Compound Poisson point processes, concentration and oracle inequalities

Huiming Zhang[1*] and Xiaoxu Wu[2]

*Correspondence:
zhanghuiming@pku.edu.cn
[1]School of Mathematical Sciences
and Center for Statistical Science,
Peking University, Beijing, China
Full list of author information is
available at the end of the article

Abstract

This note aims at presenting several new theoretical results for the compound Poisson point process, which follows the work of Zhang et al. (Insur. Math. Econ. 59:325–336, 2014). The first part provides a new characterization for a discrete compound Poisson point process (proposed by Aczél (Acta Math. Hung. 3(3):219–224, 1952)), it extends the characterization of the Poisson point process given by Copeland and Regan (Ann. Math. 37:357–362, 1936). Next, we derive some concentration inequalities for discrete compound Poisson point process (negative binomial random variable with unknown dispersion is a significant example). These concentration inequalities are potentially useful in count data regression. We give an application in the weighted Lasso penalized negative binomial regressions whose KKT conditions of penalized likelihood hold with high probability and then we derive non-asymptotic oracle inequalities for a weighted Lasso estimator.

MSC: 60G55; 60G51; 62J12

Keywords: Characterization of point processes; Poisson random measure; Concentration inequalities; Sub-gamma random variables; High-dimensional negative binomial regressions

1 Introduction

Powerful and useful concentration inequalities of empirical processes have been derived one after another for several decades. Various types of concentration inequalities are mainly based on moments condition (such as sub-Gaussian, sub-exponential and sub-Gamma) and bounded difference condition; see Boucheron et al. [4]. Its central task is to evaluate the fluctuation of empirical processes from some value (for example, the mean and median) in probability. Attention has greatly expanded in various area of research such as high-dimensional statistics, models, etc. For the Poisson or negative binomial count data regression (Hilbe [13]), it should be noted that the Poisson distribution (the limit distribution of negative binomial is Poisson) is not sub-Gaussian, but locally sub-Gaussian distribution (Chareka et al. [6]) or sub-exponential distribution (see Sect. 2.1.3 of Wainwright [34]). However, few relevant results are recognized about the concentration inequalities for a weighted sum of negative binomial (NB) random variables and its statistical applications. With applications to the segmentation of RNA-Seq data, a documental example is that Cleynen and Lebarbier [7] obtained a concentration inequality of

Springer

the sum of independent centered negative binomial random variables and its derivation hinges on the assumption that the dispersion parameter is known. Thus, the negative binomial random variable belongs to the exponential family. But, if the dispersion parameter in negative binomial random variables X is unknown in real world problems, it does not belong to the exponential family with density

$$f_X(x \mid \theta) = h(x) \exp\big(\eta(\theta)T(x) - A(\theta)\big), \tag{1}$$

where $T(x)$, $h(x)$, $\eta(\theta)$, and $A(\theta)$ are some given functions.

Whereas, it is well-known that Poisson and negative binomial distributions belong to the family of infinitely divisible distributions. Houdré [14] studies dimension-free concentration inequalities for non-Gaussian infinitely divisible random vectors with finite exponential moments. In particular, the geometric case, a special case of negative binomial, has been obtained in Houdré [14]. Via the entropy method, Kontoyiannis and Madiman [20] gives a simple derivation of the concentration inequalities for a Lipschitz function of discrete compound Poisson random variable. Nonetheless, when deriving oracle inequalities of Lasso or Elastic-net estimates (see Ivanoff et al. [16], Zhang and Jia [38]), it is strenuous to use the results of Kontoyiannis and Madiman [20], Houdré [14] to get an explicit expression of tuning parameter under the KKT conditions of the penalized negative binomial likelihood (Poisson regression is a particular case). Moreover, when the negative binomial responses are treated as sub-exponential distributions (distributions with the Cramer type condition), the upper bound of the ℓ_1-estimation error oracle inequality is determined by the tuning parameter with the rate $O(\frac{\log p}{n})$, which does not show rate-optimality in the minimax sense, although it is sharper than $O(\sqrt{\frac{\log p}{n}})$. Consider linear regressions $Y = X\beta^* + \varepsilon$ with noise $\operatorname{Var}\varepsilon = \sigma^2 I_n$ and β^* being the p-dimensional true parameter. Under the ℓ_0-constraints on β^*, Raskutti et al. [28] shows that the lower bound on the minimax prediction error for all estimators $\hat{\beta}$ is

$$\inf_{\hat{\beta}} \sup_{\|\beta^*\|_0 \le s_0} \frac{1}{n} \big\| X\hat{\beta} - X\beta^* \big\|_2^2 \ge O\left(\frac{s_0 \log(p/s_0)}{n} \right), \quad (X \text{ is an } n \times p \text{ design matrix})$$

with probability at least $1/2$. Thus, the optimal minimax rate for $\|\hat{\beta} - \beta^*\|_1$ should be proportional to the root of the rate $O(\frac{\log p}{n})$. Similar minimax lower bound for Poisson regression is provided in Jiang et al. [17]; see Chap. 4 of Rigollet and Hütter [30] for more minimax theory.

This paper is prompted by the motif of high-dimensional count regression that derives useful NB concentration inequalities which can be obtained under compound Poisson frameworks. The obtained concentration inequalities are beneficial to optimal high-dimensional inferences. Our concentration inequalities are about the sum of independent centered weighted NB random variables, which is more general than the un-weighted case in Cleynen and Lebarbier [7], which supposes that the over-dispersion parameter of NB random variables is known. However, in our case, the over-dispersion parameter can be unknown. As a by-product, we proposed a characterization for discrete compound Poisson point processes following the work in Zhang et al. [40] and Zhang and Li [39]. The paper will end up with some applications of high-dimensional NB regression. Städler et al. [32] studies the oracle inequalities for ℓ_1-penalization finite mixture of Gaussian regressions model but they didn't concern the mixture of count data regression (such as mixture

Poisson, see Yang et al. [36] for example). Here the NB regression can be approximately viewed as finite mixture of Poisson regression since NB distribution is equivalent to a continuous mixture of Poisson distributions where the mixing distribution of the Poisson rate is gamma distributed. The paper will end up with some applications of high-dimensional NB regression in terms of oracle inequalities.

The paper is organized as follows. Section 2 is an introduction to discrete compound Poisson point processes (DCPP). We give a theorem for characterizing DCPP, which is similar to the result that initial condition, stationary and independent increments, condition of jumps, together quadruply characterize a compound Poisson process, see Wang and Ji [35] and the references therein. In Sect. 3, we derive the concentration inequalities for discrete compound Poisson point process with infinite-dimensional parameters. As an application, for the optimization problem of weighted Lasso penalized negative binomial regression, we show that the true parameter version of KKT conditions hold with high probability, and the optimal data-driven weights in weighted Lasso penalty are determined. We also show the oracle inequalities for weighted Lasso estimates in NB regressions, and the convergence rate attain the minimax rate derived in references.

2 Introductions to discrete compound Poisson point process

In this section, we present the preliminary knowledge for discrete compound Poisson point process. A new characterization of the point process with five assumptions is obtained.

2.1 Definition and preliminaries

To begin with, we need to be familiar with the definition of the discrete compound Poisson distribution and its relation to the weighted sum of Poisson random variables. For more details, we refer the reader to Sect. 9.3 of Johnson et al. [18], Zhang et al. [40], Zhang and Li [39] and the references therein.

Definition 2.1 We say that Y is discrete compound Poisson (DCP) distributed if the characteristic function of Y is

$$\varphi_Y(t) = \mathrm{E}e^{\mathrm{i}tY} = \exp\left\{\sum_{i=1}^{\infty} \alpha_i \lambda \left(e^{\mathrm{i}t\theta} - 1\right)\right\} \quad (\theta \in \mathbb{R}), \tag{2}$$

where $(\alpha_1\lambda, \alpha_2\lambda, \ldots)$ are infinite-dimensional parameters satisfying $\sum_{i=1}^{\infty} \alpha_i = 1$, $\alpha_i \geq 0$, $\lambda > 0$. We denote it by $Y \sim \mathrm{DCP}(\alpha_1\lambda, \alpha_2\lambda, \ldots)$.

If $\alpha_i = 0$ for all $i > r$ and $\alpha_r \neq 0$, we say that Y is DCP distributed of order r. If $r = +\infty$, then the DCP distribution has infinite-dimensional parameters and we say it is DCP distributed of order $+\infty$. When $r = 1$, it is the well-known Poisson distribution for modeling equal-dispersed count data. When $r = 2$, we call it the Hermite distribution, which can be applied to model the over-dispersed and multi-modal count data; see Giles [11].

Equation (2) is the canonical representation of characteristic function for non-negative valued infinitely divisible random variable, i.e. a special case of the Lévy–Khinchine formula (see Chap. 2 of Sato [31] and Sect. 1.6 of Petrov [27]),

$$\mathrm{E}e^{\mathrm{i}tY} = \exp\left(a\mathrm{i}t - \frac{1}{2}\sigma^2 t^2 + \int_{\mathbb{R}\setminus\{0\}} \left(e^{\mathrm{i}tx} - 1 - \mathrm{i}tx\mathrm{I}_{|x|<1}(x)\right)\nu(dx)\right),$$

where $a \in \mathbb{R}$, $\sigma \geq 0$ and $\mathrm{I}_{\{\cdot\}}(x)$ is indicator function, the function $\nu(\cdot)$ is called the Lévy measure with restriction: $\int_{\mathbb{R} \setminus \{0\}} \min(x^2, 1)\nu(dx) < \infty$.

Let $\delta_x(t) := 1_{\{x\}}(t)$ and $\mathbb{N}^+$ be the positive integer set. If we set $\nu(t) = \sum_{x \in \mathbb{N}^+} \lambda \delta_x(t)\alpha_x$, the $\nu(\cdot)$ is the Lévy measure in the Lévy–Khinchine formula. It is easy to see that Y has a weighted Poisson decomposition: $Y = \sum_{i=1}^{\infty} iN_i$, where the N_i are independent Poisson distributed with mean $\lambda\alpha_i$. This decomposition is also called a Lévy–Ito decomposition; see Chap. 4 of Sato [31]. If $\mathrm{E}e^{-\theta Y} < \infty$ for θ in a neighbourhood of zero, the moment generating function (m.g.f.) of DCP is

$$M_Y(\theta) := \mathrm{E}e^{-\theta Y} = \exp\left\{\sum_{k=1}^{\infty} \alpha_k \lambda \left(e^{-k\theta} - 1\right)\right\} = \exp\left\{\int_0^{\infty} \left(e^{-\theta x} - 1\right)\nu(dx)\right\} \quad (|\theta| < R)$$

for some $R > 0$.

In order to define the discrete compound Poisson point process, we shall repeatedly use the concept of Poisson random measure. Let A be any measurable set on a measurable space $(E, \mathcal{A})$, a good example could be $E = \mathbb{R}^d$. In the following sections, we let $E = \mathbb{R}^d$ and denote $N(A, \omega)$ as the number of random points in set A. We introduce the Poisson point processes below, and sometimes it is called the Poisson random measure (Sato [31], Kingman [19]).

Definition 2.2 Let $(E, \mathcal{A}, \mu)$ be a measurable space with σ-finite measure μ, and $\mathbb{N}$ be the non-negative integer set. The Poisson random measure with intensity μ is a family of random variables $\{N(A, \omega)\}_{A \in \mathcal{A}}$ (defined on some probability space $(\Omega, \mathcal{F}, P)$) as the product map

$$N : \mathcal{A} \times \Omega \to \mathbb{N}$$

satisfying
1. $\forall A \in \mathcal{A}$, the $N(A, \cdot)$ is Poisson random variable on $(\Omega, \mathcal{F}, P)$ with mean $\mu(A)$;
2. $\forall \omega \in \Omega$, the $N(\cdot, \omega)$ is a counting measure on $(E, \mathcal{A})$;
3. If sets $A_1, A_2, \ldots, A_n \in \mathcal{A}$ are disjoint, then $N(A_1, \cdot), N(A_2, \cdot), \ldots, N(A_n, \cdot)$ are mutually independent.

Let $N(\mathcal{A}) := N(\mathcal{A}, \omega)$. Based on the Poisson random measure, we define the discrete compound Poisson point processes (DCPP) $\{CP(A)\}_{A \in \mathcal{A}}$ by the Lévy–Ito decomposition

$$CP(A) := \sum_{k=1}^{\infty} kN_k(A),$$

where the $N_k(A)$ are independent Poisson point process with mean measure $\mu_k(A) := \alpha_k \times \int_A \lambda(x)\,dx$.

The m.g.f. of $\{CP(A)\}_{A \in \mathcal{A}}$ is $M_{CP(A)}(\theta) = \exp\{\sum_{k=1}^{\infty} \alpha_k \int_A \lambda(x)\,dx(e^{-k\theta} - 1)\}$. We can see that $\lambda(\cdot)$ is the intensity function of the *generating Poisson point processes* $\{N_k(A)\}_{A \in \mathcal{A}}$ for each k. Aczél [1] derives the p.m.f. and he calls it the inhomogeneous composed Poisson distribution as $d = 1$. If the intensity function $\lambda(x)$ is the constant (i.e. the mean measure is a multiple of Lebesgue measure), the N is said to be a homogeneous Poisson point process.

Define the probability $P_k(A)$ by $P(N(A) = k)$. If $N(A)$ follows from a DCP point process with intensity measure $\lambda(A)$, then the probability of having k $(k \geq 0)$ points in the set A is given by

$$P_k(A) = \sum_{R(s,k)=k} \frac{\alpha_1^{s_1} \cdots \alpha_k^{s_k}}{s_1! \cdots s_k!} \left[\lambda(A)\right]^{s_1 + \cdots + s_k} e^{-\lambda(A)}, \tag{3}$$

where $R(s,k) := \sum_{t=1}^{k} t s_t$; see Aczél [1] for the proof.

