Journal of Machine Learning Research 22 (2021) 1-46 Submitted 10/19; Published 4/21

A General Framework for Empirical Bayes Estimation in Discrete Linear Exponential Family

Trambak Banerjee TRAMBAK@KU.EDU
Analytics, Information and Operations Management
University of Kansas
Lawrence, KS 66045, USA

Qiang Liu LQIANG@CS.UTEXAS.EDU
Computer Science
University of Texas at Austin
Austin, Texas 78712, USA

Gourab Mukherjee GOURAB@USC.EDU
Data Sciences and Operations
University of Southern California
Los Angeles, CA 90089, USA

Wenguang Sun WENGUANS@MARSHALL.USC.EDU
Data Sciences and Operations
University of Southern California
Los Angeles, CA 90089, USA

Editor: Jon McAuliffe

Abstract

We develop a Nonparametric Empirical Bayes (`NEB`) framework for compound estimation in the discrete linear exponential family, which includes a wide class of discrete distributions frequently arising from modern big data applications. We propose to directly estimate the Bayes shrinkage factor in the generalized Robbins' formula via solving a convex program, which is carefully developed based on a RKHS representation of the Stein's discrepancy measure. The new `NEB` estimation framework is flexible for incorporating various structural constraints into the data driven rule, and provides a unified approach to compound estimation with both regular and scaled squared error losses. We develop theory to show that the class of `NEB` estimators enjoys strong asymptotic properties. Comprehensive simulation studies as well as analyses of real data examples are carried out to demonstrate the superiority of the `NEB` estimator over competing methods.

Keywords: Asymptotic Optimality; Empirical Bayes; Power Series Distributions; Shrinkage estimation; Stein's discrepancy

1. Introduction

Shrinkage methods, exemplified by the seminal work of James and Stein (1961), have received renewed attention in modern large-scale inference problems (Efron, 2012; Fourdrinier et al., 2018). Under this setting, the classical Normal means problem has been extensively studied (Brown, 2008; Jiang and Zhang, 2009; Brown and Greenshtein, 2009; Efron, 2011; Xie et al., 2012; Weinstein et al., 2018). However, in a variety of applications, the observed data are often discrete. For instance, in the News Popularity study discussed in Section 5, the goal is to estimate the popularity of a large number of news items based on their frequencies of being shared in social media platforms such as Facebook and LinkedIn. Another important application scenario arises from genomics research, where estimating the expected number of mutations across a large number of genomic locations can help identify key drivers or inhibitors of a given phenotype of interest.

We mention two main limitations of existing shrinkage estimation methods. First, the methodology and theory developed for continuous variables, in particular for Normal means problem, may not be directly applicable to discrete models. Second, existing methods have focused on the squared error loss. However, the scaled loss (Clevenson and Zidek, 1975), which effectively reflects the asymmetries in decision making [cf. Equation (3)], is a more desirable choice for many discrete models such as Poisson, where the scaled loss corresponds to the local Kulback-Leibler distance. The scaled loss also provides a more desirable criterion in a range of sparse settings, for example, when the goal is to estimate the rates of rare outcomes in Binomial distributions (Fourdrinier and Robert, 1995). Much research is needed for discrete estimation problems under various loss functions. This article develops a general framework for empirical Bayes estimation for the discrete linear exponential (`DLE`) family, also known as the family of discrete power series distributions (Noack, 1950), under both regular and scaled squared error losses.

The `DLE` family includes a wide class of popular members such as the Poisson, Binomial, Negative Binomial and Geometric distributions. Let Y be a non-negative integer valued random variable. Then Y is said to belong to a `DLE` family if its probability mass function (pmf) is of the form

$$p(y|\theta) = \frac{a_y \theta^y}{g(\theta)}, \quad y \in \{0, 1, 2, \cdots\}, \tag{1}$$

where a_y and $g(\theta)$ are known functions such that $a_y \geq 0$ is independent of θ and $g(\theta)$ is a normalizing factor that is differentiable at every θ. Special cases of `DLE` include the Poisson(λ) distribution with $a_y = (y!)^{-1}$, $\theta = \lambda$ and $g(\theta) = \exp(\theta)$, and the Binomial(m, q) distribution with $a_y = \binom{m}{y}$, $\theta = q/(1-q)$ and $g(\theta) = (1+\theta)^m$. Suppose $Y_1, \ldots, Y_n$ obey the following hierarchical model

$$Y_i \mid \theta_i \overset{ind.}{\sim} \texttt{DLE}(\theta_i), \quad \theta_i \overset{i.i.d}{\sim} G(\cdot), \tag{2}$$

where $G(\cdot)$ is an unspecified prior distribution on θ_i. The problem of interest is to estimate $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ based on $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. Empirical Bayes (EB) approaches to this compound decision problem date back to the famous Robbins' formula (Robbins, 1956) under the Poisson model. In the terminology of Efron (2014, 2019) there are two main modeling strategies for such EB estimation, namely, g-modeling and f-modeling strategies. The main

goal under g-modeling is to model the prior distribution G of θ using, for example, Non-parametric Maximum Likelihood estimation (NPMLE) techniques (Kiefer and Wolfowitz, 1956; Laird, 1978) or by modeling G as a low dimensional exponential family (Efron, 2016). With an estimate $\hat{G}$ of G, one can then derive an estimate of θ by plugging $\hat{G}$ into the Bayes rule for various loss functions (see for example Jiang and Zhang (2009); Koenker and Mizera (2014); Gu and Koenker (2017)). The f-modeling strategy, on the other hand, first starts from a particular form of the Bayes rule and then directly estimates the unknown marginal pmf $p(\cdot)$ of Y using, for instance, the observed empirical frequencies (Robbins, 1956), the smoothness-adjusted estimator of Brown et al. (2013), kernel density estimation techniques (Brown and Greenshtein, 2009) or through maximum likelihood estimation in flexible exponential family models (Efron, 2012).

This article develops a general non-parametric empirical Bayes (`NEB`) framework for compound estimation in discrete models. We first derive generalized Robbins' formula (`GRF`) for the `DLE` model (2), and then implement `GRF` via solving a convex program which is carefully developed based on a reproducing kernel Hilbert space (RKHS) representation of Stein's discrepancy measure and leads to a class of efficient `NEB` shrinkage estimators. Our work is related to the aforementioned f-modeling strategy however, in contrast with existing f-modeling approaches that estimate $p(y)$, the proposed `NEB` estimation framework directly produces estimates of Bayes shrinkage factors that are ratios of the marginal pmf $p(y)$ and appear in the `GRF` for the `DLE` model (2). We develop theories to show that the `NEB` estimator is $\sqrt{n}$ consistent up to certain logarithmic factors and enjoys superior risk properties. Simulation studies are conducted to illustrate that the efficiency gain of the `NEB` estimator over existing approaches, such as Brown et al. (2013), Koenker and Mizera (2014); Koenker and Gu (2017), Efron (2016), is substantial in many settings.

There are several advantages of the proposed `NEB` estimation framework. First, in contrast with existing methods such as the smoothness-adjusted Poisson estimator in Brown et al. (2013), our methodology covers a much wider range of distributions and presents a unified approach to compound estimation in discrete models. Second, our proposed convex program directly produces stable estimates of optimal Bayes shrinkage factors and can easily incorporate various structural constraints into the decision rule. By contrast, the three-step estimator in Brown et al. (2013), which involves smoothing, Rao-Blackwellization and monotonicity adjustments, is complicated, computationally intensive and sometimes unstable (as the numerator and denominator of the ratio are computed separately). Third, the RKHS representation of Stein's discrepancy measure provides a new analytical tool for developing theories such as asymptotic optimality and convergence rates. Finally, the `NEB` estimation framework is robust to departures from the true model due to its utilization of a generic quadratic program that does not rely on the specific form of a particular `DLE` family. Our numerical results in Section 4 demonstrate that the `NEB` estimator has a better risk performance than competitive approaches of Efron (2011), Brown et al. (2013) and Efron (2016) under a mis-specified Poisson model.

An alternative approach to compound estimation in discrete models, as suggested and investigated by Brown et al. (2013), is to employ variance stabilizing transformations, which converts the discrete problem to a classical normal means problem. This allows estimation via Tweedie's formula for normal variables (Efron, 2011). However, there are several drawbacks of this approach compared to our `NEB` framework. First, Tweedie's formula is not

applicable to scaled error loss whereas our methodology is built upon the generalized Robbins' formula, which covers both regular and scaled squared error losses. Second, there can be information loss in conventional data processing steps such as standardization, transformation and continuity approximation. While investigating the impact of information loss on compound estimation is of great interest, it is desirable to develop methodologies directly based on generalized Robbins' formula that is specifically derived and tailored for discrete variables. Finally, our `NEB` framework provides a convenient tool for developing asymptotic theories. By contrast, convergence rates are yet to be developed for normality inducing transformations, which can be highly non-trivial.

The rest of the paper is organized as follows. In Section 2, we introduce our estimation framework while Section 3 presents a theoretical analysis of the `NEB` estimator. The numerical performance of our method is investigated using both simulated and real data in Sections 4 and 5 respectively. Section 6 concludes with a discussion. Additional technical details and proofs are relegated to the Appendices.

2. A General Framework for Compound Estimation in `DLE` Family

This section describes the proposed `NEB` framework for compound estimation in discrete models. We first introduce in Section 2.1 the generalized Robbins' formula for the `DLE` family (2), then propose in Section 2.2 a convex optimization approach for its practical implementation. Details regarding tuning parameter selection are discussed in Section 2.3.

2.1 Generalized Robbins' formula for `DLE` models

Denote $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$ to be an estimator of $\boldsymbol{\theta}$ based on $\boldsymbol{Y}$. Consider a class of loss functions

$$\ell^{(k)}(\theta_i, \delta_i) = \theta_i^{-k}(\theta_i - \delta_i)^2 \tag{3}$$

for $k \in \{0, 1\}$, where $\ell^{(0)}(\theta_i, \delta_i)$ is the usual squared error loss, and $\ell^{(1)}(\theta_i, \delta_i) = \theta_i^{-1}(\delta_i - \theta_i)^2$ corresponds to the scaled squared error loss (Clevenson and Zidek, 1975; Fourdrinier and Robert, 1995). In compound estimation, one is concerned with the average loss

$$\mathcal{L}_n^{(k)}(\boldsymbol{\theta}, \boldsymbol{\delta}) = n^{-1} \sum_{i=1}^{n} \ell^{(k)}(\theta_i, \delta_i).$$

The associated risk is denoted $\mathcal{R}_n^{(k)}(\boldsymbol{\theta}, \boldsymbol{\delta}) = \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\theta}} \mathcal{L}_n^{(k)}(\boldsymbol{\theta}, \boldsymbol{\delta})$. Let $\boldsymbol{G}(\boldsymbol{\theta})$ denote the joint distribution of $(\theta_1, \cdots, \theta_n)$. The Bayes estimator $\boldsymbol{\delta}_{(k)}^{\pi}$ that minimizes the Bayes risk $B_n^{(k)}(\boldsymbol{\theta}) = \int \mathcal{R}_n^{(k)}(\boldsymbol{\theta}, \boldsymbol{\delta}) \mathrm{d}\boldsymbol{G}(\boldsymbol{\theta})$ is given by Lemma 1.

Lemma 1 (Generalized Robbins' formula)**.** *Consider the `DLE` Model* (2)*. Let* $p(\cdot) = \int p(\cdot|\theta) \mathrm{d}G(\theta)$ *be the marginal pmf of* Y*. Define for* $k \in \{0, 1\}$*,*

$$w_p^{(k)}(y_i) = \frac{p(y_i - k)}{p(y_i + 1 - k)}, \text{ for } y_i = k, k+1, \cdots.$$

Then the Bayes estimator that minimizes the risk $B_n^{(k)}(\boldsymbol{\theta})$ *is given by* $\boldsymbol{\delta}_{(k)}^{\pi} = \{\delta_{(k),i}^{\pi}(y_i) : 1 \leq i \leq n\}$*, where*

$$\delta_{(k),i}^{\pi}(y_i) = \begin{cases} \dfrac{a_{y_i-k}/a_{y_i+1-k}}{w_p^{(k)}(y_i)}, & \text{for } y_i = k, k+1, \cdots \\ 0, & \text{for } y_i < k \end{cases}. \tag{4}$$

Remark 1. Under the squared error loss ($k = 0$) with $Y_i \mid \theta_i \sim \text{Poi}(\theta_i)$ and $a_{y_i} = (y_i!)^{-1}$, Lemma 1 yields

$$\delta_{(0),i}^{\pi}(y_i) = (y_i + 1)\frac{p(y_i + 1)}{p(y_i)}, \tag{5}$$

which recovers the classical Robbins' formula (Robbins, 1956). In contrast, under the scaled loss, we have

$$\delta_{(1),i}^{\pi}(y_i) = y_i \frac{p(y_i)}{p(y_i - 1)} \text{ for } y_i > 0 \text{ and } \delta_{(1),i}^{\pi}(y_i) = 0 \text{ otherwise.} \tag{6}$$

Under scaled error loss the estimator (5) can be much outperformed by (6) (and vice versa under the regular loss). We develop parallel results for the two types of loss functions.

Next we discuss related works for implementing Robbins' formula under the empirical Bayes (EB) estimation framework. Inspecting (4) and (5), we can view a_{y_i-k}/a_{y_i+1-k} as a naive and known estimator of θ_i. The ratio functional $w_p^{(k)}(y_i)$, which is unknown in practice, represents the optimal shrinkage factor that depends on $p(\cdot)$. Hence under the f-modeling strategy a simple EB approach, as done in the classical Robbins' formula, is to estimate $w_p^{(k)}(y)$ by plugging-in empirical frequencies: $\hat{w}_n^{(0)}(y) = \hat{p}_n(y)/\hat{p}_n(y+1)$, where $\hat{p}_n(y) = n^{-1}\sum_{i=1}^n \mathbb{I}(y_i = y)$. It is noted by Brown et al. (2013) that this plug-in estimator can be highly inefficient especially when θ_i are small. Moreover, the numerator and denominator in $w_p^{(0)}(y)$ are estimated separately, which may lead to unstable ratios. Brown et al. (2013) showed that Robbins' formula can be dramatically improved by imposing additional smoothness and monotonicity adjustments. An alternative approach is to estimate G using NPMLE under appropriate shape constraints. However, efficient estimation of G may not directly translate into an efficient estimation of the ratio functional $w_p^{(k)}(y)$. We recast the compound estimation problem as a convex program, which directly produces consistent estimates of the ratio functionals

$$\boldsymbol{w}_p^{(k)} = \left\{w_p^{(k)}(y_1), \ldots, w_p^{(k)}(y_n)\right\}$$

from data. The estimators are shown to enjoy superior numerical and theoretical properties. Unlike existing f-modeling works that are limited to squared loss and specific members in the `DLE` family, our method can handle a wide range of discrete distributions and various types of loss functions in a unified framework.

2.2 Shrinkage estimation by convex optimization

This section focuses on the scaled squared error loss ($k = 1$). Methodologies and theories for the case with the squared error loss ($k = 0$) can be derived similarly; details are provided

in Appendix A.1. We first introduce some notations and then present the NEB estimator in Definition 1.

Suppose Y is a non-negative integer-valued random variable with pmf $p(\cdot)$. Define

$$h_0^{(1)}(y) = \begin{cases} 1\ , & \text{if } y = 0 \\ 1 - w_p^{(1)}(y), & \text{if } y \in \{1, 2, \ldots\}. \end{cases} \tag{7}$$

Let $\mathcal{K}_\lambda(y, y') = \exp\{-\frac{1}{2\lambda}(y - y')^2\}$ be the positive definite Gaussian kernel with bandwidth parameter $\lambda \in \Lambda$ where Λ is a compact subset of $\mathbb{R}^+$ bounded away from 0. Given observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ from model (2), let $\boldsymbol{h}_0^{(1)} = \left\{h_0^{(1)}(y_1), \ldots, h_0^{(1)}(y_n)\right\}$. Define operators $\Delta_y \mathcal{K}_\lambda(y, y') = \mathcal{K}_\lambda(y + 1, y') - \mathcal{K}_\lambda(y, y')$ and

$$\Delta_{y,y'} \mathcal{K}_\lambda(y, y') = \Delta_{y'} \Delta_y \mathcal{K}_\lambda(y, y') = \Delta_y \Delta_{y'} \mathcal{K}_\lambda(y, y').$$

Consider the following $n \times n$ matrices, which are needed in the definition of the NEB estimator:

$$\boldsymbol{K}_\lambda = n^{-2}[\mathcal{K}_\lambda(y_i, y_j)]_{ij}, \quad \Delta \boldsymbol{K}_\lambda = n^{-2}[\Delta_{y_i} \mathcal{K}_\lambda(y_i, y_j)]_{ij}, \quad \Delta_2 \boldsymbol{K}_\lambda = n^{-2}[\Delta_{y_i, y_j} \mathcal{K}_\lambda(y_i, y_j)]_{ij}.$$

Definition 1 (NEB estimator)**.** *Consider the DLE model* (2) *with loss* $\ell^{(1)}(\theta_i, \delta_i)$*. For any fixed* $\lambda \in \Lambda$*, let* $\hat{\boldsymbol{h}}_n^{(1)}(\lambda) = \left\{\hat{h}_1^{(1)}(\lambda), \ldots, \hat{h}_n^{(1)}(\lambda)\right\}$ *be the solution to the following quadratic optimization problem:*

$$\min_{\boldsymbol{h} \in \boldsymbol{H}_n} \boldsymbol{h}^T \boldsymbol{K}_\lambda \boldsymbol{h} + 2\boldsymbol{h}^T \Delta \boldsymbol{K}_\lambda \boldsymbol{1} + \boldsymbol{1}^T \Delta_2 \boldsymbol{K}_\lambda \boldsymbol{1}, \tag{8}$$

where $\boldsymbol{H}_n = \{\boldsymbol{h} = (h_1, \ldots, h_n) : \mathcal{A}\boldsymbol{h} \preceq \boldsymbol{b},\ \mathcal{C}\boldsymbol{h} = \boldsymbol{d}\}$ *is a convex set and* $\mathcal{A}, \mathcal{C}, \boldsymbol{b}$ *and* $\boldsymbol{d}$ *are known real matrices and vectors that enforce linear constraints on the components of* $\boldsymbol{h}$*. Define* $\hat{w}_i^{(1)}(\lambda) = 1 - \hat{h}_i^{(1)}(\lambda)$*. Then the NEB estimator is given by* $\boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\lambda) = \left\{\delta_{(1),i}^{\mathsf{neb}}(\lambda) : 1 \le i \le n\right\}$*, where*

$$\delta_{(1),i}^{\mathsf{neb}}(\lambda) = \frac{a_{y_i - 1}/a_{y_i}}{\hat{w}_i^{(1)}(\lambda)}, \qquad \text{if } y_i \in \{1, 2, \ldots\},$$

and $\delta_{(1),i}^{\mathsf{neb}}(\lambda) = 0$ *if* $y_i = 0$*.*

Remark 2. In problem (8) the linear inequality constraints $\mathcal{A}\boldsymbol{h} \preceq \boldsymbol{b}$ can be used to impose structural constraints on the NEB decision rule $\boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\lambda)$. These structural constraints, which may take the form of monotonicity constraints (Brown et al., 2013; Koenker and Mizera, 2014), have been shown to be effective for stabilizing the estimator and hence improving the accuracy. For instance, a monotonicity constraint on $\delta_{(1),i}^{\mathsf{neb}}(\lambda)$ will imply $\delta_{(1),(1)}^{\mathsf{neb}}(\lambda) \ge \cdots \ge \delta_{(1),(n)}^{\mathsf{neb}}(\lambda)$ for $y_{(1)} \ge y_{(2)} \ge \cdots \le y_{(n)}$. In particular, when $Y_i \mid \theta_i \sim \mathrm{Poi}(\theta_i)$ then $\delta_{(1),i}^{\mathsf{neb}}(\lambda) = y_i/\{1 - \hat{h}_i^{(1)}(\lambda)\}$ and the monotonicity constraints in this setting will imply

$$-\hat{h}_{(i)}^{(1)}(\lambda) + \frac{y_{(i)}}{y_{(i+1)}} \hat{h}_{(i+1)}^{(1)}(\lambda) \le \frac{y_{(i)}}{y_{(i+1)}} - 1, \text{ for } 1 \le i \le (n-1)$$

These $n-1$ linear inequality constraints may be imposed with an $(n-1)\times n$ matrix $\mathcal{A}$ and an $n-1$ column vector $\boldsymbol{b}$ such that for $1 \le i \le (n-1)$ and $1 \le r \le n$,

$$\begin{aligned} \mathcal{A}(i,r) &= \begin{cases} -1, \text{ when } y_r = y_{(i)} \\ y_{(i)}/y_{(i+1)}, \text{ when } y_r = y_{(i+1)} \\ 0, \text{ otherwise} \end{cases} \\ \text{and } b_i &= y_{(i)}/y_{(i+1)} - 1. \end{aligned}$$

Moreover, when $y_i = 0$ we set $\delta^{\mathsf{neb}}_{(1),i}(\lambda) = 0$ by convention (see lemma 1). The equality constraints $\mathcal{C}\boldsymbol{h} = \boldsymbol{d}$ accommodate such boundary conditions along with instances of ties for which we require $\hat{h}^{(1)}_i(\lambda) = \hat{h}^{(1)}_j(\lambda)$ whenever $y_i = y_j$.

Next we provide some insights on why the optimization criterion (8) works; theories are developed in Section 3 to establish the properties of the `NEB` estimator rigorously. Denote $h^{(1)}_0$ and $\tilde{h}^{(1)}$ as the ratio functionals corresponding to pmfs p and $\tilde{p}$, respectively and let $\hat{\mathbb{M}}_{\lambda,n}(\boldsymbol{h}) = \boldsymbol{h}^T\boldsymbol{K}_\lambda\boldsymbol{h} + 2\boldsymbol{h}^T\Delta\boldsymbol{K}_\lambda\mathbf{1} + \mathbf{1}^T\Delta_2\boldsymbol{K}_\lambda\mathbf{1}$. Suppose Y_i are i.i.d. samples obeying $p(y)$. Theorem 1 shows that

$$\hat{\mathbb{M}}_{\lambda,n}(\tilde{\boldsymbol{h}}) = \mathbb{M}_\lambda(\tilde{\boldsymbol{h}}) + O_p\Big(\frac{\log^2 n}{n^{1/2}}\Big),$$

where $\hat{\mathbb{M}}_{\lambda,n}(\tilde{\boldsymbol{h}})$ is the objective function in (8) and $\mathbb{M}_\lambda(\tilde{\boldsymbol{h}})$, also denoted $\mathcal{S}_\lambda[\tilde{p}](p)$, is the kernelized Stein's discrepancy (KSD). Roughly speaking, the KSD measures how different one distribution p is from another distribution $\tilde{p}$, with $\mathcal{S}_\lambda[\tilde{p}](p) = 0$ if and only if $\tilde{p} = p$. A key feature of the KSD is that $\mathcal{S}_\lambda[\tilde{p}](p)$ can be equivalently represented by the discrepancy between the corresponding ratio functionals $h^{(1)}_0$ and $\tilde{h}^{(1)}$. Hence, optimizing (8) is asymptotically equivalent to finding $\tilde{h}^{(1)}$ that is as close as possible to the true underlying $h^{(1)}_0$, which corresponds to the optimal shrinkage factor in the compound estimation problem. Theorems 2 and 3 demonstrate that (8) is an effective convex program in the sense that the minimizer $\hat{\boldsymbol{h}}_n$ is $\sqrt{n}$ consistent with respect to $\boldsymbol{h}^{(1)}_0$, and the resultant `NEB` estimator converges to the Bayes estimator.

2.3 Bandwidth selection

The implementation of the quadratic program in (8) requires the choice of a tuning parameter λ in the Gaussian kernel. For practical applications, λ must be determined in a data-driven fashion. For infinitely divisible random variables (Klenke, 2014) such as Poisson variables, Brown et al. (2013) proposed a modified cross validation (MCV) method for choosing the tuning parameter. However, the MCV method cannot be applied to distributions with bounded support as they are not infinitely divisible (Sato and Ken-Iti, 1999) such as the Binomial distribution. To provide a unified estimation framework for the `DLE` family, we develop an alternative method for choosing λ. The key idea is to derive an asymptotic risk estimate $\mathsf{ARE}^{(1)}_n(\lambda)$ that serves as an approximation to the true loss $\mathcal{L}^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda))$. Then the tuning parameter is chosen to minimize $\mathsf{ARE}^{(1)}_n(\lambda)$.

The methodology based on ARE is illustrated below under the scaled loss (see definition 2) and in Appendix A.2 we provide relevant details for choosing λ under the regular squared loss $\mathcal{L}^{(0)}_n$.

