Information and Inference: A Journal of the IMA (2025) **00**, 1–42
https://doi.org/10.1093/imaiai/iaaf027

Optimal empirical Bayes estimation for the Poisson model via minimum-distance methods

SOHAM JANA*
Department of ACMS, University of Notre Dame, 102G Crowley Hall, Notre Dame, 46556, IN, USA

YURY POLYANSKIY
Department of EECS, MIT, 50 Vassar St, Cambridge, 02139, MA, USA

AND

YIHONG WU
Department of Statistics and Data Science, Yale University, 10 Hillhouse Ave., New Haven, 06511, CT, USA
*Corresponding author: soham.jana@nd.edu

[Received on 08 August 2024; revised on 04 March 2025; accepted on 04 September 2025]

The Robbins estimator is the most iconic and widely used procedure in the empirical Bayes literature for the Poisson model. On one hand, this method has been recently shown to be minimax optimal in terms of the regret (excess risk over the Bayesian oracle that knows the true prior) for various nonparametric classes of priors. On the other hand, it has been long recognized in practice that the Robbins estimator lacks the desired smoothness and monotonicity of Bayes estimators and can be easily derailed by those data points that were rarely observed before. Based on the minimum-distance distance method, we propose a suite of empirical Bayes estimators, including the classical nonparametric maximum likelihood, that outperform the Robbins method in a variety of synthetic and real data sets and retain its optimality in terms of minimax regret.

Keywords: Mixture modeling; Robbins method; Poisson mean estimation; Nonparametric estimation; NPMLE.

1. Introduction

Consider the Poisson mean estimation problem. Given observations $Y^n \triangleq (Y_1, \ldots, Y_n)$, independently distributed according to the Poisson distribution with mean parameters $\theta^n \triangleq (\theta_1, \ldots, \theta_n)$, the goal is to estimate the parameter vector under the squared error loss. It is well known in the literature that the minimax estimator need not be the best choice in practice, unless the observations are known to be generated according to the least favorable prior distribution on the parameter space. A class of shrinkage-type alternative estimators was proposed in the seminal paper of [59, 60], namely the empirical Bayes (EB) methodology. In the regular Bayes setup, which also produces estimators with shrinkage properties, one assumes that the parameter values are independently distributed according to a prior distribution G. Then the best estimator under the squared error loss (i.e., the Bayes estimator) of θ_j is given by the posterior mean $\widehat{\theta}_G(Y_j) = \mathbb{E}_G[\theta_j|Y_j]$. The EB theory proposes to bypass the assumed knowledge about G, which might be unavailable in practice, by approximating the G dependent expressions using the observations. For example, in the Poisson

© The Author(s) 2025. Published by Oxford University Press on behalf of the Institute of Mathematics and its Applications. All rights reserved.

model, given a prior distribution G on θ, the posterior mean is of the form

$$\widehat{\theta}_G(y) = \mathbb{E}_{Y\sim \mathrm{Poi}(\theta),\theta\sim G}[\theta|Y=y] = \frac{\int \theta e^{-\theta}\frac{\theta^y}{y!}G(d\theta)}{\int e^{-\theta}\frac{\theta^y}{y!}G(d\theta)} = (y+1)\frac{f_G(y+1)}{f_G(y)}. \tag{1.1}$$

Here $\mathrm{Poi}(\theta)$ denotes the Poisson distribution with mean θ and marginal density of Y_j is given by

$$f_G(y) = \int f_\theta(y)G(d\theta), \quad f_\theta(y) = e^{-\theta}\frac{\theta^y}{y!}, y \in \mathbb{Z}_+ \triangleq \{0,1,\ldots\} \tag{1.2}$$

Then, in the EB methodology, one can approximate either G or f_G from the data and plug it into the above formula. The significant achievement of the EB theory is that when the number of independent observations is large, it is possible to "borrow strength" from these independent (and seemingly unrelated) observations to achieve the asymptotically optimal Bayes risk per coordinate. Since its conception, the theory and methodology of EB has been well developed and widely applied in large-scale data analysis in practice, cf. e.g. [7, 21, 53, 63]. We refer the reader to the surveys and monographs on the theory and practice of EB [11, 18, 19, 46, 52, 69].

In particular, to motivate the use of EB methodology in the Poisson settings, we present a real data example where we produce three EB estimators that beat the minimax optimal estimator. We analyze the data on the total number of goals scored in the National Hockey League for the seasons 2017-18 and 2018-19 (the data is available at https://www.hockey-reference.com/). We consider the statistics of $n = 745$ players for whom the data were collected for both seasons. Let Y_i be the total number of goals scored by the i^{th} player in the season 2017-18. We model Y_i as independently distributed $\mathrm{Poi}(\theta_i)$ random variables, where θ_i's are independently distributed according to some prior G on $\mathbb{R}_+$. Based on the observations, we intend to predict the goals scored by each player in the 2018-19 season. Let us explain how *estimation* of θ's can be used to make *future predictions*. If we assume that player i scores $Z_i \sim \mathrm{Poi}(\theta_i)$ goals in the future year, then predictor $\widehat{Z}_i$ that minimizes mean-square error (MSE) is clearly $\mathbb{E}[Z_i|Y_i] = \mathbb{E}[\theta_i|Y_i] = \widehat{\theta}_G(Y_i)$. Thus, since the EB estimator attempts to approximate $\widehat{\theta}_G$ it can also be used as an estimator for Z_i. Note that if the prediction metric is mean *absolute* error (MAE), then the optimal predictor would be a posterior median (under G). We do not discuss in this paper EB methods for estimating posterior median and simply reuse the estimator of θ_i for MAE as well. As the number of goals Y_i-s in the data are all below 50, for the sake of computation, we can assume that the parameters θ_i-s are supported on $[0,50]$. The minimax estimator in the above Poisson settings with a squared error loss is given by the posterior mean for the least favorable prior (see Appendix A for proof). However, even though the minimax estimator is designed to perform optimally for the worst-case scenario, its average performance, particularly with real data, can be overly conservative. Hence, the main advantage of EB is that, in instances far from being the least favorable, one can typically outperform the minimax estimator by being Bayesian with a prior learned from data. Therefore, we will present a comparative study of our EB methods against the minimax estimator and show that in most of our numerical examples, including the hockey-data experiments, there are significant gains in using the EB methodology. To find the least favorable prior, we solve

$$\underset{G\in\mathscr{P}([0,h])}{\mathrm{argmax}}\ \mathbb{E}[|\theta - \mathbb{E}[\theta|Y]|^2] = \underset{G\in\mathscr{P}([0,h])}{\mathrm{argmax}} \left[\mathbb{E}_G\left[\theta^2\right] - \sum_{y=0}^{\infty}(y+1)^2\frac{(f_G(y+1))^2}{f_G(y)}\right] \tag{1.3}$$

TABLE 1 *Minimax vs EB estimators*

Methods	Robbins	Minimax	Naive	minimum-H^2	NPMLE	minimum-χ^2
RMSE	15.59	8.62	6.19	6.02	6.04	6.05
MAD	6.64	7.54	4.35	4.37	4.38	4.39

where the maximization is over $G \in \mathscr{P}([0,h])$, all priors supported on $[0,h]$. For computation, we pick $h = 50$, divide parameter space $[0,50]$ into a grid of 1000 equidistant points, and then optimize the prior using a gradient ascent algorithm. See Fig. 1 for the plot of the least favorable prior.

To emulate the Bayesian oracle, we consider EB estimators based on three methods of estimating the prior G from the data:

- Nonparametric maximum likelihood estimator (NPMLE) [32]
- Minimum squared Hellinger (H^2) distance estimator
- Minimum χ^2-distance estimator.

These methods are detailed later in Section 2. We compare their performances with the classical Robbins estimators [59, 60] (presented in (1.4)), the Minimax estimator, and the Naive estimator that directly uses the goals from season 2017-2018 to predict the goals in season 2018-2019 for the same player. The root mean squared error (RMSE) and mean absolute deviation error (MAD) for predicting the hockey goals are presented in Table 1.[1] Notably, all EB estimators based on G-estimation perform better than the minimax estimator and the Robbins estimator in both error metrics. The above three EB methods based on G-estimation also outperform the Naive estimator in terms of RMSE. The performances for the Naive method and our EB methods are similar in the MAD metric. However, note that the construction of the above EB estimators is aimed at estimating the minimum mean squared error estimator. It may be possible to construct EB estimators that mimic the Bayes estimator under absolute error loss, which could provide improved performance guarantees in the MAD metric compared to the Naive method. This is left for future directions.

Getting back to the literature review, there are two main avenues to solving the EB problem:

- f-modeling: Construct an approximate Bayes estimator by approximating the marginal density. For example, the Robbins estimator [60] is a plug-in estimate of (1.1) replacing the true f_G with the empirical distribution, leading to

$$\widehat{\theta}_j = \widehat{\theta}_{\text{Robbins}}(Y_j|Y_1,\dots,Y_n) \triangleq (Y_j+1)\frac{N(Y_j+1)}{N(Y_j)}, \quad N(y) \triangleq |\{i \in [n] : Y_i = y\}|. \tag{1.4}$$

- g-modeling: We first obtain an estimate $\widehat{G}$ of the prior G from Y^n and then apply the corresponding Bayes estimator formula $\widehat{\theta}_{\widehat{G}}(Y_j)$. Examples of $\widehat{G}$ include the celebrated NPMLE method mentioned above

$$\widehat{G} = \operatorname*{argmax}_G \frac{1}{n}\sum_{i=1}^{n} \log f_G(Y_i) \tag{1.5}$$

[1] Given data points $Y_1,\dots,Y_n$ and their predictions $\widehat{Y}_1,\dots,\widehat{Y}_n$ the RMSE is defined as $\sqrt{\frac{1}{n}\sum_{i=1}^n (Y_i-\widehat{Y}_i)^2}$ and the MAD is defined as $\frac{1}{n}\sum_{i=1}^n |Y_i-\widehat{Y}_i|$.

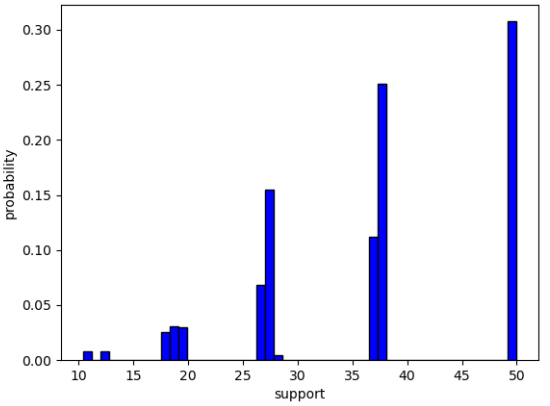

FIG. 1. Least favorable prior on [0,50]

where the maximization is over all priors on $\mathbb{R}_+$ (unconstrained NPMLE). When additional information about the prior is available (e.g., compactly supported), it is convenient to incorporate these constraints into the above optimization, leading to a constrained NPMLE.

In a nutshell, both f-modeling and g-modeling rely on an estimate of the population density f_G; the difference is that the former applies improper density estimate such as the empirical distribution or kernel density estimate (see, e.g., [8, 41, 70] for Gaussian models), while the latter applies *proper* density estimate of the form $f_{\widehat{G}}$.

In recent years, there have been significant advances in the theoretical analysis of f-modeling EB estimators for the Poisson model, specifically, the Robbins method. For compactly supported priors, [9] showed that with Poisson sampling (replacing the sample size n by $\mathrm{Poi}(n)$), the Robbins estimator achieves a $O\left(\frac{(\log n)^2}{n(\log\log n)^2}\right)$ regret for estimating each θ_i. Later [57] showed that the same bound holds with fixed sample size n and established the optimality of the Robbins estimator by proving a matching minimax lower bound. For the class of subexponential priors, for estimating each θ_i, the Robbins estimator also achieves optimal minimax regret $\Theta\left(\frac{(\log n)^3}{n}\right)$.

On the other hand, despite its simplicity and optimality, it has long been recognized that the Robbins method often produces unstable estimates in practice. This occurs particularly for that y which appears a few times or none whatsoever, so that $N(y)$ is small or zero. Thus, unless $N(y+1)$ is also small, the formula (1.4) produces exceptionally large value of $\widehat{\theta}_{\mathrm{Robbins}}(y)$. In addition, if $N(y+1)=0$ (e.g., when $y \geq \max\{Y_1,\ldots,Y_n\}$), we have $\widehat{\theta}_{\mathrm{Robbins}}(y)=0$ irrespective of any existing information about y, which is at odds with the fact that the Bayes estimator $\widehat{\theta}_G(y)$ is always monotonically increasing in y for any G [26]. These issues of the Robbins estimator have been well-documented and discussed in the literature; see, for example, [48, Section 1] and [46, Section 1.9] for a finite-sample study and [20, Section 6.1] for the destabilized behavior of Robbins estimator in practice (e.g., in analyzing insurance claims data). To alleviate the shortcomings of the Robbins estimator, a number of modifications have been proposed [9, 48] that enforce smoothness or monotonicity; nevertheless, it is unclear if they still retain the regret optimality of the Robbins

method. This raises the question of whether it is possible to construct a well-behaved EB estimator that is provably optimal in terms of regret.

In this paper, we answer this question in the positive. This is accomplished by a class of g-modeling EB estimators, which are free from the unstable behavior of the Robbins estimator, thanks to their Bayesian form, which guarantees monotonicity among many other desirable properties. The prior is learned using the *minimum-distance* method, including the NPMLE (1.5) as a special case. Introduced in the pioneering works [64, 65, 66], the minimum-distance method aims to find the best fit *in class* to the data with respect to a given distance. As such, it is well-suited for estimating the prior, and the obtained density estimate is *proper* and of the desired mixture type.

As a concrete example, we consider a simple uniform prior and compare the numerical performance of Robbins and three prototypical examples of minimum-distance estimators of G, with respect to the Kullback-Leibler (KL) divergence (i.e., the NPMLE), the Hellinger distance, and the χ^2-divergence, respectively (see Section 2.1 for the formal definitions). As evident in Fig. 2, the minimum-distance EB estimators provide a much more consistent approximation of the Bayes estimator compared to the Robbins estimator and the minimax estimator, which can be calculated using the least favorable prior obtained via (1.3) with $h = 3$. This advantage is even more pronounced for unbounded priors (cf. Fig. 6 in Section 5.4); see also Fig. 3 for a real-world example where EB methodology is applied to a prediction task with sports data. Notably, in multidimensional settings, such minimum distance based EB methodologies are difficult to implement in practice as they are computationally expensive even in fixed dimensions. However, we propose that the unidimensional EB methodology can also be employed to provide improved analyses in multidimensional setups. To demonstrate the above, we considered a regression problem based on simulations. We demonstrate that the performance of an ordinary least squares (OLS) method can be significantly improved by pre-processing the individual covariate columns using minimum-distance EB filters before supplying them to the algorithm.

The superior performance of minimum-distance EB estimators in practice is also justified by theory. In addition to characterizing their structural properties (existence, uniqueness, discreteness) in the Poisson model, we show that, under appropriate conditions on the distance functional, their regret is minimax optimal for both compactly supported and subexponential priors. This is accomplished by first proving the optimality of minimum-distance estimate for density estimation in Hellinger distance, then establishing a generic regret upper bound for g-modeling EB estimators in terms of the Hellinger error of the corresponding density estimates. We also extend the theoretical analyses to a multidimensional Poisson models.

1.1. *Related works*

Searching for a stable and smooth alternative to the classical Robbins method for the Poisson EB problem has a long history. [47] was one of the proponents of using g-modeling estimators to resolve this problem. The author considered modeling the prior using the Gamma distribution and estimated the scale and shape parameters using a χ^2-distance minimization. This is a parametric approach as opposed to the nonparametric approach in this paper. Based on the monotonicity of the Bayes estimator, [49] used non-decreasing polynomials to approximate the Bayes oracle (recently, similar isotonic regression based EB approaches have also been used to estimate the Bayes oracle in other models, e.g., see [1] for an example in the Gaussian mean estimation problem). Nonetheless, it is unclear whether these regression-based approaches for estimating the Bayes oracle can be used directly to draw any conclusions about estimating the marginal data distribution, which is also one of the primary focuses of our work. [40] proposed an iterative method of estimating the prior

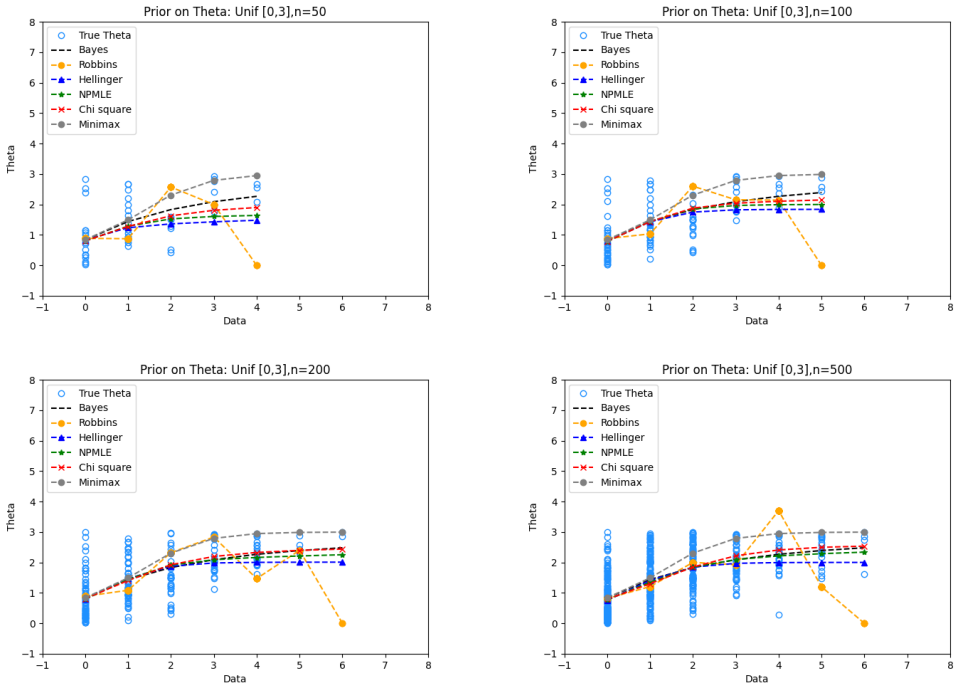

FIG. 2. Comparison of Robbins estimator with different minimum-distance EB estimators. Here the latent $\theta_i \overset{\text{i.i.d.}}{\sim} \text{Uniform}[0,3]$ and the observation $Y_i \overset{\text{ind}}{\sim} \text{Poi}(\theta_i)$, for $i = 1,\dots,n$. We plot $\widehat{\theta}(Y_i)$ against Y_i for various EB estimators $\widehat{\theta}$. For reference, we also plot the true value θ_i and the Bayes estimator $\widehat{\theta}_G(Y_i)$. The sample sizes are $n = 50, 100, 200, 500$.

by first using the empirical distribution of the training sample Y^n and then using corresponding posterior means of the θ_i's to denoise. In a similar vein, [3] assumed the existence of a density of the prior distribution and used the kernel method to approximate the prior. For a detailed exposition of other smooth EB methods, see [46]. However, none of these methods has theoretical guarantees in terms of the regret for the nonparametric class of priors considered in the present paper.

Applying NPMLE to estimate the mixture distribution has been well-studied in the literature. [32] was one of the preliminary papers to prove the consistency of the NPMLE, which was subsequently extended in [24, 29, 38, 54]; for a more recent discussion, see [13]. In the present paper, we focus on the Poisson mixture model and sharpen these results by obtaining the optimal rate of convergence for the NPMLE. In addition to the aforementioned statistical results, structural understanding of the NPMLE (existence, uniqueness, and discreteness) has been obtained in [29, 42, 43, 45, 61] for the general univariate exponential family. We extend these structural results to a class of minimum-distance estimators for Poisson mixture models following [61]. Finally, we mention the recent work [50], which explored the application of NPMLE in a related scenario of heterogeneous Poisson mixtures.

Initial work on applying NPMLE for EB estimation was carried out in [37] for the Binomial and the normal location models, and the analysis is primarily numerical. For theoretical results, [23, 70] analyzed the Hellinger risk of NPMLE-based mixture density estimates, which forms the basis of

the analysis of NPMLE for EB estimation in [30]. The resulting regret bounds, though state-of-the-art, still differ from the minimax lower bounds in [57] by logarithmic factors for both the classes of compactly supported and subgaussian priors. This is because (a) the density estimation analysis in [70] is potentially suboptimal compared to the lower bounds in [33]; (b) the Fourier-analytic reduction from the Hellinger distance for mixture density to regret in [30] is loose. In comparison, density estimation and regret bounds are optimal in this paper with exact logarithmic factors. This can be attributed to the discrete nature of the Poisson model, which allows a simple truncation-based analysis to suffice for light-tailed priors. These sharp results are also generalized from the NPMLE-based EB estimator to the minimum-distance estimators.

