Electronic Journal of Statistics
Vol. 16 (2022) 3789–3835
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2029

Poisson mean vector estimation with nonparametric maximum likelihood estimation and application to protein domain data*

Hoyoung Park† **and Junyong Park**‡

Department of Statistics, Sookmyung Women's University, Seoul, Korea

Department of Statistics, Seoul National University, Seoul, Korea
e-mail: hyparks1015@gmail.com; junyongpark@snu.ac.kr

Abstract: In this paper, we propose the nonparametric empirical Bayes (NPEB) estimator based on the nonparametric maximum likelihood estimation (NPMLE) in Poisson mean vector estimation, also known as the g-modeling in the nonparametric empirical Bayes method. Due to the recent developments of highly scalable algorithms of empirical Bayes, it is more attractive to use g-modelling, while most of the studies have focused on the performance of f-modeling in the NPEB estimator. We study the theoretical properties of the NPEB estimator of Poisson mean vector based on g-modeling combined with the NPMLE, such as the convergence rate, and compare our result with some existing studies. Our simulation studies and real data examples of protein domain data show that the estimator based on the g-modeling outperforms existing f-modeling based estimators in both computational efficiency and accuracy.

MSC2020 subject classifications: Primary 62G05; secondary 62C12, 62C25.
Keywords and phrases: Compound decision problem, empirical Bayes, nonparametric maximum likelihood estimate, Poisson distribution.

Received August 2021.

Contents

*This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A2C1A01100526).
†First supporter of the project.
‡Corresponding author.

1. Introduction

In this paper, we consider the simultaneous mean vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ estimation based on the observed data vector $\boldsymbol{X} = (X_1, \ldots, X_n)$. There are two general problems of estimation of Normal and Poisson mean vectors, and there has been tremendous effort to develop efficient estimators for both problems. Regarding the Normal mean vector, the most recent development of estimators includes the nonparametric empirical Bayes (NPEB) approach by [2] and general maximum likelihood estimation (GMLE) by [14]. These methods are shown to be better than classical linear estimators such as Stein's shrinkage estimator. [8] demonstrated two types of empirical Bayes approaches, called g-modeling and f-modeling: the former is based on estimating the prior distribution of mean values, while the latter is based on estimation of the marginal density. In fact, [2] and [14] are f-modeling and g-modeling, respectively. [14] included delicate theoretical results on g-modeling on estimation of the Normal mean vector, including risk-consistency for different scenarios. The extensive literature on the Poisson mean vector estimation has also been done along with this Normal mean vector. For example, the James-Stein type estimators in [12] and [5] and NPEB in [19], [24] and [7]. [25] and [23] originally proposed NPEB estimator based on f-modeling using estimate of the marginal probability function of X_i. In addition, [20] proposed a regularized Robbins type NPEB estimator and investigated the asymptotic performance of their regularized NPEB estimator. Meanwhile,

[3] showed the necessity of modifying the Robbin's estimator in their numerical studies and proposed modified version of the Robbin's estimator by smoothing the marginal probabilities and imposing monotonicity on the estimated decision function. Furthermore, they also considered a Normal approximation to the Poisson model.

To the best of our knowledge, most studies on estimating the Poisson mean vector have focused on the f-modeling and its theoretical properties. On the other hand, the g-modeling in the estimation of Poisson mean vector based on the nonparametric maximum likelihood estimation (NPMLE) [16] has not been investigated theoretically, although it has been used in practice, and there have been many studies on the Normal model [14, 15, 13].

Thanks to the development of some highly scalable algorithms, such as [18], the g-modeling based on the NPMLE is getting more attention in such estimation problems since it is more efficient in reflecting the structure of mean values, for example, bi-modality and sparsity, etc. With this motivation, we propose the Poisson mean vector estimator based on the g-modeling version of the nonparametric empirical Bayes method (g-NPEB). In particular, we present theoretical results of our proposed g-NPEB from the view of the convergence rate of risk and provide numerical studies to compare with methods based on f-modeling such as [20] and [3]. Inspired by many previous kinds of research such as [30, 28, 32, 14], our theoretical development and proof techniques are based on the empirical process for evaluating the properties of a maximum likelihood estimator.

In terms of application, estimation of the Poisson mean vector has been paid relatively less attention than that of the Normal mean vector. However, since counting data in genetics, such as RNAseq and protein data, are becoming more common, it is more demanding to consider accurate and computationally efficient statistical inference for those types of count data. As real data examples, we provide protein domain data sets that consist of mutation counts from different positions in the protein domain. The numbers of mutation counts for different positions are modeled using Poisson distribution, and it is of interest to estimate the intensity of mutations for each position.

This paper is organized as follows. In Section 2, we present the g-NPEB estimator based on the NPMLE to estimate the mixing distribution. Section 3 includes the main results on asymptotic properties of the estimator, such as the convergence rate of the risk of the estimator and section 4 provides the detail of the proof of the main results. Section 5 and section 6 show simulation studies and real data examples of protein domains, respectively, and we also compare with some existing methods. We provide concluding remarks and proofs in Section 7 and Appendices, respectively.

2. Simultaneous Poisson mean vector estimation

Suppose we have independent bivariate random vectors (X_i, θ_i) $i = 1, \ldots, n$ such that $X_i|\theta_i \sim Poisson(\theta_i)$ for a given θ_i under a probability measure $P_{n,\boldsymbol{\theta}}$

with deterministic $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$. When $(\theta_1, \ldots, \theta_n)$ is fixed, estimation of $\boldsymbol{\theta}$ is known as a compound decision problem in [25] On the other hand, when θ_i s are generated from an unknown distribution G, it is known as an empirical Bayes estimation problem. In the next section, we describe the relationship between a compound decision problem and empirical Bayes. Throughout this paper, we consider the compound decision problem for the unknown and deterministic parameters $\boldsymbol{\theta}$ which is also considered for Normal distribution in [14].

Based on the observed vector $(X_1, \ldots, X_n)$, one main goal is to present estimator of $\boldsymbol{\theta}$ by minimizing the square error $L_n(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n (\widehat{\theta}_i - \theta_i)^2$ or $EL_n(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ for some estimator $\widehat{\boldsymbol{\theta}}$. There have been numerous studies on this estimation problem; James-Stein type estimators by [12] and [5] nonparametric empirical Bayes estimator (NPEB) by [19] and [24] and [7]. One most typical well known estimator is the Robbins estimator, which is the nonparametric empirical Bayes type estimator proposed by [24] which is

$$\widehat{\theta} = \widehat{E}(\theta | X = x) = \frac{(x+1)\widehat{p}_G(x+1)}{\widehat{p}_G(x)}, \tag{2.1}$$

where $p_G(x) = \int p(x|\theta) dG(\theta)$ for $p(x|\theta) = \frac{e^{-\theta}\theta^x}{x!}$ and its estimator $\widehat{p}_G(x) = \frac{\#\{X_i = x\}}{n}$ is the relative frequency of $X_i = x$. The estimated Bayes rule in (2.1) depends only on estimator of the marginal probability $p_G(x)$. However, if $\widehat{p}_G(x) = 0$ or $\widehat{p}_G(x) \approx 0$ and $(x+1)\widehat{p}_G(x+1) > 0$, then the estimator (2.1) is not well defined or not stable. To avoid this, there are different ways of estimating $p_G(x)$ such as regularization of 2.1, for example [20] proposed to use

$$\widehat{\theta} = x + \frac{(x+1)\widehat{p}_G(x+1) - x\widehat{p}_G(x)}{\widehat{p}_G(x) \vee \rho}, \tag{2.2}$$

for some regularizing constant ρ and $a \vee b = \max(a, b)$. [20] investigated some asymptotic performances of (2.2) such that the regularized estimator in (2.2) obtains the risk consistency, but not any result of rate of convergence. The numerical studies in [20] demonstrates that the estimator in (2.2) achieves slow convergence rate to the optimal risk. [3] also provided the convergence rate of the risk of (2.1). In fact, [3] does not discuss the theoretical property of their proposed estimator, which is based on the tuning parameter chosen by cross-validation.

While [20] and [3] used the idea of f-modeling, we consider the NPEB using the NPMLE of G which is based on the g-modeling in [8].This idea in the simultaneous estimation problem has been used in [14] for Normal mean vector estimation. In Poisson mean vector estimation, it can be used similarly. However, the derivation of asymptotic properties is different from those in Normal mean vector estimation since the parameter of the Poisson distribution affects the variance as well as mean and the mean of the Normal distribution with fixed scale parameter is simply location parameter. It is attractive to use the g-modeling since the estimation of G is expected to provide a more flexible fitting to the various structure of mean values.

3. Main results

In this section, we provide asymptotic results for Poisson mean vector estimation based on the NPMLE. Although there have been numerous studies on simultaneous estimation of Poisson means, theoretical studies such as optimality are limited. [20] provided the risk consistency of the regularized Robbins estimator under increasing parameter space as n increases. [3] proposed the data splitting method to estimate a tuning parameter to improve the Robbins estimator. Both of these studies focus on the Robbins estimator, which is f-modeling in [8].

In particular, we discuss the asymptotic performance of the g-modeling which is based on estimating G resulting in $\widehat{p}_G(x) = p_{\widehat{G}}(x)$. Throughout this paper, we assume the true underlying G_0 of mean values has a bounded support such that

$$G_0([0, b]) = 1, \tag{3.1}$$

Moreover, we do not consider the trivial situation where the distribution G_0 degenerates to point 0. That is, there exists constant $b_1 > 0$ such that $G_0([0, b_1)) \le \delta_0$ for some $\delta_0 < 1$ which implies

$$G_0([b_1, b]) = w_0 \ge 1 - \delta_0. \tag{3.2}$$

This bounded support condition (3.1) was also used in [3] in their theoretical study. For Normal distribution, [14] allowed support to be increased as n increases while [6] considered a compact support for G. We investigate the situations where b is an unknown positive constant or increases with the number of observations n. For the latter, we develop our theory by considering cases where the order of b is up to $o\left(\frac{\log n}{\log\log n}\right)$.

Since we assume that b is unknown, we use a support for estimator $\widehat{G}_n$ such that

$$\widehat{G}_n([0, M]) = 1, \tag{3.3}$$

for an increasing sequence of M depending on n. We take M so that the support $[0, M]$ of $\widehat{G}_n$ can include $[0, b]$ of G_0 as n increases. It is worth noting that the mean parameter θ in Poisson distribution is somewhat different from that in Normal distribution because one parameter serves as both mean and variance. Therefore, there must be a constraint on the support of the Poisson distribution. Specifically, we set the maximum value b of the support only up to $o\left(\frac{\log n}{\log\log n}\right)$ to expect a similar theoretical result to that from Normal distribution [14]. Moreover, in the Normal case, there are many elegant properties (e.g., the probability density function with a sufficiently thin tail and no deformation depending on the average parameter), so a lot of previous studies have been accumulated that can be used to develop the theory of the Normal distribution model [30, 27, 33, 14]. On the contrary, in the Poisson case such nice properties cannot be expected and some careful and meticulous approaches are needed.

In the theoretical results described later, we point out a tricky part that arises from the Poisson model in Remark 4.1 and 4.2.

As in [23], the compound estimation of a vector of deterministic Poisson parameters is highly related to the Bayes estimation of a single random Poisson mean with proper prior. For deteministic $\boldsymbol{\theta}$, define the empirical distribution of θ_i s,

$$G_n(u) = \frac{1}{n}\sum_{i=1}^{n} I(\theta_i \leq u), \tag{3.4}$$

where $I(\cdot)$ is the indicator function, equal to one if the condition in parentheses is true and zero otherwise. Next, consider the following single random variable which has the Bayesian perspective structure,

$$Y|\lambda \sim P(\lambda), \quad \lambda \sim G_0. \tag{3.5}$$

If the prior distribution $G_0 = G_n$ which is the empirical distribution (3.4), then for any separable rule $\widehat{\boldsymbol{\theta}} = t(\boldsymbol{X})$, the *fundamental theorem of compound decisions* [25] implies, under the squared error loss, the compound risk of a separable rule $t(\boldsymbol{X}) = (t(X_1), \cdots, t(X_n))$ and the Bayes risk of a $t(Y)$ with prior G_n are same, i.e.,

$$E_{n,\boldsymbol{\theta}} L_n(t(\boldsymbol{X}), \boldsymbol{\theta}) = E_{G_n}(t(Y) - \lambda)^2, \tag{3.6}$$

where $L_n(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \frac{1}{n}||\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2 = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\theta}_i - \theta_i)^2$. Furthermore, the compound decision problem is not limited to the squared loss, but can also explain about the general loss function. See [33] for details. For any prior G, denote the Bayes rule as

$$t_G^* = \underset{t}{argmin} E_G(t(Y) - \lambda)^2, \tag{3.7}$$

and the minimum Bayes risk as

$$R^*(G) = E_G(t_G^*(Y) - \lambda)^2. \tag{3.8}$$

Then, out of all separable rules, the compound risk is minimized by the Bayes rule along with the empirical distribution prior G_n defined on (3.4), resulting in the oracle benchmark as,

$$R^*(G_n) = E_{n,\boldsymbol{\theta}} L_n(t_{G_n}^*(\boldsymbol{X}), \boldsymbol{\theta}) = \min_{t(\cdot)} E_{n,\boldsymbol{\theta}} L_n(t(\boldsymbol{X}), \boldsymbol{\theta}). \tag{3.9}$$

Therefore, as in [14, 13], our goal is to derive an estimator $\widehat{t}_n(\cdot)$ of the general EB oracle rule $t_{G_n}^*(\cdot)$ such that minimize the regret for the square root of the MSE $\widetilde{r}_{n,\boldsymbol{\theta}}(\widehat{t}_n)$ which is defined as,

$$\widetilde{r}_{n,\boldsymbol{\theta}}(\widehat{t}_n) = \sqrt{E_{n,\boldsymbol{\theta}} L_n(\widehat{t}_n(\boldsymbol{X}), \boldsymbol{\theta})} - \sqrt{R^*(G_n)}. \tag{3.10}$$

We now define the Bayes rule in the estimation of Poisson mean vector. For any distribution G of θ_i s, the Bayes rule is

$$t_G^*(\mathbf{X}) = (t_G^*(X_1), \ldots, t_G^*(X_n)), \tag{3.11}$$

where

$$t_G^*(x) = \frac{(x+1)p_G(x+1)}{p_G(x)}. \tag{3.12}$$

Additionally, let $g_G(x) = (x+1)p_G(x+1) - xp_G(x)$ and define the regularized Bayes rule with regularizing constant ρ,

$$t_G^*(\mathbf{X};\rho) = (t_G^*(X_1;\rho), \ldots, t_G^*(X_n;\rho)), \tag{3.13}$$

where

$$t_G^*(x;\rho) = x + \frac{g_G(x)}{p_G(x) \vee \rho}. \tag{3.14}$$

However, an oracle estimator, $t_{G_n}^*(\cdot)$ is infeasible in practice as it requires information on the unknown $\theta_1, \cdots, \theta_n$ through the empirical distribution G_n defined on (3.4). Therefore, [23] proposed the concept of the general empirical Bayes estimator, which aims to approximate the oracle rule $t_{G_n}^*(\cdot)$ without any assumptions about the prior.

Meanwhile, to derive an estimator $\widehat{t}_n(\cdot)$ of the general EB oracle rule $t_{G_n}^*$, we replace G_n with the NPMLE estimator $\widehat{G}_n$ by Kiefer-Wolfowitz in [16] satisfying

$$\widehat{G}_n = \text{argmax}_{G \in \mathcal{G}} \prod_{i=1}^n \int p(X_i|\theta) dG(\theta), \tag{3.15}$$

where $\mathcal{G}$ is a certain class of G.

As in [14, 13, 26], we allow approximate solutions to be used.

That is, the NPMLE is any solution of

$$\widehat{G}_n \in \mathcal{G}, \quad \prod_{i=1}^n \int p(X_i|\theta) d\widehat{G}_n \geq q_n \sup_{G \in \mathcal{G}} \prod_{i=1}^n \int p(X_i|\theta) dG(\theta), \tag{3.16}$$

with $q_n = \frac{e}{n^2} \wedge 1$.

Remark 3.1. We can rewrite (3.16) as

$$\frac{1}{n} \log q_n + \frac{1}{n} \sum_{i=1}^n \log \int p(X_i|\theta) d{G_n}^* \leq \frac{1}{n} \sum_{i=1}^n \log \int p(X_i|\theta) d\widehat{G}_n \tag{3.17}$$

$$\leq \frac{1}{n} \sum_{i=1}^n \log \int p(X_i|\theta) dG_n^*, \tag{3.18}$$

where G_n^* is the optimal solution which maximize $\prod_{i=1}^n \int p(X_i|\theta) dG(\theta)$ among the class $G \in \mathcal{G}$. Due to the characteristics of q_n, our solution $\widehat{G}_n$ requires that it exist within some error compared to the optimal solution G_n^*, so (3.17) and (3.18) is a relatively moderate condition. Specifically, we use this inequality (3.16) to quantify some regularized constant $\rho = \rho_n$ which connecting the g-NPEB and regularized version of g-NPEB defined defined below.

To find out $\widehat{G}_n$, we assume $\widehat{G}_n$ has the form of

$$\widehat{G}_n = \sum_{j=1}^{m} \widehat{w}_j \delta_{b_j}, \quad \widehat{w}_j \geq 0, \text{ and } \quad \sum_{j=1}^{m} \widehat{w}_j = 1, \tag{3.19}$$

for given grid points b_j, $1 \leq j \leq m$, $0 < b_1 < b_2 < \cdots < b_m = M$ and δ_u is a point mass function at u. There are various ways to determine b_i and we determine b_i based on n observed data points $X_1, \cdots, X_n$. Specifically, as in [14, 6, 9], we consider that b_i s are equally spaced apart in the interval $[X_{(1)} = \min(X_1, \ldots, X_n), X_{(n)} = \max(X_1, \ldots, X_n)]$ and proceed with simulation and case studies.

Remark 3.2. If $\theta_i = \lambda$, $i = 1, \ldots, n$, for some fixed $\lambda > 0$, [17] showed that $P_{n,\boldsymbol{\theta}}\left(X_{(n)} \in (I_n, I_n + 1)\right) \to 1$ as $n \to \infty$ where $I_n \sim \log n / \log\log n$. Therefore, considering $b = o\left(\frac{\log n}{\log\log n}\right)$ in our case, the interval support $[X_{(1)}, X_{(n)}]$ asymptotically covers the support $[0, b]$ of true G_0.

There are many approaches to solve the optimization in (3.15), for example, the EM algorithm, the convex optimization in [18] etc. In our numerical studies, we use the *Pmix* function in the R packages "`REBayes`" [11] to obtain $\widehat{G}_n$.

In our theoretical studies, we derive asymptotic results for a $\widehat{G}_n$ satisfying (3.16). In practice, we use $\widehat{G}_n$ estimated from the REBayes function which is assumed to satisfy (3.16). See [13, 26] for a similar argument.

