Sharp regret bounds for empirical Bayes and compound decision problems

Yury Polyanskiy and Yihong Wu*

September 9, 2021

Abstract

We consider the classical problems of estimating the mean of an n-dimensional normally (with identity covariance matrix) or Poisson distributed vector under the squared loss. In a Bayesian setting the optimal estimator is given by the (prior-dependent) conditional mean. In a frequentist setting various shrinkage methods were developed over the last century. The framework of empirical Bayes, put forth by Robbins [Rob56], combines Bayesian and frequentist mindsets by postulating that the parameters are independent but with an unknown prior and aims to use a fully data-driven estimator to compete with the Bayesian oracle that knows the true prior. The central figure of merit is the regret, namely, the total excess risk over the Bayes risk in the worst case (over the priors). Although this paradigm was introduced more than 60 years ago, little is known about the asymptotic scaling of the optimal regret in the nonparametric setting.

We show that for the Poisson model with compactly supported and subexponential priors, the optimal regret scales as $\Theta((\frac{\log n}{\log\log n})^2)$ and $\Theta(\log^3 n)$, respectively, both attained by the original estimator of Robbins. For the normal mean model, the regret is shown to be at least $\Omega((\frac{\log n}{\log\log n})^2)$ and $\Omega(\log^2 n)$ for compactly supported and subgaussian priors, respectively, the former of which resolves the conjecture of Singh [Sin79] on the impossibility of achieving bounded regret; before this work, the best regret lower bound was $\Omega(1)$. In addition to the empirical Bayes setting, these results are shown to hold in the compound setting where the parameters are deterministic. As a side application, the construction in this paper also leads to improved or new lower bounds for density estimation of Gaussian and Poisson mixtures.

Contents

1 Introduction **2**

1.1 Definitions and main results 3

1.2 Related work 7

1.3 Notations 8

1.4 Organization 8

*Y.P. is with the Department of EECS, MIT, Cambridge, MA, email: yp@mit.edu. Y.W. is with the Department of Statistics and Data Science, Yale University, New Haven, CT, email: yihong.wu@yale.edu. 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.

2 Preliminaries **8**
2.1 Reduction to estimating regression functions 8
2.2 Truncation of priors 9
2.3 General program for regret lower bound 10

3 Proofs of the main results **14**
3.1 Normal mean model 14
3.2 Poisson model 17

4 Discussion **21**

A Proofs for the Gaussian model **22**
A.1 Proof of Lemma 9 22
A.2 Proof for Remark 2 24

B Proofs for the Poisson model **25**

C Regret bound for the Robbins estimator **30**

D Proofs for the compound setting **36**

E Results on density estimation **38**

1 Introduction

Consider estimating an unknown parameter $\theta \in \mathbb{R}^{10}$ from a single measurement $Y \sim \mathcal{N}(\theta, I_{10})$. Suppose that the observed data is

$$Y = [1.08,\ -2.43,\ 1.52,\ -1.17,\ 1.39,\ 8.3,\ 10.02,\ 10.61,\ 9.3,\ 10.14]\,.$$

Although the worst-case reasoning suggests simply using the maximum likelihood estimator (MLE) $\widehat{\theta} = Y$, it is well-known by now that it is often advantageous to use shrinkage estimators, such as James-Stein. However, in the example shown,[1] it should also be clear that applying the James-Stein estimator on the first 5 coordinates and the rest *separately* would be even more advantageous since shrinking to a common mean of all 10 coordinates (in this bimodal situation) appears unnatural. How can one formalize derivation of such clever procedures and how can one study their fundamental limits? Both of these questions were addressed in a decision-theoretic framework proposed by Robbins [Rob51, Rob56] under the name of *empirical Bayes*. Since then a lot has been written about this paradigm and empirical Bayes methodology has been widely used in practice for large-scale data analysis, with notable applications in computational biology especially microarrays [ETST01], sports prediction [Bro08], etc. We refer to the survey articles [Cas85, Zha03, Efr21] and the monograph [Efr12] for theory and methodology for empirical Bayes.

The basic tenet and the major surprise of the empirical Bayes theory is that, when the number of independent observations is large, it is possible to "borrow strength" from these independent (and seemingly unrelated) observations so as to achieve the optimal Bayes risk per coordinate asymptotically. In this work we study the principal performance metric (regret) and its asymptotic scaling as the number of observations grows, thereby characterizing the optimal speed to approach the Bayes risk. We proceed to formal definitions.

[1] For a bigger example in the same spirit, see Efron's "two towers" example [Efr19, Figure 1].

1.1 Definitions and main results

Throughout the paper, we use the vector notation $\theta^n \triangleq (\theta_1, \ldots, \theta_n)$. Let $\{P_\theta : \theta \in \Theta \subset \mathbb{R}\}$ be a parametric family of distributions with a bi-measurable density $f_\theta(y)$ with respect to a common dominating measure ν. Conditioned on the parameters θ^n, the observations $Y_1, \ldots, Y_n$ are independently distributed according to $P_{\theta_i}, i = 1, \ldots, n$.

Following the terminology in [Rob51, Rob56, Zha97, Zha03], we consider both the *empirical Bayes* (EB) and *compound* settings, in which θ_i's are iid (with an unknown prior) or deterministic, respectively. In both settings, the goal is to estimate the parameters $\theta_1, \ldots, \theta_n$ under the quadratic loss and compete with the oracle that has extra distributional information (e.g. the prior or the empirical distribution of the parameters).

We start with the empirical Bayes setting. Let $\theta^n = (\theta_1, \ldots, \theta_n) \overset{\text{i.i.d.}}{\sim} G$. Then $Y_i \overset{\text{i.i.d.}}{\sim} P_G$, where $P_G = \int P_\theta G(d\theta)$ is the mixture under the prior (mixing distribution) G with density $f_G(y) = \int f_\theta(y) G(d\theta)$. Denote by $\mathbb{E}_G$ the expectation taken over the joint distribution of (θ^n, Y^n). The Bayes risk under the prior G is the minimum mean squared error (MMSE) of estimating $\theta \sim G$ based on a single observation $Y \sim f_\theta$:

$$\mathrm{mmse}(G) \triangleq \inf_{\widehat{\theta}} \mathbb{E}_G[(\widehat{\theta}(Y) - \theta)^2] = \mathbb{E}_G[(\widehat{\theta}_G(Y) - \theta)^2], \tag{1}$$

where

$$\widehat{\theta}_G(y) \triangleq \mathbb{E}_G[\theta | Y = y] = \frac{\int_\Theta \theta f_\theta(y) G(d\theta)}{\int_\Theta f_\theta(y) G(d\theta)} \tag{2}$$

is the Bayes estimator (conditional mean) associated with the prior G. In this paper we will be concerned with the following basic models:

- Normal mean model: $P_\theta = \mathcal{N}(\theta, 1), \theta \in \mathbb{R}$, with density $f_\theta(y) = \varphi(y - \theta)$ and $\varphi(y) \triangleq \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$. For any prior G, the corresponding Bayes estimator can be expressed using the mixture density f_G and its derivative as (Tweedie's formula [Efr11]):

$$\widehat{\theta}_G(y) = y + \frac{f_G'(y)}{f_G(y)}, \tag{3}$$

  with the second term known as the Bayes correction.

- Poisson model: $P_\theta = \mathrm{Poi}(\theta), \theta \in \mathbb{R}_+$, with probability mass function $f_\theta(y) = \frac{e^{-\theta}\theta^y}{y!}$ and $y \in \mathbb{Z}_+ \triangleq \{0, 1, \ldots\}$. The Bayes estimator takes the form:

$$\widehat{\theta}_G(y) = (y+1)\frac{f_G(y+1)}{f_G(y)}. \tag{4}$$

Both models are well studied and widely applied in practice. In addition to denoising, the normal mean model has also found usefulness in other high-dimensional problems such as principle component analysis.

The main objective of empirical Bayes is to compete with the oracle that knows the true prior and minimize the excess risk, known as the *regret* [JZ09, Efr11, Efr19], by mimicking the true Bayes estimator. This strategy is either carried out explicitly (by learning a prior from data then executing the learned Bayes estimator) or implicitly (by constructing an approximate Bayes estimator using

the marginal distribution of the observation directly).[2] Robbins' famous estimator for the Poisson model [Rob56] belongs to the latter category. Namely, we estimate θ_j by

$$\widetilde{\theta}_j \triangleq (Y_j+1)\frac{N(Y_j+1)}{N(Y_j)}\,, \quad N(y) \triangleq |\{i \in [n] : Y_i = y\}|. \tag{5}$$

which can be viewed as a plugin version of the Bayes estimator (3) replacing the marginal f_G by the empirical version. Despite the broad application of the Robbins estimator in practice, little is known about its optimality.

To study the fundamental limit of empirical Bayes estimation, let us define the optimal regret. Given a collection of priors $\mathcal{G}$, the total regret is defined as

$$\mathsf{TotRegret}_n(\mathcal{G}) \triangleq \inf_{\widehat{\theta}^n} \sup_{G\in\mathcal{G}} \left\{\mathbb{E}_G\left[\|\widehat{\theta}^n(Y_1,\ldots,Y_n) - \theta^n\|^2\right] - n\cdot \mathrm{mmse}(G)\right\}. \tag{6}$$

where $\|\cdot\|$ denotes the Euclidean norm and the infimum is taken over all estimator $\widehat{\theta}^n$ with respect to $Y_1,\ldots,Y_n$. In this paper we focus on *nonparametric* settings; for example, the class $\mathcal{G}$ consists of all priors over a compact parameter space.

Under appropriate assumptions, Robbins demonstrated it is possible to achieve a *sublinear* total regret $o(n)$, so that the excess risk per observation is amortized as the sample size n tends to infinity. A natural question is to determine the order of the optimal regret. Singh conjectured [Sin79, p. 891] that for any exponential family, bounded total regret is not possible even if the priors are compactly supported. In this work we resolve this conjecture in the positive for both normal and Poisson model. Before this work the best lower bound is $\Omega(1)$ [LGL05], which can be shown by considering a parametric family of priors.

Our main results are as follows:

Theorem 1 (Normal means problem)**.** *Consider the normal mean model: $P_\theta = \mathcal{N}(\theta,1), \theta\in\mathbb{R}$.*

- *Let $\mathcal{P}([-h,h])$ be the set of all distributions supported on $[-h,h]$. There exist constants $c_0 = c_0(h) > 0$ and n_0, such that for all $n \ge n_0$,*

$$\mathsf{TotRegret}_n(\mathcal{P}([-h,h]) \ge c_0\left(\frac{\log n}{\log\log n}\right)^2. \tag{7}$$

- *Denote by $\mathsf{SubG}(s)$ the set of all s-subgaussian distributions, namely $\mathsf{SubG}(s) = \{G : G([-t,t]^c) \le 2e^{-t^2/(2s)}, \forall t>0\}$. Then for some constant $c_1 = c_1(s) > 0$ and all $n \ge n_0$*

$$\mathsf{TotRegret}_n(\mathsf{SubG}(s)) \ge c_1 \log^2 n\,. \tag{8}$$

Theorem 2 (Poisson problem)**.** *Consider the Poisson model $P_\theta = \mathrm{Poi}(\theta), \theta\in\mathbb{R}_+$.*

- *For $h>0$, there exist constants $c_1 = c_1(h), c_2 = c_2(h)$ such that for all $n \ge n_0$,*

$$c_1\left(\frac{\log n}{\log\log n}\right)^2 \le \mathsf{TotRegret}_n(\mathcal{P}([0,h]) \le c_2\left(\frac{\log n}{\log\log n}\right)^2. \tag{9}$$

- *Denote by $\mathsf{SubE}(s)$ the set of all s-subexponential distributions on $\mathbb{R}_+$, namely $\mathsf{SubE}(s) = \{G : G([t,\infty)) \le 2e^{-t/s}, \forall t>0\}$. Then for some constants $c_3 = c_3(s), c_4 = c_4(s)$ and all $n \ge n_0$,*

$$c_3 \log^3 n \le \mathsf{TotRegret}_n(\mathsf{SubE}(s)) \le c_4 \log^3 n\,. \tag{10}$$

[2] These two approaches are referred to as g-modeling and f-modeling in [Efr14], respectively.

Furthermore, in both (9) and (10), the upper bound is achieved by the estimator (5) of Robbins.

A few remarks are in order:

1. It is known that tail assumptions (e.g. compact support or moment conditions such as subgaussianity) are necessary for achieving non-trivial regret. In fact, in the normal means problem, if we allow all priors on $\mathbb{R}$, [Zha09, Example 1] showed that $\mathsf{TotRegret}_n = n + o(n)$, which is almost as large as possible.

2. The upper bound in (9) for compactly supported priors is essentially contained in [BGR13], who analyzed the risk of the Robbins estimator with sample size n replaced by $\mathrm{Poi}(n)$, which renders the empirical counts $N(y)$'s in (5) independent for technical convenience. In Appendix C we bound the regret of the Robbins estimator without Poissonization following essentially the same strategy as in [BGR13] and fix a misstep therein.

3. The notion of *conditional regret* is defined in [Efr11, Sec. 4], which refers to the excess prediction risk of a given EB estimator over the oracle (Bayes) risk conditioned on the value y of the observation. For several parametric classes of priors and parametric EB estimators, such as Gaussian priors and the James-Stein estimator (see [Efr11, Eq. (4.10)]), the conditional regret is shown to be $\frac{c(y)}{n}$ for some explicit constant $c(y)$. Our results show that for nonparametric classes of priors, this constant $c(y)$ is not uniformly bounded and the conditional regret in the worst case is of higher order than $\frac{1}{n}$ by logarithmic factors.

4. In addition to Robbins estimator, which does not take a Bayes form, the optimal regret in (9) and (10) in the Poisson model can also achieved by a more principled approach of first estimating the prior then using the corresponding Bayes estimator. As suggested by Robbins [Rob56, Section 6], one could use Wolfowitz's minimum-distance estimator [Wol57]. This strategy indeed succeeds for a class of minimum-distance estimators, including the nonparametric MLE (NPMLE) [KW56, Lin83] which minimizes the Kullback-Leibler divergence [JPW21]. In contrast, for the Gaussian model, the current best regret upper bound is $O(\log^5 n)$ for the subgaussian (hence also compactly supported) class [JZ09] – see Section 1.2 for a more detailed discussion.

5. An interesting open question for the Poisson model is to obtain a constant-factor characterization of regret uniformly over h (or s) in $(0, \infty)$. At this point it is not clear how magnitude of h (or s) affects the regret and such characterization would clarify this.

Compound setting. In the compound estimation problem, the unknown parameter θ^n is no longer assumed to be random. Again, we are interested in estimating this deterministic vector θ^n by competing with an oracle possessing some side information about θ^n. There are two natural choices of the oracle resulting in the two definitions of the regret, that we define below. For a given θ^n let us denote its empirical distribution by

$$G_{\theta^n} \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{\theta_i}. \tag{11}$$

A simple observation is that

$$\inf_{f_1} \sum_{i=1}^n \mathbb{E}_{\theta_i}[(\theta_i - f_1(Y_i))^2] = n \cdot \mathrm{mmse}(G_{\theta^n}),$$

achieved by the Bayes estimator under the prior being the empirical distribution, namely, $\mathbb{E}_{\theta\sim G_{\theta^n}}[\theta|Y]$. This leads us to defining the (most commonly used) version of the regret in the compound setting:

$$\mathsf{TotRegretComp}_n(\Theta) \triangleq \inf_{\widehat{\theta}^n} \sup_{\theta^n\in\Theta^{\otimes n}} \left\{\mathbb{E}[\|\widehat{\theta}^n - \theta^n\|^2] - n\cdot \mathrm{mmse}(G_{\theta^n})\right\}, \tag{12}$$

which, by the observation above, corresponds to competing with an oracle who can adapt to θ^n but is restricted to use a separable estimator, namely, applying a common univariate rule f_1 to all coordinates.

An alternative notion of regret was introduced in [GR09], which can be formalized by defining

$$R_{\mathsf{oracle}}(G_{\theta^n}) = \inf_f \sup_{\sigma\in S_n} \mathbb{E}_{\sigma\theta^n}[\|\sigma\theta^n - f(Y^n)\|^2],$$

where S_n is the group of permutations on $[n]$ and $\sigma\theta^n \triangleq (\theta_{\sigma(1)},\dots,\theta_{\sigma(n)})$. One can show that the value of R_{oracle} is unchanged if one (a) replaces $\sup_\sigma$ with average over a random uniform σ; or (b) if one removes $\sup\sigma$ and instead restricts the infimum to permutation-invariant functions $f:\mathbb{R}^n\to\mathbb{R}^n$. (As [GR09] shows, the optimal f is given by the conditional expectation $\mathbb{E}[\theta^n|Y^n]$ under the uniform prior on the S_n orbit of the support of G_{θ^n}.) The corresponding notion of regret is then defined as[3]

$$\mathsf{TotRegretComp}'_n(\Theta) \triangleq \inf_{\widehat{\theta}^n} \sup_{\theta^n\in\Theta^{\otimes n}} \left\{\mathbb{E}[\|\widehat{\theta}^n - \theta^n\|^2] - R_{\mathsf{oracle}}(G_{\theta^n})\right\}, \tag{13}$$

which can be understood either as (a) competing with the oracle that has information about G_{θ^n} (but not θ^n), or (b) can adapt to θ^n but is restricted to use permutation-invariant estimators $\widehat{\theta}^n$.

It turns out that our lower bounds for the EB regret also apply to both versions of the compound regret. Indeed, denote by $\mathcal{P}(\Theta)$ the collection of all priors on the parameter space Θ. The following general lemma relates the regret in the compound setting to that of the EB setting (see Appendix D for a proof). As a result, for bounded parameter space, the $\Omega((\frac{\log n}{\log\log n})^2)$ regret lower bound in Theorem 1 and Theorem 2 also hold in the compound setting.

Proposition 3.

$$\mathsf{TotRegretComp}'_n(\Theta) \ge \mathsf{TotRegretComp}_n(\Theta) \ge \mathsf{TotRegret}_n(\mathcal{P}(\Theta)). \tag{14}$$

Extending the definitions in (12) and (13), one can define the corresponding total regret in the constrained setting, denoted by $\mathsf{TotRegretComp}_n(\mathcal{G})$ and $\mathsf{TotRegretComp}'_n(\mathcal{G})$ respectively, where the supremum is taken over $\theta^n\in\mathbb{R}^n$ whose empirical distribution belongs to a prescribed class $\mathcal{G}$. Similar to Proposition 3, one can relate the regret in the compound setting to that of the EB setting. The following result (also proved in Appendix D) provides such a reduction for the normal means problem with subgaussian priors, which shows the $\Omega(\log^2 n)$ lower bound in Theorem 1 continues to hold in the compound setting.

Proposition 4. *Consider the normal means problem. There exists a universal constant c_0 such that for any $s>0$,*

$$\mathsf{TotRegretComp}(\mathsf{SubG}(s)) \ge \mathsf{TotRegret}(\mathsf{SubG}(c_0 s)) - \frac{s}{c_0 n},$$

where $\mathsf{SubG}(s)$ denotes the collection of s-subgaussian distributions in Theorem 1.