2.2 A new characterisation of DCP point process

Based on preliminaries, we give a new characterization of DCP point process, which is an extension of Copeland and Regan [8]. A similar characterization for DCP distribution and process (not for point process in terms of random measure) is derived by Wang and Ji [35]. For more characterizations of the point process, see the monograph Last and Penrose [21] and the references therein.

Theorem 2.1 (Characterization for DCP point process) *Consider the following assumptions*:

1. $P_k(A) > 0$ *if* $0 < m(A) < \infty$, *where* $m(\cdot)$ *is the Lebesgue measure*;
2. $\sum_{k=0}^{\infty} P_k(A) = 1$;
3. $P_k(A_1 \cup A_2) = \sum_{i=0}^{k} P_{k-i}(A_1) P_i(A_2)$ *if* $A_1 \cap A_2 = \emptyset$;
4. *Let* $S_k(A) = \sum_{i=k}^{\infty} P_i(A)$, *then* $\lim_{m(A) \to 0} S_1(A) = 0$;
5. $\lim_{m(A) \to 0} \frac{P_k(A)}{S_1(A)} = \alpha_k$, *where* $\sum_{k=1}^{\infty} \alpha_k = 1$.

If $P_k(A)$ *satisfies the assumptions* 1–5, *then there exists a measure* $\nu(A)$ *(countable additivity, absolutely continuous) such that* $P_k(A)$ *is represented by the equation* (3) *for all* k *and* A.

For A being the closed interval in $\mathbb{R}$, this setting turns to the famous characterization of the Poisson process; see p. 447 of Feller [10] (The postulates for the Poisson process). The proof of Theorem 2.1 is given in the Appendix. The proof consists of showing countably additive and absolutely continuous of the intensity measure and dealing with the matrix differential equation.

As a random measure, the DCPP is a mathematical extension from discrete compound Poisson distribution and discrete compound Poisson process index by real line. It has been studied by Baraud and Birgé [2] that they use histogram type estimator to estimate the intensity of the Poisson random measure. A recent application, Das [9] proposes a nonparametric Bayesian approach to RNA-seq count data by DCP process. For a set of regions $\{A \in \mathcal{A}\}$ of interest on the genome, such as genes, exons, or junctions, the number of reads is counted, and the count in the region A_i is viewed as a DCP process for gene expression. The $P_k(A)$ satisfies assumptions 1–5 in Theorem 2.1 by the real-world setting (the process of generating the expression level for each gene in the experiment).

3 Concentration inequalities for discrete compound Poisson point process

This section is about the construction of concentration inequalities for DCPP. It has applications in high-dimensional ℓ_1-regularized negative binomial regression; see Sect. 4.

As an appetizing example, if the DCPP has finite parameters, the existing result such as Lemma 3 of Baraud and Birgé [2] can be used to obtain the concentration inequality for the sum of DCP random variables.

Lemma 3.1 (Baraud and Birgé [2]) *Let $Y_1, \ldots, Y_n$ be n independent centered random variables, and $\tau, \kappa, \{\eta_i\}_{i=1}^n$ be some positive constants.*

(a) *If $\log(\mathrm{E}e^{zY_i}) \leq \kappa \frac{z^2 \eta_i}{2(1-z\tau)}$ for all $z \in [0, 1/\tau[$, and $1 \leq i \leq n$, then*

$$P\left[\sum_{i=1}^n Y_i \geq \left(2\kappa x \sum_{i=1}^n \eta_i\right)^{1/2} + \tau x\right] \leq e^{-x} \quad \text{for all } x > 0.$$

(b) *If for $1 \leq i \leq n$ and all $z > 0$ $\log(\mathrm{E}e^{-zY_i}) \leq \kappa z^2 \eta_i / 2$, then*

$$P\left[\sum_{i=1}^n Y_i \leq -\left(2\kappa x \sum_{i=1}^n \eta_i\right)^{1/2}\right] \leq e^{-x} \quad \text{for all } x > 0.$$

Actually, the moment conditions in Lemma 3.1 is a direct application of the existing result based on Cramer-type conditions (Lemma 2.2 in Petrov [27], or sub-exponential condition (4) discussed below) for the convolutions' Poisson distribution with different rates. The centered random variable Y_i with the moment condition $\log(\mathrm{E}e^{zY}) \leq \kappa \frac{z^2 \eta_i}{2(1-z\tau)}$ in Lemma 3.1(a), is called the *sub-Gamma* random variables (see Sect. 2.4 in Boucheron et al. [4]). A random variable X with zero mean is *sub-exponential* (see Sect. 2.1.3 in Wainwright [34]) if there are non-negative parameters (ν, α) such that

$$\mathrm{E}e^{\lambda X} \leq e^{\frac{\nu^2 \lambda^2}{2}} \quad \text{for all } |\lambda| < \frac{1}{\alpha}. \tag{4}$$

Note that the sub-exponential implies the sub-gamma condition by the following inequality: $\log(\mathrm{E}e^{-zY_i}) \leq \frac{\kappa z^2 \eta_i}{2} \leq \frac{\kappa z^2 \eta_i}{2(1-z\tau)}$ for all $z \in [0, 1/\tau[$.

Proposition 3.1 *Let $Y_i \sim \mathrm{DCP}(\alpha_1(i)\lambda(i), \ldots, \alpha_r(i)\lambda(i))$ for $i = 1, 2, \ldots, n$, and $\sigma_i := \mathrm{Var}\, Y_i = \lambda(i) \sum_{k=1}^r k^2 \alpha_k(i)$, then for all $x > 0$*

$$P\left[\sum_{i=1}^n (Y_i - \mathrm{E}Y_i) \geq \sqrt{2x \sum_{i=1}^n \sigma_i^2 + rx}\right] \leq e^{-x},$$

$$P\left[\sum_{i=1}^n (Y_i - \mathrm{E}Y_i) \leq -\sqrt{2x \sum_{i=1}^n \sigma_i^2}\right] \leq e^{-x}.$$

Moreover, we have

$$P\left[\left|\sum_{i=1}^n (Y_i - \mathrm{E}Y_i)\right| \geq \sqrt{2x \sum_{i=1}^n \sigma_i^2 + rx}\right] \leq 2e^{-x}. \tag{5}$$

Proof First, in order to apply the above lemma, we need to evaluate the log-moment-generating function of centered DCP random variables. Let $\mu_i =: \mathrm{E}Y_i = \lambda(i) \sum_{k=1}^r k\alpha_k(i)$,

we have

$$\log \mathrm{E}e^{z(Y_t-\mu_i)} = -z\mu_i + \log \mathrm{E}e^{zY_i} = -z\mu_i + \log e^{\lambda(i)\sum_{k=1}^{r}\alpha_k(i)(e^{kz}-1)}$$

$$= \lambda(i)\sum_{k=1}^{r}\alpha_k(i)\big(e^{kz} - kz - 1\big).$$

Using the inequality $e^z - z - 1 \le \frac{z^2}{2(1-z)}$ for $1 > z > 0$, one derives

$$\log \mathrm{E}e^{z(Y_t-\mu_i)} \le \lambda(i)\sum_{k=1}^{r}\alpha_k(i)\frac{k^2z^2}{2(1-kz)} \le \lambda(i)\sum_{k=1}^{r}\alpha_k(i)\frac{k^2z^2}{2(1-rz)} =: \frac{\sigma_i z^2}{2(1-rz)} \quad (z > 0),$$

where $\sigma_i = \mathrm{Var}\, Y_i = \lambda(i)\sum_{k=1}^{r} k^2\alpha_k(i)$. And by applying the inequality $e^z - z - 1 \le \frac{z^2}{2}$ for $z < 0$, we get

$$\log \mathrm{E}e^{z(Y_t-\mu_i)} \le \lambda(i)\sum_{k=1}^{r}\alpha_k(i)\frac{k^2z^2}{2} =: \frac{\sigma_i z^2}{2} \quad (z < 0). \tag{6}$$

Thus (6) implies $\log \mathrm{E}e^{-z(Y_t-\mu_i)} \le \frac{\sigma_i z^2}{2}$ for $z > 0$.

Therefore, by letting $\tau = 1$, $\kappa = 1$, $\eta_i = \sigma_i^2$ in Lemma 3.1, we have (5). □

In the sequel, without employing Cramer-type conditions, we will derive a Bernstein-type concentration inequality for DCPP with infinite-dimensional parameters $r = +\infty$. Theorem 3.1 below is a weighted-sum version that is not the same as Proposition 3.1.

Theorem 3.1 *Let $f(x)$ be the measurable function such that $\int_E f^2(x)\lambda(x)\,dx < \infty$, where the $\lambda(x)$ is the intensity function of the DCPP with parameter $(\alpha_1\int_E \lambda(x)\,dx, \alpha_2\int_E \lambda(x)\,dx, \ldots)$ under the condition $\sum_{k=1}^{\infty} k\alpha_k < \infty$. Then the concentration inequality for a stochastic integral of DCPP is*

$$P\left(\int_E f(x)\left[CP(dx) - \left(\sum_{k=1}^{\infty} k\alpha_k\right)\lambda(x)\,dx\right] \ge \left(\sum_{k=1}^{\infty} k\alpha_k\right)\left[\sqrt{2yV_f} + \frac{y}{3}\|f\|_\infty\right]\right)$$

$$\le e^{-y} \quad (y > 0), \tag{7}$$

where $V_f = \int_E f^2(x)\lambda(x)\,dx$, $\|f\|_\infty$ is the supremum of $f(x)$ on E.

Let $f(x) = 1_A(x)$ in Theorem 3.1, and we have $P(Y - \mathrm{E}Y \ge \frac{\mathrm{E}Y}{\lambda}[\sqrt{2y\lambda} + \frac{y}{3}]) \le e^{-y}$ for $Y \sim$ DCP$(\alpha_1\lambda, \alpha_2\lambda, \ldots)$.

From the guide of the proof in Corollary 3.1 below, it is not hard to get the Theorem 3.1 by a linear combination of random variables which is more important in applications in Sect. 4. The proof of Theorem 3.1 is given in the Appendix. This is essentially from the fact that the infinitely divisible distributions are closed under the convolutions and scalings. Corollary 3.1 is the version for the sum of n independent random measures.

Corollary 3.1 *For a given set $\{S_i\}_{i=1}^n$ in E, if we have n independent DCPP $\{CP_i(S_i)\}_{i=1}^n$ with a series of parameters*

$$\left\{ \left(\alpha_1(i) \int_{S_i} \lambda(x)\,dx, \alpha_2(i) \int_{S_i} \lambda(x)\,dx, \dots \right) \right\}_{i=1}^n$$

correspondingly, then, for a series of measurable function $\{f_i(x)\}_{i=1}^n$, we have

$$P\left(\sum_{i=1}^n \left\{ \int_{S_i} f_i(x)\big[CP_i(dx) - \mu_i \lambda(x)\,dx\big] \right\} \geq \sum_{i=1}^n \mu_i \left(\sqrt{2yV_{i,f_i}} + \frac{y}{3}\|f_i\|_\infty \right) \right)$$

$$\leq e^{-ny} \quad (y > 0),$$

$$(8)$$

where $\mu_i := \sum_{k=1}^\infty k\alpha_k(i) < \infty$ and $V_{i,f_i} := \int_{S_i} f_i^2(x)\lambda(x)\,dx < \infty$.
Moreover, let $c_{1n} := \sum_{i=1}^n \mu_i\sqrt{2V_{i,f_i}}$, $c_{2n} := \sum_{i=1}^n \frac{\mu_i}{3}\|f_i\|_\infty$, then

$$P\left(\sum_{i=1}^n \left\{ \int_{S_i} f_i(x)\big[CP_i(dx) - \mu_i \lambda(x)\,dx\big] \right\} \geq t \right)$$

$$\leq \exp\left\{ -n\left(\sqrt{\frac{t}{c_{2n}} + \frac{c_{1n}^2}{4c_{2n}^2}} - \frac{c_{1n}}{2c_{2n}} \right)^2 \right\}.$$

$$(9)$$

The difference between Proposition 3.1 and Corollary 3.1 is that the DCP random variables in Proposition 3.1 have order $r < +\infty$ while Corollary 3.1 is related to DCP point process (a particular case of DCPP is the constant intensity $\lambda(A) = \lambda$) with order $r = +\infty$. Moreover, Proposition 3.1 require dependent variance factors σ_i^2 for DCP random variables.

Remark 1 Significantly, Corollary 3.1 is the weighted-sum version, and it improves the results in Cleynen and Lebarbier [7] who deals with the un-weighted sum of NB random variables (as a special case of DCP distribution). Moreover, Cleynen and Lebarbier's [7] result relies on Lemma 3.1, which depends on the sub-Gamma condition. It should be noted that we do not impose any bounded restriction for y in our concentration inequalities (7) and (8), while Theorem 1 of Houdré [14] and concentration inequalities of Houdré and Privault [15] require that the infinitely divisible random vector satisfies the Cramer-type conditions and the y in $P(X - EX \geq y)$ is bounded by a given constant. The moment conditions in Lemma 3.1 sometimes are difficult to check, being hard to apply in the Lasso KKT condition in Sect. 4.1 if the count data are other infinitely divisible discrete distributions. In our Theorem 3.1 and Corollary 3.1, we only need to check that the means of the DCP processes are finite and Theorem 3.1 and Corollary 3.1 do not need the Bernstein-type condition (see Theorem 2.8 in Petrov [27] and p. 27 of Wainwright [34]). In addition, our concentration result does not require the higher moments conditions proposed by Kontoyiannis and Madiman [20], they assume that $EZ^L = \sum_{k=1}^\infty k^L\alpha_k < \infty$ $(L > 1)$ about the compound distribution $P(Z = i) := \alpha_i$ for the given discrete random variable Z.

