On the Optimality of Several Algorithms on Polynomial Regression of Empicial Bayes Poisson Model

by

Benjamin Kang

SB, Mathematics and Computer Science and Engineering, MIT, 2024

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

Authored by:	Benjamin Kang Department of Electrical Engineering and Computer Science May 17, 2024
Certified by:	Yury Polyanskiy Professor of Electrical Engineering and Computer Science, Thesis Supervisor
Accepted by:	Katrina LaCurts Chair Master of Engineering Thesis Committee

On the Optimality of Several Algorithms on Polynomial Regression of Empicial Bayes Poisson Model

by

Benjamin Kang

Submitted to the Department of Electrical Engineering and Computer Science
on May 17, 2024 in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

ABSTRACT

The empirical Bayes estimator for the Poisson mixture model in [1], [2] has been an important problem studied for the past 70 years. In this thesis, we investigate extensions of this problem to estimating polynomial functions of the Poisson parameter rather than just the parameter itself. We generalize three different algorithms for estimation, specifically the Robbins estimator from [2], the NPMLE method from [3], and the ERM method from [4]. For each of these algorithms, we prove upper bounds on the minimax regret. We also prove a general lower bound that applies to any estimation algorithm for this setup. In addition to the theoretical bounds, we empirically simulate the performance of all three algorithms in relation to both the number of sample and the degree of the polynomial function we estimate.

Thesis supervisor: Yury Polyanskiy
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

First, I would like to thank Anzo Teh for being a great mentor along every step of this project. I have learned so much about this field from all of our weekly meetings and discussions about different directions to explore. Thank you also for reading drafts of my thesis and proofreading all of the huge equations in my proofs. I would also like to thank Professor Yury Polyanskiy for advising me on this thesis and helping brainstorm future applications of this work.

I am also incredibly thankful to Lauren Chen for all her love and support throughout the year. Finally, I would like to thank my family for always being by my side and getting me to where I am now.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The empirical Bayes estimator is a classic and very powerful technique used in statistics, inference, and machine learning. Such an estimator can be useful in a wide variety of models, and there are also many different techniques to calculate the Bayes estimate in these models. In this thesis, we will examine algorithms for finding an estimator which makes excess loss approach 0 when the sample size increases.

1.1 Model

We focus on mixture models with a known channel but an unknown prior as shown in Fig. 1.1. In particular, there is some unknown prior distribution π from which hidden parameters $\theta_1, \ldots, \theta_n \overset{\text{iid}}{\sim} \pi$ and a known channel γ such that observations are generated according to $X_i \sim \gamma(\theta)$.

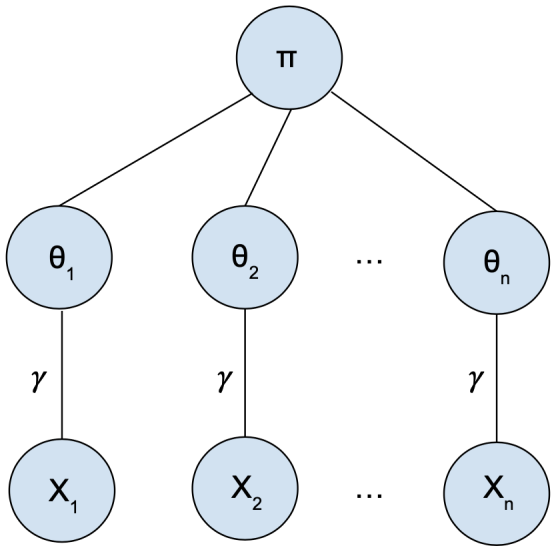

Figure 1.1: Illustration of a Mixture Model

We focus mainly on the mixture model with a Poisson channel, that is, $X_i \sim \text{Poi}(\theta_i)$.

1.2 Estimation on Poisson Mixture Model

In this thesis, we focus on the Poisson model. Past work has focused on the goal of estimating θ given sample data points $X_1, \ldots, X_n$. We would like to extend this goal further: given any smooth function r, estimate $r(\theta)$. Over the rest of this thesis, we will focus on functions of the form $r(\theta) = \theta^k$ for integers $k \geq 1$.

We can compute the marginal distribution of X to be

$$p_\pi(x) = \int e^{-\theta} \frac{\theta^x}{x!} d\pi(\theta). \tag{1.1}$$

For any given x, the Bayes estimator for θ^k that minimizes the squared error is the posterior mean, which we calculate using (1.1) to be

Definition 1 (Bayes Estimator)**.** *The Bayes estimator of θ^k for a prior π is*

$$\begin{aligned} f^*(x) &= \mathbb{E}[\theta^k | X = x] \\ &= \frac{\int \theta^k \left(e^{-\theta} \frac{\theta^x}{x!}\right) d\pi(\theta)}{p_\pi(x)} \\ &= \frac{P(x+k,k) \int e^{-\theta} \frac{\theta^{x+k}}{(x+k)!} d\pi(\theta)}{p_\pi(x)} \\ &= \frac{P(x+k,k) p_\pi(x+k)}{p_\pi(x)} \end{aligned} \tag{1.2}$$

where we let $P(N,j) = \frac{\Gamma(N+1)}{\Gamma(N-j+1)}$ be the permutation number. If $N-j+1$ is a nonpositive integer, define $P(N,j) = 0$. This notation will be useful as it appears in the Bayes estimator and thus in many of our equations. Unfortunately, we do not have access to the true distribution $\pi(\cdot)$ so we cannot calculate the exact value of $f^*(x)$, but we do have access to training samples $x_1, \ldots, x_n$. Our goal is to learn an approximation $\widehat{f}$ of the Bayes estimator. There are many metrics by which we can measure the accuracy of an estimator, but we will focus on the regret, which captures the difference between the mean squared error of the estimator $\widehat{f}$ and the true Bayes estimator f^*. Furthermore, the mean squared error of the Bayes estimator will be important in our bounds later, so we call it the mmse. The two definitions are below.

Definition 2 (mmse)**.** *Let the mmse of a prior distribution π be the expected squared error of the Bayes estimator of θ^k, specifically*

$$\mathsf{mmse}_k(\pi) \triangleq \min_f \mathbb{E}_\pi[(f(X) - \theta^k)^2] = \mathbb{E}_\pi[(f^*(X) - \theta^k)^2].$$

Definition 3 (Regret)**.** *The regret of an estimator f is*

$$\mathsf{Regret}_{\pi,k}(f) = \mathbb{E}\left[\left(f(X) - \theta^k\right)^2\right] - \mathbb{E}\left[\left(f^*(X) - \theta^k\right)^2\right] = \mathbb{E}\left[\left(f(X) - \theta^k\right)^2\right] - \mathsf{mmse}_k(\pi).$$

Expanding, we can also obtain

$$\mathsf{Regret}_{\pi,k}(f) = \mathbb{E}\left[f(X)^2 - 2\theta^k f(X) - f^*(X)^2 + 2\theta^k f^*(X)\right].$$

Since $\mathbb{E}[\theta^k|X] = f^*(X)$, the Tower rule of Expectation tells us that

$$\begin{aligned}\mathsf{Regret}_{\pi,k}(f) &= \mathbb{E}\left[\mathbb{E}\left[f(X)^2 - 2\theta^k f(X) - f^*(X)^2 + 2\theta^k f^*(X)|X\right]\right] \\ &= \mathbb{E}\left[f(X)^2 - 2f(X)f^*(X) + f^*(X)^2\right] \\ &= \mathbb{E}\left[(f(X) - f^*(X))^2\right].\end{aligned}$$

All three forms of the regret will be useful later in this thesis.

1.3 Literature Review

Empirical Bayes estimators on these mixture models have been studied for many years now. As mentioned previously, all of the related works focus on the $k = 1$ case. One of the earliest papers also studying the Poisson mixture model was introduced by Robbins[1], [2], where it was shown that the true Bayes estimator is

$$f^*(x) = \frac{(x+1)p_\pi(x+1)}{p_\pi(x)}$$

and is monotonic. Note that this matches Definition 1 with $k = 1$.

1.3.1 Robbins Estimator

Robbins[2] first proposed an estimator

$$\widehat{f}_{\text{Rob}}(x) = \frac{(x+1)N(x+1)}{N(x)}$$

where $N(x)$ represents the number of occurrences of x in the data set $x_1, \ldots, x_n$. This essentially approximates p_π with the empirical distribution. Such an approach has been called f-modeling[5], and it has been shown that the Robbins estimator achieves optimal regret for π that is bounded or subexponential[6], [7]. However, the Robbin's estimator can be very unstable (e.g. when x is large and the counts are small, small changes in counts greatly affect the estimated values)[8]. Furthermore, the Robbin's estimator may not be monotonic, which is a desired property of the empirical Bayes estimator[9]. However, there have been modified versions of this algorithm where monotonicity can be imposed without increasing its regret[10].

1.3.2 Non-parametric Maximum Likelihood Estimation

A different approach first proposed in [3] is to approximate π rather than p_π. To do this, a maximum likelihood estimator (MLE) is used. Specifically, we approximate

$$\widehat{\pi} = \operatorname*{argmax}_Q \prod_{i=1}^n p_Q(x_i)$$

and then calculate $\widehat{f} = \frac{(x+1)p_{\widehat{\pi}}(x+1)}{p_{\widehat{\pi}}(x)}$. Since no parametric form of Q is assumed, this prior is determined through Non-parametric Maximum Likelihood Estimation, which we refer to as NPMLE for the rest of the thesis. This approach has been termed g-modeling[5], and the resulting estimator has been shown to achieve optimal regret for bounded and subexponential distributions[11]. In addition, g-modeling achieves optimal regret on polynomial tailed prior distributions π, while the Robbin's estimator has been proven to be suboptimal[12].

Due to the Bayesian structure of $\widehat{f}_{\text{NPMLE}}$, the desired property of monotonicity is still preserved. Furthermore, the estimated values are more stable than those in f-modeling, and optimality is satisfied even for heavy polynomial tailed distributions while it is not satisfied with f-modeling. Experiments have also shown that the NPMLE estimator can also be useful as a preprocessing method for data analysis. However, there are tradeoffs to using this method. The optimization problem to be solved is very difficult and computationally expensive, especially when the dimensions increase as the time is exponential in d.

1.3.3 Empirical Risk Minimizer

The newest methodology proposed is an estimator proposed by [4] based on the Empirical Risk Minimizer, which we refer to as ERM in the remainder of this thesis. The ERM is an idea in statistical learning theory first introduced in [13] where an optimal hypothesis is learned by finding the hypothesis with the smallest loss over the empirical data.

Motivated by this idea, rather than approximating the prior or the posterior distribution, we can directly solve for the Bayes estimator. This is done by first noticing that the Bayes estimator minimizes the mean squared error. This naturally leads to finding the estimator which minimizes the empirical mean squared error, which is the ERM solution. Since we are minimizing over a set of estimators, we can impose constraints on the set we search over. This leads to [4, (Equation 6)], where

$$\widehat{f}_{\mathsf{erm}} \in \underset{f \in \mathcal{F}_{\text{monotone}}}{\operatorname{argmin}} \ \widehat{\mathbb{E}}[f(X) - 2Xf(X-1)].$$

In this way, the ERM estimator maintains the desired monotonicity of the Bayes estimator similar to g-modeling. The framework for calculating function is also very flexible and extra constraints or different function classes can easily be implementable. Performance guarantees in the bounded and subexponential prior case match those of both f and g-modeling. Moreover, this minimization problem can be solved using isotonic regression, so the computational cost is much lower. However, regret in the heavier tailed prior case is still unknown.