[3] Conditions under which $\mathsf{TotRegretComp}'_n = \mathsf{TotRegretComp}_n + O(1)$ are given in [GR09, Theorem 5.1], which are satisfied by bounded normal mean model [GR09, Corollary 5.2].

1.2 Related work

The field of empirical Bayes and compound decision is exceedingly rich in terms of both theoretical and methodological development. In this paper we have focused on the specific problem of *estimation* under the squared error in both the normal and Poisson model, with the main objective of determining the optimal regret for various nonparametric classes of priors. Next we review the existing literature that are directly related to our main results. Later in Section 4 we discuss a few different formulations and perspectives in empirical Bayes in connection with the present paper.

As mentioned earlier, the best regret lower bound before this work is $\Omega(1)$ [LGL05, Theorem 2.2], which is shown by a two-point argument (Le Cam's method), namely, choosing two priors which differ by $\frac{1}{\sqrt{n}}$ in mean and whose corresponding mixture distributions has statistical distance (e.g. KL divergence) $O(1/n)$. As well-known in the high-dimensional statistics literature, two-point method often fails to capture the correct dependency on the model complexity. Indeed, in order to show superconstant regret, it is necessary to consider many hypotheses and Assouad's lemma provides a principled way to do so; see Section 2.3 for a general program of proving regret lower bound.

For regret upper bound in the Poisson model, the only quantitative result appears to be [BGR13] analyzing the original Robbins estimator with Poissonized sample size, which turns out to be optimal in view of Theorem 2. For the normal mean model, a lot more is known but none meets the lower bound in Theorem 1. The state of the art is achieved by the Generalized Maximum Likelihood Empirical Bayes (GMLEB) method in [JZ09]. The idea of GMLEB is to first estimate a prior using the NPMLE, then using the (regularized) Bayes rule with this estimated prior. Various regret bounds are obtained under moment or subgaussian assumptions on the prior by a reduction from the Hellinger risk of density estimation to regret [JZ09, Theorem 3 (ii)], and leveraging density estimation results for Gaussian mixtures [GvdV01, Zha09]. In particular, for the subgaussian class, it is shown that $\mathsf{TotRegret}(\mathsf{SubG}(C)) = O(\log^5 n)$ for constant C. (In comparison, Theorem 1 shows $\mathsf{TotRegret}(\mathsf{SubG}(C)) = \Omega(\log^2 n)$.) Furthermore, these upper bounds on the regret (additive error) are extended to the compound setting [JZ09, Theorems 1 and 5] leading to attaining the Bayes risk (under the empirical distribution as the prior) up to multiplicative factor of $1 + o(1)$, referred to as *adaptive ratio optimality*, provided that the Bayes risk is at least $\Omega(\frac{\log^5 n}{n})$.

In addition to GMLEB, kernel-based EB estimator is also commonly used. Note that unlike the discrete Poisson model, there exists no unbiased estimator for the mixture density, so there is no counterpart of Robbins estimator (5) for the Gaussian model. Nevertheless, the form of the Bayes estimator (3) motivates using kernel-based estimators in place of the mixture density and its derivative. Using polynomial kernels, Singh [Sin79] obtained upper bound for general exponential families, which, for the bounded normal means problem, leads to $n^{o(1)}$ upper bound for the total regret, which is subsequently improved to $O((\log n)^8)$ in [LGL05, Example 1] by using polynomials of degree that scales logarithmically in n. On the other hand, in view of the analyticity of the Gaussian mixture, [Zha09] uses sinc kernel to achieve adaptive ratio optimality and a total regret that scales as $\sqrt{n}$. Similar result is also obtained in [BG09] using Gaussian kernel under stronger assumptions.

Another line of research aims to obtain regret bound under regularity assumptions on the prior, e.g., by imposing conditions on the Fourier transform $\widetilde{g}(\omega)$ of the prior density g. Pensky [Pen99] constructed a wavelet-based EB estimator $\widetilde{\theta}(\cdot) = \widetilde{\theta}(\cdot; Y_1, \ldots, Y_n)$, which, under the assumption that $\int |\widetilde{g}(\omega)|^2(\omega^2+1)^\alpha < \infty$ for some $\alpha > 0$, satisfies $\mathbb{E}[(\widetilde{\theta}(y) - \theta_G(y))^2] = O(\frac{(\log n)^{3/2}}{n})$ uniformly for y in any compact interval. For the related testing problem, Liang [Lia04] showed that if $\widetilde{g}$ has bounded support then the parametric rate $O(\frac{1}{n})$ is achievable by a kernel-based EB estimator. In this paper,

we only impose tail conditions on the prior rather than smoothness of the prior density.

1.3 Notations

Given a measurable space $(\mathcal{X}, \mathcal{F})$, let $\mathcal{P}(\mathcal{X})$ denote the collection of all probability measures on $(\mathcal{X}, \mathcal{F})$. Throughout this paper, $\mathcal{X}$ will be taken to be a subset of $\mathbb{R}$ and $\mathcal{F}$ the Borel σ-algebra. The Lebesgue measure on $\mathbb{R}$ is denoted by Leb. Given a measure μ on $\mathcal{X}$, let $L_2(\mu)$ denote the collection of all square-integrable functions $f : \mathcal{X} \to \mathbb{R}$ with $\|f\|_{L_2(\mu)} \triangleq (\int_{\mathcal{X}} f(x)^2 \mu(dx))^{1/2}$ and inner product $(f, g)_{L_2(\mu)} \triangleq \int_{\mathcal{X}} f(x) g(x) \mu(dx)$.

Throughout the paper we adopt the vector notation $y^n \triangleq (y_1, \ldots, y_n)$ which applies to both deterministic and random vectors. We use standard asymptotic notations: For any sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, we write $a_n \gtrsim b_n$ if $a_n \geq c b_n$ holds for all n and some absolute constant $c > 0$, $a_n \lesssim b_n$ if $a_n \gtrsim b_n$, and $a_n \asymp b_n$ if both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold; the notations $O(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$ are similarly defined. We write $a_n = o(b_n)$ or $b_n = \omega(b_n)$ or $a_n \ll b_n$ or $b_n \gg a_n$ if $a_n / b_n \to 0$ as $n \to \infty$.

1.4 Organization

The rest of the paper is organized as follows. Section 2 discusses the basic reduction for EB regret and introduces a general program for proving regret lower bounds. In Section 3 we apply this program to show the lower bounds in Theorems 1 and 2 for both the Gaussian and Poisson model, making use of the analytical tools developed in Appendices A.1 and B. Section 4 concludes the paper with a discussion of related problems and open questions. Regret upper bound for Robbins estimator is shown in Appendix C. With most of the paper confusing on the EB setting, proofs for the compound setting is deferred to Appendix D. As a side application, the construction in the present paper can be used to obtain improved or new lower bounds for *mixture density estimation*; Appendix E collects these results.

2 Preliminaries

2.1 Reduction to estimating regression functions

In this subsection we describe the basic reduction relating the total regret (6) in the empirical Bayes setting to the problem of estimating a regression function, the latter of which is closely related to the "density estimation" problem in the Gaussian or Poisson mixture model. In turn, the rest of our paper proceeds in the spirit of proving the lower bounds on the density estimation, namely by building up a large collection of densities and applying the Assouad's lemma to show estimation lower bound; cf. Section 2.3 for details.

First, in addition to the total regret (6) we define the individual regret.

$$\mathsf{Regret}_n(\mathcal{G}) \triangleq \inf_{\widehat{\theta}} \sup_{G \in \mathcal{G}} \left\{ \mathbb{E}_G \left[(\widehat{\theta}(Y_1, \ldots, Y_n) - \theta_n)^2 \right] - \mathrm{mmse}(G) \right\} . \tag{15}$$

Here $Y_1, \ldots, Y_{n-1}$ can be viewed as training data from which we learn an estimator and apply it on a fresh (unseen) data point Y_n in order to predict θ_n. For example, for the Robbins estimator $\{\widetilde{\theta}_j\}$ in (5), the learned estimator $\widehat{\theta}_{\mathrm{Robbins}}(\cdot) = \widehat{\theta}_{\mathrm{Robbins}}(\cdot; Y_1, \ldots, Y_{n-1})$ is given by

$$\widehat{\theta}_{\mathrm{Robbins}}(y) = (y+1) \frac{N_{n-1}(y+1)}{N_{n-1}(y) + 1}$$

where $N_{n-1}(y) = \sum_{i=1}^{n-1} 1\{Y_i = y\}$ is the number of occurrences of y in the first $n-1$ observations, so that $\widetilde{\theta}_n = \widehat{\theta}_{\text{Robbins}}(Y_n; Y_1, \ldots, Y_{n-1})$. Note that the extra "add one" in the denominator is reminiscent of the Laplace estimator (cf. [CT06, Sec. 13.2]); this additional smoothing ensures a nonzero denominator which is clearly needed for achieving a non-trivial regret. Similar regularization has also been applied in the Gaussian model to the maximum likelihood method [JZ09].

There is a simple relation between $\mathsf{TotRegret}$ and Regret:

Lemma 5.

$$\mathsf{Regret}_n(\mathcal{G}) = \frac{1}{n}\mathsf{TotRegret}_n(\mathcal{G}).$$

Proof. Given any $\widehat{\theta}$ for (15), define $\widetilde{\theta}^n = (\widetilde{\theta}_1, \ldots, \widetilde{\theta}_n)$ for (6) by $\widetilde{\theta}_i = \widehat{\theta}(Y_{\setminus i}, Y_i)$. By symmetry, $\mathbb{E}[(\widetilde{\theta}_i - \theta_i)^2] = \mathbb{E}[(\widehat{\theta} - \theta_n)^2]$ for all i. This shows that $\mathsf{TotRegret}_n(\mathcal{G}) \le n \cdot \mathsf{Regret}_n(\mathcal{G})$.

The other direction is less obvious. Given any $\widehat{\theta}^n = (\widehat{\theta}_1, \ldots, \widehat{\theta}_n)$ for (6), define a *randomized* estimator $\widetilde{\theta}$ for (15) as follows by setting $\widetilde{\theta} = \widehat{\theta}_i(Y_1, \ldots, Y_{i-1}, Y_n, Y_{i+1}, \ldots, Y_{n-1}, Y_i)$ with probability $\frac{1}{n}$ for each $i \in [n]$. In other words, we randomly select i uniformly at random from $[n]$, then swap Y_i and Y_n in the sample and apply the strategy $\widehat{\theta}_i$. Then by exchangeability,

$$\mathbb{E}[(\widetilde{\theta}-\theta_n)^2] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(\widehat{\theta}_i(Y_1,\ldots,Y_n,\ldots,Y_i)-Y_n)^2] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(\widehat{\theta}_i(Y_1,\ldots,Y_n)-Y_i)^2] = \frac{1}{n}\mathbb{E}[\|\widehat{\theta}^n-\theta^n\|^2].$$

This shows that $\mathsf{Regret}_n(\mathcal{G}) \le \frac{1}{n}\mathsf{TotRegret}_n(\mathcal{G})$. □

By the orthogonality principle, the average quadratic risk of a given estimator $\widehat{\theta}$ can be decomposed as

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

Thus the regret can be viewed the quadratic loss of estimating the Bayes rule as a function, with the extra twist that the error is averaged with respect to the mixture distribution P_G that depends on the underlying G. In other words,

$$\mathsf{Regret}_n(\mathcal{G}) = \inf_{\widehat{\theta}} \sup_{G\in\mathcal{G}} \mathbb{E}_G\left[\left(\widehat{\theta}(Y_1,\ldots,Y_n) - \widehat{\theta}_G(Y_n)\right)^2\right] \tag{16}$$

$$= \inf_{\widehat{\theta}} \sup_{G\in\mathcal{G}} \mathbb{E}_G\left[\|\widehat{\theta} - \widehat{\theta}_G\|^2_{L_2(f_G)}\right] \tag{17}$$

where in (17) $\widehat{\theta}(\cdot)$ is understood as $\widehat{\theta}(Y_1, \ldots, Y_{n-1}, \cdot)$. Note a conceptual shift: with these reductions we can think of $\frac{1}{n}\mathsf{TotRegret}_n$ as a measure of the quality of estimation of the regression function after receiving $n-1$ training samples.

2.2 Truncation of priors

Frequently we are interested in priors with compact support, which can be unwieldy to use in the lower bound construction. Next we provide a general truncation lemma that translates regret lower bound for priors with uniform tail bound (e.g. subgaussian or subexponential) to that for compactly supported priors:

Lemma 6. *Given $a > 0$, let $\mathcal{G} = \mathcal{P}([-a, a])$ denote all probability measures supported on the interval $[-a, a]$. Let $\mathcal{G}'$ be a collection of priors on $\mathbb{R}$, such that $\sup_{G\in\mathcal{G}'} G([-a,a]^c) \le \epsilon \le \frac{1}{2}$ for some $\epsilon = \epsilon(a)$, and $\sup_{G\in\mathcal{G}'} \mathbb{E}_G[\theta^4] \le M$. Then*

$$\mathsf{Regret}_n(\mathcal{G}) \ge \mathsf{Regret}_n(\mathcal{G}') - 6\sqrt{(M + a^4)n\epsilon}. \tag{18}$$

Proof. Define $E \triangleq \{|\theta_i| \le a, i = 1, \dots, n\}$ and notice that for any estimator $\widehat{\theta}$ taking values in $[-a, a]$ and any prior $G \in \mathcal{G}'$ we have

$$\begin{aligned}
\mathbb{E}_G[(\widehat{\theta} - \theta_n)^2] &= \mathbb{E}_G[(\widehat{\theta} - \theta_n)^2 \mathbf{1}_E] + \mathbb{E}_G[(\widehat{\theta} - \theta_n)^2 \mathbf{1}_{E^c}] \\
&\le \mathbb{E}_G[(\widehat{\theta} - \theta_n)^2 \mid E] + \sqrt{\mathbb{E}_G[(\widehat{\theta} - \theta_n)^4] \cdot \mathbb{P}_G[E^c]} \\
&\le \mathbb{E}_G[(\widehat{\theta} - \theta_n)^2 \mid E] + \sqrt{8(M + a^4) \cdot n\epsilon}. \qquad (19)
\end{aligned}$$

For any G, let G_a denote its restriction on $[-a, a]$, i.e., $G_a(A) = \frac{G(A \cap [-a,a])}{G([-a,a])}$. Then, we get the following chain of inequalities:

$$\begin{aligned}
\mathsf{Regret}_n(\mathcal{G}) &= \inf_{|\widehat{\theta}| \le a} \sup_{G \in \mathcal{G}} \mathbb{E}_G \left[(\widehat{\theta}(Y_1, \dots, Y_n) - \theta_n)^2 \right] - \mathrm{mmse}(G) \\
&\ge \inf_{|\widehat{\theta}| \le a} \sup_{G \in \mathcal{G}'} \mathbb{E}_{G_a} \left[(\widehat{\theta}(Y_1, \dots, Y_n) - \theta_n)^2 \right] - \mathrm{mmse}(G_a) \\
&= \inf_{|\widehat{\theta}| \le a} \sup_{G \in \mathcal{G}'} \mathbb{E}_G \left[(\widehat{\theta}(Y_1, \dots, Y_n) - \theta_n)^2 \Big| E \right] - \mathrm{mmse}(G_a) \\
&\overset{(a)}{\ge} \inf_{|\widehat{\theta}| \le a} \sup_{G \in \mathcal{G}'} \mathbb{E}_G \left[(\widehat{\theta}(Y_1, \dots, Y_n) - \theta_n)^2 \Big| E \right] - \frac{1}{1-\epsilon} \mathrm{mmse}(G) \\
&\overset{(b)}{\ge} \inf_{|\widehat{\theta}| \le a} \sup_{G \in \mathcal{G}'} \mathbb{E}_G \left[(\widehat{\theta}(Y_1, \dots, Y_n) - \theta_n)^2 \right] - \sqrt{8(M + a^4)(1 - (1-\epsilon)^n)} - \frac{1}{1-\epsilon} \mathrm{mmse}(G) \\
&\overset{(c)}{\ge} \mathsf{Regret}_n(\mathcal{G}') - \sqrt{8(M + a^4) n\epsilon} - 2\epsilon\sqrt{M},
\end{aligned}$$

where (a) follows from the simple fact (cf. e.g. [WV12, Lemma 2])

$$\mathrm{mmse}(G_a) \le \frac{1}{G([-a, a])} \mathrm{mmse}(G); \qquad (20)$$

(b) is from (19); (c) follows from $\mathrm{mmse}(G) \le \mathbb{E}_G[\theta^2] \le \sqrt{M}$ and $\frac{1}{1-\epsilon} \le 1 + 2\epsilon$ due to $\epsilon \le 1/2$. □

2.3 General program for regret lower bound

We describe our general method for proving regret lower bounds next. Select a prior G_0 and let $f_0 \equiv f_{G_0} = \int f_\theta G_0(d\theta)$ denote the induced mixture density. We aim to prove regret bound by restricting to priors whose likelihood ratio relative to G_0 is bounded from above and below. Specifically, we will construct a collection of perturbations around G_0 whose induced perturbations on the regression function satisfy an approximate orthogonality property as required by Assouad's lemma. To describe this program, let us first introduce the necessary notations.

Define an operator K that maps a bounded measurable function r to Kr given by

$$Kr(y) \triangleq \mathbb{E}_{G_0}[r(\theta) | Y = y] = \frac{\int r(\theta) f_\theta(y) G_0(d\theta)}{f_0(y)}. \qquad (21)$$

In other words, K is an integral operator with kernel function given by the posterior density. Notice that for the prior G defined by $dG = r(\theta) dG_0$, the corresponding mixture density is given by

$$f_G(y) = f_0(y) \cdot Kr(y). \qquad (22)$$

Next, fix an arbitrary bounded function r and for sufficiently small δ, define the distribution G_δ by

$$dG_\delta \triangleq \frac{1}{1+\delta\int r dG_0}(1+\delta r)dG_0\,.$$

Denote by

$$K\theta(y) = \mathbb{E}_{G_0}[\theta|Y=y] \tag{23}$$

to be the result of applying K to the identity function $\theta \mapsto \theta$. By the Bayes rule and using (22), the regression function corresponding to G_δ satisfies:

$$\begin{aligned}
\widehat{\theta}_{G_\delta}(y) \equiv \mathbb{E}_{G_\delta}[\theta|Y=y] &= \mathbb{E}_{G_0}\left[\theta\frac{1+\delta r(\theta)}{1+\delta Kr(y)}\middle| Y=y\right] \\
&= \frac{K\theta+\delta K(\theta r)}{1+\delta Kr}(y) \\
&= K\theta(y) + \delta\frac{K(\theta r)-(K\theta)(Kr)}{1+\delta Kr}(y) \\
&= K\theta(y) + \delta K_1 r(y) + \delta^2\frac{1}{1+\delta Kr(y)}(Kr)(y)\cdot(K_1 r)(y)\,,
\end{aligned} \tag{24}$$

where we introduced an operator

$$K_1 r \triangleq K(\theta r) - (K\theta)(Kr)\,. \tag{25}$$

Note that the mapping from r to the regression function $\widehat{\theta}_{G_\delta}$ is nonlinear, but can be expressed in terms of the linear operators K and K_1; these two operators play crucial roles in the lower bound construction. In applications below it will often be that K is a convolution-like operator, while K_1 will be a differentiation operator (followed by convolution). The emergence of differentiation may appear surprising at first, so it may be satisfying to notice from (24) an alternative definition of K_1 emphasizing the differentiation explicitly:

$$K_1 r(y) = \frac{d}{d\delta}\bigg|_{\delta=0}\mathbb{E}_{G_\delta}[\theta|Y=y]\,.$$

The main method for proving our lower bounds is contained in the following proposition. We note that similar strategy based on Assouad's lemma has been used before in [Kim14,KG20] to show lower bound for estimating Gaussian mixture density. Here the construction is more challenging in view of the non-linear relationship between the regression function and the mixture density (see (3) and (4) for example) which requires working with the kernel K and K_1 defined above. The dependency of the regression function on "high-order" information of the mixture density (such as its derivatives) suggests the rate for individual regret may be strictly slower than the density estimation rate. This turns out to be the case for the Poisson model [JPW21].