Before we give the proof of Corollary 3.1, we need some notations and a lemma. For a general point process $N(A)$ defined on $\mathbb{R}^d$, let f be any non-negative measurable function on E. The Laplace functional is defined by $L_N(f) = E[\exp\{-\int_E f(x)N(dx)\}]$, it is the stochastic integral is specified by $\int_E f(x)N(dx) =: \sum_{x_i \in \Pi \cap E} f(x_i)$, where the Π is the state

space of the point process $N(A)$. The Laplace functional for a random measure is a crucial numerical characteristics that enables us to handle stochastic integral of the point process. Moreover, we rely on the following theorem due to Campbell.

Lemma 3.2 (See Sect. 3.3 in Kingman [19]) *Let $N(A)$ be a Poisson random measure with intensity $\lambda(t)$ and a measurable function $f : \mathbb{R}^d \to \mathbb{R}$, the random sum $S = \sum_{x \in \Pi} f(x)$ is absolutely convergent with probability one if and only if the integral $\int_{\mathbb{R}^d} \min(|f(x)|, 1)\lambda(x)\,dx < \infty$. Assume this condition holds, thus we assert that, for any complex number value θ, the equation*

$$\mathrm{E}e^{\theta S} = \exp\left\{\int_{\mathbb{R}^d}\left[e^{\theta f(x)} - 1\right]\lambda(x)\,dx\right\}$$

holds if the integral on the right-hand side is convergent.

Proof Now, we can easily compute the DCPP's Laplace functional by Campbell's theorem. Set $CP_r(A) =: \sum_{k=1}^{r} kN_k(A)$. And let $\theta = -1$, that is, $\mathrm{L}_N(f) = \exp\{\int_E [e^{-f(x)} - 1]\lambda(x)\,dx\}$. By independence, it follows that

$$
\begin{aligned}
\mathrm{L}_{CP_r}(f) &= \mathrm{E}\exp\left\{-\int_E f(x)CP_r(dx)\right\} = \mathrm{E}\exp\left\{-\int_E f(x)\sum_{k=1}^{r} kN_k(dx)\right\}\\
&= \prod_{k=1}^{r}\mathrm{E}\exp\left\{-\int_E kf(x)N_k(dx)\right\}\\
&= \prod_{k=1}^{r}\exp\left\{\int_E \left[e^{-kf(x)} - 1\right]\alpha_k\lambda(x)\,dx\right\}\\
&= \exp\left\{\sum_{k=1}^{r}\int_E \left[e^{-kf(x)} - 1\right]\alpha_k\lambda(x)\,dx\right\}.
\end{aligned}
$$

The above Laplace functional of a compound random measure seems like a Lévy–Khintchine representation. Define for $\eta > 0$

$$
\begin{aligned}
D_r(\eta) := \exp\Bigg\{&\left\{\int_E \eta f(x)\left[CP_r(dx) - \left(\sum_{k=1}^{r} k\alpha_k\right)\lambda(x)\right]dx\right\}\\
&- \int_E \sum_{k=1}^{r}\left[e^{kf(x)} - k\eta f(x) - 1\right]\alpha_k\lambda(x)\,dx\Bigg\}.
\end{aligned}
\tag{10}
$$

Therefore, we obtain $\mathrm{E}D_r(\eta) = 1$.

It follows from (10) and the Markov's inequality that

$$
\begin{aligned}
P\Bigg(&\int_E \eta f(x)\left[CP_r(dx) - \left(\sum_{k=1}^{r} k\alpha_k\right)\lambda(x)\,dx\right]\\
&\geq \int_E \sum_{k=1}^{r}\left[e^{kf(x)} - k\eta f(x) - 1\right]\alpha_k\lambda(x)\,dx + y\Bigg)\\
&= P\big(D(\eta) \geq e^y\big) \leq e^{-y}.
\end{aligned}
\tag{11}
$$

Note that

$$
\begin{aligned}
e^{k\eta f(x)} - k\eta f(x) - 1 &\le \sum_{i=2}^{\infty} \frac{\eta^i}{i!} \|kf\|_{\infty}^{i-2} k^2 f^2(x) = k^2 f^2(x) \sum_{i=2}^{\infty} \frac{\eta^i}{i(i-1)\cdots 2} \|kf\|_{\infty}^{i-2} \\
&\le \frac{k^2 \eta^2 f^2(x)}{2} \sum_{i=2}^{\infty} \left(\frac{1}{3} k\eta \|f\|_{\infty} \right)^{i-2} \\
&\le \frac{k^2 \eta^2 f^2(x)}{2} \bigg/ \left(1 - \frac{1}{3} k\eta \|f\|_{\infty} \right).
\end{aligned}
\tag{12}
$$

Hence, using (12), (11) turns to

$$
\begin{aligned}
P\!\left(\int_E f(x) \left[CP_r(dx) - \left(\sum_{k=1}^r k\alpha_k \right) \lambda(x)\,dx \right] \ge \int_E \sum_{k=1}^r \alpha_k \left(\frac{1}{2} \frac{k^2 \eta f^2(x)\lambda(x)}{1-\frac{1}{3}k\eta\|f\|_{\infty}}\,dx + \frac{y}{\eta} \right) \right) \\
\le e^{-y}.
\end{aligned}
\tag{13}
$$

Let $V_f = \int_E f^2(x)\lambda(x)\,dx$, notice that

$$
\begin{aligned}
\sum_{k=1}^r \alpha_k & \left(\frac{1}{2} \frac{\eta k^2 V_f}{1-\frac{1}{3}k\eta\|f\|_{\infty}} + \frac{y}{\eta} \right) \\
&= \sum_{k=1}^r \alpha_k \left(\frac{1}{2} \frac{\eta k^2 V_f}{1-\frac{1}{3}k\eta\|f\|_{\infty}} + \frac{y}{\eta}\left(1 - \frac{1}{3}k\eta\|f\|_{\infty}\right) + \frac{ky}{3}\|f\|_{\infty} \right) \\
&\ge \sum_{k=1}^r \alpha_k \left(k\sqrt{2V_f y} + \frac{ky}{3}\|f\|_{\infty} \right) = \left(\sum_{k=1}^r k\alpha_k \right)\left[\sqrt{2V_f y} + \frac{y}{3}\|f\|_{\infty} \right].
\end{aligned}
\tag{14}
$$

Applying (14), so by minimizing η in (13), we have

$$
\begin{aligned}
P\!\left(\int_E f(x) \left[CP_r(dx) - \left(\sum_{k=1}^r k\alpha_k \right)\lambda(x)\,dx \right] \ge \left(\sum_{k=1}^r k\alpha_k \right)\left[\sqrt{2yV_f} + \frac{y}{3}\|f\|_{\infty} \right] \right) \\
\le e^{-y}.
\end{aligned}
\tag{15}
$$

Letting $r \to \infty$ in (15), then $CP_r(A) \xrightarrow{d} CP(A)$, so we finally prove the inequality (7).   □

4 Application
4.1 Application to KKT conditions
For negative binomial random variable NB, the probability mass function (p.m.f.) is

$$
P(NB = n) = \frac{\Gamma(n+s)}{\Gamma(s)n!}(1-q)^s q^n \quad \left(q \in (0,1), n \in \mathbb{N} \right),
$$

where s is a certain real value number. When s is positive integer, the negative binomial distribution reduces to Pascal distribution which is modeled as the number of failures before the s-th success in repeated mutually independent Bernoulli trials (with probability of success $p = 1 - q$).

The m.g.f. of *NB* is

$$\varphi_{NB}(t) = \exp\left\{\log\left(\frac{1-q}{1-qe^{it}}\right)^s\right\} = \exp\left\{\sum_{k=1}^{\infty} -s\log(1-q)\cdot\frac{q^k}{-k\log(1-q)}\left(e^{ikt}-1\right)\right\}. \quad (16)$$

Here the parametrization of DCP is $NB \sim \text{DCP}(\alpha_1\lambda, \alpha_2\lambda, \ldots)$ with $\lambda = -s\log(1-q)$, $\alpha_k = \frac{q^k}{-k\log(1-q)}$.

Given the data $\{Y_i, x_i\}_{i=1}^n$, the NB regression models suppose that the responses $\{Y_i\}_{i=1}^n$ is NB distributed random variables with mean $\{\mu_i\}_{i=1}^n$, and the vectors covariates $x_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p$ are p dimensional fixed (non-random) variable for simplicity; see the so-called NB2 model in Hilbe [13] for details. Based on the log-linear model in regression analysis, we usually suppose that $\log\mu_i = f_0(x_i)$ is a linear function of x_i.

Two types of high-dimensional setting for count data regression are classified as follows.

- *Sparse approximation for nonparametric regression.*
  Sometimes, it is too rough and too simple to use a linear function to approximate $\log\mu_i$ by a function of x_i, denoted as $f_0(x_i)$. The connection between $\log\mu_i$ and x_i is sometimes unknown, thus it is unlikely to be linear. The dimension of x_i would be much larger than the sample size when the f is extremely flexible and complex. In order to capture the unknown functional relation $f_0(x_i)$, we prefer to use a given dictionary (orthogonal base functions) $\mathbb{D} = \{\phi_j(\cdot)\}_{j=1}^p$ such that the linear combination $\hat{f}(x_i) = \sum_{j=1}^p \hat{\beta}_j\phi_j(x_{ij})$ is a sparse approximation to $f_0(x_i)$ which contains increasing-dimensional parameter $\beta^* := (\beta_1^*, \ldots, \beta_p^*)^T$ with $p := p_n \to \infty$ as $n \to \infty$. A crucial point for orthogonal base functions $\mathbb{D}$ is that many $f_0(x_i)$, which have no sparse representations in the non-orthogonal base, should be represented as sparse linear combinations of $\mathbb{D}$ in the orthogonal scenario; see Sect. 10 of Hastie et al. [12] for details. In practice, the advantage of the dictionary approach is that the flexible $f_0(x_i)$ may admit a good approximation sparse in some dictionary functions while not in others. For high-dimensional count data, Ivanoff et al. [16] mentioned that: "The richer the dictionary, the sparser the decomposition, so that p can be larger than n and the model becomes high-dimensional". Typical candidates of the dictionary are Fourier basis, cosine basis, Legendry base, wavelet basis (Haar basis), etc.
- *High-dimensional features in gene engineering.*
  In the big data era, one challenge case we encountered in massive data sets is that the number of given covariates p is larger than the sample size n and the responses of interest are measured as counts. For example, in gene engineering, the covariates (features) are types of genes impacting on specific count phenotypes. The NB regression is a flexible framework for the analysis of over-dispersed counts RNA-seq data; see Li [23], Mallick and Tiwari [25] for more details.

We assume that the expectation of Y_i will be related to $\phi^T(x_i)\beta$ after a log-transformation,

$$\mu_i := \exp\left\{\sum_{j=1}^p \beta_j\phi_j(x_{ij})\right\} := e^{\phi^T(x_i)\beta} =: h(x_i),$$

where $\phi(x_i) = (\phi_1(x_{i1}), \ldots, \phi_p(x_{ip}))^T$ with covariate $\{\phi_j(x_{ij})\}$ being the jth component of $\phi(x_i)$. Here $\{\phi_j(\cdot)\}_{j=1}^p$ are given transformations of bounded covariances. A trivial example is $\phi_j(x) = x$.

Let $\{Y_i\}_{i=1}^n$ be a NB random variable with failure parameters $\{q_i = \frac{\mu_i}{\theta + \mu_i}\}$ where $\theta > 0$ is an unknown dispersion parameter which can be estimated (see Hilbe [13]). Then the p.m.f.s of $\{Y_i\}_{i=1}^n$ are

$$f(y_i; \mu_i, \theta) = \frac{\Gamma(\theta + y_i)}{\Gamma(\theta) y_i!} \left(\frac{\mu_i}{\theta + \mu_i}\right)^{y_i} \left(\frac{\theta}{\theta + \mu_i}\right)^{\theta}$$

$$\propto \exp\{y_i \phi^T(x_i)\beta - (\theta + y_i)\log(\theta + e^{\phi^T(x_i)\beta})\}, \tag{17}$$

where $\mu_i > 0$ is the parameter with $\mu_i = \mathrm{E}Y_i$ and $\mathrm{Var}\, Y_i = \mu_i + \frac{\mu_i^2}{\theta} > \mathrm{E}Y_i$.

The log-likelihood of the predictor variable $\{Y_i\}_{i=1}^n$ is

$$l(\beta) =: \log\left[\prod_{i=1}^n f(Y_i; f_0(x_i), \theta)\right] = \sum_{i=1}^n \left[Y_i \phi^T(x_i)\beta - (\theta + Y_i)\log(\theta + e^{\phi^T(x_i)\beta})\right]. \tag{18}$$

Let $R(y, \beta, x) = y\phi^T(x)\beta - (\theta + y)\log(\theta + e^{\phi^T(x)\beta})$ denote the negative average empirical risk function by $\ell(\beta) = -\frac{1}{n}l(\beta) := \mathbb{P}_n R(Y, \beta, x)$.