Therefore for any $\widehat{G}_n$ satisfying (3.16), we further define the proposed estimator based on the g-modeling version of the nonparametric empirical Bayes estimator (g-NPEB) which is

$$\widehat{t}_n(\mathbf{X}) \equiv t^*_{\widehat{G}_n}(\mathbf{X}) = (t^*_{\widehat{G}_n}(X_1), \ldots, t^*_{\widehat{G}_n}(X_n)), \tag{3.20}$$

and its regularized version of g-NPEB is

$$t^*_{\widehat{G}_n}(\mathbf{X}; \rho) = (t^*_{\widehat{G}_n}(X_1; \rho), \ldots, t^*_{\widehat{G}_n}(X_n; \rho)). \tag{3.21}$$

We investigate the accuracy of $t^*_{\widehat{G}_n}(\mathbf{X})$ as an estimate of $t^*_G(\mathbf{X})$ in detail with respect to the $L_n(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \frac{1}{n}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2$ loss. The corresponding result is presented in Theorem 3.1 as the main result in this paper.

Remark 3.3. Related to the context of the empirical Bayes, the problem we consider in this paper is a special case when $G = G_n$ defined in (3.4) is considered as a prior of θ. The Bayes rule under $X|\theta \sim Poisson(\theta)$ and $\theta \sim G_n$ is $E_{G_n}(\theta|X = x)$, so this Bayes rule is the optimal among all separable rules under the loss $\frac{1}{n}||\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2$.

Before presenting our theoretical results, we define some notations that will be used throughout the paper.

Notations.

1. For any $r \in \mathbb{R}$, $\lceil r \rceil$ ($\lfloor r \rfloor$) is the smallest (largest) integer greater (less) than or equal to r.
2. $a_n \asymp b_n$ denotes $a_n/b_n + b_n/a_n = O(1)$.
3. $a_n \lesssim b_n$ denotes $a_n/b_n = O(1)$.
4. $\|f\|_h \equiv \left\{\int f^2(x)h(x)dx\right\}^{1/2}$ is the $L_2(h(x)dx)$ norm for $h \geq 0$.
5. $\|h\|_{\infty,B} \equiv \sup_{|x|\leq B} |h(x)|$ is the seminorm with $B > 0$.
6. For $\boldsymbol{x} \in \mathbb{R}^p$ and $\boldsymbol{y} \in \mathbb{R}^p$ with $p \in \mathbb{N}$
   - $\|\boldsymbol{x} - \boldsymbol{y}\|_1 = \sum_{i=1}^p |x_i - y_i|$.
   - $\|\boldsymbol{x} - \boldsymbol{y}\| = \left(\sum_{i=1}^p (x_i - y_i)^2\right)^{1/2}$.
   - $\|\boldsymbol{x} - \boldsymbol{y}\|_\infty = \sup_{1\leq i\leq p} |x_i - y_i|$.
7. For any class $\mathcal{A}$, $\#\mathcal{A}$ denotes the cardinality of $\mathcal{A}$.
8. Let $d_H(p_G, p_{G_0})$ be the Hellinger distance, i.e.,

$$d_H(p_G, p_{G_0})^2 = \sum_{x=0}^{\infty} \left(\sqrt{p_G(x)} - \sqrt{p_{G_0}(x)}\right)^2. \tag{3.22}$$

Now we provide our main theorem of this paper regarding an oracle inequality that gives upper bounds for the regret of the proposed g-NPEB as follows.

Theorem 3.1. *Let $\{X_i\}_{i=1}^n$ be independent Poisson random variables such that $X_i|\theta_i \sim Poisson(\theta_i)$ under $P_{n,\boldsymbol{\theta}}$ with a deterministic $\boldsymbol{\theta} \in [0, b]^n \subset \mathbb{R}^n$ for some unknown positive b that can have up to an order of $o\left(\frac{\log n}{\log\log n}\right)$. Additionally, let $L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n(\widehat{\theta}_i - \theta_i)^2$ be the average squared, and $t^*_{\widehat{G}_n}(\mathbf{X})$ be the proposed g-NPEB defined on (3.20) with $\widehat{G}_n$ satisfying (3.16). Then there exists some universal constants C^* such that*

$$\begin{aligned}\tilde{r}_{n,\boldsymbol{\theta}}(t^*_{\widehat{G}_n}(\mathbf{X})) &= \sqrt{E_{n,\boldsymbol{\theta}} L_n(t^*_{\widehat{G}_n}(\mathbf{X}), \boldsymbol{\theta})} - \sqrt{R^*(G_n)} \\ &\leq \left[C^* \left\{(\log n)^3 \vee (\log n)^2 (b_1)^2\right\} \epsilon_n^2\right]^{1/2}, \end{aligned} \tag{3.23}$$

where $b_1 = b \vee 1$, and $n\epsilon_n^2 = (\log n)^3$.

Proof. See Appendix. $\square$

Remark 3.4. Theorem 3.1 shows that, for $b = O(\sqrt{\log n})$, we have

$$E_{n,\boldsymbol{\theta}} L_n(t^*_{\widehat{G}_n}(\mathbf{X}), \boldsymbol{\theta}) - R^*(G_n) \leq M_0 \epsilon_n^2 (\log n)^3 = O((\log n)^6/n).$$

On the other hand, [3] claimed that the Robbins estimator in (2.1) achieves the convergence rate

$(\log n)^2/(\log\log n)^2/n$ which is a faster rate than $(\log n)^6/n$. The result in Theorem 3.1 provides the upper bound for the regret while

$(\log n)^2/(\log\log n)^2/n$ in [3] is the exact rate. There may exist a chance that the actual rate of the regret in (3.23) has a faster rate than $(\log n)^2/(\log\log n)^2/n$. Our numerical studies in section 5 demonstrate that the proposed g-NPEB $t^*_{\widehat{G}_n}(\mathbf{X})$ converges to the Bayes risk faster than two f-modelling based estimators in [3] and [20]. In fact, the Robbins estimator in (2.1) is unstable since the denominator $\widehat{p}_G(x)$ can be zero or fairly small, so (2.1) is not directly used without further modification in practice.

Regarding the proof of our Theorem 3.1, we refer to the proof process of Theorem 5 in [14], and the complete proof is represented in the appendix section. Here, we provide a simple overview of the proofs of our main results, Theorem 3.1. Since our proposed g-NPEB $t^*_{\widehat{G}_n}(\mathbf{X})$ defined in (3.20) can be unstable when the denominator of it closes to 0, we replace $t^*_{\widehat{G}_n}(\mathbf{X})$ with the regularized version $t^*_{\widehat{G}_n}(\mathbf{X};\rho)$ defined on (3.21) in our theoretical development to exclude such circumstances. To justify this replacement step, we present Proposition 4.1 below which implies $t^*_{\widehat{G}_n}(X_i) = t^*_{\widehat{G}_n}(X_i;\rho_n)$, $i = 1,\cdots,n$ with proper regularizing constant ρ_n. We define $A_n = \left\{d_H\left(p_{\widehat{G}_n}, p_{G_n}\right) \le t^*\epsilon_n\right\}$, where $t^* = t_* \vee 1$ and t_* is sufficiently large constant. The large deviation inequality in Theorem 4.2 implies $P_{n,\boldsymbol{\theta}}(A_n^c) \le 1/n$. Furthermroe, we define a $2\eta^*$ net $\left\{t^*_{H_j}(\cdot;\rho_n), j \le N\right\}$ of $\mathcal{T}_{\rho_n} \cap \{t^*_G : d_H(p_G, p_{G_n}) \le t^*\varepsilon_n\}$ under the seminorm $\|\cdot\|_{\infty,M}$ where $\mathcal{T}_\rho = \left\{t^*_G(\cdot;\rho) : G \in \mathcal{G}_{[0,M]}\right\}$ and η^* defined in Theorem 4.3. Then we show that N is manageable size by deriving the entropy bound of N in Theorem 4.3.

Next, we show $\sqrt{n}\tilde{r}_{n,\boldsymbol{\theta}}(t^*_{\widehat{G}_n}(\mathbf{X})) \le \sqrt{E_{n,\boldsymbol{\theta}}\left(\sum_{j=0}^4 |\zeta_{jn}|\right)^2}$ where

$$\zeta_{0n} = ||t_{\widehat{G}_n}(\mathbf{X}) - t_{\widehat{G}_n}(\mathbf{X};\rho_n)|| \tag{3.24}$$

$$\zeta_{1n} = \left\|t_{\widehat{G}_n}(\mathbf{X};\rho_n) - \boldsymbol{\theta}\right\| I_{A_n^c}, \tag{3.25}$$

$$\zeta_{2n} = \left\{\left\|t^*_{\widehat{G}_n}(\mathbf{X};\rho_n) - \boldsymbol{\theta}\right\| I_{A_n} - \max_{j\le N}\left\|t^*_{H_j}(\mathbf{X};\rho_n) - \boldsymbol{\theta}\right\|\right\}_+, \tag{3.26}$$

$$\zeta_{3n} = \max_{j\le N}\left\{\left\|t^*_{H_j}(\mathbf{X};\rho_n) - \boldsymbol{\theta}\right\| - E_{n,\boldsymbol{\theta}}\|t^*_{H_j}(\mathbf{X};\rho_n) - \boldsymbol{\theta}\|\right\}_+, \tag{3.27}$$

$$\zeta_{4n} = \max_{i<N}\sqrt{E_{n,\boldsymbol{\theta}}\|t^*_{H_j}(\mathbf{X};\rho_n) - \boldsymbol{\theta}\|^2} - \sqrt{nR^*(G_n)}. \tag{3.28}$$

To end the proof of Theorem 3.1, we control each of the four terms related to $\{\zeta_{jn}\}_{j=0}^4$ as follows.

0. $E_{n,\boldsymbol{\theta}}\zeta_{0n}^2$ is bounded by $C_{0,0}n\epsilon_n^2$ for some constant $C_{0,0}$ based on Lemma 4.1.
1. $E_{n,\boldsymbol{\theta}}\zeta_{1n}^2$ is bounded by $C_{0,1}n\epsilon_n^2$ for some constant $C_{0,1}$ using Theorem 4.2, Lemma 4.4 and some elementary calculations.

2. $E_{n,\boldsymbol{\theta}}\zeta_{2n}^2$ is bounded by $C_{0,2}n\epsilon_n^2$ for some constant $C_{0,2}$ using the non-random net $\left\{t^*_{H_j}\left(\cdot;\rho_n\right), j \leq N\right\}$ with size N is asymptotically less then $\left\{\log \frac{1}{\eta}\right\}^2$ from Theorem 4.3.
3. $E_{n,\boldsymbol{\theta}}\zeta_{3n}^2$ is bounded by $C_{0,3}(\log n)^2 n\epsilon_n^2$ for some constant $C_{0,3}$ using the isopherimetric inequality for Poisson process case in Lemma 4.5 in addition to Theorem 4.3.
4. $E_{n,\boldsymbol{\theta}}\zeta_{4n}^2$ is bounded by $C_{0,4}\left\{(\log n)^3 \vee (\log n)^2\left(b_1\right)^2\right\} n\epsilon_n^2$ for some constant $C_{0,4}$ using Theorem 4.1 which implies upper bound of the regret of not knowing the empirical distribution G_n.

In the following section 4, we present relevant lemmas and theorems for the upper bounds of $E_{n,\boldsymbol{\theta}}\zeta_{jn}^2$, $j = 1, \cdots, 4$ to completes the proof of Theorem 3.1.

4. Proof of Theorem 3.1

This section consists of subsections which are used to show the connection between g-NPEB in (3.20) and regularized version of g-NPEB (3.21) with specific regularized constant and computation of upper bounds for $E_{n,\boldsymbol{\theta}}\zeta_{jn}^2$. For the upper bounds of $E_{n,\boldsymbol{\theta}}\zeta_{jn}^2$, we investigate the properties of regularized Bayes rule and deliver the Hellinger consistency of the NPMLE to exploit the large deviation inequality. Finally, we demonstrate the important theoretical development by showing the entropy bound of the regularized Bayes rule.

4.1. Relationship between g-NPEB and regularized version of g-NPEB

The following proposition provides the connection between our g-NPEB in (3.20) and the regularized version of g-NPEB in (3.21). The proof of the following proposition is described in Proposition 2 of [14]. Although [14] dealt with the Normal case, they provided a general proof for all parametric families.

Proposition 4.1. *Let $\{X_i\}_{i=1}^n$ be given data. Let $\widehat{G}_n$ be an approximate NPMLE of a mixing distribution satisfying (3.16) with $q_n = \frac{e}{n^2} \wedge 1$. Then for all $j = 1, \cdots, n$,*

$$p_{\widehat{G}_n}(X_j) \geq \frac{q_n}{en}\sup_u \left\{p(X_j|u)\right\} = \frac{q_n}{en}\left\{p(X_j|X_j)\right\}. \tag{4.1}$$

Remark 4.1. Proposition 4.1 shows that the lower bound of $p_{\widehat{G}_n}(X_j)$ depends on $p(X_j|X_j)$, which is a random quantity in Poisson distribution while the lower bound in case of Normal distribution in [14] does not depend on the data. As stated in the following lemma, we can take a deterministic sequence as a low bound leads to asymptotic equivalence between $t^*_{\widehat{G}_n}(X_i)$ and $t^*_{\widehat{G}_n}(X_i;\rho_n)$.

Lemma 4.1. *Let $\{X_i\}_{i=1}^n$ be independent Poisson random variables such that $X_i|\theta_i \sim Poisson(\theta_i)$ under $P_{n,\boldsymbol{\theta}}$ with a deterministic $\boldsymbol{\theta} \in [0,b]^n \subset \mathbb{R}^n$ for some unknown positive b that can have up to an order of $o\left(\frac{\log n}{\log\log n}\right)$. From the Proposition 4.1, we show that, for $i = 1, \cdots, n$, the probability of the event in which the proposed estimator $t^*_{\widehat{G}_n}(X_i)$ and its regularized version $t^*_{\widehat{G}_n}(X_i;\rho_n)$ being different is asymptotically 0 provided $\rho_n = \frac{q_n}{en}\frac{1}{\sqrt{2\pi\log n}}$, i.e.,*

$$P_{n,\boldsymbol{\theta}}\left(t^*_{\widehat{G}_n}(X_i) \neq t^*_{\widehat{G}_n}(X_i;\rho_n)\right) \quad \leq \quad n\left(\frac{eb}{\lfloor \log n/e\rfloor}\right)^{\lfloor \log n/e\rfloor} \to 0. \tag{4.2}$$

Proof. See Appendix. □

With a proper choice of ρ_n, we use the Lemma 4.1 in the proof of Theorem 3.1.

4.2. Properties of the regularized Bayes rule

In this subsection, we show several properties of the regularized Bayes rule, which play crucial roles in the proof of Theorem 3.1. First, we show the isopherimetric inequality of the Bayes rule of our Poisson setting. Meanwhile, in [14], they used the isopherimetric inequality for Normal distributions which is presented in [28]. For Normal case, [28] used the Wiener process to construct a semigroup operator. In a similar way, we use Poisson process to generate a semigroup operator to derive the following lemma.

Lemma 4.2. *Let $\boldsymbol{Z}$ be a n-dimensional random vector with the Poisson distributions where $Z_i|\theta_i \sim Poisson(\theta_i)$, $i = 1, \cdots, n$, under $P_{n,\boldsymbol{\theta}}$ with a deterministic $\boldsymbol{\theta} \in [0,b]^n \subset \mathbb{R}^n$ for some unknown $b > 0$. Then for every Lipschitz function $f : R^n \to R$ with $||f||_{Lip} \equiv \sup_{x\neq y}\frac{|f(x)-f(y)|}{|x-y|} \leq 1$, for some $\kappa_b = (3-e)b > 0$, we have*

$$P_{n,\boldsymbol{\theta}}(f(\boldsymbol{Z}) - E_{n,\boldsymbol{\theta}}f(\boldsymbol{Z}) > \lambda) \leq \exp\left(-\frac{\kappa_b\lambda^2}{2}\right). \tag{4.3}$$

Proof. See Appendix. □

The following lemma shows the bound of $|g_G(x)|/p_G(x)$ when θ is bounded by some constant. It will be shown that the bound of $|g_G(x)|/p_G(x)$ depends on the upper bound of θ as well as ρ while the similar bound for Normal case depends only on ρ. As mentioned, this is because the parameter of Poisson distribution also affects the scale of Poisson distribution while the parameter of Normal distribution in [14] determines only location.

Lemma 4.3. *If $G([0,b]) = 1$ for some unknown $b > 0$ then we have*

$$\frac{|g_G(x)|}{p_G(x)} \leq L(p_G(x);b), \tag{4.4}$$

where $L(\rho;b) = -\log\rho + b_1$ for $b_1 = b \vee 1$.

Proof. See Appendix. □

The following lemma shows the upper bound of the absolute value distance between the regularized Bayes estimator (3.14) and the MLE estimator, which is mainly used in the proof of Lemma 4.5 presented below.

Lemma 4.4. *If $G([0,b]) = 1$ for some unknown $b > 0$ and $0 < \rho < 1$, we have*

$$|x - t^*_G(x;\rho)| = \frac{|g_G(x)|}{p_G(x) \vee \rho} \leq L(\rho; b), \tag{4.5}$$

where $L(\rho; b) = -\log \rho + b_1$ for some $b_1 = b \vee 1$.

Proof. See Appendix. □

Based on Lemma 4.2, 4.3 and 4.4, we derive the following lemma regarding the isopherimetric inequality to be used for $E_{n,\boldsymbol{\theta}}\zeta^2_{3n}$'s upper bound where ζ_{3n} is defined in (3.27)

Lemma 4.5. *Suppose that $X_i|\theta_i \sim Poisson(\theta_i)$, $i = 1, \cdots, n$, under $P_{n,\boldsymbol{\theta}}$ with a deterministic $\boldsymbol{\theta} \in [0,b]^n \subset \mathbb{R}^n$ for some unknown $b > 0$ and let $t^*_G(x;\rho)$ be the regularized Bayes rule with G satisfying $G([0,b]) = 1$ for some constant $b > 0$, then for some $0 \leq \rho < 1$, we have*

$$P_{n,\boldsymbol{\theta}}(||t^*_G(\mathbf{X};\rho) - \theta|| \geq E_{n,\boldsymbol{\theta}}||t^*_G(\mathbf{X};\rho) - \theta|| + x) \leq \exp\left(-\frac{\kappa_b x^2}{2{A_{\rho,b}}^2}\right), \tag{4.6}$$

where $A_{\rho,b} = 1 + 2L(\rho; b)$ with $L(\rho; b)$ defined in Lemma 4.4 and κ_b in the Lemma 4.2.