Proposition 7. *Fix a prior distribution G_0, constants $a, \tau, \tau_1, \tau_2, \gamma \geq 0$ and m real-valued functions $r_1, \ldots, r_m$ on Θ with the following properties.*

1. *For each i, $\|r_i\|_\infty \leq a$;*
2. *For each i, $\|Kr_i\|_{L_2(f_0)} \leq \sqrt{\gamma}$;*
3. *For any $v \in \{0, \pm 1\}^m$,*

$$\left\|\sum_{i=1}^m v_i K_1 r_i\right\|^2_{L_2(f_0)} \geq \tau\|v\|_2^2 - \tau_2 \tag{26}$$

4. For any $v \in \{0,1\}^m$, $\|\sum_{i=1}^m v_i K_1 r_i\|_{L_2(f_0)}^2 \le \tau_1^2 m$;

Then the regret over the class of priors $\mathcal{G} = \{G : |dG/dG_0 - 1| \le \frac{1}{2}\}$ satisfies

$$\mathsf{Regret}_n(\mathcal{G}) \ge C\delta^2(m(4\tau - \tau_1^2) - \tau_2)\,, \qquad \delta \triangleq \frac{1}{\max(\sqrt{n\gamma}, ma)}\,, \tag{27}$$

where $C > 0$ is an absolute constant.

Remark 1. Note that the lower bound in (27) is invariant to rescaling all r_i's by a common factor. Thus in applications, it is convenient to fix one of the parameters. For example, we will always choose $\tau = 1$. In fact, we will select r_i's so that they are *orthonormal* after the action of the kernel K_1, i.e.

$$(K_1 r_i, K_1 r_j)_{L_2(f_0)} = 1\{i \ne j\},$$

so that $\tau_1 = 1$ and $\tau_2 = 0$. In this case, the regret lower bound reads

$$\mathsf{Regret}_n \gtrsim \min\left\{\frac{m}{n\gamma}, \frac{1}{ma^2}\right\}, \tag{28}$$

which, in later applications, further simplifies to $\mathsf{Regret}_n \gtrsim \frac{m}{n\gamma}$.

Proof. Throughout the proof, $C_0, C_1, \ldots$ denote absolute constants. Let $\mu_i \triangleq \int r_i dG_0$, which satisfies $|\mu_i| \le a$ by the first assumption. For each binary string $u = (u_1, \ldots, u_m) \in \{0,1\}^m$, denote

$$r_u \triangleq \sum_{i=1}^m u_i r_i, \quad h_u \triangleq K r_u, \quad \mu_u \triangleq \sum_{i=1}^m u_i \mu_i. \tag{29}$$

We define 2^m distributions indexed by $u \in \{0,1\}^m$ as

$$dG_u \triangleq \frac{1 + \delta r_u}{1 + \delta \mu_u} dG_0, \tag{30}$$

where $\delta > 0$ is chosen so that

$$\delta m a \le \frac{1}{16}\,. \tag{31}$$

Then

$$f_u \triangleq \frac{(1 + \delta h_u) f_0}{1 + \delta \mu_u}.$$

is the mixture density induced by the prior G_u. Let $\widetilde{\mathcal{G}} = \{G_u : u \in \{0,1\}^m\}$. By (31), each G_u satisfies $\frac{1}{2} \le \frac{dG_u}{dG_0} \le \frac{3}{2}$ and hence $\widetilde{\mathcal{G}} \subset \mathcal{G}$. Abbreviate $T_u(y) \triangleq \mathbb{E}_{G_u}[\theta | Y = y]$. Then we have:

$$\begin{aligned}
\mathsf{Regret}_n(\mathcal{G}) \ge \mathsf{Regret}_n(\widetilde{\mathcal{G}}) &\overset{\text{(a)}}{=} \inf_{\widehat{T}} \sup_{u \in \{0,1\}^m} \mathbb{E}_{G_u}\left[\|\widehat{T} - T_u\|_{L^2(f_u)}^2\right] \\
&\overset{\text{(b)}}{\ge} \frac{1}{2} \inf_{\widehat{T}} \sup_{u \in \{0,1\}^m} \mathbb{E}_{G_u}\left[\|\widehat{T} - T_u\|_{L^2(f_0)}^2\right] \\
&\overset{\text{(c)}}{\ge} \frac{1}{8} \inf_{\widehat{u} \in \{0,1\}^m} \sup_{u \in \{0,1\}^m} \mathbb{E}_{G_u}\left[\|T_{\widehat{u}} - T_u\|_{L^2(f_0)}^2\right]
\end{aligned} \tag{32}$$

where (a) follows from the representation (17) for individual regret; in (b) we used the fact that $\frac{f_u}{f_0} \ge \frac{1}{2}$ to change from $L_2(f_u)$ to $L_2(f_0)$; (c) follows from the usual argument for restricting to

proper estimators, namely, for any $\widehat{T}$, $\widehat{u} = \mathrm{argmin}_{v\in\{0,1\}^m} \|\widehat{T} - T_v\|_{L^2(f_0)}$ satisfies that, for any u, $\|\widehat{T} - T_u\|_{L^2(f_0)} \le 2\|T_{\widehat{u}} - T_u\|_{L^2(f_0)}$. In the remainder of this proof all $\|\cdot\|_2$ norms are with respect to $L_2(f_0)$.

In order to apply Assouad's lemma, we next show that we have approximate orthogonality, i.e. for some ϵ and ϵ_1 (to be specified) and all $u, v \in \{0,1\}^m$ we have

$$\|T_u - T_v\|_2^2 \ge \epsilon d_{\mathrm{H}}(u,v) - \epsilon_1, \tag{33}$$

where $d_{\mathrm{H}}(u,v) \triangleq \sum |u_i - v_i|$ stands for the Hamming distance. Indeed, by (24), the regression function T_u is given by

$$T_u = K\theta + \delta K_1 r + \delta^2 \frac{h_u}{1+\delta h_u} K_1 r. \tag{34}$$

(Recall that $K\theta$ is understood as in the sense of (23).) Since K is a conditional expectation operator, we have $\|h_u\|_\infty \le \|r_u\|_\infty \le ma$. Thus together with the condition (31), which implies $1 + \delta h_u \ge 1/2$, we have the following estimate on the last term in (34):

$$\left\| \delta^2 \frac{h_u}{1+\delta h_u} K_1 r_u \right\|_2 \le 2\delta^2 ma \|K_1 r_u\|_2 \overset{\text{(a)}}{\le} 2\delta^2 m^{3/2} a \tau_1 \overset{\text{(b)}}{\le} \frac{1}{8}\delta\sqrt{m}\tau_1\,,$$

where (a) follows from the fourth assumption of the lemma, and in (b) we also used (31). Then from (34) and the triangle inequality we have

$$\|T_u - T_v\|_2 \ge \delta \|K_1(r_u - r_v)\|_2 - \frac{\delta\sqrt{m}\tau_1}{4}\,.$$

Squaring this, applying $(a-b)^2 \ge \frac{1}{2}a^2 - b^2$ we obtain $\|T_u - T_v\|_2^2 \ge \frac{1}{2}\delta^2 \|K_1(r_u - r_v)\|_2^2 - \frac{1}{16}\delta^2 m\tau_1^2$. Finally, applying the assumption (26) we obtain (33) with

$$\epsilon = \frac{1}{2}\tau\delta^2, \quad \epsilon_1 = \delta^2\left(\frac{1}{16}m\tau_1^2 + \frac{1}{2}\tau_2\right)$$

Next, consider the following identity

$$\frac{f_u}{f_0} - \frac{f_v}{f_0} = \frac{1+\delta h_u}{1+\delta\mu_u} - \frac{1+\delta h_v}{1+\delta\mu_v} = \delta\frac{h_u - h_v}{1+\delta\mu_u} + (1+\delta h_v)\frac{\delta(\mu_v - \mu_u)}{(1+\delta\mu_u)(1+\delta\mu_v)}\,.$$

Because of (31) we have $1+\delta h_v, 1+\delta\mu_u, 1+\delta\mu_v \in [\frac{1}{2}, \frac{3}{2}]$. Thus the χ^2-divergence from f_u to f_v is given by

$$\begin{aligned}
\chi^2(f_u\|f_v) &\triangleq \int d\mu f_0 \frac{(f_u/f_0 - f_v/f_0)^2}{f_v/f_0} \\
&= \delta^2 \int d\mu f_0 \frac{1+\delta\mu_v}{1+\delta h_v}\left(\frac{h_u - h_v}{1+\delta\mu_u} + \frac{\mu_v - \mu_u}{(1+\delta\mu_u)(1+\delta\mu_v)}\right)^2 \\
&\le C_0\delta^2\|h_u - h_v\|_2^2 + \delta^2(\mu_v - \mu_u)^2 \\
&\le C_1\delta^2\|h_u - h_v\|_2^2\,.
\end{aligned}$$

where the last step follows from, by Cauchy-Schwarz,

$$\mu_u - \mu_v = \int dG_0(r_u - r_v) = \int d\mu f_0 (h_u - h_v) \le \|h_u - h_v\|_2.$$

For any u and v such that $d_{\mathrm{H}}(u,v)=1$, $\|h_u-h_v\|_2^2 \le \gamma$ by the second assumption. Recall that the χ^2-divergence between n-fold product distributions is given by $\chi^2(f_u^{\otimes n}\|f_v^{\otimes n}) = (1+\chi^2(f_u\|f_v))^n-1$. We have

$$\chi^2(f_u^{\otimes n}\|f_v^{\otimes n}) \le C_2 \qquad \forall u,v : d_{\mathrm{H}}(u,v)=1\,, \tag{35}$$

provided that

$$n\delta^2\gamma \le 1 \tag{36}$$

Applying Assouad's lemma (see, e.g., [Tsy09, Theorem 2.12(iv)]), we have

$$\inf_{\widehat{u}\in\{0,1\}^m} \sup_{u\in\{0,1\}^m} \mathbb{E}_{G_u}\left[d_{\mathrm{H}}(\widehat{u},u)\right] \ge C_3 m.$$

Selecting the largest δ subject to (31) and (36), the proof is completed in view of (32) and (33). □

In proving our main results in both the Gaussian and Poisson model, we follow essentially the same high-level strategy: To apply Proposition 7, we need to find many functions $r_1,\dots,r_m$ which are mapped by the operator K into orthogonal ones. Thus, it is natural to take r_k to be the singular functions of K, i.e. those that diagonalize the self-adjoint operator K^*K. Because of the general relation between K_1 and K in (25), it will turn out that r_k's indexed by *sufficiently separated* k's are also mapped to orthogonal functions by the operator K_1, thus yielding a large subcollection of functions for Proposition 7 with $\tau=\tau_1=1$ and $\tau_2=0$. For the Gaussian model, we exactly follow this strategy by identifying K^*K as a Mehler kernel and working with its eigenfunctions (Hermite functions). For the Poisson model, it may be difficult to compute the eigenfunctions of the resulting kernel; instead, we directly construct functions that satisfy the desired orthogonality. We proceed to details.

3 Proofs of the main results

3.1 Normal mean model

In this section we prove Theorem 1. Choose the prior $G_0=\mathcal{N}(0,s)$ for $s>0$ to be specified later. Let $\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ be the standard normal density. Then $P_{G_0}=\mathcal{N}(0,1+s)$ with density $f_0(y)=\frac{1}{\sqrt{1+s}}\varphi\left(\frac{y}{\sqrt{1+s}}\right)$. Furthermore, the operators K and K_1 in (21) and (25) are given explicitly as follows:

Proposition 8.

$$Kr(y) = (r(\eta\cdot)*\varphi)(\eta y), \qquad \eta \triangleq \sqrt{\frac{s}{s+1}} \tag{37}$$

$$K_1 r(y) = \eta^2 K(r')(y)\,, \tag{38}$$

where in (37) $*$ *denotes convolution and* (38) *holds for all* r *with bounded derivative.*

Proof. To show (37), by definition we have

$$Kr(y) = \frac{1}{\eta}\int_{\mathbb{R}} dx \frac{\varphi(y-x)\varphi(\frac{x}{\sqrt{s}})}{\varphi(\frac{y}{\sqrt{1+s}})} r(x)\,.$$

Changing the integration variable to $x_1 = x/\eta$ and noticing that $\frac{\varphi(y-x)\varphi(\frac{x}{\sqrt{s}})}{\varphi(\frac{y}{\sqrt{1+s}})} = \varphi(x_1 - \eta y)$ completes the proof of (37). We note that trivially then

$$(Kx)(y) = \eta^2 y\,, \tag{39}$$

which also follows from the expression of the condition mean with a Gaussian prior. To prove (38), we first notice that it is sufficient to show that for an arbitrary (possibly non-differentiable) bounded r we have

$$K_1 r(y) = \frac{d}{dy} Kr(y)\,. \tag{40}$$

Indeed, from here for differentiable r we can simply apply derivative to the first term in the convolution in (37) and get (38). To show (40) we take derivative on the φ term of the convolution and using $\varphi'(z) = -z\varphi(z)$ to get

$$\frac{d}{dy} Kr(y) = \eta(r(\eta\cdot) * \varphi')(\eta y) = \eta \int dx r(\eta x)(x - \eta y)\varphi(x - \eta y) = (K(xr) - KxKr)(y)\,,$$

where we used (39) for the second term. This proves (40) by the definition of K_1 operator. □

In view of (38) we see that to apply Proposition 7 we need to be able to compute inner products of the form $(Kr_1, Kr_2)_{L_2(f_0)}$. To simplify these computations, we first extend (via (37)) the domain of K to $L_2(\mathbb{R}, \mathrm{Leb})$ and then introduce the self-adjoint operator $S = K^*K : L_2(\mathbb{R}, \mathrm{Leb}) \to L_2(\mathbb{R}, \mathrm{Leb})$, so that for any $f, g \in L_2(\mathrm{Leb})$ we have

$$(Kf, Kg)_{L_2(f_0)} = (Sf, g)_{L_2(\mathrm{Leb})}\,. \tag{41}$$

An explicit computation shows that S is in fact a Mehler kernel operator, i.e.

$$Sf(x) = \int f(\widetilde{x}) S(x, \widetilde{x}) d\widetilde{x}, \tag{42}$$

with (in view of (37)):

$$S(x, \widetilde{x}) = \frac{1}{\eta^2} \int_{-\infty}^{\infty} \varphi(x/\eta - \eta y)\varphi(\widetilde{x}/\eta - \eta y) f_0(y)\, dy = \lambda_1 e^{-\frac{\lambda_2}{2}(x^2 + \widetilde{x}^2 - 2\rho x\widetilde{x})}\,, \tag{43}$$

where

$$\lambda_1 = \frac{1}{2\pi s} \frac{1+s}{\sqrt{1+2s}}, \qquad \lambda_2 = \frac{(1+s)^2}{s(1+2s)}\,, \qquad \rho = \frac{s}{1+s}\,. \tag{44}$$

It is well-known that the Mehler kernel is diagonalized by the Hermite functions [Meh66]. We collect all the relevant technical details in the following lemma proved in Appendix A.1.

Lemma 9. *For every $s > 0$, there is an orthonormal basis $\{\psi_k : k = 0, 1, \ldots\}$ in $L_2(\mathrm{Leb})$ consisting of eigenfunctions of the operator S, satisfying the following:*

$$(K\psi_k, K\psi_n)_{L_2(f_0)} = \lambda_0 \mu^k 1\{k = n\} \tag{45}$$

$$(K_1\psi_k, K_1\psi_n)_{L_2(f_0)} = \begin{cases} 0, & |k-n| \geq 3\,, \\ \lambda_3(k\mu^{k-1} + (k+1)\mu^{k+1}), & k = n\,, \end{cases} \tag{46}$$

$$\|\psi_k\|_\infty \leq \sqrt{\alpha_1}\,, \tag{47}$$

where $\mu, \lambda_0, \lambda_3, \alpha_1$ are positive constants depending only on s, which, for $0 < s < 1/2$, satisfy

$$\mu \asymp s,\ \lambda_0 \asymp \frac{1}{\sqrt{s}}\,, \lambda_3 \asymp \sqrt{s}, \alpha_1 \asymp \frac{1}{\sqrt{s}}\,. \tag{48}$$

In addition, for all $s > 0$ we have $0 < \mu < 1$.

Proof of Theorem 1. In view of Lemma 5, it is equivalent to consider the individual regret Regret_n. We first consider the s-subgaussian case. Without loss of generality, assume that $s < 1/2$. By applying Proposition 7, we aim to show the following lower bound

$$\mathsf{Regret}_n(\mathsf{SubG}(2s)) \geq \frac{c}{n}\frac{\log n}{\log \frac{1}{s}}$$

for some universal constant c. The case of compact support then follows from taking $s \asymp \frac{1}{\log n}$ and truncation (Lemma 6).

Let $\mathcal{G}'_s$ be the set of distributions $\mathcal{G}$ in Proposition 7, namely, $\mathcal{G}'_s = \{G : 1/2 \leq \frac{dG}{dG_0} \leq 3/2\}$, where $G_0 = \mathcal{N}(0, s)$. It is straightforward to verify that $\mathcal{G}'_s \subset \mathsf{SubG}(2s)$ so we focus on lower bounding $\mathsf{Regret}_n(\mathcal{G}'_s)$. We select the required m perturbations as follows:

$$r_j = \xi_{m+3j}\psi_{m+3j}\,, \qquad j = 1, \ldots, m\,,$$

where

$$\xi_i \triangleq \frac{1}{\sqrt{(K_1\psi_i, K_1\psi_i)}}\,.$$

In view of Lemma 9, applying Proposition 7 with $\tau = \tau_1 = 1$ and $\tau_2 = 0$ yields

$$\mathsf{Regret}_n(\mathcal{G}'_s) \gtrsim \frac{m}{\max(n\gamma, m^2a^2)}\,, \tag{49}$$

where we can take

$$a = \max_{m \leq i \leq 4m} \xi_i \|\psi_i\|_\infty,\ \gamma = \max_{m \leq i \leq 4m} \xi_i^2 \|K\psi_i\|^2_{L_2(f_0)}\,.$$

By Lemma 9, we have

$$\xi_i^2 = \frac{1}{\lambda_3(i\mu^{i-1} + (i+1)\mu^{i+1})} \asymp \frac{1}{\sqrt{s}m\mu^{i-1}}\,.$$

Together with $\|K\psi_i\|^2 \asymp \frac{1}{\sqrt{s}}\mu^i$ we get

$$\gamma \asymp \frac{1}{m}. \tag{50}$$

Similarly, we have

$$a^2 = \max_{m \leq i \leq 4m} (\xi_i)^2 \|\psi_i\|^2_\infty \lesssim \frac{1}{sm\mu^{4m-1}} \tag{51}$$

Finally, we select m so that $n\gamma > m^2a^2$. In view of the previous estimates, this requirement is equivalent to $\frac{n}{m} \gtrsim \frac{m}{s}\mu^{1-4m} \asymp m\mu^{-4m}$. It is clear that taking $m = c\frac{\log n}{\log \frac{1}{\mu}} \asymp \frac{\log n}{\log \frac{1}{s}}$ for sufficiently small constant c fulfills this requirement. In all, from (49) we obtain the desired bound

$$\mathsf{Regret}_n(\mathcal{G}'_s) \gtrsim \frac{1}{n}\frac{\log n}{\log \frac{1}{s}}\,, \tag{52}$$

where the hidden constant is uniform in $0 < s < 1/2$.