Furthermore, [4] investigated the d-dimensional version of this problem. We formulate the problem as follows: an unknown set of vector parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ are independently sampled from a multidimensional distribution π. Then, samples $\boldsymbol{x}_i$ are generated such that $x_{ij} \sim \text{Poi}(\theta_{ij})$. We similarly aim to achieve an estimator $\widehat{\boldsymbol{f}}(\mathbf{x})$ to minimize the regret, which is

$$\mathsf{Regret}_{\pi}(\boldsymbol{f}) \triangleq \mathbb{E}\left[\left\|\widehat{\boldsymbol{f}}(\boldsymbol{x}) - \boldsymbol{\theta}\right\|^2\right] - \mathbb{E}\left[\|\boldsymbol{f}^*(\boldsymbol{x}) - \boldsymbol{\theta}\|^2\right].$$

The generalization of the one dimensional monotonicity condition becomes the following.

For each i,

$$\boldsymbol{f}^*(\boldsymbol{x}+\mathbf{e}_i) \geq \boldsymbol{f}^*(\boldsymbol{x}) \tag{1.3}$$

where $\mathbf{e}_i$ is the vector with 1 in the ith coordinate and 0 everywhere else. The ERM estimator can now be written as

$$\widehat{\boldsymbol{f}}_{\mathsf{erm}}(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{f}\in\mathcal{F}} \widehat{\mathbb{E}}\left[\|\boldsymbol{f}(\boldsymbol{x})\|^2 - 2\sum_{i=1}^{d} x_j f_j(\boldsymbol{x}-\mathbf{e}_i)\right]$$

and the class of functions $\mathcal{F}$ is all functions $\mathbb{Z}_+^d \to \mathbb{R}_+^d$ satisfying the generalized monotonicity constraint in Equation (1.3). In this setup, the time complexity of ERM will be polynomial in n and d while NPMLE can take up to $n^{\Theta(d)}$ [14], making erm the much more scalable algorithm.

Regret bounds on the multidimensional Poisson case has also been calculated for priors with bounded supports and subexponential marginals in [4]. It is not yet known if these match the lower bound, but it is conjectured that a better lower bound can be calculated and this algorithm has nearly optimal regret. Experiments in 2 dimensions have also been run showing that the multidimensional ERM estimator runs many times faster than the g-modeling approach, and the regret is also significantly lower than f-modeling.

1.3.4 Lower Bounds

In addition to the above three algorithms, lower bounds have also been investigated for estimation on the Poisson mixture model. In fact, [15] proved a lower bound matching the regret upper bounds of all three algorithms mentioned above in the case of a bounded and subexponential prior. The general idea for lower bounding is to start with a gamma prior, consider the set of distributions which are close to this prior, and show that it is difficult to learn significant information within this set of distributions.

1.3.5 Other Mixture Distributions

Although a lot of the work in this field has been focused on the Poisson case, there have been multiple papers on the normal location model. Both the f[16] and g-modeling[17] approaches have been shown to obtain a nearly optimal fast rate of regret. However, the analogous ERM estimator has only been proven to achieve a slow rate for regret in [18], and it is still unknown whether a faster rate can be achieved.

1.4 Optimal Minimax Regret

This thesis will examine the estimation of θ^k in a Poisson mixture model given two different regimes of priors: bounded and subexponential.

Definition 4 (Bounded Distribution)**.** *We use $\mathcal{P}([0,h])$ to denote the set of all probability distributions that can only achieve values in the range $[0,h]$.*

Definition 5 (Subexponential Distribution)**.** *We use* $\mathsf{SubE}(s)$ *to denote the set of all probability distributions that satisfy the tail bound* $\mathbb{P}(x \geq t) \leq 2e^{-t/s}$ *for all* $t > 0$*.*

Over the course of this thesis, we will describe multiple different estimators of θ^k by generalizing the three different algorithms (f-modeling, g-modeling, and erm). We will also prove regret bounds for each along with optimal minimax regret bounds to see which algorithms can attain them.

Theorem 1. *The minimax regrets are as follows:*

1. *For any bounded prior* $\pi \in \mathcal{P}([0, h])$*, the optimal minimax regret is*

$$\Theta_{h,k}\left(\frac{1}{n}\left(\frac{\log n}{\log\log n}\right)^{k+1}\right)$$

2. *For any bounded prior* $\pi \in \mathsf{SubE}(s)$*, the optimal minimax regret is*

$$\Theta_{s,k}\left(\frac{1}{n}(\log n)^{2k+1}\right).$$

The three algorithms from Section 1.3.1, Section 1.3.2, and Section 1.3.3 achieve the regrets shown in Table 1.1

Algorithm	Bounded Prior	Subexponential Prior
Robbins	$O_{h,k}\left(\frac{1}{n}\left(\frac{\log n}{\log\log n}\right)^{k+1}\right)$	$O_{s,k}\left(\frac{1}{n}(\log n)^{2k+1}\right)$
NPMLE	$O_{h,k}\left(\frac{1}{n}\left(\frac{\log n}{\log\log n}\right)^{k+1}\right)$	$O_{s,k}\left(\frac{1}{n}(\log n)^{2k+1}\right)$
ERM	$O_{h,k}\left(\frac{1}{n}\left(\frac{\log n}{\log\log n}\right)^{2k}\right)$	$O_{s,k}\left(\frac{1}{n}(\log n)^{2k+1}\right)$

Table 1.1: Minimax Regrets of each algorithm.

The optimal rate on the bounded priors can be achieved by a generalized version of the Robbins and NPMLE estimator, while the optimal rate on subexponential priors can be achieved by all three algorithms.

Proof. All upper bounds and the lower bound will be established in the later sections. □

Chapter 2

Upper Bound on Polynomial Robbins Regret

2.1 Modified Robbins Estimator

In this section, we introduce a natural extension to the original version of Robbin's estimator for θ. From Definition 1, we know the form of the true Bayes Estimator. Motivated by the Robbin's estimator for the estimation of θ, we can again estimate $p_\pi(x)$ using the empirical distribution, giving us the estimator

$$\widehat{f}_{\mathsf{Rob},k} = \frac{P(x+k,k)N(x+k)}{N(x)}. \tag{2.1}$$

Theorem 2. *The Robbins estimator for θ^k defined in (2.1) satisfies the following regret bounds:*

1. $$\sup_{\pi \in \mathcal{P}([0,h])} \mathsf{Regret}_{\pi,k}(\widehat{f}_{\mathsf{Rob},k}) \leq O_{h,k}\left(\frac{1}{n}\left(\frac{\log n}{\log\log n}\right)^{k+1}\right)$$

2. $$\sup_{\pi \in \mathsf{SubE}(s)} \mathsf{Regret}_{\pi,k}(\widehat{f}_{\mathsf{Rob},k}) \leq O_{s,k}\left(\frac{1}{n}(\log n)^{2k+1}\right).$$

2.2 General Regret Upper Bound via Robbins

We first start with the following lemma, which we use to bound the regret on a prior bounded by $[0,h]$.

Lemma 1. *For a prior in $\mathcal{P}([0,h])$ (here h may depend on n), the regret satisfies for some constant $c = c(k)$,*

$$\begin{aligned}&\mathsf{Regret}_{\pi,k}(\widehat{f}_{\mathsf{Rob},k})\\ \leq&\frac{c}{n}\left(\max\{h^{2k},1\}+\sum_{x\geq 1}P(x+k,k)h^k\min\left\{n^2p_\pi(x)^2,1\right\}+h^{2k}\min\left\{np_\pi(x),1\right\}\right).\end{aligned}$$

Proof. Using the last definition in Definition 3, we have

$$\begin{aligned}n\cdot\mathsf{Regret}_{\pi,k}(\widehat{f}_{\mathsf{Rob},k})&=\mathbb{E}\left[\sum_{i=1}^n(\widehat{f}_{\mathsf{Rob},k}(X_i)-f^*(X_i))^2\right]\\&=\mathbb{E}\left[\sum_{x\geq 0}N(x)P(x+k,k)^2\left(\frac{N(x+k)}{N(x)}-\frac{p_\pi(x+k)}{p_\pi(x)}\right)^2\mathbf{1}_{N(x)>0}\right]\\&=\mathbb{E}\left[\sum_{x\geq 0}\frac{\mathbf{1}_{N(x)>0}P(x+k,k)^2}{N(x)}\left(N(x+k)-\frac{p_\pi(x+k)N(x)}{p_\pi(x)}\right)^2\right].\end{aligned}\tag{2.2}$$

Similar to [4, (P1)-(P4)], we have

(P1) $N(x)\sim\mathrm{Binom}(n,p_\pi(x))$ and for some absolute constants $c',c_2>0$ [15, Lemma 16]

$$\mathbb{E}\left[\frac{\mathbf{1}_{N(x)>0}}{N(x)}\right]\leq c'\min\left\{np_\pi(x),\frac{1}{np_\pi(x)}\right\},\quad\mathbb{E}\left[\frac{\mathbf{1}_{N(x)>0}}{N(x)}(N(x)-np_\pi(x))^2\right]\leq c_2.$$

(P2) Conditioned on $N(x)$, $N(x+k)\sim\mathrm{Binom}\left(n-N(x),\frac{p_\pi(x+k)}{1-p_\pi(x)}\right)$

(P3) $f^*(x)=\frac{P(x+k,k)p_\pi(x+k)}{p_\pi(x)}=\mathbb{E}[\theta^k|x]\leq h^k$ for all $x\geq 0$,

(P4) Stirling's method entails $\frac{x^ye^{-x}}{y!}\leq\frac{y^ye^{-y}}{y!}\leq\frac{1}{\sqrt{2\pi y}}$. Therefore, $p_\pi(y)\leq\frac{1}{\sqrt{2\pi y}}$, $\quad y\geq 1$.

