1$.

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

$$\widehat{w} = \arg\min_{w \in \Theta_s} -\frac{1}{n} \sum_{i=1}^{n} y_i \log\left(\lambda_{0,i} + a_i^\top w\right) - a_i^\top w \quad (1)$$

Eqn.(1) is a convex minimization problem and can be solved efficiently by conventional optimization algorithms. It needs to be mentioned that y_i's are not identically distributed so the consistency of the constrained ML does not trivially follow from the consistency of ordinary maximum likelihood. In the next section, we will describe sufficient conditions for consistency of constrained ML estimation.

3. MAIN RESULTS

3.1. Assumptions

Under some mild conditions on the response vectors, a_i's, we can prove consistency of the estimation $\widehat{w}$ in (1):

Assumption 1. *Boundedness: All elements of the sensing matrix, A, must be bounded:*

$$0 \leq a_{i,j} \leq a_{max} < \infty$$

Assumption 2. *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 $u \in \mathbb{C}(S) \triangleq \{u \neq \mathbf{0} : \|u_S\|_1 \geq \|u_{S^c}\|_1\}$, we have:*

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

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

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 [10], [11]. In this paper, we are going to use this condition to establish the consistency of constrained ML for our highly non-linear Poisson model.

There are a number of well known results for random sensing matrices, A, for which Assumption 1 holds with high probability in terms of n [12]. For example, consider

the case that elements of A are i.i.d. samples from a sub-Gaussian 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, c_1, and c_2 are universal constants [13].

We exploit these results in our synthetic experimental results to guarantee RE condition for the randomly generated sensing matrix, A.

Our theoretical results are provided in the following sections. To simplify the theorem statements, we will use the following definition :

$$\beta \triangleq \frac{\lambda_{max}(\lambda_{min} + 2sa_{max})}{\lambda_{min}}\sqrt{\log(\lambda_{max})} \tag{3}$$

3.2. Theoretical Results

Our first result is on consistency of constrained ML estimation in Eqn. (1).

Theorem 1. *Under Assumption 1 and 2, if* $\gamma_k = \omega\left(\frac{1}{\sqrt{n}}\right)$, *then for a suitable choice of* Θ_s, *we have:*

$$\lim_{n\to\infty} \Pr\{\|\widehat{w} - w^*\|_2 \geq \epsilon\} = 0$$

Our second result is on the sample complexity of constrained ML as defined in Eqn. (1):

Theorem 2. *If Assumption 1 holds for* γ_k *and elements of* A *are bounded, and*

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

then for a suitable choice of Θ_s *and any* $0 < \delta < 1$, *we have:*

$$\Pr\{\|\widehat{w} - w^*\|_2 \geq \epsilon\} \leq \delta$$

where

$$\epsilon' \triangleq \min\left(\epsilon, \sqrt{\frac{1}{L\gamma_k}}\beta\right),$$

L *and* c *are universal constants, and* β *is as defined in Eqn.* (3).

The proof of the theorems are provided in [14].

4. NUMERICAL RESULTS

We start this section by verifying our theoretical results. After confirming the consistency of ML estimation for our specific choice of objective function, we apply this method to solve the online marketing problem we stated earlier. Throughout this section, we compare the performance of constrained ML with that of Rescaled LASSO [15]. In Rescaled Lasso, the Poisson noise is viewed as an additive Gaussian noise where the noise variance is equal to its mean to mimic "Poisson like" behavior. Our results demonstrate the superiority of Poisson model (ML method) over its Guassian counterpart (Rescaled LASSO).

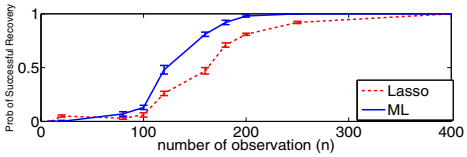

Fig. 1: *Probability of successful support recovery as a function of n for* $p = 400$, $\lambda_0 = 100$ $k = 40$, $t = 10^{-4}$ *and* $m = 100$ *(monte carlo loops).*