Proof. See Appendix. □

Remark 4.2. Lemma 4.2-4.5 includes upper bounds, which depend on b, the maximum value of θ. [14] has similar results, however they do not depend on the upper bound of θ. In particular, unlike the case of the Normal model case in [14], the upper bound of exponential inequality (4.6) for the difference between the loss and risk of regularized Bayes rules depends on the maximum value b of the support of the mean parameter. Furthermore, as b increases, the upper bound of inequality (4.6) converges to 1. So for this upper bound to be meaningful, b should not be incremented in any order, and we need to constrain the order of b properly. Throughout the paper, we show that b can increase up to $o\left(\frac{\log n}{\log\log n}\right)$ order.

4.3. Upper bound of a regularized Bayes estimator discrepancy with a misspecified prior

Now we present the Theorem 4.1, which provides the upper bound of the regret using the regularized Bayes rule (3.14) due to the lack of the knowledge of the true G_0. Theorem 4.1 provides the bound for $E_{n,\boldsymbol{\theta}}\zeta^2_{4n}$ where ζ_{4n} is defined in (3.28). First, we define the following notation $||f||_g^2$ be the L^2 norm of f with respect to g, i.e.,

$$||f||_g^2 = \sum_{x=0}^{\infty} f(x)^2 g(x). \tag{4.7}$$

Theorem 4.1. *Suppose that (3.5) holds with G_0 where G_0 is any distribution that satisfies $G_0([0,b]) = 1$ for some unknown $b > 0$. Additionally, let $t_G^*(y;\rho)$ be the regularized Bayes rule as defined on (3.14), where G is any distribution with support $[0,M]$ and M is any constant greater than $b_1 = b \vee 1$. Then, we have the following results.*

1. *Let $D = 4(M+1)$ and $L(\rho; M) = -\log\rho + M$ then for $a > M+2$, we have*

$$\begin{aligned} &\{E_{G_0}(t_G^*(Y;\rho) - \theta)^2 - R^*(G_0)\}^{1/2} \\ \leq\ & 2\sqrt{D}\left(ad_H(p_G, p_{G_0}) + \sqrt{8M^2e^{-M}\frac{(eM)^{a-2}}{(a-2)^{a-2}}}\right) \\ & +2\sqrt{2}L(\rho;M)d_H(p_G,p_{G_0}) + ||(1-p_{G_0}/\rho)_+ g_{G_0}/p_{G_0}||_{p_{G_0}}. \end{aligned}$$

2. *Let $a = C(-\log d_H(p_G, p_{G_0}))$ with constant $C > 1$ satisfying $\frac{a-2}{e^3M} > 1$, and assume that $b = o\left(\frac{-\log\rho}{\log(-\log(\rho))}\right)$. Then for constant K defined in Lemma A.7.1 we have,*

$$\begin{aligned} &\{E_{G_0}(t_G^*(Y;\rho) - \theta)^2 - R^*(G_0)\}^{1/2} \\ \leq\ & \Big\{2\sqrt{D}(C(-\log d_H(p_G,p_{G_0})) + \sqrt{K}\,(d_H(p_G,p_{G_0}))^{C-1}) \\ & +2\sqrt{2}L(\rho;M)\Big\}\, d_H(p_G,p_{G_0}) + \rho^{1/2}b_1. \end{aligned} \tag{4.8}$$

Additionally, if $D \asymp -\log d_H(p_G, p_{G_0})$, then for ϵ_0 satisfying $d_H(p_G,p_{G_0}) \leq \epsilon_0 \leq e^{-3/2}$, $\rho \leq |\log\rho|^2\epsilon_0^2$ and some constant M_0, we have

$$\begin{aligned} & E_{G_0}(t_G^*(Y;\rho)-\theta)^2 - R^*(G_0) \\ \leq\ & M_0\left(|\log\epsilon_0|^3 + |\log\epsilon_0|\,\epsilon_0^{C-2} + |\log\epsilon_0|\,L(\rho;M)^2 + |\log\rho|^2(b_1)^2\right)\epsilon_0^2. \end{aligned} \tag{4.9}$$

Proof. See Appendix. □

4.4. Hellinger consistency of the NPMLE

Our main Theorem 3.1 uses a large deviation inequality for the Hellinger distance $d_H\left(p_{\widehat{G}_n}, p_{G_n}\right)$ at a certain convergence rate ϵ_n represented in Theorem 4.2. For the large deviation inequality, we need to derive the entropy bounds of the class $\mathcal{P} \equiv \left\{p_G : G \in \mathcal{G}_{[0,M]}\right\}$ where $\mathcal{G}_{[0,M]}$ is the class of all probability distributions supported on $[0, M]$. We present Lemma 4.6 $\sim$ 4.8 for the proof of Theorem 4.2 which will be used in the upper bound of $E_{n,\boldsymbol{\theta}}\zeta_{1n}^2$..

The following Lemma 4.6 implies that we can approximates a Poisson mixture $p_G(x) = \int p(x \mid \theta)dG(\theta)$ by $p_{G^*}(x)$ where G^* is a distribution function with a finite support.

Lemma 4.6. *Let $0 < \eta < 1$ be given, and B be an any positive integer. For any distribution G with support $[0, M]$ where $M = L \log \frac{1}{\eta}$ for some constant $L \geq 1$, there exist a discrete distribution G_m on $[0, M]$ with at most $m \leq \lceil 4.32M \rceil + B + 2$ support points such that*

$$\begin{aligned} ||p_G - p_{G_m}||_{\infty,B} &\leq C_1\eta, \\ ||g_G - g_{G_m}||_{\infty,B} &\leq C_1\eta(B \vee M), \end{aligned}$$

for some constant $C_1 > 0$.

Proof. See Appendix. □

Lemma 4.7 below follows Lemma 4.6 and provides a covering bound for the family of all convolution functions with distributions bounded on $[0, M]$ for some $M > 0$. This result will be used in the proof of Theorem 4.2. We first introduce the well-known definition in the empirical process field and then state Lemma 4.7.

Definition 4.1. For any metric d, the η-covering number $N(\eta, \mathcal{P}, d)$ is the minimal number of η-balls $\{g : d(g, p) \leq \eta; p \in \mathcal{P}\}$ of radius η needed to cover the set $\mathcal{P}$.

Lemma 4.7. *Let $\eta > 0$ be given and B be an any positive integer. Define $\mathcal{P} \equiv \left\{p_G : G \in \mathcal{G}_{[0,M]}\right\}$ where $\mathcal{G}_{[0,M]}$ is the class of all probability distributions supported on $[0, M]$ with $M = L \log \frac{1}{\eta}$ for some constant $L \geq 1$. Then we have $\log N(\eta, \mathcal{P}, || \cdot ||_{\infty,B}) \lesssim (\lceil 4.32M \rceil + B + 2) \log \left(\frac{1}{\eta}\right)$.*

Proof. See Appendix. □

As the last lemma for Theorem 4.1, we provide the following Lemma 4.8 which also plays an important role in the proof of Theorem 4.2.

Lemma 4.8. *Let $X_i|\theta_i \sim Poisson(\theta_i)$, $i = 1, \dots, n$, under $P_{n,\boldsymbol{\theta}}$ with a deterministic $\boldsymbol{\theta} \in [0, b]^n \subset \mathbb{R}^n$ for some unknown $b > 0$. Then, for any positive integer B satisfying $B - 1 > b_1 = \max(1, b)$, $0 < \lambda \leq 1$ and $a > 0$,*

$$E_{n,\boldsymbol{\theta}} \left\{ \prod_{i=1}^{n} (aX_i)^{I\{X_i \geq B\}} \right\}^{\lambda} \leq \exp\left[a^{\lambda} n B^{\lambda-1} b^B e^{-b} \left(\frac{e}{B-1}\right)^{B-1} \right]. \quad (4.10)$$

Proof. See Appendix. □

We now present the following theorem for our Poisson case corresponding to Theorem 1 in [34] for Normal case. This shows the convergence rate of the hellinger distance between the NPMLE solution $\widehat{G}_n$ and the empirical distribution G_n as defined on (3.4). We consider any approximate the NPMLE satisfying

$$L_n(p_{\widehat{G}_n}, p_{G_n}) = \prod_{i=1}^{n} \frac{p_{\widehat{G}_n}(X_i)}{p_{G_n}(X_i)} \geq e^{-2t^2 n\epsilon_n^2/15}. \tag{4.11}$$

While [14] showed that the $\widehat{G}_n$ from the EM algorithm satisfies a similar condition to (4.11), [34] and [13] assume that any solution $\widehat{G}_n$ satisfying a similar condition to (4.11) can be obtained from the EM algorithm or the `REBayes` package. We also assume that well known methods such as the EM algorithm and the `REBayes` pacakge provide the solution $\widehat{G}_n$ satisfying (4.11) as in [15, 13, 26]. Then the following theorem provides the large deviation inequality for $d_H(p_{\widehat{G}_n}, p_{G_n})$. Using this result, we can ignore the event where $d_H(p_{\widehat{G}_n}, p_{G_n})$ deviates from zero and then we can derive the upper bound of $E_{n,\boldsymbol{\theta}}\zeta_{1n}^2$ where ζ_{1n}^2 is defined in (3.25).

Theorem 4.2. *Let Let $X_i|\theta_i \sim Poisson(\theta_i)$, $i = 1, \dots, n$, under $P_{n,\boldsymbol{\theta}}$ with a deterministic $\boldsymbol{\theta} \in [0, b]^n \subset \mathbb{R}^n$ for some unknown $b > 0$. Additionally, let G_n be an empirical distribution of θ_i s defined on (3.4) with satisfying $G_n([0, b]) = 1$. Then for $\widehat{G}_n$ is the NPMLE satisfying (4.11) with $\widehat{G}_n([0, M]) = 1$, there exits some universal constants t_* such that for all $t \geq t_*$,*

$$P_{n,\boldsymbol{\theta}}\{d_H(p_{\widehat{G}_n}, p_{G_n}) \geq t\epsilon_n\} \leq C_0 \exp\left(-\frac{t^2 n\epsilon_n^2}{2\log n}\right), \tag{4.12}$$

where C_0 is universal constant and $\epsilon_n = \frac{(\log n)^{3/2}}{\sqrt{n}}$.

Proof. See Appendix. □

4.5. Entropy bound of the regularized Bayes rule

Now we provide an entropy bound for collections of regularized Bayes rules used to control $E_{n,\boldsymbol{\theta}}\zeta_{2n}^2$ and $E_{n,\boldsymbol{\theta}}\zeta_{3n}^2$ where ζ_{2n} and ζ_{3n} are defined in equations (3.26) and (3.27). The challenge is that the regularized version $t^*_{\widehat{G}_n}(X_i; \rho_n)$ is not separable since the regularized version $t^*_{\widehat{G}_n}(X_i; \rho_n)$ in (3.21) with $\rho = \rho_n$ includes the NPMLE $\widehat{G}_n$ which depends on all observed data $X_1, \cdots, X_n$. We choose a class of regularized Bayes rule with a nonrandom collection, say $\{H_j\}_{j=1}^N$, which has controllable size N. We can approximate $t^*_{\widehat{G}_n}(X_i; \rho_n)$ by $t^*_{H_j}(X_i; \rho_n)$ for some j. Our solution needs an entropy bound for this class of regularized Bayes rule for Poisson distribution as [14], [26] and [13] did for Normal distribution. The following theorem gives an entropy bound for class of all regularized Bayes rules defined in (3.11).

Theorem 4.3. *For any $\rho > 0$ and $0 < \eta \leq \rho$, define $\mathcal{T}_\rho = \left\{t^*_G(\cdot;\rho) : G \in \mathcal{G}_{[0,M]}\right\}$ where $\mathcal{G}_{[0,M]}$ is the class of all probability distributions supported on $[0, M]$ with $M = L \log \frac{1}{\eta}$ for some constant $L \geq 1$. Then, for the integer $B = \lfloor M \rfloor$ the following holds,*

$$\log N(\eta^*, \mathcal{T}_\rho, \|\cdot\|_{\infty,B}) \lesssim \left\{\log \frac{1}{\eta}\right\}^2, \tag{4.13}$$

where $\eta^* = D^* \frac{\eta}{\rho}\left(\log \frac{1}{\eta} + L(\rho; M)\right)$ for some constant $D^* > 0$ and $L(\rho; M) = -\log \rho + M$.

Proof. See Appendix. □

5. Simulation studies

In this section, we present numerical studies comparing our proposed estimator g-NPEB and the estimators in [20] and [3]. [20] considered various simulation settings to compared their NPEB estimator with linear shrinkage estimators such as [12] and [21]. Numerical studies in [20] show that their NPEB outperforms linear shrinkage estimators, so we do not consider those linear shrinkage estimators in our simulations.

We use the following settings which were used in [20].

- **Simulation 1.** $G_0 \sim Gamma(10, 2)$
- **Simulation 2.** $G_0 \sim Gamma(10, 5)$
- **Simulation 3.** $G_0 \sim U(5, 10)$
- **Simulation 4.** $G_0 \sim U(1, 5)$
- **Simulation 5.** $G_0 \sim 0.8\delta(0.5) + 0.2Gamma(10, 2)$
- **Simulation 6.** $G_0 \sim 0.8\delta(0.5) + 0.2Gamma(10, 5)$
- **Simulation 7.** $G_0 \sim 0.8\delta(0.5) + 0.2U(5, 10)$
- **Simulation 8.** $G_0 \sim 0.8\delta(0.5) + 0.2U(1, 5)$

where $\delta(\cdot)$ is the Dirac-measure, $Gamma(\alpha, \beta)$ is a gamma distribution with probability density $f_G(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} I(x \geq 0)$ and $U(a, b)$ is a uniform distribution with probability density $f_U(x; a, b) = \frac{1}{b-a} I(a \leq x \leq b)$. Simulation 1-4 represent the situations that θ_i s form unimodal distributions while Simulation 5-8 bimodal distributions, which are mixture of point mass and some unimodal distribution. In particular, the latter four simulation settings may be useful in estimation when there are many small counts in the data, for example, protein domain data which will be demonstrated in our real data examples. The estimators of proposed method, [20] and [3] are referred to here as "g-NPEB", "Park" and "Brown" respectively. In addition, "Bayes" represents the oracle error rate which is theoretically the optimal error rate. Our numerical studies show that the proposed g-NPEB based on the g-modeling is efficient in detecting such multimodality of θ values compared to [3] and [20] based on f-modeling. We set the number of Poisson data observed in each simulation from $n = 200$ to 2000. For each n, we set the number of repetitions to 100.

Table 1
Result table of the Simulation $1 \sim 4$ which assume the true parameter θ_i s distributed from some unimodal distributions.

	Simulation 1.									
n	200	400	600	800	1000	1200	1400	1600	1800	2000
g-NPEB	1.87	1.78	1.74	1.72	1.73	1.73	1.71	1.70	1.70	1.70
Brown	1.85	1.79	1.76	1.74	1.75	1.75	1.73	1.73	1.72	1.73
Park	4.73	3.94	3.28	2.93	2.91	2.80	2.71	2.56	2.47	2.43
Bayes	1.69	1.69	1.69	1.69	1.69	1.69	1.69	1.69	1.69	1.69
	Simulation 2.									
n	200	400	600	800	1000	1200	1400	1600	1800	2000
g-NPEB	0.41	0.37	0.36	0.35	0.35	0.35	0.35	0.34	0.34	0.34
Brown	0.38	0.36	0.36	0.35	0.35	0.35	0.35	0.35	0.35	0.35
Park	0.99	0.78	0.66	0.63	0.61	0.58	0.57	0.55	0.53	0.54
Bayes	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34
	Simulation 3.									
n	200	400	600	800	1000	1200	1400	1600	1800	2000
g-NPEB	1.86	1.75	1.74	1.71	1.68	1.69	1.66	1.66	1.67	1.67
Brown	1.83	1.76	1.74	1.73	1.70	1.71	1.68	1.69	1.68	1.69
Park	8.78	6.14	5.27	4.68	4.15	3.98	3.64	3.35	3.30	3.23
Bayes	1.62	1.62	1.62	1.62	1.62	1.62	1.62	1.62	1.62	1.62
	Simulation 4.									
n	200	400	600	800	1000	1200	1400	1600	1800	2000
g-NPEB	1.00	0.96	0.95	0.94	0.93	0.93	0.92	0.92	0.92	0.92
Brown	1.00	0.97	0.98	0.97	0.96	0.96	0.95	0.96	0.95	0.96
Park	2.25	1.75	1.70	1.54	1.45	1.44	1.36	1.35	1.30	1.29
Bayes	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9

Table 1 and 2 show the results for Simulation 1-8. As displayed, we see that the proposed g-NPEB outperforms the other two methods. Furthermore, the risks from the proposed g-NPEB converge to the Bayes error rate faster than the other two f-modelling based estimators. Regarding computational efficiency, the proposed g-NPEB and the nonparametric empirical Bayes method in [20] are quite fast in computation. However the method proposed by [3] is computationally inefficient to be derived. Considering the risks and computational efficiency, the proposed g-NPEB based on g-modeling has superiority over two other methods in [20] and [3].

6. Real data examples

To study cancer at the molecular level, it is important to understand which somatic mutations are involved in tumor initiation or progression. [22] found that known cancer mutations tend to cluster more in specific locations than those associated with unrelated diseases. This finding implies that understanding the molecular mechanisms related to cancer may be achieved from protein domain hotspots. Protein domains are regarded as the structural and functional units of proteins, and it has huge potential to exhibit tumor variants. For our case study, we analyze the mutation data of 5,848 patients from The Cancer Genome Atlas (TCGA) `http://tcga-data.nci.nih.gov/tcga/`. These data were mapped to specific positions within protein domain models to identify

TABLE 2
Result table of the Simulation $5 \sim 8$ which assume the true parameter θ_i s distributed from some mixture of point mass and some unimodal distribution.

	Simulation 5.									
n	200	400	600	800	1000	1200	1400	1600	1800	2000
g-NPEB	0.92	0.86	0.86	0.84	0.84	0.84	0.82	0.82	0.82	0.83
Brown	1.06	1.00	0.98	0.96	0.97	0.96	0.91	0.90	0.93	0.91
Park	1.27	1.24	1.26	1.24	1.21	1.24	1.21	1.20	1.18	1.16
Bayes	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
	Simulation 6.									
n	200	400	600	800	1000	1200	1400	1600	1800	2000
g-NPEB	0.32	0.30	0.29	0.28	0.28	0.28	0.28	0.28	0.28	0.28
Brown	0.34	0.33	0.33	0.32	0.32	0.31	0.32	0.32	0.31	0.31
Park	0.48	0.44	0.42	0.41	0.40	0.40	0.39	0.39	0.38	0.38
Bayes	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.27
	Simulation 7.									
n	200	400	600	800	1000	1200	1400	1600	1800	2000
g-NPEB	0.94	0.85	0.84	0.82	0.82	0.82	0.80	0.80	0.79	0.79
Brown	1.20	1.04	1.02	1.00	1.00	0.98	0.95	0.96	0.91	0.92
Park	1.77	1.71	1.62	1.67	1.56	1.55	1.57	1.55	1.50	1.52
Bayes	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72
	Simulation 8.									
n	200	400	600	800	1000	1200	1400	1600	1800	2000
g-NPEB	0.56	0.53	0.52	0.52	0.51	0.51	0.51	0.51	0.51	0.51
Brown	0.65	0.62	0.59	0.62	0.59	0.59	0.59	0.58	0.59	0.57
Park	0.78	0.76	0.75	0.74	0.72	0.71	0.70	0.70	0.69	0.69
Bayes	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

clusters. TCGA MAF files were obtained on July 7th, 2014 for 20 cancer types. Each protein domain consists of positions that have mutation counts. We analyze five protein domains that can be obtained from the supplementary material in [10] (https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Fbiom.12779&file=biom12779-sup-0003-SuppData.txt).