Let $q=\frac{p_\pi(x+k)}{1-p_\pi(x)}$. Then by (P2) and the bias variance decomposition,

$$\begin{aligned}&\mathbb{E}\left[\left(N(x+k)-\frac{p_\pi(x+k)N(x)}{p_\pi(x)}\right)^2|N(x)\right]\\=&(n-N(x))q(1-q)+\left((n-N(x))q-\frac{p_\pi(x+k)N(x)}{p_\pi(x)}\right)^2\\\leq&n\frac{p_\pi(x+k)}{1-p_\pi(x)}+\left(\frac{p_\pi(x+k)}{(1-p_\pi(x))p_\pi(x)}\right)^2(np_\pi(x)-N(x))^2.\end{aligned}\tag{2.3}$$

Now by (P4), we see that for $x \geq 1$, $\frac{1}{1-p_\pi(x)} \leq c$ for the constant $c \triangleq \frac{\sqrt{2\pi}}{\sqrt{2\pi}-1}$. Then using (2.3), we can continue (2.2) to get

$$\begin{aligned}
&n \cdot \mathsf{Regret}_{\pi,k}(\widehat{f}_{\mathsf{Rob},k}) \\
=&\mathbb{E}\left[\sum_{x\geq 0} \frac{\mathbf{1}_{N(x)>0} P(x+k,k)^2}{N(x)} \left(n\frac{p_\pi(x+k)}{1-p_\pi(x)} + \left(\frac{p_\pi(x+k)}{(1-p_\pi(x))p_\pi(x)}\right)^2 (np_\pi(x) - N(x))^2 \right)\right] \\
=&\mathbb{E}\Bigg[\frac{\mathbf{1}_{N(0)>0}(k!)^2}{N(0)} \left(n\frac{p_\pi(k)}{1-p_\pi(0)} + \left(\frac{p_\pi(k)}{(1-p_\pi(0))p_\pi(0)}\right)^2 (np_\pi(0) - N(0))^2 \right) \\
&+ \sum_{x\geq 1} \frac{\mathbf{1}_{N(x)>0} P(x+k,k)^2}{N(x)} \left(cnp_\pi(x+k) + c^2 \left(\frac{p_\pi(x+k)}{p_\pi(x)}\right)^2 (np_\pi(x) - N(x))^2 \right)\Bigg] \\
\leq&(k!)^2 \left(\frac{c' p_\pi(k)}{(1-p_\pi(0))p_\pi(0)} + c_2 \left(\frac{p_\pi(k)}{(1-p_\pi(0))p_\pi(0)}\right)^2 \right) \\
&+ \sum_{x\geq 1} cc' P(x+k,k) f^*(x) \min\left\{n^2 p_\pi(x)^2, 1\right\} + c^2 c_2 f^*(x)^2 \min\left\{np_\pi(x), 1\right\}. \qquad (2.4)
\end{aligned}$$

Note that

$$\frac{p_\pi(k)}{(1-p_\pi(0))p_\pi(0)} = \frac{\max\left\{\frac{p_\pi(k)}{p_\pi(0)}, \frac{p_\pi(k)}{1-p_\pi(0)}\right\}}{\max\{p_\pi(0), 1-p_\pi(0)\}} \overset{(a)}{\leq} 2\max\left\{\frac{h^k}{k!}, 1\right\} \leq 2\max\{h^k, 1\} \qquad (2.5)$$

where (a) is due to (P3) ($\frac{p_\pi(k)}{p_\pi(0)} \leq \frac{h^k}{k!}$), $p_\pi(k) = 1 - \sum_{x\neq k} p_\pi(x) \leq 1 - p_\pi(0)$, and $\max\{p_\pi(0), 1-p_\pi(0)\} \geq \frac{1}{2}$. Substituting (2.5) and (P3) back into (2.4) and then dividing by n, we obtain for some constant $c_3 = c_3(k)$, the desired inequality

$$\begin{aligned}
&\mathsf{Regret}_{\pi,k}(\widehat{f}_{\mathsf{Rob},k}) \\
\leq&\frac{c_3}{n}\left(\max\{h^{2k}, 1\} + \sum_{x\geq 1} P(x+k,k) h^k \min\left\{n^2 p_\pi(x)^2, 1\right\} + h^{2k} \min\left\{np_\pi(x), 1\right\} \right).
\end{aligned}$$

□

To allow us to further bound the RHS of Lemma 1, we prove the following tail bounds on p_π to bound the min using $p_\pi(x)$ for large x.

Lemma 2. *We prove tail bounds on p_π for the two types of distributions.*

- *Let $\pi \in \mathcal{P}([0,h])$. Then for some constant $c = c(h)$ and $x_0 = \max\left(2h, c\frac{\log n}{\log\log n}\right)$,*

$$\sum_{x > x_0 + k} p_\pi(x)^2 P(x+k,k) \leq \frac{2^{k+1} h^k}{n^2}. \qquad (2.6)$$

- *Let $\pi \in subE(s)$. Then for some constant $c = c(s)$ and $x_1 = c\log n$,*

$$\sum_{x > x_1 + k} p_\pi(x)^2 P(x+k,k) \leq \frac{1}{n^2}. \qquad (2.7)$$

A proof of this lemma can be found in Appendix A.

2.3 Regret of Truncated Prior

Lemma 1 is sufficient for any bounded distributions. However, to obtain regret bounds on subexponential distributions, we need to relate its regret to the regret of a truncated version of this prior. For any distribution π, let π_h be π restricted to the range $[0, h]$. We first begin with a helpful lemma regarding the mmse of a truncated prior.

Lemma 3. *For any π and h, $\mathsf{mmse}_k(\pi_h) \le \frac{\mathsf{mmse}_k(\pi)}{\mathbb{P}_\pi(\theta \le h)}$.*

Proof. Let E be the event that $\theta \le h$ under π. Then

$$\mathsf{mmse}_k(\pi) = \min_f \mathbb{E}_\pi[(f(X) - \theta^k)^2] \ge \min_f \mathbb{E}_\pi[(f(X) - \theta^k)^2 | E]\mathbb{P}[E] = \mathsf{mmse}_k(\pi_h)\mathbb{P}(E).$$

□

In addition, we prove the following lemma which bounds the moments of θ and $X_{\max} = \max(X_1, \dots, X_n)$ for a subexponential prior. These results will help us in bounding the regrets in the tails of subexponential distributions.

Lemma 4. *Let $\pi \in \mathsf{SubE}(s)$ and $X \sim p_\pi$. There exists some constant $C(k, s)$ such that*

$$\mathbb{E}_\pi[\theta^{4k}] \le 8k(4k-1)!s^{4k} \qquad \mathbb{E}[X_{\max}^\ell] \le C(k, s)(\log n)^\ell \tag{2.8}$$

for all $\ell \le 4k$.

A proof of this lemma relies mainly on rewriting the expectation using tail probabilities and bounding those values. The exact proof is given in Appendix A.

Recall the notation of the truncated prior π_h defined before Lemma 3. With these lemmas, we are now ready to show that the regret over the truncated prior $\pi_{c_1 s \log n}$ exceeds the regret over π by at most $o_{s,k}(1/n)$.

Lemma 5. *For any estimator $\widehat{f}$ such that $\mathbb{E}[\widehat{f}(X)^4] = O_{s,k}(n^4(\log n)^{4k})$, there exists constants $c_1, c_2, c_3 > 0$ such that*

$$\mathsf{Regret}_{\pi,k}(\widehat{f}) \le \mathsf{Regret}_{\pi_{c_1 s \log n},k}(\widehat{f}) + o_{s,k}\left(\frac{1}{n}\right).$$

Proof. Let $\pi \in \mathsf{SubE}(s)$, then there exists a constant $c(s) \triangleq 11s$ by Definition 5 such that

$$\epsilon = \mathbb{P}[\theta > c(s) \log n] \le \frac{1}{n^{10}}, \quad \theta \sim \pi.$$

Let E be the event $\{\theta_i \le c(s) \log n, \forall i = 1, \dots, n\}$. By union bounding, $\mathbb{P}[E^c] \le n^{-9}$. Recall the regret from Definition 3. We obtain the following series of equations:

$$\begin{aligned}
\mathsf{Regret}_{\pi,k}(\widehat{f}) &= \mathbb{E}_\pi[(\widehat{f}(x) - \theta^k)^2] - \mathsf{mmse}_k(\pi) \\
&\le \mathbb{E}_\pi[(\widehat{f}(x) - \theta^k)^2 | E] - \mathsf{mmse}_k(\pi_{c_1 s \log n}) + \mathsf{mmse}_k(\pi_{c_1 s \log n}) \\
&\quad - \mathsf{mmse}_k(\pi) + \mathbb{E}_\pi[(\widehat{f}(x) - \theta^k)^2 \mathbf{1}_{E^c}] \\
&= \mathsf{Regret}_{\pi_{c_1 s \log n},k}(\widehat{f}) + \mathsf{mmse}_k(\pi_{c_1 s \log n}) - \mathsf{mmse}_k(\pi) + \mathbb{E}_\pi[(\widehat{f}(x) - \theta^k)^2 \mathbf{1}_{E^c}].
\end{aligned} \tag{2.9}$$

For the last term of (2.9), applying Cauchy-Schwarz gives

$$\mathbb{E}_\pi[(\widehat{f}(x)-\theta_1^k)^2\mathbf{1}_{E^c}] \le \sqrt{\mathbb{P}[E_c]\mathbb{E}_\pi[(\widehat{f}(X)-\theta^k)^4]} \le \sqrt{n^{-9}\mathbb{E}_\pi[(\widehat{f}(X)-\theta^k)^4]} \le \frac{c(s,k)}{n}$$

since $\mathbb{E}[\widehat{f}(X)^4] = O_{s,k}(n^4(\log n)^{4k})$ and $\mathbb{E}_\pi[\theta^{4k}] = O_{s,k}(1)$ by Lemma 4. For the middle two terms of (2.9), Lemma 3 tells us that

$$\mathsf{mmse}_k(\pi) \ge \mathsf{mmse}_k(\pi_{c_1 s\log n})(1-n^{-9})$$

so we have

$$\mathsf{mmse}_k(\pi_{c_1 s\log n}) - \mathsf{mmse}_k(\pi) \le \frac{n^{-9}}{1-n^{-9}}\mathsf{mmse}_k(\pi) \le 2c(k)s^{4k}n^{-9} = o_{s,k}(n^{-1})$$

by $n^{-9} \le \frac{1}{2}$ and Lemma 4. Combining these inequalities together, we obtain the desired result. □

2.4 Proof of Theorem 2

Now we are ready to apply Lemma 1 to prove the original result.

Proof. Let us first deal with the bounded prior case. Split up the summation in Lemma 1 at x_0+k from Lemma 2 and upper bound the min's to obtain

$$\begin{aligned} n\cdot\mathsf{Regret}_{\pi,k}(\widehat{f}) &\le \max\{h^{2k},1\} + \sum_{x=1}^{x_0+k}\left(P(x+k,k)h^k + h^{2k}\right) \\ &\quad + \sum_{x>x_0+k}\left(P(x+k,k)h^k n^2 p_\pi(x)^2 + h^{2k}np_\pi(x)\right) \\ &\overset{(a)}{\le} \max\{h^{2k},1\} + h^{2k}(x_0+k) + h^k(x_0+k)P(x_0+2k,k) + 2^{k+1}h^{2k} + h^{2k} \\ &\overset{(b)}{=} O_{h,k}\left(\left(\frac{\log n}{\log\log n}\right)^{k+1}\right) \end{aligned}$$

where (a) is from (2.6) and [15, (Equation 122)], and (b) is from plugging in x_0. Dividing by n yields the desired result.