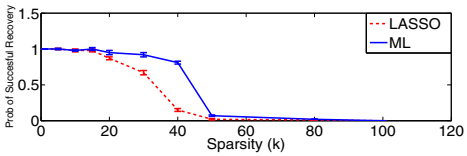

Fig. 2: *Probability of successful support recovery as a function of k for* $p = 200$, $\lambda_0 = 100$, $t < \frac{0.01}{k}$, $n = 100$, *and* $m = 100$ *(monte carlo loops).*

$$\widehat{w}_{ML} = \arg\min_{w\in\Theta_s} -\frac{1}{n}\sum_{i=1}^{n} y_i \log(\lambda_{0,i} + a_i^\top w) - a_i^\top w$$

$$\widehat{w}_{LASSO} = \arg\min_{w\in\Theta_s} \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - \lambda_{0,i} - a_i^\top w)^2}{\lambda_{0,i} + a_i^\top w} + \eta\|w\|_1$$

4.1. Synthetic Data

To verify our theoretical results and to compare the performance of constrained 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 Gaussian random variable. Due to the fact that entries in A are i.i.d. instantiations of a Gaussian random variable, this matrix satisfies RE condition with high probability and regardless of how p scales with n [5]. 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:

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

For comparison purposes we then threshold the solution by zeroing out components of $\widehat{w}_{ML}$ and $\widehat{w}_{LASSO}$ below a predefined small threshold t. We average the estimation performance over 100 Monte Carlo loops. The performance of the two methods are compared in Fig. 1 and Fig. 2. The results are compared in terms of different sparsity levels, k and number of observations, n, respectively.

Notice that the error bars in Fig. 1 and Fig. 2 indicate that our result is indeed statistically significant. Moreover, in Fig. 2, since $\|w^*\|_1$ is fixed, we have to scale the threshold, t, with sparsity level, k.

4.2. Internet Marketing Application

In this application, we need to find a small set of advertisement websites that brings the largest amount of online traffic

in the clothing market. Our assumption is that the website traffic is generated as a superposition of the traffic generated from current costumers (direct visits of the website) and the traffic redirected from advertisement websites through clicking on the ads. Although larger number of ads usually results in higher redirected traffic, the ads in different websites are not equal in their effectiveness. The amount of online traffic that an ad in a specific advertisement website brings depends on different factors such as the popularity of the advertisement website, the percentage of the target market that regularly visits the advertisement website, etc. We quantify these factors into a score, w_j, for the advertisement website j. Our goal is to find the websites with highest score.

Based on the above discussion, we model the daily website traffic, y_i, by:

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

where $\lambda_{0,i}$ models the current costumers who visit the site directly, $a_{i,j}$ is the number of ad link for the website i in the advertisement website j, and w_j is the dominance score for advertisement website j.

Our observations are daily online visits of the 50 top clothings brands. From the information provided in alexia.com, we chose the top 150 advertisement websites for these brands along with the number of ad links for each website. We recover w from constrained ML as defined in Eqn. (1).

Note that checking RE condition for a given matrix A is an NP hard problem. However, one can reasonably assume that generally fewer links are purchased from less effective advertisement websites, i.e. $A_S > A_{S^c}$. Therefore Assumption 2 is quite likely satisfied.

Before proceeding to the main results, we compare the result of ML and Lasso estimations of w^* for this problem. However, since the ground truth is not known, we use predictive likelihood test to evaluate the performances of the two methods [23]. In predictive likelihood test, we fist divide the data into two segments. We use one segment of data to estimate $\widehat{w}_{ML}$ and $\widehat{w}_{LASSO}$. Then, we use $\widehat{w}_{ML}$ and $\widehat{w}_{LASSO}$ to estimate the log likelihood of data in the second segment. Based on Bayes' rule, the model with higher predictive log likelihood is chosen . Note that this type of comparison is applicable the parameters on both models are equal. The predictive Log Likelihoods for different sparsity levels is illustrated in Fig. 3.