To give a brief explanation on the 5 protein domains we analyze, there are growth factors (cd00031), the calcium-binding domain of epidermal growth factors (cd00054), protein kinases (cd00180), ankyrin domains (cd00204), which play a role in mediating protein-protein interactions and finally RAS-Like GTPase family of genes (cd00882) are well-known for their role in regulating pathways. For the detailed description, see [10]. Figure 1 provides the histogram of mutations of each protein domain and its number of position n respectively. One typical phenomenon in those protein domain data in Figure 1 is that we see mixtures of many zero or small counts and large valued counts. Such small counts are considered background noise while large values represent mutation counts at hot spots related to some disease. When the number of mutations in each protein data is assumed to follows the Poisson model, we conjecture that the corresponding true mean values have bi-modality or multi modality which are considered in Simulation 5-8 in section 5. In this case, we assume that the mutation counts in each position follow the Poisson model independently and apply our proposed g-NPEB to estimate the mean vector of the Poisson model. We base our protein domain data analysis on two main assumptions.

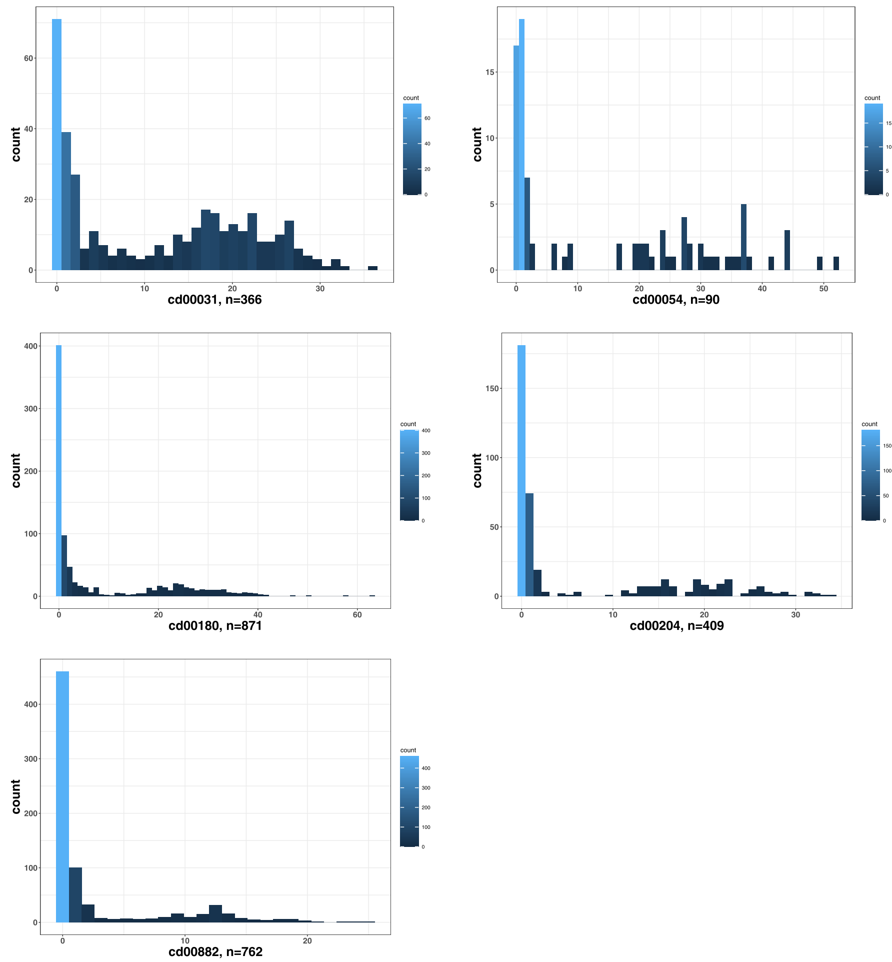

Fig 1. *Histograms of protein domain data for cd00031, cd00054, cd00180, cd00204 and cd00882 with total number of positions n.*

The first assumption is that each mutation count follows a Poisson distribution with different mean parameters, and the second is that the mutation counts are independent of each other. First of all, we would like to emphasize that the mutation counts observed in the protein domain have been studied recently, and researches that apply the Poisson or negative binomial model to count data with few prior biological knowledge are quite common [1, 4, 10]. In particular, having analyzed protein domains identical to ours, [10] applied the Zero-inflated Generalized Poisson model under the assumption that most mutation counts are independently and identically generated, but this model does not meet our goal of estimating heterogeneous mean values for different positions. Second, in a practical application of the mean vector estimation, many studies assume the independence of data without its validation. For example, in Brown's batting average in baseball data [2], many studies have been conducted under the rule of independence between players, even though it is not clear whether the performances of all players are independent of each other [2, 31, 9, 29]. If correlations among mutation counts are considered, mutation counts should be viewed as multivariate data and their mean vector can be estimated when there is many replication of data that is not currently available. If independence and Poisson distribution are seriously violated, then estimation of the mean vector is of course inaccurate, however estimation of mean vector under these assumptions of independence and Poisson distribution for count data is currently most common due to limited data and lack of methodology for multivariate data with correlations.

For comparison, we compare the f-modelling methods considered in the previous simulation study with our g-NPEB. Evaluating estimators in mean vector estimation is not feasible in general since we do not know the true mean vector in real data. However, by borrowing the insight from [3], we define the following procedure to evaluate estimators in real data:

Step 1 : Prefix the proportion parameter $p \in (0, 1)$.

Step 2 : Generate random variable $U_i|Y_i \sim Binomial(Y_i, p)$ where Y_i is the number of mutations in the protein domain data and assumed to be distributed from the $Poisson(\theta_i)$, $i = 1, \cdots, n$.

Step 3 : Define $V_i \equiv Y_i - U_i$, $i = 1, \cdots, n$.

Step 4 : Calculate $\widehat{R}_n$ defined as

$$\widehat{R}_n = \frac{1}{n}\sum_{i=1}^{n}\left(\Delta(U_i) - \frac{p}{1-p}V_i\right)^2. \tag{6.1}$$

Step 5 : Repeat 2∼4, T times and average $\widehat{R}_n$ s.

Remark 6.1. If we generate U_i and V_i as above then, marginally,

$$U_i \sim Poisson(p\theta_i) \text{ and } V_i \sim Poisson((1-p)\theta_i),$$

and they are independent given θ_i, $i = 1, \cdots, n$. Then we consider $\frac{p}{1-p}V_i$ with the mean as $p\theta_i$ to be true parameter and we estimate $p\theta_i$ as a function of U_i, $\Delta(U_i)$.

Remark 6.2. The intuition of using $\hat{R}_n$ as a criterion of the performance of $\Delta(\cdot)$ is as follows. Define the following average risks:

$$R_n = \frac{1}{n}\sum_{i=1}^{n} E_{\theta_i}\left(\Delta(U_i) - \frac{p}{1-p}V_i\right)^2, \tag{6.2}$$

$$R^*_{1,n} = \frac{1}{n}\sum_{i=1}^{n} E_{\theta_i}\left(\Delta(U_i) - p\theta_i\right)^2, \tag{6.3}$$

Table 3
Mean vector estimation results for the Protein domain data.

	$p = 0.5$		
Protein domain	g-NPEB	Brown	Park
cd00031	8.090	8.269	12.787
cd00054	12.506	12.561	18.620
cd00180	6.418	6.554	8.875
cd00204	4.457	4.726	6.916
cd00882	2.049	2.174	2.642
	$p = 0.7$		
Protein domain	g-NPEB	Brown	Park
cd00031	21.932	22.071	30.331
cd00054	31.297	31.671	40.656
cd00180	16.724	16.912	20.496
cd00204	12.533	12.692	16.329
cd00882	5.386	5.524	6.309
	$p = 0.9$		
Protein domain	g-NPEB	Brown	Park
cd00031	96.237	96.573	108.752
cd00054	126.660	128.195	139.214
cd00180	70.719	71.037	76.143
cd00204	55.137	55.476	60.660
cd00882	23.462	23.576	24.715

$$R^*_{2,n} \quad = \quad \frac{1}{n}\sum_{i=1}^{n} E_{\theta_i}\left(p\theta_i - \frac{p}{1-p}V_i\right)^2, \tag{6.4}$$

where E_{θ_i} represents the expectation given θ_i. Then, it can be shown easily that $R_n = R^*_{1,n} + R^*_{2,n}$. Note that $R^*_{2,n}$ is unrelated to the mean estimator $\Delta(U_i)$ s. Therefore, a small value of R_n implies a small value of $R^*_{1,n}$. From this result, we quantify the performance of $\Delta(U_i)$ s using the $\widehat{R}_n$ in (6.1).

Remark 6.3. There are mainly two reasons that we have an independence assumption in our protein domain analysis. First, the research on the protein domain we used has been relatively recently studied, as demonstrated in [10]. So there doesn't seem to be enough information such as the location or spatial structure of the positions in protein domain. For a similar reason, [10] assumed the independence of mutations counts for different spots of the protein domain. Second, in order to consider the correlation structure of the mean parameters in mean vector estimation, we need to replicate the data corresponding to each parameter. However, since there is no replication for the data given to us, the correlation structure cannot be considered.

Table 3 represents the average of $\widehat{R}_n$ s with repeat number $T = 500$ of our "g-NPEB", "Brown" in [3] and "Park" in [20] with proportion p =0.5, 0.7 and 0.9. In Table 3, we can see that for all protein domains and all proportions p, as with the simulation study results, our "g-NPEB" and "Brown" achieve almost similar risks, however our "g-NPEB" performs better consistently than "Brown"s method.

7. Concluding remarks

In this paper, we presented theoretical results of the estimated Bayes rule (g-NPEB) using the NPMLE, which has not been addressed. [3] stated that an advantage of their methodology based on f-modeling is that it does not require complex optimization and can derive an estimator with an elementary procedure. However, due to the development of many built-in functions of deconvolution and demixing, such as REBayes in R-package implementing [18], we believe that the g-modeling is implemented efficiently without any difficulty. Although many empirical Bayes approaches have been mainly concerned with f-modeling in estimating the mean vector of the Poisson model (e.g., [20] and [3]), our numerical studies show that each of them shows some drawbacks: [20] has slow convergence to the Bayes risk implying poor performance in finite samples, and [3] needs some significant computation time to implement their method since it requires cross-validation to select the smoothing parameter and tunning step to impose monotonicity on the decision function. On the other hand, we show that the g-modeling based estimator is as efficient as [20] in computations and achieves more accurate risk than [20] and [3].

We believe that such g-modeling based methods have more advantages in both accuracy and computational efficiency than f-modeling due to many built-in functions such as REBayes in R-package. In particular, the g-modeling based estimators for the Poisson mean vector can be potentially used in statistical inference for count data examples from genetics and bioinformatics. Especially, we assumed that mutation counts were independent of each other in protein domain analysis for several practical reasons. However, considering that protein domain research has been conducted relatively recently, the authors believe that additional data will be given in the future to reveal the biological knowledge of the mutation counts and consider the correlation of the mutation counts. Then, it is of great interest to estimate the intensity rates of mutation counts considering spatial correlations along with the development of a new estimator. We leave this as a future research.

Appendix A: Proofs

A.1. Proof of Theorem 3.1

Let $L(\rho; M) = \log(e^M/\rho) = -\log\rho + M$ and take the following

$$\rho_n = \frac{1}{n^3}\frac{1}{\sqrt{2\pi\log n}}, \quad \eta = \frac{\rho_n}{n},$$

and

$$\eta^* = D^*\frac{\eta}{\rho_n}\left(\log\frac{1}{\eta} + L(\rho_n; M)\right), \quad M = L\log\frac{1}{\eta}$$

where $L \geq 1$ and D^* are some universal constants.

Note that (4.2) holds with $\rho_n = \frac{1}{n^3}\frac{1}{\sqrt{2\pi \log n}}$ and $q_n = \frac{e}{n^2}$. Let $\{t^*_{H_j}(\cdot;\rho_n),\ j \le N\}$ be a $2\eta^*$ net of $\mathcal{T}_{\rho_n} \cap \{t^*_G : d(p_G, p_{G_n}) \le t^*\varepsilon_n\}$ under the seminorm $\|\cdot\|_{\infty,M}$ and $N = \log N(\eta^*, \mathcal{T}_{\rho_n}, \|\cdot\|_{\infty,M})$. We adopt the preliminary calculation for $\tilde{r}_{n,\boldsymbol{\theta}}((t^*_{\widehat{G}_n}(\mathbf{X}))$ from [14] such that

$$\sqrt{E_{n,\boldsymbol{\theta}} n L_n(t^*_{\widehat{G}_n}(\mathbf{X}), \boldsymbol{\theta})} \le \sqrt{nR^*(G_n)} + \sqrt{E_{n,\boldsymbol{\theta}}\left(\sum_{j=0}^{4}|\zeta_{jn}|\right)^2}, \tag{A.1}$$

where

$$\zeta_{0n} = ||t_{\widehat{G}_n}(\mathbf{X}) - t_{\widehat{G}_n}(\mathbf{X};\rho_n)||, \tag{A.2}$$

$$\zeta_{1n} = ||t_{\widehat{G}_n}(\mathbf{X};\rho_n) - \boldsymbol{\theta}|| I_{A_n^c}, \tag{A.3}$$

$$\zeta_{2n} = \left\{||t^*_{\widehat{G}_n}(\mathbf{X};\rho_n) - \boldsymbol{\theta}|| I_{A_n} - \max_{j\le N}||t^*_{H_j}(\mathbf{X};\rho_n) - \boldsymbol{\theta}||\right\}_+, \tag{A.4}$$

$$\zeta_{3n} = \max_{j\le N}\{||t^*_{H_j}(\mathbf{X};\rho_n) - \boldsymbol{\theta}|| - E_{n,\boldsymbol{\theta}}||t^*_{H_j}(\mathbf{X};\rho_n) - \boldsymbol{\theta}||\}_+, \tag{A.5}$$

$$\zeta_{4n} = \max_{j\le N}\sqrt{E_{n,\boldsymbol{\theta}}||t^*_{H_j}(\mathbf{X};\rho_n) - \boldsymbol{\theta}||^2} - \sqrt{nR^*(G_n)}, \tag{A.6}$$

for $A_n = \{d_H(p_{\widehat{G}_n}, p_{G_n}) \le t^*\epsilon_n\}$, $t^* = t_* \vee 1$. Note that $P_{n,\boldsymbol{\theta}}(A_n^c) \le \frac{1}{n}$ from Theorem 4.2 for $t^* = t_* \vee 1$. Since $E_{n,\boldsymbol{\theta}}\left(\sum_{j=0}^4 |\zeta_{jn}|\right)^2 \le 5\sum_{j=0}^4 E_{n,\boldsymbol{\theta}}\zeta_{jn}^2$ in (A.1) using $(\sum_{j=0}^4 a_j)^2 \le 5\sum_{j=0}^4 a_j^2$ for any real values of a_j s, we compute the bound of $E_{n,\boldsymbol{\theta}}\zeta_{jn}^2$ for $j = 0, 1, \ldots, 4$ as follows:

0. Note that $t_{\widehat{G}_n}(\mathbf{X};\rho_n) - t_{\widehat{G}_n}(\mathbf{X}) \ne 0$ only when $p_{\widehat{G}_n}(X_i) < \rho_n$. Define $I_n = I(p_{\widehat{G}_n}(X_i) < \rho_n)$ then,

$$\begin{aligned}
E_{n,\boldsymbol{\theta}}\zeta_{0n}^2 &\le \sum_{i=1}^n E_{n,\boldsymbol{\theta}}\left\{\left(t_{\widehat{G}_n}(X_i) - t_{\widehat{G}_n}(X_i;\rho_n)\right)^2 I_n\right\} \\
&= \sum_{i=1}^n E_{n,\boldsymbol{\theta}}\left\{\left(\frac{g_{\widehat{G}_n}(X_i)}{p_{\widehat{G}_n}(X_i)}\right)^2\left(1 - \frac{p_{\widehat{G}_n}(X_i)}{p_{\widehat{G}_n}(X_i)\vee\rho_n}\right)^2 I_n\right\} \\
&\le \sum_{i=1}^n E_{n,\boldsymbol{\theta}}\left\{\left(\frac{g_{\widehat{G}_n}(X_i)}{p_{\widehat{G}_n}(X_i)}\right)^2 I_n\right\} \\
&= \sum_{i=1}^n E_{n,\boldsymbol{\theta}}\left\{\left(\frac{\int(\theta - X_i)p(X_i|\theta)d\widehat{G}_n(\theta)}{\int p(X_i|\theta)d\widehat{G}_n(\theta)}\right)^2 I_n\right\} \\
&\le \sum_{i=1}^n E_{n,\boldsymbol{\theta}}\left\{(M + X_i)^2 I_n\right\}, \text{ provided } \widehat{G}_n([0,M]) = 1 \\
&\le \sum_{i=1}^n \left\{E_{n,\boldsymbol{\theta}}(M + X_i)^4\right\}^{1/2}\left\{P_{n,\boldsymbol{\theta}}(p_{\widehat{G}_n}(X_i) < \rho_n)\right\}^{1/2}
\end{aligned}$$

$$\leq \sum_{i=1}^{n}\left\{E_{n,\boldsymbol{\theta}}\left(M+X_i\right)^4\right\}^{1/2}\left\{n\left(\frac{eb}{\lfloor\log n/e\rfloor}\right)^{\lfloor\log n/e\rfloor}\right\}^{1/2}$$
$$\leq C_{0,0}(M)^2$$
$$\leq C_{0,0}n\epsilon_n^2. \tag{A.7}$$

Note that the third inequality from the bottom is derived from (4.2), and in the second to last inequality, we take advantage of the property that the second curly brace term vanishes to zero faster than any polynomial n. In addition, $C_{0,0}$ is universal constant and we use the fact that $E_{n,\boldsymbol{\theta}}\left(M+X_i\right)^4 \lesssim M^4$ provided $\forall i,\ \theta_i \leq b \leq M = O(\log n)$.