Now we move onto the subexponential case. By Lemma 5, it suffices to take $h = c_1 s\log n$ and bound the regret on π_h. Split up the summation in Lemma 1 at x_1+k from Lemma 2 (where $x_1 \triangleq c(s)\log n$ is as defined in Lemma 2) and upper bound the min's to obtain

$$\begin{aligned} n\cdot\mathsf{Regret}_{\pi_h,k}(\widehat{f}) &\le h^{2k} + \sum_{x=1}^{x_1+k}\left(P(x+k,k)h^k + h^{2k}\right) \\ &\quad + \sum_{x>x_1+k}\left(P(x+k,k)h^k n^2 p_\pi(x)^2 + h^{2k}np_\pi(x)\right) \\ &\overset{(a)}{\le} h^{2k} + h^{2k}(x_1+k) + h^k(x_1+k)P(x_1+2k,k) + h^k + h^{2k} \\ &\overset{(b)}{=} O_{s,k}\left((\log n)^{2k+1}\right) \end{aligned}$$

where (a) is from (2.7) and [15, (Equation 124)], and (b) is from plugging in x_1. Dividing by n yields the desired result. □

Chapter 3

Upper Bound on Polynomial NPMLE Regret

3.1 NPMLE Algorithm

In this section, we extend the NPMLE algorithm first introduced in [3] to estimate θ^k. This extension is very natural, as the estimation of the prior distribution remains the same. Using this estimated prior, we calculate its empirical Bayes estimator by applying Definition 1.

Before proving the main results, we first discuss how the the prior estimation via NPMLE. This method is a specific instance of the more general class of minimum distance estimators. These estimators are defined by a measure of distance between two distributions. There are many possibilities of distance functions, so we will focus on a specific set of functions which we call *Generalized distance functions*.

Definition 6 (*generalized distance functions*)**.** *A function $d : \mathcal{P}(\mathbb{Z}_+) \times \mathcal{P}(\mathbb{Z}_+) \to \mathbb{R}_+$ such that $d(p \parallel q) \geq 0$ with equality true if and only if $p = q$.*

Note that this includes any metrics and divergence. Then a minimum distance estimator with respect to d over a set of distributions $\mathcal{G}$ is

$$\widehat{\pi} \in \operatorname*{argmin}_{\pi \in \mathcal{G}} d(p_n^{\text{emp}} \parallel p_\pi).$$

In [11], they describe some specific examples of these distance functions correspond with well known estimators, including the NPMLE estimator we focus on:

- The NPMLE estimator corresponds to the KL-divergence $d(p \parallel q) = \mathrm{KL}(p \parallel q) = \mathbb{E}\left[\log \frac{p(x)}{q(x)}\right]$.
- The Minimum-Hellinger estimator corresponds to the squared Hellinger distance $d(p \parallel q) = H^2(p, q) = \sum \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2$.
- The Minimum-χ^2 estimator corresponds to the χ^2-divergence $d(p \parallel q) = \chi^2(p \parallel q) = \sum \frac{(p(x)-q(x))^2}{q(x)}$.

It turns out that we can actually prove regret bounds on our algorithm for not only NPMLE, but any minimum distance estimator using a generalized distance function that satisfies Assumptions 1 and 2 in [11]:

Assumption 1. *There exists a map $t : \mathcal{P}(\mathbb{Z}_+) \to \mathbb{R}$ and $\ell : \mathbb{R}^2 \to \mathbb{R}$ such that for any two distributions $p, q \in \mathcal{P}(\mathbb{Z}_+)$,*

$$d(p \parallel q) = t(p) + \sum_{x \geq 0} \ell(p(x), q(x))$$

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

Assumption 2. *There exist positive constants c_1, c_2 such that for $p, q \in \mathcal{P}(\mathbb{Z}_+)$,*

$$c_1 H^2(p, q) \leq d(p \parallel q) \leq c_2 \chi^2(p \parallel q).$$

The KL-divergence satisfies these assumptions, so the following theorem applies to the NPMLE estimator as well. The main result of this section is that a minimum distance estimator satisfies the following regret bounds:

Theorem 3. *Suppose d satisfies Assumptions 1 and 2. For a fixed h and s, the following regret bounds hold:*

1. *If $\widehat{\pi} = \operatorname{argmin}_{\pi \in \mathcal{P}([0,h])} d(p_n^{emp} \parallel p_\pi)$ and $\widehat{f}$ is the Bayes estimator for the prior $\widehat{\pi}$, then for any $n \geq 3$*

$$\sup_{\pi \in \mathcal{P}([0,h])} \mathsf{Regret}_{\pi,k}(\widehat{f}) = O_{h,k}\left(\frac{1}{n}\left(\frac{\log n}{\log\log n}\right)^{k+1}\right)$$

2. *If $\widehat{\pi} = \operatorname{argmin}_{\pi} d(p_n^{emp} \parallel p_\pi)$ and $\widehat{f}$ is the Bayes estimator for the prior $\widehat{\pi}$, then for any $n \geq 2$*

$$\sup_{\pi \in \mathsf{SubE}(s)} \mathsf{Regret}_{\pi,k}(\widehat{f}) = O_{s,k}\left(\frac{1}{n}(\log n)^{2k+1}\right).$$

3.2 General Regret Upper Bound via NPMLE

To prove Theorem 3, we will use the following more general lemma bounding the regret using the Hellinger distance. Then we can use the Hellinger distance results about the NPMLE prior estimate to prove Theorem 3.

Lemma 6. *Let π be a distribution such that $\mathbb{E}_\pi[\theta^{4k}] \leq M$ for some constant M. Then for any distribution $\widehat{\pi}$ supported on $[0, \widehat{h}]$, any $h > 0$ with $\mathbb{P}_\pi(\theta \leq h) > \frac{1}{2}$ and any $K \geq 1$,*

$$\begin{aligned}\mathsf{Regret}_{\pi,k}(\widehat{f}) \leq &\left\{12(h^{2k} + \widehat{h}^{2k}) + 48(h^k + \widehat{h}^k)K^k\right\}(H^2(p_\pi, p_{\widehat{\pi}}) + 4\mathbb{P}_\pi(\theta > h)) \\ &+ 2(h^k + \widehat{h}^k)^2 \mathbb{P}_{p_\pi}(X > K - k) + 2(1 + 2\sqrt{2})\sqrt{(M + \widehat{h}^{4k})\mathbb{P}_\pi(\theta > h)}\end{aligned}$$

where $\widehat{f}$ is the Bayes estimator for the prior $\widehat{\pi}$.

A proof of this lemma is provided in Appendix B. Now equipped with Lemma 6, we can find a suitable application to prove the original result.

3.3 Proof of Theorem 3

Proof. For any $\pi \in \mathcal{P}([0,h])$, we have $\widehat{\pi} = \operatorname{argmin}_{\pi \in \mathcal{P}([0,h])} d(p_n^{\mathrm{emp}} \parallel p_\pi)$. We apply Lemma 6 with

$$\widehat{h} = h, \quad M = h^4, \quad K = \left\lceil \frac{2(2+he)\log n}{\log\log n} + k - 1 \right\rceil.$$

As shown in the proof of [11, Theorem 2(a)], we have

$$\mathbb{P}_{p_\pi}\left(X \geq K - k + 1\right) \leq \frac{2}{n^2}, \quad \mathbb{P}_\pi(\theta > h) = 0, \quad \mathbb{E}[H^2(f^*, \widehat{f})] = \frac{C_1}{n}\left(\frac{\log n}{\log\log n}\right)$$

for some constant C_1. Thus, we have

$$\begin{aligned}
&\mathsf{Regret}_{\pi,k}(\widehat{f}) \\
\leq &\left\{ 24h^{2k} + 96h^k \left\lceil \frac{2(2+he)\log n}{\log\log n} + k - 1 \right\rceil^k \right\} \mathbb{E}[H^2(p_\pi, p_{\widehat{\pi}})] + 2(2h^k)^2 \mathbb{P}_{p_\pi}(X > K - k) \\
= &O_{h,k}\left(\frac{1}{n} \left(\frac{\log n}{\log\log n} \right)^{k+1} \right).
\end{aligned}$$

For any $\pi \in \mathsf{SubE}(s)$, we have $\widehat{\pi} = \operatorname{argmin}_\pi d(p_n^{\mathrm{emp}} \parallel p_\pi)$. By [11, Lemma 8], $\widehat{\pi}$ is supported on $[0, \widehat{h}]$ where $\widehat{h} = X_{\max}$. By the bound on $\mathbb{E}[\theta^{4k}]$ in Lemma 4, we can apply Lemma 6 with

$$h = 4s\log n, \quad M = 8k(4k-1)!s^{4k}, \quad K = \frac{2\log n}{\log\left(1 + \frac{1}{2s}\right)} + k - 1.$$

By [11, (Equation 44)] and Definition 5, we have

$$\mathbb{P}_{p_\pi}\left(X \geq \frac{2\log n}{\log\left(1+\frac{1}{2s}\right)} \right) \leq \frac{3}{2n^2}, \quad \mathbb{P}_\pi(\theta > h) \leq \frac{2}{n^4}.$$

Thus, we have

$$\begin{aligned}
\mathsf{Regret}_{\pi,k}(\widehat{f}) &\leq \mathbb{E}\left[\left\{ 12(h^{2k} + X_{\max}^{2k}) + 48(h^k + X_{\max}^k)K^k \right\} H^2(p_\pi, p_{\widehat{\pi}}) + \frac{2}{n^4} \right] \\
&\quad + 2(h^k + X_{\max}^k)^2 \frac{3}{2n^2} + (1+2\sqrt{2})\sqrt{(M + \widehat{h}^{4k})\frac{2}{n^4}} \\
&= \mathbb{E}\left[\left\{ 12(h^{2k} + X_{\max}^{2k}) + 48(h^k + X_{\max}^k)K^k \right\} H^2(p_\pi, p_{\widehat{\pi}}) \right] + O_{s,k}\left(\frac{1}{n^2}\right). \qquad (3.1)
\end{aligned}$$

It remains to bound the first term in (3.1). Splitting into cases by comapring $X_{\max}$ to $2K$ and using $X_{\max}\, H^2 \leq 2$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\{ h^{2k} + X_{\max}^{2k} + 4(h^k + X_{\max}^k)K^k \right\} H^2(p_\pi, p_{\widehat{\pi}}) \right] \\
\leq &(h^{2k} + (2K)^{2k} + 4(h^k + (2K)^k)K^k)\mathbb{E}\left[H^2(p_\pi, p_{\widehat{\pi}}) \right] \\
&+ 2\mathbb{E}\left[\left\{ h^{2k} + X_{\max}^{2k} + 4(h^k + X_{\max}^k)K^k \right\} \mathbb{I}_{X_{\max} \geq 2K} \right]. \qquad (3.2)
\end{aligned}$$

By [11, Theorem 2(b)], we know $\mathbb{E}\left[H^2(p_\pi, p_{\widehat{\pi}})\right] = O_s\left(\frac{\log n}{n}\right)$. Then plugging in both h and K tells us the first term of the RHS is $O_{s,k}\left(\frac{(\log n)^{2k+1}}{n}\right)$. By Cauchy-Schwarz, the second term of the RHS is

$$\begin{aligned}
&2\mathbb{E}\left[\left\{h^{2k} + X_{\max}^{2k} + 4(h^k + X_{\max}^k)K^k\right\}\mathbb{I}_{X_{\max}\geq 2K}\right] \\
\leq &2\sqrt{\mathbb{E}\left[\left\{h^{2k} + X_{\max}^{2k} + 4(h^k + X_{\max}^k)K^k\right\}^2\right]\mathbb{P}_{p_\pi}\left(X_{\max}\geq 2K\right)} \\
\leq &2\sqrt{\mathbb{E}\left[\left\{h^{2k} + X_{\max}^{2k} + 4(h^k + X_{\max}^k)K^k\right\}^2\right]n\mathbb{P}_{p_\pi}\left(X\geq 2K\right)} \\
\overset{(a)}{\leq} &2\sqrt{\mathbb{E}\left[\left\{4h^{4k} + 4X_{\max}^{4k} + 64(h^{2k} + X_{\max}^{2k})K^{2k}\right\}^2\right]n\frac{3}{2n^4}} \\
\overset{(b)}{=} &o_{s,k}\left(\frac{1}{n}\right)
\end{aligned}$$

where (a) follows from [11, (Equation 44)] and (b) follows from the moment bounds on $X_{\max}$ in Lemma 4. Plugging this back into (3.2) and then (3.1), we obtain

$$\mathsf{Regret}_{\pi,k}(\widehat{f}) = O_{s,k}\left(\frac{1}{n}\left(\log n\right)^{2k+1}\right).$$