Having obtained high-dimensional covariates, a principal task is to select important variables. Here, we prefer the weighted ℓ_1-penalized likelihood principle for sparse NB regression,

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\mathrm{argmin}}\left\{\ell(\beta) + \sum_{j=1}^p w_j |\beta_j|\right\}, \tag{19}$$

where $\{w_j\}_{j=1}^p$ are data-dependent weights which are supposed to be specified in the following discussion.

By using KKT conditions (i.e. the first order condition in convex optimization; see p. 68 of Bühlmann and van de Geer [5]), the $\hat{\beta}$ is a solution of (19) iff $\hat{\beta}$ satisfies first order conditions

$$\begin{cases} -\dot{\ell}_j(\hat{\beta}) = w_j \, \mathrm{sign}(\hat{\beta}_j) & \text{if } \hat{\beta}_j \neq 0, \\ |\dot{\ell}_j(\hat{\beta})| \leq w_j & \text{if } \hat{\beta}_j = 0, \end{cases} \tag{20}$$

where $\dot{\ell}_j(\hat{\beta}) := \frac{\partial \ell(\hat{\beta})}{\partial \hat{\beta}_j}$ $(j = 1, 2, \ldots, p)$.

Let $\Phi(X)$ be the $n \times p$ dictionary design matrix given by

$$\Phi(X) =: (\phi_j(x_{ij}))_{1 \leq i \leq n, 1 \leq j \leq p} = \begin{pmatrix} \phi_1(x_{11}) & \cdots & \phi_p(x_{1p}) \\ \vdots & & \vdots \\ \phi_1(x_{n1}) & \cdots & \phi_p(x_{np}) \end{pmatrix} =: (\phi^T(x_1), \ldots, \phi^T(x_n))^T.$$

The KKT conditions implies $|\dot{\ell}_j(\hat{\beta})| \leq w_j$ for all j. And

$$\dot{\ell}(\beta) := \frac{\partial \ell(\beta)}{\partial \beta} = -\frac{1}{n}\sum_{i=1}^n \phi(x_i)\left[Y_i - (\theta + Y_i)\frac{e^{\phi^T(x_i)\beta}}{\theta + e^{\phi^T(x_i)\beta}}\right]$$

$$= -\frac{1}{n}\sum_{i=1}^n \frac{\phi(x_i)(Y_i - e^{\phi^T(x_i)\beta})\theta}{\theta + e^{\phi^T(x_i)\beta}} = -\frac{1}{n}\sum_{i=1}^n \frac{\phi(x_i)(Y_i - \mathrm{E}Y_i)\theta}{\theta + \mathrm{E}Y_i}.$$

The Hessian matrix of the log-likelihood function is

$$\ddot{\ell}(\beta) := \frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i)\phi^T(x_i) \frac{\theta(\theta + Y_i)e^{\phi^T(x_i)\beta}}{(\theta + e^{\phi^T(x_i)\beta})^2}. \tag{21}$$

Let $d^* = |\{j : \beta_j^* \neq 0\}|$; the true coefficient vector β^* is defined by

$$\beta^* = \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \mathrm{E}R(Y, X, \beta). \tag{22}$$

For the optimization problem of maximizing the ℓ_1-penalized empirical likelihood, our establishment is that KKT conditions are satisfied for the estimated parameter. But here we use the true parameter version of the KKT conditions

$$\left|\dot{\ell}_j(\beta^*)\right| = \left|\frac{1}{n} \sum_{i=1}^{n} \frac{\phi_j(x_i)(Y_i - \mathrm{E}Y_i)\theta}{\theta + \mathrm{E}Y_i}\right| \leq w_j \tag{23}$$

to approximate estimated parameter version of KKT conditions.

When $\theta \to \infty$, this approach of defining data-driven weights w_j is proposed by Ivanoff et al. [16] for Poisson regression by using concentration inequality for Poisson point process version of Theorem 3.1. Similarly, in NB regression the w_j are determined such that the event (23) holds with high probability for each j by applying Corollary 3.1.

Recall the weighted Poisson decomposition of the NB point processes

$$NB(A) := \sum_{k=1}^{\infty} kN_k(A),$$

where $N_k(A)$ is an inhomogeneous Poisson point process with intensity $\alpha_k \int_A \lambda(x)\,dx$ and $\alpha_k = \frac{q^k}{-k\log(1-q)}$ defined in (16).

Let $\mu(A) := \int_A \lambda(x)\,dx$, the m.g.f. of $NB(A)$ is written as

$$M_{NB(A)}(\theta) = \exp\left\{\sum_{k=1}^{\infty} \alpha_k \mu(A)\left(e^{-k\theta} - 1\right)\right\}. \tag{24}$$

For the subsequent step, let $\tilde{\phi}_j(x_i) := \frac{\phi_j(x_i)\theta}{\theta + \mathrm{E}Y_i} \leq \phi_j(x_i)$ and we want to evaluate the complementary events of the true parameter version of the KKT conditions:

$$P\left(\left|\sum_{i=1}^{n} \frac{\phi_j(x_i)(Y_i - \mathrm{E}Y_i)\theta}{\theta + \mathrm{E}Y_i}\right| \geq w_j\right)$$

$$= P\left(\left|\sum_{i=1}^{n} \tilde{\phi}_j(x_i)(Y_i - \mathrm{E}Y_i)\right| \geq w_j\right) \quad \text{for } j = 1, 2, \ldots, p. \tag{25}$$

The aim is to find the values $\{w_j\}_{j=1}^{p}$ by the concentration inequalities we proposed. Then it is sufficient to estimate the probability in the right-hand side of the above inequality.

Let $\mu(A)$ be a Lebesgue measure for $A \in \mathbb{R}^d$ and $N_{i,k}(A)$ be Poisson random measure with rate $\frac{\alpha_k(i)\mu(A)}{\mu_i}$, where $\mu_i = \sum_{k=1}^{\infty} k\alpha_k(i)$ with $\alpha_k(i) := \frac{q_i^k}{-k\log(1-q_i)}$ and $\{q_i := \frac{\mu_i}{\theta + \mu_i}\}_{i=1}^{n}$. We

define the independent NB point processes $NB_i(A)$ as:

$$NB_i(A) =: \sum_{k=1}^{\infty} k N_{i,k}(A), \quad \text{with } E[NB_i(A)] = \sum_{k=1}^{\infty} k \frac{\alpha_k(i)\mu(A)}{\mu_i} = \mu(A)$$

where $\{N_{i,k}(A)\}$ are independent for each k, i.

For $[0,1]^d := \bigcup_{i=1}^{n} S_i$ being a disjoined union, we assume that the intensity function

$$\lambda(t) = \sum_{i=1}^{n} \frac{h(x_i)}{\mu(S_i)} I_{S_i}(t)$$

is a histogram type estimator (Baraud and Birgé [2]). It is just a piece-wise constant intensity function.

Now we will apply Corollary 3.1, note that

$$\mu(S_i) = \int_{S_i} \lambda(t)dt = h(x_i) := E Y_i$$

via the definition of $\lambda(t)$ that we defined before.

Thus, both $(NB_1(S_1),\ldots,NB_n(S_n))$ and $(Y_1,\ldots,Y_n)$ have the same law by the weighted Poisson decomposition.

Choosing $f(x) = g_j(x) = \sum_{l=1}^{n} \tilde{\phi}_j(x_l) 1_{S_l}(x)$, it derives

$$\sum_{i=1}^{n} \int_{S_i} g_j(x)NB_i(dx) = \sum_{i=1}^{n} \int_{S_i} \sum_{l=1}^{n} \tilde{\phi}_j(x_l) I_{S_l}(x) NB_i(dx) = \sum_{i=1}^{n} \tilde{\phi}_j(x_i) Y_i$$

for each j.

Assume that there exists constants C_1, C_2 such that

$$C_1 = \max_{1 \le i \le n} \mu_i = \max_{1 \le i \le n} \sum_{k=1}^{\infty} k\alpha_k(i) = O(1) \quad \text{and} \quad O(1) = \max_{1 \le i \le n} h(x_i) \le C_2. \tag{26}$$

Then

$$V_{i,g_j} = \int_{S_i} g_j^2(x)\lambda(x)dx = \int_{S_i} \sum_{l=1}^{n} \tilde{\phi}_j^2(x_l) I_{S_l}(x) \sum_{i=1}^{n} \frac{h(x_i)}{\mu(S_i)} I_{S_i}(t)\,dx$$

$$= \tilde{\phi}_j^2(x_i)h(x_i) \le C_2 \max_{1 \le i \le n} \tilde{\phi}_j^2(x_i) \le C_2 \max_{1 \le i \le n} \phi_j^2(x_i).$$

Hence

$$P\left(\frac{1}{n}\left|\sum_{i=1}^{n} \tilde{\phi}_j(\mathbf{x}_i)(Y_i - E Y_i)\right| \ge C_1\left[\sqrt{2yC_2 \max_{1 \le i \le n} \phi_j^2(x_i)} + \frac{y}{3}\|g_j(x)\|_{\infty}\right]\right)$$

$$\le P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\int_{S_i} g(x)\big[NB_i(dx) - \mu_i\lambda(x)\,dx\big]\right\}\right| \ge \frac{1}{n}\sum_{i=1}^{n} \mu_i\left[\sqrt{2yV_{i,g_j}} + \frac{y}{3}\|g_j(x)\|_{\infty}\right]\right).$$

Now we let $w_j = C_1[\sqrt{2yC_2 \max_{1 \le i \le n} \phi_j^2(x_i)} + \frac{y}{3}\|g_j(x)\|_{\infty}]$.

From (25), we apply concentration inequality in Corollary 3.1 to the last probability in the above inequality. This implies

$$
P\left(\frac{1}{n} \left| \sum_{i=1}^{n} \frac{\phi_j(\mathbf{x}_i)(Y_i - EY_i)\theta}{\theta + EY_i} \right| \geq w_j \right)
$$
$$
\leq P\left(\left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \int_{S_i} g(x)\big[NB_i(dx) - \mu_i\lambda(x)\,dx\big] \right\} \right| \geq w_j \right) \leq 2e^{-ny}.
$$

Take $y = \frac{\gamma}{n}\log p$ where $\gamma > 0$, we have

$$
P\left(\frac{1}{n} \left| \sum_{i=1}^{n} \frac{\phi_j(x_i)(Y_i - EY_i)\theta}{\theta + EY_i} \right| \geq w_j \right) \leq \frac{2}{p^\gamma}. \tag{27}
$$

The weights are proportional to

$$
w_j \propto \sqrt{\frac{2\gamma\log p}{n}} C_2\left(\max_{i=1,\dots,n} |\phi_j(x_i)| \right) + \frac{\gamma\log p}{3n} \max_{i=1,\dots,n} |\phi_j(x_i)|.
$$

Here the γ can be treated as tuning parameters. In high-dimensional statistics, we often have dimension assumption $\frac{\log p}{n} = o(1)$, $\max_{i=1,\dots,n} |\phi_j(\mathbf{x}_i)| < \infty$, the term $\frac{\gamma\log p}{3n} \times \max_{i=1,\dots,n} |\phi_j(\mathbf{x}_i)|$ is negligible in the weights, comparing to the term $\sqrt{\frac{2\gamma\log p}{n}}$.

Then, for large dimension p, the above upper bound of the probability inequality tends to zero. Therefore, the event of the KKT condition holds with high probability.

Based on (27) and some regular assumptions such as compatibility factor (a generalized restricted eigenvalue condition), we will derive the following ℓ_1-estimation error with high probability:

$$
\|\hat{\beta} - \beta^*\|_1 \leq O_p\left(d^* \sqrt{\frac{\log p}{n}} \right)
$$

in the next section.

Under the restricted eigenvalue condition proposed by Bickel et al. [3], oracle inequalities for a weighted Lasso regularized Poisson binomial regression are studied in Ivanoff et al. [16]. Since the negative binomial is neither sub-gaussian nor a member of the exponential family, our extension is not straightforward. We employ the assumption of compatibility factor which renders the proof simpler than that of restricted eigenvalue condition.

4.2 Oracle inequalities for the weighted Lasso in count data regressions

The ensuing subsections are described in two steps: The first step is to construct two lemmas to get a lower or upper bound for symmetric Bregman divergence in the weighted Lasso case. Adopting a symmetric Bregman divergence, the second step is to derive oracle inequalities by some inequality scaling tricks.

4.2.1 Case of NB regression

The symmetric Bregman divergence (sB divergence) is defined by

$$
D^s(\hat{\beta}, \beta) = (\hat{\beta} - \beta)^T\big(\dot{\ell}(\hat{\beta}) - \dot{\ell}(\beta)\big);
$$

see Nielsen and Nock [26].

The following lemma is a crucial inequality for deriving the oracle inequality under the assumption of a compatibility factor which is also used in Yu [37] for the Cox model. Let β^* be the true estimated coefficients vector defined by (22), and $H = \{j : \beta_j^* \neq 0\}$, $H^c = \{j : \beta_j^* = 0\}$. Then we have the following.

Lemma 4.1 *Let $\hat{\beta}$ be the estimate in* (19), *set $l_m := \|\dot{l}(\beta^*)\|_\infty$, and $\tilde{\beta} = \hat{\beta} - \beta^*$. Then, we have*

$$(w_{\min} - l_m)\|\tilde{\beta}_{H^c}\|_1 \leq D^s(\hat{\beta}, \beta^*) + (w_{\min} - l_m)\|\tilde{\beta}_{H^c}\|_1 \leq (w_{\max} + l_m)\|\tilde{\beta}_H\|_1,$$

where $w_{\max} = \max_{1 \leq j \leq p} w_j$, $w_{\min} = \min_{1 \leq j \leq p} w_j$.