1. We rearrange $t_{\widehat{G}_n}(\mathbf{X};\rho_n)-\boldsymbol{\theta}$ as $t_{\widehat{G}_n}(\mathbf{X};\rho_n)-\mathbf{X}-(\boldsymbol{\theta}-\mathbf{X})$ in ζ_{1n}. Then using Lemma 4.4 and $L(\rho_n;M)\leq C_0(\log n)$ for $\rho_n=\frac{1}{n^3}\frac{1}{\sqrt{2\pi\log n}}$, we have

$$\begin{aligned}
E_{n,\boldsymbol{\theta}}\zeta_{1n}^2 &\leq 2nL(\rho_n;M)^2P_{n,\boldsymbol{\theta}}(A_n^c)+2\sum_{i=1}^{n}E_{n,\boldsymbol{\theta}}(X_i-\theta_i)^2I_{A_n^c}\\
&\leq 2nL(\rho_n;M)^2P_{n,\boldsymbol{\theta}}(A_n^c)\\
&\quad +2\sum_{i=1}^{n}\sum_{x\geq 0}\min\{P_{n,\boldsymbol{\theta}}((X_i-\theta_i)^2\geq x),1/n\}\\
&\leq C_0^2(\log n)^2\\
&\quad +2\sum_{i=1}^{n}\left(\sum_{x\leq(\log n)^2}\frac{1}{n}+\sum_{x>(\log n)^2}P_{n,\boldsymbol{\theta}}((X_i-\theta_i)^2\geq x)\right).
\end{aligned}$$

Note that as $n\to\infty$, if $x>(\log n)^2$, we have $\theta_i-\sqrt{x}<0$ asymptotically. Therefore, $P_{n,\boldsymbol{\theta}}((X_i-\theta_i)^2\geq x)=P_{n,\boldsymbol{\theta}}(X_i>\theta_i+\sqrt{x})\leq \sum_{y>\theta_i+\sqrt{x}}\frac{e^{-\theta_i}\theta_i^y}{y!}\leq\frac{e^{-\theta_i}(e\theta_i)^{\sqrt{x}}}{\sqrt{x}^{\sqrt{x}}}\leq\left(\frac{eb}{\sqrt{x}}\right)^{\sqrt{x}}$. For $\log n\geq e^2b$, we have $eb/\log n\leq e^{-1}$, so $\left(\frac{eb}{\sqrt{x}}\right)^{\sqrt{x}}\leq e^{-\sqrt{x}}\leq\frac{1}{n}$ when $x>(\log n)^2$. Using this, we obtain $\sum_{x\geq(\log n)^2}P_{n,\boldsymbol{\theta}}((X_i-\theta_i)^2\geq x)\leq\sum_{x\geq(\log n)^2}\exp(-\sqrt{x})$ for $1\leq i\leq n$. Therefore, by using change of variable $\sqrt{x}=t$, we have $\sum_{x\geq(\log n)^2}e^{-\sqrt{x}}\leq C\int_{(\log n)^2}e^{-\sqrt{x}}dx=-2C_1(t+1)e^{-t}\Big|_{\log n}^{\infty}=2C_1(\frac{\log n}{n}+\frac{1}{n})$ for some constant C_1. Finally, for some constants C_0,C_1 and $C_{0,1}$, we have

$$\begin{aligned}
E_{n,\boldsymbol{\theta}}\zeta_{1n}^2 &\leq C_0^2(\log n)^2+2(\log n)^2+2C_1(\log n+1)\leq C_{0,1}(\log n)^2\\
&\leq C_{0,1}n\epsilon_n^2,
\end{aligned} \tag{A.8}$$

from $n\epsilon_n^2=(\log n)^3$.

2. For ζ_{2n}, $\left\{t_{H_j}^*\left(\cdot;\rho_n\right),j\leq N\right\}$ is a $2\eta^*$ net of

$$\mathcal{T}_{\rho_n}\cap\{t_G^*:d_H\left(p_G,p_{G_n}\right)\leq x^*\varepsilon_n\}$$

under the seminorm $\|\cdot\|_{\infty,M}$ and by Lemma 4.4, we know that

$$|t_G^*(x;\rho_n) - x| \le L_1(\rho_n; M)$$

for $G = \widehat{G}_n$ or H_j. Let $N = \log N(\eta^*, \mathcal{T}_{\rho_n}, \|\cdot\|_{\infty,M})$, then we obtain $N \lesssim \left\{\log\frac{1}{\eta}\right\}^2$ from Theorem 4.3. Therefore we have

$$\begin{aligned}\zeta_{2n}^2 &\le \min_{j\le N} ||t_{\widehat{G}_n}(X;\rho_n) - t_{H_j}(X;\rho_n)||^2 I_{A_n} \\ &\le (2\eta^*)^2 \#\{i : X_i < M\} + 4L(\rho_n;M)^2 \sum_{i=1}^n I(X_i \ge M) \quad \text{(A.9)}\end{aligned}$$

leading to

$$\begin{aligned}E_{n,\boldsymbol{\theta}}\zeta_{2n}^2 &\le n(2\eta^*)^2 + 4L(\rho_n;M)^2 \sum_{i=1}^n E_{n,\boldsymbol{\theta}} I(X_i \ge M) \\ &= n(2\eta^*)^2 + 4L(\rho_n;M)^2 \sum_{i=1}^n P_{n,\boldsymbol{\theta}}(X_i \ge M) \\ &\le n(2\eta^*)^2 + 4L(\rho_n;M)^2 \sum_{i=1}^n \sum_{x\ge M} \frac{e^{-\theta_i}\theta_i^x}{x!}.\end{aligned}$$

Note that, $\sum_{x\ge M}\frac{e^{-\theta_i}\theta_i^x}{x!} \le \sum_{x\ge M}\frac{e^{-b}b^x}{x!} \le e^{-b}\left(\frac{eb}{M}\right)^M$ for $M > b$. Additionally, $\eta^* = D^*\frac{\eta}{\rho_n}\left(\log\frac{1}{\eta} + L(\rho_n;M)\right) \le C_2\frac{(\log n)}{n}$ and $M = O(\log n)$. Furthermore, $M = O(\log n)$ implies $\left(\frac{eb}{M}\right)^M \le \frac{1}{n^2}$ for sufficiently large n and $L(\rho_n;M) = O(\log n)$. Hence with some constant C_2 and $C_{0,2}$, we have the upper bound of $E_{n,\boldsymbol{\theta}}\zeta_{2n}^2$ as follows:

$$\begin{aligned}E_{n,\boldsymbol{\theta}}\zeta_{2n}^2 &\le n(2\eta^*)^2 + 4L(\rho_n;M)^2 n e^{-b}\left(\frac{eb}{M}\right)^M \\ &\le C_2\left(n\frac{(\log n)^2}{n^2} + (\log n)^2 n\frac{1}{n^2}\right) \\ &\le C_{0,2}\epsilon_n^2. \quad \text{(A.10)}\end{aligned}$$

3. For ζ_{3n}, we use Lemma 4.5 and Theorem 4.3 and $A_{\rho_n,M} = 1 + 2L_1(\rho_n;M) = O(\log n)$ which lead to

$$\begin{aligned}E_{n,\boldsymbol{\theta}}\zeta_{3n}^2 &\le \sum_{x=0}^{\infty} P_{n,\boldsymbol{\theta}}(\zeta_{3n} > \sqrt{x}) \\ &\le \int_0^{\infty} \min\left\{1, N\exp\left(-\kappa_b x/(2A_{\rho_n,M}^2)\right)\right\} dx, \quad \kappa_b = (3-e)b \\ &= \sum_{x\le(2A_{\rho_n,M}^2\log N)/\kappa_b} 1\end{aligned}$$

$$
\begin{aligned}
& + \sum_{x > (2A^2_{\rho_n,M}\log N)/\kappa_b} N\exp(-\kappa_b x/(2A^2_{\rho_n,M})) \\
\leq\ & (2A^2_{\rho_n,M}\log N)/\kappa_b \\
& -2A^2_{\rho_n,M} N/\kappa_b \exp\left(-\kappa_b x/(2A^2_{\rho_n,M})\right)|^{\infty}_{(2A^2_{\rho_n,M}\log N)/\kappa_b} \\
\leq\ & (2A^2_{\rho_n,M}\log N)/\kappa_b \\
\lesssim\ & 2A^2_{\rho_n,M}\left\{\log\frac{1}{\eta}\right\}^2/\kappa_b \\
\leq\ & C_{0,3}(\log n)n\epsilon_n^2/\kappa_b. \\
\leq\ & C_{0,3}(\log n)n\epsilon_n^2, \qquad \text{(A.11)}
\end{aligned}
$$

where $C_{0,3}$ is universal constant and the last inequality we use the fact that $1/\kappa_b = O(1)$.

4. For ζ_{4n}, take any constant C satisfies $C\left(-\log d_H\left(p_{\widehat{G}_n}, p_{G_0}\right)\right) > 2+e^3M$. Then, with $G_0 = G_n$, $G = H_j$, $\rho = \rho_n$ $\epsilon_0 = t^*\epsilon_n \geq d_H(p_{\widehat{G}_n}, p_{G_n})$ we apply (4.9) in Theorem 4.1. Since, $|\log\epsilon_n| = O(\log n)$ from $\epsilon_n = \frac{(\log n)^{3/2}}{\sqrt{n}}$ and $L(\rho_n; M) = O(\log n)$, we obtain

$$
\begin{aligned}
E_{n,\boldsymbol{\theta}}\zeta^2_{4n} &\leq n\max_{j\leq N}\left\{E_{G_n}(t^*_{H_j}(Y;\rho_n)-\theta)^2 - R^*(G_n)\right\} \\
&\leq nM_0\Big(|\log\epsilon_0|^3 + |\log\epsilon_0|\,\epsilon_0^{C-2} \\
&\quad + |\log\epsilon_0|\,L(\rho;b)^2 + |\log\rho_n|^2(b_1)^2\Big)\epsilon_0^2, \\
&\leq C_{0,4}\left\{(\log n)^3 \vee (\log n)^2(b_1)^2\right\} n\epsilon_n^2, \qquad \text{(A.12)}
\end{aligned}
$$

where $C_{0,4}$ is universal constant and M_0 is constants defined on Theorem 4.1 and $C_{0,4}$ is universal constant.

Therefore, by combining (A.8), (A.10), (A.11) and (A.12), we have

$$
E_{n,\boldsymbol{\theta}}\left(\sum_{j=0}^{4}|\zeta_{jn}|\right)^2 \leq 5\sum_{j=0}^{4}E_{n,\boldsymbol{\theta}}\zeta^2_{jn} \leq C^*\left\{(\log n)^3 \vee (\log n)^2(b_1)^2\right\} n\epsilon_n^2 \qquad \text{(A.13)}
$$

with for some universal constant $C^* > 0$. This proves the inequality in (3.23). □

A.2. Proof of Lemma 4.1

$$
\begin{aligned}
P_{n,\boldsymbol{\theta}}\left(t^*_{\widehat{G}_n}(X_i) \neq t^*_{\widehat{G}_n}(X_i;\rho_n)\right) &= P_{n,\boldsymbol{\theta}}\left(p_{\widehat{G}_n}(X_i) < \rho_n\right) \\
&\leq P_{n,\boldsymbol{\theta}}\left(\frac{q_n}{en}\{p(X_i|X_i)\} < \rho_n\right), \text{ by (4.1)}
\end{aligned}
$$

$$
\begin{aligned}
&\le P_{n,\boldsymbol{\theta}}\left(\frac{q_n}{en}\left\{p(X_{(n)}|X_{(n)})\right\} < \rho_n\right) \\
&\le P_{n,\boldsymbol{\theta}}\left(\log n/e^{1/(6X_{(n)})} < X_{(n)}\right) \\
&\le P_{n,\boldsymbol{\theta}}\left(\log n/e < X_{(n)}\right) \qquad \text{(A.14)} \\
&\le \sum_{i=1}^{n} P_{n,\boldsymbol{\theta}}\left(X_i > \log n/e\right) \\
&\le \sum_{i=1}^{n} e^{-\theta_i}\left(\frac{e\theta_i}{\lfloor \log n/e \rfloor}\right)^{\lfloor \log n/e \rfloor} \\
&\le n\left(\frac{eb}{\lfloor \log n/e \rfloor}\right)^{\lfloor \log n/e \rfloor} \to 0, \qquad \text{(A.15)}
\end{aligned}
$$

where (A.14) is obtained by using the fact that $N! < \sqrt{2\pi N}\left(\frac{N}{e}\right)^N e^{\frac{1}{12N}}$ for any positive integer N, and (A.15) uses the upper bound of the right tail probability of a Poisson distribution.

A.3. *Proof of Lemma 4.2*

This is the corresponding to the second result for Normal distribution Z in Lemma A.2.2. in [28]. Let Z_t is Poisson process with θt, then $Z_t - Z_s \sim Poisson(\theta(t-s))$ for $t > s$. Let $X_t = xe^{Z_t - 2bt}$ where $Z_t \sim Poisson(\theta t)$. We know that Z_t is a stationary process with independent increment. Define $P_t f(x) = E(f(X_t)|Z_0 = 0) = Ef(e^{Z_t-2bt}x)$, then we show that P_t forms a semigroup ($P_0 = I$, $P_{s+t} = P_t P_s$). Without loss of generality, we assume that $E(f(X)) = 0$. It is obvious that $P_0 = I$ since $P_0 f(x) = E(f(X_0)|X_0 = x) = f(x)$. We now show $P_{t+s} = P_t P_s$. Since $Z_{s+t} - Z_t$ and Z_t are independent, we have

$$
\begin{aligned}
P_{t+s}f(x) &= E(f(e^{Z_{t+s}-2b(t+s)}x)) = EE(f(e^{Z_{t+s}-Z_t-2bs}e^{Z_t-2bt}x)|X_t = y) \\
&= EEf(e^{Z_s-2bs}y|X_t = y) = EP_s f(X_t) = P_t(P_s f(x)) = P_t P_s f(x),
\end{aligned}
$$

which shows that P_t forms a semigroup. Define $G(t) = E\exp\left(rP_t f(Z)\right)$, then $G(0) = E\exp\left(rf(Z)\right)$ and $G(\infty) = E\exp\left(rP_\infty f(Z)\right) = E\exp(r0) = 1$ since $Z_t - 2bt \to -\infty$ a.e.. We also have $-G'(t) = -rE\exp(rP_t f(Z))AP_t f(Z) = r^2 E\exp(rP_t(f(Z)))||\nabla P_t f(Z)||^2$ where $A = \frac{d}{dt}P_t|_{t=0}$ and $\frac{d}{dt}P_t = AP_t$. We can assume without loss of generality that f is differentiable with gradient $\sup_x ||\nabla f(x)|| \le 1$. See p.439 in [28] for more detail. From this,

$$
\begin{aligned}
||\nabla P_t f(Z)|| &= ||E\nabla f(e^{Z_t-2bt}Z)||e^{Z_t-2bt} \le E(e^{Z_t-2bt}) \\
&= \sum_{x=0}^{\infty} e^{x-2bt}\frac{e^{-\theta t}(\theta t)^x}{x!} = \exp(-2bt - \theta t + e\theta t) \\
&= \exp(-t(2b-(e-1)\theta)) \le \exp(-(2-(e-1))bt) = \exp(-\kappa_b t),
\end{aligned}
$$

where $\kappa_b = (3-e)b$. Since $-G'(t) \le r^2 G(t) e^{-2\kappa_b t}$, we have $(\kappa_b \log G(t))' \ge (\frac{1}{2} r^2 e^{-2\kappa_b t})'$. Since $\kappa_b \log G(\infty) = 0$ and $\frac{1}{2} r^2 e^{-2\kappa_b t} = 0$ for $t = \infty$, we have $\kappa_b \log G(0) \le \frac{1}{2} r^2$ for $t = 0$ due to $(\kappa_b \log G(t))' \ge (\frac{1}{2} r^2 e^{-2\kappa_b t})'$. Therefore we obtain $E \exp(r f(Z)) = G(0) \le \exp\left(\frac{r^2}{2\kappa_b}\right)$. Using this result, we have $P(f(Z) \ge \lambda) = P(e^{rf(Z)} \ge e^{r\lambda}) \le e^{-r\lambda} E e^{rf(Z)} \le e^{\frac{r^2}{2\kappa_b} - r\lambda}$ and then by maximization w.r.t. r, we obtain $e^{-\frac{\kappa_b \lambda^2}{2}}$. Since $Ef(X) = 0$ is assumed, it still holds for $f(Z) - E(f(Z))$ which proves this lemma. $\square$

A.4. Proof of Lemma 4.3

First, define the function $h(\zeta) = e^{\zeta}$, we proof the inequality for two cases: $x \ge 1$ and $x = 0$.

1. When $x \ge 1$, we have

$$\begin{aligned} & h\left(\frac{|g_G(x)|}{p_G(x)}\right) \\ \le\ & \exp(E(|\theta - x||x)) \le E(e^{|\theta - x|}|X = x) \\ \le\ & \frac{1}{\sqrt{2\pi x} p_G(x)} \int e^{|u-x| - u + x\log u - x \log x + x} G(du) \\ \le\ & \frac{1}{\sqrt{2\pi} p_G(x)} \int e^{|u-x| - u + x\log u - x \log x + x} G(du), \end{aligned} \tag{A.16}$$

from $\frac{1}{x!} \le \frac{1}{\sqrt{2\pi}} e^{-(x+1/2)\log x + x}$. Define $f_\theta(x) = |\theta - x| - \theta + x \log\theta - x \log x + x$. We show the upper bound of equation (A.16) for three cases: (i) $x \ge e\theta$, (ii) $\theta \le x < e\theta$ and (iii) $x < \theta$.

(i) For $x \ge e\theta$, we have $|x - \theta| = x - \theta$ which leads to $f_\theta(x) = 2x - 2\theta + x\log\theta - x\log x$. Since $f_\theta(x)$ is decreasing in x for $x \ge e\theta$, so $f_\theta(x)$ has the maximum when $x = e\theta$ which is $f_\theta(e\theta) = (e-2)\theta$. This leads to (A.16) $\le \frac{1}{\sqrt{2\pi} p_G(x)} e^{(e-2)b}$, provided $\theta \in [0, b]$.

(ii) For $\theta \le x < e\theta$, we have $f_\theta(x) = 2x - 2\theta + x\log\theta - x\log x$ and $f_\theta(x)$ is increase on $x \in [\theta, e\theta]$. Therefore, we obtain (A.16) $\le \frac{1}{\sqrt{2\pi} p_G(x)} e^{(e-2)b}$.

(iii) For $x < \theta$, we have $f_\theta(x) = x\log\theta - x\log x$. Since $f'_\theta(x) = \log\theta/e - \log x$, we see that $f_\theta(x)$ has the maximum at $x = \theta e^{-1}$, so we have $f_\theta(\theta e^{-1}) = \theta e^{-1} \le b e^{-1}$ which leads to (A.16) $\le \frac{1}{\sqrt{2\pi} p_G(x)} e^{b/e}$.