Definition 2 (ARE of $\boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda)$ in the DLE model)**.** *Suppose* $Y_i \mid \theta_i \overset{ind.}{\sim} \texttt{DLE}(\theta_i)$*. Under the loss* $\ell^{(1)}(\theta_i, \cdot)$*, an asymptotic risk estimate of the true loss of* $\boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda)$ *is*

$$\mathsf{ARE}^{(1)}_n(\lambda, \boldsymbol{y}) = \frac{1}{n}\Big\{\sum_{i=1}^n \psi_i(\lambda) - 2\sum_{i=1}^n \delta^{\mathsf{neb}}_{(1),i}(\lambda)\Big\}, \text{ where}$$

$$\psi_i(\lambda) = \{\delta^{\mathsf{neb}}_{(1),j_i}(\lambda)\}^2(a_{y_i+1}/a_{y_i}),\ y_i = 0, 1, \ldots.$$

with $j_i \in \{1, \ldots, n\}$ *such that* $y_{j_i} = y_i + 1$.

We propose the following estimate of the tuning parameter λ based on the $\mathsf{ARE}^{(1)}_n(\lambda, \boldsymbol{y})$:

$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda} \mathsf{ARE}^{(1)}_n(\lambda, \boldsymbol{y}) \tag{9}$$

In practice we recommend using $\Lambda = [10, 10^2]$, which worked well in all our simulations and real data analyses. In Section 3, we present Lemma 2 which provides asymptotic justifications for selecting λ using equation (9).

3. Theory

This section studies the theoretical properties for the `NEB` estimator under the Poisson and Binomial models. We first investigate the large-sample behavior of the KSD measure (Section 3.1), then turn to the performance of the estimated risk ratios $\hat{\boldsymbol{w}}_n$ (Section 3.2), and finally establish the consistency and risk properties of the proposed estimator $\boldsymbol{\delta}^{\mathsf{neb}}_{(1)}$ (Section 3.3). The accuracy of the ARE criteria, which are used in choosing tuning parameter λ, will also be investigated.

3.1 Theoretical properties of the KSD measure

To provide motivation and theoretical support for Definition 1, we introduce the Kernelized Stein's Discrepancy (KSD) (Liu et al., 2016; Chwialkowski et al., 2016) and discuss its connection to the quadratic program (8). While the KSD has been used in various contexts including goodness of fit tests (Liu et al., 2016; Yang et al., 2018), variational inference (Liu and Wang, 2016) and Monte Carlo integration (Oates et al., 2017), our theory on its connection to the compound estimation problem and empirical Bayes methodology is novel.

Assume that (Y, Y') are i.i.d. copies from the marginal pmf p. Consider h_0 defined in Equation (7)[1] and let $\tilde{p}$ denote a pmf on the support of Y, for which we similarly define $\tilde{h}$. The KSD, which is formally defined as

$$\mathcal{S}_\lambda[\tilde{p}](p) = \mathbb{E}_p\Big[\Big\{\tilde{h}(Y) - h_0(Y)\Big\}\mathcal{K}_\lambda(Y, Y')\Big\{\tilde{h}(Y') - h_0(Y')\Big\}\Big], \tag{10}$$

provides a discrepancy measure between p and $\tilde{p}$ in the sense that (a)

$$\mathcal{S}_\lambda[\tilde{p}](p) \geq 0 \text{ and } \mathcal{S}_\lambda[\tilde{p}](p) = 0 \text{ if and only if } p = \tilde{p},$$

1. In Section 3.1 we shall drop the superscript from $\boldsymbol{h}_0$, which is used to indicate whether the loss is scaled or regular. The simplification has no impact since the general idea holds for both types of losses and the discussion in this section focuses on the scaled loss.

and (b) informally, $\mathcal{S}_\lambda[\tilde{p}](p)$ tends to increase when there is a bigger disparity between h_0 and $\tilde{h}$ (or equivalently, between p and $\tilde{p}$).

The direct evaluation of $\mathcal{S}_\lambda[\tilde{p}](p)$ via Equation (10) is difficult because h_0 is unknown. Note that while the pmf p can be learned well from a random sample $\{Y_1, \ldots, Y_n\} \sim p$, we introduce an alternative representation of KSD, developed by Liu et al. (2016), in a reproducing kernel Hilbert space (RKHS) that does not directly involve unknown h_0. Concretely, consider a positive definite kernel function $\kappa_\lambda[\tilde{h}(u), \tilde{h}(v)]$ where

$$\kappa_\lambda[\tilde{h}(u), \tilde{h}(v)](u,v) = \tilde{h}(u)\tilde{h}(v)\mathcal{K}_\lambda(u,v) + \tilde{h}(u)\Delta_v\mathcal{K}_\lambda(u,v) + \tilde{h}(v)\Delta_u\mathcal{K}_\lambda(u,v) + \Delta_{u,v}\mathcal{K}_\lambda(u,v). \tag{11}$$

For i.i.d. copies (Y, Y') from distribution p, it can be shown that

$$\begin{aligned} \mathcal{S}_\lambda[\tilde{p}](p) &= \mathbb{E}_{(Y,Y') \overset{i.i.d.}{\sim} p}\Big[\kappa_\lambda[\tilde{h}(Y), \tilde{h}(Y')](Y, Y')\Big] \\ &= \frac{1}{n(n-1)}\mathbb{E}_p\Big[\sum_{1 \le i \neq j \le n} \kappa_\lambda[\tilde{h}(Y_i), \tilde{h}(Y_j)](Y_i, Y_j)\Big] \\ &\coloneqq \mathbb{M}_\lambda(\tilde{\boldsymbol{h}}), \end{aligned} \tag{12}$$

where $\{Y_1, \ldots, Y_n\}$ is a random sample from p. It can be similarly shown that $\mathbb{M}_\lambda(\tilde{\boldsymbol{h}}) = 0$ if and only if $\tilde{\boldsymbol{h}} = \boldsymbol{h}_0$. Substituting the empirical distribution $\hat{p}_n$ in place of the pmf p in (12), we obtain the following empirical evaluation scheme for $\mathcal{S}_\lambda[\tilde{p}](p)$

$$\mathcal{S}_\lambda[\tilde{p}](\hat{p}_n) = \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \kappa_\lambda[\tilde{h}(y_i), \tilde{h}(y_j)](y_i, y_j). \tag{13}$$

Note that (13) is exactly the objective function $\hat{\mathbb{M}}_{\lambda,n}(\tilde{\boldsymbol{h}})$ of the quadratic program (8).

The empirical representation of KSD (13) provides an extremely useful tool for solving the discrete compound decision problem under the EB estimation framework. A key observation is that the kernel function $\kappa_\lambda[\tilde{h}(u), \tilde{h}(v)](u,v)$ depends on $\tilde{p}$ only through $\tilde{h}$. Meanwhile, the EB implementation of the generalized Robbins' formula [cf. Equations (4) and (7)] essentially boils down to the estimation of h_0. Hence, if $\mathcal{S}_\lambda[\tilde{p}](\hat{p}_n)$ is asymptotically equal to $\mathcal{S}_\lambda[\tilde{p}](p)$, then minimizing $\mathcal{S}_\lambda[\tilde{p}](\hat{p}_n)$ with respect to the unknowns $\tilde{\boldsymbol{h}} = \Big\{\tilde{h}(y_1), \ldots, \tilde{h}(y_n)\Big\}$ is effectively the process of finding an $\tilde{h}$ that is as close as possible to h_0, which yields an asymptotically optimal solution to the EB estimation problem. Therefore our formulation of the `NEB` estimator $\boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda)$ would be justified as long as we can establish the asymptotic consistency of the sample criterion $\mathcal{S}_\lambda[\tilde{p}](\hat{p}_n)$ around the population criterion $\mathcal{S}_\lambda[\tilde{p}](p)$ uniformly over λ (Theorem 1).

Our analysis in this and the following sections will be based on the hierarchical model of equation (2): $Y_i \mid \theta_i \overset{ind.}{\sim} \mathtt{DLE}(\theta_i)$, $\theta_i \overset{i.i.d}{\sim} G(\cdot)$ where $G(\cdot)$ is an unspecified prior distribution on θ_i. In this setup the marginal pmf of Y is $p(y) \coloneqq P(Y = y) = \int p(y|\theta)\mathrm{d}G(\theta)$. We impose the following regularity conditions that are needed in our technical analysis.

(A1) $\mathbb{E}_p|\kappa_\lambda[\tilde{h}(U), \tilde{h}(V)](U,V)|^2 < \infty$ for all $\lambda \in \Lambda$ where Λ is a compact subset of $\mathbb{R}^+$ bounded away from 0.

(A2) For some $\epsilon \in (0,1)$, $\mathbb{E}_G\{\exp(\epsilon\theta)\} < \infty$ where the expectation is taken with respect to the prior distribution G of θ.

(A3) For any function g that satisfies $0 < \|g\|_2^2 < \infty$, there exists a constant $c > 0$ such that $\underline{\lim}_{n\to\infty} \sum_{y,y'=0}^{n} g(y)\mathcal{K}_\lambda(y,y')g(y') > c\|g\|_2^2$ for every $\lambda \in \Lambda$.

Remark 3. Assumption (A1) is a moment condition on the kernel function related to V-statistics; see, for example, Serfling (2009). Assumption (A2) is a moment condition on the prior distribution G. In particular, it ensures that with high probability $\max(Y_1,\ldots,Y_n) \leq \log n$ as $n \to \infty$. This idea is formalized in Lemma 4 in Appendix B. It is likely that assumption (A2) can be further relaxed but we do not seek the full generality here. Assumption (A3) is a standard condition which ensures that the KSD $\mathcal{S}_\lambda[\tilde{p}](p)$ is a valid discrepancy measure (Liu et al., 2016; Chwialkowski et al., 2016).

Theorem 1. *If $\tilde{p}$ is a probability mass function on the support of Y then, under Assumptions (A1) and (A2), we have*

$$\sup_{\lambda\in\Lambda} \Big|\hat{\mathbb{M}}_{\lambda,n}(\tilde{\boldsymbol{h}}) - \mathbb{M}_\lambda(\tilde{\boldsymbol{h}})\Big| = O_p\Big(\frac{\log^2 n}{\sqrt{n}}\Big).$$

In the context of our compound estimation framework, Theorem 1 is significant because it guarantees that the empirical version of the KSD measure given by $\hat{\mathbb{M}}_{\lambda,n}(\tilde{\boldsymbol{h}})$ is asymptotically close to its population counterpart $\mathbb{M}_\lambda(\tilde{\boldsymbol{h}})$ uniformly in $\lambda \in \Lambda$. Moreover, along with the fact that $\mathbb{M}_\lambda(\boldsymbol{h}_0) = 0$, Theorem 1 establishes that $\hat{\mathbb{M}}_{\lambda,n}(\boldsymbol{h})$ is the appropriate criteria to minimize with respect to $\boldsymbol{h} \equiv \tilde{\boldsymbol{h}}$. In Theorem 2, we further show that the resulting estimator of the ratio functionals $\boldsymbol{w}_p^{(1)}$ from equation (8) are consistent.

3.2 Theoretical properties of $\hat{\boldsymbol{w}}_n$

The optimization problem in (8) is defined over a convex set $\boldsymbol{H}_n = \{\boldsymbol{h} = (h_1,\ldots,h_n) : \mathcal{A}\boldsymbol{h} \preceq \boldsymbol{b},\ \mathcal{C}\boldsymbol{h} = \boldsymbol{d}\}$ which is a subset of $\mathbb{R}^n$. However, the dimension of $\boldsymbol{H}_n$, denoted by $\texttt{dim}(\boldsymbol{H}_n)$, is usually much smaller than n. Consider the Binomial case where $Y_i|q_i \sim \texttt{Bin}(m_i, q_i)$ with $q_i \in (0,1)$, $m_i \leq m < \infty$ and $\theta_i = q_i/(1-q_i)$. Here $\texttt{dim}(\boldsymbol{H}_n)$ is at most m since $\max(Y_1,\ldots,Y_n) \leq m$. While the boundedness of the support is not always available outside the Binomial case, in most practical applications it is reasonable to assume that the distribution of θ_i has some finite moments, which ensures that $\texttt{dim}(\boldsymbol{H}_n)$ grows slower than $\log n$; see Assumption (A2). In Lemma 4 we make this precise. Moreover, as discussed in remark 2, the linear inequality constraints $\mathcal{A}\boldsymbol{h} \preceq \boldsymbol{b}$ impose structural constraints on $\boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\lambda)$. For the ensuing discussion and following Brown et al. (2013), we let these structural constraints to take the form of monotonicity constraints on the `NEB` decision rule. Since the Binomial and the Poisson models have the monotone likelihood ratio property, $\boldsymbol{\delta}_{(1)}^{\pi}$ is monotone and so $\boldsymbol{h}_0 \in \boldsymbol{H}_n$. The next theorem establishes the asymptotic consistency of $\hat{\boldsymbol{w}}_n^{(1)}(\lambda)$.

Theorem 2. *Let $\mathcal{K}_\lambda(\cdot,\cdot)$ be the positive definite Gaussian kernel with bandwidth parameter $\lambda \in \Lambda$. If $\lim_{n\to\infty} c_n n^{-1/2}\log^2 n = 0$ then, under Assumptions (A1) - (A3), we have for any $\lambda \in \Lambda$,*

$$\lim_{n\to\infty} \mathbb{P}\left\{\frac{1}{n}\Big\|\hat{\boldsymbol{w}}_n^{(1)}(\lambda) - \boldsymbol{w}_p^{(1)}\Big\|_2^2 \geq c_n^{-1}\epsilon\right\} = 0,\ \textit{for any } \epsilon > 0,$$

where $\hat{\boldsymbol{w}}_n^{(1)}(\lambda) = 1 - \hat{\boldsymbol{h}}_n^{(1)}(\lambda)$.

Theorem 2 shows that under the scaled squared error loss, $\hat{\boldsymbol{w}}_n^{(1)}(\lambda)$, the optimizer of quadratic problem in equation (8), is a consistent estimator of $\boldsymbol{w}_p^{(1)}$, the optimal shrinkage factor in the Bayes rule (Lemma 1). In particular, the aforementioned consistency result is related to the theoretical analysis of minimum KSD estimators in Barp et al. (2019). While Barp et al. (2019) establish almost sure convergence and asymptotic normality of such minimum KSD estimators, the analysis in this section is geared towards studying the asymptotic optimality of the proposed NEB estimator in the sense of Theorem 3 below. The proof of Theorem 2 is available in Appendix B.3 which also includes relevant details for proving a companion result under the regular squared error loss.

Remark 4. The estimation framework in Definition 1 may be used for producing consistent estimators for any member in the DLE family. This allows the corresponding NEB estimator to cover a much wider class of discrete distributions than previously proposed. Compared to the existing methods of Efron (2011) and Brown et al. (2013), our proposed NEB estimation framework is robust against departures from the true data generating process. This is due to the fact that the quadratic optimization problem in (8) does not rely on the specific form of the distribution of $Y|\theta$, and that the shrinkage factors are estimated in a non-parametric fashion. The robustness of the estimator is corroborated by our numerical results in Section 4.

3.3 Properties of the NEB estimator

In this section we discuss the risk properties of the NEB estimator. Let

$$\rho_n^{(1)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\lambda)) = \mathcal{L}_n^{(1)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\lambda)) - \frac{1}{n}\sum_{i=1}^{n}\theta_i.$$

We begin with Lemma 2 which shows that uniformly in $\lambda \in \Lambda$, the gap between $\mathsf{ARE}_n^{(1)}(\lambda)$ and $\mathbb{E}\{\rho_n^{(1)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\lambda))\}$ is asymptotically negligible. This justifies our proposed methodology for choosing the tuning parameter λ in Section 2.3. In the following lemma, we let c_n be a sequence satisfying $\lim_{n\to\infty} c_n n^{-1/4}\log^3 n = 0$.

Lemma 2. *Under Assumptions (A1) - (A3), we have*

$$(1). \quad c_n \sup_{\lambda\in\Lambda}\left|\mathsf{ARE}_n^{(1)}(\lambda, \boldsymbol{Y}) - \rho_n^{(1)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\lambda))\right| = o_p(1).$$

$$(2). \quad c_n \sup_{\lambda\in\Lambda}\left|\mathsf{ARE}_n^{(1)}(\lambda, \boldsymbol{Y}) - \mathbb{E}\{\rho_n^{(1)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\lambda))\}\right| = o_p(1);$$

In Appendices B.5 and B.6 we prove Lemma 2 for the Binomial and Poisson models under both scaled squared error and squared error losses.

To analyze the quality of the data-driven bandwidth $\hat{\lambda}$ [cf. Equation (9)], we consider an oracle loss estimator $\boldsymbol{\delta}_{(1)}^{\mathsf{or}} := \boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\lambda_1^{\mathsf{orc}})$, where

$$\lambda_1^{\mathsf{orc}} := \arg\min_{\lambda\in\Lambda} \mathcal{L}_n^{(1)}\left\{\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\lambda)\right\}.$$

The oracle bandwidth λ_1^{orc} is not available in practice since it requires the knowledge of unknown $\boldsymbol{\theta}$. However, it provides a benchmark for assessing the effectiveness of the data-driven bandwidth selection procedure in Section 2.3. The following lemma shows that the loss of $\boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\hat{\lambda})$ converges in probability to the loss of $\boldsymbol{\delta}_{(1)}^{\mathsf{or}}$.

Lemma 3. *Under Assumptions (A1) - (A3), if* $\lim_{n\to\infty} c_n n^{-1/4}\log^2 n = 0$*, then for both the Poisson and Binomial models, we have*

$$\lim_{n\to\infty} \mathbb{P}\Big[\mathcal{L}_n^{(1)}\left\{\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\hat{\lambda})\right\} \geq \mathcal{L}_n^{(1)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\mathsf{or}}) + c_n^{-1}\epsilon\Big] = 0 \text{ for any } \epsilon > 0.$$

Obviously, the estimator $\boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\lambda_1^{\mathsf{orc}})$ is lower bounded by the risk of the optimal solution $\boldsymbol{\delta}_{(1)}^{\pi}$ (Lemma 1). Next we study the *asymptotic optimality* of $\boldsymbol{\delta}_{(1)}^{\mathsf{neb}}$, which aims to provide decision theoretic guarantees on $\boldsymbol{\delta}_{(1)}^{\mathsf{neb}}$ in relation to $\boldsymbol{\delta}_{(1)}^{\pi}$. Theorem 3 establishes the optimality theory by showing that (a) the average squared error between $\boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\hat{\lambda})$ and $\boldsymbol{\delta}_{(1)}^{\pi}$ is asymptotically small, and (b) the `NEB` estimator is asymptotically as good as the corresponding Bayes estimator in terms of expected loss.

Theorem 3. *Under the conditions of Theorem 2, if* $\lim_{n\to\infty} c_n n^{-1/2}\log^4 n = 0$*, then for both the Poisson and Binomial models, we have*

$$\frac{c_n}{n}\Big\|\boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\hat{\lambda}) - \boldsymbol{\delta}_{(1)}^{\pi}\Big\|_2^2 = o_p(1).$$

Furthermore, under the same conditions, we have,

$$\lim_{n\to\infty} \mathbb{E}\Big[\mathcal{L}_n^{(1)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\hat{\lambda})) - \mathcal{L}_n^{(1)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\pi})\Big] = 0.$$

In Appendix A.2, we discuss the counterpart to Theorem 3 under the squared error loss $\mathcal{L}_n^{(0)}$.

4. Numerical Results

In this section we first discuss, in Section 4.1, the implementation details of the convex program (8) and bandwidth selection process (9) (see also (19) in Appendix A.2). Then we investigate the numerical performance of the `NEB` estimator for Poisson, Binomial and Negative Binomial compound decision problems, respectively in Sections 4.2,4.3 and 4.4. In each case, we consider both regular and scaled squared losses. Our numerical results demonstrate that the efficiency gain of the `NEB` estimator over competitive methods is substantial in many settings.

We have developed an R package, `npeb`, to implement the `NEB` estimator in definition 1 (and definition 3 in Appendix A.1). Moreover, the R code that reproduces the numerical results in this section can be downloaded from the following link: `https://github.com/trambakbanerjee/DLE_paper`.

4.1 Implementation Details

For a fixed λ we use the R-package `CVXR` (Fu et al., 2017) to solve the optimization problem in Equations (8) (and (15) in Appendix A.1). As discussed in remark 2 of section 2.2, under the scaled squared error loss ($k = 1$) the linear inequality constraints, given by $\boldsymbol{\mathcal{A}h} \preceq \boldsymbol{b}$, ensure that the resulting decision rule $\boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda)$ is monotonic, while the equality constraints $\boldsymbol{\mathcal{C}h} = \boldsymbol{d}$ handle boundary cases that involve $y_i = 0$ and ties. Moreover, since $w^{(1)}_p(y) > 0$, the inequality constraints also ensure that $h_i < 1$ whenever $y_i > 0$. Implementation under the squared error loss ($k = 0$) follows along similar lines and the inequality constraints in this case ensure that $h_i + y_i > 0$ whenever $y_i \geq 0$.

A data-driven choice of the tuning parameter λ is obtained by first solving problems (8) and (15) over a grid of λ values, i.e. $\{\lambda_1, \ldots, \lambda_s\}$, and then computing the corresponding asymptotic risk estimate $\texttt{ARE}^{(k)}_n(\lambda_j)$ for $j = 1, \ldots, s$. Then λ is chosen according to

$$\hat{\lambda}_k \coloneqq \underset{\lambda \in \{\lambda_1, \ldots, \lambda_s\}}{\arg\min} \ \texttt{ARE}^{(k)}_n(\lambda),$$

where $k \in \{0, 1\}$. For all simulations and real data analyses considered in this paper, we have fixed $s = 10$ and employed an equi-spaced grid over $[10, 10^2]$.

4.2 Simulations: Poisson Distribution

In this section we consider the Poisson compound decision problem and generate $Y_i \mid \theta_i \overset{ind.}{\sim} \texttt{Poi}(\theta_i)$ for $i = 1, \ldots, n$. We vary n from 500 to 5000 in increments of 500 and simulate θ_i from the following four different scenarios:

Scenario 1: $\theta_i \overset{i.i.d.}{\sim} \texttt{Unif}(0.5, 15)$.

Scenario 2: $\theta_i \overset{i.i.d.}{\sim} \texttt{Gamma}(10, 2)$.

In the next two scenarios we consider departures from the usual Poisson model and simulate our data from the Conway-Maxwell-Poisson distribution (Shmueli et al., 2005) $\texttt{CMP}(\theta_i, \nu)$. The `CMP` distribution is a generalization of some well-known discrete distributions. With $\nu < 1$, `CMP` represents a discrete distribution that has longer tails than the Poisson distribution with parameter θ_i.

Scenario 3: We simulate $\theta_i \overset{i.i.d.}{\sim} 0.5\ \delta_{\{10\}} + 0.5\ \texttt{Gamma}(5, 2)$ for each i and let

$$Y_i \mid \theta_i \overset{ind.}{\sim} 0.8\ \texttt{Poi}(\theta_i) + 0.2\ \texttt{CMP}(\theta_i, \nu),$$

where we fix $\nu = 0.8$ for the `CMP` distribution.

Scenario 4.1: In this scenario we conduct estimation under the scaled squared error loss. We let $\theta_i \overset{i.i.d.}{\sim} 0.5\ \delta_{\{5\}} + 0.5\ \delta_{\{15\}}$, $\nu_i | \theta_i = 0.8\ \mathbb{I}(\theta_i = 5) + 1\ \mathbb{I}(\theta_i = 15)$ and simulate Y_i from the `CMP` distribution with parameters θ_i and ν_i. Thus, about half of the samples arise from a Poisson distribution with mean 15 while remaining are realizations from a $\texttt{CMP}(5, 0.8)$.

Scenario 4.2: We consider estimation under the squared error loss and let $\boldsymbol{\theta}$ to be an equi-spaced vector of length n in $[1, 5]$. We simulate Y_i from the `CMP` distribution with parameters θ_i and ν fixed at 0.8.

For each scenario, the following competing estimators of θ_i are considered:

1. the proposed estimator, denoted `NEB` and the oracle `NEB` estimator $\boldsymbol{\delta}^{\mathsf{or}}_{(k)} := \boldsymbol{\delta}^{\mathsf{neb}}_{(k)}(\lambda^{\mathsf{orc}})$, denoted `NEB OR`;
2. the estimator of Poisson means from Brown et al. (2013), denoted `BGR`;
3. Tweedie's formula for the Poisson model, denoted `TF OR`;
4. Tweedie's formula for the Normal means problem based on transformed data, denoted `TF Gauss`. The approach using transformation was suggested by Brown et al. (2013).
5. the estimator of Poisson means from Koenker and Gu (2017), denoted `KM`;
6. the estimator of Poisson means based on the g-modeling approach of Efron (2016), denoted `Deconv`.