Moreover, in the event $\{l_m \leq \frac{\xi-1}{\zeta+1} w_{\min}\}$ for any $\xi > 1$ and $\|\Phi(X)\|_\infty \leq K$, with probability $1 - p^{1-O(1)r^2}$, we have

$$\|\tilde{\beta}_{H^c}\|_1 \leq \frac{\xi w_{\max}}{w_{\min}} \|\tilde{\beta}_H\|_1. \tag{28}$$

Proof The left inequality is obtained by $D^s(\hat{\beta}, \beta^*) = (\hat{\beta} - \beta^*)^T[\dot{\ell}(\hat{\beta}) - \dot{l}(\beta^*)] \geq 0$ due to convexity of the function $\ell(\beta)$. Note that

$$\begin{aligned}
\tilde{\beta}^T\{\dot{\ell}(\beta^* + \tilde{\beta}) - \dot{\ell}(\beta^*)\} &= \sum_{j \in H^c} \tilde{\beta}_j \dot{\ell}_j(\beta^* + \tilde{\beta}) + \sum_{j \in H} \tilde{\beta}_j \dot{\ell}_j(\beta^* + \tilde{\beta}) + \tilde{\beta}^T(-\dot{\ell}(\beta^*)) \\
&\leq \sum_{j \in H^c} -w_j \hat{\beta}_j \operatorname{sign}(\hat{\beta}_j) + \sum_{j \in H} w_j|\tilde{\beta}_j| + \|\tilde{\beta}\|_1 l_m \\
&= \sum_{j \in H^c} -w_j|\hat{\beta}_j| + \sum_{j \in H} w_j|\tilde{\beta}_j| + l_m\|\tilde{\beta}_H\|_1 + l_m\|\tilde{\beta}_{H^c}\|_1 \\
&\leq \sum_{j \in H^c} -w_{\min}|\hat{\beta}_j| + \sum_{j \in H} w_{\max}|\tilde{\beta}_j| + l_m\|\tilde{\beta}_H\|_1 + l_m\|\tilde{\beta}_{H^c}\|_1 \\
&= (l_m - w_{\min})\|\tilde{\beta}_{H^c}\|_1 + (w_{\max} + l_m)\|\tilde{\beta}_H\|_1.
\end{aligned}$$

where the second last inequality follows from KKT conditions and dual norm inequality.

Thus we obtain (28). In fact, inequality (28) turns to

$$\frac{2w_{\min}}{\xi + 1}\|\tilde{\beta}_{H^c}\|_1 \leq D^s(\hat{\beta}, \beta^*) + \frac{2w_{\min}}{\xi + 1}\|\tilde{\beta}_{H^c}\|_1 \leq \frac{2\xi w_{\max}}{\xi + 1}\|\tilde{\beta}_H\|_1 \tag{29}$$

due to $w_{\min} - l_m \geq \frac{2w_{\min}}{\xi+1}$, $w_{\max} + l_m \leq \frac{2\xi}{\xi+1} w_{\max}$.

For the event $\{\|\dot{\ell}(\beta^*)\|_\infty \leq \frac{\xi-1}{\xi+1} w_{\min}\}$, applying the NB concentration inequality (9), we have

$$\begin{aligned}
&P\left(\|\dot{\ell}(\beta^*)\|_\infty \geq \frac{\xi - 1}{\xi + 1} w_{\min}\right) \\
&\leq pP\left(\left|\sum_{i=1}^n \frac{\phi_1(x_i)\theta}{(\theta + EY_i)}(Y_i - EY_i)\right| \geq \frac{\xi - 1}{\xi + 1} n w_{\min}\right) \\
&\leq p\exp\left\{-n\left(\sqrt{\frac{\xi - 1}{\xi + 1} \cdot \frac{n w_{\min}}{c_{2n}} + \frac{c_{1n}^2}{4c_{2n}^2}} - \frac{c_{1n}}{2c_{2n}}\right)^2\right\}
\end{aligned}$$

$$
= p \exp\left\{ -n\left(\frac{\xi - 1}{\xi + 1} \cdot \frac{nw_{\min}}{c_{2n}} \middle/ \sqrt{\frac{\xi - 1}{\xi + 1} \cdot \frac{nw_{\min}}{c_{2n}} + \frac{c_{1n}^2}{4c_{2n}^2}} + \frac{c_{1n}}{2c_{2n}} \right)^2 \right\}
$$

$$
\sim p \exp\left\{ -nw_{\min}^2 O(1) \right\} \sim p^{1 - O(1)r^2}.
$$

where the second last $\sim$ is due to $c_{1n} = O(n) = c_{2n}$ by the assumption (26) and the the last $\sim$ is by letting $w_{\min} = O(r\sqrt{\frac{\log p}{n}})$.

Here $r > 0$ is a suitable tuning parameter, such that the event $\{\|\dot{\ell}(\beta^*)\|_\infty \le \frac{\xi-1}{\xi+1} w_{\min}\}$ holds with probability tending to 1 as $p, n \to \infty$. □

In order to derive Theorem 4.1, we adopt a crucial inequality by using the similar approach in Lemma 5 of Zhang and Jia [38], which is the technique of Taylor's expansion for convex log-likelihood functions.

Lemma 4.2 *Let the Hessian matrix $\ddot{\ell}(\beta)$ be given in* (21), *and, for $\delta \in \mathbb{R}^p$, assume the identifiability condition: $\phi^T(x_i)(\beta + \delta) = \phi^T(x_i)\beta$ implies $\phi^T(x_i)\delta = 0$. Then we have*

$$
D^s(\beta + \delta, \beta) \ge \delta^T \ddot{\ell}(\beta)\delta e^{-2\|\Phi(X)\|_\infty \|\delta\|},
$$

where $\|\Phi(X)\|_\infty = \max\{|\phi_j(x_{ij})|; 1 \le i \le n, 1 \le j \le p\}$.

By the result in the second parts of Theorem 4.1, in the event $\{\|\dot{\ell}(\beta^*)\|_\infty \le \frac{\xi-1}{\xi+1}\lambda_{\min}\}$, the error of estimate $\tilde{\beta} = \hat{\beta} - \beta^*$ belongs to the following cone set:

$$
S(\eta, H) = \left\{ b \in \mathbb{R}^p : \|b_{H^c}\|_1 \le \eta\|b_H\|_1 \right\} \tag{30}
$$

with $\eta = \frac{\xi w_{\max}}{w_{\min}}$.

Let $d^*(\beta^*) =: |H(\beta^*)|$ with $H(\beta^*) =: \{j : \beta_j^* \ne 0, \beta^* = (\beta_1^*, \dots, \beta_j^*, \dots, \beta_p^*) \in \mathbb{R}^p\}$. Sometimes we write $H(\beta^*)$ as H and $d^*(\beta^*)$ as d^* if there is no ambiguity. Define the compatibility factor (see van de Geer [33]) of a $p \times p$ nonnegative-definite matrix Σ as

$$
C(\eta, H, \Sigma) = \inf_{0 \ne b \in S(\eta, H)} \frac{(d^o(b))^{1/2}(b^T \Sigma b)^{1/2}}{\|b_H\|_1} > 0 \quad (\eta \in \mathbb{R}),
$$

where $S(\eta, H) = \{b \in \mathbb{R}^p : \|b_{H^c}\|_1 \le \eta\|b_H\|_1\}$ is the cone condition.

The compatibility factor is strongly linked to the restricted eigenvalue proposed by Bickel et al. [3],

$$
RE(\eta, H, \Sigma) = \inf_{0 \ne b \in S(\eta, H)} \frac{(b^T \Sigma b)^{1/2}}{\|b\|_2} > 0 \quad (\eta \in R). \tag{31}
$$

Actually, by the Cauchy–Schwarz inequality, it is evident that $[C(\eta, H, \Sigma)]^2 \ge [RE(\eta, H, \Sigma)]^2$. In the proof of Theorem 4.1, the upper bounds of the oracle inequalities are sharper if we use a compatibility factor rather than a restricted eigenvalue.

Next, let $C(\eta, H) = C(\xi, H, \ddot{\ell}(\beta^*))$, then we have the following theorem which is analogous to Zhang and Jia [38] who give oracle inequalities for the Elastic-net estimate in NB regression.

Theorem 4.1 *Suppose* $\|\Phi(X)\|_\infty \leq K$ *with some* $K > 0$ *and let* $\tau = \frac{K(\xi+1)d^* w_{\max}^2}{2w_{\min}[C(\xi w_{\max}/w_{\min}, H)]^2} \leq \frac{1}{2}e^{-1}$ *with assumption* (26). *Setting* $w_{\min} = O(r\sqrt{\frac{\log p}{n}})$ *where* $r > 0$ *is a suitable tuning parameter, then in the event* $\{\|\dot{\ell}(\beta^*)\|_\infty \leq \frac{\xi-1}{\xi+1}w_{\min}\}$ *with probability tending to* 1 *as* $p, n \to \infty$, *we have*

$$\|\hat{\beta} - \beta^*\|_1 \leq \frac{e^{2a_\tau}(\xi+1)d^* w_{\max}^2}{2w_{\min}[C(\xi w_{\max}/w_{\min}, H)]^2} = O_p\left(d^* \frac{w_{\max}^2}{w_{\min}}\right),$$

where $a_\tau \leq \frac{1}{2}$ *is the smaller solution of the equation* $ae^{-2a} = \tau$. *And*

$$D^s(\hat{\beta}, \beta^*) \leq \frac{4e^{2a_\tau}\xi^2 d^* w_{\max}^2}{(\xi+1)^2[C(\xi w_{\max}/w_{\min}, H)]^2}.$$

Proof Let $\tilde{\beta} = \hat{\beta} - \beta^* \neq 0$ and $b = \tilde{\beta}/\|\tilde{\beta}\|_1$, and observe that $\ell(\beta^* + bx)$ is a convex function in x since $\ell(\beta)$ is convex. From the proof of (29) in Lemma 4.1, we have

$$b^T[\dot{l}(\beta^* + bx) - \dot{\ell}(\beta^*)] + \frac{2w_{\min}}{\xi+1}\|b_{H^C}\|_1 \leq \frac{2\xi w_{\max}}{\xi+1}\|b_H\|_1 \tag{32}$$

holding for $x \in [0, \|\tilde{\beta}\|_1]$.

Since $\ell(\beta)$ is a convex in β, $b^T[\dot{\ell}(\beta + bx) - \dot{\ell}(\beta)]$ is increasing in x) and the function $b \in S(\xi w_{\max}/w_{\min}, H)$. By the assumption, we have $\|\Phi(X)\|_\infty\|xb\|_1 \leq Kx$ by noticing that $\|b\|_1 = 1$. And assuming $\|\Phi(x)\|_\infty \leq K$, we obtain the following by Lemma 4.2:

$$xb^T[\dot{\ell}(\beta^* + bx) - \dot{\ell}(\beta^*)] \geq x^2 e^{-2\|\Phi(X)\|_\infty\|xb\|}b^T\ddot{\ell}(\beta)b \geq x^2 e^{-2Kx}b^T\ddot{\ell}(\beta)b,$$

this implies

$$b^T[\dot{\ell}(\beta^* + bx) - \dot{\ell}(\beta^*)] \geq xe^{-2Kx}b^T\ddot{\ell}(\beta)b. \tag{33}$$

For the Hessian matrix evaluated at the true coefficient β^*, the compatibility factor is written as $C(\eta, H) =: C(\eta, H, \ddot{\ell}(\beta^*))$. From the definition of the compatibility factor and the inequality (33), we have

$$Kxe^{-2Kx}[C(\xi w_{\max}/w_{\min}, H)]^2\|b_H\|^2/d^* \leq Kxe^{-2Kx}b^T\ddot{\ell}(\beta)b$$

$$\text{(by (33))} \leq Kb^T[\dot{\ell}(\beta^* + bx) - \dot{\ell}(\beta^*)]$$

$$\text{(by (29))} \leq K\left(\frac{2\xi w_{\max}}{\xi+1}\|b_H\|_1 - \frac{2w_{\min}}{\xi+1}\|b_{H^c}\|_1\right)$$

$$= K\left[\frac{2\xi w_{\max}}{\xi+1}\|b_H\|_1 - \frac{2w_{\min}}{\xi+1}(1 - \|b_H\|_1)\right]$$

$$\leq K\left[\frac{2}{\xi+1}(\xi w_{\max} + w_{\max})\|b_H\|_1 - \frac{2w_{\min}}{\xi+1}\right]$$

$$= K\left[2w_{\max}\|b_H\|_1 - \frac{2\xi w_{\min}}{\xi+1}\right]$$

$$\leq \frac{K(\xi+1)\|b_H\|_1^2 w_{\max}^2}{2w_{\min}},$$

where the last inequality is due to $\frac{2w_{\min}}{\xi+1} + \frac{(\xi+1)\|b_H\|_1 w_{\max}^2}{2w_{\min}} \geq 2w_{\max}\|b_H\|_1$.

Then we have

$$Kxe^{-2Kx} \le \frac{K(\xi + 1)d^* w_{\max}^2}{2w_{\min}[C(\xi w_{\max}/w_{\min}, H)]^2} =: \tau \tag{34}$$

for any $x \in [0, \|\tilde{\beta}\|_1]$.

So with constrain $x \in [0, \|\tilde{\beta}\|_1]$, let a_τ, b_τ, $(a_\tau < 0.5 < b_\tau)$ be the two solutions of the equation $\{z : ze^{-2z} = \tau\}$. Let $S = [0, \|\tilde{\beta}\|_1] \cap \{(-\infty, a_\tau] \cup [b_\tau, +\infty)\}$ satisfy (34).