From the results of (i), (ii) and (iii) for $x \ge 1$ and $e - 2 \ge e^{-1}$, we have

$$h\left(\frac{|g_G(x)|}{p_G(x)}\right) \le \frac{e^{(e-2)b}}{\sqrt{2\pi} p_G(x)} \le \frac{e^b}{p_G(x)}, \tag{A.17}$$

which means

$$\frac{|g_G(x)|}{p_G(x)} \le -\log\left(\frac{p_G(x)}{e^b}\right) \le L_1(p_G(x)). \tag{A.18}$$

2. When $x = 0$, we have $h(\frac{|g_G(0)|}{p_G(0)}) \le E(e^{\theta}|x = 0) = \frac{1}{p_G(0)} \int dG(\theta) = \frac{1}{p_G(0)}$. This leads to

$$\frac{|g_G(0)|}{p_G(0)} \le -\log p_G(0) \le -\log p_G(0) + b \le L(p_G(0); b). \tag{A.19}$$

Combining (A.18) and (A.19), we prove the lemma for $x \ge 0$. □

A.5. Proof of Lemma 4.4

If $\rho < p_G(x)$, then we obtain

$$|x - t_G^*(x;\rho)| = \frac{|g_G(x)|}{p_G(x)} \le L(p_G(x); b) \le L(\rho; b), \tag{A.20}$$

from Lemma 4.3 and the fact that $L(\rho; b)$ is decreasing in ρ.

If $p_G(x) \le \rho < 1$, then we obtain

$$|x - t_G^*(x,\rho)| \le \frac{|g_G(x)|}{\rho} = \frac{p_G(x)}{\rho}\frac{|g_G(x)|}{p_G(x)} \le \frac{p_G(x)L(p_G(x); b)}{\rho}, \tag{A.21}$$

from Lemma 4.3. Since $yL_1(y)$ is increasing for $\rho < 1$, we have

$$p_G(x)L(p_G(x); b) \le \rho L(\rho; b)$$

which results in

$$\begin{aligned} |x - t_G^*(x,\rho)| = \frac{|g_G(x)|}{p_G(x) \vee \rho} \le \frac{1}{\rho} p_G(x)L(p_G(x); b) &\le \frac{1}{\rho}\rho L(\rho; b) \\ &= L(\rho; b). \end{aligned} \tag{A.22}$$

Combining (A.20) and (A.22), we prove (4.5).

□

A.6. Proof of Lemma 4.5

Using the result $|t_G^*(x,\rho) - x| = \frac{|g_G(x)|}{p_G(x)\vee\rho} \le L(\rho; b)$ from Lemma 4.4,

$$\begin{aligned} |t_G^*(x+1;\rho) - t_G^*(x;\rho)| &\le 1 + |t_G^*(x+1;\rho) - (x+1)| + |t_G^*(x;\rho) - x| \\ &\le 1 + 2L(\rho; b). \end{aligned}$$

Without loss of generality, we assume $x < y$ and then we obtain

$$\begin{aligned} |t_G^*(y;\rho) - t_G^*(x;\rho)| &\le \sum_{i=x}^{y-1} |t_G^*(i+1;\rho) - t_G^*(i;\rho)| \le (1 + 2L(\rho; b))|y - x| \\ &\equiv A_{\rho,b}|y - x|, \end{aligned}$$

for $A_{\rho,b} = 1 + 2L(\rho; b)$. Let $h(\mathbf{x}) = ||t^*_G(\mathbf{x};\rho) - \boldsymbol{\theta}||$, then
$|h(\mathbf{x}) - h(\mathbf{y})| \leq ||t^*_G(\mathbf{x};\rho) - t^*_G(\mathbf{y};\rho)|| \leq A_{\rho,b}||\mathbf{x} - \mathbf{y}||$ leading to

$$\begin{aligned} & P_{n,\boldsymbol{\theta}}(||t^*_G(\mathbf{X};\rho) - t^*_G(\mathbf{Y};\rho)|| \geq E_{n,\boldsymbol{\theta}}||t^*_G(\mathbf{X};\rho) - t^*_G(\mathbf{Y};\rho)|| + x) \\ \leq\ & \exp\left(-\frac{\kappa_b x^2}{2A^2_{\rho,b}}\right), \end{aligned}$$

by Lemma 4.2.

□

A.7. Proof of Theorem 4.1

We first demonstrate the following two lemmas, Lemma A.7.1 and Lemma A.7.2, which will be used to prove Theorem 4.1.

Lemma A.7.1. *Let G_0 satisfying $G_0([0,b]) = 1$ and $G([0,M]) = 1$ with $M > \max(1,b)$ then we have*

$$\begin{aligned} & \sum_{x=0}^{\infty} \frac{(g_G(x) - g_{G_0}(x))^2}{p_G(x)\vee\rho + p_{G_0}(x)\vee\rho} \\ \leq\ & 4(M+1)\left(a^2 d_H(p_G, p_{G_0})^2 + 8M^2 e^{-M}\frac{(eM)^{a-2}}{(a-2)^{a-2}}\right), \qquad (A.23) \end{aligned}$$

for $a > M + 2$.

In particular, if M satisfies $\frac{a-2}{e^3 M} > 1$ as $a \to \infty$, then

$$\sum_{x=0}^{\infty} \frac{(g_G(x) - g_{G_0}(x))^2}{p_G(x)\vee\rho + p_{G_0}(x)\vee\rho} \leq 4(M+1)\left(a^2 d_H(p_G,p_{G_0})^2 + Ke^{-2a}\right), \quad (A.24)$$

for some constant $K > 0$.

Proof. Using $(g_G(x)-g_{G_0}(x))^2 \leq 2(x+1)^2(p_G(x+1)-p_{G_0}(x+1))^2+2x^2(p_G(x)-p_{G_0}(x))^2$, we have

$$\begin{aligned} \sum_{x=0}^{\infty} \frac{(g_G(x) - g_{G_0}(x))^2}{p_G(x)\vee\rho + p_{G_0}(x)\vee\rho} \quad \leq \quad & 2\sum_{x=0}^{\infty} \frac{(x+1)^2(p_G(x+1) - p_{G_0}(x+1))^2}{p_G(x)\vee\rho + p_{G_0}(x)\vee\rho} \\ & +2\sum_{x=0}^{\infty} \frac{x^2(p_G(x) - p_{G_0}(x))^2}{p_G(x)\vee\rho + p_{G_0}(x)\vee\rho}. \end{aligned}$$

We first show $\frac{p_G(x+1)\vee\rho + p_{G_0}(x+1)\vee\rho}{p_G(x)\vee\rho + p_{G_0}(x)\vee\rho} \leq M$.

For this, it is enough to show that $\frac{p_G(x+1)\vee\rho}{p_G(x)\vee\rho} \leq M$ and similary for G_0. There are four different cases which are

1. if $p_G(x+1) < \rho$ and $p_G(x) < \rho$, $\frac{p_G(x+1)\vee\rho}{p_G(x)\vee\rho} = \frac{\rho}{\rho} = 1$,

2. if $p_G(x+1) \geq \rho$ and $p_G(x) < \rho$, $\frac{p_G(x+1)\vee\rho}{p_G(x)\vee\rho} = \frac{p_G(x+1)}{\rho} \leq M$, since $p_G(x+1) \leq \frac{M}{x+1} p_G(x) < M\rho < M$,
3. if $p_G(x+1) < \rho$ and $p_G(x) \geq \rho$, $\frac{p_G(x+1)\vee\rho}{p_G(x)\vee\rho} = \frac{\rho}{p_G(x)} \leq 1$,
4. if $p_G(x+1) \geq \rho$ and $p_G(x) \geq \rho$, $\frac{p_G(x+1)\vee\rho}{p_G(x)\vee\rho} = \frac{p_G(x+1)}{p_G(x)} \leq M$.

From these four cases, we have $p_G(x+1)\vee\rho \leq M(p_G(x)\vee\rho)$ and $p_{G_0}(x+1)\vee\rho \leq M(p_{G_0}(x)\vee\rho)$ leading to

$$\frac{1}{p_G(x)\vee\rho + p_{G_0}(x)\vee\rho} \leq \frac{M}{p_G(x+1)\vee\rho + p_{G_0}(x+1)\vee\rho}. \tag{A.25}$$

We also have

$$\begin{aligned}
&\sum_{x=0}^{\infty} \frac{(g_G(x) - g_{G_0}(x))^2}{p_G(x)\vee\rho + p_{G_0}(x)\vee\rho} \\
\leq\ & 2M \sum_{x=0}^{\infty} \frac{(x+1)^2(p_G(x+1) - p_{G_0}(x+1))^2}{p_G(x+1)\vee\rho + p_{G_0}(x+1)\vee\rho} \\
& + 2\sum_{x=0}^{\infty} \frac{x^2(p_G(x) - p_{G_0}(x))^2}{p_G(x)\vee\rho + p_{G_0}(x)\vee\rho} \\
\leq\ & 2(M+1)\sum_{x=0}^{\infty} \frac{x^2(p_G(x) - p_{G_0}(x))^2}{p_G(x)\vee\rho + p_{G_0}(x)\vee\rho}.
\end{aligned} \tag{A.26}$$

By using

$$\frac{(\sqrt{p_G(x)} + \sqrt{p_{G_0}(x)})^2}{p_G(x)\vee\rho + p_{G_0}(x)\vee\rho} \leq \frac{2(p_G(x) + p_{G_0}(x))}{p_G(x)\vee\rho + p_{G_0}(x)\vee\rho} \leq 2, \tag{A.27}$$

we have

$$\begin{aligned}
(\text{A.26}) \leq\ & 4(M+1)\sum_{x=0}^{\infty} x^2(\sqrt{p_G(x)} - \sqrt{p_{G_0}(x)})^2 \\
=\ & 4(M+1)\Bigg(a^2 \sum_{x<a} \left(\sqrt{p_G(x)} - \sqrt{p_{G_0}(x)}\right)^2 \\
& \qquad + \sum_{x\geq a} x^2(\sqrt{p_G(x)} - \sqrt{p_{G_0}(x)})^2\Bigg) \\
\leq\ & 4(M+1)\Bigg(a^2 d_H(p_G, p_{G_0})^2 \\
& \qquad + 2\sum_{x\geq a} x^2(p_G(x) + p_{G_0}(x))\Bigg).
\end{aligned} \tag{A.28}$$

To obtain the upper bound of (A.28), we first find out bounds of $\sum_{x\geq a} x^2 p_G(x)$ and $\sum_{x\geq a} x^2 p_{G_0}(x)$ in (A.28). By using $\frac{e^{-\theta}\theta^x}{x!} \leq \frac{e^{-M}M^x}{x!}$ for $x \geq a > M+2$, we have

$$\begin{aligned}
\sum_{x\geq a} x^2 p_G(x) &\leq \int_{[0,M]} \sum_{x\geq a} x^2 \frac{e^{-\theta}\theta^x}{x!} dG(\theta) \leq \sum_{x\geq a} x^2 \frac{e^{-M}M^x}{x!} \\
&\leq \sum_{x\geq a} x(x-1)\frac{e^{-M}M^x}{x!} + \sum_{x\geq a} x\frac{e^{-M}M^x}{x!} \\
&\leq M^2 \sum_{x\geq a} \frac{e^{-M}M^{x-2}}{(x-2)!} + M\sum_{x\geq a}\frac{e^{-M}M^{x-1}}{(x-1)!} \\
&\leq 2M^2 P(X \geq a-2|\theta = M) \\
&\leq 2M^2 \sum_{x\geq a-2} \frac{e^{-M}M^x}{x!} \\
&\leq 2M^2 e^{-M}\frac{(eM)^{a-2}}{(a-2)^{a-2}},
\end{aligned}$$

where the last inequality is from $P(X \geq x) \leq \frac{e^{-\theta}(e\theta)^x}{x^x}$ for $X \sim Poisson(\theta)$. Similarly, we also have $\sum_{x\geq a} x^2 p_{G_0}(x) \leq 2M^2 e^{-M}\frac{(eM)^{a-2}}{(a-2)^{a-2}}$. Therefore we have

$$\begin{aligned}
&\sum_{x=0}^{\infty} \frac{(g_G(x) - g_{G_0}(x))^2}{p_G(x)\vee\rho + p_{G_0}(x)\vee\rho} \\
\leq\ & 4(M+1)\left(a^2 d_H(p_G, p_{G_0})^2 + 8M^2 e^{-M}\frac{(eM)^{a-2}}{(a-2)^{a-2}}\right),
\end{aligned}$$

which proves (A.23).

To prove (A.24), by using $8M^2e^{-M} \leq K$ for some constant $K > 0$, if M satisfies the following condition

$$\frac{a-2}{M} > e^3, \tag{A.29}$$

then we have

$$\begin{aligned}
\frac{a-2}{M} > e^3 &\iff \frac{a-2}{e^3 M} > 1 \\
&\iff e^2 \leq \left(\sqrt{\frac{a-2}{e^3M}}\right)^{a-2} \quad \text{for sufficiently large } a \\
&\iff e^a \leq \left(\sqrt{\frac{a-2}{eM}}\right)^{a-2} \\
&\iff \left(\sqrt{\frac{eM}{a-2}}\right)^{a-2} \leq e^{-a},
\end{aligned}$$

which leads to

$$\sum_{x=0}^{\infty} \frac{(g_G(x) - g_{G_0}(x))^2}{p_G(x)\vee\rho + p_{G_0}(x)\vee\rho}$$

$$\leq \quad 4(M+1)\left(a^2 d_H(p_G, p_{G_0})^2 8M^2 e^{-M}\frac{(eM)^{a-2}}{(a-2)^{a-2}}\right)$$
$$\leq \quad 4(M+1)\left(a^2 d_H(p_G, p_{G_0})^2 + Ke^{-2a}\right). \quad \square \tag{A.30}$$

Lemma A.7.2. *Define* $\|f\|_h = \left\{\int f^2(x)h(x)dx\right\}^{1/2}$. *For* $\rho \to 0$, *assume that the support of the distribution* G_0 *is* $[0,b]$ *with* $b = o\left(\frac{-\log\rho}{\log(-\log\rho)}\right)$. *Then we have,*

$$||(1 - p_{G_0}/\rho)_+ g_{G_0}/p_{G_0}||_{p_{G_0}} = O\left(\rho^{1/2} b_1\right), \tag{A.31}$$

where $b_1 = b \vee 1$.

Proof. Note that $p_{G_0}(x) = \int p(x|\theta)dG_0(\theta)$ is decreasing as $x \to 0$ or $x \to \infty$. Therefore, for sufficiently small ρ, the following two sets $\{x : 0 \leq x \leq c_1\}$ and $\{x : x \geq c_2\}$ with some constants c_1 and c_2 depending on ρ can be considered as candidates for the asymptotically upper set of the set $\{x : p_{G_0}(x) < \rho\}$. However, in the former case, $p_{G_0}(x) = \int p(x|\theta)dG_0(\theta)$ has a as a lower bound as $p(0|b) = e^{-b}$, which is always greater than ρ due to the assumption of $b = o\left(\frac{-\log\rho}{\log(-\log\rho)}\right)$. Thus, for sufficiently small ρ, we have $\{x : p_{G_0}(x) < \rho\} \subseteq \{x : x \geq C_\rho\}$ for some constant C_ρ depending on ρ.

From now on, we figure out the specific form of C_ρ. Recall that we only consider the distribution G_0 which does not degenerate at 0. That is there exists constant $b_1 > 0$ such that $G_0([0, b_1)) \leq \delta_0$ for some $\delta_0 < 1$. Therefore, $G_0([b_1, b]) = w_0 \geq 1 - \delta_0$. We consider the following two cases according to the relationship between x and b. We mainly use the following facts:

1. $b = o\left(\frac{-\log\rho}{\log(-\log\rho)}\right)$ implies $b\log b = o(-\log\rho)$
2. For any $x \in \mathbb{R}$, define $f_x : [0,b] \to \mathbb{R}$ such that $f_x(\theta) = \theta^x e^{-\theta}$. Then $f_x(\theta)$ is increasing (decreasing) function when $\theta \leq x$ ($\theta > x$).

Case 1. $x > b$

We show that the set $\{x : p_{G_0}(x) < \rho\}$ is asymptotically subset of $\{x : x\log x > -\log\rho\}$. Note that

$$w_0 b_1^x e^{-b_1} \leq \int \theta^x e^{-\theta} dG_0(\theta) \leq b^x e^{-b}.$$

Therefore using the fact that $\log x! = O(x\log x)$ for $x \to \infty$, we derive the following

$$\begin{aligned} \{x : p_{G_0}(x) < \rho\} \quad &\subseteq \quad \{x : w_0 b_1^x e^{-b_1}/x! < \rho\} \\ &= \quad \{\log x! - x\log b_1 + b_1 - \log w_0 > -\log\rho\} \\ &= \quad \{\log x!(1 + o(1)) > -\log\rho\} \\ &\asymp \quad \{x : x\log x > -\log\rho\}. \end{aligned} \tag{A.32}$$

Case 2. $x \leq b$

In this case we show that $p_{G_0}(x) > \rho$.

Note that

$$w_0 p(\lfloor b \rfloor | b_1) \quad \gtrsim \quad \rho, \tag{A.33}$$

$$\begin{aligned}(\because) - \log w_0 p(\lfloor b \rfloor | b_1) &= \log \lfloor b \rfloor ! - \lfloor b \rfloor \log b_1 + b_1 - \log w_0 \\ &\asymp b \log b \\ &= o(-\log \rho) < -\log \rho,\end{aligned}$$

and

$$w_0 p(0|b) = w_0 e^{-b} \gtrsim \rho. \tag{A.34}$$

Combining (A.33) and (A.34)

$$p_{G_0}(x) \geq w_0 \min\left(p(\lfloor b \rfloor | b_1), p(0|b)\right) \gtrsim \rho. \tag{A.35}$$

Therefore for the **Case 2.** $x \leq b$, the set $\{x : p_{G_0}(x) < \rho\}$ is asymptotically empty set.

$$\{x : p_{G_0}(x) < \rho\} \asymp \phi. \tag{A.36}$$

Thus from the results (A.32) and (A.36), we conclude that $p_{G_0}(x) < \rho$ only happens only when $x > b$ and the set of x satisfying $p_{G_0}(x) < \rho$ becomes a subset of the set $\{x : x \log x > -\log \rho\}$ asymptotically. Define $C_\rho = \frac{-\log \rho}{\log(-\log \rho)}(1+o(1))$ then $x \log x \sim -\log \rho$ when $x = C_\rho$. Therefore, it can be shown asymptotically that

$$\{x : p_{G_0}(x) < \rho\} \approx \{x : x \geq C_\rho\}. \tag{A.37}$$

Therefore,

$$\begin{aligned}||(1 - p_{G_0}/\rho)_+ g_{G_0}/p_{G_0}||^2_{p_{G_0}} &\leq \sum_{x : p_{G_0}(x) < \rho} \left(\frac{g_{G_0}(x)}{p_{G_0}(x)}\right)^2 p_{G_0}(x) \\ &\leq \sum_{x : p_{G_0}(x) < \rho} x^2 p_{G_0}(x) \\ &\leq \sum_{x : p_{G_0}(x) < \rho} (b^2 p_{G_0}(x-2) + b p_{G_0}(x-1)) \\ &\leq 2(b \vee 1)^2 \sum_{x : p_{G_0}(x) < \rho} p(x-2|b) \\ &\leq 2(b_1)^2 \sum_{x > C_\rho - 2} p(x|b) \\ &\leq 2(b_1)^2 \frac{(eb)^{C_\rho - 2}}{(C_\rho - 2)^{C_\rho - 2}}, \\ &\sim 2(b_1)^2 e^{-C_\rho(\log C_\rho - \log(eb))} \sim 2(b_1)^2 \rho,\end{aligned}$$

leading to $||(1 - p_{G_0}/\rho)_+ g_{G_0}/p_{G_0}||_{p_{G_0}} = O\left(\rho^{1/2} b_1\right)$ as $\rho \to 0$. □

Now the proof of Theorem 4.1 is directly demonstrated by the previous two lemmas, Lemma A.7.1 and Lemma A.7.2.