The risk performance of the `TF OR` method relies heavily on the choice of a suitable bandwidth parameter $h > 0$. We use the oracle loss estimate h^{orc}, which is obtained by minimizing the true loss $\mathcal{L}^{(0)}_n$. The `TF Gauss` methodology is only applicable for the Normal means problem, and uses a variance stabilization transformation on Y_i to get $Z_i = 2\sqrt{Y_i + 0.25}$. The Z_i are then treated as approximate Normal random variables with mean μ_i and variances 1. To estimate the normal means μ_i we rely on g-modeling and use NPMLE. Finally, θ_i are estimated as $0.25\hat{\mu_i}^2$. It is important to note that along with the `NEB` estimator, `BGR` and `TF OR` are based on f-modeling while the rest in the preceding list of six competitors are based on g-modeling. Moreover, `BGR`, `TF OR` and `TF Gauss` only focus on the regular squared error loss $\mathcal{L}^{(0)}_n$. Nevertheless, in our simulation we assess the performance of these estimators for estimating $\boldsymbol{\theta}$ under both $\mathcal{L}^{(0)}_n$ and $\mathcal{L}^{(1)}_n$.

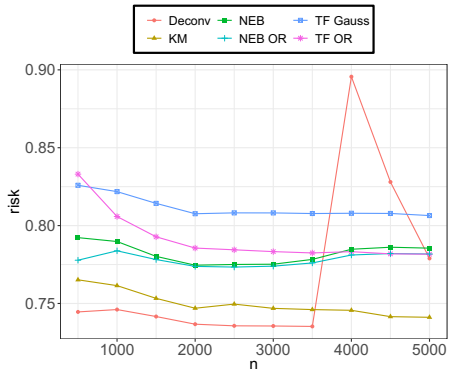
Table 1: Poisson compound decision problem under scaled squared error loss: Risk ratios $\mathcal{R}^{(1)}_n(\boldsymbol{\theta}, \cdot)/\mathcal{R}^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(1)})$ at $n = 5000$ for estimating $\boldsymbol{\theta}$.

	Scenario			
Method	1	2	3	4.1
`KM`	0.94	1.00	1.11	1.00
`Deconv`	1.00	1.06	1.03	1.11
`TF Gauss`	1.03	1.03	1.23	1.18
`TF OR`	1.00	1.02	1.28	1.10
`BGR`	1.22	1.07	1.28	1.25
`NEB`	1.00	1.00	1.00	1.00
`NEB OR`	1.00	1.00	0.98	1.00

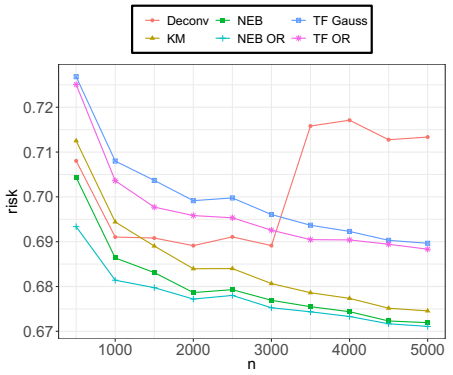
Table 2: Poisson compound decision problem under squared error loss: Risk ratios $\mathcal{R}^{(0)}_n(\boldsymbol{\theta}, \cdot)/\mathcal{R}^{(0)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(0)})$ at $n = 5000$ for estimating $\boldsymbol{\theta}$.

	Scenario			
Method	1	2	3	4.2
`KM`	1.00	1.01	1.59	1.21
`Deconv`	1.02	1.08	1.43	1.21
`TF Gauss`	1.00	1.01	1.51	1.08
`TF OR`	1.07	1.03	1.66	1.12
`BGR`	1.01	1.02	1.55	1.15
`NEB`	1.00	1.00	1.00	1.00
`NEB OR`	1.00	1.00	0.90	1.00

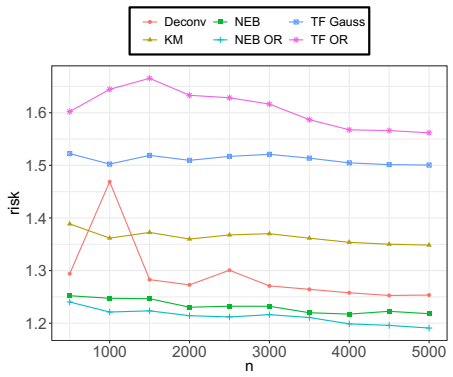
The performances of these six estimators are presented in figures 1 and 2 wherein the risk $\mathcal{R}^{(k)}_n(\boldsymbol{\theta}, \cdot)$ is estimated using 50 Monte Carlo repetitions for varying n. Tables 1 and 2



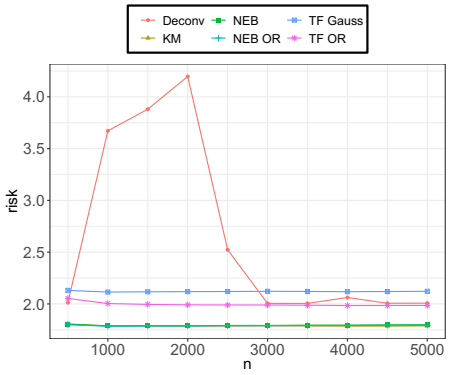
(a) Scenario 1: Estimation of $\boldsymbol{\theta}$ under loss $\mathcal{L}_n^{(1)}$ where $\theta_i \overset{iid.}{\sim}$ Unif$(0.5, 15)$.



(b) Scenario 2: Estimation of $\boldsymbol{\theta}$ under loss $\mathcal{L}_n^{(1)}$ where $\theta_i \overset{i.i.d.}{\sim}$ `Gamma`$(10, 2)$.



(c) Scenario 3: Estimation of $\boldsymbol{\theta}$ under loss $\mathcal{L}_n^{(1)}$ where $\theta_i \overset{iid.}{\sim} 0.5\ \delta_{\{10\}} + 0.5$ `Gamma`$(5, 2)$.

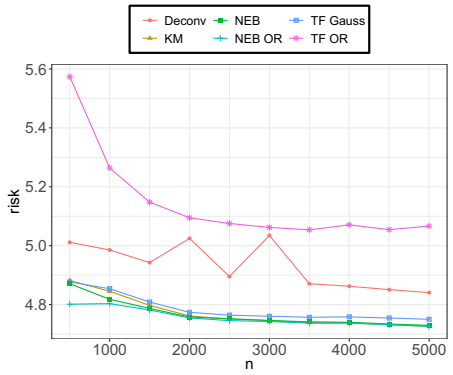


(d) Scenario 4.1: Estimation of $\boldsymbol{\theta}$ under loss $\mathcal{L}_n^{(1)}$ where $\boldsymbol{\theta}$ is an equi-spaced vector of length n in $[1, 5]$ and $Y_i|\theta_i \overset{ind.}{\sim}$ `CMP`$(\theta_i, 0.8)$.

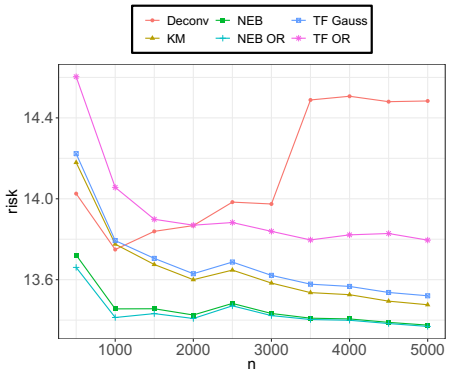
Figure 1: Poisson compound decision problem under scaled squared error loss: Risk estimates of the various estimators for scenarios 1, 2, 3 and 4.1.

report the ratios $\mathcal{R}_n^{(k)}(\boldsymbol{\theta}, \cdot)/\mathcal{R}_n^{(k)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(k)}^{\mathsf{neb}})$ of the average risks at $n = 5000$ and for $k = 1, 0$ respectively, where a risk ratio bigger than 1 indicates a smaller estimation risk for the `NEB` estimator. For `BGR` the modified cross validation approach of choosing the bandwidth parameter was extremely slow in our simulations and we therefore report its risk performance only at $n = 5000$.

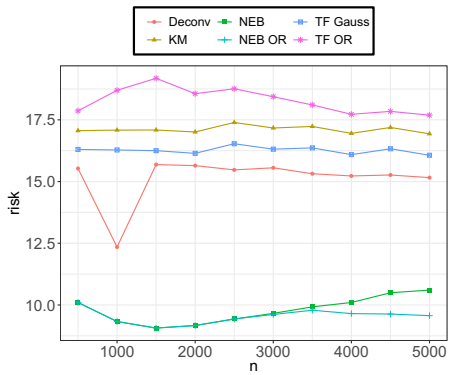
Figure 1 and table 1 present the risk performances of the competing estimators under the scaled squared error loss. Under scenarios 1 and 2 all estimators, with the exception of `BGR` in scenario 1 (table 1), exhibit competitive risk performance. For scenarios 3 and 4.1, which represent departures from the Poisson model, the `NEB` estimator demonstrates a substantially better performance than `TF Gauss, TF OR` and `BGR`. We note that `KM` and `Deconv` are competitive in scenarios 4.1 and 3, respectively, which indicates that along with the `NEB` estimator these g-modeling based approaches are potentially robust to misspecifications of the Poisson model considered in scenarios 3 and 4.1. Figure 1 reveals that the risk profile of `Deconv` is affected by its poorer estimates of $\boldsymbol{\theta}$ at various sample sizes



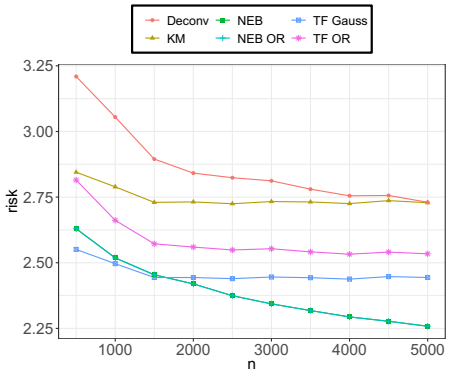
(a) Scenario 1: Estimation of $\boldsymbol{\theta}$ under loss $\mathcal{L}_n^{(0)}$ where $\theta_i \overset{iid.}{\sim} \texttt{Unif}(0.5, 15)$.



(b) Scenario 2: Estimation of $\boldsymbol{\theta}$ under loss $\mathcal{L}_n^{(0)}$ where $\theta_i \overset{i.i.d.}{\sim} \texttt{Gamma}(10, 2)$.



(c) Scenario 3: Estimation of $\boldsymbol{\theta}$ under loss $\mathcal{L}_n^{(0)}$ where $\theta_i \overset{iid.}{\sim} 0.5\ \delta_{\{10\}} + 0.5\ \texttt{Gamma}(5, 2)$.



(d) Scenario 4.2: Estimation of $\boldsymbol{\theta}$ under loss $\mathcal{L}_n^{(0)}$ where $\boldsymbol{\theta}$ is an equi-spaced vector of length n in $[1, 5]$ and $Y_i|\theta_i \overset{ind.}{\sim} \texttt{CMP}(\theta_i, 0.8)$.

Figure 2: Poisson compound decision problem under squared error loss: Risk estimates of the various estimators for scenarios 1, 2, 3 and 4.2.

and especially at the smaller sample sizes for scenarios 3 and 4.1. This behavior continues to appear even when the number of Monte Carlo repetitions are increased.

The risk performance of the competing estimators under the squared error loss is presented in figure 2 and table 2. Under scenarios 1 and 2 all estimators continue to exhibit a competitive performance. `BGR`, in particular, demonstrates a substantially improved performance now that estimation is conducted under squared error loss. Scenarios 3 and 4.2 consider departures from the Poisson model and in these settings the `NEB` estimator has a substantially better risk performance than all other competing methods considered here. We note that in scenarios 3, 4.1 and 4.2 the `NEB` estimator is robust to departures from the Poisson model. Proposition 7 in Barp et al. (2019) guarantees that, in general, the influence function of minimum KSD estimators, such as the `NEB` estimator, is bounded under data corruption and the behavior of the proposed `NEB` estimator in scenarios 3, 4.1 and 4.2 is potentially an example of such robustness property of minimum KSD estimators.

4.3 Simulations: Binomial Distribution

In this section we consider the Binomial compound decision problem and generate $Y_i \mid q_i \overset{ind.}{\sim} \mathcal{B}in(m_i,\ q_i)$ for $i = 1, \ldots, n$. We vary n from 500 to 5000 in increments of 500 and simulate $\theta_i = q_i/(1-q_i)$ from the following four different scenarios:

Scenario 1: $q_i \overset{i.i.d.}{\sim} \mathsf{Unif}(0.1, 0.7)$ and $m_i = 10$.

Scenario 2: $\theta_i \overset{i.i.d.}{\sim} (1/3)\ \delta_{\{0.5\}} + (1/3)\ \delta_{\{1\}} + (1/3)\ \delta_{\{2\}}$ and $m_i = 10$.

Scenario 3: $q_i \overset{i.i.d.}{\sim} \mathsf{Beta}(1, 6)$ and $m_i = 10$.

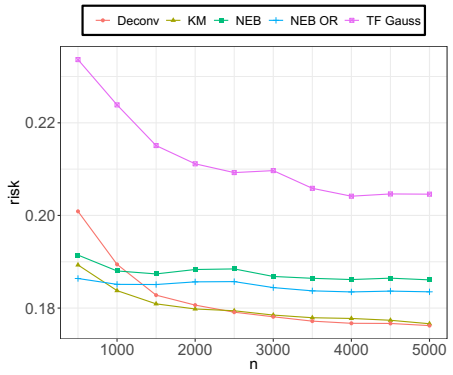
Scenario 4: $\theta_i \overset{i.i.d.}{\sim} \mathsf{Exp}(2)$ and $m_i = 5$.

Unlike scenarios 1 and 3, the data generating process in scenarios 2 and 4 directly sample the odds. Moreover the compound estimation problem in scenarios 3 and 4 is challenging because in these settings the distribution of θ_i has a mean that is substantially smaller in magnitude to the mean of θ_i in scenarios 1 and 2. For example in scenario 2 the mean of θ_i is about 1.16 while that in scenario 4 is 0.5. We consider the following five competing estimators of θ_i:

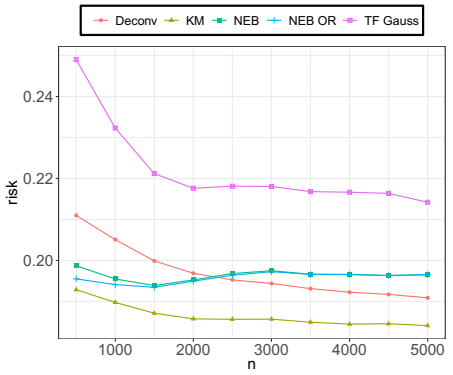
1. the proposed estimator `NEB` and its oracle version `NEB OR`;
2. Tweedie's formula for Binomial log odds, denoted `TF OR`;
3. Tweedie's formula for the Normal means problem based on transformed data, denoted `TF Gauss`.
4. the estimator of Binomial odds from Koenker and Gu (2017), denoted `KM`;
5. the estimator of Binomial odds from Efron (2016), denoted `Deconv`.

For `TF OR`, analogous to the Poisson case, we continue to use the oracle loss estimate h^{orc} as a choice for the bandwidth parameter. Since the `TF Gauss` methodology is only applicable for the Normal means problem, it uses a variance stabilization transformation on Y_i to get $Z_i = \mathsf{arcsin}\sqrt{(Y_i + 0.25)/(m_i + 0.5)}$. The Z_i are then treated as approximate Normal random variables with mean μ_i, variances $(4m_i)^{-1}$, and estimate of the means μ_i's are obtained using NPMLE. Finally, q_i is estimated as $\{\sin(\hat{\mu}_i)\}^2$. We note that the competitors `TF OR` and `TF Gauss` to our `NEB` estimator do not directly estimate the odds θ_i. For instance under the squared error loss, `TF Gauss` estimates the success probabilities q_i while `TF OR` estimates $\log \theta_i$. Nevertheless, in this simulation experiment we assess the performance of these two estimators for estimating the odds under both squared error loss and its scaled version. The simulation results are presented in Figures 3 and 4 wherein the risks of various estimators are calculated by averaging over 50 Monte Carlo repetitions for varying n. Tables 3 and 4 report the risk ratios $\mathcal{R}_n^{(k)}(\boldsymbol{\theta}, \cdot)/\mathcal{R}_n^{(k)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(k)}^{\mathsf{neb}})$ at $n = 5000$ and for $k = 1, 0$ respectively, where a risk ratio bigger than 1 indicates a smaller estimation risk for the `NEB` estimator.

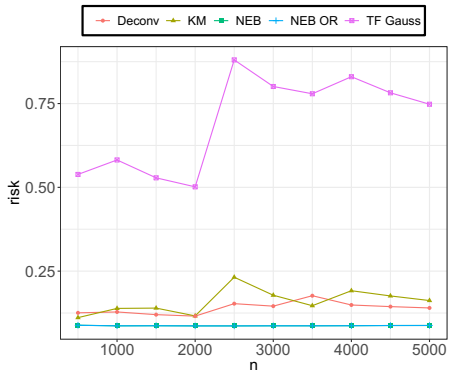
Under the scaled squared error loss (figure 3 and table 3) `KM` and `Deconv` demonstrate a superior risk performance for scenarios 1 and 2 while the `NEB` outperforms them for the



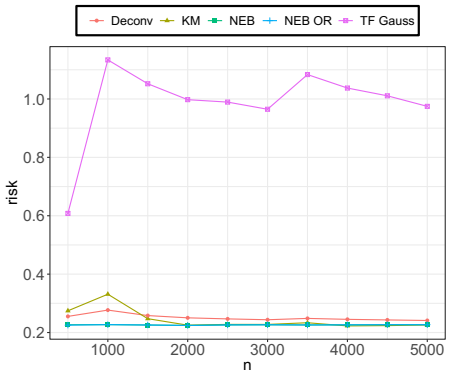
(a) Scenario 1: Estimation of odds $\boldsymbol{\theta}$ under loss $\mathcal{L}_n^{(1)}$ where $q_i \overset{i.i.d.}{\sim} \mathsf{Unif}(0.1, 0.7)$ and $m_i = 10$.



(b) Scenario 2: Estimation of odds $\boldsymbol{\theta}$ under loss $\mathcal{L}_n^{(1)}$ where $\theta_i \overset{i.i.d.}{\sim} (1/3)\ \delta_{\{0.5\}} + (1/3)\ \delta_{\{1\}} + (1/3)\ \delta_{\{2\}}$ and $m_i = 10$.



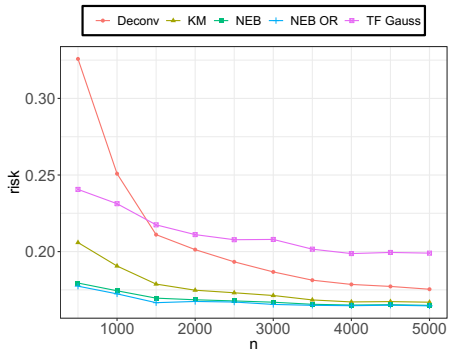
(c) Scenario 3: Estimation of odds $\boldsymbol{\theta}$ under loss $\mathcal{L}_n^{(1)}$ where $q_i \overset{i.i.d.}{\sim} \mathsf{Beta}(1, 6)$ and $m_i = 10$.



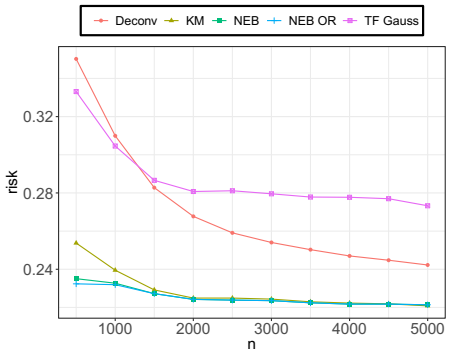
(d) Scenario 4: Estimation of odds $\boldsymbol{\theta}$ under loss $\mathcal{L}_n^{(1)}$ where $\theta_i \overset{i.i.d.}{\sim} \mathsf{Exp}(2)$ and $m_i = 5$.

Figure 3: Binomial compound decision problem under scaled squared error loss: Risk estimates of the various estimators for Scenarios 1 to 4.

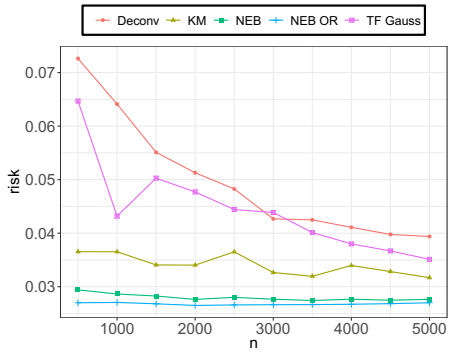
challenging settings of scenarios 3 and 4. The two Tweedie's formula based estimators, `TF Gauss` and `TF OR`, exhibit relatively poorer performance which is not surprising because these two estimators are designed to estimate q_i and $\log \theta_i$ under loss $\mathcal{L}_n^{(0)}$. For the squared error loss (figure 4 and table 4) the simulation results reveal that with the exception of scenario 3, the `NEB` estimator and `KM` demonstrate competitive risk performance. Scenario 3, along with scenario 4, is a challenging setting wherein the mean of the distribution of θ_i is substantially smaller in magnitude to the mean of θ_i in scenarios 1 and 2. Across the four scenarios, `TF OR` exhibits the poorest performance and appears to suffer from the fragmented approach of estimating the gradient of the log density $\log p(y)$ wherein $p(y)$ and its first derivative with respect to y are estimated separately using a Gaussian kernel with common bandwidth h^{orc}. Between the two g-modeling based approaches considered in this section, `Deconv` exhibits a relatively poorer risk performance than `KM` and for scenario 3 in particular the average risk of `Deconv` is substantially larger.



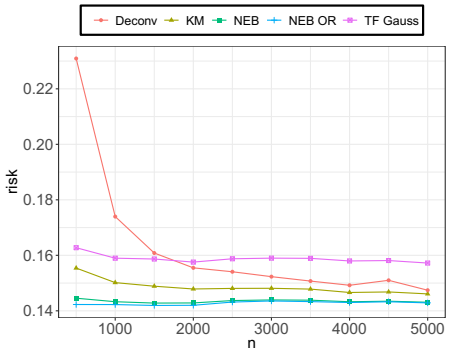
(a) Scenario 1: Estimation of odds $\boldsymbol{\theta}$ under loss $\mathcal{L}_n^{(0)}$ where $q_i \overset{i.i.d.}{\sim} \mathsf{Unif}(0.1, 0.7)$ and $m_i = 10$.



(b) Scenario 2: Estimation of odds $\boldsymbol{\theta}$ under loss $\mathcal{L}_n^{(0)}$ where $\theta_i \overset{i.i.d.}{\sim} (1/3)\ \delta_{\{0.5\}} + (1/3)\ \delta_{\{1\}} + (1/3)\ \delta_{\{2\}}$ and $m_i = 10$.



(c) Scenario 3: Estimation of odds $\boldsymbol{\theta}$ under loss $\mathcal{L}_n^{(0)}$ where $q_i \overset{i.i.d.}{\sim} \mathsf{Beta}(1, 6)$ and $m_i = 10$.



(d) Scenario 4: Estimation of odds $\boldsymbol{\theta}$ under loss $\mathcal{L}_n^{(0)}$ where $\theta_i \overset{i.i.d.}{\sim} \mathsf{Exp}(2)$ and $m_i = 5$.

Figure 4: Binomial compound decision problem under squared error loss: Risk estimates of the various estimators for Scenarios 1 to 4.

Table 3: The Binomial compound decision problem under scaled squared error loss: Risk ratios $\mathcal{R}_n^{(1)}(\boldsymbol{\theta}, \cdot)/\mathcal{R}_n^{(1)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\mathsf{neb}})$ at $n = 5000$ for estimating $\boldsymbol{\theta}$.

	Scenario			
Method	1	2	3	4
`KM`	0.95	0.94	1.85	1.00
`Deconv`	0.95	0.97	1.59	1.06
`TF Gauss`	1.01	1.09	8.52	4.30
`TF OR`	> 10	> 10	> 10	> 10
`NEB`	1.00	1.00	1.00	1.00
`NEB OR`	0.99	1.00	1.00	1.00

Table 4: The Binomial compound decision problem under the squared error loss: Risk ratios $\mathcal{R}_n^{(0)}(\boldsymbol{\theta}, \cdot)/\mathcal{R}_n^{(0)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(0)}^{\mathsf{neb}})$ at $n = 5000$ for estimating $\boldsymbol{\theta}$

	Scenario			
Method	1	2	3	4
`KM`	1.01	1.00	1.15	1.02
`Deconv`	1.06	1.09	1.42	1.03
`TF Gauss`	1.21	1.23	1.27	1.10
`TF OR`	> 10	> 10	> 10	> 10
`NEB`	1.00	1.00	1.00	1.00
`NEB OR`	1.00	1.00	0.98	1.00

4.4 Simulations: Negative Binomial Distribution

In this section we investigate the numerical performance of the NEB estimator for compound decision problems involving the Negative Binomial (NB) distribution. We generate observations $Y_i \mid q_i \overset{ind.}{\sim} \texttt{NBinom}(r_i, q_i)$ for $i = 1, \ldots, n$ and vary n from 500 to 5000 in increments of 500. Here the goal is to estimate $\theta_i = 1 - q_i$ and we consider the following three different scenarios for simulating q_i for $i = 1, \ldots, n$:

Scenario 1: $q_i \overset{i.i.d.}{\sim} 0.4\ \delta_{\{0.5\}} + 0.6\ \texttt{Beta}(1, 1)$ and fix $r_i = 3$.