A brief look at Table I shows how w's have changed dramatically over time. To study this change closely, we estimated w's for different advertisement websites from 2004 to 2013. We group the Social networks, such as facebook.com, twitter.com, pinterest.com, etc, together to study the effect of Social Media Marketing (SSM). We also group search engines, such as google.com, yahoo.com, bing.com, etc, together to represent Search Engine Marketing (SEM). We add the scores of the corresponding websites in each group. Fig. 4 demonstrate the dynamics of SSM and SEM, the most controversial forms of online marketing, over time [16]. Although SEM has been thought to be the most powerful media marketing tool, recent empirical studies show the growing influence of SSM during the last couple of years [18]. The gigantic size of social media coupled with the relatively low cost per impression and the so called word of mouth have made SSM a powerful marketing tool. Our results confirms the significant influence of SSM relative to SEM since 2012.

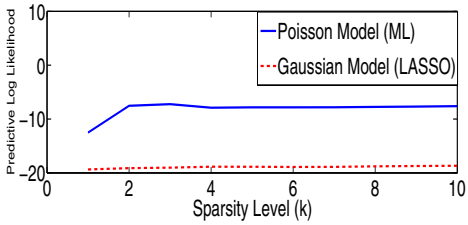

Fig. 3: *The Predictive Held-out Log Likelihood Comparison. Bayes' rule suggest that the Poisson distribution used in constrained ML approach is a better model than Gaussian distribution used in rescaled LASSO.*

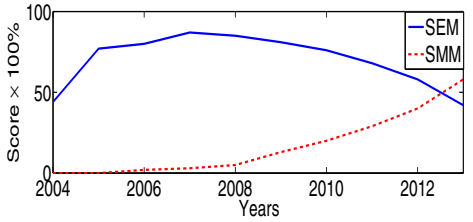

Fig. 4: *Dynamics of SEM and SSM over time for clothing market. This figure gives a quantitative comparison of the two most controversial methods of online marketing.*

Table 1: *Top advertisement websites for clothing brands in 2013*

Backward link	Estimated weights in 2013	Estimated weights in 2008
Twitter.com	0.18	0.02
Facebook.com	0.18	0.02
Pinterest.com	0.14	0.00
Amazon.com	0.28	0.22
Google.com	0.15	0.52
Bing.com	0.05	0.00
Yahoo.com	0.00	0.11

5. CONCLUSIONS

We have provided convergence guarantees for the solution of ML decoder for heterogeneous Poisson model with high dimensional sparse underlying parameter. We showed that Restricted Eigenvalue (RE) condition, which has been originally used to prove the consistency of sparse linear models, is a sufficient condition to obtain consistency and convergence results for our non-linear problem.

After verifying our theoretical results empirically, we applied an ML estimator to our online marketing application. Our results provide a quantitative analysis of the dynamics of online marketing.

6. REFERENCES

[1] D. L. Snyder, *Random Point Processes*, Wiley-Interscience Publication, 1976.

[2] Z. T. Harmany, R. F. Marcia, and R. M. Willet, "Sparse poisson intensity reconstruction algorithms," in *SSP*, 2009.

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

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

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

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

[7] 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.

[8] 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.

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

[10] P. J. Bickel, Ya, acov Ritov, and Alexandre B. Tsybakov, "Simultaneous analysis of lasso and dantzig selector," *Annals of Statistics, Volume 37, Number 4 (2009), 1705-1732*, 2009.

[11] 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.

[12] 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.

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

[14] V. Saligrama D. Motamedvaziri, M. H. Rohban, "Sparse signal recovery under poisson statistics," *arxiv*, 2013.

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

[16] 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.

[17] Kappe F. Trattner, C., "Social stream marketing on facebook: A case study," *International Journal of Social and Humanistic Computing*, 2013.

[18] M. Wasiq A. Bashar, I. Ahmad, "Effectiveness of social media as a marketing tool: Anempirical study," *International Journal of Marketing, Financial Services and Management Research*, 2012.

[19] W. K. Newey, "Uniform convergence in probability and stochastic equicontinuity," *Econometrica*, 1991.

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

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

[22] A. Takeshi, "Advanced econometrics," *Harvard University Press*, 1985.

[23] D. M. Blei and P. I. Frazier, "Distance dependent chinese restaurant processes," *Journal of Machine Learning Research*, 2011.

[24] R. E. Kass and A. E. Raftery, "Bayes factor," *Journal of the American Statistical Association, Vol. 90, No. 430. (Jun., 1995), pp. 773-795*, 1995.