The minimum distance based analysis for the Poisson model we consider here can be easily extended to other members of the exponential family of distribution. In a recent follow-up work [28], it was shown that the Bayes estimator for various distributions, including the Geometric distribution, Negative Binomial distribution, and Exponential distribution, can be represented similarly in terms of simple functions of the prior distribution. Our minimum distance methodology can then be extended to estimate the prior distribution, and similar regret analysis can be generalized to these discrete models.

1.2. *Organization*

The rest of the paper is organized as follows. In Section 2, we introduce the class of minimum distance estimators and identify conditions on the distance function that guarantee the minimizer's existence and uniqueness. The theoretical guarantees in density estimation and regret are presented in Theorem 2 and Theorem 3 therein. The proof sketches of these results are presented in Section 3 and Section 4, respectively. In Section 5, we present an algorithm for computing minimum-distance estimators in the one-dimensional setting. We also study their numerical performance in empirical Bayes estimation with simulated and real datasets. In Section 6, we mention our theoretical results in a multi-dimensional setting. For all the other related details of proofs, see the Appendix.

1.3. *Notations*

Denote by $\mathbb{Z}_+$ (resp. $\mathbb{R}_+$) the set of non-negative integers (resp. real numbers). For a Borel measurable subset $\Theta \subset \mathbb{R}$, let $\mathscr{P}(\Theta)$ be the collection of all probability measures on Θ. For any $\theta \in \mathbb{R}_+$ let δ_θ denote the Dirac measure at θ. Denote by $\mathsf{SubE}(s)$ the set of all s-subexponential distributions on $\mathbb{R}_+$: $\mathsf{SubE}(s) = \left\{G : G([t,\infty)]) \le 2e^{-t/s}, \forall t > 0\right\}$. Let $Y_i \sim \mathrm{Poi}(\theta_i)$ for $i = 1,\dots,n$ and $Y \sim \mathrm{Poi}(\theta)$, with $\theta_1,\dots,\theta_n,\theta \overset{\text{i.i.d.}}{\sim} G$. This also implies $Y_1,\dots,Y_n,Y \overset{\text{i.i.d.}}{\sim} f_G$ where f_G is the mixture distribution defined in (1.2). Let $\mathbb{E}_G$ and $\mathbb{P}_G$ denote the expectation and probability where the true mixing distribution is G. Define $Y_{\max} = \max_{i\in[n]} Y_i, Y_{\min} = \min_{i\in[n]} Y_i$.

2. Problem formulation and results

2.1. *Minimum-distance estimators*

Denote by $\mathscr{P}(\mathbb{Z}_+)$ the collection of probability distributions (pmfs) on $\mathbb{Z}_+$. We call $\mathsf{dist} : \mathscr{P}(\mathbb{Z}_+) \times \mathscr{P}(\mathbb{Z}_+) \to \mathbb{R}_+$ a *generalized distance* if $\mathsf{dist}(p\|q) \ge 0$ for any $p,q \in \mathscr{P}(\mathbb{Z}_+)$, with equality if and only if $p = q$. Note that any metric or f-divergence [14] qualifies as a generalized distance.

The minimum-distance[2] methodology aims to find the closest fit in the model class to the data. While it is widely used and well-studied in parametric models [4, 5, 6, 51, 55], it is also useful in nonparametric settings such as mixture models. Denote by

$$p_n^{\mathsf{emp}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i} \tag{2.1}$$

the empirical distribution of the sample $Y_1, \dots, Y_n$. The minimum-distance estimator for the mixing distribution with respect to dist, over some target class of distributions $\mathscr{G}$, is

$$\widehat{G} \in \operatorname*{argmin}_{Q \in \mathscr{G}} \mathsf{dist}(p_n^{\mathsf{emp}} \| f_Q). \tag{2.2}$$

Note that in our analysis we also allow $\mathscr{G} = \mathscr{P}(\mathbb{R}_+)$, the set of all probability distributions on the nonnegative real line. In such cases, for ease of notations, we will replace $\operatorname{argmin}_{Q\in\mathscr{P}(\mathbb{R}_+)} \mathsf{dist}(p_n^{\mathsf{emp}} \| f_Q)$ with $\operatorname{argmin}_Q \mathsf{dist}(p_n^{\mathsf{emp}} \| f_Q)$. Primary examples of minimum-distance estimators considered in this paper include the following

- Maximum likelihood: $\mathsf{dist}(p\|q) = \mathrm{KL}(p\|q) \triangleq \sum_{y\ge 0} p(y) \log \frac{p(y)}{q(y)}$ is the KL divergence. In this case, one can verify that the minimum-KL estimator coincides with the NPMLE (1.5).
- Minimum-Hellinger estimator: $\mathsf{dist}(p\|q) = H^2(p,q) \triangleq \sum_{y\ge 0} \left(\sqrt{p(y)} - \sqrt{q(y)}\right)^2$ is the squared Hellinger distance.
- Minimum-χ^2 estimator: $\mathsf{dist}(p\|q) = \chi^2(p\|q) \triangleq \sum_{y\ge 0} \frac{(p(y)-q(y))^2}{q(y)}$ is the χ^2-divergence.

Note that there are other minimum-distance estimators previously studied for the Gaussian mixture model, e.g., those respect to L_p-distance of the CDFs, aiming at estimating the mixing distribution [12, 15, 17, 25]. These are outside the scope of the theory developed here.

In general, the solution to (2.2) need not be unique; nevertheless, for the Poisson mixture model, the uniqueness is guaranteed provided that the generalized distance dist admits the following decomposition:

Assumption 1 *There exist maps $t : \mathscr{P}(\mathbb{Z}_+) \to \mathbb{R}$ and $\ell : \mathbb{R}^2 \to \mathbb{R}$ such that for any two distributions $q_1, q_2 \in \mathscr{P}(\mathbb{Z}_+)$*

$$\mathsf{dist}\,(q_1\|q_2) = t(q_1) + \sum_{y\ge 0} \ell(q_1(y), q_2(y)),$$

where $b \mapsto \ell(a,b)$ is strictly decreasing and strictly convex for $a > 0$ and $\ell(0,b) = 0$ for $b \ge 0$.

The following theorem guarantees the existence, uniqueness, and discreteness of both unconstrained and support-constrained minimum-distance estimators. For the special case of unconstrained NPMLE this result was previously shown by [61] and later extended to all one-dimensional exponential family [45].

[2] We adopt this conventional terminology even when dist need not be a distance.

Theorem 1 *Let* dist *satisfy Assumption 1. Let p be a probability distribution on* $\mathbb{Z}_+$ *with support size m. Then for any* $h > 0$*, the constrained solution* $\operatorname{argmin}_{Q\in\mathscr{P}([0,h])} \mathsf{dist}(p\|f_Q)$ *exist uniquely and is a discrete distribution with support size at most m. Furthermore, the same conclusion also applies to the unconstrained solution* $\operatorname{argmin}_{Q\in\mathscr{P}(\mathbb{R}_+)} \mathsf{dist}(p\|f_Q)$*, which in addition is supported on* $[\min_{i=1,\dots,m} y_i, \max_{i=1,\dots,m} y_i]$*, where* $\{y_1,\dots,y_m\}$ *is the support of p.*

To analyze the statistical performance of minimum-distance estimators, we impose the following regulatory condition on the generalized distance dist:

Assumption 2 *There exist absolute constants* $c_1, c_2 > 0$ *such that for pmfs* q_1, q_2 *on* $\mathbb{Z}_+$

$$c_1 H^2(q_1,q_2) \le \mathsf{dist}(q_1\|q_2) \le c_2\chi^2(q_1\|q_2) \tag{2.3}$$

Major examples of generalized distance satisfying Assumptions 1 and 2 include the KL divergence, squared Hellinger distance, and χ^2-divergence. This follows from noting that $2H^2 \le \mathrm{KL} \le \chi^2$ and each of them satisfies the decomposition Assumption 1: for squared Hellinger $t \equiv 2, \ell(a,b) = -2\sqrt{ab}$, for KL divergence $t \equiv 0, \ell(a,b) = a\log\frac{a}{b}$, for χ^2-divergence $t \equiv -1, \ell(a,b) = \frac{a^2}{b}$. On the other hand, total variation (TV) satisfies neither Assumption 1 nor 2 so the theory in the present paper does not apply to the minimum-TV estimator.

Remark 1 *Before proceeding further, note the following argument in the context of Theorem 1 (which is a deterministic result) to exclude* $G = \delta_0$ *as a possible choice in the analysis, and the situation where all* y_i*-s are zero. The remark applies for the rest of the paper to exclude* $G = \delta_0$ *as a possible choice in the analysis. Whenever we need to divide with* $f_G(y)$*, the choice* $G = \delta_0$ *will lead to technical difficulties as* f_G *will be degenerate at zero as well. Such division occurs, for example, in the proof of Theorem 2, where we use a bound based on the* χ^2*-divergence between the empirical distribution and* f_G *to control the behavior of the general minimum distance estimator. Consider the following cases:*

- $y_1 = \ldots = y_n = 0$*. Then clearly* $\widehat{G} = \delta_0$ *is the unique NPMLE solution.*
- $y_i > 0$ *for some* $i \in [n]$*. Then clearly* δ_0 *is not the data generating distribution. As a result, we may assume* f_G *to be fully supported on* $\mathbb{Z}_+$ *for priors G in the analysis.*

2.2. *Main results*

In this section we state the statistical guarantee for the minimum-distance estimator $\widehat{G}$ defined in the previous section. Our main results are two-fold (both minimax optimal):

1. Density estimation, in terms of the Hellinger distance $f_{\widehat{G}}$ and the true mixture f_G;
2. Empirical Bayes, in terms of the regret of the Bayes estimator with the learned prior $\widehat{G}$.

As mentioned in Section 1, the regret analysis in fact relies on bounding the density estimation error. We start with the result for density estimation. Recall from Section 1.3 $\mathscr{P}([0,h])$ and $\mathsf{SubE}(s)$ denote the class of compactly supported and subexponential priors respectively.

Theorem 2 (Density estimation) *Let* dist *satisfy Assumption 1 and Assumption 2. Suppose that* $\widehat{G}$ *is the unconstrained minimum distance estimator*

$$\widehat{G} = \operatorname*{argmin}_{Q \in \mathscr{P}(\mathbb{R}^+)} \mathsf{dist}(p_n^{\mathsf{emp}} \| f_Q). \tag{2.4}$$

Then there exist constants c_1, c_2 *such that the following holds.*

(a) $\sup_{G \in \mathscr{P}[0,h]} \mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\right] \le \frac{c_1}{n} \cdot \min\left\{\max\{1,h\}\frac{\log n}{\log\log n}, h + \log n\right\}$ *for any* $n \ge 3$.

(b) $\sup_{G \in \mathsf{SubE}(s)} \mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\right] \le \frac{c_2 \max\{1,s\}}{n} \log n$ *for any* $n \ge 2$.

Remark 2 *(i) If the prior G is allowed to be any distribution on* $\mathbb{R}^+$*, then neither density estimation nor empirical Bayes estimation is possible. This fact is well-known for the Gaussian mixture model [62, 70]. To see this, fix a quadratically spaced grid* $\{\theta_1, \ldots, \theta_n\}$ *where* $\theta_i = i^2 \cdot (\log n)^{10}$*. Consider a prior* $G = \frac{1}{n}\sum_{i=1}^n$ *Uniform*$(\theta_i, \theta_i + \alpha_i)$*, where* α_i*'s are iid drawn from Uniform*$(0,1)$*. In other words, the prior G is a uniform mixture over n clusters each of which has an* $O(1)$ *spread. The quadratic grid is chosen so that the spacing* $\theta_{i+1} - \theta_i$ *far exceeds the standard deviation* $\sqrt{\theta_i}$ *so that with high probability we know which cluster each data point* Y_i *is drawn from. However, there is no enough information to estimate the parameters* α_i*'s consistently because on average we only observe one sample for each cluster. As such, in order to obtain uniform error bound as in Theorem 2 (and later in Theorem 3 for regret), it is necessary to restrict the priors to a subclass.*

(ii) It has been shown recently in [57, Theorem 21] that for any constant h, s*, the minimax squared Hellinger density estimation errors are at least* $\Omega\left(\frac{\log n}{n \log\log n}\right)$ *and* $\Omega\left(\frac{\log n}{n}\right)$ *for priors in the class* $\mathscr{P}([0,h])$ *and* $\mathsf{SubE}(s)$*, respectively. This establishes the minimax optimality of our minimum-distance density estimates.*

(iii) In the shape-constrained density estimation literature, e.g., [36], there has also been some interest in replacing the maximum likelihood fitting criteria with other divergences. In this context, the motivation was the desire to impose weaker concavity constraints than log-concavity while still preserving the underlying convexity of the variational formulation of the problem. Thus, in that setting, divergences were dictated by the form of the concavity constraints. Such convexity assumption is often a necessary criterion to guarantee the validity of divergences; for example, in the case of general Bregman divergences [27, 58], the functional density power divergences [58], and the convexity of the loss function might result in practical benefits, such as the efficiency of minimum divergence estimators [44, Section 2],[2]. In contrast, our work imposes such convexity constraints on the divergences mainly to guarantee the uniqueness and existence of the estimators. Our theoretical analysis does not require convexity constraints as we primarily aim to establish the finite sample error rate of our minimum distance estimators. Using different convexity structures may result in different multiplicative constants in the minimax error rates, which is beyond the scope of our current work. A detailed study of the convexity structures might help to differentiate between the performances of the minimum-distance estimators, and we leave it for future directions.

(iv) Our current theoretical results become vacuous when h approaches a similar order of magnitude to n. However, this is mostly due to technicalities in the proof of our upper bound where we aimed to perform a uniform analysis for all priors over $[0,h]$ *for a constant h. We can revise our analysis to achieve consistency guarantees of the estimators for larger values of h, however*

the general treatment is beyond the scope of the current work. As an example, we detail below the performance guarantee of the unconstrained minimum distance estimator $\widehat{G}$ (the estimator $\widehat{G}$ does not know G or h) when the data-generating prior G is degenerate at some h, as suggested by one of the reviewers. Our analysis includes the the case when h can be larger than n, but significantly smaller than $n^2/\log n$. Note that the degenerate prior corresponds to the data distribution $Y_1 \ldots, Y_n \overset{iid}{\sim} \mathrm{Poi}(h)$. We will show that the minimum distance-based estimator $f_{\widehat{G}}$ of f_G should consistently estimate $\mathrm{Poi}(h)$. Revisiting the proof of Theorem 2 and using that the unconstrained minimum distance estimator $\widehat{G}$ is supported on $[Y_{\min}, Y_{\max}]$, we get

$$\begin{aligned} H^2(f_G, f_{\widehat{G}}) &\le 2\left(H^2(p_n^{\mathsf{emp}}, f_{\widehat{G}}) + H^2(p_n^{\mathsf{emp}}, f_G)\right) \\ &\le \frac{2}{c_1}\left(\mathsf{dist}(p_n^{\mathsf{emp}} \| f_{\widehat{G}}) + \mathsf{dist}(p_n^{\mathsf{emp}} \| f_G)\right) \le \frac{4}{c_1}\mathsf{dist}(p_n^{\mathsf{emp}} \| f_G) \le \frac{4c_2}{c_1}\chi^2(p_n^{\mathsf{emp}} \| f_G), \end{aligned}$$

and hence, for K_1, K_2 to be chosen later, we have

$$\begin{aligned} \mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\right] &= \mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\mathbf{1}_{\{Y_{\min} < K_1 \text{ or } Y_{\max} > K_2\}}\right] + \mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\mathbf{1}_{\{K_1 < Y_{\min} \le Y_{\max} < K_2\}}\right] \\ &\le 4\mathbb{P}\left[Y_{\min} < K_1 \text{ or } Y_{\max} > K_2\right] + \frac{4c_2}{c_1}\mathbb{E}\left[\chi^2(p_n^{\mathsf{emp}} \| f_G)\mathbf{1}_{\{K_1 < Y_{\min} \le Y_{\max} < K_2\}}\right] \\ &\le 4n\mathbb{P}\left[Y < K_1\right] + 4n\mathbb{P}\left[Y > K_2\right] + \frac{4c_2}{c_1}\mathbb{E}\left[\chi^2(p_n^{\mathsf{emp}} \| f_G)\mathbf{1}_{\{K_1 < Y_{\min} \le Y_{\max} < K_2\}}\right] \end{aligned} \tag{2.5}$$

for a random variable $Y \sim \mathrm{Poi}(h)$, where the last inequality used union bounds. We can bound the rightmost term in the above expression as

$$\begin{aligned} &\mathbb{E}\left[\chi^2(p_n^{\mathsf{emp}} \| f_G)\mathbf{1}_{\{K_1 < Y_{\min} \le Y_{\max} < K_2\}}\right] \\ &= \sum_{y \in [K_1, K_2]} \frac{\mathbb{E}\left[(p_n^{\mathsf{emp}}(y) - f_G(y))^2\right]}{f_G(y)} + \sum_{y < K_1 \text{ or } y > K_2} f_G(y)\mathbb{P}[K_1 < Y_{\min} \le Y_{\max} < K_2] \\ &= \sum_{y \in [K_1, K_2]} \frac{1 - f_G(y)}{n} + \mathbb{P}[K_1 < Y_{\min} \le Y_{\max} < K_2] \cdot \mathbb{P}_{Y \sim \mathrm{Poi}(h)}\left[Y < K_1 \text{ or } Y > K_2\right] \\ &\le \frac{K_2 - K_1}{n} + \mathbb{P}_{Y \sim \mathrm{Poi}(h)}\left[Y < K_1\right] + \mathbb{P}_{Y \sim \mathrm{Poi}(h)}\left[Y > K_2\right]. \end{aligned} \tag{2.6}$$

In view of (2.5) the last display implies

$$\mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\right] \le \frac{4c_2(K_2 - K_1)}{c_1 n} + (4n + \frac{4c_2}{c_1})(\mathbb{P}\left[Y < K_1\right] + \mathbb{P}\left[Y > K_2\right]) \tag{2.7}$$

We choose

$$K_1 = \max\{0, h - 2\sqrt{h\log n}\}, \quad K_2 = h + 3\sqrt{h\log n}.$$

Then we can use Lemma 10 to bound $\mathbb{P}\left[Y < K_1\right], \mathbb{P}[Y > K_2]$ as long as $h \ge c\log n$ for a large enough c. The application gives us $\mathbb{P}\left[Y < K_1\right], \mathbb{P}[Y > K_2] \le \frac{2}{n^2}$. Plugging the above choice in

(2.7), *we get for a constant* $c > 0$

$$\mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\right] \leq \frac{c\sqrt{h \log n}}{n}.$$

This implies that the minimum distance estimator is consistent given h *is significantly smaller than* $n^2/\log n$.

Next, we turn to the problem of estimating $\theta_1, \ldots, \theta_n$ from $Y_1, \ldots, Y_n$, under the squared error loss, using the empirical Bayes methodology. In this work, we study the estimation guarantees of the g-modeling type estimators. Notably, to produce an estimator $\widehat{\theta}_j$ of θ_j, we use the observations $Y^{-j} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_n)$ to approximate G and then plug it in the formula of the Bayes estimator $\widehat{\theta}_{\widehat{G}}(Y_j)$ in (1.1). Given any class of distributions $\mathscr{G}$ and any distribution estimator strategy characterized by $\widehat{G}$, define the total regret as its worst-case excess risk over the Bayes error:

$$\mathsf{TotRegret}_n(\widehat{G}; \mathscr{G}) \triangleq \sup_{G \in \mathscr{G}} \left\{ \mathbb{E}_G \left[\|\widehat{\theta}^n(Y^n) - \theta^n\|^2 - n \cdot \mathrm{mmse}(G) \right] \right\}, \tag{2.8}$$

$$\widehat{\theta}_j = \widehat{\theta}_{\widehat{G}(Y^{-j})}(Y_j), \quad j = 1, \ldots, n. \tag{2.9}$$

where $\mathrm{mmse}(G)$ denotes the minimum mean squared error of estimating $\theta \sim G$ based on a single observation $Y \sim f_\theta$, i.e., the Bayes risk

$$\mathrm{mmse}(G) \triangleq \inf_{\widehat{\theta}} \mathbb{E}_G \left[\left(\widehat{\theta}(Y) - \theta \right)^2 \right] = \mathbb{E}_G \left[(\widehat{\theta}_G(Y) - \theta)^2 \right]. \tag{2.10}$$

In addition, one can define the problem of quantifying the individual regret for the estimator $\widehat{G}$

$$\mathsf{Regret}_n(\widehat{G}; \mathscr{G}) \triangleq \sup_{G \in \mathscr{G}} \left\{ \mathbb{E}_G \left[(\widehat{\theta}_n(Y^n) - \theta_n)^2 \right] - \mathrm{mmse}(G) \right\}, \quad \widehat{\theta}_n(Y^n) = \widehat{\theta}_{\widehat{G}(Y^{n-1})}(Y_n). \tag{2.11}$$

Here $Y_1, \ldots, Y_{n-1}$ can be viewed as training data which is used to learn the estimator, and then we apply it on a fresh (unseen) data point Y_n to predict θ_n. Getting back to the loss function under consideration, it is not difficult to see that the total regret with sample size n can be bounded from above using n times the individual regret with training sample size $n-1$

$$\mathsf{TotRegret}_n(\widehat{G}; \mathscr{G}) \leq n \cdot \mathsf{Regret}_n(\widehat{G}; \mathscr{G}). \tag{2.12}$$

In view of the above inequality of the total and individual regret functions, we limit ourselves to studying individual regret only, as this will suffice to achieve the desired optimal rates.