Finally, we prove the lower bound (7) for priors supported on $[-h, h]$. We choose $s = \frac{c}{\log n}$ in (52) with $c > 0$ to be specified shortly. Recall that every $G \in \mathcal{G}'_s$ satisfies $\frac{dG}{dG_0} \leq 3/2$, where $G_0 = \mathcal{N}(0, s)$. Thus, for every G we have $\mathbb{E}_{\theta\sim G}[\theta^4] \lesssim s^4$, while $G[|\theta| > h] \leq 3e^{-\frac{h^2}{2s}}$. Applying Lemma 6 we obtain then

$$\mathsf{Regret}_n(\mathcal{P}([-h, h])) \geq \mathsf{Regret}_n(\mathcal{G}'_s) - O(1)\sqrt{ne^{-\frac{h^2}{2s}}s^4}\,.$$

By selecting c sufficiently small depending on h, we can ensure the second term is at most $O(n^{-2})$ so that (52) implies the required lower bound of $\Omega_h(\frac{1}{n}(\frac{\log n}{\log\log n})^2)$. □

Remark 2 (Optimality of the construction)**.** Consider $s = 1$. Can we get a lower bound better than $\Omega(\frac{\log^2 n}{n})$ by selecting other functions r_i (and still with the same base prior G_0) in Proposition 7? We will argue that the answer is negative, at least if the latter proposition is applied with $\tau = \tau_1 = 1$ and $\tau_2 = 0$ as we did above. Indeed, suppose we found m, a, γ such that there are functions $r_1, \dots, r_m$ such that $\|r_i\|_\infty \le a$, $(K_1 r_i, K_1 r_j)_{L_2(f_0)} = 1\{i \neq j\}$, and $\|K r_i\|^2_{L_2(f_0)} \le \gamma$. Then Proposition 7 yields the regret lower bound (28), namely, $\mathsf{Regret}_n \gtrsim \min\{\frac{m}{n\gamma}, \frac{1}{ma^2}\}$. We will show in Appendix A.2 that for any such collection of functions it must be that $m \lesssim \log a$ and $\gamma \gtrsim \frac{1}{\log a}$. This implies that the resulting lower bound is at most on the order of $\min\{\frac{\log^2 a}{n}, \frac{1}{ma^2}\} \lesssim \min\{\frac{\log^a}{n}, \frac{1}{ma^2}\} \lesssim \frac{\log^2 n}{n}$, and therefore, to be interesting we need to have $a^2 \lesssim \frac{n}{\log^2 n}$, or $\log a \lesssim \log n$.

Remark 3 (Metric properties of Gaussian convolution)**.** We can rephrase the content of Lemma 9 and the previous remark also as an interesting analytical fact. Consider a convolution operator $T : L_\infty \to L_2(\mathcal{N}(0,1))$ defined as

$$Tf = f * \mathcal{N}(0, \lambda)\,.$$

We want to understand the largest number of orthonormal functions that one can find in the image under T of an L_∞-ball $B_\infty(0, a) = \{f : \|f\|_\infty \le a\}$ as $a \to \infty$. As an attempt, one may try working with sinusoids, which are eigenfunctions of the convolution; however, selecting appropriately normalized sinusoids only yields $\Theta(\sqrt{\log a})$ such functions. At the same time, Lemma 9 gives $\Theta(\log a)$ functions, which, as shown in Remark 2, is in fact optimal.

3.2 Poisson model

Consider $Y \sim \mathrm{Poi}(\theta)$, where θ takes values in $\mathbb{R}_+$ and Y takes values in $\mathbb{Z}_+$. As before, let G_0 denote a prior on θ and f_0 the corresponding probability mass function (PMF) for Y. In order to have tractable expressions for K and K_1 it seems natural to select G_0 to be a conjugate prior, so we record some observations:

- G_0 is Gamma(α, β), with density

$$G_0(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \qquad x \in \mathbb{R}_+\,. \tag{53}$$

- f_0 is negative binomial, with PMF

$$f_0(y) = \binom{y+\alpha-1}{y} \left(\frac{\beta}{1+\beta}\right)^\alpha \left(\frac{1}{1+\beta}\right)^y \qquad y \in \mathbb{Z}_+ \tag{54}$$

- The posterior distribution of θ given $Y = y$ is Gamma$(y+\alpha, \beta+1)$ and thus the operator (21) can be written as

$$Kr(y) = \int_0^\infty r(x) K(x, y) dx\,,$$

  with the kernel function explicitly being given as

$$K(x, y) = \frac{(1+\beta)^{y+\alpha+1}}{\Gamma(y+\alpha)} x^{y+\alpha-1} e^{-(1+\beta)x}\,. \tag{55}$$

In the special case of $\alpha = 1$, G_0 is an exponential distribution with density $g_0(x) = \beta e^{-\beta x}$ and the induced Poisson mixture is geometric distribution with PMF $f_0(y) = \frac{\beta}{1+\beta}(\frac{1}{1+\beta})^y$.

As before, we first derive a useful differential expression for the K_1 operator.

Proposition 10 (From K_1 to K). *Let $G_0 = \text{Gamma}(\alpha, \beta)$, $\alpha > 0$, then for all $r = r(x)$, smooth and bounded on $\mathbb{R}_+$, we have*

$$K_1 r = \frac{1}{1+\beta} K(xr')\,. \tag{56}$$

Conversely, suppose that for some G_0, which is smooth on $[0,\infty)$ and strictly positive on $(0,\infty)$, and all smooth bounded r on $\mathbb{R}_+$ we have

$$K_1 r = cK(xr') \tag{57}$$

for some constant c. Then $0 < c < 1$ and $G_0 = \text{Gamma}(\alpha, \frac{1}{c} - 1)$ for some $\alpha > 0$.

Proof. Fix an arbitrary smooth prior density G_0 strictly positive on $\mathbb{R}_+$, and denote by $f_0(y)$ the induced mixture PMF. The kernel of operator K is given by

$$K(x,y) = \frac{G_0(x)}{f_0(y)} e^{-x} \frac{x^y}{y!}\,.$$

We have then the following easily verified identities:

$$K(xg)(y) = \frac{f_0(y+1)}{f_0(y)}(y+1)Kg(y+1) \tag{58}$$

$$\partial_x K(x, y+1) = K(x,y)\frac{f_0(y)}{f_0(y+1)} w(x,y) \qquad w(x,y) \triangleq 1 + (\partial_x \ln G_0(x) - 1)\frac{x}{y+1} \tag{59}$$

$$f_0(y+1)K(x,y+1) = f_0(y)K(x,y)\frac{x}{y+1} \tag{60}$$

From (58) we have

$$K(xr')(y) = \frac{f_0(y+1)}{f_0(y)}(y+1)K(r')(y+1)\,.$$

On the other hand, since r is bounded, integrating by parts and noting that $K(0, y+1) = 0$ for all $y \geq 0$ we obtain[4]

$$K(r')(y+1) = \int_{\mathbb{R}_+} dx r'(x) K(x, y+1) = -\int_{\mathbb{R}_+} dx\, r(x) \partial_x K(x, y+1)\,.$$

Invoking (59) we overall get

$$K(xr')(y) = -(y+1)\int dx r(x) K(x,y) w(x,y)\, dx\,. \tag{61}$$

Next, we apply (58) to get

$$Kx(y) = \frac{(y+1)f_0(y+1)}{f_0(y)}\,.$$

Then, from (25) we obtain that

$$K_1 r(y) = K(xr)(y) - \frac{(y+1)f_0(y+1)}{f_0(y)} Kr(y)\,.$$

Applying (58) to the first term we get

$$K_1 r(y) = \frac{(y+1)f_0(y+1)}{f_0(y)}(Kr(y+1) - Kr(y))\,. \tag{62}$$

[4] $K(\infty, y+1) = 0$ may not hold. Is integration by parts OK? Should we also require G_0 to be bounded to be safe?

Via (61) and (62), we observe that (57) is equivalent to

$$\frac{(y+1)f_0(y+1)}{f_0(y)}\int_{\mathbb{R}_+} dx r(x)(K(x,y+1)-K(x,y)) = -c\int dx r(x)(y+1)K(x,y)w(x,y)\,,$$

which needs to hold for all $r(x)$ and $y \geq 0$. Thus, (57) is equivalent to

$$\frac{f_0(y+1)}{f_0(y)}(K(x,y+1)-K(x,y)) = -cw(x,y)K(x,y)\,.$$

Apply (60) to replace $K(x,y+1)$ with $K(x,y)$ to get:

$$\left(\frac{x}{y+1}-\frac{f_0(y+1)}{f_0(y)}\right)K(x,y) = -cw(x,y)K(x,y)\,.$$

Since $G_0(x) > 0$ for all $x > 0$, we also get that $K(x,y) > 0$ and, thus, overall (57) is equivalent to

$$x(1+c(\partial_x \ln G_0(x)-1)) = (y+1)\left(-c+\frac{f_0(y+1)}{f_0(y)}\right)\,. \tag{63}$$

From here we first notice that when $G_0 = \text{Gamma}(\alpha,\beta)$ and f_0 is the negative binomial both sides evaluate to $\frac{\alpha-1}{1+\beta}$ if $c = \frac{1}{1+\beta}$. This proves (56) and the first part of the proposition. For the second part, notice that in order for (63) to hold for all x, y both sides must be constants, implying that that $f_0(y)$ is a negative binomial PMF and, thus, the G_0 is a Gamma distribution. □

The converse part of Proposition 10 justifies yet again our focus on the Gamma prior G_0. In Appendix B we proceed by (almost) diagonalizing the self-adjoint compact integral operator K^*K. The result of this analysis is summarized in the following.

Lemma 11. *Fix $\delta > 0$. Let $G_0 = \text{Gamma}(\alpha,\beta)$. Then there exist absolute positive constants C, m_0 such that for all $m \geq m_0$, $\beta \geq 2$ and $\alpha \geq 4m$, there are functions $r_1,\ldots,r_m$ with the following properties:*

$$\|K_1 r_j\|^2_{L_2(f_0)} = 1\,, \tag{64}$$

$$(Kr_j, Kr_i)_{L_2(f_0)} = (K_1 r_j, K_1 r_i)_{L_2(f_0)} = 0 \quad \forall i \neq j \tag{65}$$

$$\|Kr_j\|^2_{L_2(f_0)} \leq \frac{C\beta}{\alpha m}\,, \tag{66}$$

$$\|r_j\|_\infty \leq \sqrt{\frac{\beta}{\alpha}}e^{C(m\log\beta+\alpha)}\,. \tag{67}$$

Lemma 12. *In the special case of $\alpha = 1$ (so that G_0 is the exponential distribution with mean $1/\beta$) for any fixed $\beta > 0$ there exists a constant $C = C(\beta) > 0$ such that for all $m \geq 1$ there exist functions $r_1,\ldots,r_m$ satisfying (64), (65) and*

$$\|Kr_j\|^2_{L_2(f_0)} \leq \frac{C}{m^2}\,, \tag{68}$$

$$\|r_j\|_\infty \leq e^{Cm}\,. \tag{69}$$

Proofs of Lemmas 11 and 12 are found in Appendix B. Assuming these lemmas we are in position to prove the lower bound in Theorem 2. (The upper bound is proved by analyzing the Robbins estimator in Appendix C.) At the high level, for the lower bound in the compactly supported case

we will choose the base prior $G_0 = \text{Gamma}(\alpha, \beta)$ with $\alpha = \Theta(\log n)$ and $\beta = \Theta(\log n)$. In this case, G_0 is concentrated on a constant mean with variance $\Theta(\frac{1}{\log n})$; this is conceptually similar to the construction for the Gaussian model with bounded means where we choose $G_0 = N(0, \Theta(\frac{1}{\log n}))$. For the subexponential class, however, we choose an exponential G_0 (with $\alpha = 1$ and $\beta = \Theta(1)$), which is analogous to $G_0 = N(0, \Theta(1))$ in the Gaussian model with subgaussian priors.

Proof of Theorem 2: Lower bound. We start with the compactly supported case. Let $\mathcal{G} = \mathcal{P}([0, h])$ denote the collection of all priors on $[0, h]$. First, in view of Lemma 5 the lower bound is equivalent to proving that

$$\mathsf{Regret}_n(\mathcal{G}) \gtrsim \frac{1}{n}\left(\frac{\log n}{\log\log n}\right)^2 .$$

We aim to apply Proposition 7 with functions $r_1, \ldots, r_m$ chosen from Lemma 11 applied with

$$m = c_1 \frac{\log n}{\log\log n}, \quad \alpha = c_1 \log n\,, \quad \beta = c_2\alpha,$$

with constants $c_1, c_2, c_3 > 0$ to be specified later based on h. Properties (64) and (65) ensure that $\tau_2 = 0$, $\tau = \tau_1 = 1$ and $\gamma = \frac{Cc_2}{m}$ for some absolute constant C (in the notation of the Proposition 7). Setting $\mathcal{G}' = \{G : |\frac{dG}{dG_0} - 1| \le 1/2\}$, Proposition 7 implies that

$$\mathsf{Regret}_n(\mathcal{G}') \gtrsim \frac{m^2}{n} \asymp \frac{1}{n}\left(\frac{\log n}{\log\log n}\right)^2 , \tag{70}$$

provided that

$$m\sqrt{c_2}e^{C(m\log\beta+\alpha)} < \sqrt{n\gamma} = \sqrt{Cc_2 n/m}\,. \tag{71}$$

Note that as $n \to \infty$, the left side is $n^{2Cc_1+o(1)}$ and the right side is $n^{1/2+o(1)}$. Thus, taking $c_1 = \frac{1}{8C}$ ensures that, for all sufficiently large $n \ge n_0$, condition (71) and, in turn, the bound (70) hold.

To translate (70) to regret on the set of priors $\mathcal{G} = \{G : \text{supp}(G) \subset [0, h]\}$ we apply Lemma 6 (with $a = h$). Note that for any $G \in \mathcal{G}'$ we have $G[\theta > h] \le 2G_0[\theta > h]$. For $G_0 = \text{Gamma}(\alpha, \beta)$ with mean $\frac{\alpha}{\beta} = \frac{1}{c_2}$, Chernoff bound yields

$$G_0[\theta > h] \le \inf_{t>0} e^{-th}\left(1 - \frac{t}{\beta}\right)^{-\alpha} = e^{-h\beta+\alpha+\alpha\log\left(\frac{h\beta}{\alpha}\right)} = e^{-\alpha(hc_2+1-\log(hc_2))}, \tag{72}$$

provided that $h > \frac{1}{c_2}$. Recall that $\alpha = c_1 \log n$. Thus, choosing $c_2 = \kappa/h$ with $\kappa = \kappa(c_1) > 1$ such that $\kappa + 1 - \log(\kappa) = \frac{4}{c_1}$, we get $G[\theta > h] \le 2n^{-4}$. On the other hand, we have

$$\mathbb{E}_{G_0}[\theta^4] = \frac{1}{\beta^4}\int_0^\infty \frac{y^{\alpha+3}}{\Gamma(\alpha)}e^{-y}dy = \frac{\Gamma(\alpha+4)}{\Gamma(\alpha)\beta^4} \asymp c_2^{-4}\,.$$

Overall, from (70) and Lemma 6 we obtain that

$$\mathsf{Regret}_n(\mathcal{G}) \gtrsim \frac{1}{n}\left(\frac{\log n}{\log\log n}\right)^2 - O(n^{-3/2})\,,$$

completing the proof of the lower bound (9).

Next we address the subexponential case. Recall that $\mathsf{SubE}(s)$ denotes the collection of all priors G such that $G(X \ge t) \le 2e^{-t/s}$. Choose the base prior G_0 to be $\Gamma(1, \beta)$, namely, the exponential distribution with parameter β. Then $\mathcal{G}' = \{G : |\frac{dG}{dG_0} - 1| \le 1/2\} \subset \mathsf{SubE}(s)$. The rest of the argument is the same by applying Proposition 7 with Lemma 12 in place of Lemma 11. Specifically, let $\alpha = 1, \beta = s, m = c\log n$. Thanks to Lemma 12, we may apply Proposition 7 with $\tau_2 = 0$, $\tau = \tau_1 = 1$ and $\gamma = \frac{C}{m^2}$, where $C = C(\beta)$. Choose c such that $me^{Cm} \le \sqrt{n/m}$, we obtain $\mathsf{Regret}(\mathsf{SubE}(s)) \ge \mathsf{Regret}(\mathcal{G}') \gtrsim \frac{m^3}{n} \asymp \frac{\log^3 n}{n}$, with hidden constants depending on s only. □

4 Discussion

In this paper we proved regret lower bounds for nonparametric empirical Bayes estimation. For the Poisson model (Theorem 2), this is achieved by the Robbins estimator (5) which certifies its optimality. For the Gaussian model (Theorem 1), the rate for the optimal regret is open, but given Remark 2 it is likely that the improvement of the upper bounds is the bottleneck.

Our program for proving regret lower bound (provided by Proposition 7) is general and applies to arbitrary models. However, its execution depends on the choice of a good base prior, such that the resulting kernels K and K_1, cf.(21) and (25), are easy to handle. Namely, the lower bound requires understanding the decay of the eigenvalues and the growth of the eigenfunctions of these kernels. As such, this is the main difficulty of extending our work to more general models, e.g. exponential families considered in [Sin79, Pen99, LGL05].

To end this paper, we discuss related problems in empirical Bayes with the hope of offering different perspectives and identifying open questions.

Restricted EB. In this paper we adopt the general EB formulation where the goal is to approach the Bayes risk (1). In comparison, restricted EB aims to approach the minimum risk in a given class of estimators, such as linear or thresholding estimators. This perspective provides EB interpretations for a number of commonly used methodologies for denoising or model selection, including James-Stein estimator, FDR, AIC/BIC, etc; see [EM72, GF00, JS04, JZ09]. It is of interest to understand the regret in restricted EB settings. Notably, the reduction to individual regret in Section 2.1 no longer holds.