Scenario 2: $q_i \overset{i.i.d.}{\sim} (1/3)\ \delta_{\{0.5\}} + (1/3)\ \delta_{\{0.7\}} + (1/3)\delta_{\{0.9\}}$ and fix $r_i = 5$.

Scenario 3: $q_i \overset{i.i.d.}{\sim} \texttt{Beta}(5, 2)$ and fix $r_i = 10$.

In scenarios 2 and 3 the median θ_i is substantially smaller than 0.5 which represents a challenging estimation setting for the following competing estimators:

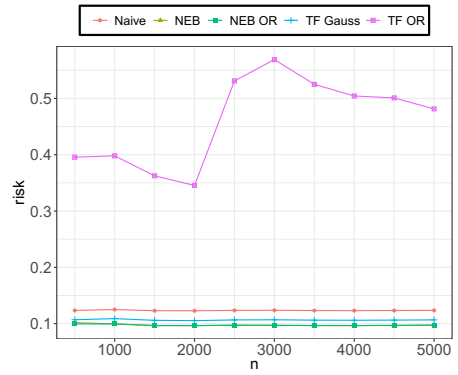
1. the proposed estimator, denoted NEB and the oracle version NEB OR;
2. Tweedie's formula for $\log \theta_i$ under the NB model, denoted TF OR;
3. Tweedie's formula for the Normal means problem based on transformed data, denoted TF Gauss;
4. the naive estimator $1 - (r_i - 1)/(r_i + Y_i - 1)$ of θ_i where $(r_i - 1)/(r_i + Y_i - 1)$ is the minimum variance unbiased estimator (MVUE) of q_i.

We continue to use the oracle loss estimate h^{orc} as the bandwidth choice for TF OR. For TF Gauss we use a variance stabilization transformation on Y_i to get $Z_i = 2\ \mathsf{arcsin}\sqrt{Y_i/r_i}$. The Z_i are then treated as approximate Normal random variables with mean μ_i and variances $1/r_i$. To estimate the normal means μ_i we rely on g-modeling and use NPMLE. Finally, θ_i are estimated as $1 - \{1 + [\texttt{sinh}(0.5\hat{\mu}_i)]^2\}^{-1}$. It is important to note that unlike the NEB estimator, the remaining competing estimators only focus on the regular squared error loss $\mathcal{L}_n^{(0)}$. Nevertheless, in our simulation we assess the performance of these estimators for estimating $\boldsymbol{\theta}$ under both $\mathcal{L}_n^{(0)}$ and $\mathcal{L}_n^{(1)}$.

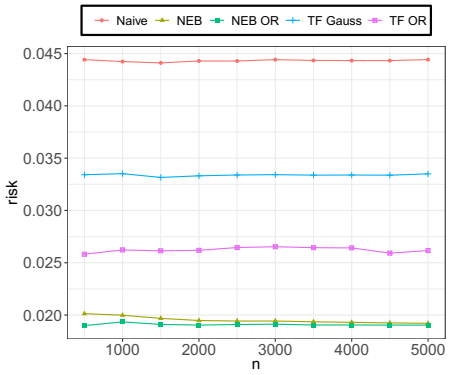
Figure 5 and tables 5, 6 report the performance of the competing estimators of $\boldsymbol{\theta}$ for the NB compound estimation problem. Under the squared error loss table 6 and right panel of figure 5 reveal that across all sample sizes performance of the NEB estimator is substantially better than the competing estimators considered in this experiment. In this setting TF OR is the next best while the naive estimator of $\boldsymbol{\theta}$ is outperformed by the three shrinkage estimators. Under the scaled squared error loss (table 5 and left panel of figure 5), the NEB estimator continues to be better than the competing estimators although the performance of TF Gauss is impressive given that it is based on Normality transformed data under the usual squared error loss.

5. Real Data Analyses

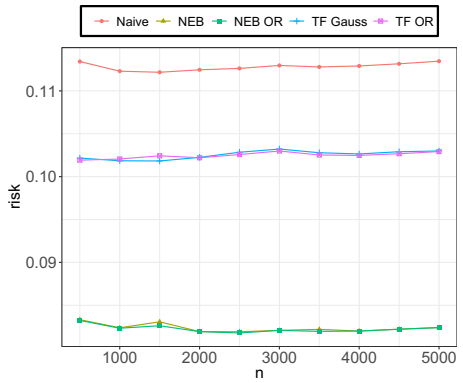
This section illustrates two real data applications that use the proposed method for estimating Juvenile Delinquency rates from Poisson models and news popularity from Binomial models.



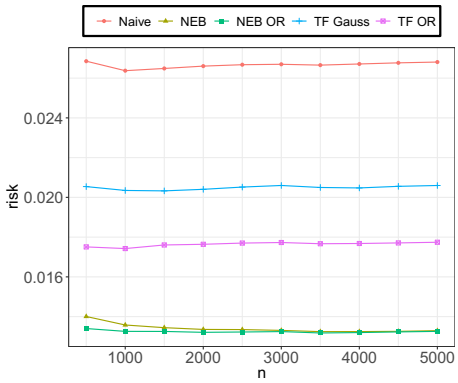
(a) Scenario 1: Estimation of $\theta_i = 1 - q_i$ under loss $\mathcal{L}_n^{(1)}$ where $q_i \overset{i.i.d.}{\sim} 0.4\ \delta_{\{0.5\}} + 0.6\ \texttt{Beta}(1,1)$ and $r_i = 3$.



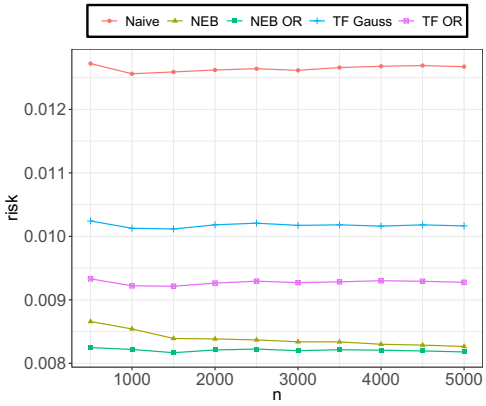
(b) Scenario 1: Estimation of $\theta_i = 1 - q_i$ under loss $\mathcal{L}_n^{(0)}$ where $q_i \overset{i.i.d.}{\sim} 0.4\ \delta_{\{0.5\}} + 0.6\ \texttt{Beta}(1,1)$ and $r_i = 3$.



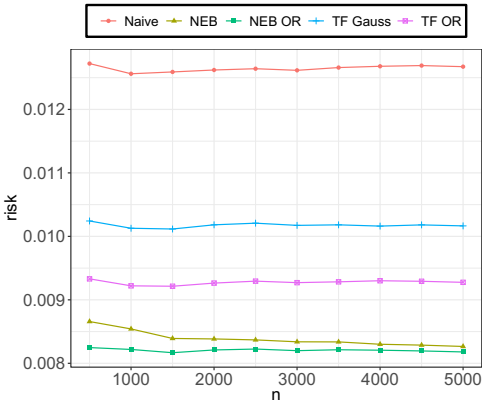
(c) Scenario 2: Estimation of $\theta_i = 1 - q_i$ under loss $\mathcal{L}_n^{(1)}$ where $q_i \overset{i.i.d.}{\sim} (1/3)\ \delta_{\{0.5\}} + (1/3)\ \delta_{\{0.7\}} + (1/3)\delta_{\{0.9\}}$ and $r_i = 5$.



(d) Scenario 2: Estimation of $\theta_i = 1 - q_i$ under loss $\mathcal{L}_n^{(0)}$ where $q_i \overset{i.i.d.}{\sim} (1/3)\ \delta_{\{0.5\}} + (1/3)\ \delta_{\{0.7\}} + (1/3)\delta_{\{0.9\}}$ and $r_i = 5$.



(e) Scenario 3: Estimation of $\theta_i = 1 - q_i$ under loss $\mathcal{L}_n^{(1)}$ where $q_i \overset{i.i.d.}{\sim} \texttt{Beta}(5,2)$ and $r_i = 10$.



(f) Scenario 3: Estimation of $\theta_i = 1 - q_i$ under loss $\mathcal{L}_n^{(0)}$ where $q_i \overset{i.i.d.}{\sim} \texttt{Beta}(5,2)$ and $r_i = 10$.

Figure 5: Negative Binomial compound decision problem: Risk estimates of the various estimators for Scenarios 1 to 3.

Table 5: The NB compound decision problem under scaled squared error loss: Risk ratios $\mathcal{R}_n^{(1)}(\boldsymbol{\theta},\cdot)/\mathcal{R}_n^{(1)}(\boldsymbol{\theta},\boldsymbol{\delta}_{(1)}^{\mathsf{neb}})$ at $n = 5000$ for estimating $\boldsymbol{\theta}$.

	Scenario		
Method	1	2	3
`Naive`	1.27	1.38	1.23
`TF Gauss`	1.09	1.25	1.14
`TF OR`	4.94	1.25	1.11
`NEB`	1.00	1.00	1.00
`NEB OR`	1.00	1.00	1.00

Table 6: The NB compound decision problem under the squared error loss: Risk ratios $\mathcal{R}_n^{(0)}(\boldsymbol{\theta},\cdot)/\mathcal{R}_n^{(0)}(\boldsymbol{\theta},\boldsymbol{\delta}_{(0)}^{\mathsf{neb}})$ at $n = 5000$ for estimating $\boldsymbol{\theta}$.

	Scenario		
Method	1	2	3
`Naive`	2.31	2.02	1.53
`TF Gauss`	1.74	1.55	1.23
`TF OR`	1.36	1.33	1.12
`NEB`	1.00	1.00	1.00
`NEB OR`	0.99	1.00	0.99

5.1 Estimation of Juvenile Delinquency rates

We consider an application for analysis of the Uniform Crime Reporting Program (UCRP) Database (US Department of Justice and Federal Bureau of Investigation, 2014) that holds county-level counts of arrests and offenses ranging from robbery to weapons violations in 2012. The database is maintained by the National Archive of Criminal Justice Data (NACJD) and is one of the most widely used database for research related to factors that affect juvenile delinquency (JD) across the United States (see for example (Aizer and Doyle Jr, 2015; Damm and Dustmann, 2014; Koski et al., 2018)). A preliminary and important goal in these analyses is to estimate the JD rates based on observed arrest data and determine the counties that are amongst the worst or least affected. However with almost 3,000 counties being evaluated the observed JD counts are susceptible to selection bias, wherein some of the data points are in the extremes merely by chance and traditional estimators may underestimate or overestimate the corresponding delinquency rates, especially in counties with fewer total number of arrests across all age groups.

For the purpose of our analyses, we use the 2012 UCRP data that spans $n = 3,178$ counties in the U.S. and consider estimating the JD rate θ_i for county $i = 1,\ldots,n$. The observed data for county i in the year 2012 is the pair (y_{i1}, m_{i1}) which represent, respectively, the number of juvenile arrests and total arrests in county i during that year. We assume that $Y_{i1} \mid m_{i1},\theta_i \overset{ind.}{\sim} \texttt{Poi}(m_{i1}\theta_i)$ and use the following six competing estimators of $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_n)$ from section 4.2: `NEB`, `BGR`, `KM`, `TF OR`, `TF Gauss` and `Deconv`. To assess the performance of the aforementioned estimators we consider predicting the 2014 county level JD counts $\boldsymbol{Y}_2 = (Y_{12},\ldots,Y_{n2})$ and compare their prediction performance under both $\mathcal{L}_n^{(0)}$ and $\mathcal{L}_n^{(1)}$ losses. In particular for any estimate $\hat{\delta}_i$ of θ_i, the 2014 predicted JD counts are $\hat{\boldsymbol{Y}}_2 = (\hat{\delta}_1 m_{12},\ldots,\hat{\delta}_n m_{n2})$ where m_{i2} is the total number of arrests in county i during 2014. The prediction performance of $\hat{\boldsymbol{\delta}}$ is then evaluated under loss $\mathcal{L}_n^{(k)}(\boldsymbol{Y}_2,\hat{\boldsymbol{Y}}_2)$ for $k \in \{0,1\}$. The data were cleaned prior to any analyses which ensured that all counties in the year 2012 had at least one arrest (juvenile or not). This resulted in $n = 2803$ counties where all methods are applied to. Let $\hat{\boldsymbol{Y}}_{2,(k)}^{\mathsf{neb}}$ denote the n vector of predicted JD counts for 2014 using $\boldsymbol{\delta}_{(k)}^{\mathsf{neb}}$. Table 7 reports the loss ratios $\mathcal{L}_n^{(k)}(\boldsymbol{Y}_2,\hat{\boldsymbol{Y}}_2)/\mathcal{L}_n^{(k)}(\boldsymbol{Y}_2,\hat{\boldsymbol{Y}}_{2,(k)}^{\mathsf{neb}})$ where a ratio bigger than 1 indicates a smaller prediction loss for $\boldsymbol{\delta}_{(k)}^{\mathsf{neb}}$. We see that under the scaled squared

Table 7: Loss ratios of the competing methods for predicting $\boldsymbol{Y}_2$.

$(n = 2,803)$	Loss ratios	
Method	$k = 1$	$k = 0$
`BGR`	1.09	0.97
`KM`	1.04	1.01
`Deconv`	3.18	1.01
`TF Gauss`	1.07	1.00
`TF OR`	1.19	1.08
`NEB`	1.00	1.00

error loss ($k = 1$) all five competing estimators to `NEB` exhibit loss ratios bigger than 1 while `BGR` outperforms all others under the squared error loss ($k = 0$). Under this loss, however, the `NEB` estimator continues to provide a better prediction accuracy than `TF OR` and demonstrates a competitive performance against `KM`, `Deconv` and `TF Gauss`.

5.2 News popularity in social media platforms

Journalists and editors often face the critical task of assessing the popularity of various news items and determining which articles are likely to become popular; hence existing content generation resources can be efficiently managed and optimally allocated to avenues with maximum potential. Due to the dynamic nature of the news articles, popularity is usually measured by how quickly the article propagates (frequency) and the number of readers that the article can reach (severity) through social media platforms like Twitter, Youtube, Facebook and LinkedIn. As such predicting these two aspects of popularity based on early trends is extremely valuable to journalists and content generators (Bandari et al., 2012). In this section, we assess the popularity of several news items based on their frequency

Table 8: Loss ratios of the competing methods for estimating $\boldsymbol{\theta}$. News article genre: `Economy` and social media: `Facebook`

$(n = 3,972)$	Loss Ratios	
Method	$k = 1$	$k = 0$
`NEB`	1.00	1.00
`KM`	6.98	> 10
`Deconv`	> 10	> 10
`TF Gauss`	4.13	3.33
`TF OR`	> 10	> 10

Table 9: Loss ratios of the competing methods for estimating $\boldsymbol{\theta}$. News article genre: `Microsoft` and social media: `LinkedIn`

$(n = 3,850)$	Loss Ratios	
Method	$k = 1$	$k = 0$
`NEB`	1.00	1.00
`KM`	> 10	> 10
`Deconv`	> 10	> 10
`TF Gauss`	9.26	7.34
`TF OR`	> 10	> 10

of propagation and analyze a dataset from Moniz and Torgo (2018) that holds 48 hours worth of social media feedback data on a large collection of news articles since the time of first publication. For the purposes of our analysis, we consider two popular genres of news from this data set: `Economy` and `Microsoft`, and examine how frequently these articles

were shared in Facebook and LinkedIn, respectively, over a period of 48 hours from the time of their first publication. Each news article in the data has a unique identifier and 16 consecutive time intervals, each of length 180 minutes, to detect whether the article was shared at least once in that time interval. Let $Z_{ij} = 1$ if article i was shared in time interval j and 0 otherwise, where $i = 1, \ldots, n$ and $j = 1, \ldots, 16$. Suppose $q_{ij} \in [0, 1]$ denotes the probability that news article i is shared in interval j. Note that in general q_{ij} depends on j since the popularity of any news article evolves with time and therefore Z_{ij} are not independently distributed for $j = 1, \ldots, 16$. However for the purposes of this analysis, we let $q_{ij} = q_i$ for $j = 1, \ldots, 16$ and assume that for each i, Z_{ij} are independent realizations from $\texttt{Ber}(q_i)$. It then follows that $Y_{i1} = \sum_{j=1}^{8} Z_{ij} \overset{ind.}{\sim} \texttt{Bin}(8, q_i)$. To assess the popularity of article i we estimate its odds of sharing in the remaining 8 time intervals ($j = 9, \ldots, 16$) and consider the following 5 estimators from section 4.3: `NEB`, `KM`, `Deconv`, `TF Gauss` and `TF OR`. Tables 8 and 9 report the loss ratios $\mathcal{L}_n^{(k)}(\boldsymbol{\theta}, \boldsymbol{\delta})/\mathcal{L}_n^{(k)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(k)}^{\mathsf{neb}})$ for any estimator $\boldsymbol{\delta}$ of $\boldsymbol{\theta}$ where $\theta_i = Y_{i2}/(8 - Y_{i2})$ and $Y_{i2} = \sum_{j=9}^{16} Z_{ij}$. We observe that all four competitors to the `NEB` estimator exhibit loss ratios substantially bigger than 1 under both the losses. The relatively poorer performance of `KM` and `Deconv` in this example stems from the fact that in this application $Y_{i1} = 8$ for several news articles. For those news articles both `KM` and `Deconv` return disproportionately bigger estimates of θ_i which explains their relatively larger estimation loss.

6. Discussion

In this paper we propose a Nonparametric Empirical Bayes framework for compound estimation in the discrete linear exponential family. The proposed estimator is consistent and presents a unified framework for compound estimation in the `DLE` family by estimating the Bayes shrinkage factors via a convex program that can easily incorporate various structural constraints, such as monotonicity, into the data driven decision rule. Our numerical evidence suggests that across many settings the `NEB` estimator has a substantially better risk performance than the competing approaches considered here.

We conclude this article with two open issues. First, in large scale compound estimation problems, one is often interested in constructing confidence intervals for the EB shrinkage estimators. Recall that for any coordinate i, $\hat{\delta}_{(k),i}^{\mathsf{neb}}$ is non-linear and a biased estimate of θ_i. While the CLT for minimum KSD estimators in Barp et al. (2019) will be important for deriving the asymptotic distribution of the `NEB` estimator, the main challenge in constructing confidence intervals lies in accounting for the bias in $\hat{\delta}_{(k),i}^{\mathsf{neb}}$ for estimating θ_i. In the absence of any information on the prior G, it is not immediately clear how to accurately characterize this bias. Notable recent developments include "de-biasing" the EB estimator (Ignatiadis and Wager, 2019) or assuming a Normal distribution on G but using a carefully constructed larger critical value to account for the bias due to shrinkage (Armstrong et al., 2020). Secondly, the `NEB` estimation framework handles both regular and scaled squared error losses and it is desirable to construct such empirical Bayes shrinkage estimators for other asymmetric losses such as the Linex loss (Varian, 1975) and the Generalized absolute loss function (Koenker and Bassett Jr, 1978). As part of our future research, we will be interested in pursuing these aforementioned directions.

Acknowledgments

We thank the Action Editor and three anonymous referees whose comments have substantially improved the quality of the paper. G. Mukherjee's work was partly supported by the NSF grant DMS-1811866. W. Sun's work was partly supported by the NSF grant DMS-1712983. Q. Liu's work is supported in part by NSF CRII 1830161 and NSF CAREER 1846421.

Supplementary Material for "EB Estimation in Discrete Linear Exponential Family"

In this supplement, we first present in Appendix A the results for the NEB estimator under the squared loss, then in Appendix B we provide the proofs and technical details of all theories in the main text and Appendix A.

A. Results Under the Squared Error Loss

A.1 The NEB estimator

In this section we discuss the estimation of $\boldsymbol{w}_p^{(0)}$ that appear in lemma 1 under the usual squared error loss ($k = 0$). Let Y be a non-negative integer-valued random variable with probability mass function (pmf) p and define

$$h_0^{(0)}(y) = \frac{y+1}{w_p^{(0)}(y)} - y \ , y \in \{0\} \cup \mathbb{N} \tag{14}$$

Suppose $\mathcal{K}_\lambda(y, y') = \exp\{-0.5\lambda^{-1}(y-y')^2\}$ be the positive definite Gaussian kernel with bandwidth parameter $\lambda \in \Lambda$ where Λ is a compact subset of $\mathbb{R}^+$ bounded away from 0. Given observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ from model (2), let $\boldsymbol{h}_0^{(0)} = (h_0^{(0)}(y_1), \ldots, h_0^{(0)}(y_n))$ and define the following $n \times n$ matrices: $n^2\boldsymbol{K}_\lambda = [\mathcal{K}_\lambda(y_i, y_j)]_{ij}$, $n^2\Delta\boldsymbol{K}_\lambda = [\Delta_{y_i}\mathcal{K}_\lambda(y_i, y_j + 1)]_{ij}$ and $n^2\Delta_2\boldsymbol{K}_\lambda = [\Delta_{y_i,y_j}\mathcal{K}_\lambda(y_i, y_j)]_{ij}$ where $\Delta_y\mathcal{K}_\lambda(y, y') = \mathcal{K}_\lambda(y+1, y') - \mathcal{K}_\lambda(y, y')$ and $\Delta_{y,y'}\mathcal{K}_\lambda(y, y') = \Delta_{y'}\Delta_y\mathcal{K}_\lambda(y, y') = \Delta_y\Delta_{y'}\mathcal{K}_\lambda(y, y')$.

Definition 3 (NEB estimator of θ_i)**.** *Consider the DLE Model* (2) *with loss* $\ell^{(0)}(\theta_i, \delta_i)$*. For a fixed* $\lambda \in \Lambda$*, let* $\hat{w}_i^{(0)}(\lambda) = (y_i + 1)/(y_i + \hat{h}_i^{(0)}(\lambda))$ *and* $\hat{\boldsymbol{h}}_n^{(0)}(\lambda) = \left\{\hat{h}_1^{(0)}(\lambda), \ldots, \hat{h}_n^{(0)}(\lambda)\right\}$ *be the solution to the following quadratic optimization problem:*

$$\min_{\boldsymbol{h} \in \boldsymbol{H}_n} \boldsymbol{h}^T\boldsymbol{K}_\lambda\boldsymbol{h} + 2\boldsymbol{h}^T\Delta\boldsymbol{K}_\lambda\boldsymbol{y} + \boldsymbol{y}^T\Delta_2\boldsymbol{K}_\lambda\boldsymbol{y}, \tag{15}$$

where $\boldsymbol{H}_n = \{\boldsymbol{h} = (h_1, \ldots, h_n) : \mathcal{A}\boldsymbol{h} \preceq \boldsymbol{b},\ \mathcal{C}\boldsymbol{h} = \boldsymbol{d}\}$ *is a convex set and* $\mathcal{A}, \mathcal{C}, \boldsymbol{b}, \boldsymbol{d}$ *are known real matrices and vectors that enforce linear constraints on the components of* $\boldsymbol{h}$*. Then the NEB estimator for a fixed* λ *is given by* $\boldsymbol{\delta}_{(0)}^{\mathsf{neb}}(\lambda) = \left\{\delta_{(0),i}^{\mathsf{neb}}(\lambda) : 1 \le i \le n\right\}$*, where*

$$\delta_{(0),i}^{\mathsf{neb}}(\lambda) = \frac{a_{y_i}/a_{y_i+1}}{\hat{w}_i^{(0)}(\lambda)}, \qquad \text{if } y_i \in \{0, 1, 2, \ldots\}$$

Remark 5. In problem (15) the linear inequality constraints $\mathcal{A}\boldsymbol{h} \preceq \boldsymbol{b}$ can be used to impose structural constraints on the NEB decision rule $\boldsymbol{\delta}^{\mathsf{neb}}_{(0)}(\lambda)$. The structural constraints may take the form of monotonicity constraints so that $\delta^{\mathsf{neb}}_{(0),(1)}(\lambda) \geq \cdots \geq \delta^{\mathsf{neb}}_{(0),(n)}(\lambda)$ for $y_{(1)} \geq y_{(2)} \geq \cdots \leq y_{(n)}$. In particular, when $Y_i \mid \theta_i \sim \text{Poi}(\theta_i)$ then $\delta^{\mathsf{neb}}_{(0),i}(\lambda) = y_i + \hat{h}^{(0)}_i(\lambda)$ and the monotonicity constraints in this setting will imply

$$-\hat{h}^{(0)}_{(i)}(\lambda) + \hat{h}^{(0)}_{(i+1)}(\lambda) \leq y_{(i)} - y_{(i+1)}, \text{ for } 1 \leq i \leq (n-1)$$

These $n-1$ linear inequality constraints may be imposed with an $(n-1) \times n$ matrix $\mathcal{A}$ and an $n-1$ column vector $\boldsymbol{b}$ such that for $1 \leq i \leq (n-1)$ and $1 \leq r \leq n$,

$$\begin{aligned} \mathcal{A}(i,r) &= \begin{cases} -1, \text{ when } y_r = y_{(i)} \\ 1, \text{ when } y_r = y_{(i+1)} \\ 0, \text{ otherwise} \end{cases} \\ \text{and } b_i &= y_{(i)} - y_{(i+1)}. \end{aligned}$$

The equality constraints $\mathcal{C}\boldsymbol{h} = \boldsymbol{d}$ may accommodate instances of ties for which we require $\hat{h}^{(0)}_i(\lambda) = \hat{h}^{(0)}_j(\lambda)$ whenever $y_i = y_j$.