Now, we are in a position to describe the main results for empirical Bayes estimation. For an ease of notation, suppose that given a fresh sample $Y \sim \mathrm{Poi}(\theta)$, where θ is generated from an unknown prior distribution G, we want to predict the value of θ in the squared error loss and training sample to construct the estimator $\widehat{G}$ is given by $Y_1, \ldots, Y_n$. Given any estimator $\widehat{G}$ of G we

define the regret of the empirical Bayes estimate $\widehat{\theta}_{\widehat{G}}$ as

$$\begin{aligned}\mathsf{Regret}(\widehat{G};G) &= \mathbb{E}_G\left[\left(\widehat{\theta}_{\widehat{G}}(Y)-\theta\right)^2\right]-\mathrm{mmse}(G)\\ &\overset{\text{(a)}}{=} \mathbb{E}_G\left[\left(\widehat{\theta}_{\widehat{G}}(Y)-\widehat{\theta}_G(Y)\right)^2\right]=\mathbb{E}_G\left[\sum_{y\geq 0}(\widehat{\theta}_{\widehat{G}}(y)-\widehat{\theta}_G(y))^2 f_G(y)\right],\end{aligned} \tag{2.13}$$

where the identity (a) followed by using the orthogonality principle: the average risk of any estimator $\widehat{\theta}$ can be decomposed as

$$\mathbb{E}_G[(\widehat{\theta}-\theta)^2]=\mathrm{mmse}(G)+\mathbb{E}_G[(\widehat{\theta}-\widehat{\theta}_G)^2]. \tag{2.14}$$

Similarly we define the maximum regret of $\widehat{G}$ over the class of data generating distributions $\mathscr{G}$

$$\mathsf{Regret}(\widehat{G};\mathscr{G})=\sup_{G\in\mathscr{G}}\mathsf{Regret}(\widehat{G};G). \tag{2.15}$$

Then we have the following estimation guarantees.

Theorem 3 (Empirical Bayes) *Let* dist *satisfy Assumption 1 and Assumption 2. Suppose that $\widehat{G}$ is the unconstrained minimum distance estimator given in (2.4). Then there exist constants c_1, c_2 such that the following holds*

(a) $\mathsf{Regret}(\widehat{G};\mathscr{P}([0,h]))\leq \frac{c_1}{n}\cdot\min\left\{\max\{1,h\}\frac{\log n}{\log\log n}, h+\log n\right\}^3$ *for any* $n\geq 3$

(b) $\mathsf{Regret}(\widehat{G};\mathsf{SubE}(s))\leq \frac{c_2\cdot\max\{1,s^3\}}{n}(\log n)^3$

In addition, if the data generating distribution G is supported on $[0,h]$ for a constant $h>0$, then the constrained minimum distance estimator $\widetilde{G}$ with access to h achieves an improved risk guarantee

$$\widetilde{G}=\operatorname*{argmin}_{Q\in\mathscr{P}[0,h]}\mathsf{dist}(p_n^{\mathsf{emp}}\|f_Q),\ \mathsf{Regret}(\widehat{G};\mathscr{P}([0,h]))\leq\frac{c_1}{n}\cdot\min\left\{\max\{1,h\}\frac{\log n}{\log\log n}, h+\log n\right\}^2.$$

Remark 3 *(i) As mentioned in Section 1, for fixed h and s, the above establishes the regret optimality of minimum-distance EB estimators by matching the minimax lower bounds recently shown in [57, Theorem 1]. This minimax optimality was only known for the f-modeling-based Robbins estimator.*

(ii) When dist *is the KL divergence, the minimum-distance estimator $\widehat{G}=\operatorname{argmin}_Q \mathsf{KL}(p\|q)$ is the NPMLE. This follows from the expansion*

$$KL(p_n^{\mathsf{emp}}\|f_Q)=\sum_{y\geq 0}p_n^{\mathsf{emp}}(y)\log\frac{p_n^{\mathsf{emp}}(y)}{f_Q(y)}=\sum_{y\geq 0}p_n^{\mathsf{emp}}(y)\log p_n^{\mathsf{emp}}(y)-\frac{1}{n}\sum_{i=1}^n\log f_Q(Y_i).$$

(iii) Theorem 3 holds for approximate solutions. Consider the following approximate minimum-distance estimators $\widehat{G}$, over some target class of distributions $\mathscr{G}$, that satisfies

$$\mathsf{dist}(p_n^{\mathsf{emp}}\|f_{\widehat{G}})\leq\inf_{Q\in\mathscr{G}}\mathsf{dist}(p_n^{\mathsf{emp}}\|f_Q)+\delta. \tag{2.16}$$

for some $\delta>0$. Then the regret bound for the bounded prior case (resp. subExponential data generating prior case) continues to hold if $\delta\lesssim\frac{\log n}{n\log\log n}$ (resp. $\frac{\log n}{n}$). Note that $\widehat{G}$ is the NPMLE

over $\mathscr{G}$ if $\delta = 0$ and dist *is given by KL divergence. In case of NPMLE,* (2.16) *translates to an approximate likelihood maximizer $\widehat{G}$ such that*

$$\frac{1}{n}\sum_{i=1}^{n}\log f_{\widehat{G}}(Y_i) \geq \operatorname*{argmax}_{G\in\mathscr{G}}\frac{1}{n}\sum_{i=1}^{n}\log f_G(Y_i) - \delta.$$

This type of results is well-known in the literature, see, for example, [30, 70] for the normal location-mixture model.

(iv) Similar to Remark 2, our results become trivial when h is of similar order to n, and this is primarily due to our proof strategy, which attempts to provide a uniform analysis for all priors over $[0,h]$ for a constant h. The estimators in practice perform well in many scenarios when h is large. To demonstrate the above, as suggested by one of the reviewers, we consider the prior distribution degenerate at $h = 50$ and performed a simulation study with $n = 50$ samples. We present below the performance of the unconstrained minimum distance estimators, in terms of the average RMSE and the average MAD metric, out of 25 repetitions. Notably, out estimators

TABLE 2 *Performance of the unconstrained estimators with $h = n = 50$*

Methods	Robbins	Naive	minimum-H^2	NPMLE	minimum-χ^2
Average RMSE	43.94	6.94	0.84	1.25	2.14
Average MAD	32.91	5.56	0.83	1.05	1.72

outperform both the Robbins estimator and the naive estimator, that uses the data point Y_i to estimate θ_i-s (which are now identically 50).

3. Proof for density estimation

The proof of Theorem 2 is based on a simple truncation idea. It is straightforward to show that the density estimation error for any minimum distance estimator can be bounded from above, within a constant factor, by the expected squared Hellinger distance between the empirical distribution p_n^{emp} and the data-generating distribution f_G, which is further bounded by the expected χ^2-distance. The major contribution to $\chi^2(p_n^{\mathsf{emp}}\|f_G)$ comes from the "effective support" of f_G, outside of which the total probability is $o(\frac{1}{n})$. For the the prior classes $\mathscr{P}([0,h])$ and $\mathsf{SubE}(s)$, the Poisson mixture f_G is effectively supported on $\{0,\dots,O(\frac{\log n}{\log\log n})\}$ and $\{0,\dots,O(\log n)\}$. Each point in the effective support contributes $\frac{1}{n}$ to $\chi^2(p_n^{\mathsf{emp}}\|f_G)$ from which our results follow.

Proof of Theorem 2 For any integer $K \geq 1$ and distribution G denote

$$\varepsilon_K(G) \triangleq \mathbb{P}[Y \geq K] = \sum_{y=K}^{\infty} f_G(y) \tag{3.1}$$

Note that dist satisfies Assumption 2, namely (2.3). We first prove the following general inequality

$$\mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\right] \leq \frac{4c_2}{c_1}\frac{K}{n} + \left(\frac{4c_2}{c_1} + 2n\right)\varepsilon_K(G). \tag{3.2}$$

Using the inequality $(a+b)^2 \le 2(a^2+b^2)$, and as $\widehat{G}$ is the minimizer we get

$$H^2(f_G, f_{\widehat{G}}) \le \left(H(p_n^{\mathsf{emp}}, f_{\widehat{G}}) + H(p_n^{\mathsf{emp}}, f_G)\right)^2 \le 2[H^2(p_n^{\mathsf{emp}}, f_{\widehat{G}}) + H^2(p_n^{\mathsf{emp}}, f_G)]$$
$$\le \frac{2}{c_1}(\mathsf{dist}(p_n^{\mathsf{emp}} \| f_{\widehat{G}}) + \mathsf{dist}(p_n^{\mathsf{emp}} \| f_G)) \le \frac{4}{c_1}\mathsf{dist}(p_n^{\mathsf{emp}} \| f_G). \tag{3.3}$$

Define $Y_{\max} \triangleq \max_{i=1}^n Y_i$. as before. Then, bounding $\frac{1}{c_2}d$ by χ^2 we get the following chain

$$\frac{1}{c_2}\mathbb{E}\left[\mathsf{dist}(p_n^{\mathsf{emp}} \| f_G)\mathbf{1}_{\{Y_{\max}<K\}}\right] \le \mathbb{E}\left[\chi^2(p_n^{\mathsf{emp}} \| f_G)\mathbf{1}_{\{Y_{\max}<K\}}\right] = \sum_{y\ge 0}\frac{\mathbb{E}\left[(p_n^{\mathsf{emp}}(y) - f_G(y))^2\mathbf{1}_{\{Y_{\max}<K\}}\right]}{f_G(y)}$$
$$\overset{\text{(a)}}{=} \sum_{y<K}\frac{\mathbb{E}\left[(p_n^{\mathsf{emp}}(y) - f_G(y))^2\mathbf{1}_{\{Y_{\max}<K\}}\right]}{f_G(y)} + \sum_{y\ge K} f_G(y)\mathbb{P}[Y_{\max} < K],$$

where the last equality follows from the fact that under $\{Y_{\max} < K\}$ we have $p_n^{\mathsf{emp}}(y) = 0$ for any $y \ge K$. Using $\mathbb{E}[p_n^{\mathsf{emp}}(y)] = f_G(y)$ and, thus, $\mathbb{E}[(p_n^{\mathsf{emp}}(y) - f_G(y))^2] = \mathrm{Var}(p_n^{\mathsf{emp}}(y)) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(\mathbf{1}_{\{Y_i=y\}}) = \frac{f_G(y)(1-f_G(y))}{n}$ we continue the last display to get

$$\frac{1}{c_2}\mathbb{E}\left[\mathsf{dist}(p_n^{\mathsf{emp}} \| f_G)\mathbf{1}_{\{Y_{\max}<K\}}\right] \le \sum_{y<K}\frac{\mathbb{E}\left[(p_n^{\mathsf{emp}}(y) - f_G(y))^2\mathbf{1}_{\{Y_{\max}<K\}}\right]}{f_G(y)} + \varepsilon_K(G)(1-\varepsilon_K(G))^n$$
$$\le \sum_{y<K}\frac{\mathbb{E}\left[(p_n^{\mathsf{emp}}(y) - f_G(y))^2\right]}{f_G(y)} + \varepsilon_K(G)(1-\varepsilon_K(G))^n$$
$$\overset{\text{(b)}}{=} \frac{1}{n}\sum_{y<K}(1-f_G(y)) + \varepsilon_K(G)(1-\varepsilon_K(G))^n \le \frac{K}{n} + \varepsilon_K(G). \tag{3.4}$$

Using the union bound and the fact $H^2 \le 2$ we have $\mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\mathbf{1}_{\{Y_{\max}\ge K\}}\right] \le 2\mathbb{P}[Y_{\max} \ge K] \le 2n\varepsilon_K(G)$. Combining this with (3.3) and (3.4) yields

$$\mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\right] \le \mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\mathbf{1}_{\{Y_{\max}<K\}}\right] + \mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\mathbf{1}_{\{Y_{\max}\ge K\}}\right]$$
$$\le \frac{4}{c_1}\mathbb{E}\left[\mathsf{dist}(p_n^{\mathsf{emp}} \| f_G)\mathbf{1}_{\{Y_{\max}<K\}}\right] + 2n\varepsilon_K(G) \le \frac{4c_2}{c_1}\frac{K}{n} + \left(\frac{4c_2}{c_1} + 2n\right)\varepsilon_K(G),$$

which completes the proof of (3.2).

To complete the proof of the theorem we need to choose the value of K such that $\varepsilon_K(G) \lesssim \frac{1}{n^2}$. This is done slightly differently for each of the two different classes of priors:

(a) Pick $K = \min\left\{\left\lceil\frac{5(he^2+2)\log n}{\log\log n}\right\rceil, he^2 + 5\log n\right\}$. Then using Lemma 11 with $a = 4$ we get the desired result.

(b) Let $G \in \mathsf{SubE}(s)$. Choose $K = \max\left\{1, \frac{2}{\log\left(1+\frac{1}{2s}\right)}\right\}\log n$. Then Lemma 12 in Appendix D implies that $\varepsilon_K(G) \le \frac{3}{n^2}$. Plugging this in (3.2) completes the proof.

□

4. Proof of regret upper bound

4.1. *General regret upper bound via density estimation*

The proof of Theorem 3 relies on relating the regret in EB estimation to estimating the mixture density in the Hellinger distance. This idea has been previously noted in [30, Theorem 3] for the Gaussian location models using Fourier analysis and an ingenious induction argument. Here the analysis turns out to be much simpler thanks in part to the discreteness of the Poisson model and the light tail of the prior, leading to the following deterministic result which is crucial for proving the regret optimality of minimum-distance EB estimators.

Lemma 4 *Let G be a distribution such that $\mathbb{E}_G[\theta^4] \le M$ for some constant M. Then for any distribution $\widehat{G}$ supported on $[0,\widehat{h}]$, any $h > 0$ with $G([0,h]) > \frac{1}{2}$ and any $K \ge 1$,*

$$\begin{aligned}\mathsf{Regret}(\widehat{G};G) \le \left\{12(h^2+\widehat{h}^2)+48(h+\widehat{h})K\right\}\left(H^2(f_G,f_{\widehat{G}})+4G((h,\infty))\right)\\ +2(h+\widehat{h})^2\varepsilon_K(G)+2(1+2\sqrt{2})\sqrt{(M+\widehat{h}^4)G((h,\infty))}\end{aligned}$$

where $\mathsf{Regret}(\widehat{G};G)$ *and* $\varepsilon_K(G)$ *were defined in* (3.1) *and (2.13) respectively.*

Note that $\widehat{G}$ in the above statement denotes, with a slight abuse of notation, an arbitrary deterministic prior and to compute the regret for the random estimator $\widehat{G}$ we will need to consider the expectation. This will be clarified in the proof of Theorem 3 later on. We provide a sketch of the proof here (see Appendix C for the full proof.) It is relatively easy to bound the regret if the corresponding Bayes estimator is also bounded, which is the case if the prior G is compactly supported. Otherwise, one can consider its restriction G_h on $[0,h]$ defined by $G_h(\cdot) = \frac{G(\cdot\cap[0,h])}{G([0,h])}$. The truncation error can be controlled using properties of the mmse as follows:

$$\mathsf{Regret}(\widehat{G};G) \le \mathsf{Regret}(\widehat{G};G_h) + \frac{(1+2\sqrt{2})\sqrt{(M+\widehat{h}^4)G((h,\infty))}}{G([0,h])}. \tag{4.1}$$

Then we use the structure of the Bayes estimator (1.1) in the Poisson model to relate $\mathsf{Regret}(\widehat{G};G_h)$ to the squared Hellinger distance between f_{G_h} and $f_{\widehat{G}}$

$$\mathsf{Regret}(\widehat{G};G_h) \le \left\{6(h^2+\widehat{h}^2)+24(h+\widehat{h})K\right\}H^2(f_{G_h},f_{\widehat{G}})+(h+\widehat{h})^2\varepsilon_K(G_h), \tag{4.2}$$

for any $K \ge 0$. We then show that $\varepsilon_K(G_h)$ and $H^2(f_{G_h},f_{\widehat{G}})$ satisfies

$$\varepsilon_K(G_h) \le 2\varepsilon_K(G), \quad H^2(f_{\widehat{G}},f_{G_h}) \le 2\left\{H^2(f_G,f_{\widehat{G}})+4G((h,\infty))\right\}.$$

Replacing these bounds in (4.2) we get the desired result.

4.2. *Proof of Theorem 3*

For rest of the section, let $C_1, C_2, \ldots$ denote constants independent of h, s, d as required. For Part (a), recall that $\widehat{G} = \operatorname{argmin}_{Q\in\mathscr{P}(\mathbb{R}^+)} \mathsf{dist}\left(p_n^{\mathsf{emp}} \| f_Q\right)$ is the unconstrained minimum distance estimator.

To apply Lemma 4, set

$$\widehat{h} = Y_{\max}, \quad K = \min\left\{\left\lceil \frac{5(he^2+2)\log n}{\log\log n}\right\rceil, he^2 + 5\log n\right\}, \quad M = h^4.$$

Then, in view of Lemma 11 we get that

$$\mathbb{P}\left[\widehat{h} > K\right] \le n \cdot \mathbb{P}[Y_1 > K] \le \frac{2}{n^4}. \tag{4.3}$$

For any $G \in \mathscr{P}([0,h])$ we have from the proof of Theorem 2(a)

$$\varepsilon_K(G) \le \frac{C_1}{n^3}, \quad G((h,\infty)) = 0, \quad \mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\right] \le C_1 \frac{K}{n}.$$

Then Lemma 4 yields the required bound

$$\begin{aligned}\mathsf{Regret}(\widehat{G};G) &\le C_3\left\{h^2 + K^2 + (h+K)K\right\}\mathbb{E}\left[H^2\left(f_G, f_{\widehat{G}}\right)\right] + C_4\mathbb{E}\left[\left(h^2 + \widehat{h}^2 + (h+\widehat{h})K\right)\mathbf{1}_{\{\widehat{h}>K\}}\right] \\ &\quad + \frac{C_4\mathbb{E}\left[(h+\widehat{h})^2\right]}{n^3} \\ &\overset{(a)}{\le} C_5\left(\frac{K^3}{n} + \sqrt{\mathbb{E}\left[h^4 + \widehat{h}^4 + K^4\right]\cdot\mathbb{P}[\widehat{h} > K]}\right) \overset{(b)}{\le} \frac{C_6K^3}{n},\end{aligned}$$

where (a) followed from the Cauchy-Schwarz inequality, and (b) from (4.3) and Lemma 11.

To achieve the result involving the constrained optimizer, we use $\widehat{h} = h$. Following the proof of Theorem 2, note that we only required the optimality property of the estimator of G in (3.3), and the above equation holds true for the constrained estimator $\widetilde{G}$ as well. Hence we can show

$$\sup_{G\in\mathscr{P}[0,h]} \mathbb{E}\left[H^2(f_G, f_{\widetilde{G}})\right] \le \frac{c_1}{n}\cdot\min\left\{\max\{1,h\}\frac{\log n}{\log\log n}, h + \log n\right\}$$

as well. In view of Lemma 4 we get $\mathsf{Regret}(\widehat{G};G) \le C_6\left\{h^2 + hK\right\}\mathbb{E}\left[H^2\left(f_G, f_{\widetilde{G}}\right)\right] + \frac{C_7h^2}{n^3} \le C_8\frac{hK^2}{n}$.