Multiple testing. In addition to estimation, multiple testing and variable selection [GF00, YL05] are important formulations in empirical Bayes. As a specific problem akin to ours, consider the normal means problem where for each θ_i we aim to test the hypothesis $H_{0,i} : \theta_i \le \theta_0$ and $H_{1,i} : \theta_i > \theta_0$. For any test procedure $\delta \in \{0, 1\}$, a meaningful loss function is $R(\theta, \delta) \triangleq \delta(\theta_0 - \theta)1\{\theta \le \theta_0\} + (1-\delta)(\theta - \theta_0)1\{\theta > \theta_0\}$. An $O((\log n)^{3/2})$ upper bound on the total regret is shown in [Lia00] (see also [LG02]). It will be interesting to obtain regret lower bound for such testing problems. Note that for this non-quadratic loss, the reduction (17) to estimating regression function does not hold.

Compound regret. As shown in Proposition 3 and Proposition 4, the EB regret lower bounds in this paper can be extended to the compound setting. In comparison, regret upper bounds are less understood. For the Poisson model, it is not obvious whether the analysis of Robbins estimator in Appendix C applies to the compound regret in (12). For the Gaussian model, converting the multiplicative bound of [JZ09] into additive form yields a regret that scales at least as $\sqrt{n}$. As such, achieving a logarithmic compound regret is currently open.

Sequential version. It is of practical interest to consider the sequential version of the EB or compound estimation problem (see, e.g., [Han57, Sam65, VR66, Gil68]), in which case the estimation of each parameter θ_t can only depend on the data received so far. To this end, consider the *accumulative regret*, which is a variant of the total regret (6)

$$\mathsf{AccRegret}_n(\mathcal{G}) \triangleq \inf_{\widehat{\theta}^n} \sup_{G \in \mathcal{G}} \left\{ \mathbb{E}_G \left[\|\widehat{\theta}^n(Y_1, \ldots, Y_n) - \theta^n\|^2 \right] - n \cdot \mathrm{mmse}(G) \right\}. \tag{73}$$

with the causality constraint that $\widehat{\theta}_t = \widehat{\theta}_t(Y_1, \ldots, Y_t)$ for each $t = 1, \ldots, n$. The following simple result relates the *accumulative regret* to individual regret. For example, for bounded normal means problem, Theorem 1 shows that the accumulative regret is between $(\frac{\log n}{\log \log n})^2$ and $\frac{(\log n)^3}{(\log \log n)^2}$.

Lemma 13.

$$n \cdot \mathsf{Regret}_n(\mathcal{G}) \le \mathsf{AccRegret}_n(\mathcal{G}) \le \sum_{k=1}^{n} \mathsf{Regret}_k(\mathcal{G}).$$

The lower bound follows by dropping the causality constraint and using Lemma 5. The upper bound simply follows from applying the optimal estimator that achieves the individual regret Regret_k for sample size k. It is unclear whether this strategy is optimal, because the worst-case "hyperprior" (prior on the set of priors) for each individual regret depends on the sample size, while in (73) the prior is frozen throughout all sample sizes.

ℓ_p-ball and sparsity. For the normal mean model Theorem 1 provides a regret lower bound for subgaussian priors. It is of great interest to consider the moment constraint due to its connection to sparsity and adaptive estimation. Consider the EB setting, where the prior G has bounded pth absolute moment, namely, $\mathcal{M}_{p,\alpha} \triangleq \{G : \int |\theta|^p G(d\theta) \le \alpha^p\}$. [JZ09, Theorem 3] showed that for this class the individual regret is at most $\mathsf{Regret}(\mathcal{M}_{p,\alpha}) = \widetilde{O}((\alpha/n)^{\frac{p}{1+p}})$. (Here and below tilde hides logarithmic factors.) Determining the tightness of this bound is an open question. In comparison, under the same moment condition, the minimax squared Hellinger risk for density estimation is also $\widetilde{O}((\alpha/n)^{\frac{p}{1+p}})$ (see [Zha09, Theorem 1] or [JZ09, Theorem 4]). This dependency on n is optimal within logarithmic factors as shown by [KG20, Theorem 2.3] for constant α.

The connection to sparse estimation lies in the compound setting. Suppose the mean vector belongs to the ℓ_p-ball $\Theta_{n,p,\alpha} = \{\theta^n \in \mathbb{R}^n : \sum_{i=1}^n |\theta_i|^p \le n\alpha\}$, where $0 < p < 2$. The minimax squared error of estimating θ^n on the basis of $Y^n \sim \mathcal{N}(\theta^n, I)$ is known to be $\widetilde{\Theta}(n\alpha^p)$ (see [DJ94, Theorem 3] for precise characterizations). [JZ09, Theorem 2] showed that EB estimator based on NPMLE can adaptively achieve the minimax risk over $\Theta_{n,p,\alpha}$ up to an $1 + o(1)$ multiplicative factor, provided that $\alpha = \widetilde{\Omega}(n^{-1/p})$; this condition can be explained by comparing the regret $(\alpha/n)^{\frac{p}{1+p}}$ with the minimax risk per coordinate α^p. Using arguments similar to Proposition 4 one can show that the regret in the compound setting is at least that in the EB setting. Thus proving regret lower bound sheds light on to what extent can one adapt to the radius of the ℓ_p-ball and the optimality of NPMLE-based schemes.

A Proofs for the Gaussian model

A.1 Proof of Lemma 9

Let $\{H_k\}$ denote the monic orthogonal polynomial with respect to the standard normal density φ, i.e. $H_k(y) = (-1)^k e^{y^2/2} \frac{d^n}{dy^n} e^{-y^2/2}$, so that where $H_0(y) = 1, H_1(y) = y, H_2(y) = y^2 - 1, \ldots$.

We start by recalling the following facts about Hermite polynomials that will be used in the proof:[5]

- Orthogonality [GR07, 7.374]:

$$\int \varphi(x) H_i(x) H_j(x) = j! 1\{i = j\}. \tag{74}$$

- Three-term recursion [GR07, 8.952.2]:

$$x H_k(x) = H_{k+1}(x) + k H_{k-1}(x). \tag{75}$$

[5] These facts are adapted from [GR07, Chap. 7 and 8]. Therein, the Hermite polynomials, denoted here by $\mathbf{H}_k$, are orthogonal with respect to e^{-x^2}. Thus $H_k(x) = 2^{-k/2} \mathbf{H}_k(x/\sqrt{2})$.

- Derivatives [GR07, 8.952.1]:
$$H_k'(y) = kH_{k-1}(y), \tag{76}$$
- Cramér's inequality [GR07, 8.954.2]:
$$|H_k(y)| \le \kappa\sqrt{k!}e^{\frac{y^2}{4}}, \tag{77}$$
where $\kappa \approx 1.086$ is an absolute constant.
- Mehler formula [Meh66]: For any $0 < \mu < 1$ and $u, v \in \mathbb{R}$ we have
$$\sum_{k=0}^{\infty} \frac{\mu^k}{k!} H_k(u)H_k(v)\varphi(u)\varphi(v) = \frac{1}{2\pi\sqrt{1-\mu^2}} e^{-\frac{1}{2(1-\mu^2)}(u^2+v^2-2\mu uv)}. \tag{78}$$

From (78) by dividing both sides by $\sqrt{\varphi(u)\varphi(v)}$ we obtain

$$\sum_{k=0}^{\infty} \frac{\mu^k}{k!} g_k(u)g_k(v) = \frac{1}{\sqrt{2\pi(1-\mu^2)}} e^{-\frac{a_1}{2}(u^2+v^2-2c_1uv)}, \qquad a_1 \triangleq \frac{1}{2}\frac{1+\mu^2}{1-\mu^2},\ c_1 \triangleq \frac{2\mu}{1+\mu^2}, \tag{79}$$

with $g_k(u) \triangleq H_k(u)\sqrt{\varphi(u)}$, which forms a complete set of orthogonal functions in $L_2(\text{Leb})$, cf. (74). Our goal is to find constants $(\mu, \alpha_1, \alpha_2)$ such that

$$\alpha_2 \sum_k \frac{\mu^k}{k!} g_k(\alpha_1 x)g_k(\alpha_1 \widetilde{x}) = S(x, \widetilde{x}), \tag{80}$$

with $S(x, \widetilde{x})$ defined in (43), in which case $\{g_k(\alpha_1 x)\}$ will be the orthogonal eigenbasis of the S operator in (42). By comparing (79) and the definition of $S(x, \widetilde{x})$, we need to fulfill

$$c_1 = \rho, \quad a_1\alpha_1^2 = \lambda_2$$

where $\rho = \frac{s}{1+s}, \lambda_1 = \frac{1}{2\pi s}\frac{1+s}{\sqrt{1+2s}}$ and $\lambda_2 = \frac{(1+s)^2}{s(1+2s)}$ as in (44). Using (79), we find

$$\mu = \frac{\rho}{1+\rho_1}, \alpha_1 = \sqrt{2\lambda_2\rho_1}, \alpha_2 = \lambda_1\sqrt{2\pi(1-\mu^2)}, \quad \rho_1 \triangleq \sqrt{1-\rho^2}, \tag{81}$$

which satisfies the desired (80); in particular, we have $0 < \mu < 1$. Finally, setting

$$\psi_k(x) \triangleq \frac{g_k(\alpha_1 x)}{\|g_k(\alpha_1 \cdot)\|_{L_2(\text{Leb})}} = \sqrt{\frac{\alpha_1}{k!}} H_k(\alpha_1 x)\sqrt{\varphi(\alpha_1 x)} \tag{82}$$

completes the construction of the orthonormal eigenbasis of S. Indeed, (80) is then rewritten as

$$S(x, \widetilde{x}) = \sum_{k=0}^{\infty} \lambda_0\mu^k\psi_k(x)\psi_k(\widetilde{x}), \tag{83}$$

where

$$\lambda_0 \triangleq \frac{\alpha_2}{\alpha_1} = \frac{1}{\sqrt{2\pi s(1+\rho_1)}}, \tag{84}$$

where we used $1-\mu^2 = \frac{2\rho_1}{1+\rho_1}$ to simplify the expression. In turn, (83) implies that

$$S\psi_k = \lambda_0\mu^k\psi_k, \tag{85}$$

i.e. that $\{\psi_k\}$ is the orthonormal eigenbasis for S.

We proceed to checking properties (45)-(47) of the ψ_k claimed in the statement of the lemma. In view of (41), (45) is just a restatement of (85). For (46) we notice that

$$(K_1\psi_k, K_1\psi_n)_{L_2(f_0)} \overset{(38)}{=} \eta^4 (K\psi_k', K\psi_n')_{L_2(f_0)} \overset{(41)}{=} \eta^4 (S\psi_k', \psi_n')_{L_2(\text{Leb})} .$$

On the other hand, differentiating (82) and using (76) and (75) we get

$$\begin{aligned}
\psi_k'(x) &= \sqrt{\frac{\alpha_1}{k!}}\sqrt{\varphi(\alpha_1 x)}\left[\alpha_1 H_k'(\alpha_1 k) - H_k(\alpha_1 x)\frac{\alpha_1^2}{2}x\right] \\
&= \sqrt{\frac{\alpha_1}{k!}}\sqrt{\varphi(\alpha_1 x)}\left[\alpha_1 k H_{k-1}(\alpha_1 x) - \frac{\alpha_1}{2}(H_{k+1}(\alpha_1 x) + kH_{k-1}(\alpha_1 x))\right] \\
&= \frac{\alpha_1}{2}[\sqrt{k}\psi_{k-1}(x) - \sqrt{k+1}\psi_{k+1}] .
\end{aligned}$$

This implies that in the basis ψ_k the operator $K_1^* K_1$ is tridiagonal, i.e.

$$(K_1\psi_k, K_1\psi_m)_{L_2(f_0)} = \lambda_3 \cdot \begin{cases} -\sqrt{(k-1)k}\mu^{k-1}, & m = k-2 \\ k\mu^{k-1} + (k+1)\mu^{k+1}, & m = k \\ -\sqrt{(k+1)(k+2)}\mu^{k+1}, & m = k+2 \\ 0, & \text{o/w} \end{cases} , \tag{86}$$

where

$$\lambda_3 = \lambda_0 \frac{\alpha_1^2 \eta^4}{4} = \frac{1}{\sqrt{8\pi}}\sqrt{\frac{s}{1+\rho_1}}\frac{\rho_1}{1+2s} . \tag{87}$$

This proves (46).

To show (47), we simply apply (77) to (82)

$$|\psi_k(x)| \le \kappa\sqrt{\frac{\alpha_1}{k!}}\sqrt{k!}e^{\frac{\alpha_1 x^2}{4}}\sqrt{\varphi(\alpha_1 x)} = \frac{\kappa}{(2\pi)^{1/4}}\sqrt{\alpha_1} \le \sqrt{\alpha_1} .$$

(In other words, L_2-normalized Hermite functions are uniformly bounded.)

Finally, (48) follows from the expressions of $\mu, \lambda_0, \lambda_3, \alpha_1$ in (81), (84), and (87), by noting that $\rho = s(1+o(1))$ and $\rho_1 \to 1$ as $s \to 0$.

A.2 Proof for Remark 2

In this appendix we justify the claim about m and γ in Remark 2. Namely, suppose there exist functions $r_1, \ldots, r_m$ such that $\|r_i\|_\infty \le a$, $(K_1 r_i, K_1 r_j)_{L_2(f_0)} = 1\{i \ne j\}$, and $\|Kr_i\|^2_{L_2(f_0)} \le \gamma$. Then we must have $m \lesssim \log a$ and $\gamma \gtrsim \frac{1}{\log a}$.

To this end, we develop r_i's in the basis $\{\psi_k\}$ as follows:

$$r_i = \sum_k \rho_{i,k}\psi_k ,$$

where $\{\psi_k\}$ are (dilated) Hermite functions in (82) and satisfy Lemma 9. It turns out that Hermite functions have polynomial-size L_p-norms, namely [ADSMT12, Theorem 2.1]:

$$\|\psi_k\|_{L_p(\text{Leb})} \asymp \alpha_1^{-\frac{1}{p}} k^{\frac{2-p}{4p}} , \qquad 0 < p < 4 , \tag{88}$$

where by (48) $\alpha_1 \asymp 1$ since $s = 1$. Thus, $|\rho_{i,k}| = |(r_i, \psi_k)| \le \|r_i\|_\infty \|\psi_k\|_1 \lesssim ak^{1/4}$. Now consider

$$(K_1 r_i, K_1 r_j)_{L_2(f_0)} = \sum_{k_1,k_2} \rho_{i,k_1}\rho_{j,k_2}(K_1\psi_{k_1}, K_1\psi_{k_2})\,.$$

From (86) and the estimate $|\rho_{i,k}| \lesssim ak^{1/4}$ we conclude that in the preceding sum the total contribution of the terms with $k_1, k_2 \ge k_0$ can be bounded as $\lesssim a^2 k_0^{O(1)}\mu^{k_0}$. Thus, selecting $k_0 \asymp \log a$ we can define truncated versions

$$\widetilde{r}_i = \sum_{k \le k_0} \rho_{i,k}\psi_k\,.$$

For these m functions we have

$$(K_1\widetilde{r}_i, K_1\widetilde{r}_j)_{L_2(f_0)} = 1\{i \ne j\} + o(1)$$

as $a \to \infty$. So on one hand, the set of m functions $\{K_1\widetilde{r}_i, i \in [m]\}$ is almost orthonormal, and on the other hand it is contained in the span of functions $\{K_1\psi_j, 0 \le j \le k_0\}$. This implies $m \le k_0 + 1 \lesssim \log a$.

Next, we show the claim for γ. A similar truncation argument (this time leveraging the diagonal structure of K in (45)) shows that $\|K\widetilde{r}_i\|^2_{L_2(f_0)} = (1 + o(1))\|Kr_i\|^2_{L_2(f_0)}$. Since $\widetilde{r}_i \in V \triangleq \mathrm{span}\{\psi_1, \ldots, \psi_{k_0}\}$ we have

$$\frac{1 + o(1)}{\gamma} \le \max_{f \in V} \frac{\|K_1 f\|^2_{L_2(f_0)}}{\|Kf\|^2_{L_2(f_0)}} =: \xi\,.$$

Standard manipulations then show that ξ equals maximal eigenvalue of the operator $S^{-\frac{1}{2}}K_1^* K_1 S^{-\frac{1}{2}}$ restricted to the subspace V. This $k_0 \times k_0$ matrix is (as can be seen from (45) and (86)) tridiagonal with k-th row's non-zero entries being $\{-\sqrt{(k-1)k}, \frac{k}{\mu} + \mu(k+1), -\sqrt{(k+1)(k+2)}\}$. Bounding the maximal eigenvalue by the maximal absolute row-sum, we get $\xi \lesssim k_0 \lesssim \log a$, as required.

B Proofs for the Poisson model

In this appendix we prove supporting Lemmas 11 and 12 for the Poisson model. Similar to the proof of Lemma 9, we first derive the explicit expression of the operator $S = K^*K : L_2(\mathbb{Z}_+, f_0) \to L_2(\mathbb{R}_+, \mathrm{Leb})$ that satisfies

$$(Kf, Kg)_{L_2(\mathbb{Z}_+, f_0)} = (Sf, g)_{L_2(\mathbb{R}_+, \mathrm{Leb})}\,. \tag{89}$$

.

Proposition 14 (Kernel under the Gamma prior)**.** *Let $G_0 = \mathrm{Gamma}(\alpha, \beta)$ and f_0 be the induced negative binomial distribution (54). Let K denote the operator in (21) which acts as $K : L_2(\mathbb{R}_+, \mathrm{Leb}) \to L_2(\mathbb{Z}_+, f_0)$. Then the operator $S \triangleq K^*K : L_2(\mathbb{R}_+, \mathrm{Leb}) \to L_2(\mathbb{R}_+, \mathrm{Leb})$ satisfies $Sr(x) = \int_0^\infty S(x, x')r(x')dx'$ with kernel function given by*

$$S(x, x') = C(\alpha, \beta)e^{-(x+x')(1+\beta)}((1+\beta)xx')^{\frac{\alpha-1}{2}} I_\nu(2\sqrt{(1+\beta)xx'}), \quad \nu \triangleq \alpha - 1\,, \tag{90}$$

where $C(\alpha, \beta) = \frac{(1+\beta)^3\beta^\alpha}{\Gamma(\alpha)}$ and $I_\nu(z)$ is the modified Bessel function.