Theorem 4. *Let $\mathcal{K}_\lambda(\cdot,\cdot)$ be the positive definite Gaussian kernel with bandwidth parameter $\lambda \in \Lambda$. If $\lim_{n\to\infty} c_n n^{-1/2} \log^4 n = 0$ then, under assumptions $(A1)-(A3)$, we have for any $\lambda \in \Lambda$,*

$$\lim_{n\to\infty} \mathbb{P}\Big[\frac{1}{n}\Big\|\hat{\boldsymbol{w}}^{(0)}_n(\lambda) - \boldsymbol{w}^{(0)}_p\Big\|_2^2 \geq c_n^{-1}\epsilon\Big] = 0, \text{ for any } \epsilon > 0$$

where $\hat{\boldsymbol{w}}^{(0)}_n(\lambda) = [(Y_i+1)/(\hat{h}^{(0)}_i(\lambda) + Y_i)]_i$.

We now provide some motivation behind the minimization problem in definition 3 for estimating the ratio functionals $\boldsymbol{w}^{(0)}_p$. Let $\hat{\mathbb{M}}_{\lambda,n}(\boldsymbol{h}) = \boldsymbol{h}^T\boldsymbol{K}_\lambda\boldsymbol{h} + 2\boldsymbol{h}^T\Delta\boldsymbol{K}_\lambda\boldsymbol{y} + \boldsymbol{y}^T\Delta_2\boldsymbol{K}_\lambda\boldsymbol{y}$ be the objective function in equation (15). Suppose $\tilde{p}$ be a probability mass function on the support of Y and define

$$\mathcal{S}_\lambda[\tilde{p}](p) = \mathbb{E}_p\Big[(\tilde{h}^{(0)}(Y) - h^{(0)}_0(Y))\mathcal{K}_\lambda(Y+1, Y'+1)(\tilde{h}^{(0)}(Y') - h^{(0)}_0(Y'))\Big] \tag{16}$$

where $h^{(0)}_0, \tilde{h}^{(0)}$ are as defined in equation (14) and Y, Y' are i.i.d copies from the marginal distribution that has mass function p. $\mathcal{S}_\lambda[\tilde{p}](p)$ in equation (16) is the Kernelized Stein's Discrepancy (KSD) measure that can be used to distinguish between two distributions with mass functions $p, \tilde{p}$ such that $\mathcal{S}_\lambda[\tilde{p}](p) \geq 0$ and $\mathcal{S}_\lambda[\tilde{p}](p) = 0$ if and only if $p = \tilde{p}$ (Liu et al., 2016; Chwialkowski et al., 2016; Yang et al., 2018). Moreover for i.i.d. copies (Y, Y') from p, it can be shown that

$$\mathcal{S}_\lambda[\tilde{p}](p) = \frac{1}{n(n-1)}\mathbb{E}_p\Big[\sum_{1\leq i\neq j\leq n} \kappa_\lambda[\tilde{h}^{(0)}(Y_i), \tilde{h}^{(0)}(Y_j)](Y_i, Y_j)\Big]$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is a random sample from the marginal distribution with mass function p and under the squared error loss $\kappa_\lambda[\tilde{h}^{(0)}(u), \tilde{h}^{(0)}(v)](u,v)$ is the positive definite

kernel function

$$\tilde{h}^{(0)}(u)\tilde{h}^{(0)}(v)\mathcal{K}_\lambda(u,v)+\tilde{h}^{(0)}(u)v\Delta_v\mathcal{K}_\lambda(u+1,v)+\tilde{h}^{(0)}(v)u\Delta_u\mathcal{K}_\lambda(u,v+1)+uv\Delta_{u,v}\mathcal{K}_\lambda(u,v). \tag{17}$$

An empirical evaluation scheme for $\mathcal{S}_\lambda[\tilde{p}](p)$ is given by $\mathcal{S}_\lambda[\tilde{p}](\hat{p}_n)$ where

$$\mathcal{S}_\lambda[\tilde{p}](\hat{p}_n)=\frac{1}{n^2}\sum_{i,j=1}^{n}\kappa_\lambda[\tilde{h}^{(0)}(y_i),\tilde{h}^{(0)}(y_j)](y_i,y_j) \tag{18}$$

where $\hat{p}_n$ is the empirical CDF. Note that $\kappa_\lambda[\tilde{h}^{(0)}(u),\tilde{h}^{(0)}(v)](u,v)$ in equation (17) involves $\tilde{p}$ only through $\tilde{h}^{(0)}$ and may analogously be denoted by $\kappa_\lambda[\tilde{h}(u),\tilde{h}(v)](u,v)$ where we have dropped the superscript from $\tilde{h}$ that indicates that the loss in question is the regular squared error loss. This slight abuse of notation is harmless as the discussion in this section is geared towards the squared error loss only.

Under the compound estimation framework of model (2), our goal is to estimate $h_0^{(0)}$. To do that we minimize $\mathcal{S}_\lambda[\tilde{p}](\hat{p}_n)$ in equation (18) with respect to the unknowns $\tilde{\boldsymbol{h}}=(\tilde{h}(y_1),\ldots,\tilde{h}(y_n))$. Note that $\mathcal{S}_\lambda[\tilde{p}](\hat{p}_n)$ is exactly the objective function $\hat{\mathbb{M}}_{\lambda,n}(\tilde{\boldsymbol{h}})$ of the quadratic program (15) with optimisation variables $h_i\equiv\tilde{h}(y_i)$ for $i=1,\ldots,n$.

A.2 Bandwidth choice and asymptotic properties

We propose the following asymptotic risk estimate $\mathsf{ARE}_n^{(0)}(\lambda)$ of the true loss of $\boldsymbol{\delta}_{(0)}^{\mathsf{neb}}(\lambda)$ in the DLE model (2).

Definition 4 (ARE of $\boldsymbol{\delta}_{(0)}^{\mathsf{neb}}(\lambda)$ in the DLE model)**.** *Suppose* $Y_i\mid\theta_i\overset{ind.}{\sim}\texttt{DLE}\,(\theta_i)$. *Under the loss* $\ell^{(0)}(\theta_i,\cdot)$ *an asymptotic risk estimate of the true loss of* $\boldsymbol{\delta}_{(0)}^{\mathsf{neb}}(\lambda)$ *is*

$$\mathsf{ARE}_n^{(0)}(\lambda,\boldsymbol{y})=\frac{1}{n}\Big\{\sum_{i=1}^{n}[\delta_{(0),i}^{\mathsf{neb}}(\lambda)]^2-2\sum_{i=1}^{n}\psi_i(\lambda)\Big\}$$

where

$$\psi_i(\lambda)=\delta_{(0),j_i}^{\mathsf{neb}}(\lambda)(a_{y_i-1}/a_{y_i}),\ \ y_i=1,2\ldots$$

with $j_i\in\{1,\ldots,n\}$ *such that* $y_{j_i}=y_i-1$.

An estimate of the tuning parameter λ based on $\mathsf{ARE}_n^{(0)}(\lambda,\boldsymbol{y})$ is given by:

$$\hat{\lambda}=\underset{\lambda\in\Lambda}{\arg\min}\,\mathsf{ARE}_n^{(0)}(\lambda,\boldsymbol{y}) \tag{19}$$

where a choice of $\Lambda=[10,10^2]$ worked well in the simulations and real data analyses of sections 4 and 5. Lemma 2 continues to provide the large-sample properties of the proposed $\mathsf{ARE}_n^{(0)}$ criteria for the Poisson and Binomial distributions provided c_n is a sequence that satisfies $\lim_{n\to\infty}c_nn^{-1/4}\log^4 n=0$.

To analyze the quality of the estimates $\hat{\lambda}$ obtained from equation (19), we consider an oracle loss estimator $\boldsymbol{\delta}_{(0)}^{\mathsf{or}}:=\boldsymbol{\delta}_{(0)}^{\mathsf{neb}}(\lambda_0^{\mathsf{orc}})$ where

$$\lambda_0^{\mathsf{orc}}=\underset{\lambda\in\Lambda}{\arg\min}\,\mathcal{L}_n^{(0)}(\boldsymbol{\theta},\boldsymbol{\delta}_{(0)}^{\mathsf{neb}}(\lambda))$$

and Lemma 3 establishes the asymptotic optimality of $\hat{\lambda}$ obtained from equation (19). In theorem 5 below we provide decision theoretic guarantees on the `NEB` estimator and show that the average squared error between $\boldsymbol{\delta}^{\mathsf{neb}}_{(0)}(\hat{\lambda})$ and $\boldsymbol{\delta}^{\pi}_{(0)}$ is asymptotically small.

Theorem 5. *Under the conditions of Theorem 4, if* $\lim_{n\to\infty} c_n n^{-1/2}\log^6 n = 0$ *then, for the Poisson and the Binomial model,*

$$\frac{c_n}{n}\Big\|\boldsymbol{\delta}^{\mathsf{neb}}_{(0)}(\hat{\lambda}) - \boldsymbol{\delta}^{\pi}_{(0)}\Big\|_2^2 = o_p(1).$$

Furthermore, under the same conditions, we have,

$$\lim_{n\to\infty}\mathbb{E}\Big[\mathcal{L}^{(0)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(0)}(\hat{\lambda})) - \mathcal{L}^{(0)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(0)})\Big] = 0.$$

B. Technical Details and Proofs

We will begin this section with some notations and then state three lemmata that will be used in proving the statements discussed in Section 3.

Let $c_0, c_1, \ldots$ denote some generic positive constants which may vary in different statements. Let $\mathcal{D}_n = \{0, 1, 2, \ldots, \lceil C\log n\rceil\}$ where $\lceil x\rceil$ denotes the smallest integer greater or equal to x. Given a random sample $(Y_1, \ldots, Y_n)$ from model (2) denote B_n to be the event $\{\max_{1\le i\le n} Y_i \le C\log n\}$ where C is the constant given by lemma 4 below under assumption $(A2)$.

Lemma 4. *Assumption* $(A2)$ *implies that with probability tending to* 1 *as* $n\to\infty$,

$$\max(Y_1, \ldots, Y_n) \le C\log n$$

where $C > 0$ *is a constant depending on* ϵ.

Our next lemma below is a statement on the pointwise Lipschitz stability of the optimal solution $\hat{\boldsymbol{h}}^{(k)}_n(\lambda)$ under perturbations on the parameter $\lambda\in\Lambda$. See, for example, Bonnans and Shapiro (2013) for general results on the stability and sensitivity of parametrized optimization problems.

Lemma 5. *Let* $\hat{\boldsymbol{h}}^{(k)}_n(\lambda_0)$ *be the solution to problems* (8) *and* (15)*, respectively, for* $k\in\{0,1\}$ *and for some* $\lambda_0\in\Lambda$. *Then, under Assumption* $(A3)$*, there exists a constant* $L > 0$ *such that for any* $\lambda\in\Lambda$ *the solution* $\hat{\boldsymbol{h}}^{(k)}_n(\lambda)$ *to problems* (8) *and* (15) *satisfies*

$$\Big\|\hat{\boldsymbol{h}}^{(k)}_n(\lambda) - \hat{\boldsymbol{h}}^{(k)}_n(\lambda_0)\Big\|_2 \le L|\lambda-\lambda_0|$$

Lemma 6. *Suppose* $Y \mid \theta \overset{ind.}{\sim} \texttt{DLE}\,(\theta)$. *Then the following hold:*

$$\mathbb{E}_{Y|\theta}\Big\{[\delta^{\pi}_{(1)}(Y+1)]^2\Big(\frac{a_{Y+1}}{a_Y}\Big) - \frac{[\delta^{\pi}_{(1)}(Y)]^2}{\theta}\Big\} = 0 \text{ and}$$
$$\mathbb{E}_{Y|\theta}\Big\{\delta^{\pi}_{(0)}(Y-1)\Big(\frac{a_{Y-1}}{a_Y}\Big) - \theta\delta^{\pi}_{(0)}(Y)\Big\} = 0.$$

The proofs of Lemmata 4, 5 and 6 are available in appendix B.8.

B.1 Proof of Lemma 1

First note that for any coordinate i, the integrated Bayes risk of an estimator $\delta_{(k),i}$ of θ_i is $\sum_{y_i}\int p(y_i|\theta_i)\ell_n^{(k)}(\theta_i,\delta_{(k),i})\mathrm{d}G(\theta_i)$ which is minimized with respect to $\delta_{(k),i}$ if for each y_i, $\delta_{(k),i}(y_i)$ is defined as

$$\delta^{\pi}_{(k),i}(y_i)=\underset{\delta_{(k),i}}{\arg\min}\int p(y_i|\theta_i)\ell_n^{(k)}(\theta_i,\delta_{(k),i})\mathrm{d}G(\theta_i)$$

However, $\int p(y_i|\theta_i)\ell_n^{(k)}(\theta_i,\delta_{(k),i})\mathrm{d}G(\theta_i)$ is a minimum with respect to $\delta_{(k),i}$ when

$$\delta^{\pi}_{(k),i}(y_i)=\frac{\int p(y_i|\theta_i)\theta_i^{1-k}\mathrm{d}G(\theta_i)}{\int p(y_i|\theta_i)\theta_i^{-k}\mathrm{d}G(\theta_i)}$$

The result then follows by noting that $p(y_i-k)=\int a_{y_i-k}\theta_i^{y_i-k}/g(\theta_i)\mathrm{d}G(\theta_i)$, and $p(y_i+1-k)=\int a_{y_i+1-k}\theta_i^{y_i+1-k}/g(\theta_i)\mathrm{d}G(\theta_i)$ for $y_i=k,k+1,\ldots$.

B.2 Proof of Theorem 1

Define $\overline{\mathbb{M}}_\lambda(\tilde{\boldsymbol{h}})=\sum_{i,j\in\mathcal{D}_n}\kappa_\lambda[\tilde{h}(i),\tilde{h}(j)](i,j)P(Y=i)P(Y=j)$ and re-write $\hat{\mathbb{M}}_{\lambda,n}(\tilde{\boldsymbol{h}})$ as

$$\hat{\mathbb{M}}_{\lambda,n}(\tilde{\boldsymbol{h}})=\frac{1}{n^2}\sum_{i,j=1}^{n}\kappa_\lambda[\tilde{h}(i),\tilde{h}(j)](i,j)\mathcal{C}_{ij},$$

where $\mathcal{C}_{ij}$ is the number of pairs (Y_r,Y_s) in the sample that has $Y_r=i,Y_s=j$ and $P(Y=i)=\int p(i|\theta)\mathrm{d}G(\theta)$. Now, we have

$$\sup_{\lambda\in\Lambda}\Big|\hat{\mathbb{M}}_{\lambda,n}(\tilde{\boldsymbol{h}})-\mathbb{M}_\lambda(\tilde{\boldsymbol{h}})\Big|\le\sup_{\lambda\in\Lambda}\Big|\hat{\mathbb{M}}_{\lambda,n}(\tilde{\boldsymbol{h}})-\overline{\mathbb{M}}_\lambda(\tilde{\boldsymbol{h}})\Big|+\sup_{\lambda\in\Lambda}\Big|\mathbb{M}_\lambda(\tilde{\boldsymbol{h}})-\overline{\mathbb{M}}_\lambda(\tilde{\boldsymbol{h}})\Big| \tag{20}$$

Consider the first term on the right hand side of the inequality in equation (20). Let $P_i\coloneqq P(Y=i)$ and note that assumption (A2) and lemma 4 imply

$$\begin{aligned}\mathbb{E}_p\sup_{\lambda\in\Lambda}\Big|\hat{\mathbb{M}}_{\lambda,n}(\tilde{\boldsymbol{h}})-\overline{\mathbb{M}}_\lambda(\tilde{\boldsymbol{h}})\Big|&\le\sum_{i,j\in\mathcal{D}_n}\mathbb{E}_p\Big[\sup_{\lambda\in\Lambda}\Big|\kappa_\lambda[\tilde{h}(i),\tilde{h}(j)](i,j)\Big|\Big|\frac{\mathcal{C}_{ij}}{n^2}-P_iP_j\Big|\Big]\Big\{1+o(1)\Big\}\\&\le\sum_{i,j\in\mathcal{D}_n}\Big\{\mathbb{E}_p\Big[\sup_{\lambda\in\Lambda}\Big|\kappa_\lambda[\tilde{h}(i),\tilde{h}(j)](i,j)\Big|\Big]^2\mathbb{E}_p\Big|\frac{\mathcal{C}_{ij}}{n^2}-P_iP_j\Big|^2\Big\}^{1/2}\Big\{1+o(1)\Big\}.\end{aligned} \tag{21}$$

In equation (21) above, $\mathbb{E}_p|n^{-2}\mathcal{C}_{ij}-P_iP_j|^2$ is $O(1/n)$. Moreover, assumption (A1) together with the compactness of Λ and the continuity of $\kappa_\lambda[\tilde{h}(i),\tilde{h}(j)](i,j)$ with respect to λ imply that $\mathbb{E}_p[\sup_{\lambda\in\Lambda}|\kappa_\lambda[\tilde{h}(i),\tilde{h}(j)](i,j)|]^2<\infty$. Thus $\mathbb{E}_p\sup_{\lambda\in\Lambda}|\hat{\mathbb{M}}_{\lambda,n}(\tilde{\boldsymbol{h}})-\overline{\mathbb{M}}_\lambda(\tilde{\boldsymbol{h}})|$ is $O(\log^2 n/\sqrt{n})$.

Now consider the second term on the right hand side of the inequality in equation (20) and note that it is bounded above by the following tail sums

$$2\sum_{i\in\mathcal{D}_n,j\notin\mathcal{D}_n}\sup_{\lambda\in\Lambda}|\kappa_\lambda[\tilde{h}(i),\tilde{h}(j)](i,j)|P_iP_j+\sum_{i,j\notin\mathcal{D}_n}\sup_{\lambda\in\Lambda}|\kappa_\lambda[\tilde{h}(i),\tilde{h}(j)](i,j)|P_iP_j.$$

But from assumption (A1), $\mathbb{E}_p \sup_{\lambda\in\Lambda} |\kappa_\lambda[\tilde{h}(U), \tilde{h}(V)](U,V)| < \infty$ and together with assumption (A2) and proof of lemma 4, it follows that the terms in the display above are $O(n^{-\nu})$ for some $\nu > 1/2$.

Now fix an $\epsilon > 0$ and let $c_n = \sqrt{n}/\log^2 n$. Since $\mathbb{E}_p \sup_{\lambda\in\Lambda} |\hat{\mathbb{M}}_{\lambda,n}(\tilde{\boldsymbol{h}}) - \overline{\mathbb{M}}_\lambda(\tilde{\boldsymbol{h}})|$ is $O(\log^2 n/\sqrt{n})$ there exists a finite constant $M > 0$ and an N_1 such that $c_n \mathbb{E}_p \sup_{\lambda\in\Lambda} |\hat{\mathbb{M}}_{\lambda,n}(\tilde{\boldsymbol{h}}) - \overline{\mathbb{M}}_\lambda(\tilde{\boldsymbol{h}})| \le M$ for all $n \ge N_1$. Moreover since $\sup_{\lambda\in\Lambda} |\mathbb{M}_\lambda(\tilde{\boldsymbol{h}}) - \overline{\mathbb{M}}_\lambda(\tilde{\boldsymbol{h}})| \to 0$ as $n \to \infty$, there exists an N_2 such that $\sup_{\lambda\in\Lambda} |\mathbb{M}_\lambda(\tilde{\boldsymbol{h}}) - \overline{\mathbb{M}}_\lambda(\tilde{\boldsymbol{h}})| \le M/c_n$ for all $n \ge N_2$. Thus with $t = 4M/\epsilon$, we have $\mathbb{P}(c_n \sup_{\lambda\in\Lambda} |\hat{\mathbb{M}}_{\lambda,n}(\tilde{\boldsymbol{h}}) - \mathbb{M}_\lambda(\tilde{\boldsymbol{h}})| > t) < \epsilon$ for all $n \ge \max(N_1, N_2)$ which suffices to prove the desired result.

B.3 Proofs of Theorems 2 and 4

We will first prove Theorem 2. Note that from equation (7),

$$\Big\|\hat{\boldsymbol{w}}_n^{(1)}(\lambda) - \boldsymbol{w}_p^{(1)}\Big\|_2^2 = \Big\|\hat{\boldsymbol{h}}_n^{(1)}(\lambda) - \boldsymbol{h}_0^{(1)}\Big\|_2^2.$$

Now from assumption $(A3)$ and for any $\epsilon > 0$, there exists a $\delta > 0$ such that for any $\lambda \in \Lambda$,

$$\mathbb{P}\Big[\frac{c_n}{n}\Big\|\hat{\boldsymbol{h}}_n^{(1)}(\lambda) - \boldsymbol{h}_0^{(1)}\Big\|_2^2 \ge \epsilon\Big] \le \mathbb{P}\Big[c_n\Big\{\mathbb{M}_\lambda(\hat{\boldsymbol{h}}_n^{(1)}) - \mathbb{M}_\lambda(\boldsymbol{h}_0^{(1)})\Big\} \ge \delta\Big].$$

But the right hand side is upper bounded by the sum of $\mathbb{P}\Big[c_n\Big\{\mathbb{M}_\lambda(\hat{\boldsymbol{h}}_n^{(1)}) - \hat{\mathbb{M}}_{\lambda,n}(\hat{\boldsymbol{h}}_n^{(1)})\Big\} \ge \delta/3\Big]$, $\mathbb{P}\Big[c_n\Big\{\hat{\mathbb{M}}_{\lambda,n}(\hat{\boldsymbol{h}}_n^{(1)}) - \hat{\mathbb{M}}_{\lambda,n}(\boldsymbol{h}_0^{(1)})\Big\} \ge \delta/3\Big]$ and $\mathbb{P}\Big[c_n\Big\{\hat{\mathbb{M}}_{\lambda,n}(\boldsymbol{h}_0^{(1)}) - \mathbb{M}_\lambda(\boldsymbol{h}_0^{(1)})\Big\} \ge \delta/3\Big]$. From theorem 1, the first and third terms go to zero as $n \to \infty$ while the second term is zero since $\hat{\mathbb{M}}_{\lambda,n}(\hat{\boldsymbol{h}}_n^{(1)}) \le \hat{\mathbb{M}}_{\lambda,n}(\boldsymbol{h}_0^{(1)})$ as $\boldsymbol{h}_0^{(1)} \in \boldsymbol{H}_n$. This proves the statement of theorem 2.

To prove theorem 4 first note that from equation (14),

$$\Big\|\hat{\boldsymbol{w}}_n^{(0)}(\lambda) - \boldsymbol{w}_p^{(0)}\Big\|_2^2 = \sum_{i=1}^n \Big\{\frac{Y_i + 1}{[Y_i + \hat{h}_{n,i}^{(0)}(\lambda)][Y_i + h_{0,i}^{(0)}]}\Big\}^2 \Big\{\hat{h}_{n,i}^{(0)}(\lambda) - h_{0,i}^{(0)}\Big\}^2.$$

From assumption (A2) and Lemma 4, there exists a constant $c_0 > 0$ such that for large n, $\max_{1\le i\le n}(Y_i + 1) \le c_0 \log n$ with high probability. Moreover for $i = 1, \cdots, n$, since $\hat{w}_{n,i}^{(0)}(\lambda) > 0$ for every $\lambda \in \Lambda$ and $w_{p,i}^{(0)} > 0$, equation (14) implies $\hat{h}_{n,i}^{(0)}(\lambda) + Y_i > 0$ and $h_{0,i}^{(0)} + Y_i > 0$. Thus, conditional on the event $\{\max_{1\le i\le n}(Y_i + 1) \le c_0 \log n\}$ we have for some constant $c_1 > 0$,

$$\Big\|\hat{\boldsymbol{w}}_n^{(0)}(\lambda) - \boldsymbol{w}_p^{(0)}\Big\|_2^2 \le c_1 \log^2 n \Big\|\hat{\boldsymbol{h}}_n^{(0)}(\lambda) - \boldsymbol{h}_0^{(0)}\Big\|_2^2$$

and for any $\epsilon > 0$,

$$\mathbb{P}\Big[\frac{c_n}{n\log^2 n}\Big\|\hat{\boldsymbol{w}}_n^{(0)}(\lambda) - \boldsymbol{w}_p^{(0)}\Big\|_2^2 \ge \epsilon\Big] \le \mathbb{P}\Big[c_1\frac{c_n}{n}\Big\|\hat{\boldsymbol{h}}_n^{(0)}(\lambda) - \boldsymbol{h}_0^{(0)}\Big\|_2^2 \ge \epsilon\Big].$$

The proof of the statement of theorem 4 then follows from the proof of theorem 2 above and lemma 4.