For Part (b), we choose

$$h = 4s\log n, \quad K = \max\left\{1, \frac{2}{\log\left(1+\frac{1}{2s}\right)}\right\}\log n, \quad M = 12s^4. \tag{4.4}$$

Since G is s-subexponential, we have (see Lemma 12 in Appendix D for details)

$$\mathbb{E}_G[\theta^4] \le M, \quad G((h,\infty)) \le \frac{2}{n^4}, \quad \varepsilon_K(G) \le \frac{3}{n^2}, \quad \varepsilon_{2K}(G) \le \frac{3}{n^4}. \tag{4.5}$$

In view of Lemma 9 in Appendix B we get that $\widehat{G}$ is supported on $[0,\widehat{h}]$ where $\widehat{h} = Y_{\max}$. Then Lemma 4 and $(\mathbb{E}_G[Y_{\max}^2])^2 \le \mathbb{E}_G[Y_{\max}^4] \le \max\{1,s^4\}(\log n)^4$ (see Appendix D for a proof) implies

$$\mathsf{Regret}(\widehat{G};G) \le \mathbb{E}\left[\left\{6(h^2 + Y_{\max}^2) + 24K(h+Y_{\max})\right\}H^2(f_G, f_{\widehat{G}})\right] + \frac{C_5}{n}. \tag{4.6}$$

Next we bound the expectation in the last display. Using the fact that $H^2 \leq 2$, we get

$$\begin{aligned}
&\mathbb{E}\left[\left\{(h^2+Y_{\max}^2)+4K(h+Y_{\max})\right\}H^2(f_G,f_{\widehat{G}})\right] \\
&\leq (h^2+4Kh+12K^2)\mathbb{E}\left[H^2\left(f_G,f_{\widehat{G}}\right)\right]+2\mathbb{E}\left[\left\{(h^2+Y_{\max}^2)+4K(h+Y_{\max})\right\}\mathbf{1}_{\{Y_{\max}\geq 2K\}}\right] \qquad (4.7)
\end{aligned}$$

Using Theorem 2 we get that the first part on the right of the above inequality is bounded as

$$(h^2+4Kh+12K^2)\mathbb{E}\left[H^2\left(f_G,f_{\widehat{G}}\right)\right] \leq c_0 \max\left\{1,s^3\right\}\frac{(\log n)^3}{n}$$

for some absolute constant $c_0 > 0$. For the second term in (4.7) we use Cauchy-Schwarz inequality and union bound to get

$$\begin{aligned}
&\mathbb{E}\left[\left\{6(h^2+Y_{\max}^2)+24K(h+Y_{\max})\right\}\mathbf{1}_{\{Y_{\max}\geq 2K\}}\right] \\
&\leq \sqrt{\mathbb{E}\left[\left\{6(h^2+Y_{\max}^2)+24K(h+Y_{\max})\right\}^2\right]\mathbb{P}_G\left[Y_{\max}\geq 2K\right]} \\
&\overset{(a)}{\leq} \sqrt{\mathbb{E}\left[\left\{6(h^2+Y_{\max}^2)+24K(h+Y_{\max})\right\}^2\right]n\varepsilon_{2K}(G)} \\
&\overset{(b)}{\leq} \frac{6}{n^{3/2}}\sqrt{\mathbb{E}\left[\left\{4(h^4+Y_{\max}^4)+16K^2(h^2+Y_{\max}^2)\right\}\right]} \leq \frac{1}{n}.
\end{aligned}$$

where (a) followed from (4.5) and (b) followed for large enough n. Plugging the bounds back in (4.7) and in view of (4.6), we complete the proof.

5. Numerical experiments

In this section, we analyze the performances of the empirical Bayes estimators based on the minimum-H^2, the minimum-χ^2, and the minimum-KL divergence estimator (i.e., the NPMLE). We compare them against the Robbins estimator and also draw comparisons among their individual performances. Unlike the Robbins estimator, the minimum-distance based estimators do not admit a closed form solution. Our algorithm to compute the solution is closely related to the vertex direction method (VDM) algorithms for finding NPMLE [42, 45], specialized for the Poisson family and modified to work with the generalized distance we considered. In the case of the NPMLE, the convergence of the VDM method to the unique optimizer is well-known [22, 67], and the algorithms for finding the other minimum dist-distance estimators are expected to show similar convergence guarantees as well. Additionally, thanks to the Poisson density, the first-order optimality condition takes on a polynomial form, allowing us to use existing root-finding algorithms for polynomials to update the support points of the solution. See [61] for a similar VDM-type algorithm for Poisson mixtures and [34, 35] for discretization-based algorithms.

5.1. *First-order optimality condition and algorithm*

In the numerical experiments we focus on the unconstrained minimum-distance estimator $\widehat{G} = \operatorname{argmin}_Q \mathsf{dist}(p_n^{\mathsf{emp}} \| f_Q)$, which is a discrete distribution (Theorem 1). For any $\theta \in \mathbb{R}_+$ let δ_θ denote the Dirac measure at θ. Suppose that the support of p_n^{emp} be $\{y_1,\ldots,y_m\}$. The optimality of $\widehat{G}$

implies that for all $\theta, \varepsilon \in [0,1]$ we have $\mathsf{dist}(p_n^{\mathsf{emp}} \| f_{\widehat{G}}) \le \mathsf{dist}(p_n^{\mathsf{emp}} \| f_{(1-\varepsilon)\widehat{G}+\varepsilon\delta_\theta})$, leading to the first-order optimality condition $\frac{d}{d\varepsilon}\mathsf{dist}(p_n^{\mathsf{emp}} \| f_{(1-\varepsilon)\widehat{G}+\varepsilon\delta_\theta})\Big|_{\varepsilon=0} \ge 0$, namely

$$D_{\widehat{G}}(\theta) \triangleq \sum_{i=1}^{m} \frac{d}{df}\ell(p_n^{\mathsf{emp}}(y_i), f)\Big|_{f=f_{\widehat{G}}(y_i)} (f_\theta(y_i) - f_{\widehat{G}}(y_i)) \ge 0. \tag{5.1}$$

Averaging the left hand side over $\theta \sim \widehat{G}$, we get $\int D_{\widehat{G}}(\theta) d\widehat{G}(\theta) = 0$. This implies that each θ in the support of $\widehat{G}$ satisfies $D_{\widehat{G}}(\theta) = 0$. Taking derivative on both sides of the equation $D_{\widehat{G}}(\theta) = 0$ with respect to θ we get that the atoms of $\widehat{G}$ satisfies the following polynomial equation in θ

$$\sum_{i=1}^{m} w_i(\widehat{G}) \left(y_i \theta^{y_i - 1} - \theta^{y_i}\right) = 0, \quad w_i(\widehat{G}) = \left\{ \frac{d}{df}\ell(p_n^{\mathsf{emp}}(y_i), f)\Big|_{f=f_{\widehat{G}}(y_i)} \right\} / y_i!.$$

Iterating the above conditions leads to following algorithm for computing $\widehat{G}$.

Algorithm 1 Computing the minimum dist-distance estimators

Input: Data points $Y_1, \ldots, Y_n$. Target distribution $G_{\boldsymbol{\theta},\boldsymbol{\mu}} = \sum_j \mu_j \delta_{\theta_j}$. Divergence dist with $t - \ell$ decomposition $\mathsf{dist}(q_1 \| q_2) = t(q_1) + \sum_{y \ge 0} \ell(q_1(y), q_2(y))$. Initialization of $(\boldsymbol{\theta}, \boldsymbol{\mu})$. Tolerance η_1, η_2 and number of iterations N.

Steps:

1: Calculate empirical distribution p_n^{emp}. Obtain the set of distinct sample entries $\{y_1, \ldots, y_m\}$.
2: **for** N iterations **do**
3: $\quad$ *newroots* $= \left\{\theta : \theta \ge 0, \sum_{i=1}^{m} w_i(G_{\boldsymbol{\theta},\boldsymbol{\mu}}) \left(y_i \theta^{y_i - 1} - \theta^{y_i}\right) = 0\right\}$.
4: $\quad$ Combine $\boldsymbol{\theta}$ and *newroots* and denote the new vector as $\boldsymbol{\theta}'$.
5: $\quad$ Merge entries of $\boldsymbol{\theta}'$ that are within η_1 distance of each other.
6: $\quad$ Find $\mathrm{argmin}_{\widetilde{\boldsymbol{\mu}}} \sum_{i=1}^{m} \ell(p_n^{\mathsf{emp}}(y_i), f_{G_{\boldsymbol{\theta}', \widetilde{\boldsymbol{\mu}}}}(y_i))$, via gradient descent with initialization $\widetilde{\boldsymbol{\mu}} = \boldsymbol{\mu}$.
7: $\quad$ Remove entries of $\boldsymbol{\theta}'$ and $\boldsymbol{\mu}'$ at locations of $\boldsymbol{\mu}'$ that are less than η_2 and re-normalize $\boldsymbol{\mu}'$.
8: $\quad$ $(\boldsymbol{\theta}, \boldsymbol{\mu}) \leftarrow (\boldsymbol{\theta}', \boldsymbol{\mu}')$.
9: **end for**

Output: $(\boldsymbol{\theta}, \boldsymbol{\mu})$.

We apply this algorithm for finding the minimum-distance estimators in the following examples. In all our experiments we used $\eta_1 = 0.01, \eta_2 = 0.001$. We set the maximum number of iterations N to be 15 as the outputs of the algorithm in all our simulations were observed to converge by then. We choose the initialization for $\boldsymbol{\theta}$ as the uniform grid of size 1000 over the interval $[0, Y_{\max}]$, with a uniform initial probability assignment $\boldsymbol{\mu}$.

5.2. *Real-data analysis: Prediction of hockey goals*

Here we extend our study related to the National Hockey League as mentioned in Section 1. In Fig. 3 we first plot the actual data, where for each true data point, its x-axis represents the number of goals scored by a particular player in the 2017-18 season (denoted as "Past") and the y-axis of the data represents the goals by the same player in the 2018-19 season (denoted as "Future"). Then,

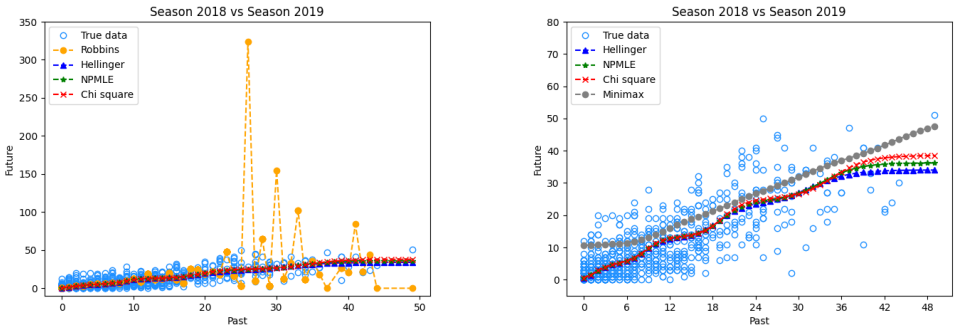

FIG. 3. Prediction of hockey goals with empirical Bayes, comparing Robbins and minimum-distance estimators. On the right panel, the Robbins estimator is replaced by the minimax estimator.

for each possible value of the goal scored in the 2017-18 season, we plot the EB estimators based on the Robbins method, the minimum H^2, the minimum-χ^2 distance estimator, and the NPMLE, and the *gold standard* minimax estimator.

The left panel shows that a large number of individuals exist for whom the Robbins estimator produces unstable predictions that are significantly worse than those of all minimum-distance methods. This difference is significant for the values of scored goals, which have lower sample representations. Thus, on the right panel, we omit the Robbins estimator and provide a more detailed comparison of the three minimum-distance estimators, which shows that their behavior is mostly comparable except near the tail end of the data points. We also present the comparison with the minimax estimator given by the conditional estimator of the least favorable prior in Fig. 1, and our plot shows that the EB estimators align more closely to the actual data.

Interestingly, all three estimators seem to do shrinkage towards several fixed values. There could be several explanations for this multi-modality. One is that different clusters correspond to different player positions (defense, winger, center). The other is that clusters correspond to the line of the player (different lines get different amounts of ice time). To test this hypothesis, we also redid on Fig. 3 the estimation for each position separately. Since the multi-modality is retained, we conclude that the second option is more likely to be the real explanation. In addition, we also compared the four goal-prediction methods based on different EB estimators and the minimax estimator based on the least favorable prior on $[0,50]$ across the possible playing positions: defender, center, and winger. Similar to before, we used the Poisson model and tried to predict the goal scoring for the year 2019 using the goal scoring data from the year 2018 for players in each playing position separately. As expected, the minimum distance methodology provides more stable and accurate estimates than the estimates based on the Robbins method and minimax strategy. The plots showing the closeness of the predictions to the actual number of goals for the different EB methods are provided in Fig. 4.

5.3. *Application of EB methods for filtering*

In this section, we demonstrate the application of the EB methodology to data cleaning. We propose to show that given a standard statistical methodology, incorporating an EB-based filter on the data before feeding it to the algorithm, can significantly improve the existing performance guarantees. For our analysis we use simulated data based on multivariate linear models.

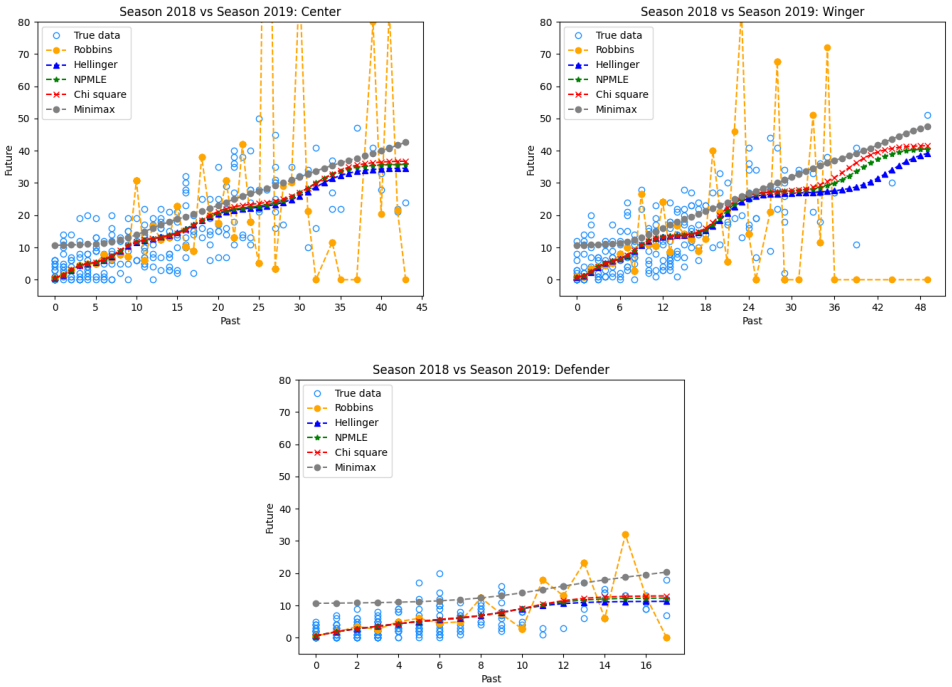

FIG. 4. Prediction of hockey goals at different playing positions.

For this simulation study, we assume that the observed data $y_1,\dots,y_n$ are independently generated via a linear model

$$y_i \overset{\text{ind.}}{=} \boldsymbol{\theta}_i\beta, \quad \boldsymbol{\theta}_i = \{\theta_{ij}\}_{j=1}^{d} \in \mathbb{R}_+^d, \beta \in \mathbb{R}^d, \quad i = 1,\dots,n.$$

In addition, we also assume that the observer does not directly see the data generating θ_i-s. Instead, we can only observe a Poissonized version X_i of θ_i, given as

$$\boldsymbol{X}_i = \{X_{ij}\}_{j=1}^{d}, \quad X_{ij} \overset{\text{ind.}}{\sim} \text{Poi}(\theta_{ij}), \quad i = 1,\dots,n, \quad j = 1,\dots,d.$$

In other words, each coordinate of $\boldsymbol{X}_i$ is generated independently according to a Poisson channel, with mean being the corresponding coordinate of $\boldsymbol{\theta}_i$. Then we pose the following question:

Upon observing $\{(y_i,\boldsymbol{X}_i)\}_{i=1}^n$, can we achieve better error guarantees if we apply the EB filters to the covariates before running the ordinary least squares (OLS) methodology?

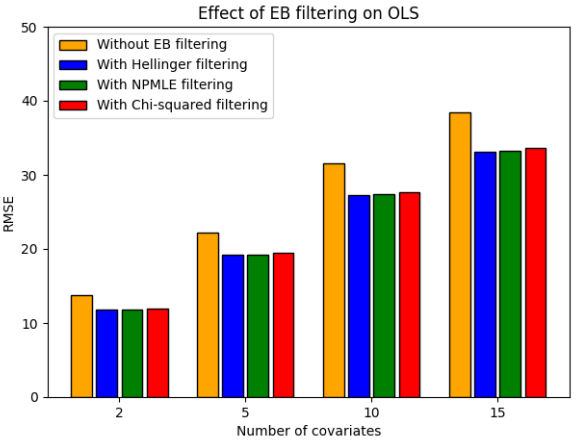

FIG. 5. Improving results for OLS using EB filtering

To answer the problem, we will show that EB-based one-dimensional data filters, applied separately on each of the covariates before running the OLS, can significantly improve the prediction of y if the $\boldsymbol{\theta}_i$-s values are coming from a multivariate mixture model. Our process for generating the data $\{(y_i, \boldsymbol{X}_i)\}_{i=1}^n$ is outlined as follows. To generate each coordinate of $\boldsymbol{\theta}_i$, we randomly generated entries from a uniform mixture of Gaussian distributions, with means $[2, 8, 16, 32]$ and standard deviation 1, and then took the absolute values. Each coordinate of the regression coefficient β was selected uniformly from $[-5, 5]$. The number of covariates (i.e., d) are varied in the set $[2, 5, 10, 15]$. A total of 1200 samples $(y_1, ..., y_{1200})$ were generated. To assess the goodness of the fit, we compute $\widehat{y}$ and the corresponding root mean squared error (RMSE). To examine the effect of the one-dimensional EB filters on the covariates, we used NPMLE, Hellinger-based, and Chi-squared distance based one dimensional EB filters. The RMSE prediction errors were compared with and without the filtering. In all the simulations, the EB filtering improved the result. The plot of errors is presented in Fig. 5, and the errors are reported (along with standard deviations inside parenthesis) in Table 3.

TABLE 3 *Performance of EB filtering*

Covariates	Without filtering	H^2 filter	NPMLE filter	χ^2 filter
2	13.75 (0.149)	11.857 (0.131)	11.867 (0.131)	11.993 (0.132)
5	22.25 (0.145)	19.262 (0.128)	19.277 (0.128)	19.497 (0.129)
10	31.621 (0.143)	27.327 (0.126)	27. 349 (0.126)	27.670 (0.128)
15	38.412 (0.145)	33.163 (0.129)	33.195 (0.129)	33.578 (0.130)

5.4. *More simulation studies*

In this subsection, we test more priors in addition to the uniform prior in Fig. 2, including discrete priors and priors with unbounded support. In Section 5.2 we see that the three minimum-distance

estimators performed similarly. However, the question arises whether the best choice among the minimum-distance EB methods can be argued when some information about the prior is available. With the specific goal of differentiating the three minimum-distance estimators among themselves, we carry out simulation studies at the end of this section using different priors.

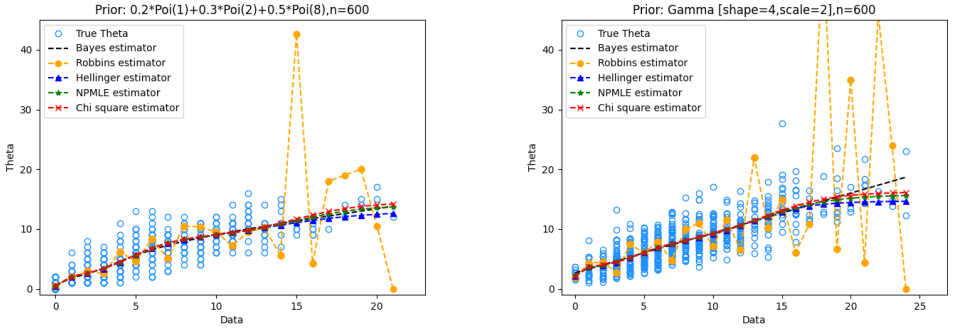

FIG. 6. Robbins vs. minimum-distance estimators: Unbounded priors

For comparing the EB methods in the discrete setup we choose the prior G to be $0.2\text{Poi}(1) + 0.3\text{Poi}(2) + 0.5\text{Poi}(8)$ and for the continuous unbounded setup we choose the prior G to be the Gamma distribution with scale parameter 2 and shape parameter 4, i.e., with prior density $f(x) = \frac{1}{96}x^3e^{-\frac{x}{2}}$. In both of the cases we simulate $\{\theta_i\}_{i=1}^{600}$ independently from the prior distribution and correspondingly generate data $Y_i \sim \text{Poi}(\theta_i)$. For each of the priors we calculate the Bayes estimator numerically (denoted by the black dashed line in the plots). Then, from the generated datasets, we compute the Robbins estimator, the NPMLE-based EB estimator, the H^2-distance-based EB estimator, and the χ^2-distance-based EB estimator. All the estimators are then plotted against θ and the data (Fig. 6). As expected, the Robbins estimator shows high deviation from the true θ values in many instances whereas the minimum-distance based estimators are much more stable.