Note that $b^T[\dot{\ell}(\beta + bx) - \ell(\beta)]$ is an increasing function of x, then S cannot be a disjoint union of the two intervals by the restriction condition in (32), thus S is a closed interval $x \in [0, \tilde{x}]$. The fact that all $x \in [0, \|\tilde{\beta}\|_1]$ implies $x \in [0, \tilde{x}]$, therefore we have

$$Kx \le K\tilde{x} \le a_\tau.$$

Using (34) again, the end point $\tilde{x}$ satisfies $K\tilde{x}e^{-2K\tilde{x}} \le \tau$. Then, for $\forall x \in [0, \tilde{x}]$, we have

$$\|\tilde{\beta}\|_1 = \|bx\|_1 = \|b\|_1 x = x \le \tilde{x} \le \frac{a_\tau}{K} = \frac{e^{2a_\tau}\tau}{K} = \frac{e^{2a_\tau}(\xi + 1)d^* w_{\max}^2}{2w_{\min}[C(\xi w_{\max}/w_{\min}, H)]^2}.$$

It remains to show the oracle inequality of $D^s(\hat{\beta}, \beta^*)$. Let $\delta = bx$ in Lemma 4.2. Using the definition of compatibility factor, and observing that $a_\tau \ge Kx$ and $Kx \ge \|\Phi(X)\|_\infty \|\delta\|_1$, one derives

$$e^{-2a_\tau}\left[C(\xi w_{\max}/w_{\min}, H)\right]^2 \|\tilde{\beta}_H\|_1^2/d^* \le e^{-2a_\tau}\tilde{\beta}^T\ddot{\ell}(\beta^*)\tilde{\beta} \le e^{-2Kx}\tilde{\beta}^T\ddot{i}(\beta^*)\tilde{\beta}$$
$$\le e^{-2\|\Phi(X)\|_\infty\|\delta\|_1}\tilde{\beta}^T\ddot{\ell}(\beta^*)\tilde{\beta}.$$

Applying Lemma 4.2, one immediately deduces that

$$e^{-2a_\tau}\left[C(\xi w_{\max}/w_{\min}, H)\right]^2 \|\tilde{\beta}_H\|_1^2/d^* \le e^{-2\|\Phi(X)\|_\infty\|\delta\|_1}\tilde{\beta}^T\ddot{\ell}(\beta^*)\tilde{\beta} \le D^s(\hat{\beta}, \beta^*)$$
$$\le \frac{2\xi w_{\max}}{\xi + 1}\|\tilde{\beta}_H\|_1,$$

where the last inequality is from (32).

Canceling the term $\|\tilde{\beta}_H\|_1$ on the both sides of the inequality above, we get

$$\|\tilde{\beta}_H\|_1 \le \frac{2e^{2a_\tau}\xi d^* w_{\max}}{(\xi + 1)[C(\xi w_{\max}/w_{\min}, H)]^2}.$$

Consequently,

$$D^s(\hat{\beta}, \beta^*) \le \frac{2\xi w_{\max}}{\xi + 1}\|\tilde{\beta}_H\|_1 \le \frac{4e^{2a_\tau}\xi^2 d^* w_{\max}^2}{(\xi + 1)^2[C(\xi w_{\max}/w_{\min}, H)]^2}. \qquad \square$$

Theorem 4.1 indicates that, if we presume $d^* = O(1)$, the l_1-estimation error bound is of the order $\sqrt{\frac{\log p}{n}}$. And this convergence rate is rate-optimal in the minimax sense; see Raskutti et al. [28] and Li et al. [22]. Under some regularity conditions, the weighted Lasso estimates are expected to have the consistency property in the sense of the ℓ_1-norm when the dimension increases as order $\exp(o(n))$. Unlike the MLE estimator with convergence

rate $\frac{1}{\sqrt{n}}$, in the high-dimensional case we have to multiply $\sqrt{\log p}$ to the rate of MLE as the price to pay.

4.2.2 Case of Poisson regression

In this section, we use the same procedure as in Sect. 4.2.1 to prove the case of Poisson regression with weighted Lasso penalty. The derivation of this type of oracle inequalities is simpler than that in Ivanoff et al. [16]. Similar to Lemma 5 of Zhang and Jia [38], the proof is based on the following key lemma.

Lemma 4.3 *The Hessian matrix and gradient of the log-likelihood function in the Poisson regression model are $\dot{\ell}(\beta) = -\frac{1}{n}\sum_{i=1}^{n}\phi(x_i)[Y_i - e^{\phi^T(x_i)\beta}]$, $\ddot{\ell}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\phi^T(x_i)e^{\phi^T(x_i)\beta}$. Assume the identifiability condition: $\phi^T(x_i)(\beta + \delta) = \phi^T(x_i)\beta$ implies $\phi^T(x_i)\delta = 0$, then we have*

$$D^s(\beta + \delta, \beta) \geq \delta^T \ddot{\ell}(\beta)\delta e^{-\|\Phi(X)\|_\infty \|\delta\|_1},$$

where $\|\Phi(X)\|_\infty = \max\{|\phi_j(x_{ij})|; 1 \leq i \leq n, 1 \leq j \leq p\}$.

Proof Without loss of generality, we assume that $\phi^T(x_i)\delta \neq 0$. By the expression of $\dot{\ell}(\beta)$ and $\ddot{\ell}(\beta)$, one deduces

$$\delta^T\big[\dot{\ell}(\beta + \delta) - \dot{\ell}(\beta)\big] = \delta^T \cdot \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\big(e^{\phi^T(x_i)(\beta+\delta)} - e^{\phi^T(x_i)\beta}\big)$$

$$= \delta^T \cdot \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)e^{\phi^T(x_i)\beta} \cdot \frac{e^{\phi^T(x_i)\delta} - 1}{\phi^T(x_i)\delta - 0} \cdot \phi^T(x_i)\delta$$

$$\geq \delta^T \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\phi^T(x_i)e^{\phi^T(x_i)\beta}\delta \cdot e^{-\|\Phi(X)\|_\infty \|\delta\|_1},$$

where the last inequality is from $\frac{e^x - e^y}{x-y} \geq e^{-|x|\vee|y|}$. □

Theorem 4.2 *For a weighted Lasso estimator $\hat{\beta}$ in a Poisson regression, suppose $\|\Phi(X)\|_\infty \leq K$ with assumption (26) and let $\tau = \frac{K(\xi+1)d^*w_{\max}^2}{2w_{\min}[C(\xi w_{\max}/w_{\min}, H)]^2} \leq e^{-1}$ with a certain $K > 0$. Then, in the event $\|\dot{\ell}(\beta^*)\|_\infty \leq \frac{\xi-1}{\xi+1}w_{\min}$, we have*

$$\|\hat{\beta} - \beta^*\|_1 \leq \frac{e^{a_\tau}(\xi+1)d^*w_{\max}^2}{2w_{\min}[C(\xi w_{\max}/w_{\min}, H)]^2},$$

where $a_\tau \leq 1$ is the smaller solution of the equation $ae^{-a} = \tau$. Also,

$$D^s(\hat{\beta}, \beta^*) \leq \frac{4e^{a_\tau}\xi^2 d^*w_{\max}^2}{(\xi+1)^2[C(\xi w_{\max}/w_{\min}, H)]^2}.$$

Proof Note that a_τ is the small solution of $ze^{-z} = \tau$ in Poisson regression. By Lemma 4.1 and Lemma 4.3, the proof is almost exactly the same as the proof of Theorem 4.1. □

5 Simulations

This section aims to compare the performances of the un-weighted Lasso and weighted Lasso for NB regression on simulated data sets. We use the R package `mpath` with the function `glmregNB` to fit the un-weighted Lasso estimator of NB regression. The R package `lbfgs` is employed to do a weighted ℓ_1-penalized optimization for NB regression. For weighted Lasso, we first run an un-weighted Lasso estimator and find the optimal tuning parameter λ_{op} by cross-validation. The actual weights we use are the standardized weights given by

$$\tilde{w}_j := p \frac{w_j}{\sum_{j=1}^{p} w_j}.$$

And then we solve the optimization problem

$$\hat{\beta}^{(t)} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \left\{ -\ell(\beta) + \lambda_{\text{op}} \sum_{j=1}^{p} \tilde{w}_j |\beta_j| \right\}. \tag{35}$$

Here we first apply the function `cv.glmregNB()` to do a 10-fold cross-validation to obtain the optimal penalty parameter λ. For comparison, we use the function `glmregNB` to do un-weighted Lasso for NB regression and use the estimated θ as the initial value for our weighted Lasso algorithms.

For simulations, we simulate 100 random data sets. By optimization with suitable λ, we obtain the model with the parameter $\hat{\beta}_{\lambda_{\text{opt}}}$. Then we estimate the performance of the ℓ_1 estimation error $\|\beta^* - \hat{\beta}_{\lambda_{\text{opt}}}\|_1$ and the prediction error $\|X_{\text{test}}\beta^* - X_{\text{test}}\hat{\beta}_{\lambda_{\text{opt}}}\|_2$ on the test data (X_{test} of size n_{test}) by the average of the 100-times errors.

For each simulation $n_{\text{train}} = 100$, $n_{\text{test}} = 200$. We mainly adopt the simulation setting: The predictor variables X are randomly drawn from $N_p(0, \Sigma)$, where Σ has elements $\rho_{|j-l|}$ ($j, l = 1, 2, \ldots, p$). The correlation among predictor variables is controlled by ρ with $\rho = 0.5$ and 0.8, respectively. We assign the true vector as

$$\beta^* = (1.25, -0.95, 0.9, -1.1, 0.6, \underbrace{0, \ldots, 0}_{p-5}).$$

The simulation results are displayed in Table 1. The table shows that the proposed weighted Lasso estimators are more accurate than the un-weighted Lasso estimators. With the help of the optimal weights, it reflects that the controlling of KKT conditions by a data-dependent tuning parameter is able to improve the accuracy of the estimation in aspects of the ℓ_1-estimation, square prediction errors. It can be also seen that the increasing p will hinder the ℓ_1-estimation error by the curse of dimensionality.

6 Summary and discussion

The paper contributes to three parts. In the first part, we provide a characterization of discrete compound Poisson point processes (DCPP) in the same fashion as it has been done in 1936 by Arthur H. Copeland and Francis Regan for the Poisson process. The second part is focused on deriving concentration inequalities for DCPP, which are applied in the third part to derive optimally oracle inequalities of the weighted Lasso estimation in high-dimensional NB regression.

Table 1 Means of ℓ_1 error and prediction error

Method	$\|\hat{\beta} - \beta^*\|_1$	Prediction error
$p = 20, \rho = 0.5$		
Lasso	1.902	7.161
Weight Lasso	1.624	6.147
$p = 20, \rho = 0.8$		
Lasso	3.001	7.155
Weight Lasso	2.917	6.979
$p = 100, \rho = 0.5$		
Lasso	4.137	15.452
Weight Lasso	3.905	13.866
$p = 100, \rho = 0.8$		
Lasso	4.543	10.866
Weight Lasso	4.552	10.695
$p = 200, \rho = 0.5$		
Lasso	4.467	17.533
Weight Lasso	4.373	15.854
$p = 200, \rho = 0.8$		
Lasso	4.467	17.533
Weight Lasso	4.373	15.854

Oracle inequalities for discrete distributions are statistically useful in both non-asymptotic and asymptotical analyses of count data regression. The motivation for deriving new concentration inequalities for DCP processes is that we want to show that the KKT conditions of the NB regressions with Lasso penalty (wuth a tuning parameter λ) is a high-probability event. The KKT conditions are based on the concentration equality for a centralized weighted sum of NB random variables. However, existing concentration results are not friendly applicable to the aim of optimal inference procedure. The optimal inference procedure here is to choose an optimal tuning parameter for Lasso estimates in high-dimensional NB regressions, which lead to the minimax optimal convergence for the estimator by considering the oracle inequalities for the ℓ_1-estimation error.

In the future, it is of interest to study the concentration inequalities for other extended Poisson distributions in regression analysis such as the Conway–Maxwell–Poisson distribution; see Li et al. [22] for the existing probability properties. In another direction, our proposed concentration inequalities would be desirable to be extended to establish oracle inequalities for penalized projection estimators studied by Reynaud-Boure [29], by considering the intensity of some inhomogeneous compound Poisson processes.

Appendix: Proof of Theorem 2.1

A.1 Proof of Theorem 2.1

The proof is divided into two parts. In the first part, we show that $\nu(A)$ is countably additive and absolutely continuous. Moreover, in the second part, we will show Eq. (3).

Proof Step 1.

(1) Note that $P_k(A) > 0$ and $\sum_{k=0}^{\infty} P_k(A) = 1$ by assumption 1, we have $0 < P_k(A) < 1$ whenever $0 < m(A) < \infty$. And from assumption 2, we get $P_0(A) = 1$ if $m(A) = 0$. It follows that $0 < P_0(A) \le 1$ if $0 < m(A) < \infty$. By assumptions 2 and 4, one has $\lim_{m(A) \to 0} \nu(A) = \lim_{m(A) \to 0} -\log[1 - S_1(A)] = 0$. Therefore, $\nu(\cdot)$ is absolutely continuous.

(2) Let $\nu(A) := -\log P_0(A)$. For $A_1 \cap A_2 = \emptyset$, it implies by assumption 3 that

$$\nu(A_1 \cup A_2) = -\log P_0(A_1 \cup A_2) = -\log\left[P_0(A_1) \cdot P_0(A_2)\right] = \nu(A_1) + \nu(A_2),$$

which shows that $\nu(\cdot)$ is finitely additive.