B.4 Proofs of Theorems 3 and 5

We will prove Theorem 3 while the proof of Theorem 5 will follow using similar arguments and Theorem 4.

Note that,

$$\Big\|\boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\hat{\lambda}) - \boldsymbol{\delta}^{\pi}_{(1)}\Big\|_2^2 = \sum_{i=1}^{n}\Big[\frac{a_{Y_i-1}/a_{Y_i}}{\hat{w}^{(1)}_{n,i}(\hat{\lambda})w^{(1)}_{p,i}}\Big]^2\Big[\hat{w}^{(1)}_{n,i}(\hat{\lambda}) - w^{(1)}_{p,i}\Big]^2.$$

Now, $\hat{w}^{(1)}_{n,i}(\lambda) > 0$ for every $\lambda \in \Lambda$ and $w^{(1)}_{p,i} > 0$. This fact along with assumption (A2) and lemma 4 imply that there exists a constant $c_0 > 0$ such that $||\boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\hat{\lambda}) - \boldsymbol{\delta}^{\pi}_{(1)}||_2^2 \leq c_0 \log^2 n \|\hat{\boldsymbol{w}}^{(1)}_n(\hat{\lambda}) - \boldsymbol{w}^{(1)}_p\|_2^2$. The first result thus follows from the above inequality and Theorem 2.

To prove the second part of the theorem, note that $|\mathcal{L}^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(1)}) - \mathcal{L}^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\hat{\lambda}))|$ equals

$$\Big|\sqrt{\mathcal{L}^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(1)})} - \sqrt{\mathcal{L}^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\hat{\lambda}))}\Big|\Big|\sqrt{\mathcal{L}^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(1)})} + \sqrt{\mathcal{L}^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\hat{\lambda}))}\Big|$$

and Triangle inequality implies

$$\begin{aligned}\Big|\sqrt{\mathcal{L}^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(1)})} - \sqrt{\mathcal{L}^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\hat{\lambda}))}\Big| &\leq \Big|\sqrt{\frac{1}{n}\sum_{i=1}^{n}\Big(\delta^{\mathsf{neb}}_{(1),i}(\hat{\lambda}) - \delta^{\pi}_{(1),i}\Big)^2/\theta_i}\,\Big| \\ &\leq \frac{c_0}{\sqrt{n}}\Big\|\boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\hat{\lambda}) - \boldsymbol{\delta}^{\pi}_{(1)}\Big\|_2 = O_p\Big(\frac{\log^2 n}{n^{1/4}}\Big) \quad (22)\end{aligned}$$

from the first part of theorem 3. Thus, it follows from equation (22) that

$$\sqrt{\mathcal{L}^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\hat{\lambda}))} \leq \sqrt{\mathcal{L}^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(1)})} + O_p\Big(\frac{\log^2 n}{n^{1/4}}\Big)$$

and

$$\Big|\mathcal{L}^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(1)}) - \mathcal{L}^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\hat{\lambda}))\Big| \leq 4\sqrt{\mathcal{L}^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(1)})}\,\Big|\sqrt{\mathcal{L}^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(1)})} - \sqrt{\mathcal{L}^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\hat{\lambda}))}\Big| \quad (23)$$

Now from assumption (A2) and the proof of Lemma 4, $\mathcal{L}^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(1)}) \leq c_1 \log^2 n$ for some constant $c_1 > 0$. Therefore, together with equations (22) and (23) we have

$$\Big|\mathcal{L}^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(1)}) - \mathcal{L}^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\hat{\lambda}))\Big| = O_p\Big(\frac{\log^3 n}{n^{1/4}}\Big).$$

Define $Z_n(\hat{\lambda}) = \mathcal{L}^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(1)}) - \mathcal{L}^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\hat{\lambda}))$. We have already shown that $Z_n(\hat{\lambda}) \to 0$ in probability as $n \to \infty$. Moreover, under the Poisson model with $\delta^{\mathsf{neb}}_{(1),i}(\hat{\lambda}) \leq Y_i/d_1$ and $\delta^{\pi}_{(1),i} \leq Y_i/w^{(1)}_p(Y_i)$ where $w^{(1)}_p(Y_i) > 0$, we have for some positive constants c_0, c_1

$$|Z_n(\hat{\lambda})| \leq \frac{1}{n}\sum_{i=1}^{n}\Big\{c_0 Y_i + c_1 Y_i^2\Big\} := U_n. \quad (24)$$

Now, under the Poisson model and from assumption $(A2)$, $\sup_n \mathbb{E}(U_n^{1+\gamma}) < \infty$ for some $\gamma > 0$. Thus $\{U_n\}$ is uniformly integrable. Therefore, from equation (24), $\{Z_n(\hat{\lambda})\}$ is uniformly integrable and along with the fact that $Z_n(\hat{\lambda}) \to 0$ in probability as $n \to \infty$, we have $\mathbb{E}|Z_n(\hat{\lambda})| \to 0$ as $n \to \infty$. This proves the desired result under the Poisson model. For the Binomial model, with $m < \infty$, the result continues to hold since $\delta^{\mathsf{neb}}_{(1),i}(\lambda) \le m/d_1$ and $\delta^{\pi}_{(1),i} \le m/w_p^{(1)}(Y_i)$ where $w_p^{(1)}(Y_i) > 0$. Thus, $|Z_n(\hat{\lambda})| < \infty$ and so $\{Z_n(\hat{\lambda})\}$ is still uniformly integrable.

B.5 Proof of Lemma 2 - Binomial model

We will first prove the two statements of lemma 2 under the scaled squared error loss and conditional on the event B_n which is the event that $\{\max_{1\le i\le n} Y_i \le C \log n\}$. Under assumption $(A2)$, lemma 4 guarantees that B_n holds with high probability. Throughout the proof, we shall denote $d_1 := \inf_{\lambda\in\Lambda} \inf_{1\le i\le n}(1-\hat{h}^{(1)}_{n,i}(\lambda)) > 0$, $d_2 := \inf_{\lambda\in\Lambda} \inf_{1\le i\le n} \hat{w}^{(0)}_{n,i}(\lambda) > 0$ and assume $m < \infty$. Moreover, we will use the fact that under the Binomial model, $|\hat{h}^{(k)}_{n,i}(\lambda)| < \infty$ uniformly in $\lambda \in \Lambda$. This is a consequence of $d_1 > 0,\ d_2 > 0$ and $m < \infty$. The proof for the squared error loss will follow from similar arguments and we will highlight only the important steps.

Proof of statement 1 (Binomial model) for the scaled squared error loss $(k = 1)$

First note that under the Binomial model, $y_i \le m$ and $a_{y_i+1}/a_{y_i} = (m - y_i)/(y_i + 1) \le m$. Now,

$$\sup_{\lambda\in\Lambda}\Big|\mathsf{ARE}^{(1)}_n(\lambda, \boldsymbol{Y}) - \rho^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda))\Big| = \sup_{\lambda\in\Lambda}\frac{1}{n}\Big|\sum_{i=1}^{n}\Big\{[\delta^{\mathsf{neb}}_{(1),j_i}(\lambda)]^2\Big(\frac{a_{y_i+1}}{a_{y_i}}\Big) - \frac{[\delta^{\mathsf{neb}}_{(1),i}(\lambda)]^2}{\theta_i}\Big\}\Big|$$

$$\begin{aligned}
&\le \sup_{\lambda\in\Lambda}\frac{m}{n}\Big|\sum_{i=1}^{n}\Big\{[\delta^{\mathsf{neb}}_{(1),j_i}(\lambda)]^2 - [\delta^{\pi}_{(1),j_i}]^2\Big\}\Big| + \sup_{\lambda\in\Lambda}\frac{c_0}{n}\Big|\sum_{i=1}^{n}\Big\{[\delta^{\pi}_{(1),i}]^2 - [\delta^{\mathsf{neb}}_{(1),i}(\lambda)]^2\Big\}\Big| \\
&+ \frac{1}{n}\Big|\sum_{i=1}^{n}\Big\{[\delta^{\pi}_{(1),j_i}]^2\Big(\frac{a_{y_i+1}}{a_{y_i}}\Big) - \frac{[\delta^{\pi}_{(1),i}]^2}{\theta_i}\Big\}\Big| := T_1 + T_2 + T_3. \qquad (25)
\end{aligned}$$

Here we have used the fact that $a_{y_i+1}/a_{y_i} \le m$ and since $\theta_i > 0$, $1/\theta_i < c_0$ for some positive constant c_0. Consider the term T_3 in equation (25) above and define

$$V_i = [\delta^{\pi}_{(1),j_i}]^2\Big(\frac{a_{y_i+1}}{a_{y_i}}\Big) - \frac{[\delta^{\pi}_{(1),i}]^2}{\theta_i}.$$

Note that from lemma 6, $\mathbb{E}V_i = 0$. Moreover, V_i are independent and $\mathbb{E}|V_i|^2 < \infty$ since $|V_i| \le c_1 m^3$ for some constant $c_1 > 0$. The bound on $|V_i|$ follows from the fact that for any i, $\delta^{\pi}_{(1),i} := \delta^{\pi}_{(1),i}(y_i) \le m/w_p^{(1)}(y_i)$ where $w_p^{(1)}(y_i) > 0$. So, T_3 is $O_p(n^{-1/2})$.

We now consider the second term T_2 in equation (25) and define $Z_n(\lambda) = n^{-1}\sum_{i=1}^n\{\delta^{\pi}_{(1),i} - \delta^{\mathsf{neb}}_{(1),i}(\lambda)\}$. Note that under the Binomial model $\delta^{\mathsf{neb}}_{(1),i}(\lambda) \le m/d_1$ and so for any $\lambda \in \Lambda$

$$\frac{c_0}{n}\Big|\sum_{i=1}^n\Big\{[\delta^{\pi}_{(1),i}]^2 - [\delta^{\mathsf{neb}}_{(1),i}(\lambda)]^2\Big\}\Big| \le c_2\Big|Z_n(\lambda)\Big| \le \frac{c_2}{n}\Big\|\boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda) - \boldsymbol{\delta}^{\pi}_{(1)}\Big\|_1 \le \frac{c_2}{\sqrt{n}}\Big\|\boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda) - \boldsymbol{\delta}^{\pi}_{(1)}\Big\|_2 \tag{26}$$

for some constant $c_2 > 0$. The last term in the inequality above is $O_p(\log^2 n/n^{1/4})$ from the first part of theorem 3. Next for a perturbation λ' of λ such that $(\lambda, \lambda') \in \Lambda \coloneqq [\lambda_l, \lambda_u]$, we will bound the increments $|Z_n(\lambda) - Z_n(\lambda')|$. To that effect, note that

$$n\Big|Z_n(\lambda) - Z_n(\lambda')\Big| \le \Big\|\boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda) - \boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda')\Big\|_1 \le \frac{m}{d_1^2}\Big\|\hat{\boldsymbol{h}}^{(1)}_n(\lambda) - \hat{\boldsymbol{h}}^{(1)}_n(\lambda')\Big\|_1.$$

Now from lemma 5 we know that

$$\|\hat{\boldsymbol{h}}^{(1)}_n(\lambda) - \hat{\boldsymbol{h}}^{(1)}_n(\lambda')\|_1 \le n^{1/2}c^{-1}|\lambda - \lambda'| \sup_{\boldsymbol{h}\in N_\delta(\hat{\boldsymbol{h}}^{(1)}_n(\lambda'))} \|\nabla^2_{\hat{\boldsymbol{h}}^{(1)}_n(\lambda),\lambda}\hat{\mathbb{M}}_{\lambda',n}(\boldsymbol{h}) + o(1)\|_2.$$

However, under the Binomial model with $m < \infty$, $|\hat{h}^{(1)}_{n,i}(\lambda)| < \infty$ uniformly in $\lambda \in \Lambda$. Thus, the supremum in the display above is finite. So,

$$\Big|Z_n(\lambda) - Z_n(\lambda')\Big| \le \frac{c_3}{\sqrt{n}}|\lambda - \lambda'|.$$

Thus $Z_n(\lambda)$ is Lipschitz continuous in $\lambda \in \Lambda$ and, along with the compactness of Λ, it implies that there exists a $(\lambda_1, \lambda_2) \in \Lambda$ such that $\sup_{\lambda\in\Lambda} Z_n(\lambda) = Z_n(\lambda_1)$ and $\inf_{\lambda\in\Lambda} Z_n(\lambda) = Z_n(\lambda_2)$. Therefore, taking the supremum with respect λ in equation (26) and using the first part of theorem 3 suffice to prove that T_2 is $O_p(\log^2 n/n^{1/4})$. Finally, the first term T_1 in equation (26) is $O_p(\log^2 n/n^{1/4})$ which follows using similar arguments for the term T_2. Therefore, we have the desired result that $\sup_{\lambda\in\Lambda}|\mathsf{ARE}^{(1)}_n(\lambda, \boldsymbol{Y}) - \rho^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda))|$ is $O_p(\log^2 n/n^{1/4})$.

Proof of statement 2 (Binomial model) for the scaled squared error loss $(k = 1)$

From Triangle inequality, $\sup_{\lambda\in\Lambda}|\mathsf{ARE}^{(1)}_n(\lambda, \boldsymbol{Y}) - \mathbb{E}\rho^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda))| \le T_1 + T_2 + T_3 + T_4$ where $T_1 \coloneqq \sup_{\lambda\in\Lambda}|\mathsf{ARE}^{(1)}_n(\lambda, \boldsymbol{Y}) - \rho^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda))|$, $T_2 \coloneqq \sup_{\lambda\in\Lambda}|\rho^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda)) - \rho^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(1)})|$, $T_3 \coloneqq |\rho^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(1)}) - \mathbb{E}\rho^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(1)})|$ and $T_4 \coloneqq \sup_{\lambda\in\Lambda}|\mathbb{E}\rho^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(1)}) - \mathbb{E}\rho^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda))|$.

From statement 1 of lemma 2, T_1 is $O_p(\log^2 n/n^{1/4})$. Moreover, under the Binomial model $|\rho^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(1)})| \le c_0 m^2$ and so T_3 is $O_p(n^{-1/2})$.

We will now consider the term T_2. Define $Z_n(\lambda) = \rho^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda)) - \rho^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(1)})$ and note that for the Binomial model the proof of the second part of theorem 3 implies that $|Z_n(\lambda)|$ is $O_p(\log^2 n/n^{1/4})$ for any $\lambda \in \Lambda$. Moreover, for $(\lambda, \lambda') \in \Lambda$

$$\Big|Z_n(\lambda) - Z_n(\lambda')\Big| \le \frac{c_0}{n}\Big\|\boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda) - \boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda')\Big\|_1 \le \frac{c_1}{n}\Big\|\hat{\boldsymbol{h}}^{(1)}_n(\lambda) - \hat{\boldsymbol{h}}^{(1)}_n(\lambda')\Big\|_1.$$

Now using Lemma 5 and the fact that under the Binomial model with $m < \infty$, $|\hat{h}^{(1)}_{n,i}(\lambda)| < \infty$ uniformly in $\lambda \in \Lambda$, we conclude

$$\Big|Z_n(\lambda) - Z_n(\lambda')\Big| \leq \frac{c_3}{\sqrt{n}}|\lambda - \lambda'|.$$

Thus $Z_n(\lambda)$ is Lipschitz continuous in $\lambda \in \Lambda$ and, along with the compactness of Λ, it implies that there exists a $(\lambda_1, \lambda_2) \in \Lambda$ such that $\sup_{\lambda\in\Lambda} Z_n(\lambda) = Z_n(\lambda_1)$ and $\inf_{\lambda\in\Lambda} Z_n(\lambda) = Z_n(\lambda_2)$. Therefore, T_2 is $O_p(\log^2 n/n^{1/4})$.

Our desired result will follow if we can now show that $T_4 \to 0$ as $n \to \infty$. To do that we consider the proof of the second part of theorem 3 which shows that for any $\lambda \in \Lambda$, $Z_n(\lambda) \to 0$ in probability as $n \to \infty$. Moreover, under the Binomial model with $\delta^{\mathsf{neb}}_{(1),i}(\lambda) \leq m/d_1$ and $\delta^{\pi}_{(1),i} \leq m/w^{(1)}_p(y_i)$ where $w^{(1)}_p(y_i) > 0$, we have $|Z_n(\lambda)| < \infty$. Thus $\{Z_n(\lambda)\}$ is uniformly integrable and along with the fact that $Z_n(\lambda) \to 0$ in probability as $n \to \infty$, we have $\mathbb{E}|Z_n(\lambda)| \to 0$ as $n \to \infty$. Therefore, for any $\lambda \in \Lambda$ we have shown that $|\mathbb{E}\rho^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(1)}) - \mathbb{E}\rho^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda))| \to 0$ as $n \to \infty$. To prove the result uniformly in λ we note that $Z_n(\lambda)$ is Lipschitz continuous in $\lambda \in \Lambda$ and Λ is compact. So $T_4 \leq \mathbb{E}\sup_{\lambda\in\Lambda}|Z_n(\lambda)| = \mathbb{E}|Z_n(\lambda^*)|$ where $\lambda^* \in \Lambda$ is such that $\sup_{\lambda\in\Lambda}|Z_n(\lambda)| = |Z_n(\lambda)^*|$. Thus $T_4 \to 0$ as $n \to \infty$ which completes our proof.

Proof of statement 1 (Binomial model) for the squared error loss $(k = 0)$

Under the Binomial model, $y_i \leq m$ and $a_{y_i-1}/a_{y_i} = y_i/(m - y_i + 1) \leq m$. Now,

$$\sup_{\lambda\in\Lambda}\Big|\mathsf{ARE}^{(0)}_n(\lambda, \boldsymbol{Y}) - \rho^{(0)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(0)}(\lambda))\Big| = \sup_{\lambda\in\Lambda}\frac{2}{n}\Big|\sum_{i=1}^{n}\Big\{\delta^{\mathsf{neb}}_{(0),i}(\lambda)\theta_i - \delta^{\mathsf{neb}}_{(0),j_i}(\lambda)\Big(\frac{a_{y_i-1}}{a_{y_i}}\Big)\Big\}\Big|$$

$$\begin{aligned}
&\leq \frac{2}{n}\sup_{\lambda\in\Lambda}\Big|\sum_{i=1}^{n}\theta_i\Big\{\delta^{\mathsf{neb}}_{(0),i}(\lambda) - \delta^{\pi}_{(0),i}\Big\}\Big| + \frac{2}{n}\Big|\sum_{i=1}^{n}\Big\{\delta^{\pi}_{(0),i}\theta_i - \delta^{\pi}_{(0),j_i}\Big(\frac{a_{y_i-1}}{a_{y_i}}\Big)\Big\}\Big| \\
&+ \frac{2m}{n}\sup_{\lambda\in\Lambda}\Big|\sum_{i=1}^{n}\Big\{\delta^{\pi}_{(0),j_i} - \delta^{\mathsf{neb}}_{(0),j_i}(\lambda)\Big\}\Big| := T_1 + T_2 + T_3. \qquad (27)
\end{aligned}$$

Consider the term T_2 in equation (27) above and define

$$V_i = \delta^{\pi}_{(0),i}\theta_i - \delta^{\pi}_{(0),j_i}\Big(\frac{a_{y_i-1}}{a_{y_i}}\Big).$$

Note that from lemma 6 and conditional on θ_i, $\mathbb{E}_{Y_i|\theta_i}V_i = 0$. Moreover, V_i are independent and $|V_i| \leq c_0\theta_i$ for some constant $c_0 > 0$. The bound on $|V_i|$ follows from the fact that for any i, $\delta^{\pi}_{(0),i} := \delta^{\pi}_{(0),i}(y_i) \leq m/w^{(0)}_p(y_i)$ where $w^{(0)}_p(y_i) > 0$. Thus, applying Hoeffding's inequality to $n^{-1}|\sum_{i=1}^{n}V_i|$ we get that $T2$ is $O_p(\|\boldsymbol{\theta}\|_2/n)$. Now, from assumption $(A2)$ the distribution of θ has finite second moments which implies $T2$ is $O_p(n^{-1/2})$.

We now consider the second term T_1 in equation (27) and define $Z_n(\lambda) = n^{-1}\sum_{i=1}^{n}\theta_i\{\delta^{\mathsf{neb}}_{(0),i}(\lambda) - \delta^{\pi}_{(0),i}\}$. For any $\lambda \in \Lambda$, we have

$$\frac{2}{n}\Big|\sum_{i=1}^{n}\theta_i\Big\{\delta^{\mathsf{neb}}_{(0),i}(\lambda) - \delta^{\pi}_{(0),i}\Big\}\Big| \leq c_3\Big|Z_n(\lambda)\Big| \leq \frac{c_3}{n}\big\|\boldsymbol{\theta}\big\|_2\,\Big\|\boldsymbol{\delta}^{\mathsf{neb}}_{(0)}(\lambda) - \boldsymbol{\delta}^{\pi}_{(0)}\Big\|_2 \qquad (28)$$

for some constant $c_3 > 0$. From assumption $(A2)$ and the proof of theorem 5, the last term in the inequality above is $O_p(\log^3 n/n^{1/4})$. Next for a perturbation λ' of λ such that $(\lambda, \lambda') \in \Lambda \coloneqq [\lambda_l, \lambda_u]$, we will bound the increments $|Z_n(\lambda) - Z_n(\lambda')|$. To that effect, note that

$$n\Big|Z_n(\lambda) - Z_n(\lambda')\Big| \leq \big\|\boldsymbol{\theta}\big\|_2 \Big\|\boldsymbol{\delta}^{\mathsf{neb}}_{(0)}(\lambda) - \boldsymbol{\delta}^{\pi}_{(0)}\Big\|_2 \leq c_4\big\|\boldsymbol{\theta}\big\|_2\Big\|\hat{\boldsymbol{h}}^{(0)}_n(\lambda) - \hat{\boldsymbol{h}}^{(0)}_n(\lambda')\Big\|_2.$$

Now from lemma 5 we know that

$$\|\hat{\boldsymbol{h}}^{(0)}_n(\lambda) - \hat{\boldsymbol{h}}^{(0)}_n(\lambda')\|_2 \leq c^{-1}|\lambda - \lambda'| \sup_{\boldsymbol{h} \in N_\delta(\hat{\boldsymbol{h}}^{(0)}_n(\lambda'))} \|\nabla^2_{\hat{\boldsymbol{h}}^{(0)}_n(\lambda),\lambda} \hat{\mathbb{M}}_{\lambda',n}(\boldsymbol{h}) + o(1)\|_2,$$

and the supremum in the display above is finite since $|\hat{h}^{(0)}_{n,i}(\lambda)| < \infty$ uniformly in $\lambda \in \Lambda$. So, from assumption $(A2)$ and Lemma 4

$$\Big|Z_n(\lambda) - Z_n(\lambda')\Big| \leq c_5 \frac{\log n}{\sqrt{n}}|\lambda - \lambda'|,$$

for n sufficiently large. Thus $Z_n(\lambda)$ is Lipschitz continuous in $\lambda \in \Lambda$ and, along with the compactness of Λ, it implies that there exists a $(\lambda_1, \lambda_2) \in \Lambda$ such that $\sup_{\lambda\in\Lambda} Z_n(\lambda) = Z_n(\lambda_1)$ and $\inf_{\lambda\in\Lambda} Z_n(\lambda) = Z_n(\lambda_2)$. Therefore, taking the supremum with respect λ in equation (28) and using the first part of theorem 5 suffice to prove that T_1 is $O_p(\log^3 n/n^{1/4})$. Finally, the third term T_3 in equation (28) is $O_p(\log^3 n/n^{1/4})$ which follows using similar arguments for the term T_1. Therefore, we have the desired result that $\sup_{\lambda\in\Lambda} |\mathsf{ARE}^{(0)}_n(\lambda, \boldsymbol{Y}) - \rho^{(0)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(0)}(\lambda))|$ is $O_p(\log^3 n/n^{1/4})$.