To differentiate the different minimum-distance based EB methods we analyze the effect of the tail properties of the prior in the simulations below. Consider the exponential distribution parameterized by scale (α) and with density $g_\alpha(x) = \frac{1}{\alpha}e^{-x/\alpha}$. Note that the higher values of α generate distributions with heavier tails. We consider three values of α: 0.3,1.05 and 2. For each α we estimate the training regret for sample sizes n in the range $[50, 300]$. Given sample $Y_1, \ldots, Y_n$ from the mixture distribution with prior G we define the training regret for any estimator $\widehat{G}$ of G as $\mathbb{E}_G[\frac{1}{n}\sum_{i=1}^n(\widehat{\theta}_G(Y_i) - \widehat{\theta}_{\widehat{G}}(Y_i))^2]$. We compute the Bayes estimator $\widehat{\theta}_G(y)$ numerically for each y. For every pair (α, n) we replicate the following experiment independently 10,000 times for each minimum-distance method:

- Generate $\{\theta_i\}_{i=1}^n$ and $Y_i \sim \text{Poi}(\theta_i)$,
- Calculate $\widehat{G}$ using minimum-distance method,
- Calculate prediction error $\mathsf{E}(Y^n) = \frac{1}{n}\sum_{i=1}^n(\widehat{\theta}_G(Y_i) - \widehat{\theta}_{\widehat{G}}(Y_i))^2$.

Then we take the average of $\mathsf{E}(Y^n)$ values from all the 10,000 replications to estimate the training error. For each α and each minimum distance method, at every n we also estimate the 95% confidence interval as $[\overline{\mathsf{E}(Y^n)} \mp 0.0196 * \mathsf{sd}(\mathsf{E}(Y^n))]$ where $\overline{\mathsf{E}(Y^n)}$ and $\mathsf{sd}(\mathsf{E}(Y^n))$ define respectively

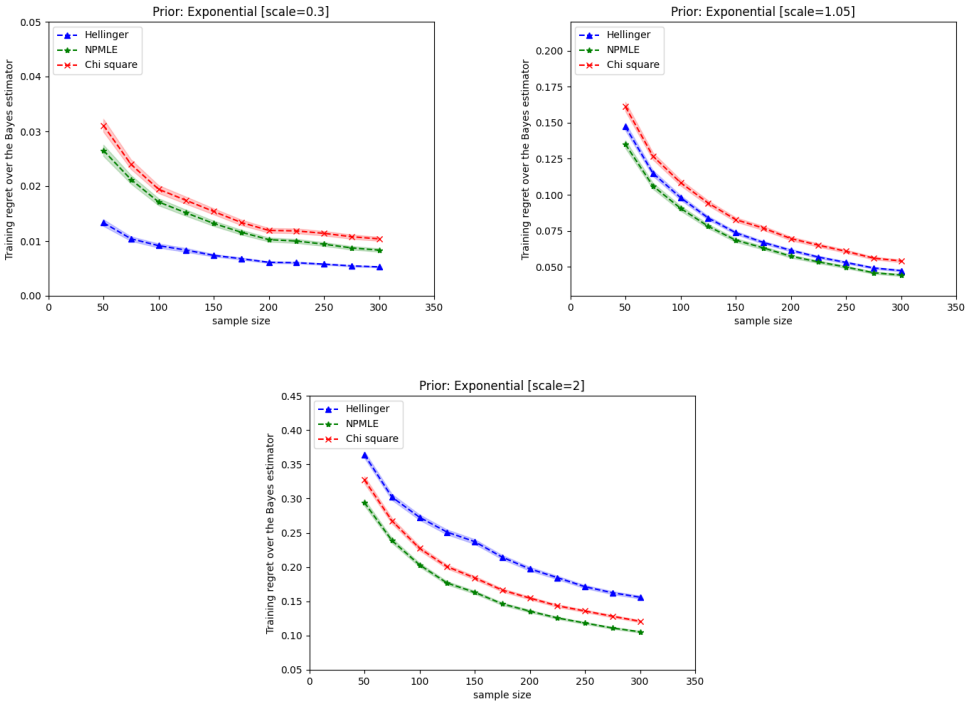

FIG. 7. Comparison of minimum-distance estimators

the sample mean and the sample standard deviation of the $\mathsf{E}(Y^n)$ values over the 10,000 independent runs. Below we plot the training regrets and their 95% confidence bands against the training sample sizes (Fig. 7). We observe that that minimum-H^2 based estimator outperforms the other estimators when the scale of the exponential distribution is small. As the tails of the prior distributions become heavier, the performance of the minimum-H^2 based estimator gets worse and the NPMLE based estimator comes out as a better choice.

6. An extension of the results in multiple dimensions

For a clarity of notations, in this section we use the bold fonts to denote vectors, e.g., $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_d), \boldsymbol{\theta}_i = (\theta_{i1},\ldots,\theta_{id}), \boldsymbol{Y} = (Y_1,\ldots,Y_d), \boldsymbol{Y}_i = (Y_{i1},\ldots,Y_{id}), \boldsymbol{y} = (y_1,\ldots,y_d)$, etc. Let G be a prior distribution on $\mathbb{R}_+^d$. We will study an extension of our minimum distance estimators (2.2) in the d-dimension coordinate-wise independent Poisson model: Consider the following data-generating process [10, 31]

$$\boldsymbol{\theta}_i \overset{\text{i.i.d.}}{\sim} G, \quad Y_{ij} \overset{\text{ind.}}{\sim} \text{Poi}(\theta_{ij}), \quad i = 1,\ldots,n, \quad j = 1,\ldots,d. \tag{6.1}$$

Note that the marginal distribution of the multidimensional Poisson mixture is given by

$$f_G(\boldsymbol{y}) = \int_{\boldsymbol{\theta}} \prod_{i=1}^{d} e^{-\theta_i} \frac{\theta_i^{y_i}}{y_i!} dG(\boldsymbol{\theta}), \quad \boldsymbol{y} \in \mathbb{Z}_+^d.$$

To construct the minimum distance estimator we use the same minimization principle as in (2.2), where we will specify the prior class $\mathscr{G}$ used for optimization when we state the results. Next we construct the EB estimator. Denote by $\widehat{\boldsymbol{\theta}}_G$ the Bayes estimator, whose j-th coordinate $\widehat{\theta}_{G,j}$ is

$$\widehat{\theta}_{G,j}(\boldsymbol{y}) = \mathbb{E}_G[\theta_j|\boldsymbol{y}] = \frac{\int_{\boldsymbol{\theta}} \theta_j \prod_{j=1}^d e^{-\theta_j} \frac{\theta_j^{y_j}}{y_j!} dG(\boldsymbol{\theta})}{f_G(\boldsymbol{y})} = (y_j+1)\frac{f_G(\boldsymbol{y}+\boldsymbol{e}_j)}{f_G(\boldsymbol{y})}, \quad j=1,\dots,d,$$

where $\boldsymbol{e}_j$ denote the j-th coordinate vector. Suppose that $\widehat{G}$ gives us an estimate of the prior distribution G and consider the corresponding empirical Bayes estimator $\widehat{\boldsymbol{\theta}}_{\widehat{G}} = (\widehat{\theta}_{\widehat{G},1},\dots,\widehat{\theta}_{\widehat{G},d})$. Similar to (2.13), let us define the regret of any plug-in estimator based on a prior estimate $\widehat{G}$ as

$$\mathsf{Regret}(\widehat{G},G) = \mathbb{E}_G\left[\|\widehat{\boldsymbol{\theta}}_{\widehat{G}}(\boldsymbol{Y}) - \widehat{\boldsymbol{\theta}}_G(\boldsymbol{Y})\|^2\right] = \mathbb{E}_G\left[\sum_{\boldsymbol{y}\in\mathbb{Z}_+^d} \|\widehat{\boldsymbol{\theta}}_{\widehat{G}}(\boldsymbol{y}) - \widehat{\boldsymbol{\theta}}_G(\boldsymbol{y})\|^2 f_G(\boldsymbol{y})\right], \tag{6.2}$$

where $\boldsymbol{Y} \sim f_G$ is a test point independent from the training sample $\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_n \overset{\text{i.i.d.}}{\sim} f_G$ ($\mathbb{Z}_+^d$ denotes the set of all d-dimensional vectors with non-negative integer coordinates). We will prove regret bounds for the minimum distance estimator of the form (2.2) where the dist function satisfies the following regularity assumption.

Assumption 3 *There exist maps $\boldsymbol{t} : \mathscr{P}(\mathbb{Z}_+^d) \to \mathbb{R}$ and $\ell : \mathbb{R}^2 \to \mathbb{R}$ such that for any two distributions $q_1, q_2 \in \mathscr{P}(\mathbb{Z}_+^d)$*

$$\mathsf{dist}\,(q_1\|q_2) = \boldsymbol{t}(q_1) + \sum_{\boldsymbol{y}\in\mathbb{Z}_+^d} \ell(q_1(\boldsymbol{y}), q_2(\boldsymbol{y})),$$

where $b \mapsto \ell(a,b)$ is strictly decreasing and strictly convex for $a > 0$ and $\ell(0,b) = 0$ for $b \geq 0$.

Assumption 4 *There exist absolute constants $c_1, c_2 > 0$ such that for pmf-s q_1, q_2 on $\mathbb{Z}_+^d$*

$$c_1 H^2(q_1,q_2) \leq \mathsf{dist}(q_1\|q_2) \leq c_2 \chi^2(q_1\|q_2). \tag{6.3}$$

Assumption 3 and Assumption 4 are identical to Assumption 1 and Assumption 2 respectively. The loss functions such as the Kullback-Leibler divergence, squared Hellinger distance, Chi-squared divergence satisfy the above assumptions, similarly as in the one-dimensional case. We have the following results.

Theorem 5 *Let* dist *satisfy Assumption 3 and Assumption 4. Suppose that $\widehat{G}$ is the unconstrained minimum distance estimator*

$$\widehat{G} = \operatorname*{argmin}_{Q\in\mathscr{P}(\mathbb{R}_+^d)} \mathsf{dist}(p_n^{\mathsf{emp}}\|f_Q). \tag{6.4}$$

Then there exist constants c_1, c_2 such that the following holds

(i) $\sup_{G\in\mathscr{P}([0,h]^d)} \mathbb{E}\left[H^2(f_{\widehat{G}}, f_G)\right] \le \frac{(c_1K)^d}{n}$, *where* $K = \min\left\{\max\{1,h\}\frac{\log n}{\log\log n}, h+\log n\right\}$;

(ii) If the data generating prior G belongs to a class $\mathscr{G}$ *where all marginals of* $G\in\mathscr{G}$ *belong to the* $\mathsf{SubE}(s)$ *class of distributions for some* $s>0$, *then*

$$\sup_{G\in\mathscr{G}} \mathbb{E}\left[H^2(f_{\widehat{G}}, f_G)\right] \le \frac{(c_2\max\{1,s\})^d}{n}(\log(n))^d.$$

Theorem 6 *Suppose that the assumptions in Theorem 5 hold true. Then the following regret bounds hold for the unconstrained estimator (6.4) whenever* $n \ge d$ ($c_1, c_2 > 0$ *below are constants):*

(i) $\mathsf{Regret}(\widehat{G}; \mathscr{P}([0,h]^d)) \le \frac{d(c_1K)^{d+2}}{n}$, *where* $K = \min\left\{\max\{1,h\}\frac{\log n}{\log\log n}, h+\log n\right\}$;

(ii) If the data generating prior G belongs to a class $\mathscr{G}$ *where all marginals of* $G\in\mathscr{G}$ *belong to the* $\mathsf{SubE}(s)$ *class of distributions for some* $s>0$, *then*

$$\mathsf{Regret}(\widehat{G};\mathscr{G}) \le \frac{d\,(c_2\max\{1,s\})^{d+2}}{n}(\log(n))^{d+2}.$$

In addition, in the case when the data generating distribution G is supported on $[0,h]^d$, *then the constrained minimum distance estimator with access h achieves improved risk guarantee*

$$\widetilde{G} = \operatorname*{argmin}_{Q\in\mathscr{P}[0,h]^d} \mathsf{dist}(p_n^{\mathsf{emp}} \| f_Q), \quad \mathsf{Regret}(\widetilde{G}; \mathscr{P}([0,h]^d)) \le \frac{dc_1^{d+2}\max\{1,h\}}{n}K^{d+1}.$$

The proofs of the above results are provided in Appendix E below. We conjecture these regret bounds in Theorem 6 are nearly optimal and factors like $(\log n)^d$ are necessary. A rigorous proof of matching lower bound for Theorem 6 will likely involve extending the regret lower bound based on Bessel kernels in [57] to multiple dimensions; this is left for future work.

Data availability

The real data set on hockey goals that we used is available at https://www.hockey-reference.com/. Our code is available at https://github.com/janasoham/codes_public/tree/main/mindist_poisson_eb.

Acknowledgment

Y. Polyanskiy is supported in part by the MIT-IBM Watson AI Lab, and the NSF Grants CCF-1717842, CCF-2131115. Y. Wu is supported in part by the NSF Grant CCF-1900507, NSF CAREER award CCF-1651588, and an Alfred Sloan fellowship.

References

1. Alton Barbehenn and Sihai Dave Zhao. A nonparametric regression alternative to empirical bayes approaches to simultaneous estimation. *arXiv preprint arXiv:2205.00336*, 2022.
2. Ayanendranath Basu, Hiroyuki Shioya, and Chanseok Park. *Statistical inference: the minimum distance approach*. CRC press, 2011.
3. G Kemble Bennett and HF Martz. A continuous empirical Bayes smoothing technique. *Biometrika*, 59(2):361–368, 1972.

4. Rudolf Beran. Minimum Hellinger distance estimates for parametric models. *The annals of Statistics*, pages 445–463, 1977.
5. Joseph Berkson. Maximum likelihood and minimum χ^2 estimates of the logistic function. *Journal of the American statistical association*, 50(269):130–162, 1955.
6. E Bolthausen. Convergence in distribution of minimum-distance estimators. *Metrika*, 24(1):215–227, 1977.
7. Lawrence D Brown. In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *The Annals of Applied Statistics*, 2(1):113–152, 2008.
8. 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.
9. 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.
10. LD Brown and RH Farrell. Complete class theorems for estimation of multivariate poisson means and related problems. *The Annals of Statistics*, pages 706–726, 1985.
11. George Casella. An introduction to empirical Bayes data analysis. *The American Statistician*, 39(2):83–87, 1985.
12. Jiahua Chen. Optimal rate of convergence for finite mixture models. *The Annals of Statistics*, 23:221–233, 1995.
13. Jiahua Chen. Consistency of the MLE under mixture models. *Statistical Science*, 32(1):47–63, 2017.
14. I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.
15. JJ Deely and RL Kruse. Construction of sequences estimating the mixing distribution. *The Annals of Mathematical Statistics*, 39(1):286–288, 1968.
16. Alex Dytso, H Vincent Poor, Ronit Bustin, and Shlomo Shamai. On the structure of the least favorable prior distributions. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1081–1085. IEEE, 2018.
17. David Edelman. Estimation of the mixing distribution for a normal mean with applications to the compound decision problem. *The Annals of Statistics*, 16(4):1609–1622, 1988.
18. Bradley Efron. Two modeling strategies for empirical Bayes estimation. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(2):285, 2014.
19. Bradley Efron. Empirical Bayes: Concepts and Methods. 2021. http://statweb.stanford.edu/~ckirby/brad/papers/2021EB-concepts-methods.pdf.
20. Bradley Efron and Trevor Hastie. *Computer Age Statistical Inference, Student Edition: Algorithms, Evidence, and Data Science*, volume 6. Cambridge University Press, 2021.
21. Bradley Efron, Robert Tibshirani, John D Storey, and Virginia Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American statistical association*, 96(456):1151–1160, 2001.
22. Valerii Vadimovich Fedorov. *Theory of optimal experiments*. Elsevier, 1972.
23. S. Ghosal and A.W. van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics*, 29(5):1233–1263, 2001.
24. James Heckman and Burton Singer. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica: Journal of the Econometric Society*, pages 271–320, 1984.
25. Philippe Heinrich and Jonas Kahn. Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics*, 46(6A):2844–2870, 2018.
26. JC van Houwelingen and Th Stijnen. Monotone empirical Bayes estimators for the continuous one-parameter exponential family. *Statistica Neerlandica*, 37(1):29–43, 1983.
27. Soham Jana and Ayanendranath Basu. A characterization of all single-integral, non-kernel divergence estimators. *IEEE Transactions on Information Theory*, 65(12):7976–7984, 2019.

28. Soham Jana, Yury Polyanskiy, Anzo Z Teh, and Yihong Wu. Empirical bayes via erm and rademacher complexities: the poisson model. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5199–5235. PMLR, 2023.
29. Nicholas P Jewell. Mixtures of exponential distributions. *The annals of statistics*, pages 479–484, 1982.
30. Wenhua Jiang and Cun-Hui Zhang. General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647–1684, 2009.
31. Iain Johnstone. Admissible estimation, dirichlet principles and recurrence of birth-death chains on $\mathbb{Z}_+^p$. *Probability theory and related fields*, 71:231–269, 1986.
32. 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.
33. Arlene KH Kim. Minimax bounds for estimation of Normal mixtures. *bernoulli*, 20(4):1802–1818, 2014.
34. Roger Koenker and Jiaying Gu. Rebayes: an r package for empirical bayes mixture methods. *Journal of Statistical Software*, 82:1–26, 2017.
35. 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.
36. Roger Koenker and Ivan Mizera. Shape constrained density estimation via penalized rényi divergence. *Statistical Science*, 33(4):510–526, 2018.
37. Nan M Laird. Empirical Bayes estimates using the nonparametric maximum likelihood estimate for the prior. *Journal of Statistical Computation and Simulation*, 15(2-3):211–220, 1982.
38. Diane Lambert and Luke Tierney. Asymptotic properties of maximum likelihood estimates in the mixed Poisson model. *The Annals of Statistics*, pages 1388–1399, 1984.
39. Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.
40. Glen H Lemon and Richard G Krutchkoff. An empirical Bayes smoothing technique. *Biometrika*, 56(2):361–365, 1969.
41. Jianjun Li, Shanti S Gupta, and Friedrich Liese. Convergence rates of empirical Bayes estimation in exponential family. *Journal of statistical planning and inference*, 131(1):101–115, 2005.
42. Bruce G Lindsay. The geometry of mixture likelihoods: a general theory. *The annals of statistics*, pages 86–94, 1983.
43. Bruce G Lindsay. The geometry of mixture likelihoods, part II: the Exponential family. *The Annals of Statistics*, 11(3):783–792, 1983.
44. Bruce G Lindsay. Efficiency versus robustness: the case for minimum hellinger distance and related methods. *The annals of statistics*, 22(2):1081–1114, 1994.
45. Bruce G Lindsay. Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–163. JSTOR, 1995.
46. Johannes S Maritz and T Lwin. *Empirical Bayes methods*. Chapman and Hall/CRC, 2018.
47. JS Maritz. Smooth empirical Bayes estimation for one-parameter discrete distributions. *Biometrika*, 53(3-4):417–429, 1966.
48. JS Maritz. On the smooth empirical Bayes approach to testing of hypotheses and the compound decision problem. *Biometrika*, 55(1):83–100, 1968.
49. JS Maritz. Empirical bayes estimation for the Poisson distribution. *Biometrika*, 56(2):349–359, 1969.
50. Zhen Miao, Weihao Kong, Ramya Korlakai Vinayak, Wei Sun, and Fang Han. Fisher-pitman permutation tests based on nonparametric Poisson mixtures with application to single cell genomics. *arXiv preprint arXiv:2106.03022*, 2021.
51. PW Millar. A general approach to the optimality of minimum distance estimators. *Transactions of the American Mathematical Society*, 286(1):377–418, 1984.
52. Carl N Morris. Parametric empirical Bayes inference: theory and applications. *Journal of the American statistical Association*, 78(381):47–55, 1983.