Proof of Theorem 4.1. We denote $d_H(p_G, p_{G_0})$ as d_H for notation simplicity.

1. Let $w_* = 1/(p_G \vee \rho + p_{G_0} \vee \rho)$. We have

$$\begin{aligned}
& \{E_{G_0}(t^*_G(Y;\rho) - \theta)^2 - R^*(G_0)\}^{1/2} \\
= \quad & ||g_G/(p_G \vee \rho) - g_{G_0}/(p_{G_0} \vee \rho)||_{p_{G_0}} \\
\leq \quad & 2||g_G - g_{G_0}||_{w_*} + 2L(\rho;M)\sqrt{2}d_H \\
& + ||(1 - p_{G_0}/\rho)_+ g_{G_0}/p_{G_0}||_{p_{G_0}} \\
\leq \quad & 2\sqrt{D}\left(a^2 d_H^2 + 8M^2 e^{-M}\frac{(eM)^{a-2}}{(a-2)^{a-2}}\right)^{1/2} \\
& + 2\sqrt{2}L(\rho;M)d_H + ||(1 - p_{G_0}/\rho)_+ g_{G_0}/p_{G_0}||_{p_{G_0}} \\
\leq \quad & 2\sqrt{D}\left(a d_H + \sqrt{8M^2 e^{-M}\frac{(eM)^{a-2}}{(a-2)^{a-2}}}\right) \\
& + 2\sqrt{2}L(\rho;M)d_H + ||(1 - p_{G_0}/\rho)_+ g_{G_0}/p_{G_0}||_{p_{G_0}}. \qquad \text{(A.38)}
\end{aligned}$$

Here we use the result

$$||g_G - g_{G_0}||^2_{w_*} \leq 4(M+1)\left(a^2 d_H(p_G, p_{G_0})^2 + 8M^2 e^{-M}\frac{(eM)^{a-2}}{(a-2)^{a-2}}\right),$$

from Lemma A.7.1

2. (4.8) can be derived directly from the equation (A.30), (A.38) and Lemma A.7.2.
3. For (4.9), note that $D = 4(M+1) \asymp -\log d_H(p_G, p_{G_0})$ implies there is constant C_0 such that

$$\begin{aligned}
& \left\{E_{G_0}\left(t^*_G(Y;\rho) - \theta\right)^2 - R^*(G_0)\right\}^{1/2} \\
\leq \quad & \left\{2\sqrt{D}\left(C(-\log d_H) + \sqrt{K}(d_H)^{C-1}\right) + 2\sqrt{2}L(\rho;M)\right\} d_H \\
& + \rho^{1/2} b_1 \\
\leq \quad & C_0\left\{(-\log d_H)^{3/2} + (-\log d_H)^{1/2}\left((d_H)^{C-1} + L(\rho;b)\right)\right\} d_H \\
& \quad + \rho^{1/2} b_1. \qquad \text{(A.39)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
& E_{G_0}\left(t^*_G(Y;\rho) - \theta\right)^2 - R^*(G_0) \\
\leq \quad & 2\Bigg\{3C_0^2\Bigg((-\log d_H)^3 + (-\log d_H)(d_H)^{2C-2} \\
& \quad + (-\log d_H)L(\rho;M)^2\Bigg)(d_H)^2 + \rho(b_1)^2\Bigg\}
\end{aligned}$$

$$\leq \quad M_0\Big(|\log \epsilon_0|^3 + |\log \epsilon_0|\, \epsilon_0^{C-2} + |\log \epsilon_0|\, L(\rho; M)^2 + |\log \rho|^2 (b_1)^2 \Big) \epsilon_0^2, \tag{A.40}$$

where the last inequality above can be derived from the fact that $(-\log y)^3 y^2$, $(-\log y) y^{2C}$, and $(-\log y) y^2$ are increasing with respect to y for $y \leq e^{-3/2}$, and $\rho \leq |\log \rho|^2 \epsilon_0^2$.

□

A.8. Proof of Lemma 4.6

Define $I = [0, M]$ and notice that $[0, M]$ is the support of G_m and G. We also define $k^* = \lceil 4.32M \rceil$. From the Carathéodory's theorem, there exists a discrete distribution function G_m for each G such that the G_m with support on $[0, M]$ and no more than $m = \lceil 4.32M \rceil + B + 2$ support points satisfying

$$\int_I u^j G(du) = \int_I u^j G_m(du), \tag{A.41}$$

for $j = 0, 1, \cdots, k^* + B$. Hence the total number of moments of two distributions must be equal is $k^* + B + 1$. By using the Taylor expansion of $e^{-\theta}$ at $\theta = 0$, we have

$$e^{-\theta} = \sum_{j=0}^{k-1} \frac{(-1)^j \theta^j}{j!} + R_k, \quad |R_k| \leq \frac{\theta^k}{k!}. \tag{A.42}$$

Using the above result, we can define the approximation error between the probability mass function of Poisson distribution with parameter θ and its Taylor expansion function of order $k-1$ as follows,

$$\mathrm{Rem_k}(x, \theta) \equiv \frac{\theta^x e^{-\theta}}{x!} - \frac{\theta^x}{x!} \sum_{j=0}^{k-1} \frac{(-1)^j \theta^j}{j!}. \tag{A.43}$$

Then for $x = 0, 1, \cdots, B$ and $\theta \in [0, M]$,

$$\begin{aligned} \Big|\mathrm{Rem_{k^*}}(x, \theta)\Big| &\leq \frac{\theta^x}{x!} \frac{\theta^{k^*}}{k^*!} = \frac{\theta^x e^{-\theta}}{x!} \frac{\theta^{k^*} e^{\theta}}{k^*!} \\ &\leq \frac{M^{k^*} e^M}{k^*!} && \text{(A.44)} \\ &\leq \frac{1}{\sqrt{2\pi k^*}} \left(\frac{eM}{k^*}\right)^{k^*} e^M && \text{(A.45)} \\ &\leq \frac{1}{2} e^{-M}, && \text{(A.46)} \end{aligned}$$

where in (A.44) we use the fact that $p(x|\theta) = \frac{\theta^x e^{-\theta}}{x!} \leq 1$ and $\theta \leq M$, in (A.45) the well-known factorial inequality $\sqrt{2\pi k^*} (k^*/e)^{k^*} e^{1/(12k^*+1)} < k^*!$, and in (A.46) the assumption $k^* \geq 4.32M$ which implies $(eM/k^*)^{k^*} \leq e^{-2M}$.

Therefore,

$$\begin{aligned}\|p_G(x) - p_{G_m}(x)\|_{\infty,B} &= \sup_{x\le B}\left|\int_I \frac{\theta^x}{x!}\sum_{j=0}^{k^*-1}\frac{(-1)^j\theta^j}{j!}d(G-G_m)(\theta)\right| \\ &\quad +2\sup_{x\le B,\theta\le M}\left|\mathrm{Rem}(x,\theta)\right| \\ &\le e^{-M} \le C_1\eta. \end{aligned} \tag{A.47}$$

where $C_1 > 0$ is an universal constant.

Similarly, we obtain a bound for $||g_{G_m} - g_G||_{\infty,B}$ as follows:

$$g_G(x) = \int((x+1)p(x+1|\theta) - xp(x|\theta))G(d\theta) = \int p(x|\theta)\,(\theta - x)\,G(d\theta).$$

For $x \le B$, we have

$$|\theta - x|Rem_{k^*}(x,\theta) \le (B\vee M)\frac{1}{2}e^{-M}. \tag{A.48}$$

Therefore, we have

$$||g_G(x) - g_{G_m}(x)||_{\infty,B} \le C_1\eta(B\vee M). \quad \square \tag{A.49}$$

A.9. Proof of Lemma 4.7

Let $\mathcal{P}_{\mathcal{G}'} \subset \mathcal{P}$ be the set of all distribution $\int p(x|u)dG'(u)$ where G' has at most $m \le \lceil 4.32\mathrm{M}\rceil + B + 2$ support points on $[0, M]$. By Lemma 4.6, $\mathcal{P}_{\mathcal{G}'}$ is an η-net over $\mathcal{P}$ with distance $\|\cdot\|_{\infty,M}$. Therefore, with $\|\cdot\|_{\infty,M}$, an η-net over $\mathcal{P}_{\mathcal{G}'}$ is an 2η-net over $\mathcal{P}$.

Next we define the followings:

$$\eta\text{-covering set } \{\boldsymbol{W}_1^m,\cdots,\boldsymbol{W}_{N_w}^m\} \text{ of the } L \text{ dimensional simplex } \mathcal{W}^m \\ \text{with } \|\cdot\|_1 \text{ s.t } N_w = N(\eta,\mathcal{W}^m,\|\cdot\|_1) \text{ then } N_w \lesssim \left(\frac{5}{\eta}\right)^m,$$

and

$$\eta\text{-covering set } \{\boldsymbol{U}_1^m,\cdots,\boldsymbol{U}_{N_u}^m\} \text{ of } [0,M]^m \text{ with } \|\cdot\|_\infty \\ \text{s.t } N_u = N(\eta,[0,M]^m,\|\cdot\|_\infty) \text{ then } N_u \lesssim \frac{1}{\eta}\left(\frac{M}{\eta}\right)^m. \tag{A.50}$$

For any $p_G \in \mathcal{P}$, by the Lemma 4.6, there exists $p_{G'} \in \mathcal{P}_{\mathcal{G}'}$ such that $\|p_G - p_{G'}\|_{\infty,M} \lesssim \eta$ with $G'(u) = \sum_{j=1}^m w_j I(u \le u_j)$.

Define weight and location parameters W^m and U^m of G' denoted by

$$\boldsymbol{W}^m = (w_1,\cdots,w_m), \quad \boldsymbol{U}^m = (u_1,\cdots,u_m).$$

Then, there exist

$$\widetilde{\boldsymbol{W}}^m (= (\widetilde{w}_1, \cdots, \widetilde{w}_m)) \in \{\boldsymbol{W}_1^m, \cdots, \boldsymbol{W}_{N_w^m}\},$$

and

$$\widetilde{\boldsymbol{U}}^m (= (\widetilde{u_1}, \cdots, \widetilde{u_m})) \in \{\boldsymbol{U}_1^m, \cdots, \boldsymbol{U}_{N_u}^m\}$$

such that $\|\boldsymbol{W}^m - \widetilde{\boldsymbol{W}}^m\|_1 \leq \eta$ and $\|\boldsymbol{U}^m - \widetilde{\boldsymbol{U}}^m\|_\infty \leq \eta$.

We define $\mathcal{P}_{\mathcal{G}''} \subset \mathcal{P}$ which is the set of all distributions $\int p(x|u)dG''(u)$ where G'' is constructed of a combination of

$$\{\boldsymbol{U}_1^m, \cdots, \boldsymbol{U}_{N_u}^m\} \text{ and } \{\boldsymbol{W}_1^m, \cdots, \boldsymbol{W}_{N_w}^m\}.$$

Then the cardinality of the $\mathcal{P}_{\mathcal{G}''}$, $\#\mathcal{P}_{\mathcal{G}''}$ has an upper bound which is

$$\#\mathcal{P}_{\mathcal{G}''} \lesssim \frac{1}{\eta}\left(\frac{5}{\eta}\right)^m \times \left(\frac{M}{\eta}\right)^m. \tag{A.51}$$

Now we show that there exists $p_{G''} \in \mathcal{P}_{\mathcal{G}''}$ such that

$$\|p_{G'} - p_{G''}\|_\infty \lesssim \eta.$$

First, let us construct the following two distributions:

$$H''(u) = \sum_{j=1}^m w_j I(\widetilde{u}_j \leq u), \quad G''(u) = \sum_{j=1}^m \widetilde{w}_j I(\widetilde{u}_j \leq u).$$

Note that

$$\begin{aligned} \left\|\int p(x|u)d(G' - H'')(u)\right\|_{\infty,B} &\leq \left\|\sum_{j=1}^m w_j \left\{p(x|u_j) - p(x|\widetilde{u}_j)\right\}\right\|_{\infty,B} \\ &\leq \|\boldsymbol{U} - \widetilde{\boldsymbol{U}}\|_\infty \times \sup_{x\leq B, u\leq M}\left|\frac{\partial}{\partial u}p(x|u)\right| \\ &\leq \eta \times \sup_{x\leq B, u\leq M}\left|\frac{\partial}{\partial u}p(x|u)\right|, \end{aligned} \tag{A.52}$$

and

$$\begin{aligned} \left\|\int p(x|u)d(H'' - G'')(u)\right\|_{\infty,B} &\leq \left\|\sum_{j=1}^m (w_j - \widetilde{w}_j)p(x|\widetilde{u}_i)\right\|_{\infty,B} \\ &\leq \|\boldsymbol{W} - \widetilde{\boldsymbol{W}}\|_1 \times \sup_{x,u\in[0,M]}|p(x|u)| \\ &\leq \eta \times \sup_{x\leq B, u\leq M}|p(x|u)|. \end{aligned} \tag{A.53}$$

We can easily show $\sup_{x\le B, u\le M}\left|\frac{\partial}{\partial u}p(x|u)\right| \le 1$ in (A.52) from

$$\frac{\partial}{\partial u}\{p(x|u)\} = \begin{cases} -\exp(-\theta) & x=0, \\ p(x-1|\theta)-p(x|\theta) & x\ge 1. \end{cases}$$

Similarly, we also obtain $\sup_{x\le B,u\le M}|p(x|u)| \le 1$ since $p(x|u)\le 1$. Therefore, the sup norm between $p_{G'}$ and $p_{G''}$ is

$$\begin{aligned} \|p_{G'}-p_{G''}\|_{\infty,B} &\le \left\|\int p(x|u)d(G'-H'')(u)\right\|_{\infty,B} \\ &\quad + \left\|\int p(x|u)d(H''-G'')(u)\right\|_{\infty,B} \\ &\le \eta\times \sup_{x\le B,u\le M}\left|\frac{\partial}{\partial u}p(x|u)\right| + \eta\times\sup_{x\le B,u\le M}|p(x|u)| \\ &\lesssim \eta. \end{aligned} \tag{A.54}$$

Using $\#\mathcal{P}_{\mathcal{G}''}\lesssim \frac{1}{\eta}\left(\frac{M}{\eta}\right)^m \times \left(\frac{5}{\eta}\right)^m$ in (A.51) with

$$m \le \lceil 4.32\mathrm{M}\rceil + B + 2$$

we derive the following result by Lemma 4.6 and $M \asymp \log\frac{1}{\eta}$:

$$\log N\left(\eta, \mathcal{P}, \|\cdot\|_{\infty,B}\right) \lesssim \left(\lceil 4.32\mathrm{M}\rceil + B+2\right)\log\left(\frac{1}{\eta}\right). \quad \square \tag{A.55}$$

A.10. Proof of Lemma 4.8

We first have

$$\begin{aligned} E_{n,\boldsymbol{\theta}}\left\{\prod_{i=1}^{n}(aX_i)^{I\{X_i\ge B\}}\right\}^{\lambda} &\le \prod_{i=1}^{n}(1+a^\lambda E_{n,\boldsymbol{\theta}}X_i^\lambda I(X_i\ge B)) \\ &\le \exp\left\{a^\lambda n E_{n,\boldsymbol{\theta}}X^\lambda I(X\ge B)\right\}. \end{aligned}$$

In (4.10), using $B\ge 1$ and $\theta_i\in[0,b]$, for some $b>0$,
we have $X^\lambda \le \frac{X}{B^{1-\lambda}}$ for $B\ge 1$ which leads to

$$\begin{aligned} E_{n,\boldsymbol{\theta}}(X^\lambda I(X\ge B)) &\le B^{\lambda-1}E_{n,\boldsymbol{\theta}}(XI(X\ge B)) \\ &\le B^{\lambda-1}b\sum_{x\ge B-1}\frac{e^{-\theta}\theta^x}{x!} \le B^{\lambda-1}b\frac{e^{-b}(eb)^{B-1}}{(B-1)^{B-1}}. \end{aligned}$$

From this, we have $E_{n,\boldsymbol{\theta}}(X^\lambda I(X\ge B)) \le B^{\lambda-1}b^Be^{-b}\left(\frac{e}{B-1}\right)^{B-1}$ and we prove the lemma.

$\square$

A.11. Proof of Theorem 4.2

The proof adopts the idea of Theorem 1 in [34], however there exist some different techniques due to using Poisson probability function $p(x|\theta) = \frac{e^{-\theta}\theta^x}{x!}$ instead of the Normal density. Let $\eta = \frac{1}{n^4}$ and $M = L\log\frac{1}{\eta}$ with some constant $L \geq 1$ as in Lemma 4.6. Additionally, let $B = \lfloor M \rfloor$, and define $p^*(x) \equiv \eta I(x \leq B) + (\eta B^2)x^{-2}I(x > B)$. We first consider any approximate solution of the NPMLE satisfying $L_n(p_{\widehat{G}_n}, p_{G_n}) = \prod_{i=1}^n \{p_{\widehat{G}_n}(X_i)/p_{G_n}(X_i)\} \geq e^{-2t^2 n\epsilon_n^2/15}$. We also let $\{p_j, j \leq N\}$ be an η-net of $\mathcal{P}$ under $||\cdot||_{\infty,B}$ with $N \equiv N(\eta, \mathcal{P}, ||\cdot||_{\infty,B})$ where $\mathcal{P}$ defined on Lemma 4.7.