Proof. With this choice of G_0 the kernel $K(x,y)$ and $f_0(y)$ are given by (55) and (54), respectively. Then

$$\begin{aligned} S(x,x') &= \sum_{y\geq 0} f_0(y)K(x,y)K(x',y) \\ &= \sum_{y\geq 0} \binom{y+\alpha-1}{y} \left(\frac{\beta}{1+\beta}\right)^{\alpha} \left(\frac{1}{1+\beta}\right)^{y} \frac{(1+\beta)^{2(y+\alpha+1)}}{\Gamma(y+\alpha)^2} (xx')^{y+\alpha-1} e^{-(1+\beta)(x+x')} \\ &\overset{(a)}{=} (1+\beta)^3 \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-(1+\beta)(x+x')} \sum_{y\geq 0} \frac{((1+\beta)xx')^{y+\alpha-1}}{y!\Gamma(y+\alpha)} \\ &\overset{(b)}{=} (1+\beta)^3 \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-(1+\beta)(x+x')} ((1+\beta)xx')^{\frac{\alpha-1}{2}} I_{\alpha-1}(2\sqrt{(1+\beta)xx'}) \,, \end{aligned}$$

where in (a) we used $\binom{y+\alpha-1}{y} = \frac{\Gamma(y+\alpha)}{y!\Gamma(\alpha)}$ and in (b) we applied the following identity [GR07, (8.445)]: for any $\nu \in \mathbb{R}$ and $z \in \mathbb{R}$ the modified Bessel functions satisfies

$$I_\nu(z) = \sum_{y=0}^{\infty} \frac{1}{y!\Gamma(y+\nu+1)} \left(\frac{z}{2}\right)^{\nu+2y} .$$

□

The easiest way to proceed would be to diagonalize the S operator, i.e. to solve the equation $S\Gamma_\lambda = \lambda\Gamma_\lambda$. Letting $\Gamma_\lambda(x) \equiv \phi_\lambda(\sqrt{b}x)$, $b = 2\sqrt{1+\beta}$, this equation is equivalent to

$$\frac{2\sqrt{1+\beta}\beta^\alpha}{\Gamma(\alpha)} \int_{\mathbb{R}_+} e^{-b(u^2+s^2)/4} (us)^{\alpha-1} I_{\alpha-1}(us)\phi_\lambda(s)s\,ds = \lambda\phi_\lambda(u) \,.$$

Unfortunately, we were not able to solve this (except for $\alpha = 1$, in which case the solution found below is an actual eigenbasis; see (102) below.) Instead, we will find a collection of functions Γ_i with the property that[6]

$$(S\Gamma_i, \Gamma_j) = 0, \quad \forall i \neq j \,. \tag{91}$$

Our method to do so is based on the following observation. Suppose there is a decomposition

$$S(x,y) = w_1(x)w_1(y) \sum_n a_n f_n(x) f_n(y) \,, \tag{92}$$

for some strictly positive weight function w_1, such that the system of functions $\{f_n\}$ is orthogonal in $L_2(\mathbb{R}_+, w)$. Then, the system of functions

$$\Gamma_n(x) \triangleq \frac{w(x)}{w_1(x)} f_n(x) \,, \tag{93}$$

satisfies (91). Indeed,

$$(S\Gamma_i, \Gamma_j) = \sum_n a_n \int dx \int dy\, w(x)w(y)f_i(x)f_n(x)f_j(y)f_n(y) = a_i \|f_i\|^4_{L_2(w)} 1\{i=j\} \,. \tag{94}$$

[6]Here and below, unless specified otherwise, inner products $(\cdot,\cdot)$ are with respect to $L_2(\mathbb{R}_+, \mathrm{Leb})$.

To get the required decomposition (92) we recall the Hardy-Hille summation formula (cf. [GR07, 8.976.1]): For all $x, y \in \mathbb{R}, \nu > -1$ and $|z| < 1$ we have

$$\sum_{n\geq 0} n! \frac{L_n^\nu(x) L_n^\nu(y) z^n}{\Gamma(n+\nu+1)} = \frac{(xyz)^{-\nu/2}}{1-z} \exp\left(-(x+y)\frac{z}{1-z}\right) I_\nu\left(\frac{2\sqrt{xyz}}{1-z}\right), \tag{95}$$

where L_n^ν are a family of generalized Laguerre polynomials, cf. [GR07, (8.970)] for the definition. We aim to apply this identity with $x \leftarrow \gamma_2 x$ and $y \leftarrow \gamma_2 x'$, where $\gamma_2 > 0$ is a parameter to be chosen shortly, to get an expression for $I_\nu(2\sqrt{(1+\beta)xx'})$. This forces the choice of z as a solution of

$$\frac{\sqrt{z}}{1-z}\gamma_2 = \sqrt{1+\beta}\,. \tag{96}$$

With this z we get from (90) and (95) the following expansion

$$S(x,x') = w_1(x) w_1(x') \sum_n a_n L_n^\nu(\gamma_2 x) L_n^\nu(\gamma_2 x')\,, \tag{97}$$

where $w_1(x) \triangleq (\gamma_2 x)^\nu e^{-\gamma_3 x}$, $\gamma_3 \triangleq 1+\beta-\gamma_2\frac{z}{1-z}$ and

$$a_n = C(\alpha,\beta)(1+\beta)^{\frac{\nu}{2}} z^{\frac{2n+\nu}{2}} (1-z)\gamma_2^{-\nu} \frac{n!}{\Gamma(n+\nu+1)}\,. \tag{98}$$

We recall the orthogonality relation for Laguerre polynomials [AS64, 22.2.12]:

$$\int_{\mathbb{R}_+} dx x^\nu e^{-x} L_m^\nu(x) L_n^\nu(x) = \frac{\Gamma(n+\nu+1)}{n!} 1\{m=n\}. \tag{99}$$

Thus, setting $w(x) = (\gamma_2 x)^\nu e^{-\gamma_2 x}$ we get that $L_n^\nu(\gamma_2 x)$ is an orthogonal system in $L_2(\mathbb{R}_+, w)$. Consequently, from (93) we get our required system of functions

$$\Gamma_n(x) \triangleq e^{-\gamma_1 x} L_n^\nu(\gamma_2 x)\,, \quad \gamma_1 \triangleq \frac{\gamma_2}{1-z} - 1 - \beta \tag{100}$$

satisfying, in view of (94), (98), and (99),

$$(S\Gamma_n, \Gamma_m) = b_n 1\{n=m\}\,, \quad b_n \triangleq C_2(\alpha,\beta) z^n \frac{\Gamma(n+\alpha)}{n!} \tag{101}$$

and

$$C_2(\alpha,\beta) \triangleq C(\alpha,\beta)(1-z)((1+\beta)z)^{\frac{\alpha-1}{2}}\gamma_2^{-\alpha-1}$$

with $C(\alpha,\beta) = \frac{(1+\beta)^3\beta^\alpha}{\Gamma(\alpha)}$ given in Proposition 14. We note that in the special case of $\alpha = 1$, we have $\nu = 0$ and $\Gamma_n(x) = e^{-\gamma_1 x} L_n(2\gamma_1 x)$ with L_n being the usual Laguerre polynomials. Using (97) and (99), one can verify that $S(x,x') = \sum_{n\geq 0} a_n \Gamma_n(x)\Gamma_n(x')$ and

$$S\Gamma_k = \frac{a_k}{\gamma_1}\Gamma_k. \tag{102}$$

In other words, for $\alpha = 1$, $\{\Gamma_k\}$ are indeed the eigenbasis of the operator S; however, this is not true when $\alpha \neq 1$ (which will be the case of proving the compactly supported case of Theorem 2).

To proceed further, we need to recall the following properties of Laguerre polynomials:

- Exponential growth [AS64, 22.14.13]: for all degrees n and all $\nu, x \geq 0$,

$$|L_n^\nu(x)| \leq e^{x/2}\binom{n+\nu}{n} \triangleq e^{x/2}\frac{\Gamma(n+\nu+1)}{n!\Gamma(\nu+1)}. \tag{103}$$

- Recurrence (cf. [AS64, 22.8 and 22.7]):

$$x\frac{d}{dx}L_n^\nu(x) = nL_n^\nu(x) - (n+\nu)L_{n-1}^\nu(x) \tag{104}$$

$$xL_n^\nu(x) = (2n+\nu+1)L_n(x) - (n+1)L_{n+1}(x) - (n+\nu)L_{n-1}(x) \tag{105}$$

To complete the construction we still need to choose $\gamma_2 > 0$. We do so in a way that permits us to prove the L_∞-boundedness of $\Gamma_n(x)$ functions by way of (103). That is, we set

$$\gamma_1 = \gamma_2/2\,.$$

Together with (96) and (100) this forces the following choice of $\gamma_1, \gamma_2, \gamma_3, z$:

$$z = (\sqrt{1+\beta} - \sqrt{\beta})^2, \quad \gamma_1 = \gamma_3 = \sqrt{\beta(1+\beta)}, \quad \gamma_2 = 2\sqrt{\beta(1+\beta)}\,. \tag{106}$$

We note that in order for (95) to hold we also need to make sure $|z| < 1$, which is indeed the case for all $\beta > 0$.

Similarly to how we defined $S = K^*K$ in (89), we also define $S_1 = K_1^*K_1$. We summarize properties of the constructed set of functions in the following Lemma.

Lemma 15. *The system of functions* (100) *with choice* (106) *satisfies*

$$\|\Gamma_k\|_\infty \leq \binom{k+\alpha}{k} \tag{107}$$

$$(S\Gamma_k, \Gamma_k) = b_k \tag{108}$$

$$(S_1\Gamma_k, \Gamma_k) = \frac{b_k}{4(1+\beta)^2}\left\{\alpha^2 + (k+1)(k+\alpha)z + (k+\alpha-1)k\frac{1}{z}\right\}, \tag{109}$$

with b_k defined in (101)*. Furthermore,*

$$(S\Gamma_k, \Gamma_j) = 0, \quad k \neq j \tag{110}$$

and

$$(S_1\Gamma_k, \Gamma_j) = 0 \qquad \forall |k-j| \geq 3 \tag{111}$$

Proof. We note that (108) and (110) have been established in (101) already, and (107) in (103).

To prove identities involving S_1, note that using Proposition 10 we get

$$(S_1\Gamma_k, \Gamma_j) = \frac{1}{(1+\beta)^2}(K(x\Gamma_k'), K(x\Gamma_j'))_{L_2(f_0)} = \frac{1}{(1+\beta)^2}(S(x\Gamma_k'), x\Gamma_j')\,. \tag{112}$$

Denoting for convenience $\gamma \equiv \gamma_1 = \gamma_2/2$ such that $\Gamma_k(x) = e^{-\gamma x}L_n^\nu(2\gamma x)$ and recalling that $\nu = \alpha - 1$ we have from the recurrence (104)–(105) of generalized Laguerre polynomials:

$$\begin{aligned}
x\Gamma_k'(x) &= e^{-\gamma x}\left\{-\gamma x L_k^\nu(2\gamma x) + 2\gamma x (L_k^\nu)'(2\gamma x)\right\} \\
&= e^{-\gamma x}\left\{-\frac{1}{2}\left((2k+\nu+1)L_k^\nu - (k+1)L_{k+1}^\nu - (k+\nu)L_{k-1}^\nu\right) + \left(kL_k - (k+\nu)L_{k-1}^\nu\right)\right\}(2\gamma x) \\
&= e^{-\gamma x}\left\{-\frac{\alpha}{2}L_k^\nu + \frac{k+1}{2}L_{k+1}^\nu - \frac{k+\nu}{2}L_{k-1}^\nu\right\}(2\gamma x) \\
&= -\frac{\alpha}{2}\Gamma_k(x) + \frac{k+1}{2}\Gamma_{k+1}(x) - \frac{k+\nu}{2}\Gamma_{k-1}(x).
\end{aligned} \tag{113}$$

If $|k-j|\geq 3$, (111) follows from (112) and the orthogonality property (101). Finally, using (113) and (101), (109) follows from

$$\begin{aligned}
&(S_1\Gamma_k,\Gamma_k)=\frac{1}{(1+\beta)^2}(S(x\Gamma_k'),x\Gamma_k')\\
&=\frac{1}{4(1+\beta)^2}\left\{\alpha^2 b_k+(k+1)^2 b_{k+1}+(k+\alpha-1)^2 b_{k-1}\right\}\\
&=\frac{C_2(\alpha,\beta)z^k}{4(1+\beta)^2}\left\{\alpha^2\frac{\Gamma(k+\alpha)}{k!}+(k+1)^2\frac{\Gamma(k+\alpha+1)}{(k+1)!}z+(k+\alpha-1)^2\frac{\Gamma(k+\alpha-1)}{(k-1)!}\frac{1}{z}\right\}\\
&=\frac{b_k}{4(1+\beta)^2}\left\{\alpha^2+(k+1)(k+\alpha)z+(k+\alpha-1)k\frac{1}{z}\right\}.
\end{aligned}$$

□

We are now in a position to prove Lemma 11.

Proof of Lemma 11. Fix m and select

$$r_k=\frac{1}{\sqrt{(S_1\Gamma_k,\Gamma_k)}}\Gamma_k,\qquad k\in\mathcal{K}=\{m,m+3,\dots,4m\}.$$

This ensures that conditions (64) and (65) are satisfied. Recall that $z=\frac{1}{(\sqrt{\beta+1}+\sqrt{\beta})^2}$ and the assumption that $\beta\geq 2$. We have

$$m\leq k\leq 4m,\quad \frac{1}{6\beta}\leq z\leq\frac{1}{4\beta}\leq\frac{1}{8}.$$

Denote $\delta=\alpha/\beta$. To show (66) notice that from (101) and (109) we have for any $k\in\mathcal{K}$

$$\|Kr_k\|^2_{L^2(f_0)}=(Sr_k,r_k)=\frac{(S\Gamma_k,\Gamma_k)}{(S_1\Gamma_k,\Gamma_k)}=\frac{4(1+\beta)^2}{\alpha^2+(k+1)(k+\alpha)z+(k+\alpha-1)\frac{k}{z}}\asymp\frac{1}{\delta m+\delta^2},$$

since $\alpha^2+(k+1)(k+\alpha)z+(k+\alpha-1)k\frac{1}{z}\asymp\alpha^2+m\alpha\beta$, using the assumption of $m\leq\alpha$.

To show (67), we use (107) to get

$$\max_{k\in\mathcal{K}}\|r_k\|_\infty\lesssim\max_{k\in\mathcal{K}}\sqrt{\frac{1}{\delta m b_k}}\binom{k+\alpha}{k}. \tag{114}$$

Using (101) and $\gamma_2=2\sqrt{\beta(1+\beta)}$, we have

$$\begin{aligned}
b_k&=(1+\beta)^3(1-z)\cdot\beta^\alpha\cdot(z(1+\beta))^{\frac{\alpha-1}{2}}\cdot\gamma_2^{-\alpha-1}\cdot z^k\binom{k+\alpha-1}{k}\\
&=AB^{\alpha/2}z^k\binom{k+\alpha-1}{k}
\end{aligned}$$

where $B\triangleq\frac{z\beta^2(1+\beta)}{\gamma_2^2}=\frac{z\beta}{4}\geq\frac{1}{24}$ and $A\triangleq\frac{(1+\beta)^3(1-z)}{2\sqrt{z\beta}}\geq\frac{1}{2}$ for all $\beta\geq 2$. Thus we get $b_k\geq\frac{1}{2}5^{-\alpha}z^k$ and (using $k\leq 4m$)

$$\binom{k+\alpha}{k}^2 b_k^{-1}\leq 2\cdot 5^\alpha(6\beta)^k 4^{k+\alpha}\leq\exp(C'(\alpha+m\log\beta))$$

for some absolute constant C'. Thus, from (114) we have shown that

$$\max_{k\in\mathcal{K}} \|r_k\|_\infty \le \sqrt{\frac{\beta}{\alpha}} e^{C(m\log\beta+\alpha)}\,,$$

for some absolute constant C. This completes the proof of (67). □

Proof of Lemma 12. For $\alpha = 1$ and fixed $\beta > 0$ (not dependent on m) we select functions r_j as in the proof of Lemma 11. This time, however, we have $C(\alpha,\beta) \asymp C_2(\alpha,\beta) \asymp z \asymp \gamma_1 \asymp \gamma_2 \asymp 1$, $b_k \asymp z^k$. This implies via (108)–(109) that

$$\|Kr_j\|^2_{L_2(f_0)} = \frac{(S\Gamma_k,\Gamma_k)}{(S_1\Gamma_k,\Gamma_k)} \asymp \frac{1}{m^2}\,,$$

since $(\alpha^2 + (k+1)(k+\alpha)z + (k+\alpha-1)k\frac{1}{z}) \asymp m^2$. This proves (68). To show (69) we notice that from (107) and $\alpha = 1$:

$$\|r_k\|_\infty = \frac{\|\Gamma_k\|_\infty}{\sqrt{(S_1\Gamma_k,\Gamma_k)}} \lesssim \frac{m}{\sqrt{z^{4m}m^2}} = z^{-2m}\,.$$

Since $z < 1$ we get the estimate (69). □

C Regret bound for the Robbins estimator

In this appendix we prove the upper bound part of Theorem 2 for the Poisson model by analyzing the Robbins estimator $\widetilde{\theta}^n$ in (5). We first consider the compact supported case and show that for any prior G on $[0,h]$,

$$\mathbb{E}_G\left[\|\widetilde{\theta}^n - \theta^n\|^2\right] - n\cdot \mathrm{mmse}(G) \le c_2\left(\frac{\log n}{\log\log n}\right)^2\,, \tag{115}$$

for some constant $c_2 = c_2(h)$. To this end, we mostly follow the method of [BGR13] fixing a mistake there[7] and adapting the proof to fixed sample size n (as opposed to $\mathrm{Poi}(n)$ in [BGR13]). First, we consider the case of bounded support. For any fixed prior G and its induced Poisson mixture f, recall the Bayes optimal estimator in (4), namely

$$\widehat{\theta}_G(y) = (y+1)\frac{f(y+1)}{f(y)}\,. \tag{116}$$

Then the total regret of Robbins' estimator under the prior G is given by

$$\begin{aligned}
R(G,\widetilde{\theta}^n) &\triangleq \sum_{i=1}^n \left(\mathbb{E}[(\widetilde{\theta}_i - \theta_i)^2] - \mathbb{E}[(\widehat{\theta}_G(Y_i) - \theta_i)^2]\right) \overset{(a)}{=} \sum_{i=1}^n \mathbb{E}[(\widetilde{\theta}_i - \widehat{\theta}_G(Y_i))^2] \\
&\overset{(b)}{=} \sum_{y=0}^\infty \mathbb{E}\left[N(y)(y+1)^2\left(\frac{N(y+1)}{N(y)} - \frac{f(y+1)}{f(y)}\right)^2 1\{N(y)>0\}\right] \\
&= \sum_{y=0}^\infty \mathbb{E}\left[\frac{1\{N(y)>0\}(y+1)^2}{N(y)}\left(N(y+1) - \frac{f(y+1)N(y)}{f(y)}\right)^2\right]
\end{aligned}$$

[7] In the first display equation in the proof of [BGR13, Theorem 1], the last identity does not hold.

where (a) follows from $\mathbb{E}[\theta_i|Y_i] = \widehat{\theta}_G(Y_i)$; in (b) we recall the definition of $N(y) = \sum_{i=1}^n 1\{Y_i = y\}$ from (5).