53. Bhagwant Persaud, Bo Lan, Craig Lyon, and Ravi Bhim. Comparison of empirical Bayes and full Bayes approaches for before–after road safety evaluations. *Accident Analysis & Prevention*, 42(1):38–43, 2010.
54. Johann Pfanzagl. Consistency of maximum likelihood estimators for certain nonparametric families, in particular: mixtures. *Journal of Statistical Planning and Inference*, 19(2):137–158, 1988.
55. David Pollard. The minimum distance method of testing. *Metrika*, 27(1):43–70, 1980.
56. G Pólya and G Szegö. *Problems and Theorems in Analysis II,(reprint ed.)*. Springer, Heidelburg, 1998.
57. Yury Polyanskiy and Yihong Wu. Sharp regret bounds for empirical Bayes and compound decision problems. *arXiv preprint arXiv:2109.03943*, 2021.
58. Souvik Ray, Subrata Pal, Sumit Kumar Kar, and Ayanendranath Basu. Characterizing the functional density power divergence class. *IEEE Transactions on Information Theory*, 69(2):1141–1146, 2022.
59. Herbert Robbins. Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, pages 131–149. University of California Press, 1951.
60. Herbert Robbins. An Empirical Bayes Approach to Statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1956.
61. Leopold Simar. Maximum likelihood estimation of a compound Poisson process. *The Annals of Statistics*, 4(6):1200–1209, 1976.
62. Ananda Theertha Suresh, Alon Orlitsky, Jayadev Acharya, and Ashkan Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. *Advances in Neural Information Processing Systems*, 27, 2014.
63. Jay M Ver Hoef. Parametric empirical Bayes methods for ecological applications. *Ecological Applications*, 6(4):1047–1055, 1996.
64. J Wolfowitz. Estimation by the minimum distance method. *Annals of the Institute of Statistical Mathematics*, 5(1):9–23, 1953.
65. J Wolfowitz. Estimation by the minimum distance method in nonparametric stochastic difference equations. *The Annals of Mathematical Statistics*, 25(2):203–217, 1954.
66. Jacob Wolfowitz. The minimum distance method. *The Annals of Mathematical Statistics*, pages 75–88, 1957.
67. Henry P Wynn. The sequential generation of d-optimum experimental designs. *The Annals of Mathematical Statistics*, 41(5):1655–1664, 1970.
68. Anru R Zhang and Yuchen Zhou. On the non-asymptotic and sharp lower tail bounds of random variables. *Stat*, 9(1):e314, 2020.
69. Cun-Hui Zhang. Compound decision theory and empirical Bayes methods. *Annals of Statistics*, pages 379–390, 2003.
70. Cun-Hui Zhang. Generalized maximum likelihood estimation of normal mixture densities. *Statistica Sinica*, pages 1297–1318, 2009.

Appendix: Optimal empirical Bayes estimation for the Poisson model via minimum-distance methods

Soham Jana, Yury Polyanskiy and Yihong Wu

A. Least favorable prior and minimax estimator in the Poisson setup

Theorem 7 *Let $h > 0$ be finite and $\mathscr{P}([0,h])$ denote the set of all probability distribution supported on $[0,h]$. Consider the minimax objective*

$$\inf_{\widehat{\theta}} \sup_{\theta\in[0,h]} \mathbb{E}\left[(\theta - \widehat{\theta}(Y))^2\right],$$

where $Y|\theta \sim \mathrm{Poi}(\theta)$. Then the least favorable prior is a discrete distribution. In addition, the minimax estimator for the objective

$$\inf_{\widehat{\theta}_1,\ldots,\widehat{\theta}_n} \sup_{\theta_1,\ldots,\theta_n\in[0,h]} \mathbb{E}\left[\sum_{i=1}^{n}(\theta_i - \widehat{\theta}_i(Y^n))^2\right], \quad Y^n = \{Y_i\}_{i=1}^n, Y_i|\theta_i \overset{ind.}{\sim} \mathrm{Poi}(\theta_i), i \in [n],$$

is the conditional mean of the above least favorable distribution given $\{Y_1,\ldots,Y_n\}$.

Proof We will use results from [16] with the notations

$$\theta = X, \quad P_{Y|\theta} = \mathrm{Poi}(\theta).$$

The discreteness of the least favorable prior follows from [16, Proposition 6]. Let $\widehat{G}$ be the least favorable prior and $\widehat{\theta}_{\widehat{G}}(Y)$ be the conditional mean of the least favorable prior evaluated at Y. We show that $\left\{\widehat{\theta}_{\widehat{G}}(Y_i)\right\}_{i=1}^n$ is the minimax estimator. Note that in view of [16, Theorem 6] we get

$$\mathbb{E}_G\left[\left(\theta - \widehat{\theta}_{\widehat{G}}(Y)\right)^2\right] \le \mathbb{E}_{\widehat{G}}\left[\left(\theta - \widehat{\theta}_{\widehat{G}}(Y)\right)^2\right], \quad G \in \mathscr{P}([0,h]), Y|\theta \sim \mathrm{Poi}(\theta).$$

This implies

$$\sup_{\theta_1,\ldots,\theta_n} \sum_{i=1}^{n}\left[\left(\theta_i - \widehat{\theta}_{\widehat{G}}(Y_i)\right)^2\right] \le \mathbb{E}_{\theta_i\sim\widehat{G}, i\in[n]} \sum_{i=1}^{n}\left[\left(\theta_i - \widehat{\theta}_{\widehat{G}}(Y_i)\right)^2\right], \quad Y_i|\theta_i \sim \mathrm{Poi}(\theta_i), i \in [n].$$

In view of the standard inequality that supremum over $\theta_1,\ldots\theta_n$ is greater than the expected value with respect to $\theta_i \sim \widehat{G}, i \in [n]$, we get

$$\sup_{\theta_1,\ldots,\theta_n} \sum_{i=1}^{n}\left[\left(\theta_i - \widehat{\theta}_{\widehat{G}}(Y_i)\right)^2\right] = \mathbb{E}_{\theta_i\sim\widehat{G}, i\in[n]} \sum_{i=1}^{n}\left[\left(\theta_i - \widehat{\theta}_{\widehat{G}}(Y_i)\right)^2\right], \quad Y_i|\theta_i \sim \mathrm{Poi}(\theta_i), i \in [n].$$

As $\left\{\widehat{\theta}_{\widehat{G}}(Y_1),\ldots,\widehat{\theta}_{\widehat{G}}(Y_n)\right\}$ is the Bayes estimator with respect to the prior $\theta_i \overset{\text{ind.}}{\sim} \widehat{G}, i \in n$, given the data generating model $Y_i|\theta \sim \mathrm{Poi}(\theta_i), i \in [n]$, we use [39, Section 5.1, Theorem 1.4] to conclude that $\left\{\widehat{\theta}_{\widehat{G}}(Y_1),\ldots,\widehat{\theta}_{\widehat{G}}(Y_n)\right\}$ is a minimax estimator. □

B. Proof of Theorem 1

We first prove the result for the constrained solution $\operatorname{argmin}_{Q\in\mathscr{P}([0,h])}\mathsf{dist}(p\|f_Q)$. As mentioned towards the end of the proof, this also implies the desired result for the unconstrained solution. Suppose that p is supported on $\{y_1,\dots,y_m\}\subset\mathbb{Z}_+$. Define

$$S \triangleq \{(f_Q(y_1),\dots,f_Q(y_m)) : Q\in\mathscr{P}([0,h])\}, \tag{B.1}$$

where $f_Q(y)=\mathbb{E}_{\theta\sim Q}[f_\theta(y)]$ is the probability mass function of the Poisson mixture (1.2), and $f_\theta(y)=e^{-\theta}\theta^y/y!$. We claim that S is convex and compact.[3] The convexity follows from definition. For compactness, note that S is bounded since $\sup_{\theta\geq 0}f_\theta(y)=e^{-y}y^y/y!$, so it suffices to check S is closed. Let $(f'_1,\dots,f'_m)\in\mathbb{R}_+^m$ be the limiting point of $(f_{Q_k}(y_1),\dots,f_{Q_k}(y_m))$ for some sequence $\{Q_k\}$ in $\mathscr{P}([0,h])$. By Prokhorov's theorem, there is a subsequence $\{Q_{k_\ell}\}$ that converges weakly to some $Q'\in\mathscr{P}([0,h])$. Since $\theta\mapsto f_\theta(y)$ is continuous and bounded, we have $f'_j=f_{Q'}(y_j)$ for all j. In other words, S is closed.

Next, define $v:S\to\mathbb{R}$ by $v(f_1,\dots,f_m)=\sum_{i=1}^m\ell(p(y_i),f_i)$. By Assumption 1, the value of the min-distance optimization can be written as

$$\min_{Q\in\mathscr{P}([0,h])}\mathsf{dist}(p\|f_Q)=t(p)+\min_{(f_1,\dots,f_m)\in S}v(f_1,\dots,f_m). \tag{B.2}$$

Furthermore, by assumption $\ell(0,b)\equiv 0$ and $b\mapsto\ell(a,b)$ is strictly convex for $a>0$. Thus v is strictly convex. Therefore, there exists a unique point $(f_1^*,\dots,f_m^*)\in S$ that achieves the minimum on the right side of (B.2). Thus, the left side has a minimizer $\widehat{G}\in\mathscr{P}([0,h])$ that satisfies

$$f_{\widehat{G}}(y_j)=f_j^*, j=1,\dots,m \tag{B.3}$$

It remains to show that the above representation is unique at the special point $(f_1^*,\dots,f_m^*)$; this argument relies on the specific form of the Poisson density. Let $\widehat{G}$ be one such minimizer. By the first-order optimality condition (see (5.1) in Section 5.1),

$$\begin{aligned} &D_{\widehat{G}}(\theta)=\sum_{i=1}^m a_i(f_\theta(y_i)-f_i^*)\geq 0, \quad \forall 0\leq\theta\leq h;\\ &D_{\widehat{G}}(\theta)=0, \quad \text{for } \widehat{G}\text{-almost every } \theta, \end{aligned} \tag{B.4}$$

where $a_i\triangleq\frac{d}{df}\ell(p(y_i),f)|_{f=f_i^*}<0$, since ℓ is strictly decreasing in the second coordinate and $f_i^*>0$. Define

$$b_i=\frac{a_i}{\sum_{i=1}^m a_i f_i^*}>0.$$

[3] In this case, S is in fact the closed convex hull of the set $\{(f_\theta(y_1),\dots,f_\theta(y_m)):\theta\in[0,h]\}$.

As ℓ is strictly decreasing in second coordinate, $\frac{d}{df}\ell(p(y_i),f) < 0$ for all $f \in \mathbb{R}_+, i = 1,\ldots,m$. Using this, we rearrange (B.4) to get

$$\sum_{i=1}^{m} \frac{b_i}{y_i!}\theta^{y_i} \le e^{\theta}, \forall \theta \in [0,h],$$
$$\sum_{i=1}^{m} \frac{b_i}{y_i!}\theta^{y_i} = e^{\theta} \text{ for each } \theta \text{ in the support of } \widehat{G}. \tag{B.5}$$

Then the following lemma shows that the support of $\widehat{G}$ has at most m points.

Lemma 8 *Suppose that $\sum_{i=1}^{m} \beta_i \theta^{y_i} \le e^{\theta}$ for all $\theta \in [0,h]$ where $\beta_i \in \mathbb{R}$ and $h > 0$. Then the number of solutions to $\sum_{i=1}^{m} \beta_i \theta^{y_i} = e^{\theta}$ in $\theta \in [0,h]$ is at most m.*

Proof The proof is a modification of [61, Lemma 3.1(2)], which deals with the specific case $h = \infty$. Recall the following version of Descartes' rule of signs [56, Part V, Problem 38 and 40]: Consider an entire function (i.e., a power series whose radius of convergence is infinity) $\phi(x) = a_0 + a_1 x + a_2 x^2 + \ldots$ with real coefficients. Let r be the number of strictly positive zeros of ϕ counted with their multiplicities and let s be the number of sign changes[4] in the sequence $a_0, a_1, \ldots$. Then $r \le s$. We apply this fact to the function

$$\phi(\theta) = \sum_{i=1}^{m} \beta_i \theta^{y_i} - e^{\theta} = \sum_{j=0}^{\infty} a_j \theta^j,$$

where

$$a_j = \begin{cases} \beta_i - \frac{1}{y_i!} & j = y_i, i = 1,\ldots,m \\ -\frac{1}{j!} < 0 & \text{else} \end{cases}$$

Case 1: Suppose that 0 is a root of $\phi(\cdot)$. Then $a_0 = 0$. As there are at most $m-1$ positive coefficients in $a_0, a_1, \ldots$, there can be at most $2(m-1)$ sign changes, which implies at most $2(m-1)$ positive roots of s counting multiplicities. Note that, as $\phi(\theta)\mathbf{1}_{\{\theta \in (0,h)\}} \le 0$ and s is an entire function, each root of s inside $(0,h)$ has multiplicity at least 2. Suppose that m_h is the multiplicity of h as a root of $\phi(\cdot)$, which we define to be 0 when h is not a root. This means that the total number of distinct roots in $(0,h)$ is at most the largest integer before $(2(m-1)-m_h)/2$. If h is not a root, then the number of distinct roots in $(0,h)$ is at most $m-1$. If h is a root, then its multiplicity is at least 1, and hence, the number of distinct roots in $(0,h)$ is at most $m-2$. Hence, there are at most m many distinct roots in $[0,h]$.

Case 2: Suppose that 0 is not a root of $\phi(\cdot)$. As there are at most m positive coefficients in $a_0, a_1, \ldots$, there can be at most $2m$ sign changes, which implies at most $2m$ positive roots counting multiplicities. By a similar argument as in the previous case, the total number of distinct roots in $(0,h)$ is at most the largest integer before $(2m-m_h)/2$. If h is not a root, then the number of

[4] The number of sign changes is the number of pairs $0 \le i < j$ such that $a_i a_j < 0$ and either $j = i+1$ or $a_k = 0$ for all $i < k < j$.

distinct roots in $(0,h)$ is at most m. If h is a root, then the number of distinct roots in $(0,h)$ is at most $m-1$. Hence, in total, there are at most m distinct roots in $[0,h]$. $\square$

Suppose that there are $r(\leq m)$ different θ_i's (denote them by $\theta_1 \ldots, \theta_r$) for which (B.5) holds. This implies given any optimizer $\widehat{G}$ its atoms form a subset of $\{\theta_1 \ldots, \theta_r\}$. Let w_j be the weight $\widehat{G}$ puts on θ_j. Then in view of (B.3) we get that

$$\sum_{j=1}^{r} w_j e^{-\theta_j} \theta_j^{y_i} = f_i^* y_i!, \quad i = 1, \ldots, r.$$

The matrix $\{\theta_j^{y_i} : j = 1, \ldots, r, i = 1, \ldots, m\}$ has full column rank, and hence the vector $(w_1, \ldots, w_r)$ can be solve uniquely. This also implies the uniqueness of the optimizer $\widehat{G}$. This finishes the proof for the constrained solution.

Next we argue for the unconstrained minimizer $\operatorname{argmin}_Q \mathsf{dist}(p\|f_Q)$. In view of Lemma 9 below, we get that the unconstrained minimum-distance estimator is supported on $[0,h]$ with $h = \max_{i=1,\ldots,m} y_i$. Then, from the above proof for $\operatorname{argmin}_{Q \in \mathscr{P}([0,h])} \mathsf{dist}(p\|f_Q)$, the existence and uniqueness of the unconstrained estimator follow.

Lemma 9 *Let* dist *satisfy Assumption 1 and let p be a probability distribution on* $\mathbb{Z}_+$ *with support* $\{y_1, \ldots, y_m\}$. *Then the minimizer* $\operatorname{argmin}_Q \mathsf{dist}(p\|f_Q)$ *is supported on the interval* $[y_{\min}, y_{\max}]$, *where* $y_{\min} = \min_{i=1,\ldots,m} y_i, y_{\max} = \max_{i=1,\ldots,m} y_i$.

Proof Let Q be a distribution with $Q([0,y_{\min})) + Q((y_{\max},\infty)) > 0$. Define another distribution $\widetilde{Q}$ by

$$\widetilde{Q}(\cdot) = Q([0,y_{\min}))\delta_{y_{\min}}(\cdot) + Q(\cdot \cap [y_{\min}, y_{\max}]) + Q((y_{\max},\infty))\delta_{y_{\max}}(\cdot).$$

In other words, $\widetilde{Q}$ moves the masses of Q on the intervals $[0,y_{\min})$ (resp. $(y_{\max},\infty)$) to the point $y_{\min}$ (resp. $y_{\max}$). As $f_\theta(y)$ is strictly increasing in $\theta \in [0,y)$ and strictly decreasing in $\theta \in (y,\infty)$ we get for each $i = 1, \ldots, n$

$$\begin{aligned}
f_Q(y_i) &= \int f_\theta(y_i) dQ(\theta) \\
&= \int_{0 \leq \theta < y_{\min}} f_\theta(y_i) dQ(\theta) + \int_{y_{\min} \leq \theta \leq y_{\max}} f_\theta(y_i) dQ(\theta) + \int_{y_{\max} > \theta} f_\theta(y_i) dQ(\theta) \\
&< Q([0,y_{\min}))f_{y_{\min}}(y_i) + \int_{y_{\min} \leq \theta \leq y_{\max}} f_\theta(y_i) dQ(\theta) + Q((y_{\max},\infty))f_{y_{\max}}(y_i) \\
&= \int f_\theta(y_i) d\widetilde{Q}(\theta) = f_{\widetilde{Q}}(y_i).
\end{aligned}$$

Hence, by Assumption 1, we get

$$\begin{aligned}
\mathsf{dist}(p\|f_Q) &= t(p) + \sum_{y \geq 0} \ell(p(y), f_Q(y)) \\
&\overset{(a)}{=} t(p) + \sum_{y: p(y)>0} \ell(p(y), f_Q(y)) \overset{(b)}{>} t(p) + \sum_{y: p(y)>0} \ell\left(p(y), f_{\widetilde{Q}}(y)\right) = \mathsf{dist}(p\|f_{\widetilde{Q}}),
\end{aligned} \tag{B.6}$$

where (a) follows from $\ell(0,\cdot)=0$; (b) follows as the function $b\mapsto\ell(a,b)$ is strictly decreasing. In other words, given any Q with $Q([0,y_{\min}))+Q((y_{\max},\infty))>0$ we can produce $\widetilde{Q}$ supported on $[y_{\min},y_{\max}]$ such that $\mathsf{dist}(p\|f_{\widetilde{Q}})<\mathsf{dist}(p\|f_Q)$. Hence, the claim follows. □

C. Proof of Lemma 4

Let $\theta\sim G, Y|\theta\sim f_\theta$. Then for any $\widehat{G}$ independent of Y, we can write $\mathsf{Regret}(\widehat{G};G)=\sum_{y=0}^{\infty}(\widehat{\theta}_{\widehat{G}}(y)-\widehat{\theta}_G(y))^2 f_G(y)=\mathbb{E}_G\left[\left(\widehat{\theta}_{\widehat{G}}(Y)-\widehat{\theta}_G(Y)\right)^2\right]$; cf. (2.13). Fix $h>0$ and note the following

- $\mathrm{mmse}(G)=\mathbb{E}[(\widehat{\theta}_G-\theta)^2]\geq\mathbb{P}[\theta\in[0,h]]\ \mathbb{E}_{\theta\sim G}[(\widehat{\theta}_G-\theta)^2|\theta\in[0,h]]\geq\mathbb{P}[\theta\in[0,h]]\,\mathrm{mmse}(G_h)$
- $\mathrm{mmse}(G)\leq\sqrt{\mathbb{E}_G[\theta^4]}\leq\sqrt{M}$, and
- For any fixed distribution $\widehat{G}$

$$\begin{aligned}
\mathbb{E}_G\left[(\widehat{\theta}_{\widehat{G}}(Y)-\theta)^2\right]&\leq\mathbb{E}_G\left[(\widehat{\theta}_{\widehat{G}}(Y)-\theta)^2\mathbf{1}_{\{\theta\leq h\}}\right]+\mathbb{E}_G\left[(\widehat{\theta}_{\widehat{G}}(Y)-\theta)^2\mathbf{1}_{\{\theta>h\}}\right]\\
&\overset{(a)}{\leq}\mathbb{E}_G\left[(\widehat{\theta}_{\widehat{G}}(Y)-\theta)^2\middle|\,\theta\leq h\right]+\sqrt{\mathbb{E}_G\left[(\widehat{\theta}_{\widehat{G}}(Y)-\theta)^4\right]\mathbb{E}_G\left[\mathbf{1}_{\{\theta>h\}}\right]}\\
&\overset{(b)}{\leq}\mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y)-\theta)^2\right]+\sqrt{8(\widehat{h}^4+\mathbb{E}_G[\theta^4])G((h,\infty))}\\
&=\mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y)-\theta)^2\right]+\sqrt{8(\widehat{h}^4+M)G((h,\infty))}.
\end{aligned}\tag{C.1}$$

where step (a) followed by Cauchy-Schwarz inequality and step (b) followed as $(x+y)^4\leq 8(x^4+y^4)$ for any $x,y\in\mathbb{R}$.