Let $p_{0,j} \in \mathcal{P}$ satisfying $d_H(p_{0,j}, p_{G_n}) \geq t\epsilon_n$, $\|p_{0,j} - p_j\|_{\infty,B} \leq \eta$ if they exist and define $J \equiv \{j \leq N; \exists p_{0,j}\}$. Then for any $p \in \mathcal{P}$ with $d_H(p, p_{G_n}) \geq t\epsilon_n$, there exists $j \in J$ such that

$$p(x) \leq \begin{cases} p_{0,j}(x) + 2\eta = p_{0,j}(x) + 2p^*(x), & x \leq B \\ 1, & x > B. \end{cases} \tag{A.56}$$

Following the proof of Theorem 1 in Zhang (2009), we obtain

$$\begin{aligned} & P_{n,\boldsymbol{\theta}}\{d_H(p_{\widehat{G}_n}, p_{G_n}) \geq t\epsilon_n\} \\ \leq\ & P_{n,\boldsymbol{\theta}}\left(\sup_{j\in J}\prod_{i=1}^n \frac{p_{0,j}(X_i) + 2p^*(X_i)}{p_{G_n}(X_i)} \geq e^{-4t^2 n\epsilon_n^2/5}\right) \\ & + P_{n,\boldsymbol{\theta}}\left(\prod_{X_i > B}\frac{1}{2p^*(X_i)} \geq e^{2t^2 n\epsilon_n^2/3}\right). \end{aligned} \tag{A.57}$$

We consider the first term in equation (A.57). For $\epsilon_n = \frac{(\log n)^{3/2}}{\sqrt{n}}$, Lemma 4.7 implies,

$$\begin{aligned} \log N + n\sqrt{4\eta B} &\leq C_1(\log n)^2 + 2\sqrt{L\log n}, \text{ some constants } C_1 \text{ and } L \geq 1. \\ &\leq \frac{nt^2\epsilon_n^2}{20}, \text{ for sufficiently large } t. \end{aligned} \tag{A.58}$$

Note that,

$$\begin{aligned} & P_{n,\boldsymbol{\theta}}\left(\prod_{i=1}^n \frac{p_{0,j}(X_i) + 2p^*(X_i)}{p_{G_n}(X_i)} \geq e^{-4t^2 n\epsilon_n^2/5}\right) \\ \leq\ & e^{2t^2 n\epsilon_n^2/5}\prod_{i=1}^n E_{n,\boldsymbol{\theta}}\sqrt{\frac{p_{0,j}(X_i) + 2p^*(X_i)}{p_{G_n}(X_i)}} \\ \leq\ & \exp\left\{\frac{2nt^2\epsilon_n^2}{5} + \sum_{i=1}^n E_{n,\boldsymbol{\theta}}\left(\sqrt{\frac{\{p_{0,j}(X_i) + 2p^*(X_i)\}}{p_{G_n}(X_i)}} - 1\right)\right\} \\ =\ & \exp\left\{\frac{2nt^2\epsilon_n^2}{5} + n\left(\int\sqrt{(p_{0,j} + 2p^*)\, p_{G_n}} - 1\right)\right\} \end{aligned}$$

$$\leq \quad \exp\left\{-\frac{nt^2\epsilon_n^2}{10} + n\sqrt{4\eta B}\right\}, \tag{A.59}$$

where the last inequality above is derived from $\int\sqrt{(p_{0,j}+2p^*)\,p_{G_n}} - 1 \leq -\frac{d_H^2(p_{0,j},p_{G_n})}{2} + \sqrt{2\int p^*}$, $-\frac{d_H^2(p_{0,j}p_{G_n})}{2} \leq -\frac{t^2\epsilon_n^2}{2}$ and $\sqrt{2\int p^*} \leq \sqrt{4\eta B}$. Therefore, we obtain the bound of the first term in (A.57) as follows:

$$\begin{aligned} P_{n,\boldsymbol{\theta}}\left(\sup_{j\in J}\prod_{i=1}^{n}\frac{p_{0,j}(X_i)+2p^*(X_i)}{p_{G_n}(X_i)} \geq e^{-4t^2n\epsilon_n^2/5}\right) \quad &\leq \quad e^{\log N + n\sqrt{4\eta B} - \frac{nt^2\epsilon_n^2}{10}} \\ &\leq \quad e^{-\frac{nt^2\epsilon_n^2}{20}}. \end{aligned} \tag{A.60}$$

Next, consider the second term in equation (A.57). By Lemma 4.8 with $a = n^2/B$ and $\lambda = \frac{1}{\log n}$, we know that

$$\begin{aligned} & P_{n,\boldsymbol{\theta}}\left(\prod_{X_i>B}\frac{1}{2p^*(X_i)} \geq e^{2t^2n\epsilon_n^2/3}\right) \\ \leq \quad & e^{-2nt^2\epsilon_n^2/(3\log n)} E_{n,\boldsymbol{\theta}}\left\{\prod_{X_i>B}\frac{n^2}{B}X_i\right\}^{1/\log n} \\ \leq \quad & e^{-2nt^2\epsilon_n^2/(3\log n)} \exp\left\{\left(\frac{n^2}{B}\right)^{1/\log n} nB^{\frac{1}{\log n}-1} b^B e^{-b} \right. \\ & \left. \times\left(\frac{e}{B-1}\right)^{B-1}\right\}. \end{aligned} \tag{A.61}$$

The second exponential term in (A.61) can be handled using the following fact,

$$\begin{aligned} & \left(\frac{n^2}{B}\right)^{1/\log n} nB^{\frac{1}{\log n}-1} b^B e^{-b}\left(\frac{e}{B-1}\right)^{B-1} \\ = \quad & \exp\left\{2 + \log n + \log\left(\frac{b}{B}\right) - b - (B-1)\log\left(\frac{B-1}{eb}\right\}\right\} \\ \leq \quad & \exp\left\{2 + \log n - (B-1)\log\left(\frac{B-1}{eb}\right)\right\} \\ \leq \quad & \exp\left\{3 + (4L\log n - 1) - (B-1)\log\left(\frac{B-1}{eb}\right)\right\} \\ = \quad & \exp\left\{3 - (B-1)\log\left(\frac{B-1}{e^2 b}\right)\right\}, \end{aligned}$$

for sufficiently large n, equivalently

$$\exp\left\{\left(\frac{n^2}{B}\right)^{1/\log n} nB^{\frac{1}{\log n}-1} b^B e^{-b}\left(\frac{e}{B-1}\right)^{B-1}\right\} \leq \exp\left(\exp(3)\right).$$

Therefore, we obtain the bound of the second term in (A.57) as follows:

$$P_{n,\boldsymbol{\theta}}\left(\prod_{X_i\geq M}\frac{1}{2p^*(X_i)}\geq e^{2nt^2\epsilon_n^2/3}\right) \leq e^{-(\frac{2t^2}{3}-K)\frac{n\epsilon_n^2}{\log n}} \leq e^{-(\frac{t^2}{3})\frac{n\epsilon_n^2}{\log n}}, \tag{A.62}$$

where $K=\frac{\exp(3)}{(\log n)^2}$ hence the last term is derived by $K\leq t^2/3$ for sufficiently large n. By combining (A.60) and (A.62) we finally obtain the result.

□

A.12. Proof of Theorem 4.3

We show that for any $G\in\mathcal{G}_{[0,M]}$, there exists H such that $t^*_G(\cdot;\rho)$ is approximated by $t^*_H(\cdot;\rho)$ where $H(u)=\sum_{j=1}^m w_j I\left(u_j\leq u\right)$ is a discrete distribution on $[0,M]$ with at most $m\leq(\lceil 4.32\mathrm{M}\rceil+M+2)$ support points. First using Lemma 4.4 and 4.6, we have

$$\begin{aligned}
&\|t^*_G(x;\rho)-t^*_H(x;\rho)\|_{\infty,B}\\
\leq\;& \left\|\frac{g_G(x)}{p_G(x)\vee\rho}-\frac{g_H(x)}{p_G(x)\vee\rho}\right\|_{\infty,B}+\left\|\frac{g_H(x)}{p_G(x)\vee\rho}-\frac{g_H(x)}{p_H(x)\vee\rho}\right\|_{\infty,B}\\
\leq\;& \frac{1}{\rho}\|g_G(x)-g_H(x)\|_{\infty,B}+\frac{L(\rho;M)}{\rho}\|p_G(x)-p_H(x)\|_{\infty,B}\\
\leq\;& \frac{\eta}{\rho}\left(C_1L(\rho;b)+C_2(-\log\eta)\right),
\end{aligned}\tag{A.63}$$

for some constant $C_1>0$ and $C_2>0$.

Similar to Lemma 4.7, let

$$H'(u)=\sum_{j=1}^m w_jI\left(\widetilde{u}_j\leq u\right),\ \text{and}\quad H''(u)=\sum_{j=1}^m\widetilde{w}_jI\left(\widetilde{u}_j\leq u\right),\tag{A.64}$$

where $\max\limits_{1\leq j\leq m}|\widetilde{u}_j-u_j|\leq\eta$ and $|\sum_{j=1}^m(w_j-\widetilde{w}_j)|\leq\eta$. Then, we obtain

$$\|p_H-p_{H'}\|_{\infty,B}\leq C_1^*\eta,\quad \|g_H-g_{H'}\|_{\infty,B}\leq C_2^*\eta,$$

for $C_1^*=\|\sup\limits_u\frac{\partial}{\partial u}\{p(x|u)\}\|_{\infty,B}$ and $C_2^*=\|\sup\limits_u\frac{\partial}{\partial u}\{(x-u)p(x|u)\}\|_{\infty,B}$, where C_1^* is uniformly bounded. Next, we consider the following equation,

$$\begin{aligned}
&\frac{\partial}{\partial u}\{(x-u)p(x|u)\}\\
=\;&\begin{cases}\exp(-u)(u-1) & x=0,\\ -(2x+1)p(x|u)+(x+1)p(x+1|u)+xp(x-1|u) & x\geq 1,\end{cases}
\end{aligned}$$

which implies $C_2^* = O(M)$ provided $x \le B \le M$. Therefore, we obtain

$$\begin{aligned}\|t_H^*(x;\rho) - t_{H'}^*(x;\rho)\|_{\infty,B} &\le \frac{1}{\rho}\|g_H(x) - g_{H'}(x)\|_{\infty,B} \\ &\quad + \frac{L(\rho;M)}{\rho}\|p_H(x) - p_{H'}(x)\|_{\infty,B} \\ &\le \frac{\eta}{\rho}\left(C_2^* + C_1^* L(\rho;b)\right). \qquad \text{(A.65)}\end{aligned}$$

Furthermore, from H' and H'' in (A.64), we also have

$$\|p_{H'} - p_{H''}\|_{\infty,B} \le \eta, \quad \|g_{H'} - g_{H''}\|_{\infty,B} \le C_3^* \eta,$$

where $C_3^* = \|\sup_u \{g(x|u)\}\|_{\infty,B}$ for

$$g(x|u) = \{(x-u)p(x|u)\} = \begin{cases} -u\exp(-u) & x = 0, \\ xp(x|u) - (x+1)p(x+1|u) & x \ge 1. \end{cases}$$

Therefore, we obtain $C_3^* = O(M)$ provided $x \le B \le M$.

Also,

$$\begin{aligned}\|t_{H'}^*(x;\rho) - t_{H''}^*(x;\rho)\|_{\infty,B} &\le \frac{1}{\rho}\|g_{H'}(x) - g_{H''}(x)\|_{\infty,B} \\ &\quad + \frac{L(\rho;M)}{\rho}\|p_{H'}(x) - p_{H''}(x)\|_{\infty,B} \\ &\le \frac{\eta}{\rho}\left(C_3^* + L(\rho;M)\right). \qquad \text{(A.66)}\end{aligned}$$

Using triangular inequality for $||\cdot||_{\infty,B}$ and equations (A.63), (A.65) and (A.66), we derive

$$\begin{aligned} & \|t_G^*(x;\rho) - t_{H''}^*(x;\rho)\|_{\infty,B} \\ \le\ & \|t_G^*(x;\rho) - t_H(x;\rho)\|_{\infty,B} + \|t_H(x;\rho) - t_{H'}^*(x;\rho)\|_{\infty,B} \\ & + \|t_{H'}^*(x;\rho) - t_{H''}^*(x;\rho)\|_{\infty,B} \\ \le\ & \frac{\eta}{\rho}\left\{C_2^* + C_3^* + C_2(-\log\eta) + L(\rho;M)(C_1 + C_1^* + 1)\right\} \\ \le\ & D^* \frac{\eta}{\rho}\left(\log\frac{1}{\eta} + L(\rho;M)\right) \\ \equiv\ & \eta^*, \qquad \text{(A.67)}\end{aligned}$$

for a universal constant D^*, where the last inequality holds since we use the fact that $M \asymp \log\frac{1}{\eta}$, C_1, C_2, and C_1^* are uniformly bounded and $C_2^* = O(M)$ and $C_3^* = O(M)$.

Also, as in Lemma 4.7, H'' can be chosen from the set with cardinality asymptotically less than $\frac{1}{\eta}\left(\frac{M}{\eta}\right)^m \times \left(\frac{5}{\eta}\right)^m$ and $m \asymp \left(\log\frac{1}{\eta}\right)$ leading to

$$\log N\left(\eta^*, \mathcal{T}_\rho, \|\cdot\|_{\infty,B}\right) \quad \lesssim \quad \left\{\log\frac{1}{\eta}\right\}^2, \tag{A.68}$$

provided $M \asymp \log\frac{1}{\eta}$.

□

Acknowledgments

We are grateful to two referees and the editor, whose valuable suggestions and comments have greatly improved the presentation of the paper. Research of J. Park was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A2C1A01100526).

References

[1] Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Nature Precedings* 1–1.
[2] Brown, L. D. and Greenshtein, E. (2009). Nonparametric Empirical Bayes and Compound Decision Approaches to Estimation of a High-Dimensional Vector of Normal Means. *The Annals of Statistics* **37** 1685–1704. MR2533468
[3] Brown, L. D., Greenshtein, E. and Ritov, Y. (2013). The Poisson Compound Decision Problem Revisited. *Journal of the American Statistical Association* **108** 741–749. MR3174656
[4] Choi, H., Gim, J., Won, S., Kim, Y. J., Kwon, S. and Park, C. (2017). Network analysis for count data with excess zeros. *BMC genetics* **18** 1–10.
[5] Clevenson, M. L. and Zidek, J. V. (1975). Simultaneous Estimation of the Means of Independent Poisson Laws. *Journal of the American Statistical Association* **70** 698–705. MR0394962
[6] Dicker, L. H. and Zhao, S. D. (2016). High-dimensional classification via nonparametric empirical Bayes and maximum likelihood inference. *Biometrika* **103** 21–34. MR3465819
[7] Efron, B. (2003). Robbins, empirical Bayes and microarrays. *The Annals of Statistics* **31** 366–378. MR1983533
[8] Efron, B. (2014). Two modeling strategies for empirical Bayes estimation. *Statistical science: a review journal of the Institute of Mathematical Statistics* **29** 285–301. MR3264543
[9] Feng, L. and Dicker, L. H. (2018). Approximate nonparametric maximum likelihood for mixture models: A convex optimization approach to fitting arbitrary multivariate mixing distributions. *Computational Statistics & Data Analysis* **122** 80–91. MR3765816

[10] Gauran, I. I. M., Park, J., Lim, J., Park, D., Zylstra, J., Peterson, T., Kann, M. and Spouge, J. L. (2018). Empirical null estimation using zero-inflated discrete mixture distributions and its application to protein domain data. *Biometrics* **74** 458–471. MR3825332

[11] Gu, J. and Koenker, R. (2017). Rebayes: An R package for empirical Bayes mixture methods cemmap working paper No. CWP37/17, London.

[12] Hudson, H. M. and Tsui, K.-W. (1981). Simultaneous Poisson Estimators for a Priori Hypotheses about Means. *Journal of the American Statistical Association* **76** 182–187. MR0608191

[13] Jiang, W. (2020). On general maximum likelihood empirical Bayes estimation of heteroscedastic IID normal means. *Electronic Journal of Statistics* **14** 2272–2297. MR4109006

[14] Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics* **37** 1647–1684. MR2533467

[15] Jiang, W. and Zhang, C.-H. (2016). GENERALIZED LIKELIHOOD RATIO TEST FOR NORMAL MIXTURES. *Statistica Sinica* **26** 955–978. MR3559938

[16] Kiefer, J. and Wolfowitz, J. (1956). Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters. *The Annals of Mathematical Statistics* **27** 887–906. MR0086464

[17] Kimber, A. (1983). A note on Poisson maxima. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **63** 551–552. MR0705624

[18] Koenker, R. and Mizera, I. (2014). Convex Optimization, Shape Constraints, Compound Decisions, and Empirical Bayes Rules. *Journal of the American Statistical Association* **109** 674–685. MR3223742

[19] Maritz, J. S. (1969). Empirical Bayes estimation for the Poisson distribution. *Biometrika* **56** 349–359. MR0258180

[20] Park, J. (2012). Nonparametric empirical Bayes estimator in simultaneous estimation of Poisson means with application to mass spectrometry data. *Journal of Nonparametric Statistics* **24** 245-265. MR2885836

[21] Peng, J. (1975). Simultaneous Estimation of the Parameters of Independent Poisson Distributions Technical Report No. No. 78, Stanford University, Department of Statistics. MR2625889

[22] Peterson, T. A., Gauran, I. I. M., Park, J., Park, D. and Kann, M. G. (2017). Oncodomains: A protein domain-centric framework for analyzing rare variants in tumor samples. *PLOS Computational Biology* **13** e1005428.

[23] Robbins, H. (1956). An Empirical Bayes Approach to Statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* 157–163. University of California Press, Berkeley, Calif. MR0084919

[24] Robbins, H. (1977). Prediction and estimation for the compound Poisson distribution. *Proceedings of the National Academy of Sciences* **74** 2670–2671. MR0451479

[25] Robbins, H. et al. (1951). Asymptotically subminimax solutions of com-

pound statistical decision problems. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*. The Regents of the University of California. MR0044803

[26] Saha, S. and Guntuboyina, A. (2020). On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising. *The Annals of Statistics* **48** 738–762. MR4102674

[27] van de Geer, S. (2003). Asymptotic theory for maximum likelihood in nonparametric mixture models. *Computational Statistics & Data Analysis* **41** 453–464. MR1973724

[28] van der Vaart, A. W. and Wellner, J. (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media. MR1385671

[29] Weinstein, A., Ma, Z., Brown, L. D. and Zhang, C.-H. (2018). Group-Linear Empirical Bayes Estimates for a Heteroscedastic Normal Mean. *Journal of the American Statistical Association* **113** 698–710. MR3832220

[30] Wong, W. H. and Shen, X. (1995). Probability Inequalities for Likelihood Ratios and Convergence Rates of Sieve MLES. *The Annals of Statistics* **23** 339–362. MR1332570

[31] Xie, X., Kou, S. C. and Brown, L. D. (2012). SURE Estimates for a Heteroscedastic Hierarchical Model. *Journal of the American Statistical Association* **107** 1465–1479. MR3036408

[32] Zhang, C.-H. (1997). EMPIRICAL BAYES AND COMPOUND ESTIMATION OF NORMAL MEANS. *Statistica Sinica* **7** 181–193. MR1441153

[33] Zhang, C.-H. (2003). Compound decision theory and empirical Bayes methods: invited paper. *The Annals of Statistics* **31** 379–390. MR1983534

[34] Zhang, C.-H. (2009). GENERALIZED MAXIMUM LIKELIHOOD ESTIMATION OF NORMAL MIXTURE DENSITIES. *Statistica Sinica* **19** 1297–1318. MR2536157