Conditioning on $N(y)$ we have that $N(y+1) \sim \text{Binom}(n - N(y), \frac{f(y+1)}{1-f(y)})$. Thus, by the (conditional) bias-variance decomposition we have (denoting temporarily $q \triangleq \frac{f(y+1)}{1-f(y)}$):

$$\mathbb{E}\left[\left(N(y+1) - \frac{f(y+1)N(y)}{f(y)}\right)^2 \middle| N(y)\right] = (n - N(y))q(1-q) + \left((n-N(y))q - \frac{f(y+1)}{f(y)}N(y)\right)^2$$
$$\le n\frac{f(y+1)}{1-f(y)} + \left(\frac{f(y+1)}{(1-f(y))f(y)}\right)^2 (nf(y) - N(y))^2 .$$

We now notice a simple fact about the Poisson mixture density $f(y)$: Since for every $x > 0$, $\frac{x^y e^{-x}}{y!} \le \frac{y^y e^{-y}}{y!} \le \frac{1}{\sqrt{2\pi y}}$ (Stirling's), we have

$$f(y) \le \frac{1}{\sqrt{2\pi y}} \qquad \forall y > 0 .$$

Thus, in all terms except $y = 0$ we have $\frac{1}{1-f(y)} \lesssim 1$ and we can write

$$R(G, \widetilde{\theta}^n) \lesssim I_0 + \sum_{y=1}^{\infty} I_1(y) + I_2(y) \tag{117}$$

where

$$I_0 \triangleq \frac{f(1)}{1-f(0)} n v_1(n, f(0)) + \left(\frac{f(1)}{(1-f(0))f(0)}\right)^2 v_2(n, f(0))$$
$$I_1(y) \triangleq (y+1)^2 n f(y+1) v_1(n, f(y))$$
$$I_2(y) \triangleq \left((y+1)\frac{f(y+1)}{f(y)}\right)^2 v_2(n, f(y)) ,$$

and the two functions v_1, v_2 are defined in the next lemma (whose proof is given at the end of this appendix):

Lemma 16. *Let $B \sim Binom(n,p)$ and define*

$$v_1(n,p) \triangleq \mathbb{E}[B^{-1} 1\{B > 0\}] , \quad v_2(n,p) \triangleq \mathbb{E}[B^{-1} 1\{B > 0\}(B - np)^2] .$$

There exist absolute constants $c_1, c_2 > 0$ such that for all $n \ge 1, p \in [0,1]$ we have

$$v_1(n,p) \le c_1 \min\left(np, \frac{1}{np}\right) \le c_1, \quad v_2(n,p) \le c_2 \min(1, np) \le c_2 .$$

Assuming this lemma, we proceed to bounding the terms in (117). For I_0, we obtain by bounding $v_1(n, f(0)) \lesssim \frac{1}{nf(0)}$ and $v_2 \lesssim 1$ that

$$I_0 \lesssim \frac{f(1)}{(1-f(0))f(0)} + \left(\frac{f(1)}{(1-f(0))f(0)}\right)^2 . \tag{118}$$

Now, notice that $f(1) \le 1 - f(0)$ and thus $\frac{f(1)}{(1-f(0))f(0)} \le \frac{1}{f(0)}$. To get another bound, we need to make a key observation: for every $y \ge 0$, $\widehat{\theta}_G(y) = \mathbb{E}_G[\theta | Y = y]$ defined in (116) belongs to the interval $[0, h]$ (since G is supported on $[0, h]$); in other words,

$$(y+1)\frac{f(y+1)}{f(y)} \le h \qquad \forall y \ge 0\,. \tag{119}$$

In particular, applying this with $y = 0$ to bound $\frac{f(1)}{f(0)} \le h$ we get, overall, that

$$\frac{f(1)}{(1-f(0))f(0)} \le \min\left(\frac{1}{f(0)}, \frac{h}{1-f(0)}\right) \lesssim \max(1, h)\,.$$

From (118), thus, we get

$$I_0 \lesssim \max(1, h)\,. \tag{120}$$

Next, we bound $I_1(y)$ via Lemma 16 and (119) as

$$\begin{aligned} I_1(y) &\lesssim (y+1)^2 n f(y+1) \min\left(nf(y), \frac{1}{nf(y)}\right) \\ &= (y+1)\left((y+1)\frac{f(y+1)}{f(y)}\right)\min\left((nf(y))^2, 1\right) \le h(y+1)\min\left(1, (nf(y))^2\right)\,. \end{aligned} \tag{121}$$

To proceed, we need another lemma (also proved at the end of this appendix):

Lemma 17. *Let f be a Poisson mixture with mixing distribution G.*

- *Suppose G is supported on $[0, h]$. Then for some constants $c = c(h)$ and $y_0 = \max(2h, c\frac{\log n}{\log\log n})$,*

$$\sum_{y \ge y_0} f(y) \le \frac{1}{n}\,, \tag{122}$$

$$\sum_{y > y_0} f(y)^2(y+1) \le \frac{4h}{n^2} \tag{123}$$

- *Suppose G satisfies $G[X > x] \le ae^{-bx}$ for all $x > 0$. Then for some constant $c_2 = c_2(a, b)$ and $y_1 = c_2 \log n$,*

$$\sum_{y \ge y_1} f(y) \le \frac{1}{n}\,, \tag{124}$$

$$\sum_{y > y_1} f(y)^2(y+1) \le \frac{1}{n^2} \tag{125}$$

Thus, choosing y_0 as above we can sum (121) to get (in the sequel all constants depend on h)

$$\begin{aligned} \sum_{y \ge 1} I_1(y) &\lesssim \sum_{y \le y_0}(y+1) + n^2 \sum_{y > y_0} f(y)^2(y+1) \\ &\lesssim y_0^2 + 1 \asymp \left(\frac{\log n}{\log\log n}\right)^2\,. \end{aligned} \tag{126}$$

Similarly, we have

$$\sum_{y\ge 1} I_2(y) \overset{\text{(a)}}{\le} h^2 \sum_{y\ge 1} v_2(n, f(y))$$
$$\overset{\text{(b)}}{\lesssim} h^2 \sum_{y\ge 1} \min(1, nf(y)) \tag{127}$$
$$\overset{\text{(c)}}{\le} \sum_{y<y_0} 1 + n \sum_{y\ge y_0} f(y) \le y_0 + 1 \asymp \frac{\log n}{\log\log n}\,, \tag{128}$$

where (a) is by (119), (b) is Lemma 16, (c) is by taking y_0 as in Lemma 17 and applying (123).

Substituting (120), (126), and (128) into (117), we obtain

$$R(G, \widetilde{\theta}^n) \le I_0 + \sum_{y\ge 1} I_1(y) + I_2(y) \lesssim \left(\frac{\log n}{\log\log n}\right)^2,$$

completing the proof of (115).

Next, we proceed to the subexponential case. Fix a prior $G \in \mathsf{SubE}(s)$, namely, $G((t,\infty)) \le 2e^{-t/s}$ for all $t > 0$. In the rest of the proof all constants depend on s. We first show that it is possible to replace G with a truncated prior $G'[X \in \cdot] = G[X \in \cdot | X \le c\log n]$ for a suitably large $c = c(s) > 0$. Indeed, observe that the Robbins estimator for any coordinate j satisfies:

$$\widetilde{\theta}_j = (Y_j + 1)\frac{N(Y_j+1)}{N(Y_j)} \le (Y_j+1)n\,. \tag{129}$$

This is because, crucially, $N(Y_j) \ge 1$ by definition. Since $\mathbb{E}[Y_j^4] = \int G(d\theta)(\theta^4 + 6\theta^3 + 7\theta^2 + \theta) \lesssim 1$, we have

$$\mathbb{E}[\widetilde{\theta}_j^4] \le n^4 \mathbb{E}[(Y_j+1)^4] \lesssim n^4. \tag{130}$$

Next, since the G is subexponential, we can select c so large that

$$\epsilon = G[X > c\log n] \le \frac{1}{n^{10}}\,.$$

Now, denote as in the proof of Lemma 6 the event $E = \{\theta_j \le c\log n, j = 1,\dots,n\}$. Recall that $\mathbb{E}_G$ denotes expectation with respect to $\theta_i \overset{iid}{\sim} G$ and $Y_i \sim \mathrm{Poi}(\theta_i)$. Then

$$\frac{1}{n}R(G,\widetilde{\theta}^n) = \mathbb{E}_G[(\widetilde{\theta}_1 - \theta_1)^2] - \mathrm{mmse}(G)$$
$$\le \mathbb{E}_G[(\widetilde{\theta}_1 - \theta_1)^2 | E] - \mathrm{mmse}(G') + \mathrm{mmse}(G') - \mathrm{mmse}(G) + \mathbb{E}_G[(\widetilde{\theta}_1 - \theta_1)^2 1_{E^c}]$$
$$= \frac{1}{n}R(G', \widetilde{\theta}^n) + \mathrm{mmse}(G') - \mathrm{mmse}(G) + \mathbb{E}_G[(\widetilde{\theta}_1 - \theta_1)^2 1_{E^c}]\,, \tag{131}$$

where the first identity follows from symmetry. For the last term in (131) we have from Cauchy-Schwarz and $\mathbb{P}[E^c] \le n\epsilon$:

$$\mathbb{E}_G[(\widetilde{\theta}_1 - \theta_1)^2 1_{E^c}] \le \sqrt{n\epsilon}\sqrt{\mathbb{E}_G[(\widetilde{\theta}_1 - \theta_1)^4}} \lesssim \sqrt{n\epsilon}n^2\,,$$

where in the last step we used (130) and $\mathbb{E}_G[\theta_1^4] \lesssim 1$ (since G is subexponential).

For the second term, as in (20) we use $\mathrm{mmse}(G') \le \frac{1}{1-\epsilon}\mathrm{mmse}(G)$ so that

$$\mathrm{mmse}(G') - \mathrm{mmse}(G) \le \frac{\epsilon}{1-\epsilon}\mathrm{mmse}(G) \lesssim \epsilon\,.$$

Altogether, we get from (131) that

$$R(G, \widetilde{\theta}^n) \le R(G', \widetilde{\theta}^n) + O(1/n)\,,$$

Thus, it is sufficient to analyze the regret of Robbins under the prior G'. Note that G' satisfies the assumptions in the second part of Lemma 17 with constants a, b depending on s only. According to (117) we need to bound I_0, $\sum_{y\ge1} I_1(y)$ and $\sum_{y\ge2} I_2(y)$. We set $h = c\log n$ (which is the support of G') and get from (120) the bound $I_0 \lesssim \log n$. For I_1 we have from (121) the following estimate

$$\sum_{y\ge1} I_1(y) \lesssim h\sum_{y\ge1}(y+1)\min(1, nf(y))^2 \le h\left(\sum_{y=1}^{y_1}(y+1) + n^2\sum_{y>y_1}(y+1)f^2(y)\right)\,,$$

where $y_1 \asymp \log n$ is defined in Lemma 17. Then the first sum is proportional to $\log^2 n$ and the second, in view of (125), is at most $\frac{1}{n^2}$. Overall we get

$$\sum_{y\ge1} I_1(y) \lesssim h(y_1^2+1) \asymp \log^3 n\,.$$

Similarly, from (127) we have

$$\sum_{y\ge1} I_2(y) \lesssim h^2\sum_{y\ge1}\min(1, nf(y)) \le h^2\left(\sum_{y\le y_1} 1 + n\sum_{y>y_1} f(y)\right)\,.$$

The first sum equals y_1 and the second is bounded by $\frac{1}{n}$ (via (124)), so we get

$$\sum_{y\ge1} I_2(y) \lesssim h^2(y_1+1) \asymp \log^3 n\,.$$

Altogether, we obtain the claimed $O(\log^3 n)$ regret bound for the subexponential class.

Remark 4. In the analysis of Robbins estimator, if we denote by y the effective support of the Poisson mixture (i.e. the $1-\frac{1}{n^{10}}$ quantile) and by h the effective support of the mixing distribution, then the total regret is at most $O(hy(h+y))$. In the compact support case $h \asymp 1$ and $y \asymp \frac{\log n}{\log\log n}$, while in the subexponential case $h \asymp y \asymp \log n$. This explains the rates in (9) and (10), respectively.

Proof of Lemma 16. Note that $B^{-1}1\{B>0\} \le \frac{2}{B+1}1\{B>0\}$. Furthermore, we have

$$\begin{aligned}
\mathbb{E}[(B+1)^{-1}1\{B>0\}] &= \sum_{k=1}^{n}\frac{n!}{(k+1)!(n-k)!}p^k(1-p)^{n-k}\\
&= \frac{1}{p(n+1)}\sum_{k=1}^{n}\binom{n+1}{k+1}p^{k+1}(1-p)^{n-k}\\
&\le \frac{1}{pn}\mathbb{P}[\mathrm{Binom}(n+1,p)\ge 2] \le \frac{1}{np}\,.
\end{aligned}$$

For the case of $np \le 1$,

$$\mathbb{E}[B^{-1}1\{B>0\}] \le \mathbb{P}[B \ge 1] \le \sum_{k=1}^{n} \frac{(np)^k}{k!} \le np \cdot \mathbb{P}[\mathrm{Poi}(np) \ge 0] \le np.$$

Overall, we have

$$\mathbb{E}[(B+1)^{-1}1\{B>0\}] \le \min\left(np, \frac{1}{np}\right), \tag{132}$$

and the desired bound $v_1(n,p) \equiv \mathbb{E}[B^{-1}1\{B>0\}] \le 2\min(np, \frac{1}{np})$.

For v_2, we again bound $B^{-1}1\{B>0\} \le \frac{2}{B+1}1\{B>0\}$, and also notice that $(B-np)^2 = (B+1-(n+1)p-(1-p))^2 \le 2(B+1-(n+1)p)^2 + 2(1-p)^2$. This implies

$$v_2(n,p) \le 2\mathbb{E}[(B+1)^{-1}1\{B>0\}\{((n+1)p-(B+1))^2 + (1-p)^2\}]$$

Notice that by (132) expectation of the second quadratic term can be bounded as $\lesssim (1-p)^2$. Thus, we focus on the first term, for which we have the following expression in terms of $\widetilde{B} \sim \mathrm{Binom}(n+1,p)$:

$$\begin{aligned}
&\mathbb{E}[(B+1)^{-1}1\{B>0\}\{((n+1)p-(B+1))^2 \\
&= \sum_{k=1}^{n} \frac{n!}{(k+1)!(n-k)!} p^k(1-p)^{n-k}(k+1-(n+1)p)^2 \\
&= \frac{1}{(n+1)p}\mathbb{E}[1\{2 \le \widetilde{B} \le n+1\}((n+1)p-\widetilde{B})^2] \\
&\le \frac{1}{(n+1)p}\mathrm{Var}[\widetilde{B}] = 1-p
\end{aligned}$$

This proves $v_2(n,p) \le 2$. On the other hand, suppose $np < 1$. Then for any $B \ge 1$ we have $(B-np)^2 \le B^2$ and, consequently,

$$v_2(n,p) \le \mathbb{E}[1\{B>0\}B] \le np\,.$$

And, hence, we have $v_2 \le 2\min(1, np)$. □

Proof of Lemma 17. First, let us consider the compactly supported case. For any $y > h > x$ we have $\frac{x^y e^{-x}}{y!} \le \bar{f}(y) \triangleq \frac{h^y e^{-h}}{y!}$, implying that $f(y) = \mathbb{E}_{X\lambda_2}[\frac{X^y e^{-X}}{y!}] \le \bar{f}(y)$. Note that $\bar{f}(y)$ is monotonically decreasing in y for $y > h$. Furthermore, we have for any $y_0 \ge 2h$ that

$$\begin{aligned}
\sum_{y \ge y_0} \bar{f}(y) &= \frac{h^{y_0}e^{-h}}{y_0!} \sum_{m \ge 0} \frac{y_0!}{(y_0+m)!} h^m \\
&\le \bar{f}(y_0) \sum_{m \ge 0} \left(\frac{h}{y_0}\right)^m \le 2\bar{f}(y_0).
\end{aligned} \tag{133}$$

Furthermore, we have

$$\begin{aligned}\sum_{y>y_0} \bar{f}(y)^2(y+1) &\overset{(a)}{\le} 2\sum_{y>y_0} y\bar{f}(y)^2 \\ &\overset{(b)}{=} 2h\sum_{y>y_0} \bar{f}(y-1)\bar{f}(y)^2 \\ &\overset{(c)}{\le} 2h\bar{f}(y_0)\sum_{y>y_0} \bar{f}(y) \\ &\overset{(d)}{\le} 4h\bar{f}(y_0)^2\,,\end{aligned}$$

where in (a) we used the fact that $y \ge 1$, in (b) the identity $y\bar{f}(y) = h\bar{f}(y-1)$, in (c) the monotonicity of $\bar{f}$, and in (d) we applied (133). Thus, selecting $y_0 > c(h)\frac{\log n}{\log\log n}$ yields $\bar{f}(y_0) \le \frac{1}{n}$ and completes the proof of (122) and (123).

For the subexponential case, let us denote by $\bar{F}(x) = G[X > x]$ and recall $\bar{F}(x) \le ae^{-bx}$. Then from integration by parts we have

$$\begin{aligned}f(y) = \mathbb{E}_{X\lambda_2}\left[\frac{X^y e^{-X}}{y!}\right] &= \frac{1}{y!}\int_0^\infty (y-x)x^{y-1}e^{-x}\bar{F}(x)dx \\ &\le \frac{1}{y!}\int_0^\infty yx^{y-1}e^{-x}\bar{F}(x)dx \\ &\le \frac{a}{(y-1)!}\int_0^\infty x^{y-1}e^{-(1+b)x}dx \\ &= a(1+b)^{-y} \triangleq \widetilde{f}(y)\,.\end{aligned}$$

The estimates (124) and (125) then easily follow from the properties of geometric distribution $\widetilde{f}$. □

D Proofs for the compound setting

Proof of Proposition 3. As noted in [GR09], the first inequality of (14) follows from $R_{\mathsf{oracle}}(G_{\theta^n}) \le n\mathrm{mmse}(G_{\theta^n})$ since $f(Y^n) = (f_1(Y_1), \dots, f_1(Y_n))$ is permutation-invariant. For the second, by the concavity of $G \mapsto \mathrm{mmse}(G)$ (see [WV12, Corollary 1]), we have

$$\mathbb{E}_{\theta^n \overset{iid}{\sim} G}\mathrm{mmse}(G_{\theta^n}) \le \mathrm{mmse}(G)\,, \tag{134}$$

and, therefore,

$$\mathbb{E}_G[\|\theta^n - \widehat{\theta}^n(Y^n)\|^2] - n\mathrm{mmse}(G) \le \mathbb{E}_G[\|\theta^n - \widehat{\theta}^n(Y^n)\|^2 - n\mathrm{mmse}(G_{\theta^n})].$$

Taking $\inf_{\widehat{\theta}^n}\sup_G$ on both sides proves the second inequality in (14). □

Proof of Proposition 4. The proof follows the proof strategy for the second inequality in (14) and the fact that the empirical distribution drawn from a subgaussian distribution is subgaussian (with multiplicative adjustment of constants) with probability $1 - 1/\mathsf{poly}(n)$.