Using these we get

$$\begin{aligned}
\mathsf{Regret}(\widehat{G};G)&=\mathbb{E}_G\left[(\widehat{\theta}_{\widehat{G}}(Y)-\theta)^2\right]-\mathrm{mmse}(G)\\
&\leq\mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y)-\theta)^2\right]-\mathrm{mmse}(G_h)+\mathrm{mmse}(G_h)-\mathrm{mmse}(G)+\sqrt{8(\widehat{h}^4+M)G((h,\infty))}\\
&\leq\mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y)-\widehat{\theta}_{G_h}(Y))^2\right]+\left(\frac{1}{G([0,h])}-1\right)\mathrm{mmse}(G)+\sqrt{8(\widehat{h}^4+M)G((h,\infty))}\\
&\leq\mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y)-\widehat{\theta}_{G_h}(Y))^2\right]+\frac{G((h,\infty))}{G([0,h])}\sqrt{M}+\sqrt{8(\widehat{h}^4+M)G((h,\infty))}\\
&\leq\mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y)-\widehat{\theta}_{G_h}(Y))^2\right]+\frac{(1+2\sqrt{2})\sqrt{(\widehat{h}^4+M)G((h,\infty))}}{G([0,h])}.
\end{aligned}\tag{C.2}$$

Next we bound the first term. Fix $K \geq 1$. Using $\widehat{\theta}_{G_h}(y) \leq h, \widehat{\theta}_{\widehat{G}}(y) \leq \widehat{h}$ we have

$$
\begin{aligned}
&\mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y) - \widehat{\theta}_{G_h}(Y))^2 \mathbf{1}_{\{Y \leq K-1\}}\right] \\
&= \sum_{y=0}^{K-1} (y+1)^2 f_{G_h}(y) \left(\frac{f_{\widehat{G}}(y+1)}{f_{\widehat{G}}(y)} - \frac{f_{G_h}(y+1)}{f_{G_h}(y)}\right)^2 \\
&\overset{(a)}{\leq} \sum_{y=0}^{K-1} (y+1)^2 f_{G_h}(y) \left\{ 3\left(\frac{f_{\widehat{G}}(y+1)}{f_{\widehat{G}}(y)} - \frac{2f_{\widehat{G}}(y+1)}{f_{G_h}(y)+f_{\widehat{G}}(y)}\right)^2 + 3\left(\frac{f_{G_h}(y+1)}{f_{G_h}(y)} - \frac{2f_{G_h}(y+1)}{f_{G_h}(y)+f_{\widehat{G}}(y)}\right)^2 \right. \\
&\quad \left. + 3\left(\frac{2f_{G_h}(y+1) - 2f_{\widehat{G}}(y+1)}{f_{G_h}(y)+f_{\widehat{G}}(y)}\right)^2 \right\} \\
&\leq 3\sum_{y=0}^{K-1} \left\{ \left(\frac{(y+1)f_{\widehat{G}}(y+1)}{f_{\widehat{G}}(y)}\right)^2 \frac{(f_{G_h}(y) - f_{\widehat{G}}(y))^2}{f_{G_h}(y)+f_{\widehat{G}}(y)} + \left(\frac{(y+1)f_{G_h}(y+1)}{f_{G_h}(y)}\right)^2 \frac{(f_{G_h}(y) - f_{\widehat{G}}(y))^2}{f_{G_h}(y)+f_{\widehat{G}}(y)} \right. \\
&\quad \left. + 4(y+1)^2 \frac{(f_{G_h}(y+1) - f_{\widehat{G}}(y+1))^2}{f_{G_h}(y)+f_{\widehat{G}}(y)} \right\} \\
&= 3(\{\widehat{\theta}_{G_h}(y)\}^2 + \{\widehat{\theta}_{\widehat{G}}(y)\}^2) \sum_{y=0}^{K-1} \frac{(f_{G_h}(y) - f_{\widehat{G}}(y))^2}{f_{G_h}(y)+f_{\widehat{G}}(y)} + 12\sum_{y=0}^{K-1} (y+1)^2 \frac{(f_{G_h}(y+1) - f_{\widehat{G}}(y+1))^2}{f_{G_h}(y)+f_{\widehat{G}}(y)} \\
&\leq 3(h^2 + \widehat{h}^2) \sum_{y=0}^{K-1} \frac{(f_{G_h}(y) - f_{\widehat{G}}(y))^2}{f_{G_h}(y)+f_{\widehat{G}}(y)} + 12\sum_{y=0}^{K-1} (y+1)^2 \frac{(f_{G_h}(y+1) - f_{\widehat{G}}(y+1))^2}{f_{G_h}(y)+f_{\widehat{G}}(y)}
\end{aligned}
$$

where (a) followed from $(x+y+z)^2 \leq 3(x^2+y^2+z^2)$ for any $x,y,z \in \mathbb{R}$. Using $(\sqrt{f_{G_h}(x)} + \sqrt{f_{\widehat{G}}(x)})^2 \leq 2(f_{G_h}(x) + f_{\widehat{G}}(x))$ for $x = y, y+1$ we continue the last display to get

$$
\begin{aligned}
&\mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y) - \widehat{\theta}_{G_h}(Y))^2 \mathbf{1}_{\{Y \leq K-1\}}\right] \\
&\leq 6(h^2 + \widehat{h}^2) \sum_{y=0}^{K-1} (\sqrt{f_{G_h}(y)} - \sqrt{f_{\widehat{G}}(y)})^2 \\
&\quad + 24K \max_{y=0}^{K-1} \frac{(y+1)f_{G_h}(y+1) + (y+1)f_{\widehat{G}}(y+1)}{f_{G_h}(y)+f_{\widehat{G}}(y)} \sum_{y=0}^{K-1} (\sqrt{f_{G_h}(y+1)} - \sqrt{f_{\widehat{G}}(y+1)})^2 \\
&\leq \left(6(h^2+\widehat{h}^2) + 24(h+\widehat{h})K\right) H^2(f_{\widehat{G}}, f_{G_h}).
\end{aligned}
$$

Again using $\widehat{\theta}_{G_h}(y) \leq h, \widehat{\theta}_{\widehat{G}}(y) \leq \widehat{h}$ we bound $\mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y) - \widehat{\theta}_{G_h}(Y))^2 \mathbf{1}_{\{Y \geq K\}}\right]$ by $(h+\widehat{h})^2 \varepsilon_K(G_h)$. Combining this with the last display we get

$$
\mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y) - \widehat{\theta}_{G_h}(Y))^2\right] \leq \left\{6(h^2+\widehat{h}^2) + 24(h+\widehat{h})K\right\} H^2(f_{\widehat{G}}, f_{G_h}) + (h+\widehat{h})^2 \varepsilon_K(G_h).
$$

In view of above continuing (C.2) we have

$$\begin{aligned}\mathsf{Regret}(\widehat{G};G) \leq &\left\{6(h^2+\widehat{h}^2)+24(h+\widehat{h})K\right\}H^2(f_{\widehat{G}},f_{G_h}) \\ &+(h+\widehat{h})^2\varepsilon_K(G_h)+\frac{(1+2\sqrt{2})\sqrt{(M+\widehat{h}^4)G((h,\infty))}}{G([0,h])}. \end{aligned} \tag{C.3}$$

Using triangle inequality and $(x+y)^2 \leq 2(x^2+y^2)$ we get

$$H^2(f_{\widehat{G}},f_{G_h}) \leq 2\left\{H^2(f_G,f_{\widehat{G}})+H^2(f_{G_h},f_G)\right\}. \tag{C.4}$$

Note that

$$H^2(f_{G_h},f_G) \leq 2\mathrm{TV}(f_{G_h},f_G) \leq 2\mathrm{TV}(G_h,G) = 4G((h,\infty)).$$

where TV denotes the total variation and the middle inequality applies the data-processing inequality [14] and the last equality followed as

$$\begin{aligned}\mathrm{TV}(G_h,G) &= \int_0^h |dG_h(\theta)-dG(\theta)| + \int_h^\infty dG(\theta) \\ &= \left(\frac{1}{G([0,h])}-1\right)\int_0^h dG(\theta)+G((h,\infty)) = 2G((h,\infty)).\end{aligned}$$

Then, combining (C.3), (C.4) and using $\varepsilon_K(G_h) \leq \frac{\varepsilon_K(G)}{G([0,h])}$ we get the desired bound

$$\begin{aligned}\mathsf{Regret}(\widehat{G};G) \leq &\left\{12(h^2+\widehat{h}^2)+48(h+\widehat{h})K\right\}\left(H^2(f_{\widehat{G}},f_G)+4G((h,\infty))\right) \\ &+(h+\widehat{h})^2\frac{\varepsilon_K(G)}{G([0,h])}+\frac{(1+2\sqrt{2})\sqrt{(M+\widehat{h}^4)G((h,\infty))}}{G([0,h])}.\end{aligned}$$

D. Auxiliary results

Lemma 10 *Let $Y \sim \mathrm{Poi}(h)$. Then we have*

$$\mathbb{P}[Y > h+y] \leq e^{-\frac{y^2}{4h}}\ \forall\, 0<y<\frac{h}{2}, \quad \mathbb{P}[Y<h-y] \leq e^{-\frac{5y^2}{9h}}\ \forall\, 0<y<\frac{h}{3}$$

Proof The proof of the above result follows from [68, Section] with the inequality $\log(1+t) \geq (t-2t^2/3)\mathbf{1}_{\left\{t\in(-\frac{1}{3},0)\right\}}+(t-t^2/2)\mathbf{1}_{\left\{t\in(0,\frac{1}{2})\right\}}$. □

Lemma 11 *Let $h>0$ and $G \in \mathscr{P}([0,h])$. If $\{Y_i\}_{i=1}^n \overset{i.i.d.}{\sim} f_G$, then the following are satisfied given any $t \geq 0, a \geq 1, n \geq 3$ and $K = \min\left\{\frac{a(he^2+2)\log n}{\log\log n}, he^2+a\log n\right\}$.*

$$\mathbb{P}\left[Y_1 > K+t\right] \leq \frac{2e^{-t}}{n^a}, \quad \mathbb{E}[Y_{\max}^4] \leq 5K^4+8.$$

Proof Let $G \in \mathscr{P}([0,h])$. As $p(\theta) = e^{-\theta}\theta^y$ is increasing in $\theta \in [0,y], y > 0$, for $\ell > 2h$,

$$\mathbb{P}[Y_1 \geq \ell] = \sum_{y=\ell}^{\infty} \int_0^h \frac{e^{-\theta}\theta^y}{y!} G(d\theta) \leq \sum_{y=\ell}^{\infty} \frac{e^{-h}h^y}{y!} \leq \frac{h^\ell}{\ell!} \sum_{y-\ell=0}^{\infty} \left(\frac{h}{\ell}\right)^{y-\ell} \leq 2\left(\frac{he}{\ell}\right)^{\ell}. \tag{D.1}$$

where the last inequality followed using $\ell! \geq \left(\frac{\ell}{e}\right)^{\ell}$ from the Stirling's formula. To get to our results, first, let $K_1 = \frac{a(he^2+2)\log n}{\log\log n}, a \geq 1$. Using the fact $\log\log\log n < \frac{\log\log n}{2}$ for all $n \geq 3$ we continue the last display with $\ell = K_1 + t$ to get

$$\begin{aligned}
\mathbb{P}[Y_1 \geq K_1 + t] &\leq 2\left(\frac{he}{K_1+t}\right)^{K_1}\left(\frac{he}{K_1+t}\right)^{t} \leq 2\left(\frac{he}{K_1}\right)^{K_1} e^{-t} \\
&\leq 2e^{-t}\left(\frac{\log\log n}{\log n}\right)^{\frac{a(he^2+2)\log n}{\log\log n}} \\
&\leq 2e^{-t-(\log\log n - \log\log\log n)\frac{a(he^2+2)\log n}{\log\log n}} \leq 2e^{-t-a\log n} \leq \frac{2e^{-t}}{n^a},
\end{aligned} \tag{D.2}$$

as required. Next, considering $K_2 = he^2 + a\log n$ we continue (D.1) with $\ell = K_2 + t$ to get

$$\mathbb{P}[Y_1 \geq K_2 + t] \leq 2\left(\frac{he}{K_2}\right)^{K_2}\left(\frac{he}{K_2+t}\right)^{t} \leq 2\left(1 - \frac{a\log n}{K_2}\right)^{K_2} e^{-t} \leq 2e^{-a\log n}e^{-t} \leq \frac{2e^{-t}}{n^a}. \tag{D.3}$$

Choosing $K = \min\{K_1, K_2\}$ we get the desired result.

Next we bound $\mathbb{E}[Y_{\max}^4]$. For any nonnegative integer valued random variable Z, using

$$\mathbb{E}[Z^4] = \sum_{z\geq 1} z^4 \mathbb{P}[Z=z] \leq 4\sum_{z\geq 1}\sum_{k=1}^{z} k^3 \mathbb{P}[Z=z] = 4\sum_{k\geq 1} k^3 \sum_{z\geq k} \mathbb{P}[Z=z] = 4\sum_{k\geq 1} k^3 \mathbb{P}[Z\geq k] \tag{D.4}$$

we have

$$\begin{aligned}
\mathbb{E}\left[(Y_{\max})^4\right] &\leq 4\sum_{y} y^3 \mathbb{P}\left[Y_{\max} > y\right] \\
&\leq 4K^4 + n\sum_{y\geq K+1} y^3 \mathbb{P}\left[Y_1 > y\right] \\
&= 4K^4 + n\sum_{t\geq 1} (K+t)^3 \mathbb{P}[Y_1 > K+t] \\
&\overset{(a)}{\leq} 4K^4 + \frac{4}{n^4}\sum_{t\geq 1}(K^3+t^3)e^{-t} \leq 4K^4 + \frac{K^3}{4}\sum_{t\geq 1} e^{-t} + \sum_{t\geq 1} t^3 e^{-t} \leq 5K^4 + 8,
\end{aligned}$$

where (a) followed from $(x+y)^3 \leq 4(x^3+y^3), x, y \geq 0$. □

Lemma 12 *Given any $s > 0$ and $G \in \mathsf{SubE}(s)$, the following are satisfied.*

(i) If $\theta \sim G$, *then* $\mathbb{E}[\theta^4] \le 12s^4$.
(ii) If $\{Y_i\}_{i=1}^n \overset{i.i.d.}{\sim} f_G$, *then for any* $K \ge 1$

$$\mathbb{P}[Y_1 \ge K] \le 3e^{-K\log\left(1+\frac{1}{2s}\right)}, \quad \mathbb{E}\left[Y_{\max}^4\right] \le \frac{64(\log n)^4 + 90}{\left(\log\left(1+\frac{1}{2s}\right)\right)^4}. \tag{D.5}$$

Proof To prove (i) we note that for any $s, M > 0$ using integral by parts we have $\int_0^M x^3 e^{-\frac{x}{s}} dx = \left[-se^{-\frac{x}{s}}(6s^3 + 6s^2x + 3sx^2 + x^3)\right]_0^M$. Then we get using the definition of $\mathsf{SubE}(s)$ tail probabilities and with limit as $M \to \infty$

$$\mathbb{E}[\theta^4] = 4\int_0^\infty y^3 \mathbb{P}[\theta > y]\, dy \le 2\int_0^\infty y^3 e^{-\frac{y}{s}} dy = 2\lim_{M\to\infty}\left[-se^{-\frac{y}{s}}(6s^3 + 6s^2y + 3sy^2 + y^3)\right]_0^M \le 12s^4.$$

The proof of the property (ii) is as follows. Using $\mathbb{E}_{Z\sim\mathrm{Poi}(\theta)}\left[e^{Zt}\right] = e^{\theta(e^t-1)}, t > 0$ and denoting $c(s) = \log\frac{1+2s}{2s}$ we have

$$\begin{aligned}
&\mathbb{E}\left[e^{Y_1 c(s)}\right] = \mathbb{E}_{\theta\sim G}\left[\mathbb{E}_{Y_1\sim\mathrm{Poi}(\theta)}\left[e^{Y_1 c(s)}\middle|\,\theta\right]\right] = \mathbb{E}_G\left[e^{\frac{\theta}{2s}}\right] = \int_0^\infty e^{\theta/2s}\, G(d\theta) \\
&= 1 + \int_{\theta=0}^\infty \int_{x=0}^\theta \frac{e^{x/2s}}{2s} dx\, G(d\theta) = 1 + \int_{x>0} \frac{e^{x/2s}}{2s}\, G([x,\infty))dx \overset{(a)}{\le} 1 + \int_{x>0} \frac{e^{-x/2s}}{s} dx \le 3
\end{aligned}$$

where (a) followed by using tail bound for $\mathsf{SubE}(s)$ distribution G. In view of Markov inequality

$$\mathbb{P}[Y_1 \ge K] \le \mathbb{E}\left[e^{Y_1 c(s)}\right] e^{-c(s)K} \le 3e^{-K\log\left(1+\frac{1}{2s}\right)}.$$

The expectation term is bounded as below. Pick L large enough such that $v(y) = y^3 e^{-y\log\left(1+\frac{1}{2s}\right)}$ is decreasing for all $y \ge L$. Then we can bound $\sum_{y=L+1}^\infty v(y) \le \int_{y>L} v(y)dy$. Then, using the last inequality for such $L > 0$ and (D.4)

$$\begin{aligned}
\mathbb{E}\left[(Y_{\max})^4\right] &\le 4\sum_y y^3 \mathbb{P}[Y_{\max} > y] \\
&\le 4L^4 + n\sum_{y\ge L+1} y^3 \mathbb{P}[Y_1 > y] \\
&\le 4L^4 + 3n\int_{y\ge L} y^3 e^{-y\log\left(1+\frac{1}{2s}\right)} dy \\
&\overset{(a)}{\le} 4L^4 + \frac{3n}{\left\{\log\left(1+\frac{1}{2s}\right)\right\}^4} \int_{z>L\log\left(1+\frac{1}{2s}\right)} z^3 e^{-z} dz \\
&\overset{(b)}{\le} 4L^4 + \frac{45n}{\left\{\log\left(1+\frac{1}{2s}\right)\right\}^4} \int_{z>L\log\left(1+\frac{1}{2s}\right)} e^{-z/2} dz \le 4L^4 + \frac{90ne^{-\frac{L}{2}\log\left(1+\frac{1}{2s}\right)}}{\left\{\log\left(1+\frac{1}{2s}\right)\right\}^4},
\end{aligned}$$

where (a) followed from a change of variable, (b) followed using $x^3 \le 15e^{-\frac{x}{2}}$ for any $x > 0$. Choosing $L = \frac{2\log n}{\log\left(1+\frac{1}{2s}\right)}$ we get the desired result. □

E. Proofs of the multidimensional results

E.1. *Density estimation in multiple dimensions*

We will assume $d \le n$, as the results are vacuous otherwise. The proof of Theorem 5 is based on a similar truncation idea as in the proof of Theorem 2. For this section, we will use the notation

$$K = \begin{cases} \min\left\{\max\{1,h\}\frac{\log n}{\log\log n}, h+\log n\right\}, & G \in \mathscr{P}([0,h]^d, \\ \max\{1,s\}\log n, & \text{marginals of } G \text{ are in } \mathsf{SubE}(s), \end{cases} \tag{E.1}$$

unless we specify it differently. We first note the following result.