□

Chapter 4

Upper Bound on Polynomial ERM Regret

4.1 ERM Algorithm

In this section, we will extend the ERM algorithm discovered in [4] to estimating θ^k. Note that the empirical Bayes estimator f^* satisfies

$$f^* = \operatorname*{argmin}_f \mathbb{E}\left[\left(f(X) - \theta^k\right)^2\right] = \operatorname*{argmin}_f \mathbb{E}\left[f(X)^2 - 2\theta^k f(X) + \theta^{2k}\right]. \tag{4.1}$$

If we can have this depend on just X and not θ, then we can obtain an unbiased estimate of this expression by using the empirical samples of X. First, θ^{2k} does not depend on the estimator, so removing it does not affect the minimization. Furthermore,

$$\begin{aligned}
\mathbb{E}\left[\theta^k f(X)\right] &= \int \sum_{x=0}^{\infty} e^{-\theta} \frac{\theta^x}{x!} f(x)\theta^k d\pi(\theta) \\
&= \int \sum_{x=0}^{\infty} e^{-\theta} \frac{\theta^{x+k}}{(x+k)!} P(x+k,k) f(x) d\pi(\theta) \\
&\overset{(a)}{=} \int \sum_{x=0}^{\infty} e^{-\theta} \frac{\theta^x}{x!} P(x,k) f(x-k) d\pi(\theta) \\
&= \mathbb{E}[P(X,k) f(X-k)].
\end{aligned} \tag{4.2}$$

where (a) follows from shifting the summation to start at k and adding in the terms from $0, \ldots, k-1$ since they are all 0. Substituting (4.2) into (4.1), we obtain

$$f^* = \operatorname*{argmin}_f \mathbb{E}\left[f(X)^2 - 2P(X,k) f(X-k)\right]. \tag{4.3}$$

As with the Bayes estimator for θ, we can show that f^* is a monotonic nondecreasing function. This is useful because it allows us to solve (4.3) over the class of monotonic functions and apply isotonic regression just as in [4].

Lemma 7. *$f^*(x)$ is an increasing function*

Proof. First, note that

$$\frac{\int e^{-\theta}\theta^{x+1}d\pi(\theta)}{\int e^{-\theta}\theta^{x}d\pi(\theta)}$$

is increasing over integer x since Cauchy-Schwartz tells us that

$$\frac{\int e^{-\theta}\theta^{x+2}d\pi(\theta)}{\int e^{-\theta}\theta^{x+1}d\pi(\theta)} \geq \frac{\int e^{-\theta}\theta^{x+1}d\pi(\theta)}{\int e^{-\theta}\theta^{x}d\pi(\theta)}.$$

Thus, we find

$$\begin{aligned}
&f^*(x+1) \geq f^*(x) \\
\iff &\frac{(x+k+1)p_\pi(x+k+1)}{p_\pi(x+1)} \geq \frac{(x+1)p_\pi(x+k)}{p_\pi(x)} \\
\iff &\frac{(x+k+1)\int e^{-\theta}\frac{\theta^{x+k+1}}{(x+k+1)!}d\pi(\theta)}{\int e^{-\theta}\frac{\theta^{x+1}}{(x+1)!}d\pi(\theta)} \geq \frac{(x+1)\int e^{-\theta}\frac{\theta^{x+k}}{(x+k)!}d\pi(\theta)}{\int e^{-\theta}\frac{\theta^{x}}{x!}d\pi(\theta)} \\
\iff &\frac{\int e^{-\theta}\theta^{x+k+1}d\pi(\theta)}{\int e^{-\theta}\theta^{x+1}d\pi(\theta)} \geq \frac{\int e^{-\theta}\theta^{x+k}d\pi(\theta)}{\int e^{-\theta}\theta^{x}d\pi(\theta)} \\
\iff &\frac{\int e^{-\theta}\theta^{x+k+1}d\pi(\theta)}{\int e^{-\theta}\theta^{x+k}d\pi(\theta)} \geq \frac{\int e^{-\theta}\theta^{x+1}d\pi(\theta)}{\int e^{-\theta}\theta^{x}d\pi(\theta)}.
\end{aligned}$$

□

Since f^* is increasing, using this function class in (4.3) and replacing with the empirical expectation, our ERM-based estimator is

$$\widehat{f}_{\mathsf{erm},k}(x) = \operatorname*{argmin}_{f\in\mathcal{F}_{\text{monotone}}} \widehat{\mathbb{E}}\left[f(X)^2 - 2P(X,k)f(X-k)\right]. \tag{4.4}$$

Although there is no unique solution to this minimization problem since there are $\widehat{f}$ is only uniquely determined for values that appear in our empirical expectation. We choose to take $\widehat{f}$ which is a step function that can only change at values where it is determined. An explicit solution to (4.4) may be calculated via [4, Lemma 1]. We now show that our empirical estimator is always bounded $X_{\max}^k$, which will help us bound the complexity of our function class.

Lemma 8. *Let $\widehat{f}_{\mathsf{erm},k}$ be the estimator defined in (4.4). Then* $\max \widehat{f}_{\mathsf{erm},k}(x) \leq X_{\max}^k$.

Proof. As $\widehat{f}_{\mathsf{erm},k}$ is monotonic, and $\widehat{f}_{\mathsf{erm},k}(x) = \widehat{f}_{\mathsf{erm}}(X_{\max})$ for all $x > X_{\max}$, it is sufficient to bound $\widehat{f}_{\mathsf{erm},k}(X_{\max})$. By [4, Lemma 1], there exists some i^* for which

$$\widehat{f}_{\mathsf{erm},k}(X_{\max}) = \frac{\sum_{i=i^*}^{X_{\max}} P(i+k,k)N(i+k)}{\sum_{i=i^*}^{X_{\max}} N(i)} = \frac{\sum_{i=i^*+k}^{X_{\max}} P(i,k)N(i)}{\sum_{i=i^*}^{X_{\max}} N(i)} \leq P(X_{\max},k) \leq X_{\max}^k$$

since $N(x) = 0$ for all $x > X_{\max}$. □

The main result of this section is the following theorem.

Theorem 4. *The ERM estimator for θ^k satisfies the following regret bounds:*

1.
$$\sup_{\pi \in \mathcal{P}([0,h])} \mathsf{Regret}_{\pi,k}(\widehat{f}_{\mathsf{erm},k}) \leq \frac{C \max\{1, h^{2k}\}}{n} \left(\frac{\log n}{\log \log n}\right)^{2k}, \quad k \geq 2$$

2.
$$\sup_{\pi \in SubE(s)} \mathsf{Regret}_{\pi,k}(\widehat{f}_{\mathsf{erm},k}) \leq \frac{C \max\{1, s^{2k+1}\} (\log n)^{2k+1}}{n}.$$

4.2 Rademacher Symmetrization

In order to prove these bounds, we use the following lemma which is a generalization of [4, Theorem 3] that allows us to bound the regret using Rademacher random variables.

Lemma 9. *Let $\mathcal{F}$ be a convex function class that contains the Bayes estimator f^*. Let $X_1, \ldots, X_n$ be a training sample drawn iid from p_π, $\epsilon_1, \ldots, \epsilon_n$ an independent sequence of iid Rademacher random variables, and $\widehat{f}$ the corresponding ERM solution. Then for any function class $\mathcal{F}_{p_n}$ depending on the empirical distribution $p_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ that includes $\widehat{f}$ and f_* we have*

$$\mathsf{Regret}_{\pi,k}(\widehat{f}) \leq \frac{3}{n} T_1(n) + \frac{2}{n} T_2(n)$$

where

$$T_1(n) = \mathbb{E}\left[\sup_{f \in \mathcal{F}_{p_n} \cup \mathcal{F}_{p'_n}} \sum_{i=1}^n (\epsilon_i - \frac{1}{6})(f^*(X_i) - f(X_i))^2\right] \tag{4.5}$$

and

$$\begin{aligned} T_2(n) = \mathbb{E}\Bigg[\sup_{f \in \mathcal{F}_{p_n} \cup \mathcal{F}_{p'_n}} \sum_{i=1}^n \Bigg\{ & 2\epsilon_i (f^*(X_i)(f^*(X_i) - f(X_i)) \\ & - P(X_i, k)(f^*(X_i - k) - f(X_i - k))) \\ & - \frac{1}{4}(f^*(X_i) - f(X_i))^2 \Bigg\}\Bigg] \end{aligned} \tag{4.6}$$

where $\mathcal{F}_{p'_n}$ is defined with an independent copy of $X_1, \ldots, X_n$.

Proof. Define

$$R(f) = \mathbb{E}[f(X)^2 - 2P(X,k) f(X-k)], \quad \widehat{R}(f) = \widehat{\mathbb{E}}[f(X)^2 - 2P(X,k) f(X-k)]. \tag{4.7}$$

Note that $\widehat{f}$ is defined as the function that minimizes $\widehat{R}(f)$. Since $\mathcal{F}$ is convex, $(1-\epsilon)\widehat{f} + \epsilon h \in \mathcal{F}$ for all $h \in \mathcal{F}$. Thus, $\widehat{R}(\widehat{f}) \leq \widehat{R}((1-\epsilon)\widehat{f} + \epsilon h)$. This means the gradient of the RHS with respect to ϵ is nonnegative when evaluated at 0. Plugging into (4.7), we obtain

$$\begin{aligned} & \frac{\partial}{\partial \epsilon} \widehat{R}((1-\epsilon)\widehat{f} + \epsilon h) \\ =& 2\widehat{\mathbb{E}}[(h(X) - \widehat{f}(X))((1-\epsilon)\widehat{f}(X) + \epsilon h(X)) - P(X,k)(h(X-k) - \widehat{f}(X-k))] \end{aligned}$$

and plugging in $\epsilon = 0$, we have

$$2\widehat{\mathbb{E}}[(h(X) - \widehat{f}(X))\widehat{f}(X) - P(X,k)(h(X-k) - \widehat{f}(X-k))] \geq 0.$$

This can be rearranged to obtain

$$\widehat{R}(h) - \widehat{R}(\widehat{f}) - \widehat{\mathbb{E}}[(h(X) - \widehat{f}(X))^2] \geq 0. \tag{4.8}$$