Based on the two arguments in *Step* 1, we have shown that $\nu(\cdot)$ is countable additive.

Step 2. For convenience, we first consider the case that $A = I_{\mathbf{a},\mathbf{b}} = (\mathbf{a}, \mathbf{b}]$ is a d-dimensional interval on $\mathbb{R}^d$, and then extend it to any measurable sets. Let $T(x) =: \nu(I_{0,x})$, so $T(x)$ is a d dimensional coordinately monotone transformation with

$$T(\mathbf{0}) = \mathbf{0}, \qquad T(x) \le T\left(x'\right) \quad \text{iff} \quad x \le x',$$

where the componentwise order symbol (partially ordered symbol) "$\le$" means that each coordinate in x is less than or equal to the corresponding coordinate of x' (for example, with $d = 2$, we have $(1, 2) \le (3, 4)$). This is by the fact that $\nu(\cdot)$ is absolutely continuous and finite additive, which implies $\nu(I_{\mathbf{0},\mathbf{0}}) = \lim_{m(I_{\mathbf{0},x}) \to 0} \nu(I_{\mathbf{0},x}) = 0$.

The definition of the minimum set of the componentwise ordering relation lays the base for our study; see Sect. 4.1 of Dinh The Luc [24]. Let A be a nonempty set in $\mathbb{R}^{+d}$. A point $\mathbf{a} \in A$ is called a *Pareto minimal point* of the set A if there is no point $a' \in A$ such that $\mathbf{a}' \ge \mathbf{a}$ and $\mathbf{a}' \ne \mathbf{a}$. The sets of Pareto minimal points of A is denoted by $\mathrm{PMin}(A)$. Next, we define the generalized inverse $T^{-1}(t)$ (like some generalizations of the inverse matrix, it is not unique, we just choose one) by

$$x = (x_1, \ldots, x_d) = T^{-1}(t) =: \mathrm{PMin}\left\{\mathbf{a} \in \mathbb{R}^{+d} : t \le T(\mathbf{a})\right\}.$$

When $d = 1$, the generalized inverse $T^{-1}(t)$ is just a similar version of the quantile function.

By Theorem 4.1.4 in Dinh The Luc [24], if A is a nonempty compact set, then it has a Pareto minimal point. Set $x = T^{-1}(t)$, $x + \boldsymbol{\xi} = T^{-1}(t + \tau)$ and define $\varphi_k(\tau, t) = P_k(I_{x, x+\xi})$ for $t = T(x)$, $\tau = T(x + \boldsymbol{\xi}) - T(x)$.

Consequently, by part one, finite additivity of $\nu(\cdot)$, it implies that

$$\tau = T(x + \boldsymbol{\xi}) - T(x) = \nu(I_{0, x+\xi}) - \nu(I_{0, x}) = \nu(I_{x, x+\xi}) = -\log P_0(I_{x, x+\xi}).$$

Thus, we have $\varphi_0(\tau, t) = e^{-\tau}$. Let $\Phi_k(\tau, t) = \sum_{i=k}^{\infty} \varphi_i(\tau, t)$. Due to assumption 5, it follows that

$$\alpha_k = \lim_{\tau \to 0} \frac{\varphi_k(\tau, t)}{\Phi_1(\tau, t)} = \lim_{\tau \to 0} \frac{\varphi_k(\tau, t)}{\tau} \quad \text{for } k \ge 1. \tag{36}$$

Applying assumption 3, we get $\varphi_k(\tau + h, t) = \sum_{i=0}^{k} \varphi_{k-i}(\tau, t)\varphi_i(h, t + \tau)$. Subtracting $\varphi_k(\tau, t)$ and dividing by h ($h > 0$) in the above equation, then

$$\frac{\varphi_k(\tau + h, t) - \varphi_k(\tau, t)}{h} = \varphi_k(\tau, t) \cdot \frac{\varphi_0(h, t + \tau) - 1}{h} + \sum_{i=1}^{k} \varphi_{k-i}(\tau, t) \cdot \frac{\varphi_i(h, t + \tau)}{h}. \tag{37}$$

Here the meaning of h in $\varphi_i(h, t + \tau)$ is that

$$h = T(x + \boldsymbol{\xi} + \boldsymbol{\eta}) - T(x + \boldsymbol{\xi}) = \nu(I_{0, x+\xi+\eta}) - \nu(I_{0, x+\xi}) = \nu(I_{x+\xi, x+\xi+\eta}),$$

where $x + \boldsymbol{\xi} + \boldsymbol{\eta} = T^{-1}(t + \tau + h)$ and $x + \boldsymbol{\xi} = T^{-1}(t + \tau)$.

Let $h \to 0$ in (37). By using assumption 5 and its conclusion (36), it implies that

$$\varphi'_k(\tau, t) =: \frac{\partial}{\partial \tau}\varphi_k(\tau, t) = -\varphi_k(\tau, t) + \alpha_1\varphi_{k-1}(\tau, t) + \alpha_2\varphi_{k-2}(\tau, t) + \cdots + \alpha_k\varphi_0(\tau, t), \quad (38)$$

which is a difference–differential equation.

To solve (38), we need to write it in matrix form,

$$\frac{\partial \mathbf{P}_k(\tau, t)}{\partial \tau} := \begin{pmatrix} \varphi'_k(\tau, t) \\ \vdots \\ \varphi'_1(\tau, t) \\ \varphi'_0(\tau, t) \end{pmatrix} = \begin{pmatrix} -1 & \alpha_1 & \cdots & \alpha_k \\ & \ddots & \ddots & \vdots \\ & & & \alpha_1 \\ & & \ddots & -1 \end{pmatrix} \begin{pmatrix} \varphi_k(\tau, t) \\ \vdots \\ \varphi_1(\tau, t) \\ \varphi_0(\tau, t) \end{pmatrix} =: \mathbf{Q}\mathbf{P}_k(\tau, t).$$

The general solution is

$$\mathbf{P}_k(\tau, t) = e^{\mathbf{Q}\tau} \cdot \mathbf{c} =: \sum_{m=0}^{\infty} \frac{\mathbf{Q}^m}{m!}\tau^m \cdot \mathbf{c},$$

where $\mathbf{c}$ is a constant vector. To specify $\mathbf{c}$, we observe that

$$(0, \dots, 0, 1)^{\mathrm{T}} = \lim_{\tau \to 0} \mathbf{P}_k(\tau, t) = \lim_{\tau \to 0} e^{\mathbf{Q}\tau}\mathbf{c} = \mathbf{c}.$$

Hence, the $\mathbf{Q}$ can be written as

$$\mathbf{Q} = -\mathbf{I}_{k+1} + \alpha_1\mathbf{N} + \alpha_2\mathbf{N}^2 + \cdots + \alpha_k\mathbf{N}^k,$$

where $\mathbf{N} =: \begin{pmatrix} \mathbf{0}_{k\times 1} & \mathbf{I}_k \\ 0 & \mathbf{0}_{1\times k} \end{pmatrix}$. As $i \geq k+1$, we have $\mathbf{N}^i = \mathbf{0}$.

We use the expansion of the power of the multinomial

$$\frac{1}{m!}\left(-\mathbf{I}_{k+1} + \sum_{i=1}^{k} \alpha_i\mathbf{N}^i\right)^m = \sum_{s_0+s_1+\cdots+s_k=m} \frac{(-\mathbf{I}_{k+1})^{s_0}\alpha_1^{s_1}\cdots\alpha_k^{s_k}}{s_0!s_1!\cdots s_k!}\mathbf{N}^{s_1+2s_2+\cdots+ks_k}.$$

Then

$$\mathbf{P}_k(\tau, t) = \sum_{m=0}^{\infty} \sum_{s_0+s_1+\cdots+s_k=m} \frac{(-\mathbf{I}_{k+1})^{s_0}\alpha_1^{s_1}\cdots\alpha_k^{s_k}}{s_0!s_1!\cdots s_k!}\mathbf{N}^{s_1+2s_2+\cdots+ks_k}\tau^{s_0+s_1+\cdots+s_k}\mathbf{c}$$

$$= \sum_{m=0}^{\infty} \sum_{s_0=0}^{m} \tau^{s_0}\frac{(-1)^{s_0}}{s_0!} \sum_{s_1+\cdots+s_k=m-s_0} \frac{\alpha_1^{s_1}\cdots\alpha_k^{s_k}}{s_1!\cdots s_k!}\mathbf{N}^{s_1+2s_2+\cdots+ks_k}\tau^{s_1+\cdots+s_k}\mathbf{c}$$

$$= \sum_{s_0=0}^{\infty} \sum_{m-s_0=0}^{\infty} \tau^{s_0}\frac{(-1)^{s_0}}{s_0!} \sum_{s_1+\cdots+s_k=m-s_0} \frac{\alpha_1^{s_1}\cdots\alpha_k^{s_k}}{s_1!\cdots s_k!}\tau^{s_1+\cdots+s_k}\mathbf{N}^{s_1+2s_2+\cdots+ks_k}\mathbf{c}$$

$$(\text{let } r = m - s_0) = \sum_{s_0=0}^{\infty} \tau^{s_0}\frac{(-1)^{s_0}}{s_0!} \sum_{r=0}^{\infty} \sum_{s_1+\cdots+s_k=r} \frac{\alpha_1^{s_1}\cdots\alpha_k^{s_k}}{s_1!\cdots s_k!}\tau^{s_1+\cdots+s_k}\mathbf{N}^{s_1+2s_2+\cdots+ks_k}\mathbf{c}$$

$$= e^{-\tau}\sum_{r=0}^{\infty} \sum_{s_1+\cdots+s_k=r} \frac{\alpha_1^{s_1}\cdots\alpha_k^{s_k}}{s_1!\cdots s_k!}\tau^{s_1+\cdots+s_k}\mathbf{N}^{s_1+2s_2+\cdots+ks_k}\mathbf{c}$$

$$= \sum_{l=0}^{\infty} \sum_{R(s,k)=l} \frac{\alpha_1^{s_1} \cdots \alpha_k^{s_k}}{s_1! \cdots s_k!} \tau^{s_1 + \cdots + s_k} e^{-\tau} \mathbf{N}^l \mathbf{c}$$

$$= \sum_{l=0}^{k} \sum_{R(s,k)=l} \frac{\alpha_1^{s_1} \cdots \alpha_k^{s_k}}{s_1! \cdots s_k!} \tau^{s_1 + \cdots + s_k} e^{-\tau} \mathbf{N}^l \mathbf{c}, \tag{39}$$

where $R(s,k) =: \sum_{t=1}^{k} t s_t$ and the last equality is obtained by the fact that $\mathbf{N}^l \mathbf{c} = (0, \ldots, 0, 1, \underbrace{0, \ldots, 0)^T}_{l}$ and $\mathbf{N}^l = \mathbf{0}$ as $l \geq k+1$.

Choose the first element in (39). We show that $\varphi_k(\tau, t) = \sum_{R(s,k)=k} \frac{\alpha_1^{s_1} \cdots \alpha_k^{s_k}}{s_1! \cdots s_k!} \tau^{s_1 + \cdots + s_k} e^{-\tau}$. By the relation of $\varphi_k(\tau, t)$ and $P_k(A)$, we obtain

$$P_k(I_{0,\xi}) = \sum_{R(s,k)=k} \frac{\alpha_1^{s_1} \cdots \alpha_k^{s_k}}{s_1! \cdots s_k!} \left[\nu(I_{0,\xi}) \right]^{s_1 + \cdots + s_k} \cdot e^{-\nu(I_{0,\xi})}. \tag{40}$$

Step 3. We need to show the countable additivity in terms of (40). Let A_1, A_2 be two disjoint intervals, we first show that

$$P_k(A_1 \cup A_2) = \sum_{R(s,k)=k} \frac{\alpha_1^{s_1} \cdots \alpha_k^{s_k}}{s_1! \cdots s_k!} \left[\nu(A_1 \cup A_2) \right]^{s_1 + \cdots + s_k} e^{-\nu(A_1 \cup A_2)}, \tag{41}$$

which means that (40) is closed under a finite union of the A_i.

By assumption 3, to show (41), it is sufficient to verify $\sum_{i=0}^{k} P_{k-i}(A_1) P_i(A_2)$ equals the right-hand side of (41). From the m.g.f. of the DCP distribution with complex θ, we have

$$M_{A_i}(\theta) =: \mathrm{E} e^{-\theta N(A_i)} = \exp\left\{ \lambda(A_i) \sum_{a=1}^{\infty} \alpha_a \left(e^{-a\theta} - 1 \right) \right\} \quad \text{for } i = 1, 2.$$

And by the definition of m.g.f. and assumption 3,

$$M_{A_1 \cup A_2}(\theta) = \sum_{k=0}^{+\infty} P_k(A_1 \cup A_2) e^{-k\theta} = \sum_{k=0}^{+\infty} \left(\sum_{i=0}^{k} P_{k-i}(A_1) P_i(A_2) \right) e^{-k\theta}$$

$$= \left(\sum_{j=0}^{+\infty} P_j(A_1) e^{-j\theta} \right) \left(\sum_{s=0}^{+\infty} P_s(A_2) e^{-s\theta} \right)$$

$$= M_{A_1}(\theta) M_{A_2}(\theta) = \exp\left\{ \left[\nu(A_1) + \nu(A_1) \right] \sum_{a=1}^{\infty} \alpha_a \left(e^{-a\theta} - 1 \right) \right\}$$

$$= \exp\left\{ \left[\nu(A_1 + A_2) \right] \sum_{a=1}^{\infty} \alpha_a \left(e^{-a\theta} - 1 \right) \right\}$$

$$= \sum_{k=0}^{\infty} \sum_{R(s,k)=k} \frac{\alpha_1^{s_1} \cdots \alpha_k^{s_k}}{s_1! \cdots s_k!} \left[\nu(A_1 \cup A_2) \right]^{s_1 + \cdots + s_k} e^{-\nu(A_1 \cup A_2)} \mathrm{e}^{-k\theta}.$$

Thus, we show (41).