Lemma 13 *There exists absolute constant $\widetilde{c}_1$ such that the following holds. If $Y \sim f_G$ such that $G \in \mathscr{P}([0,h]^d)$ or all the marginals of G belong to $\mathsf{SubE}(s)$, then $\mathbb{P}\left[\boldsymbol{Y} \notin [0,\widetilde{c}_1 K]^d\right] \le \frac{d}{n^{10}}$.*

Proof From the proof of Lemma 11 and Lemma 12, we get that there exists a constant $\widetilde{c}_1$ such that with probability at least $1-\frac{1}{n^{10}}$ all the coordinates of the random variable $\boldsymbol{Y}$ lie within $[0,\widetilde{c}_1 K]$. Then using a union bound over all the coordinates we achieve the desired result. □

Proof of Theorem 5 Suppose that the dist function, for which we compute the minimum distance estimator, satisfies Assumption 4. Then, using a proof strategy identical to proving the result (3.3) in the one-dimensional case, we get

$$H^2(f_G, f_{\widehat{G}}) \le \frac{2}{c_1}(\mathsf{dist}(p_n^{\mathsf{emp}}\|f_{\widehat{G}}) + \mathsf{dist}(p_n^{\mathsf{emp}}\|f_G)) \le \frac{4}{c_1}\mathsf{dist}(p_n^{\mathsf{emp}}\|f_G).$$

In view of Assumption 4 we bound $\frac{1}{c_2}\mathsf{dist}$ by χ^2 and use $R = \widetilde{c}_1 K$ to get the following

$$\begin{aligned}
&\frac{1}{c_2}\mathbb{E}\left[\mathsf{dist}(p_n^{\mathsf{emp}}\|f_G)\mathbf{1}_{\{\boldsymbol{Y}_i\in[0,R]^d\ \forall i=1,\ldots,n\}}\right] \\
&\le \mathbb{E}\left[\chi^2(p_n^{\mathsf{emp}}\|f_G)\mathbf{1}_{\{\boldsymbol{Y}_i\in[0,R]^d\ \forall i=1,\ldots,n\}}\right] = \sum_{\boldsymbol{y}}\frac{\mathbb{E}\left[(p_n^{\mathsf{emp}}(\boldsymbol{y})-f_G(\boldsymbol{y}))^2\mathbf{1}_{\{\boldsymbol{Y}_i\in[0,R]^d\ \forall i=1,\ldots,n\}}\right]}{f_G(\boldsymbol{y})} \\
&\overset{(a)}{=} \sum_{\boldsymbol{y}\in[0,R]^d}\frac{\mathbb{E}\left[(p_n^{\mathsf{emp}}(\boldsymbol{y})-f_G(\boldsymbol{y}))^2\mathbf{1}_{\{\boldsymbol{Y}_i\in[0,R]^d\ \forall i=1,\ldots,n\}}\right]}{f_G(\boldsymbol{y})} + \sum_{\boldsymbol{y}\notin[0,R]^d} f_G(\boldsymbol{y})\mathbb{P}[\boldsymbol{Y}_i\in[0,R]^d\ \forall i=1,\ldots,n] \\
&\le \sum_{\boldsymbol{y}\in[0,R]^d}\frac{\mathbb{E}\left[(p_n^{\mathsf{emp}}(\boldsymbol{y})-f_G(\boldsymbol{y}))^2\right]}{f_G(\boldsymbol{y})} + \mathbb{P}_{\boldsymbol{Y}\sim f_G}[\boldsymbol{Y}\notin[0,R]^d] \\
&\overset{(b)}{\le} \frac{1}{n}\sum_{\boldsymbol{y}\in[0,R]^d}(1-f_G(\boldsymbol{y})) + \frac{d}{n^{10}} \le \frac{2(R+1)^d}{n}.
\end{aligned} \tag{E.2}$$

where (a) followed as $\{\boldsymbol{Y}_i \in [0,R]^d\ \forall i = 1,\ldots,n\}$ implies $p_n^{\mathsf{emp}}(\boldsymbol{y}) = 0$ for any $y \notin [0,R]^d$; and (b) follows from $\mathbb{E}[p_n^{\mathsf{emp}}(\boldsymbol{y})] = f_G(\boldsymbol{y})$ and, thus, $\mathbb{E}[(p_n^{\mathsf{emp}}(\boldsymbol{y})-f_G(\boldsymbol{y}))^2] = \mathrm{Var}(p_n^{\mathsf{emp}}(\boldsymbol{y})) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(\mathbf{1}_{\{\boldsymbol{Y}_i=\boldsymbol{y}\}}) = \frac{f_G(\boldsymbol{y})(1-f_G(\boldsymbol{y}))}{n}$, and due to the choice of R with Lemma 13.

Using the union bound and the fact $H^2 \leq 2$ we have

$$\mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\mathbf{1}_{\{\boldsymbol{Y}_i \notin [0,R]^d \text{ for some } i \in \{1,\ldots,n\}\}}\right] \leq \frac{2d}{n^9}.$$

Combining this with (E.2) yields

$$\begin{aligned}\mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\right] &\leq \mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\mathbf{1}_{\{\boldsymbol{Y}_i \in [0,R]^d \ \forall i=1,\ldots,n\}}\right] + \mathbb{E}\left[H^2(f_G, f_{\widehat{G}})\mathbf{1}_{\{\boldsymbol{Y}_i \notin [0,R]^d \text{ for some } i \in \{1,\ldots,n\}\}}\right] \\ &\leq \frac{4(R+1)^d}{n},\end{aligned}$$

which completes the proof. □

E.2. *Regret bounds in multiple dimensions*

Proof of Theorem 6 We first note that it suffices to only prove the case where the data generating G satisfies $G \in \mathscr{P}([0,h]^d)$. To prove the case where the marginals of G belong to $\mathsf{SubE}(s)$, it suffices to choose $h = \widetilde{c}\max\{1,s\}\log n$ for a large enough constant $\widetilde{c} > 0$, as the following argument shows. Using the property of the Poisson mixture and the result on the support of the unconstrained NPMLE $\widehat{G}$ for the one dimensional case in Lemma 9, as $f_{\boldsymbol{\theta}}(\boldsymbol{y}) = \prod_{j=1}^d f_{\theta_j}(y_j)$, we get

1. $\widehat{G}$ is supported on $[0, \max_{j=1}^d \max_{i=1}^n Y_{ij} + 1]^d$, which itself is a subset of $[0, \widetilde{c}\max\{1,s\}\log n]^d$ with probability at least $1 - \frac{d}{n^9}$ for a large enough constant $\widetilde{c} > 0$.
2. As a result of the above and Lemma 13, we get that with probability at least $1 - \frac{d}{n^9}$, each coordinate of $\widehat{\boldsymbol{\theta}}_{\widehat{G}}$ lies in the interval $[0, \widetilde{c}\max\{1,s\}\log n]$.

Hence, using arguments similar to the one dimensional case in (C.2) we can argue the following.

Lemma 14 *For any G with marginals in $\mathsf{SubE}(s)$ and an estimate $\widehat{G}$ supported on $[0,\widehat{h}]^d$,*

$$\mathsf{Regret}(\widehat{G}, G) \leq \mathsf{Regret}(\widehat{G}, G_h) + O\left(\frac{\widehat{h}^2 + s^2}{n^2}\right), \quad h = \widetilde{c}\max\{1,s\}\log n,$$

where G_h denote its restriction of G on the hypercube $[0,h]^d$, i.e., $G_h[\boldsymbol{\theta} \in \cdot] = G[\boldsymbol{\theta} \in \cdot | \boldsymbol{\theta} \in [0,h]^d]$.

We will prove the above result at the end of this section. In view of this, it suffices to bound $\mathsf{Regret}(\widehat{G}, G_h)$ to get the desired regret upper bound. To bound $\mathsf{Regret}(\widehat{G}, G_h) = \mathbb{E}_{G_h}\left[\|\widehat{\boldsymbol{\theta}}_{\widehat{G}}(\boldsymbol{Y}) - \widehat{\boldsymbol{\theta}}_{G_h}(\boldsymbol{Y})\|^2\right]$ we use the following decomposition that is similar to the decomposition in the proof of the one dimensional case. Note that $\widehat{G}$ is supported on $[0,\widehat{h}]^d$, where $\widehat{h} = h$ when $\widehat{G}$ is chosen to be supported over $[0,h]^d$ or as $\max_{j=1}^d \max_{i=1}^n Y_{ij} + 1$ when $\widehat{G}$ is obtained by performing an unconstrained optimization. Note that in the later case, as we argued above, $\widehat{h}$ is

bounded from above by $\widetilde{cs}\log n$ with a probability $1-\frac{d}{n^9}$. Hence, we have

$$
\begin{aligned}
&\mathbb{E}_{G_h}\left[\|\widehat{\boldsymbol{\theta}}_{\widehat{G}}(\boldsymbol{Y})-\widehat{\boldsymbol{\theta}}_{G_h}(\boldsymbol{Y})\|^2\mathbf{1}_{\{\boldsymbol{Y}\in[0,R]^d\}}\right]\\
&=\sum_{\boldsymbol{y}\in[0,R]^d}\sum_{j=1}^{d}(y_j+1)^2 f_{G_h}(\boldsymbol{y})\left(\frac{f_{\widehat{G}}(\boldsymbol{y}+\boldsymbol{e}_j)}{f_{\widehat{G}}(\boldsymbol{y})}-\frac{f_{G_h}(\boldsymbol{y}+\boldsymbol{e}_j)}{f_{G_h}(\boldsymbol{y})}\right)^2\\
&\overset{\text{(a)}}{\le}\sum_{\boldsymbol{y}\in[0,R]^d}\sum_{j=1}^{d}(y_j+1)^2 f_{G_h}(\boldsymbol{y})\left\{3\left(\frac{f_{\widehat{G}}(\boldsymbol{y}+\boldsymbol{e}_j)}{f_{\widehat{G}}(\boldsymbol{y})}-\frac{2f_{\widehat{G}}(\boldsymbol{y}+\boldsymbol{e}_j)}{f_{G_h}(\boldsymbol{y})+f_{\widehat{G}}(\boldsymbol{y})}\right)^2+3\left(\frac{f_{G_h}(\boldsymbol{y}+\boldsymbol{e}_j)}{f_{G_h}(\boldsymbol{y})}-\frac{2f_{G_h}(\boldsymbol{y}+\boldsymbol{e}_j)}{f_{G_h}(\boldsymbol{y})+f_{\widehat{G}}(\boldsymbol{y})}\right)^2\right.\\
&\quad\left.+3\left(\frac{2f_{G_h}(\boldsymbol{y}+\boldsymbol{e}_j)-2f_{\widehat{G}}(\boldsymbol{y}+\boldsymbol{e}_j)}{f_{G_h}(\boldsymbol{y})+f_{\widehat{G}}(\boldsymbol{y})}\right)^2\right\}\\
&\le 3\sum_{\boldsymbol{y}\in[0,R]^d}\sum_{j=1}^{d}\left\{\left(\frac{(y_j+1)f_{\widehat{G}}(\boldsymbol{y}+\boldsymbol{e}_j)}{f_{\widehat{G}}(\boldsymbol{y})}\right)^2\frac{(f_{G_h}(\boldsymbol{y})-f_{\widehat{G}}(\boldsymbol{y}))^2}{f_{G_h}(\boldsymbol{y})+f_{\widehat{G}}(\boldsymbol{y})}+\left(\frac{(y_j+1)f_{G_h}(\boldsymbol{y}+\boldsymbol{e}_j)}{f_{G_h}(\boldsymbol{y})}\right)^2\frac{(f_{G_h}(\boldsymbol{y})-f_{\widehat{G}}(\boldsymbol{y}))^2}{f_{G_h}(\boldsymbol{y})+f_{\widehat{G}}(\boldsymbol{y})}\right.\\
&\quad\left.+4(y_j+1)^2\frac{(f_{G_h}(\boldsymbol{y}+\boldsymbol{e}_j)-f_{\widehat{G}}(\boldsymbol{y}+\boldsymbol{e}_j))^2}{f_{G_h}(\boldsymbol{y})+f_{\widehat{G}}(\boldsymbol{y})}\right\}\\
&\le 3(h^2+\widehat{h}^2)\sum_{\boldsymbol{y}\in[0,R]^d}\sum_{j=1}^{d}\frac{(f_{G_h}(\boldsymbol{y})-f_{\widehat{G}}(\boldsymbol{y}))^2}{f_{G_h}(\boldsymbol{y})+f_{\widehat{G}}(\boldsymbol{y})}+12\sum_{\boldsymbol{y}\in[0,R]^d}\sum_{j=1}^{d}(y_j+1)^2\frac{(f_{G_h}(\boldsymbol{y}+\boldsymbol{e}_j)-f_{\widehat{G}}(\boldsymbol{y}+\boldsymbol{e}_j))^2}{f_{G_h}(\boldsymbol{y})+f_{\widehat{G}}(\boldsymbol{y})}
\end{aligned}
$$

where (a) followed from $(x+y+z)^2\le 3(x^2+y^2+z^2)$ for any $x,y,z\in\mathbb{R}$. Using $(\sqrt{f_{G_h}(\boldsymbol{x})}+\sqrt{f_{\widehat{G}}(\boldsymbol{x})})^2\le 2(f_{G_h}(\boldsymbol{x})+f_{\widehat{G}}(\boldsymbol{x}))$ for $\boldsymbol{x}=\boldsymbol{y},\boldsymbol{y}+\boldsymbol{e}_j$ we continue the last display to get

$$
\begin{aligned}
&\mathbb{E}_{G_h}\left[\|\widehat{\boldsymbol{\theta}}_{\widehat{G}}(\boldsymbol{Y})-\widehat{\boldsymbol{\theta}}_{G_h}(\boldsymbol{Y})\|^2\mathbf{1}_{\{\boldsymbol{Y}\in[0,R]^d\}}\right]\\
&\le 6(h^2+\widehat{h}^2)\sum_{\boldsymbol{y}\in[0,R]^d}\sum_{j=1}^{d}(\sqrt{f_{G_h}(\boldsymbol{y})}-\sqrt{f_{\widehat{G}}(\boldsymbol{y})})^2\\
&\quad+24R\max_{\boldsymbol{y}\in[0,R]^d}\sum_{j=1}^{d}\frac{(y_j+1)f_{G_h}(\boldsymbol{y}+\boldsymbol{e}_j)+(y_j+1)f_{\widehat{G}}(\boldsymbol{y}+\boldsymbol{e}_j)}{f_{G_h}(\boldsymbol{y})+f_{\widehat{G}}(\boldsymbol{y})}\sum_{\boldsymbol{y}\in[0,R]^d}\sum_{j=1}^{d}(\sqrt{f_{G_h}(\boldsymbol{y}+\boldsymbol{e}_j)}-\sqrt{f_{\widehat{G}}(\boldsymbol{y}+\boldsymbol{e}_j)})^2\\
&\le d\left(6(h^2+\widehat{h}^2)+24(h+\widehat{h})R\right)H^2(f_{\widehat{G}},f_{G_h}),
\end{aligned}
$$

where the last inequality followed as

$$
\begin{aligned}
&\sum_{j=1}^{d}\frac{(y_j+1)f_{G_h}(\boldsymbol{y}+\boldsymbol{e}_j)+(y_j+1)f_{\widehat{G}}(\boldsymbol{y}+\boldsymbol{e}_j)}{f_{G_h}(\boldsymbol{y})+f_{\widehat{G}}(\boldsymbol{y})}\sum_{\boldsymbol{y}\in[0,R]^d}\sum_{j=1}^{d}(\sqrt{f_{G_h}(\boldsymbol{y}+\boldsymbol{e}_j)}-\sqrt{f_{\widehat{G}}(\boldsymbol{y}+\boldsymbol{e}_j)})^2\\
&\le\sum_{j=1}^{d}\frac{(y_j+1)f_{G_h}(\boldsymbol{y}+\boldsymbol{e}_j)+(y_j+1)f_{\widehat{G}}(\boldsymbol{y}+\boldsymbol{e}_j)}{f_{G_h}(\boldsymbol{y})+f_{\widehat{G}}(\boldsymbol{y})}\sum_{\boldsymbol{y}\in[0,R+1]^d}(\sqrt{f_{G_h}(\boldsymbol{y})}-\sqrt{f_{\widehat{G}}(\boldsymbol{y})})^2\le H^2(f_{\widehat{G}},f_{G_h}).
\end{aligned}
$$

Again, using $\widehat{\boldsymbol{\theta}}_{G_h}(\boldsymbol{y}) \le h, \widehat{\boldsymbol{\theta}}_{\widehat{G}}(\boldsymbol{y}) \le \widehat{h}$ we bound $\mathbb{E}_{G_h}\left[(\widehat{\boldsymbol{\theta}}_{\widehat{G}}(\boldsymbol{Y}) - \widehat{\boldsymbol{\theta}}_{G_h}(\boldsymbol{Y}))^2 \mathbf{1}_{\{\boldsymbol{Y} \notin [0,R]^d\}}\right]$ by $(h + \widehat{h})^2 \varepsilon_R(G_h)$, where $\varepsilon_R(G_h) = \mathbb{P}_{G_h}\left[\boldsymbol{Y} \notin [0,R]^d\right]$. Combining this with the last display we get

$$\mathbb{E}_{G_h}\left[\|\widehat{\boldsymbol{\theta}}_{\widehat{G}}(\boldsymbol{Y}) - \widehat{\boldsymbol{\theta}}_{G_h}(\boldsymbol{Y})\|^2\right] \le d\left\{6(h^2 + \widehat{h}^2) + 24(h+\widehat{h})R\right\} H^2(f_{\widehat{G}}, f_{G_h}) + (h+\widehat{h})^2 \varepsilon_R(G_h).$$

Finally we take expectation on both sides with respect to the training sample, $\widehat{G}$, and $\widehat{h}$.

(i) The proof of the result for $G \in \mathscr{P}([0,h]^d)$, with both unconstrained and constrained estimator (which has knowledge of h) is very similar to the proof of Theorem 3 and is omitted here.

(ii) In the case of subexponential marginals of G, we pick $h = \widetilde{c}\max\{1,s\}\log n$, for a large enough constant $\widetilde{c}$. Hence, the support parameter $\widehat{h}$ of $\widehat{G}$ as well as the support of each coordinate of $\boldsymbol{\theta}$ is bounded from above by h with a high probability. Using the high probability bound on $\widehat{h}$ and the bound on $\mathbb{E}\left[H^2(f_{\widehat{G}}, f_{G_h})\right]$ as in Theorem 5, the analysis for the bounded prior setup applies even though $\widehat{G}$ is obtained via an unconstrained optimization. Finally, using the bound on $\varepsilon_R(G_h)$ as in Lemma 13 we get the result.

This finishes the proof. □

Proof of Lemma 14 We first note the multi-dimensional version of (C.1) as follows:

$$\begin{aligned}
\mathbb{E}_G\left[\|\widehat{\boldsymbol{\theta}}_{\widehat{G}}(\boldsymbol{Y}) - \boldsymbol{\theta}\|^2\right] &\le \mathbb{E}_G\left[(\widehat{\boldsymbol{\theta}}_{\widehat{G}}(\boldsymbol{Y}) - \boldsymbol{\theta}\|^2 \mathbf{1}_{\{\boldsymbol{\theta} \in [0,h]^d\}}\right] + \mathbb{E}_G\left[\|\widehat{\boldsymbol{\theta}}_{\widehat{G}}(\boldsymbol{Y}) - \boldsymbol{\theta}\|^2 \mathbf{1}_{\{\boldsymbol{\theta} \notin [0,h]^d\}}\right] \\
&\overset{(a)}{\le} \mathbb{E}_G\left[\|\widehat{\boldsymbol{\theta}}_{\widehat{G}}(\boldsymbol{Y}) - \boldsymbol{\theta}\|^2 \middle| \boldsymbol{\theta} \in [0,h]^d\right] + \sqrt{\mathbb{E}_G\left[\|\widehat{\boldsymbol{\theta}}_{\widehat{G}}(\boldsymbol{Y}) - \boldsymbol{\theta}\|^4\right] \mathbb{E}_G\left[\mathbf{1}_{\{\boldsymbol{\theta} \notin [0,h]^d\}}\right]} \\
&\overset{(b)}{\le} \mathbb{E}_{G_h}\left[\|\widehat{\boldsymbol{\theta}}_{\widehat{G}}(\boldsymbol{Y}) - \boldsymbol{\theta}\|^2\right] + O\left(\sqrt{d^4(\widehat{h}^4 + s^4)\frac{d}{n^9}}\right),
\end{aligned}$$

where step (a) followed by Cauchy-Schwarz inequality, step (b) followed as

(i) $(x+y)^4 \le 8(x^4+y^4)$ for any $x, y \in \mathbb{R}$

(ii) each coordinate of $\widehat{\boldsymbol{\theta}}_{\widehat{G}}(\boldsymbol{Y})$ is bounded by $\widehat{h}$

(iii) $\mathbb{E}[\theta_j^4] \le O(\max\{1, s^4\}), j = 1, \ldots, d$ by Lemma 12

(iv) $\mathbb{P}[\boldsymbol{\theta} \notin [0,h]^d] \le \frac{d}{n^9}$ for a large enough $\widetilde{c}$.

Then, similar to (C.2) in the one-dimensional case, the following equation applies (note that here $\widehat{G}$ is supported on $[0, \widetilde{c}\max\{1,s\}\log n]$ with a high probability):

$$\mathsf{Regret}(\widehat{G}; G) \le \mathsf{Regret}(\widehat{G}; G_h) + \mathrm{mmse}(G_h) - \mathrm{mmse}(G) + O\left(\frac{\widehat{h}^2 + s^2}{n^2}\right).$$

Note that we have $\mathrm{mmse}(G_h) \le \frac{\mathrm{mmse}(G)}{\mathbb{P}\left[\boldsymbol{\theta} \in [0,h]^d\right]}$ from the following

$$\mathrm{mmse}(G) = \mathbb{E}[\|\widehat{\boldsymbol{\theta}}_G - \boldsymbol{\theta}\|^2] \ge \mathbb{P}\left[\boldsymbol{\theta} \in [0,h]^d\right] \; \mathbb{E}_{\boldsymbol{\theta} \sim G}[\|\widehat{\boldsymbol{\theta}}_G - \boldsymbol{\theta}\|^2 | \boldsymbol{\theta} \in [0,h]^d] \ge \mathbb{P}\left[\boldsymbol{\theta} \in [0,h]^d\right] \mathrm{mmse}(G_h).$$

As $\mathrm{mmse}(G) \le d$ (given that the naive estimation of $\boldsymbol{Y}$ achieves an expected loss of d) and $\mathbb{P}\left[\boldsymbol{\theta} \notin [0,h]^d\right] \le \frac{1}{n^8}$, we get the desired result. □