Now since $\mathsf{Regret}_{\pi,k}(\widehat{f}) = R(\widehat{f}) - R(f^*)$, using (4.7) and (4.8) gives

$$\begin{aligned}
&\mathsf{Regret}_{\pi,k}(\widehat{f}) \\
\leq &\mathbb{E}[R(\widehat{f}) - R(f^*) + \widehat{R}(f^*) - \widehat{R}(\widehat{f}) - \widehat{\mathbb{E}}[(f^*(X) - \widehat{f}(X))^2]] \\
= &\mathbb{E}[R(\widehat{f}) - R(f^*) - \mathbb{E}[(f^*(X) - \widehat{f}(X))^2] + (\widehat{R}(f^*) - \widehat{R}(\widehat{f}) + \widehat{\mathbb{E}}[(f^*(X) - \widehat{f}(X))^2]) \\
&\quad + \mathbb{E}[(f^*(X) - \widehat{f}(X))^2] - 2\widehat{\mathbb{E}}[(f^*(X) - \widehat{f}(X))^2]] \\
= &\mathbb{E}\Bigg[\widehat{\mathbb{E}}[2f^*(X)(f^*(X) - \widehat{f}(X)) - 2P(X,k)(f^*(X-k) - \widehat{f}(X-k))]) \\
&\quad - \mathbb{E}[2f^*(X)(f^*(X) - \widehat{f}(X)) - 2P(X,k)(f^*(X-k) - \widehat{f}(X-k))] \\
&\quad - \frac{1}{4}\left(\mathbb{E}[(f^*(X) - \widehat{f}(X))^2] + \widehat{\mathbb{E}}[(f^*(X) - \widehat{f}(X))^2]\right)\Bigg] \qquad (4.9) \\
&\quad + \left[\frac{5}{4}\mathbb{E}[(f^*(X) - \widehat{f}(X))^2] - \frac{7}{4}\widehat{\mathbb{E}}[(f^*(X) - \widehat{f}(X))^2]\right]. \qquad (4.10)
\end{aligned}$$

Using the symmetrization result in [4, Lemma 3], we can upper bound (4.9) with

$$\frac{2}{n}\mathbb{E}\left[\sup_{f \in \mathcal{F}_{p_n} \cup \mathcal{F}_{p'_n}} \sum_{i=1}^{n} \left\{2\epsilon_i(f^*(X_i)(f^*(X_i) - f(X_i)) - P(X_i,k)(f^*(X_i-k) - f(X_i-k))) - \frac{1}{4}(f^*(X_i) - f(X_i))^2\right\}\right] \tag{4.11}$$

by selecting

$$Tf(x) = -\left(2f^*(x)(f^*(x) - f(x)) - 2P(x,k)(f^*(x-k) - f(x-k))\right)$$

and $Uf(x) = \frac{1}{4}(f^*(x) - f(x))^2$, and (4.10) with

$$\frac{2}{n}\mathbb{E}\left[\sup_{f \in \mathcal{F}_{p_n} \cup \mathcal{F}_{p'_n}} \sum_{i=1}^{n} \frac{3}{2}\epsilon_i(f^*(X_i) - f(X_i))^2 - \frac{1}{4}(f^*(X_i) - f(X_i))^2\right] \tag{4.12}$$

by selecting $Tf(x) = \frac{3}{2}(f^*(x) - f(x))^2$ and $Uf(x) = \frac{1}{4}(f^*(x) - f(x))^2$. Combining the upper bounds in (4.11) and (4.12) proves the lemma. □

4.3 Bounding Rademacher Complexities

Define the function class depending on the samples

$$\mathcal{F}_* \triangleq \{f : f \text{ is monotone}, f(X_{\max}) \le \max\{X_{\max}, f^*(X_{\max})\}\}.$$

Note that by Lemma 8, it will contain both $\widehat{f}_{\text{erm}}$ and f^*. Define $\mathcal{F}'_*$ analogously for an independent set of samples. Then we will apply Lemma 9 with $\mathcal{F}_{p_n} = \mathcal{F}_*$ and $\mathcal{F}_{p'_n} = \mathcal{F}'_*$. Although this function class depends on f^* which we do not know, we can still use it for our theoretical analysis. For the rest of the proof, we consider a slightly generalized version of (4.5) and (4.6):

$$T_1(b,n) = \mathbb{E}\left[\sup_{f \in \mathcal{F}_* \cup \mathcal{F}'_*} \sum_{i=1}^{n} (\epsilon_i - \frac{1}{b})(f^*(X_i) - f(X_i))^2\right]$$

and

$$\begin{aligned} T_2(b,n) = \mathbb{E}\Bigg[\sup_{f \in \mathcal{F}_* \cup \mathcal{F}'_*} \sum_{i=1}^{n} \Bigg\{ & 2\epsilon_i (f^*(X_i)(f^*(X_i) - f(X_i)) \\ & - P(X_i,k)(f^*(X_i - k) - f(X_i - k))) \\ & - \frac{1}{b}(f^*(X_i) - f(X_i))^2 \Bigg\}\Bigg]. \end{aligned}$$

Next, given some tail bounds on the distribution p_π and moment bounds on the maximum sample, we prove bounds on the expressions T_1 and T_2.

Lemma 10. *Let $\pi \in \mathcal{P}[0,h]$ where h is a constant or $s \log n$ for some constant s. Let $M := M(n,h) > \max\{h,k\}$ be such that*

- $\sup_{\pi \in \mathcal{P}[0,h]} \mathbb{P}_{X \in p_\pi}[X > M] \le \frac{1}{n^7}$.
- *For $X_i \overset{iid}{\sim} p_\pi, \mathbb{E}\left[X_{\max}^\ell\right] \le c(\ell) M^\ell$ for $\ell \le 2k$ and constant c.*

Then

$$T_1(b,n) \le c_0(b)\left(\max\{1, h^{2k}\}M + M^{2k}\right)$$

and

$$T_2(b,n) \le c_0(b)\left(\max\{1, h^{2k}\}M + \max\{1, h^k\}M^{k+1}\right).$$

Note that these tail bounds are satisfied by bounded and subexponential priors, which are the ones that we consider later. The proof of this lemma is in Appendix C. Just as in Lemma 1, Lemma 10 provides a regret bound on a bounded prior. We again require Lemma 5 to extend our results to a subexponential prior.

4.4 Proof of Theorem 4

Now with these lemmas, we are able to combine them to obtain bounds on the regret of our algorithm over bounded and subexponential priors.

Proof. For the case of constant h, we may choose $M = \max\{c_1, c_2 h\} \frac{\log n}{\log\log n}$ due to [4, Lemma 10 and 12]. Then by Lemma 9 and Lemma 10, the regret is $O_k\left(\frac{\max\{1,h^{2k}\}}{n}\left(\frac{\log n}{\log\log n}\right)^{2k}\right)$ for $k \geq 2$.
For subexponential π, by Lemma 8 we have $\mathbb{E}[\widehat{f}(X)^4] \leq \mathbb{E}[X_{\max}^{4k}] = O_{s,k}((\log n)^{4k})$. By Lemma 5, it suffices to bound $\mathsf{Regret}_{\pi_{c_1 s \log n},k}$. By [4, Lemma 11 and 12], we may choose $M = \max\{c_1, c_2 s\} \log n$. Then by Lemma 9 and Lemma 10, the regret is $O_k\left(\frac{\max\{1,s^{2k+1}\}(\log n)^{2k+1}}{n}\right)$. □

Remark 1. *Note that in the bounded prior case, our ERM algorithm does not use the bounds of the prior. We conjecture that if only consider our estimator over the set of monotonic functions within the range of the prior, then we may obtain a tight bound with exponent $k+1$ instead of $2k$. This is because the $2k$ exponent only appears in the bound for (C.14) due to $f(X_{\max})^2 \leq X_{\max}^{2k}$. If we cap f, we can actually write $f(X_{\max})^2 \leq h^{2k}$, and we may be able to avoid the $X_{\max}^{2k}$ term entirely.*

Chapter 5

Lower Bound on Polynomial Regret

In this section, we prove that for any algorithm, there will exist some prior for which the regret matches the that of the NPMLE algorithm, implying that it is asymptotically optimal. Formally, we prove the following theorem.

Theorem 5. *Consider the Poisson mixture model. For any $h > 0$ and $s > 0$, the regret of the optimal estimator satisfies the following lower bounds:*

1. $$\inf_{\widehat{f}} \sup_{\pi \in \mathcal{P}([0,h])} \mathsf{Regret}_{\pi,k}(\widehat{f}) = \Omega_{h,k}\left(\frac{1}{n}\left(\frac{\log n}{\log\log n}\right)^{k+1}\right)$$

2. $$\inf_{\widehat{f}} \sup_{\pi \in SubE(s)} \mathsf{Regret}_{\pi,k}(\widehat{f}) = \Omega_{s,k}\left(\frac{1}{n}(\log n)^{2k+1}\right)$$

5.1 Setup for a General Lower Bound

We first set up a generalization to [15, Proposition 7] on establishing functional (namely F) of θ (in our case, we are interested in the case $F(\theta) \triangleq \theta^k$. Assume (for sake of simplicity) that F is continuously differentiable everywhere on $\theta \geq 0$.

To start with, fix a distribution G_0. We follow the recipe of [15, (Equation 21)] and define the operation K bringing function r to Kr given by

$$Kr(x) \triangleq \mathbb{E}_{G_0}[r(\theta) \mid X = x] = \frac{\int r(\theta) f_\theta(x) G_0(d\theta)}{f_0(x)} \tag{5.1}$$

where $f_0 = \int f_\theta G_0(d\theta)$. Also fix an arbitrary bounded function r, consider the distribution G_δ given by the small perturbation

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

We now consider what happens as we consider $\mathbb{E}_{G_\delta}[F(\theta)|X=x]$. To start, by (5.1), we have $KF(x) = \mathbb{E}_{G_0}[F(\theta)|X=x]$ (i.e. the base distribution). Then similar to [15, (Equation 24)] we may obtain

$$\mathbb{E}_{G_\delta}[F(\theta)|X=x] = KF(\theta)(x) + \delta K_F r(x) + \delta^2 \frac{1}{1+\delta Kr(x)}(Kr)(y)\cdot(K_F r)(x) \tag{5.2}$$

where the operation K_F is defined as (modified from [15, (Equation 25)]).

$$K_F r \triangleq K(Fr) - (KF)(Kr)$$

Note also the following identity, again can be generalized from [15, (Equation 25)]

$$K_F r(x) = \frac{d}{d\delta}\,|_{\delta=0}\,\mathbb{E}_{G_\delta}[F(\theta)\mid X=x]$$

With this, we are ready to establish the following 'general recipe' of functional estimation $F(\theta)$.

Lemma 11. *Fix a prior distribution G_0, constants $\alpha, \tau, \tau_1, \tau_2, \gamma \geq 0$ and m real-valued functions $r_1, \ldots, r_m$ on Θ with the following properties,*

$$\begin{aligned}
\|r_q\|_\infty &\leq a \quad \forall q \\
\|Kr_q\|_{L_2(f_0)} &\leq \sqrt{\gamma} \quad \forall q \\
\left\|\sum_{i=1}^m v_i K_F r_i\right\|^2_{L_2(f_0)} &\geq \tau\|v\|_2^2 - \tau_2 \quad \forall v \in \{0,\pm 1\}^m \\
\left\|\sum_{i=1}^m v_i K_F r_i\right\|^2_{L_2(f_0)} &\leq \tau_1^2 m \quad \forall v \in \{0,\pm 1\}^m.
\end{aligned}$$

Then the optimal regret in $F(\theta)$ estimation over the class of priors $\mathcal{G} = \{G : |\frac{dG}{dG_0} - 1| \leq \frac{1}{2}\}$ satisfies

$$\inf_{\widehat{f}}\sup_{\pi\in\mathcal{G}} \mathsf{Regret}_{\pi,k}(\widehat{f}) \geq C\delta^2(m(4\tau - \tau_1^2) - \tau_2), \quad \delta \triangleq \frac{1}{\max(\sqrt{n\gamma}, ma)}$$

for some constant $C > 0$.