Finally, let $E_n = \bigcup_{i=1}^{n} A_i$, $E = \bigcup_{i=1}^{\infty} A_i$ for disjoint intervals $A_1, A_2, \ldots$. We use the following lemma in Copeland and Regan [8].

Lemma A.1 *If we have assumptions* 1–4, *let* $e_n = \{E \setminus (E \cap E_n)\} \cup \{E_n \setminus (E \cap E_n)\}$. *Then*

$$\lim_{n\to\infty} m(e_n) = 0 \quad implies \quad \lim_{n\to\infty} P_k(E_n) = P_k(E).$$

To see this, since $e_n = \bigcup_{i=n+1}^{\infty} A_i$, we have $m(e_n) = \sum_{i=n+1}^{\infty} m(A_i) \to 0$ as $n \to \infty$. Thus, we obtain the countable additivity with respect to (40). □

A.2 Proof of Corollary 3.1

Let $CP_{i,r_i}(A) =: \sum_{k=1}^{r_i} k N_{i,k}(A)$, where $\{N_{i,k}(A)\}_{i=1}^{n}$ are independent Poisson point process with intensity $\alpha_k(i) \int_A \lambda(x)\,dx$ for each fixed k. By Campbell's theorem, it follows that

$$\mathrm{E}\exp\left\{-\sum_{i=1}^{n}\int_{S_i} f_i(x) CP_{i,r_i}(dx)\right\} = \mathrm{E}\exp\left\{-\sum_{i=1}^{n}\int_{S_i} f_i(x) \sum_{k=1}^{r_i} k N_{i,k}(dx)\right\}$$

$$= \prod_{i=1}^{n}\prod_{k=1}^{r_i} \mathrm{E}\exp\left\{-\int_{S_i} k f_i(x) N_{i,k}(dx)\right\}$$

$$= \prod_{i=1}^{n}\prod_{k=1}^{r_i} \exp\left\{\int_{S_i} [e^{-k f_i(x)} - 1] \alpha_k(i)\lambda(x)\,dx\right\}$$

$$= \exp\left\{\sum_{i=1}^{n}\sum_{k=1}^{r_i}\int_{S_i} [e^{-k f_i(x)} - 1] \alpha_k(i)\lambda(x)\,dx\right\}. \qquad (42)$$

For $\eta > 0$, define the normalized exponential transform of the stochastic integral by $D_{\{r_i\}}(\eta)$:

$$\log D_{\{r_i\}}(\eta) := \eta \sum_{i=1}^{n}\int_{S_i} f_i(x)\left[CP_{i,r_i}(dx) - \sum_{k=1}^{r_i} k \alpha_k(i)\lambda(x)\right]dx$$

$$- \int_{S_i}\sum_{k=1}^{r_i} [e^{k f_i(x)} - k\eta f_i(x) - 1] \alpha_k(i)\lambda(x)\,dx. \qquad (43)$$

Therefore, we obtain $\mathrm{E}D_{\{r\}}(\eta) = 1$ by (42).

It follows from (43) and Markov's inequality that

$$P\left(\eta \sum_{i=1}^{n}\int_{S_i} f_i(x)\left[CP_{i,r_i}(dx) - \sum_{k=1}^{r_i} k \alpha_k(i)\lambda(x)\right]dx \right.$$

$$\left. \geq \sum_{i=1}^{n}\int_{S_i}\sum_{k=1}^{r_i} [e^{k f_i(x)} - k\eta f_i(x) - 1] \alpha_k(i)\lambda(x)\,dx + ny\right)$$

$$= P\left(D_{\{r\}}(\eta) \geq e^{ny}\right) \leq e^{-ny}. \qquad (44)$$

By inequality (12), Eq. (29) implies

$$P\left(\sum_{i=1}^{n}\int_{S_i} f_i(x)\left[CP_{i,r_i}(dx) - \sum_{k=1}^{r_i} k \alpha_k(i)\lambda(x)\right]dx\right.$$

$$\geq \sum_{i=1}^{n} \int_{S_i} \sum_{k=1}^{r_i} \alpha_k(i) \left(\frac{1}{2} \frac{k^2 \eta f_i^2(x) \lambda(x)}{1 - \frac{1}{3} k \eta \|f_i\|_\infty} \, dx + \frac{y}{\eta} \right)$$

$$\leq e^{-ny}. \tag{45}$$

Let $V_{i,f_i} := \int_{S_i} f_i^2(x) \lambda(x) \, dx$ for $i = 1, 2, \ldots, n$. By a basic inequality, we have

$$\sum_{i=1}^{n} \int_{S_i} \sum_{k=1}^{r_i} \alpha_k(i) \left(\frac{1}{2} \frac{k^2 \eta f_i^2(x) \lambda(x)}{1 - \frac{1}{3} k \eta \|f_i\|_\infty} \, dx + \frac{y}{\eta} \right) = \sum_{i=1}^{n} \sum_{k=1}^{r_i} \alpha_k(i) \left(\frac{1}{2} \frac{\eta k^2 V_{i,f_i}}{1 - \frac{1}{3} k \eta \|f_i\|_\infty} + \frac{y}{\eta} \right)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{r_i} \alpha_k(i) \left(\frac{1}{2} \frac{\eta k^2 V_{i,f_i}}{1 - \frac{1}{3} k \eta \|f_i\|_\infty} + \frac{y}{\eta} \left(1 - \frac{1}{3} k \eta \|f_i\|_\infty \right) + \frac{ky}{3} \|f_i\|_\infty \right)$$

$$\geq \sum_{i=1}^{n} \sum_{k=1}^{r_i} \alpha_k(i) \left(k\sqrt{2 V_{i,f_i} y} + \frac{ky}{3} \|f_i\|_\infty \right) = \sum_{i=1}^{n} \sum_{k=1}^{r_i} k \alpha_k(i) \left[\sqrt{2 V_{i,f_i} y} + \frac{y}{3} \|f_i\|_\infty \right]. \tag{46}$$

Optimizing η in (45) by Eq. (46), we have

$$P\left(\sum_{i=1}^{n} \int_{S_i} f_i(x) \left[CP_{i,r_i}(dx) - \sum_{k=1}^{r_i} k\alpha_k(i) \lambda(x) \right] dx \geq \sum_{i=1}^{n} \sum_{k=1}^{r_i} k\alpha_k(i) \left(\sqrt{2 y V_{i,f_i}} + \frac{y}{3} \|f_i\|_\infty \right) \right)$$

$$\leq e^{-ny}. \tag{47}$$

For $i = 1, 2, \ldots$, let $r_i \to \infty$ in (47), then $CP_{i,r_i}(A) \xrightarrow{d} CP_i(A)$. It implies the concentration inequality (8).

Define $c_{1n} := \sum_{i=1}^{n} \mu_i \sqrt{2 V_{i,f_i}}$, $c_{2n} := \sum_{i=1}^{n} \frac{\mu_i}{3} \|f_i\|_\infty$, let $t = c_{1n}\sqrt{y} + c_{2n} y$, and we obtain the expression of y by solving a quadratic equation.

Availability of data and materials
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
The authors completed the paper and approved the final manuscript.

Author details
[1] School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing, China. [2] Mathematics Department, Rutgers University, Piscataway, USA.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References
1. Aczél, J.: On composed Poisson distributions, III. Acta Math. Hung. **3**(3), 219–224 (1952)
2. Baraud, Y., Birgé, L.: Estimating the intensity of a random measure by histogram type estimators. Probab. Theory Relat. Fields **143**(1–2), 239–284 (2009)
3. Bickel, P.J., Ritov, Y.A., Tsybakov, A.B.: Simultaneous analysis of Lasso and Dantzig selector. Ann. Stat. **37**, 1705–1732 (2009)
4. Boucheron, S., Lugosi, G., Massart, P.: Concentration Inequalities: A Nonasymptotic Theory of Independence. Oxford University Press, London (2013)
5. Bühlmann, P., van de Geer, S.A.: Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, Berlin (2011)
6. Chareka, P., Chareka, O., Kennendy, S.: Locally sub-Gaussian random variables and the strong law of large numbers. Atl. Electron. J. Math. **1**(1), 75–81 (2006)
7. Cleynen, A., Lebarbier, E.: Segmentation of the Poisson and negative binomial rate models: a penalized estimator. ESAIM Probab. Stat. **18**, 750–769 (2014)
8. Copeland, A.H., Regan, F.: A postulational treatment of the Poisson law. Ann. Math. **37**, 357–362 (1936)
9. Das, A.: Design and Analysis of Statistical Learning Algorithms which Control False Discoveries. Doctoral dissertation, Universität zu Köln (2018)
10. Feller, W.: An Introduction to Probability Theory and Its Applications, vol. I, 3rd edn. Wiley, New York (1968)
11. Giles, D.E.: Hermite regression analysis of multi-modal count data. Econ. Bull. **30**(4), 2936–2945 (2010)
12. Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, Boca Raton (2015)
13. Hilbe, J.M.: Negative Binomial Regression, 2nd edn. Cambridge University Press, Cambridge (2011)
14. Houdré, C.: Remarks on deviation inequalities for functions of infinitely divisible random vectors. Ann. Probab. **30**, 1223–1237 (2002)
15. Houdré, C., Privault, N.: Concentration and deviation inequalities in infinite dimensions via covariance representations. Bernoulli **8**(6), 697–720 (2002)
16. Ivanoff, S., Picard, F., Rivoirard, V.: Adaptive Lasso and group-Lasso for functional Poisson regression. J. Mach. Learn. Res. **17**(55), 1–46 (2016)
17. Jiang, X., Raskutti, G., Willett, R.: Minimax optimal rates for Poisson inverse problems with physical constraints. IEEE Trans. Inf. Theory **61**(8), 4458–4474 (2015)
18. Johnson, N.L., Kemp, A.W., Kotz, S.: Univariate Discrete Distributions, 3rd edn. Wiley, New York (2005)
19. Kingman, J.F.C.: Poisson Processes. Oxford University Press, London (1993)
20. Kontoyiannis, I., Madiman, M.: Measure concentration for compound Poisson distributions. Electron. Commun. Probab. **11**, 45–57 (2006)
21. Last, G., Penrose, M.D.: Lectures on the Poisson Process. Cambridge University Press, Cambridge (2017)
22. Li, B., Zhang, H., He, J.: Some characterizations and properties of COM-Poisson random variables. Commun. Stat., Theory Methods (2019). https://doi.org/10.1080/03610926.2018.1563164
23. Li, Q.: Bayesian Models for High-Dimensional Count Data with Feature Selection. Doctoral dissertation, Rice University (2016)
24. Luc, D.T.: Multiobjective Linear Programming: An Introduction. Springer, Berlin (2016)
25. Mallick, H., Tiwari, H.K.: EM adaptive LASSO a multilocus modeling strategy for detecting SNPs associated with zero-inflated count phenotypes. Front. Genet. **7**, Article 32 (2016)
26. Nielsen, F., Nock, R.: Sided and symmetrized Bregman centroids. IEEE Trans. Inf. Theory **55**(6), 2882–2904 (2009)
27. Petrov, V.V.: Limit Theorems of Probability Theory: Sequences of Independent Random Variables. Clarendon, Oxford (1995)
28. Raskutti, G., Wainwright, M.J., Yu, B.: Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. IEEE Trans. Inf. Theory **57**(10), 6976–6994 (2011)
29. Reynaud-Bouret, P.: Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. Probab. Theory Relat. Fields **126**(1), 103–153 (2003)
30. Rigollet, P., Hütter, J.C.: High dimensional statistics (2019). http://www-math.mit.edu/~rigollet/PDFs/RigNotes17.pdf
31. Sato, K.: Lévy Processes and Infinitely Divisible Distributions, revised edn. Cambridge University Press, Cambridge (2013)
32. Städler, N., Bühlmann, P., van de Geer, S.A.: ℓ_1-penalization for mixture regression models. Test **19**, 209–256 (2010)
33. van de Geer, S.A.: The deterministic lasso. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) Zürich (2007)
34. Wainwright, M.J.: High-Dimensional Statistics: A Non-asymptotic Viewpoint, vol. 48. Cambridge University Press, Cambridge (2019)
35. Wang, Y.H., Ji, S.: Derivations of the compound Poisson distribution and process. Stat. Probab. Lett. **18**(1), 1–7 (1993)
36. Yang, X., Zhang, H., Wei, H., Zhang, S.: Sparse density estimation with measurement errors. arXiv preprint (2019). arXiv:1911.0621
37. Yu, Y.: High-dimensional variable selection in Cox model with generalized Lasso-type convex penalty (2010). https://people.maths.bris.ac.uk/~yy15165/index_files/Cox_generalized_convex.pdf
38. Zhang, H., Jia, J.: Elastic-net regularized high-dimensional negative binomial regression: consistency and weak signals detection. arXiv preprint (2017). arXiv:1712.03412
39. Zhang, H., Li, B.: Characterizations of discrete compound Poisson distributions. Commun. Stat., Theory Methods **45**(22), 6789–6802 (2016)
40. Zhang, H., Liu, Y., Li, B.: Notes on discrete compound Poisson model with applications to risk theory. Insur. Math. Econ. **59**, 325–336 (2014)