Proof of statement 2 (Binomial model) for the squared error loss $(k = 0)$

From Triangle inequality, $\sup_{\lambda\in\Lambda} |\mathsf{ARE}^{(0)}_n(\lambda, \boldsymbol{Y}) - \mathbb{E}\rho^{(0)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(0)}(\lambda))| \leq T_1 + T_2 + T_3 + T_4$ where $T_1 \coloneqq \sup_{\lambda\in\Lambda} |\mathsf{ARE}^{(0)}_n(\lambda, \boldsymbol{Y}) - \rho^{(0)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(0)}(\lambda))|$, $T_2 \coloneqq \sup_{\lambda\in\Lambda} |\rho^{(0)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(0)}(\lambda)) - \rho^{(0)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(0)})|$, $T_3 \coloneqq |\rho^{(0)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(0)}) - \mathbb{E}\rho^{(0)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(0)})|$ and $T_4 \coloneqq \sup_{\lambda\in\Lambda} |\mathbb{E}\rho^{(0)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(0)}) - \mathbb{E}\rho^{(0)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(0)}(\lambda))|$.

Under squared error loss, statement 1 of lemma 2 implies T_1 is $O_p(\log^3 n/n^{1/4})$. Moreover, under the Binomial model, $|\rho^{(0)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(0)})| \leq c_0 n^{-1}\|\boldsymbol{\theta}\|_1$. Thus, applying Hoeffding's inequality to T_3 and using assumption $(A2)$ we get that T_3 is $O_p(n^{-1/2})$.

We will now consider the term T_2. Define $Z_n(\lambda) = \rho^{(0)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(0)}(\lambda)) - \rho^{(0)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(0)})$ and note that for the Binomial model, the proof of the second part of theorem 5 implies that $|Z_n(\lambda)|$ is $O_p(\log^3 n/n^{1/4})$ for any $\lambda \in \Lambda$. Moreover, for $(\lambda, \lambda') \in \Lambda$

$$\Big|Z_n(\lambda) - Z_n(\lambda')\Big| \leq \frac{c_0}{n}\big\|\boldsymbol{\theta}\big\|_2 \Big\|\boldsymbol{\delta}^{\mathsf{neb}}_{(0)}(\lambda) - \boldsymbol{\delta}^{\mathsf{neb}}_{(0)}(\lambda')\Big\|_2 \leq \frac{c_1}{n}\big\|\boldsymbol{\theta}\big\|_2 \Big\|\hat{\boldsymbol{h}}^{(0)}_n(\lambda) - \hat{\boldsymbol{h}}^{(0)}_n(\lambda')\Big\|_2.$$

Now under the Binomial model and along with Lemmata 4 and 5, we conclude

$$\Big|Z_n(\lambda) - Z_n(\lambda')\Big| \leq c_2 \frac{\log n}{\sqrt{n}}|\lambda - \lambda'|$$

for n sufficiently large. Thus $Z_n(\lambda)$ is Lipschitz continuous in $\lambda \in \Lambda$ and, along with the compactness of Λ, it implies that there exists a $(\lambda_1, \lambda_2) \in \Lambda$ such that $\sup_{\lambda\in\Lambda} Z_n(\lambda) = Z_n(\lambda_1)$ and $\inf_{\lambda\in\Lambda} Z_n(\lambda) = Z_n(\lambda_2)$. Therefore, T_2 is $O_p(\log^3 n/n^{1/4})$.

Our desired result will follow if we can now show that $T_4 \to 0$ as $n \to \infty$. We already know from the proof of the second part of theorem 5 that for any $\lambda \in \Lambda$, $Z_n(\lambda) \to 0$ in probability as $n \to \infty$. Moreover, under the Binomial model with $\delta^{\mathsf{neb}}_{(0),i}(\lambda) \le c_3 m$ and $\delta^{\pi}_{(0),i} \le m/w^{(0)}_p(y_i)$ where $w^{(0)}_p(y_i) > 0$, we have

$$|Z_n(\lambda)| \le \frac{c_4}{n}\sum_{i=1}^{n} |\theta_i| \coloneqq U_n.$$

Now, from assumption $(A2)$, $\sup_n \mathbb{E}(U_n^{1+\gamma}) < \infty$ for some $\gamma > 0$. Thus $\{U_n\}$ is uniformly integrable. Therefore, $\{Z_n(\lambda)\}$ is uniformly integrable and along with the fact that $Z_n(\lambda) \to 0$ in probability as $n \to \infty$, we have $\mathbb{E}|Z_n(\lambda)| \to 0$ as $n \to \infty$. Therefore, for any $\lambda \in \Lambda$ we have shown that $|\mathbb{E}\rho^{(0)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\pi}_{(0)}) - \mathbb{E}\rho^{(0)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(0)}(\lambda))| \to 0$ as $n \to \infty$. To prove the result uniformly in λ we note that $Z_n(\lambda)$ is Lipschitz continuous in $\lambda \in \Lambda$ and Λ is compact. So $T_4 \le \mathbb{E}\sup_{\lambda\in\Lambda} |Z_n(\lambda)| = \mathbb{E}|Z_n(\lambda^*)|$ where $\lambda^* \in \Lambda$ is such that $\sup_{\lambda\in\Lambda} |Z_n(\lambda)| = |Z_n(\lambda)^*|$. Thus $T_4 \to 0$ as $n \to \infty$ which completes our proof.

B.6 Proof of Lemma 2 - Poisson model

Here we will prove the two statements of lemma 2 under the Poisson model. As in the Binomial case, the statements will be proved first under the scaled squared error loss and conditional on the event B_n which is the event that $\{\max_{1\le i\le n} Y_i \le C \log n\}$. Under assumption $(A2)$, lemma 4 guarantees that B_n holds with high probability. Throughout the proof, we will denote $d_1 \coloneqq \inf_{\lambda\in\Lambda} \inf_{1\le i\le n}(1 - \hat{h}^{(1)}_{n,i}(\lambda)) > 0$ and $d_2 \coloneqq \inf_{\lambda\in\Lambda} \inf_{1\le i\le n} \hat{w}^{(0)}_{n,i}(\lambda) > 0$. Moreover, in the proof we will use $|\hat{h}^{(k)}_{n,i}(\lambda)| < c_0 \log n$ uniformly in $\lambda \in \Lambda$ which is a consequence of lemma 4. The proof for the squared error loss will follow from similar arguments and we will highlight only the important steps.

Proof of statement 1 (Poisson model) for the scaled squared error loss $(k = 1)$

First note that under the Poisson model $a_{Y_i+1}/a_{Y_i} = 1/(Y_i + 1) \le 1$. Now,

$$\sup_{\lambda\in\Lambda} \Big|\mathsf{ARE}^{(1)}_n(\lambda, \boldsymbol{Y}) - \rho^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda))\Big| = \sup_{\lambda\in\Lambda} \frac{1}{n}\Big|\sum_{i=1}^{n}\Big\{[\delta^{\mathsf{neb}}_{(1),j_i}(\lambda)]^2\Big(\frac{a_{y_i+1}}{a_{y_i}}\Big) - \frac{[\delta^{\mathsf{neb}}_{(1),i}(\lambda)]^2}{\theta_i}\Big\}\Big|$$

$$\begin{aligned} &\le \sup_{\lambda\in\Lambda} \frac{1}{n}\Big|\sum_{i=1}^{n}\Big\{[\delta^{\mathsf{neb}}_{(1),j_i}(\lambda)]^2 - [\delta^{\pi}_{(1),j_i}]^2\Big\}\Big| + \sup_{\lambda\in\Lambda} \frac{c_1}{n}\Big|\sum_{i=1}^{n}\Big\{[\delta^{\pi}_{(1),i}]^2 - [\delta^{\mathsf{neb}}_{(1),i}(\lambda)]^2\Big\}\Big| \\ &+ \frac{1}{n}\Big|\sum_{i=1}^{n}\Big\{[\delta^{\pi}_{(1),j_i}]^2\Big(\frac{a_{y_i+1}}{a_{y_i}}\Big) - \frac{[\delta^{\pi}_{(1),i}]^2}{\theta_i}\Big\}\Big| \coloneqq T_1 + T_2 + T_3. \end{aligned} \tag{29}$$

Here we have used the fact that $a_{y_i+1}/a_{y_i} \leq 1$ and since $\theta_i > 0$, $1/\theta_i < c_1$ for some positive constant c_1. Consider the term T_3 in equation (29) above and define

$$V_i = [\delta^{\pi}_{(1),j_i}]^2\Big(\frac{a_{y_i+1}}{a_{y_i}}\Big) - \frac{[\delta^{\pi}_{(1),i}]^2}{\theta_i}.$$

Note that from lemma 6, $\mathbb{E}V_i = 0$. Moreover, V_i are independent and $|V_i| \leq c_2 \log^2 n$ for some constant $c_2 > 0$. The bound on $|V_i|$ follows from the fact that for any i and conditional on B_n, $\delta^{\pi}_{(1),i} := \delta^{\pi}_{(1),i}(y_i) \leq c_0 \log n / w^{(1)}_p(y_i)$ where $w^{(1)}_p(y_i) > 0$. Thus, applying Hoeffding's inequality to $n^{-1}|\sum_{i=1}^n V_i|$ we get that $T3$ is $O_p(\log^2 n/\sqrt{n})$.

We now consider the second term T_2 in equation (29) and define $Z_n(\lambda) = n^{-1}\sum_{i=1}^n \{\delta^{\pi}_{(1),i} - \delta^{\mathsf{neb}}_{(1),i}(\lambda)\}$. Note that under the Poisson model $\delta^{\mathsf{neb}}_{(1),i}(\lambda) \leq Y_i/d_1$ and so for any $\lambda \in \Lambda$

$$\frac{c_1}{n}\Big|\sum_{i=1}^n \Big\{[\delta^{\pi}_{(1),i}]^2 - [\delta^{\mathsf{neb}}_{(1),i}(\lambda)]^2\Big\}\Big| \leq c_3 \log n \Big|Z_n(\lambda)\Big| \leq \frac{c_3 \log n}{n}\Big\|\boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda) - \boldsymbol{\delta}^{\pi}_{(1)}\Big\|_1 \tag{30}$$

for some constant $c_3 > 0$. In equation (30), we have used the fact that under assumption $(A2)$ and lemma 4, $Y_i \leq c_0 \log n$ with high probability. Moreover, the last term in the inequality in (30) is $O_p(\log^3 n/n^{1/4})$ since $\|\boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda) - \boldsymbol{\delta}^{\pi}_{(1)}\|_1 \leq n^{1/2}\|\boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda) - \boldsymbol{\delta}^{\pi}_{(1)}\|_2$ and $n^{-1/2}\|\boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda) - \boldsymbol{\delta}^{\pi}_{(1)}\|_2$ is $O_p(\log^2 n/n^{1/4})$ from the first part of theorem 3. Next for a perturbation λ' of λ such that $(\lambda, \lambda') \in \Lambda$, we will bound the increments $|Z_n(\lambda) - Z_n(\lambda')|$. To that effect, note that conditional on B_n

$$n\Big|Z_n(\lambda) - Z_n(\lambda')\Big| \leq \Big\|\boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda) - \boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda')\Big\|_1 \leq \frac{\log n}{d_1^2}\Big\|\hat{\boldsymbol{h}}^{(1)}_n(\lambda) - \hat{\boldsymbol{h}}^{(1)}_n(\lambda')\Big\|_1.$$

Now from lemma 5 we know that

$$\|\hat{\boldsymbol{h}}^{(1)}_n(\lambda) - \hat{\boldsymbol{h}}^{(1)}_n(\lambda')\|_1 \leq n^{1/2}c^{-1}|\lambda - \lambda'| \sup_{\boldsymbol{h} \in N_\delta(\hat{\boldsymbol{h}}^{(1)}_n(\lambda))} \|\nabla^2_{\boldsymbol{h}_n,\lambda}\hat{\mathbb{M}}_{\lambda,n}(\boldsymbol{h}) + o(1)\|_2.$$

However, under the Poisson model, assumption $(A2)$ and lemma 4, $|\hat{h}^{(1)}_{n,i}(\lambda)| < c_4 \log n$ uniformly in $\lambda \in \Lambda$. Thus for n sufficiently large, the supremum in the display above is much smaller than $\log n$. So,

$$\Big|Z_n(\lambda) - Z_n(\lambda')\Big| \leq c_5 \frac{\log^2 n}{\sqrt{n}}|\lambda - \lambda'|.$$

Thus $Z_n(\lambda)$ is Lipschitz continuous in $\lambda \in \Lambda$ and, along with the compactness of Λ, it implies that there exists a $(\lambda_1, \lambda_2) \in \Lambda$ such that $\sup_{\lambda\in\Lambda} Z_n(\lambda) = Z_n(\lambda_1)$ and $\inf_{\lambda\in\Lambda} Z_n(\lambda) = Z_n(\lambda_2)$. Therefore, taking the supremum with respect λ in equation (30) and using the first part of theorem 3 suffice to prove that T_2 is $O_p(\log^3 n/n^{1/4})$. Finally, the first term T_1 in equation (30) is $O_p(\log^3 n/n^{1/4})$ which follows using similar arguments for the term T_2. Therefore, we have the desired result that $\sup_{\lambda\in\Lambda} |\mathsf{ARE}^{(1)}_n(\lambda, \boldsymbol{Y}) - \rho^{(1)}_n(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathsf{neb}}_{(1)}(\lambda))|$ is $O_p(\log^3 n/n^{1/4})$.

Proof of statement 2 (Poisson model) for the scaled squared error loss $(k = 1)$

From Triangle inequality, $\sup_{\lambda\in\Lambda} |\mathsf{ARE}_n^{(1)}(\lambda, \boldsymbol{Y}) - \mathbb{E}\rho_n^{(1)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\lambda))| \le T_1 + T_2 + T_3 + T_4$ where $T_1 \coloneqq \sup_{\lambda\in\Lambda} |\mathsf{ARE}_n^{(1)}(\lambda, \boldsymbol{Y}) - \rho_n^{(1)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\lambda))|$, $T_2 \coloneqq \sup_{\lambda\in\Lambda} |\rho_n^{(1)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\lambda)) - \rho_n^{(1)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\pi})|$, $T_3 \coloneqq |\rho_n^{(1)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\pi}) - \mathbb{E}\rho_n^{(1)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\pi})|$ and $T_4 \coloneqq \sup_{\lambda\in\Lambda} |\mathbb{E}\rho_n^{(1)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\pi}) - \mathbb{E}\rho_n^{(1)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\lambda))|$.

From statement 1 of lemma 2, T_1 is $O_p(\log^3 n/n^{1/4})$ for the Poisson model. Moreover, under the Poisson model and conditional on B_n, $|\rho_n^{(1)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\pi})| \le c_0 \log^2 n$. Thus, applying Hoeffding's inequality to T_3 we get that T_3 is $O_p(\log^2 n/n^{-1/2})$.

We will now consider the term T_2. Define $Z_n(\lambda) = \rho_n^{(1)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\lambda)) - \rho_n^{(1)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\pi})$ and note that from the proof of the second part of theorem 3 $|Z_n(\lambda)|$ is $O_p(\log^3 n/n^{1/4})$ for any $\lambda \in \Lambda$. Moreover, for $(\lambda, \lambda') \in \Lambda$ and conditional on B_n

$$\Big|Z_n(\lambda) - Z_n(\lambda')\Big| \le \frac{c_0 \log n}{n} \Big\|\boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\lambda) - \boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\lambda')\Big\|_1 \le \frac{c_1 \log^2 n}{n} \Big\|\hat{\boldsymbol{h}}_n^{(1)}(\lambda) - \hat{\boldsymbol{h}}_n^{(1)}(\lambda')\Big\|_1.$$

Thus, from Lemma 5, the Poisson model and assumption $(A2)$ we have

$$\Big|Z_n(\lambda) - Z_n(\lambda')\Big| \le c_3 \frac{\log^3 n}{\sqrt{n}} |\lambda - \lambda'|.$$

Therefore, $Z_n(\lambda)$ is Lipschitz continuous in $\lambda \in \Lambda$ and, along with the compactness of Λ, it implies that there exists a $(\lambda_1, \lambda_2) \in \Lambda$ such that $\sup_{\lambda\in\Lambda} Z_n(\lambda) = Z_n(\lambda_1)$ and $\inf_{\lambda\in\Lambda} Z_n(\lambda) = Z_n(\lambda_2)$. Thus, T_2 is $O_p(\log^3 n/n^{1/4})$.

Our desired result will follow if we can now show that $T_4 \to 0$ as $n \to \infty$. To do that we consider the proof of the second part of theorem 3 which shows that for any $\lambda \in \Lambda$, $Z_n(\lambda) \to 0$ in probability as $n \to \infty$. Moreover, under the Poisson model with $\delta_{(1),i}^{\mathsf{neb}}(\lambda) \le Y_i/d_1$ and $\delta_{(1),i}^{\pi} \le Y_i/w_p^{(1)}(Y_i)$ where $w_p^{(1)}(Y_i) > 0$, we have for some positive constants c_0, c_1

$$|Z_n(\lambda)| \le \frac{1}{n} \sum_{i=1}^{n} \Big\{c_0 Y_i + c_1 Y_i^2\Big\} \coloneqq U_n. \tag{31}$$

Now, the Poisson model and assumption $(A2)$ imply that $\sup_n \mathbb{E}(U_n^{1+\gamma}) < \infty$ for some $\gamma > 0$. Thus $\{U_n\}$ is uniformly integrable. Therefore, from equation (31), $\{Z_n(\lambda)\}$ is uniformly integrable and along with the fact that $Z_n(\lambda) \to 0$ in probability as $n \to \infty$, we have $\mathbb{E}|Z_n(\lambda)| \to 0$ as $n \to \infty$. So, for any $\lambda \in \Lambda$ we have shown that $|\mathbb{E}\rho_n^{(1)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\pi}) - \mathbb{E}\rho_n^{(1)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(1)}^{\mathsf{neb}}(\lambda))| \to 0$ as $n \to \infty$. To prove the result uniformly in λ we note that $Z_n(\lambda)$ is Lipschitz continuous in $\lambda \in \Lambda$ and Λ is compact. Therefore, $T_4 \le \mathbb{E}\sup_{\lambda\in\Lambda} |Z_n(\lambda)| = \mathbb{E}|Z_n(\lambda^*)|$ where $\lambda^* \in \Lambda$ is such that $\sup_{\lambda\in\Lambda} |Z_n(\lambda)| = |Z_n(\lambda)^*|$. Thus $T_4 \to 0$ as $n \to \infty$ which completes our proof.

Proof of statement 1 (Poisson model) for the squared error loss $(k = 0)$

First note that under the Poisson model $a_{Y_i-1}/a_{Y_i} = Y_i$. Now,

$$\sup_{\lambda\in\Lambda} \Big|\mathsf{ARE}_n^{(0)}(\lambda, \boldsymbol{Y}) - \rho_n^{(0)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(0)}^{\mathsf{neb}}(\lambda))\Big| = \sup_{\lambda\in\Lambda} \frac{2}{n} \Big| \sum_{i=1}^{n} \Big\{\delta_{(0),i}^{\mathsf{neb}}(\lambda)\theta_i - \delta_{(0),j_i}^{\mathsf{neb}}(\lambda)\Big(\frac{a_{y_i-1}}{a_{y_i}}\Big)\Big\}\Big|$$

$$\leq \frac{2}{n}\sup_{\lambda\in\Lambda}\Big|\sum_{i=1}^{n}\theta_i\Big\{\delta^{\mathsf{neb}}_{(0),i}(\lambda)-\delta^{\pi}_{(0),i}\Big\}\Big|+\frac{2}{n}\Big|\sum_{i=1}^{n}\Big\{\delta^{\pi}_{(0),i}\theta_i-\delta^{\pi}_{(0),j_i}\Big(\frac{a_{y_i-1}}{a_{y_i}}\Big)\Big\}\Big|$$

$$+ \frac{c_1\log n}{n}\sup_{\lambda\in\Lambda}\Big|\sum_{i=1}^{n}\Big\{\delta^{\pi}_{(0),j_i}-\delta^{\mathsf{neb}}_{(0),j_i}(\lambda)\Big\}\Big| \coloneqq T_1+T_2+T_3. \tag{32}$$

Here we have used the fact that we have used the fact that conditional on event B_n, $a_{Y_i-1}/a_{Y_i}\leq Y_i\leq c_0\log n$ for some positive constant c_0. Consider the term T_2 in equation (32) above and define

$$V_i=\delta^{\pi}_{(0),i}\theta_i-\delta^{\pi}_{(0),j_i}\Big(\frac{a_{y_i-1}}{a_{y_i}}\Big).$$

Note that from lemma 6 and conditional on θ_i, $\mathbb{E}_{Y_i|\theta_i}V_i=0$. Moreover, V_i are independent and $|V_i|\leq c_2\theta_i\log^2 n$ for some constant $c_2>0$. The bound on $|V_i|$ follows from the fact that for any i and conditional on B_n, $\delta^{\pi}_{(0),i}\coloneqq\delta^{\pi}_{(0),i}(y_i)\leq c_3\log n/w^{(0)}_p(y_i)$ where $w^{(0)}_p(y_i)>0$. Thus, applying Hoeffding's inequality to $n^{-1}|\sum_{i=1}^{n}V_i|$ we get that $T2$ is $O_p(\|\boldsymbol{\theta}\|_2\log^2 n/n)$. Now, from assumption $(A2)$ the distribution of θ has finite second moments which implies $T2$ is $O_p(\log^2 n/\sqrt{n})$.

We now consider the T_1 in equation (32) and define

$$Z_n(\lambda)=n^{-1}\sum_{i=1}^{n}\theta_i\{\delta^{\mathsf{neb}}_{(0),i}(\lambda)-\delta^{\pi}_{(0),i}\}.$$

Note that for any $\lambda\in\Lambda$,

$$\frac{2}{n}\Big|\sum_{i=1}^{n}\theta_i\Big\{\delta^{\mathsf{neb}}_{(0),i}(\lambda)-\delta^{\pi}_{(0),i}\Big\}\Big|\leq c_0\Big|Z_n(\lambda)\Big|\leq\frac{c_0}{n}\|\boldsymbol{\theta}\|_2\,\Big\|\boldsymbol{\delta}^{\mathsf{neb}}_{(0)}(\lambda)-\boldsymbol{\delta}^{\pi}_{(0)}\Big\|_2 \tag{33}$$

for some constant $c_0>0$. From assumption $(A2)$ and the proof of theorem 5, the last term in the inequality in (33) is $O_p(\log^3 n/n^{1/4})$. Next for a perturbation λ' of λ such that $(\lambda,\lambda')\in\Lambda$, we will bound the increments $|Z_n(\lambda)-Z_n(\lambda')|$. To that effect, note that conditional on event B_n

$$n\Big|Z_n(\lambda)-Z_n(\lambda')\leq\|\boldsymbol{\theta}\|_2\,\Big\|\boldsymbol{\delta}^{\mathsf{neb}}_{(0)}(\lambda)-\boldsymbol{\delta}^{\mathsf{neb}}_{(0)}(\lambda')\Big\|_2\leq c_1\log n\|\boldsymbol{\theta}\|_2\,\Big\|\hat{\boldsymbol{h}}^{(0)}_n(\lambda)-\hat{\boldsymbol{h}}^{(0)}_n(\lambda')\Big\|_2.$$

Now from lemma 5 we know that

$$\|\hat{\boldsymbol{h}}^{(0)}_n(\lambda)-\hat{\boldsymbol{h}}^{(0)}_n(\lambda')\|_2\leq c^{-1}|\lambda-\lambda'|\sup_{\boldsymbol{h}\in N_\delta(\hat{\boldsymbol{h}}^{(0)}_n(\lambda))}\|\nabla^2_{\boldsymbol{h}_n,\lambda}\hat{\mathbb{M}}_{\lambda,n}(\boldsymbol{h})+o(1)\|_2,$$

and for n sufficiently large, the supremum in the display above is much smaller than $\log n$. So, with assumption $(A2)$

$$\Big|Z_n(\lambda)-Z_n(\lambda')\Big|\leq c_2\frac{\log^2 n}{\sqrt{n}}|\lambda-\lambda'|.$$

Thus $Z_n(\lambda)$ is Lipschitz continuous in $\lambda\in\Lambda$ and, along with the compactness of Λ, it implies that there exists a $(\lambda_1,\lambda_2)\in\Lambda$ such that $\sup_{\lambda\in\Lambda}Z_n(\lambda)=Z_n(\lambda_1)$ and $\inf_{\lambda\in\Lambda}Z_n(\lambda)=$

$Z_n(\lambda_2)$. Therefore, taking the supremum with respect λ in equation (33) and using the first part of theorem 5 suffice to prove that T_1 is $O_p(\log^3 n/n^{1/4})$. Finally, the third term T_3 in equation (33) is $O_p(\log^3 n/n^{1/4})$ which follows using similar arguments for the term T_1. Therefore, we have the desired result that $\sup_{\lambda\in\Lambda}|\mathsf{ARE}_n^{(0)}(\lambda,\boldsymbol{Y})-\rho_n^{(0)}(\boldsymbol{\theta},\boldsymbol{\delta}_{(0)}^{\mathsf{neb}}(\lambda))|$ is $O_p(\log^3 n/n^{1/4})$.