Proof. The proof follows exactly the proof of [15, Lemma 7] but instead using $T_u(x) \triangleq \mathbb{E}_{G_u}[F(\theta)|X=x]$ and K_F in place of K_1. In particular, the lattices for Assouad's lemma is also defined exactly as in [15, (Equation 29)]: define $\mu_i \triangleq \int r_i dG_0$, $\delta > 0$ chosen with $\delta \leq \frac{1}{16ma}$, and for each $u \in \{0,1\}^m$,

$$r_u \triangleq \sum_{i=1}^n u_i r_i \qquad h_u \triangleq Kr_u \qquad \mu_u \triangleq \sum_{i=1}^m u_i\mu_i \qquad dG_u \triangleq \frac{1+\delta r_u}{1+\delta\mu_u}dG_0 \qquad f_u \triangleq \frac{(1+\delta h_u)f_0}{1+\delta\mu_u}$$

where f_u is the mixture density induced by the prior G_u.

Let $\widetilde{\mathcal{G}} = \{G_u : u \in \{0,1\}^m\}$. The construction guarantees $\widetilde{\mathcal{G}} \subseteq \mathcal{G}$, and $\frac{1}{2} \leq \frac{dG_u}{dG_0} \leq \frac{3}{2}$. Then exactly like [15]:

$$\begin{aligned}
\inf_{\widehat{T}} \sup_{\pi \in \mathcal{G}} \mathsf{Regret}_{\pi,k}(\widehat{T}) &\geq \inf_{\widehat{T}} \sup_{\pi \in \widetilde{\mathcal{G}}} \mathsf{Regret}_{\pi,k}(\widehat{T}) \\
&= \inf_{\widehat{T}} \sup_{u \in \{0,1\}^m} \mathbb{E}_{G_u} ||\widehat{T} - T_u||^2_{L^2(f_u)} \\
&\overset{(a)}{\geq} \inf_{\widehat{T}} \sup_{u \in \{0,1\}^m} \frac{1}{2} \mathbb{E}_{G_u} ||\widehat{T} - T_u||^2_{L^2(f_0)} \\
&\overset{(b)}{\geq} \inf_{\widehat{u} \in \{0,1\}^m} \sup_{u \in \{0,1\}^m} \frac{1}{8} \mathbb{E}_{G_u} ||T_{\widehat{u}} - T_u||^2_{L^2(f_0)}
\end{aligned}$$

where (a) uses $f_u \geq \frac{1}{2} f_0$ and (b) is due to the following triangle inequality argument: if u_1, u_2 are such that

$$\inf_{\widehat{u} \in \{0,1\}^m} \sup_{u \in \{0,1\}^m} \mathbb{E}_{G_u} ||T_{\widehat{u}} - T_u||^2_{L^2(f_0)} = \mathbb{E}_{G_{u_2}} ||T_{u_1} - T_{u_2}||^2_{L^2(f_0)}$$

then for an arbitrary estimator $\widehat{T}$:

$$\sup_{u \in \{0,1\}^m} \mathbb{E}_{G_u} ||\widehat{T} - T_u||^2_{L^2(f_0)} \geq \frac{1}{4} \max\{\mathbb{E}_{G_{u_1}} ||T_{\widehat{u}} - T_{u_1}||^2_{L^2(f_0)}, \mathbb{E}_{G_{u_2}} ||T_{\widehat{u}} - T_{u_2}||^2_{L^2(f_0)}\}$$

Now we consider the following property of T_u, due to (5.2) and that $Kr = h_u$

$$T_u = KF + \delta K_F r + \delta^2 \frac{h_u}{1 + \delta h_u} \cdot K_F r$$

and by our assumptions:

$$||\delta^2 \frac{h_u}{1 + \delta h_u} \cdot K_F r||_2 \leq 2\delta^2 ma ||K_F r_u||_2 \leq 2\delta^2 m^{3/2} a \tau_1 \leq \frac{1}{8} \delta a \sqrt{m} \tau_1$$

Again by triangle inequality:

$$||T_u - T_v||_2 \geq \delta ||K_F(r_u - r_v)||_2 - \frac{\delta \sqrt{m} \tau_1}{4}$$

Thus by using $(a-b)^2 \geq \frac{1}{2} a^2 - b^2$ this translates into

$$||T_u - T_v||_2^2 \geq \frac{1}{2} \delta^2 ||K_F(r_u - r_v)||_2^2 - \frac{\delta^2 m \tau_1^2}{16} \geq \frac{1}{2} \delta^2 (\tau d_H(u,v) - \tau_2) - \frac{1}{16} \delta^2 m \tau_1^2$$

Finally, we quote directly from the proof of [15, Lemma 7] to get that for some constant C_1,

$$\chi^2(f_u || f_v) \leq C_1 \delta^2 ||h_u - h_v||_2^2 \gamma \leq C_1 \delta$$

where we used the assumption $||h_u - h_v||^2 \leq \gamma$. Note that for $d_H(u,v) = 1, ||h_u - h_v||^2 \leq \gamma$ by our assumption. Finally, for some constant C_2, if $n\delta^2\gamma \leq 1$ we have:

$$\chi^2(f_u^{\otimes n} || f_v^{\otimes n}) \leq C_2, \forall u, v : d_H(u,v) = 1$$

Thus by Assouad's lemma we have

$$\inf_{\widehat{u}\in\{0,1\}^m} \sup_{u\in\{0,1\}^m} \mathbb{E}_{G_u}[d_H(\widehat{u},u)] \geq C_3 m$$

for some constant C_3.

To summarize, by choosing δ with $\delta^2 = \frac{1}{\max(n\gamma, m^2a^2)}$ this gives us

$$\inf_{\widehat{u}\in\{0,1\}^m} \sup_{u\in\{0,1\}^m} \mathbb{E}_{G_u}[||T_u - T_v||^2] \geq \frac{1}{2}\delta^2(m(C_3\tau - \frac{1}{8}\tau_1^2) - \tau_2)$$

□

5.2 Results on the Poisson Model

Next, we consider how to incorporate this into the Poisson model. In general the approach is still the same as that of [15]: we consider G_0 as the Gamma prior Gamma(α,β). Define, now, $K_k(r) = K_F(r)$ when $F(x) = x^k$. Here we consider the following generalization of [15, Proposition 10]:

Lemma 12. *Consider the Gamma prior with parameter (α,β) PMF $G_0(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$. Then for a function r such that $r^{(j)}$ is bounded for $j < k$, $K_k(r)$ satisfies the following property:*

$$K_k r = \sum_{j=1}^{k} \frac{(-1)^{j+1}}{(1+\beta)^j}\binom{k}{j} K(x^k r^{(j)})$$

This lemma allows us to relate the operator K_k to K. The proof is given in Appendix D. Using this result, we can construct a suitable set of functions that gives the desired result when used in Lemma 11. The analysis will be continued in Appendix D, but we state the results below, which are generalizations [15, Lemma 11, Lemma 12].

Lemma 13. *Fix $\delta > 0$. Let $G_0 = Gamma(\alpha,\beta)$. Then there exist absolute positive constants C, m_0 such that for all $m \geq m_0, \beta \geq 2, \alpha \geq (2k+2)m$, there exist functions $r_1,\ldots,r_m$ such that*

$$\|K_k r_j\|^2_{L_2(f_0)} = 1 \qquad \forall j = 1,\cdots,m, \tag{5.3}$$

$$(Kr_j, Kr_i)_{L_2(f_0)} = (K_k r_j, K_k r_i)_{L_2(f_0)} = 0 \qquad \forall i \neq j, \tag{5.4}$$

$$\|Kr_j\|^2_{L_2(f_0)} \leq \frac{C\beta^k}{\alpha^k m^k} \qquad \forall j = 1,\cdots,m, \tag{5.5}$$

$$\|r_j\|_\infty \leq \sqrt{\frac{\beta^k}{\alpha^k}} e^{C(m\log\beta+\alpha)} \qquad \forall j = 1,\cdots,m. \tag{5.6}$$

Lemma 14. *Let $G_0 = Gamma(\alpha,\beta)$ where $\alpha = 1$ and $\beta > 0$ is fixed, there exists some constant $C(\beta) > 0$ such that for all $m \geq 1$ there exist functions $r_1,\ldots,r_m$ such that (5.3), (5.4), and for all $j = 1,\cdots,m$,*

$$\|Kr_j\|^2_{L_2(f_0)} \leq \frac{C}{m^{2k}}, \tag{5.7}$$

$$\|r_j\|_\infty \leq m^{1-k}e^{Cm}. \tag{5.8}$$

5.3 Proof of Theorem 5

Now using the functions constructed in Lemma 13 and Lemma 14 in Lemma 11, we can finally prove a lower bound.

Proof. For the set of bounded priors $\mathcal{P}([0,h])$, we apply Lemma 11 by using the functions generated by Lemma 13 with

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

where $c_1, c_2 > 0$ to be specified later based on h. Note that (5.3) and (5.4) ensure that $\tau = \tau_1 = 1$ and $\tau_2 = 0$. Furthermore, (5.5) gives $\gamma = \frac{Cc_2^k}{m^k}$ for some absolute constant C as defied in (5.5). (5.6) gives

$$a = c_2^{k/2} e^{C(\alpha + m\log\beta)} = c_2^{k/2} e^{(2Cc_1 + o(1))\log n} = O_{h,k}(n^{2Cc_1 + o(1)}).$$

If we pick $c_1 = \frac{1}{8C}$, then $ma = O_{h,k}(n^{1/4+o(1)})$ while $\sqrt{n\gamma} = O_{h,k}(n^{1/2-o(1)})$ so $\delta = \frac{1}{\sqrt{n\gamma}}$. Applying Lemma 11, we have

$$\inf_{\widehat{f}} \sup_{\pi \in \mathcal{G}} \mathsf{Regret}_{\pi,k}(\widehat{f}) \geq \frac{3C}{n\gamma} m = \frac{3c_1^{k+1}}{c_2^k n} \left(\frac{\log n}{\log\log n}\right)^{k+1} \tag{5.9}$$

where $\mathcal{G} = \left\{G : \left|\frac{dG}{dG_0} - 1\right| \leq \frac{1}{2}\right\}$. Now we relate the regret over $\mathcal{G}$ to the regret over $\mathcal{P}([0,h])$ using the following lemma.