Fix $s > 0$ and $G \in \mathsf{SubG}(s)$. Let $\theta_1, \dots, \theta_n \overset{\text{i.i.d.}}{\sim} G$ and denote the empirical distribution $G_{\theta^n} = \frac{1}{n}\sum_{i=1}^n \delta_{\theta_i}$. We first show that there exists an absolute constant C_0 such that

$$\mathbb{P}\left[G_{\theta^n} \in \mathsf{SubG}(C_0 s)\right] \ge 1 - n^{-4}. \tag{135}$$

Indeed, by the equivalent characterization of subgaussian constant (see [Ver18, Proposition 2.5.2]), it suffices to show

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\exp\left(\frac{\theta_i^2}{C_1 s}\right) > 2\right] \le n^{-4}. \tag{136}$$

for some absolute constant C_1. Since $\mathbb{P}[|\theta_i| \ge t] \le 2e^{-t^2/(2s)}$ by assumption, there exists C_1 such that iid random variables $X_i \triangleq \exp\left(\frac{\theta_i^2}{C_1 s}\right)$ satisfy $\mathbb{E}[X] \le 1.1$ and $\mathbb{E}\left[X_i^8\right] \le 1.2$. From Rosenthal's inequality [Ros70, Theorem 3] we obtain $\mathbb{E}[(\sum_i X_i - \mathbb{E}[X])^8] \lesssim n^4$, and then from Markov's inequality $\mathbb{P}[\frac{1}{n}\sum_i X_i > 2] \lesssim n^{-4}$, concluding the proof of (136).

Next recall that for a given class of priors $\mathcal{G}$, the total regret in the compound setting is defined as

$$\mathsf{TotRegretComp}_n(\mathcal{G}) \triangleq \inf_{\widehat{\theta}^n} \sup_{\theta^n \in \Theta(\mathcal{G})} \left\{\mathbb{E}_{\theta^n}[\|\widehat{\theta}^n(Y^n) - \theta^n\|^2] - n \cdot \mathrm{mmse}(G_{\theta^n})\right\},$$

where $\Theta(\mathcal{G}) \triangleq \{\theta^n : G_{\theta^n} \in \mathcal{G}\}$. A simple observation is that for $\mathsf{TotRegretComp}_n(\mathsf{SubG}(s))$, it is sufficient to restrict to $\widehat{\theta}^n = \widehat{\theta}^n(Y^n)$ such that $\|\widehat{\theta}^n\| \le \sqrt{C_2 sn}$ deterministically for some absolute constant C_2. This simply follows from the fact that for any $\theta^n \in \Theta(\mathsf{SubG}(s))$, $\|\theta^n\| \le \sqrt{C_2 sn}$, so that for any estimator $\widehat{\theta}^n$, the modification $\widetilde{\theta}^n \triangleq \mathrm{argmin}\{\|\widehat{\theta}^n - x\| : \|x\| \le \sqrt{C_2 sn}\}$ improves $\widehat{\theta}^n$ *pointwise*, in the sense that $\|\widetilde{\theta}^n - \theta^n\| \le \|\widehat{\theta}^n - \theta^n\|$ for every $\theta^n \in \Theta(\mathsf{SubG}(s))$.

Finally, we show

$$\mathsf{TotRegretComp}_n(\mathsf{SubG}(C_0 s)) \ge \mathsf{TotRegret}_n(\mathsf{SubG}(s)) - O(sn^{-1}) \tag{137}$$

To this end, fix any estimator $\widehat{\theta}^n$ such that $\|\widehat{\theta}^n\| \le \sqrt{C_2 C_0 sn}$. For any $G \in \mathsf{SubG}(s)$, let $\theta_1, \ldots, \theta_n \overset{\text{i.i.d.}}{\sim} G$. Then

$$\begin{aligned}
&\sup_{\theta^n \in \Theta(\mathsf{SubG}(C_0 s))} \left\{\mathbb{E}_{\theta^n}[\|\widehat{\theta}^n(Y^n) - \theta^n\|^2] - n \cdot \mathrm{mmse}(G_{\theta^n})\right\} \\
&\ge \mathbb{E}\left[\|\widehat{\theta}^n - \theta^n\|^2 - n \cdot \mathrm{mmse}(G_{\theta^n}) \mid G_{\theta^n} \in \mathsf{SubG}(C_0 s)\right] \\
&\ge \mathbb{E}[\|\widehat{\theta}^n - \theta^n\|^2 - n \cdot \mathrm{mmse}(G_{\theta^n})] - \mathbb{E}[\|\widehat{\theta}^n - \theta^n\|^2 1\{G_{\theta^n} \notin \mathsf{SubG}(C_0 s)\}] \\
&\overset{(a)}{\ge} \mathbb{E}[\|\widehat{\theta}^n - \theta^n\|^2] - n \cdot \mathrm{mmse}(G) - \sqrt{\mathbb{E}[\|\widehat{\theta}^n - \theta^n\|^4]\mathbb{P}\left[G_{\theta^n} \notin \mathsf{SubG}(C_0 s)\right]} \\
&\overset{(b)}{\ge} \mathbb{E}[\|\widehat{\theta}^n - \theta^n\|^2] - n \cdot \mathrm{mmse}(G) - O(\frac{s}{n}),
\end{aligned} \tag{138}$$

where in (a) we apply (134) again for the first term and Cauchy-Schwarz for the second; (b) follows from (135) and the following bounds

$$\begin{aligned}
\mathbb{E}[\|\widehat{\theta}^n - \theta^n\|^4] &\le 8(\mathbb{E}[\|\widehat{\theta}^n\|^4] + \mathbb{E}[\|\theta^n\|^4] \\
&\overset{(c)}{\le} 8((C_2 C_0 sn)^2 + n^2\mathbb{E}^2[\theta^2] + n\mathrm{Var}[\theta^2]) \\
&\lesssim n^2 s^2
\end{aligned}$$

and in (c) we used the fact that $\|\widehat{\theta}^n\| \le \sqrt{C_2 C_0 sn}$. Taking the infimum on both sides of (138) over $\widehat{\theta}^n$ subject to the norm constraint followed by a supremum over $G \in \mathsf{SubG}(s)$ over the right side, we complete the proof of (137). Since $\mathsf{TotRegret}(\mathsf{SubG}(c_0 s))$ is increasing with c_0, the statement (137) is equivalent to that of the Theorem upon taking $c_0 = \frac{1}{KC_0}$ with a suitably large K. □

E Results on density estimation

The construction in the present paper can be used to yield improved or new lower bounds for mixture density estimation. Denote by the minimax squared Hellinger risk for density estimation over the mixture class $\{f_G : G \in \mathcal{G}\}$, namely,

$$R_n(\mathcal{G}) \triangleq \inf_{\widehat{f}} \sup_{G \in \mathcal{G}} \mathbb{E}_G[H^2(f_G, \widehat{f})],$$

where $\widehat{f}$ is measurable with respect to the sample $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} f_G$ and the Hellinger distance between densities f and g with respect to a dominating measure ν is denoted by $H(f,g) \triangleq (\int(\sqrt{f} - \sqrt{g})^2 d\nu)^{1/2}$.

General program. The following result is a counterpart of Proposition 7. Since we no longer dealing with regression functions, the program only involves the operator K in (21) not the higher-order operator K_1.

Proposition 18. *Fix a prior distribution G_0, constants $a, \tau, \tau_2, \gamma \geq 0$ and m functions $r_1, \ldots, r_m$ on $\mathcal{X}$ with the following properties.*

1. *For each i, $\|r_i\|_\infty \leq a$;*
2. *For each i, $\|Kr_i\|_{L_2(f_0)} \leq \sqrt{\gamma}$;*
3. *For any $v \in \{0, \pm 1\}^m$,*
$$\left\| \sum_{i=1}^m v_i K r_i \right\|_{L_2(f_0)}^2 \geq \tau \|v\|_2^2 - \tau_2; \tag{139}$$
4. *For each i, $\int r_i dG_0 = 0$.*

Then the minimax H^2-risk for density estimation over the class of priors $\mathcal{G} = \{G : |dG/dG_0 - 1| \leq \frac{1}{2}\}$ satisfies

$$R_n(\mathcal{G}) \geq C\delta^2(m\tau - \tau_2)\,, \qquad \delta \triangleq \frac{1}{\max(\sqrt{n\gamma}, ma)}\,, \tag{140}$$

where $C > 0$ is an absolute constant.

Proof. The proof is almost identical to that of Proposition 7 based on Assouad's lemma, so we only point out the major differences. By the fourth assumption, $\mu_i = \int r_i dG_0 = 0$, so we simply have $dG_u = (1 + \delta r_u) dG_0$ and $f_u = (1 + \delta h_u) f_0$, where $h_u = Kr_u$. By (31), we have $\frac{1}{2} \leq f_u/f_0 \leq \frac{3}{2}$, so for any $u, v \in \{0,1\}^m$,

$$H^2(f_u, f_v) \asymp \int \frac{(f_u - f_v)^2}{f_0} = \delta^2 \|K(r_u - r_v)\|_{L_2(f_0)}^2 \geq \delta^2(\tau d_{\mathrm{H}}(u,v) - \tau_2).$$

Then the conclusion follows from applying Assouad's lemma as in the proof of Proposition 7. □

Truncation. Lemma 6 controls the effect of truncation on the regret. The counterpart for density estimation is as follows.

Lemma 19. *Under the setting of Lemma 6,*

$$R_n(\mathcal{G}) \geq R_n(\mathcal{G}') - 8n\sqrt{\epsilon}. \tag{141}$$

Proof. Indeed, using G_a as defined in the proof of Lemma 6,

$$\begin{aligned}
R_n(\mathcal{G}) &= \inf_{\widehat{f}} \sup_{G\in\mathcal{G}} \mathbb{E}_G[H^2(\widehat{f}, f_G)] \\
&\geq \inf_{\widehat{f}} \sup_{G\in\mathcal{G}'} \mathbb{E}_{G_a}[H^2(\widehat{f}, f_{G_a})] \\
&\overset{(a)}{\geq} \inf_{\widehat{f}} \sup_{G\in\mathcal{G}'} \mathbb{E}_G[H^2(\widehat{f}, f_{G_a})] - 2n\epsilon \\
&\overset{(b)}{\geq} \inf_{\widehat{f}} \sup_{G\in\mathcal{G}'} \mathbb{E}_G[H^2(\widehat{f}, f_G)] - 6H(f_G, f_{G_a}) - 2n\epsilon \\
&\overset{(c)}{\geq} R_n(\mathcal{G}') - 8n\sqrt{\epsilon},
\end{aligned}$$

where (a) follows from the following counterpart of (19) since $H^2 \leq 2$: $\mathbb{E}_G[H^2(f_G, \widehat{f})] \leq \mathbb{E}_G[H^2(f_G, \widehat{f})|E] + 2n\epsilon = \mathbb{E}_{G_a}[H^2(f_G, \widehat{f})] + 2n\epsilon$; (b) uses the triangle inequality and boundedness of Hellinger distance; (c) is due to $H^2(f_G, f_{G_a}) \leq H^2(G, G_a) \leq \mathrm{TV}(G, G_a) = \epsilon$, with the first step via the data processing inequality. □

Applications As a concrete application, we reuse the construction in Section 3.1 and Appendix A.1 to derive a minimax lower bound next for estimating Gaussian mixture densities. In particular, the lower bound $\Omega(\frac{\log n}{n})$ for s-subgaussian priors is previously shown in [Kim14, KG20] for sufficiently large s, an assumption which in fact can be dropped; the lower bound $\Omega(\frac{\log n}{n \log\log n})$ for bounded means appears to be new.

Theorem 20 (Gaussian mixture estimation)**.** *Consider the setting of Theorem 1.*

- *(Compactly supported case). For any $h > 0$, there exists constants $c_0 = c_0(h) > 0$ and n_0, such that for all $n \geq n_0$,*

$$R_n(\mathcal{P}([-h, h]) \geq \frac{c_0}{n} \frac{\log n}{\log\log n}\,. \tag{142}$$

- *(Subgaussian case.) For any $s > 0$, there exists constants $c_1 = c_1(s) > 0$ and n_0, such that for all $n \geq n_0$,*

$$R_n(\mathsf{SubG}(s)) \geq \frac{c_1}{n} \log n\,. \tag{143}$$

Proof. We show

$$R_n(\mathsf{SubG}(s)) \geq \frac{c \log n}{n} \tag{144}$$

for some absolute constant c. Taking $s = \frac{c(h)}{\log n}$ and applying the truncation argument in (141) yields (142) for the compactly supported case.

To prove (144), we apply Proposition 18 with $r_k = \frac{\psi_k}{\|K\psi_k\|_{L_2(f_0)}}, k = 1, \ldots, m$, where $\{\psi_k\}$ is the class of functions from Lemma 9, so that $\tau = \gamma = 1, \tau_2 = 0$ and $a = \sqrt{\frac{\alpha_1}{\lambda_0 \mu^m}}$. Since $\mu \asymp s$, choosing $m = \frac{c' \log n}{\log \frac{1}{s}}$ for small constant c' ensures $n\gamma \geq (ma)^2$, leading to the lower bound of $\Omega(\frac{m}{n})$. □

The counterpart for the Poisson model is as follows. The proof is almost identical to that of Theorem 20 using the construction in Lemma 11 and Lemma 12, and is hence omitted. It is not hard to show that the lower bounds in (145) and (146) are in fact optimal [JPW21].

Theorem 21 (Poissoin mixture estimation)**.** *Consider the setting of Theorem 2.*

- *(Compactly supported case.) For any $h > 0$, there exists constants $c_0 = c_0(h) > 0$ and n_0, such that for all $n \geq n_0$,*
$$R_n(\mathcal{P}([0,h]) \geq \frac{c_0}{n} \frac{\log n}{\log \log n}\,. \tag{145}$$
- *(Subexponential case.) For any $s > 0$, there exists constants $c_1 = c_1(s) > 0$ and n_0, such that for all $n \geq n_0$,*
$$R_n(\mathsf{SubE}(s)) \geq \frac{c_1}{n} \log n\,. \tag{146}$$

References

[ADSMT12] AI Aptekarev, JS Dehesa, P Sánchez-Moreno, and DN Tulyakov. Asymptotics of l_p-norms of Hermite polynomials and Rényi entropy of Rydberg oscillator states. *Contemp. Math*, 578:19–29, 2012.

[AS64] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Courier Corporation, 1964.

[BG09] 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.

[BGR13] 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.

[Bro08] 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.

[Cas85] George Casella. An introduction to empirical Bayes data analysis. *The American Statistician*, 39(2):83–87, 1985.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of information theory, 2nd Ed.* Wiley-Interscience, New York, NY, USA, 2006.

[DJ94] D.L. Donoho and I.M. Johnstone. Minimax risk over l_p-balls for l_q-error. *Probability Theory and Related Fields*, 99(2):277–303, 1994.

[Efr11] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

[Efr12] Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

[Efr14] Bradley Efron. Two modeling strategies for empirical Bayes estimation. *Statistical Science*, 29(2):285–301, 2014.

[Efr19] Bradley Efron. Bayes, oracle Bayes and empirical Bayes. *Statistical science*, 34(2):177–201, 2019.

[Efr21] Bradley Efron. Empirical Bayes: Concepts and methods. 2021. http://statweb.stanford.edu/~ckirby/brad/papers/2021EB-concepts-methods.pdf.

[EM72] Bradley Efron and Carl Morris. empirical Bayes on vector observations: An extension of stein's method. *Biometrika*, 59(2):335–347, 1972.

[ETST01] 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.

[GF00] Edward I George and Dean P Foster. Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747, 2000.

[Gil68] Dennis C Gilliland. Sequential compound estimation. *The Annals of Mathematical Statistics*, 39(6):1890–1904, 1968.

[GR07] I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals Series and Products*. Academic Press, New York, NY, seventh edition, 2007.

[GR09] Eitan Greenshtein and Ya'acov Ritov. Asymptotic efficiency of simple decisions for the compound decision problem. *Lecture Notes-Monograph Series*, pages 266–275, 2009.

[GvdV01] 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.

[Han57] James Hannan. Approximation to Bayes risk in repeated play. In *Contributions to the Theory of Games (AM-39), Volume III*, pages 97–140. Princeton University Press, 1957.

[JPW21] Soham Jana, Yury Polyanskiy, and Yihong Wu. Regret optimality of minimum-distance based empirical Bayes method for the Poisson model. *Preprint*, 2021.

[JS04] Iain M Johnstone and Bernard W Silverman. Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–1649, 2004.

[JZ09] Wenhua Jiang and Cun-Hui Zhang. General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647–1684, 2009.

[KG20] Arlene KH Kim and Adityanand Guntuboyina. Minimax bounds for estimating multivariate gaussian location mixtures. *arXiv preprint arXiv:2012.00444*, 2020.

[Kim14] Arlene KH Kim. Minimax bounds for estimation of normal mixtures. *Bernoulli*, 20(4):1802–1818, 2014.

[KW56] 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.

[LG02] Jianjun Li and Shanti S Gupta. Empirical Bayes tests based on kernel sequence estimation. *Statistica Sinica*, pages 1061–1072, 2002.

[LGL05] 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.

[Lia00] Ta Chen Liang. On an empirical Bayes test for a normal mean. *The Annals of Statistics*, 28(2):648–655, 2000.

[Lia04] Ta Chen Liang. On optimal convergence rate of empirical Bayes tests. *Statistics & probability letters*, 68(2):189–198, 2004.

[Lin83] Bruce G Lindsay. The geometry of mixture likelihoods: a general theory. *The Annals of Statistics*, 11(1):86–94, 1983.

[Meh66] F Gustav Mehler. Ueber die Entwicklung einer Function von beliebig vielen Variablen nach Laplaceschen Functionen höherer Ordnung. *Journal für die Reine und Angewandte Mathematik*, 66:161–176, 1866.

[Pen99] Marianna Pensky. Nonparametric empirical Bayes estimation via wavelets. In *Bayesian Inference in Wavelet-Based Models*, pages 323–340. Springer, 1999.

[Rob51] Herbert Robbins. Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. The Regents of the University of California, 1951.

[Rob56] 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.

[Ros70] Haskell P Rosenthal. On the subspaces of $l^p (p > 2)$ spanned by sequences of independent random variables. *Israel Journal of Mathematics*, 8(3):273–303, 1970.

[Sam65] Ester Samuel. Sequential compound estimators. *The Annals of Mathematical Statistics*, pages 879–889, 1965.

[Sin79] R. S. Singh. empirical Bayes estimation in lebesgue-exponential families with rates near the best possible rate. *The Annals of Statistics*, 7(4):890–902, 1979.

[Tsy09] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Verlag, New York, NY, 2009.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[VR66] John Van Ryzin. The sequential compound decision problem with m× n finite loss matrix. *The Annals of Mathematical Statistics*, pages 954–975, 1966.

[Wol57] J. Wolfowitz. The minimum distance method. *The Annals of Mathematical Statistics*, pages 75–88, 1957.

[WV12] Yihong Wu and Sergio Verdú. Functional properties of MMSE and mutual information. *IEEE Trans. Inf. Theory*, 58(3):1289 – 1301, Mar. 2012.

[YL05] Ming Yuan and Yi Lin. Efficient empirical Bayes variable selection and estimation in linear models. *Journal of the American Statistical Association*, 100(472):1215–1225, 2005.

[Zha97] Cun-Hui Zhang. Empirical bayes and compound estimation of normal means. *Statistica Sinica*, 7(1):181–193, 1997.

[Zha03] Cun-Hui Zhang. Compound decision theory and empirical Bayes methods. *The Annals of Statistics*, 31(2):379–390, 2003.

[Zha09] Cun-Hui Zhang. Generalized maximum likelihood estimation of normal mixture densities. *Statistica Sinica*, pages 1297–1318, 2009.