Proof of statement 2 (Poisson model) for the squared error loss $(k=0)$

From Triangle inequality, $\sup_{\lambda\in\Lambda}|\mathsf{ARE}_n^{(0)}(\lambda,\boldsymbol{Y})-\mathbb{E}\rho_n^{(0)}(\boldsymbol{\theta},\boldsymbol{\delta}_{(0)}^{\mathsf{neb}}(\lambda))|\le T_1+T_2+T_3+T_4$ where $T_1 \coloneqq \sup_{\lambda\in\Lambda}|\mathsf{ARE}_n^{(0)}(\lambda,\boldsymbol{Y})-\rho_n^{(0)}(\boldsymbol{\theta},\boldsymbol{\delta}_{(0)}^{\mathsf{neb}}(\lambda))|$, $T_2 \coloneqq \sup_{\lambda\in\Lambda}|\rho_n^{(0)}(\boldsymbol{\theta},\boldsymbol{\delta}_{(0)}^{\mathsf{neb}}(\lambda))-\rho_n^{(0)}(\boldsymbol{\theta},\boldsymbol{\delta}_{(0)}^{\pi})|$, $T_3 \coloneqq |\rho_n^{(0)}(\boldsymbol{\theta},\boldsymbol{\delta}_{(0)}^{\pi})-\mathbb{E}\rho_n^{(0)}(\boldsymbol{\theta},\boldsymbol{\delta}_{(0)}^{\pi})|$ and $T_4 \coloneqq \sup_{\lambda\in\Lambda}|\mathbb{E}\rho_n^{(0)}(\boldsymbol{\theta},\boldsymbol{\delta}_{(0)}^{\pi})-\mathbb{E}\rho_n^{(0)}(\boldsymbol{\theta},\boldsymbol{\delta}_{(0)}^{\mathsf{neb}}(\lambda))|$.

Under squared error loss, statement 1 of lemma 2 implies T_1 is $O_p(\log^3 n/n^{1/4})$. Moreover, under the Poisson model and assumption $(A2)$, $|\rho_n^{(0)}(\boldsymbol{\theta},\boldsymbol{\delta}_{(0)}^{\pi})| \le c_0 n^{-1}\log^2 n\|\boldsymbol{\theta}\|_1$. Thus, applying Hoeffding's inequality to T_3 and using assumption $(A2)$ we get that T_3 is $O_p(\log^2 n/\sqrt{n})$.

We will now consider the term T_2. Define $Z_n(\lambda)=\rho_n^{(0)}(\boldsymbol{\theta},\boldsymbol{\delta}_{(0)}^{\mathsf{neb}}(\lambda))-\rho_n^{(0)}(\boldsymbol{\theta},\boldsymbol{\delta}_{(0)}^{\pi})$ and note that from the second part of theorem 5 $|Z_n(\lambda)|$ is $O_p(\log^4 n/n^{1/4})$ for any $\lambda\in\Lambda$. Moreover, for $(\lambda,\lambda')\in\Lambda$

$$\Big|Z_n(\lambda)-Z_n(\lambda')\Big| \le \frac{1}{n}\Big\|\boldsymbol{\delta}_{(0)}^{\mathsf{neb}}(\lambda)-\boldsymbol{\delta}_{(0)}^{\mathsf{neb}}(\lambda')\Big\|_2\Big\{\|\boldsymbol{\theta}\|_2+c_0\|\boldsymbol{Y}\|_2\Big\}.$$

Now conditional on the event B_n, $\|\boldsymbol{Y}\|_2 \le c_1\sqrt{n}\log n$ and assumption $(A2)$ implies that with high probability $\|\boldsymbol{\theta}\|_2/\sqrt{n} \le c_2\sqrt{\log n}$. Thus, for n sufficiently large

$$\Big|Z_n(\lambda)-Z_n(\lambda')\Big| \le \frac{c_3\log n}{\sqrt{n}}\Big\|\boldsymbol{\delta}_{(0)}^{\mathsf{neb}}(\lambda)-\boldsymbol{\delta}_{(0)}^{\mathsf{neb}}(\lambda')\Big\|_2 \le \frac{c_4\log^2 n}{\sqrt{n}}\Big\|\hat{\boldsymbol{h}}_n^{(0)}(\lambda)-\hat{\boldsymbol{h}}_n^{(0)}(\lambda')\Big\|_2,$$

and together with Lemma 5, we have

$$\Big|Z_n(\lambda)-Z_n(\lambda')\Big| \le c_5\frac{\log^3 n}{\sqrt{n}}|\lambda-\lambda'|.$$

Thus $Z_n(\lambda)$ is Lipschitz continuous in $\lambda\in\Lambda$ and, along with the compactness of Λ, it implies that there exists a $(\lambda_1,\lambda_2)\in\Lambda$ such that $\sup_{\lambda\in\Lambda}Z_n(\lambda)=Z_n(\lambda_1)$ and $\inf_{\lambda\in\Lambda}Z_n(\lambda)=Z_n(\lambda_2)$. Therefore, T_2 is $O_p(\log^4 n/n^{1/4})$.

Our desired result will follow if we can now show that $T_4\to 0$ as $n\to\infty$. We already know from the proof of the second part of theorem 5 that for any $\lambda\in\Lambda$, $Z_n(\lambda)\to 0$ in probability as $n\to\infty$. Moreover, under the Poisson model with $\delta_{(0),i}^{\mathsf{neb}}(\lambda)\le (Y_i+1)/d_2$ and $\delta_{(0),i}^{\pi}\le (Y_i+1)/w_p^{(0)}(y_i)$ where $w_p^{(0)}(y_i)>0$, we have

$$|Z_n(\lambda)| \le \frac{1}{n}\sum_{i=1}^{n}\Big\{c_0\theta_i Y_i+c_1 Y_i^2\Big\} \coloneqq U_n.$$

Now, the Poisson model and assumption $(A2)$ imply that $\sup_n\mathbb{E}(U_n^{1+\gamma})<\infty$ for some $\gamma>0$. Thus $\{U_n\}$ is uniformly integrable. Therefore, $\{Z_n(\lambda)\}$ is uniformly integrable and

along with the fact that $Z_n(\lambda) \to 0$ in probability as $n \to \infty$, we have $\mathbb{E}|Z_n(\lambda)| \to 0$ as $n \to \infty$. Therefore, for any $\lambda \in \Lambda$ we have shown that $|\mathbb{E}\rho_n^{(0)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(0)}^{\pi}) - \mathbb{E}\rho_n^{(0)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(0)}^{\mathsf{neb}}(\lambda))| \to 0$ as $n \to \infty$. To prove the result uniformly in λ we note that $Z_n(\lambda)$ is Lipschitz continuous in $\lambda \in \Lambda$ and Λ is compact. So $T_4 \leq \mathbb{E}\sup_{\lambda\in\Lambda}|Z_n(\lambda)| = \mathbb{E}|Z_n(\lambda^*)|$ where $\lambda^* \in \Lambda$ is such that $\sup_{\lambda\in\Lambda}|Z_n(\lambda)| = |Z_n(\lambda)^*|$. Thus $T_4 \to 0$ as $n \to \infty$ which completes our proof.

B.7 Proof of Lemma 3

The statement of this lemma follows from part (1) of Lemma 2. First note that by definition $\mathsf{ARE}_n^{(k)}(\hat{\lambda}, \boldsymbol{Y}) \leq \mathsf{ARE}_n^{(k)}(\lambda_k^{\mathsf{orc}}, \boldsymbol{Y})$. So for any $\epsilon > 0$ and $k \in \{0, 1\}$, the probability $\mathbb{P}\Big[\mathcal{L}_n^{(k)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(k)}^{\mathsf{neb}}(\hat{\lambda})) \geq \mathcal{L}_n^{(k)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(k)}^{\mathsf{or}}) + c_n^{-1}\epsilon\Big]$ is bounded above by

$$\mathbb{P}\Big[\mathcal{L}_n^{(k)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(k)}^{\mathsf{neb}}(\hat{\lambda})) - \mathsf{ARE}_n^{(k)}(\hat{\lambda}, \boldsymbol{Y}) \geq \mathcal{L}_n^{(k)}(\boldsymbol{\theta}, \boldsymbol{\delta}_{(k)}^{\mathsf{or}}) - \mathsf{ARE}_n^{(k)}(\lambda_k^{\mathsf{orc}}, \boldsymbol{Y}) + c_n^{-1}\epsilon\Big].$$

The above display converges to 0 by part (1) of Lemma 2.

B.8 Proofs of Lemmata 4, 5 and 6

Proof of Lemma 4

First note that from assumption $(A2)$ and for some $\delta > 0$, $\theta \leq \epsilon^{-(1+\delta)}\log n$ with high probability. We will now prove the statement of lemma 4 for the case when $Y_i|\theta_i \overset{ind.}{\sim} \mathtt{Poi}(\theta_i)$. For distributions with bounded support, like the Binomial model, the lemma follows trivially.

Under the Poisson model, we have $\mathbb{P}(Y_i \geq \theta_i + t) \leq \exp\{-0.5t^2/(\theta_i + t)\}$ for any $t > 0$. The above inequality follows from an application of Bennett inequality to the Poisson MGF (see Pollard (2015)). Now consider $\mathbb{P}(\max_{i=1,\ldots,n} Y_i \leq \theta_i + t)$ and note that since Y_i are all independent, this probability is given by $\prod_{i=1}^n[1 - \exp\{-0.5t^2/(\theta_i + t)\}]$. Take $t = s\log n$ where $s^2/\{s + \epsilon^{-(1+\delta)}\} > 4$. Then with $\theta_i \leq \epsilon^{-(1+\delta)}\log n$, the above probability is bounded above by $a_n = \{1 - n^{-(1+\nu)}\}^n$ for some $\nu > 0$. As $n \to \infty$, $a_n \to 1$ which proves the statement of the lemma.

Proof of Lemma 5

We begin with some remarks on the optimization problems (8) and (15). Note that the feasible set $\boldsymbol{H}_n$ in equation (8) (and (15)) is compact and independent of λ. Moreover, the optimization problem in definitions 1 and 3 is convex. Consequently, (i) for all $\lambda \in \Lambda$, the optimization takes place in a compact set, and (ii) the optimal solution set corresponding to any $\lambda \in \Lambda$ is a singleton, $\{\hat{\boldsymbol{h}}_n^{(k)}(\lambda)\}$. Now fix an $\epsilon > 0$. Then for any $\lambda \in N_\epsilon(\lambda_0) \cap \Lambda$ there exists a $\delta > 0$ such that the optimal solution $\boldsymbol{h}_n := \hat{\boldsymbol{h}}_n^{(k)}(\lambda) \in N_\delta(\hat{\boldsymbol{h}}_n^{(k)}(\lambda_0))$ and $\hat{\mathbb{M}}_{\lambda,n}\{\boldsymbol{h}_n\} - \hat{\mathbb{M}}_{\lambda,n}\{\hat{\boldsymbol{h}}_n^{(k)}(\lambda_0)\} \leq 0$. Moreover, we can re-write $\hat{\mathbb{M}}_{\lambda_0,n}\{\boldsymbol{h}_n\} - \hat{\mathbb{M}}_{\lambda_0,n}\{\hat{\boldsymbol{h}}_n^{(k)}(\lambda_0)\}$ as

$$\hat{\mathbb{M}}_{\lambda_0,n}\{\boldsymbol{h}_n\} - \hat{\mathbb{M}}_{\lambda,n}\{\boldsymbol{h}_n\} - \hat{\mathbb{M}}_{\lambda_0,n}\{\hat{\boldsymbol{h}}_n^{(k)}(\lambda_0)\} + \hat{\mathbb{M}}_{\lambda,n}\{\hat{\boldsymbol{h}}_n^{(k)}(\lambda_0)\} + \hat{\mathbb{M}}_{\lambda,n}\{\boldsymbol{h}_n\} - \hat{\mathbb{M}}_{\lambda,n}\{\hat{\boldsymbol{h}}_n^{(k)}(\lambda_0)\}$$

The last term in the display above is negative and thus we can upper bound $\hat{\mathbb{M}}_{\lambda_0,n}\{\boldsymbol{h}_n\} - \hat{\mathbb{M}}_{\lambda_0,n}\{\hat{\boldsymbol{h}}_n^{(k)}(\lambda_0)\}$ by

$$\hat{\mathbb{M}}_{\lambda_0,n}\{\boldsymbol{h}_n\} - \hat{\mathbb{M}}_{\lambda,n}\{\boldsymbol{h}_n\} - \hat{\mathbb{M}}_{\lambda_0,n}\{\hat{\boldsymbol{h}}_n^{(k)}(\lambda_0)\} + \hat{\mathbb{M}}_{\lambda,n}\{\hat{\boldsymbol{h}}_n^{(k)}(\lambda_0)\}$$

Now apply the mean value theorem with respect to $\boldsymbol{h}_n$ to the function $\hat{\mathbb{M}}_{\lambda_0,n}\{\boldsymbol{h}_n\} - \hat{\mathbb{M}}_{\lambda,n}\{\boldsymbol{h}_n\}$ in the display above and notice that $\hat{\mathbb{M}}_{\lambda_0,n}\{\boldsymbol{h}_n\} - \hat{\mathbb{M}}_{\lambda_0,n}\{\hat{\boldsymbol{h}}_n^{(k)}(\lambda_0)\}$ is bounded above by

$$\Big[\nabla_{\boldsymbol{h}_n}\Big\{\hat{\mathbb{M}}_{\lambda_0,n}(\bar{\boldsymbol{h}}_n) - \hat{\mathbb{M}}_{\lambda,n}(\bar{\boldsymbol{h}}_n)\Big\}\Big]^T\Big[\boldsymbol{h}_n - \hat{\boldsymbol{h}}_n^{(k)}(\lambda_0)\Big]$$

where $\bar{\boldsymbol{h}}_n = \hat{\boldsymbol{h}}_n^{(k)}(\lambda_0) + \tau\{\boldsymbol{h}_n - \hat{\boldsymbol{h}}_n^{(k)}(\lambda_0)\}$ for some $\tau \in (0,1)$ and $\nabla_{\boldsymbol{h}}\hat{\mathbb{M}}_{\lambda,n}(\boldsymbol{h})$ is the partial derivative of $\hat{\mathbb{M}}_{\lambda,n}(\boldsymbol{h})$ with respect to $\boldsymbol{h}$. Using $\nabla_{\boldsymbol{h}_n}[\hat{\mathbb{M}}_{\lambda_0,n}(\boldsymbol{h}_n) - \hat{\mathbb{M}}_{\lambda,n}(\boldsymbol{h}_n)] = \nabla^2_{\boldsymbol{h}_n,\lambda}\hat{\mathbb{M}}_{\lambda_0,n}(\boldsymbol{h}_n)(\lambda - \lambda_0) + o(|\lambda - \lambda_0|)$ we get

$$\hat{\mathbb{M}}_{\lambda_0,n}\{\boldsymbol{h}_n\} - \hat{\mathbb{M}}_{\lambda_0,n}\{\hat{\boldsymbol{h}}_n^{(k)}(\lambda_0)\} \leq \sup_{\boldsymbol{h}\in N_\delta(\hat{\boldsymbol{h}}_n^{(k)}(\lambda_0))}\Big[\Big\|\nabla^2_{\boldsymbol{h}_n,\lambda}\hat{\mathbb{M}}_{\lambda_0,n}(\boldsymbol{h}) + o(1)\Big\|_2\Big]\Big|\lambda - \lambda_0\Big|\Big\|\boldsymbol{h}_n - \hat{\boldsymbol{h}}_n^{(k)}(\lambda_0)\Big\|_2$$

Moreover assumption $(A3)$ implies that

$$\hat{\mathbb{M}}_{\lambda_0,n}\{\boldsymbol{h}_n\} - \hat{\mathbb{M}}_{\lambda_0,n}\{\hat{\boldsymbol{h}}_n^{(k)}(\lambda_0)\} \geq c\Big\|\boldsymbol{h}_n - \hat{\boldsymbol{h}}_n^{(k)}(\lambda_0)\Big\|_2^2$$

The desired result thus follows from the above two displays with

$$L = \sup_{\boldsymbol{h}\in N_\delta(\hat{\boldsymbol{h}}_n^{(k)}(\lambda_0))} \|\nabla^2_{\boldsymbol{h}_n,\lambda}\hat{\mathbb{M}}_{\lambda_0,n}(\boldsymbol{h}) + o(1)\|_2/c.$$

Proof of Lemma 6

To prove the first statement of lemma 6, note that from equation (4)

$$\delta^\pi_{(1)}(y) = \frac{a_{y-1}/a_y}{w_p^{(1)}(y)}, \text{ for } y \geq 1.$$

Now let $V(y) = a_{y-1}/(a_y[w_p^{(1)}(y)]^2)$. Then,

$$\mathbb{E}_{Y|\theta}\Big[\delta^\pi_{(1)}(Y)\Big]^2 = \mathbb{E}_{Y|\theta}\Big[\frac{a_{Y-1}}{a_Y}V(Y)\Big] = \sum_{y=1}^{\infty}\frac{a_{y-1}}{a_y}V(y)\frac{a_y\theta^y}{g(\theta)} = \theta\sum_{y=0}^{\infty}V(y+1)\frac{a_y\theta^y}{g(\theta)} = \theta\mathbb{E}_{Y|\theta}\Big[V(Y+1)\Big],$$

where $V(y+1) = a_y/(a_{y+1}[w_p^{(1)}(y+1)]^2) = [\delta^\pi_{(1)}(y+1)]^2(a_{y+1}/a_y)$. This proves the first statement of lemma 6.

To prove the second statement, note that from equation (4)

$$\delta^\pi_{(0)}(y) = \frac{a_y/a_{y+1}}{w_p^{(0)}(y)}, \text{ for } y \geq 0.$$

Let $V(y+1) = a_y/[a_{y+1}w_p^{(0)}(y)] = \delta^\pi_{(0)}(y)$. Then,

$$\theta\mathbb{E}_{Y|\theta}\Big[V(Y+1)\Big] = \sum_{y=0}^{\infty}\frac{a_y}{a_{y+1}}V(y+1)\frac{a_{y+1}\theta^{y+1}}{g(\theta)} = \sum_{y=0}^{\infty}\frac{a_{y-1}}{a_y}V(y)\frac{a_y\theta^y}{g(\theta)} = \mathbb{E}_{Y|\theta}\Big[\frac{a_{Y-1}}{a_Y}V(Y)\Big],$$

where $a_{-1} = 0$ and $(a_{y-1}/a_y)V(y) = (a_{y-1}/a_y)\delta^\pi_{(0)}(y-1)$. This proves the second statement of lemma 6.

References

Anna Aizer and Joseph J Doyle Jr. Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges. *The Quarterly Journal of Economics*, 130(2): 759–803, 2015.

Timothy B Armstrong, Michal Kolesár, and Mikkel Plagborg-Møller. Robust empirical bayes confidence intervals. *arXiv preprint arXiv:2004.03448*, 2020.

Roja Bandari, Sitaram Asur, and Bernardo A Huberman. The pulse of news in social media: Forecasting popularity. *ICWSM*, 12:26–33, 2012.

Alessandro Barp, Francois-Xavier Briol, Andrew Duncan, Mark Girolami, and Lester Mackey. Minimum stein discrepancy estimators. In *Advances in Neural Information Processing Systems*, pages 12964–12976, 2019.

J Frédéric Bonnans and Alexander Shapiro. *Perturbation analysis of optimization problems*. Springer Science & Business Media, 2013.

Lawrence D Brown. In-season prediction of batting averages: A field test of empirical bayes and bayes methodologies. *The Annals of Applied Statistics*, pages 113–152, 2008.

Lawrence D Brown and Eitan Greenshtein. Nonparametric empirical bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics*, pages 1685–1704, 2009.

Lawrence D Brown, Eitan Greenshtein, and Ya'acov Ritov. The poisson compound decision problem revisited. *Journal of the American Statistical Association*, 108(502):741–749, 2013.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. JMLR: Workshop and Conference Proceedings, 2016.

M Lawrence Clevenson and James V Zidek. Simultaneous estimation of the means of independent poisson laws. *Journal of the American Statistical Association*, 70(351a): 698–705, 1975.

Anna Piil Damm and Christian Dustmann. Does growing up in a high crime neighborhood affect youth criminal behavior? *American Economic Review*, 104(6):1806–32, 2014.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

Bradley Efron. Two modeling strategies for empirical bayes estimation. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(2):285, 2014.

Bradley Efron. Empirical bayes deconvolution estimates. *Biometrika*, 103(1):1–20, 2016.

Bradley Efron. Bayes, oracle bayes and empirical bayes. *Statistical Science*, 34(2):177–201, 2019.

Dominique Fourdrinier and Christian P Robert. Intrinsic losses for empirical bayes estimation: A note on normal and poisson cases. *Statistics & probability letters*, 23(1):35–44, 1995.

Dominique Fourdrinier, William E Strawderman, and Martin T Wells. *Shrinkage Estimation*. Springer, 2018.

Anqi Fu, Balasubramanian Narasimhan, and Stephen Boyd. Cvxr: An r package for disciplined convex optimization. *arXiv preprint arXiv:1711.07582*, 2017.

Jiaying Gu and Roger Koenker. Empirical bayesball remixed: Empirical bayes methods for longitudinal data. *Journal of Applied Econometrics*, 32(3):575–599, 2017.

Nikolaos Ignatiadis and Stefan Wager. Bias-aware confidence intervals for empirical bayes analysis. *arXiv preprint arXiv:1902.02774*, 2019.

William James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379, 1961.

Wenhua Jiang and Cun-Hui Zhang. General maximum likelihood empirical bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647–1684, 2009.

Jack Kiefer and Jacob Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906, 1956.

Achim Klenke. *Probability Theory: A Comprehensive Course*, pages 331–349. Springer London, 2014.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

Roger Koenker and Jiaying Gu. Rebayes: An r package for empirical bayes mixture methods. *Journal of Statistical Software*, 82(1):1–26, 2017.

Roger Koenker and Ivan Mizera. Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association*, 109 (506):674–685, 2014.

Susan V Koski, David Bowers, and SE Costanza. State and institutional correlates of reported victimization and consensual sexual activity in juvenile correctional facilities. *Child and Adolescent Social Work Journal*, 35(3):243–255, 2018.

Nan Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811, 1978.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances In Neural Information Processing Systems*, pages 2378–2386, 2016.

Qiang Liu, Jason D Lee, and Michael I Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

Nuno Moniz and Luís Torgo. Multi-source social feedback of online news feeds. *arXiv preprint arXiv:1801.07055*, 2018.

Albert Noack. A class of random variables with discrete distributions. *The Annals of Mathematical Statistics*, 21(1):127–132, 1950.

Chris J Oates, Mark Girolami, and Nicolas Chopin. Control functionals for monte carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):695–718, 2017.

David Pollard. A few good inequalities. *available at: http://www.stat.yale.edu/~pollard/Books/Mini/Basic.pdf*, 2015.

Herbert Robbins. An empirical bayes approach to statistics. Technical report, COLUMBIA UNIVERSITY New York City United States, 1956.

Ken-iti Sato and Sato Ken-Iti. *Lévy processes and infinitely divisible distributions*. Cambridge university press, 1999.

Robert J Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

Galit Shmueli, Thomas P Minka, Joseph B Kadane, Sharad Borle, and Peter Boatwright. A useful distribution for fitting discrete data: revival of the conway–maxwell–poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54 (1):127–142, 2005.

US Department of Justice and Federal Bureau of Investigation. Uniform Crime Reporting Program Data: County-Level Detailed Arrest and Offense Data, United States, 2012. *Inter-University Consortium for Political and Social Research Ann Arbor, MI*, 2014. doi: https://doi.org/10.3886/ICPSR35019.v1.

Hal R Varian. A bayesian approach to real estate assessment. *Studies in Bayesian econometric and statistics in Honor of Leonard J. Savage*, pages 195–208, 1975.

Asaf Weinstein, Zhuang Ma, Lawrence D Brown, and Cun-Hui Zhang. Group-linear empirical bayes estimates for a heteroscedastic normal mean. *Journal of the American Statistical Association*, pages 1–13, 2018.

Xianchao Xie, SC Kou, and Lawrence D Brown. Sure estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association*, 107(500):1465–1479, 2012.

Jiasen Yang, Qiang Liu, Vinayak Rao, and Jennifer Neville. Goodness-of-fit testing for discrete distributions via stein discrepancy. In *International Conference on Machine Learning*, pages 5561–5570, 2018.