Lemma 15. *Given $h > 0$, let $\mathcal{G}$ be a collection of priors on $\mathbb{R}_{\geq 0}$ such that $\sup_{\pi\in\mathcal{G}} \mathbb{P}(\theta > h) \leq \epsilon \leq \frac{1}{2}$ for some ϵ and $\sup_{\pi\in\mathcal{G}} \mathbb{E}_\pi[\theta^{4k}] \leq M$. Then*

$$\inf_{\widehat{f}} \sup_{\pi \in \mathcal{P}([0,h])} \mathsf{Regret}_{\pi,k}(\widehat{f}) \geq \inf_{\widehat{f}} \sup_{\pi \in \mathcal{G}} \mathsf{Regret}_{\pi,k}(\widehat{f}) - 6\sqrt{(M + h^{4k})n\epsilon}. \tag{5.10}$$

Proof. Let E be the event that $\theta_i \leq h$ for all θ. For any estimator $\widehat{f}$ taking values in $[0, h^k]$ and any prior $\pi \in \mathcal{G}'$,

$$\begin{aligned}
\mathbb{E}_\pi[(\widehat{f}(X) - \theta^k)^2] &= \mathbb{E}_\pi[(\widehat{f}(X) - \theta)^2 \mathbf{1}_E] + \mathbb{E}_\pi[(\widehat{f}(X) - \theta)^2 \mathbf{1}_{E^C}] \\
&\leq \mathbb{E}_\pi[(\widehat{f}(X) - \theta^k)^2 | E] + \sqrt{\mathbb{E}_\pi[(\widehat{f}(X) - \theta)^4] \mathbb{P}_\pi(E^C)} \\
&\leq \mathbb{E}_\pi[(\widehat{f}(X) - \theta^k)^2 | E] + \sqrt{8(M + h^{4k})n\epsilon}
\end{aligned} \tag{5.11}$$

Since all distributions in $\mathcal{P}([0,h])$ have support in $[0,h]$, the optimal estimator will also be

in the support. Then we get the following inequalities:

$$\begin{aligned}
\inf_{\widehat{f}} \sup_{\pi\in\mathcal{P}([0,h])} \mathsf{Regret}_{\pi,k}(\widehat{f}) &= \inf_{\widehat{f}\in[0,h]} \sup_{\pi\in\mathcal{P}([0,h])} \mathbb{E}_\pi[(\widehat{f}(X)-\theta^k)^2] - \mathsf{mmse}_k(\pi) \\
&\geq \inf_{\widehat{f}\in[0,h]} \sup_{\pi\in\mathcal{G}} \mathbb{E}_{\pi_h}[(\widehat{f}(X)-\theta^k)^2] - \mathsf{mmse}_k(\pi_h) \\
&\geq \inf_{\widehat{f}\in[0,h]} \sup_{\pi\in\mathcal{G}} \mathbb{E}_{\pi}[(\widehat{f}(X)-\theta^k)^2|E] - \mathsf{mmse}_k(\pi_h) \\
&\overset{(a)}{\geq} \inf_{\widehat{f}\in[0,h]} \sup_{\pi\in\mathcal{G}} \mathbb{E}_{\pi}[(\widehat{f}(X)-\theta^k)^2] - \sqrt{8(M+h^{4k})n\epsilon} - \frac{1}{1-\epsilon}\mathsf{mmse}_k(\pi) \\
&\geq \inf_{\widehat{f}} \sup_{\pi\in\mathcal{G}} \mathsf{Regret}_{\pi,k}(\widehat{f}) - \sqrt{8(M+h^{4k})n\epsilon} - \frac{\epsilon}{1-\epsilon}\mathsf{mmse}_k(\pi) \\
&\overset{(b)}{\geq} \inf_{\widehat{f}} \sup_{\pi\in\mathcal{G}} \mathsf{Regret}_{\pi,k}(\widehat{f}) - \sqrt{8(M+h^{4k})n\epsilon} - 2\epsilon\sqrt{M}
\end{aligned}$$

where (a) follows from (5.11) and Lemma 3 and (b) follows from $\epsilon \leq \frac{1}{2}$ and $\mathsf{mmse}_k(\pi) \leq \mathbb{E}_\pi[\theta^{2k}] \leq \sqrt{M}$. We can combine the last two terms to obtain (5.10). □

By the proof of [15, Theorem 2], we can choose c_2 such that $\mathbb{P}(G \geq h) \leq 2n^{-4}$ for $G \in \mathcal{G}$. Furthermore, it is well known the moments of the Gamma distribution are

$$\mathbb{E}_{G_0}[\theta^{4k}] = \frac{\Gamma(\alpha+4k)}{\beta^{4k}\Gamma(\alpha)} \asymp c_2^{-4k} = O_{h,k}(1)$$

so $\sup_{\pi\in\mathcal{G}} \mathbb{E}_\pi[\theta^{4k}] = O_{h,k}(1)$. Now using (5.9) and Lemma 15 with $\epsilon = n^{-4}$ and constant M, we have

$$\inf_{\widehat{f}} \sup_{\pi\in\mathcal{P}([0,h])} \mathsf{Regret}_{\pi,k}(\widehat{f}) \geq \frac{3c_1^{k+1}}{c_2^k n}\left(\frac{\log n}{\log\log n}\right)^{k+1} - O_{h,k}(n^{-3/2}) = \Omega_{h,k}\left(\frac{1}{n}\left(\frac{\log n}{\log\log n}\right)^{k+1}\right).$$

Now we move onto the subexponential case. If we choose $\alpha = 1$ and $\beta = s$, $G_0 = \mathrm{Expo}(s)$ so $\mathbb{P}_{G_0}(\theta \geq t) \leq e^{-t/s}$. Thus, for all $G \in \mathcal{G} = \left\{G : \left|\frac{dG}{dG_0} - 1\right| \leq \frac{1}{2}\right\}$, $\mathbb{P}_G(\theta \geq t) \leq 2e^{-t/s}$ so $\mathcal{G} \subseteq \mathsf{SubE}(s)$. Now we apply Lemma 11 by using the functions generated by Lemma 14 with $m = c\log n$. Again, (5.3) and (5.4) ensure that $\tau = \tau_1 = 1$ and $\tau_2 = 0$. Furthermore, (5.7) gives $\gamma = \frac{C}{m^{2k}}$ and (5.8) gives

$$a = m^{1-k}e^{C(\alpha+m\log\beta)} = (c\log n)^{1-k}e^{(Cc\log s\log n)+C} = O_{s,k}(n^{Cc\log s+o(1)}).$$

If we pick $c = \frac{1}{4C\log s}$, then $ma = O_{s,k}(n^{1/4+o(1)})$ while $\sqrt{n\gamma} = O_{s,k}(n^{1/2-o(1)})$ so $\delta = \frac{1}{\sqrt{n\gamma}}$. Applying Lemma 11, and using $\mathcal{G} \subseteq \mathsf{SubE}(s)$, we have

$$\inf_{\widehat{f}} \sup_{\pi\in\mathrm{SubE}(s)} \mathsf{Regret}_{\pi,k}(\widehat{f}) \geq \inf_{\widehat{f}} \sup_{\pi\in\mathcal{G}} \mathsf{Regret}_{\pi,k}(\widehat{f}) \geq \frac{3C}{n\gamma}m = \Omega_{s,k}\left(\frac{1}{n}(\log n)^{2k+1}\right).$$

□

Chapter 6

Simulations

In addition to the theoretical regret bounds we have proven in the previous sections, we also simulated them on some sampled data sets. In particular, we calculate the regret over the following set of priors:

- Exponential Distribution with mean 0.4 and 0.7
- Uniform distribution over $[0, 2]$ and $[0, 3]$.

The true Bayes estimator for each is calculated as follows:

- For an exponential distribution with mean λ, the Bayes estimator is

$$f^*(x) = \frac{P(x+k, k)\lambda^k}{(1+\lambda)^k}.$$

- For a uniform distribution with range $[0, M]$, the Bayes estimator is

$$f^*(x) = \frac{(x+k)! - \Gamma(x+k+1, M)}{x! - \Gamma(x+1, M)}$$

  where Γ is the incomplete Gamma function.

Since these values can be calculated explicitly, we are able to calculate the regret. We then examine how each regret evolves with both n and k. To see the relationship with n, we fix $k = 2$ and simulate $n = 100, 200, \ldots, 1000$. To see the relationship with k, we fix $n = 100$ and simulate $k = 2, 3, 4, 5$. To obtain less noisy regret estimates, we run 10000 trials for each setup and take the mean regret over all trials. The graph of regrets is shown in Fig. 6.1.

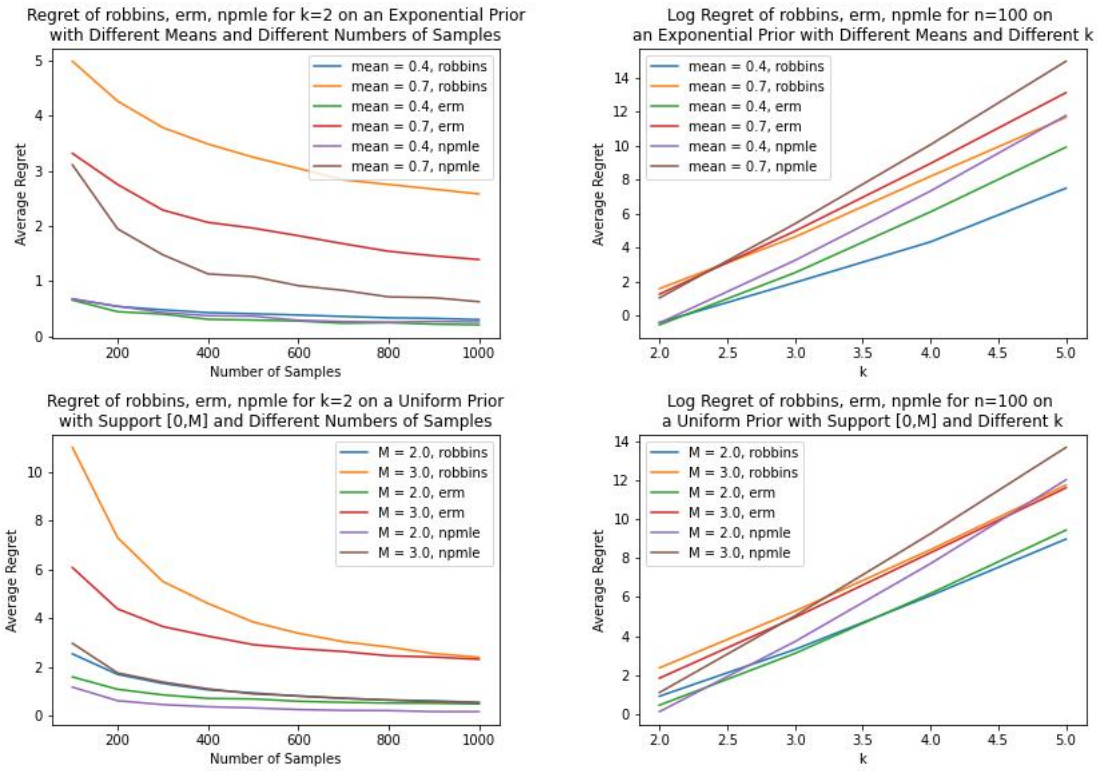

Figure 6.1: Regret of the three algorithms with respect to k and n across the four different prior distributions.

In the graphs with respect to n, we see that the decrease in the regret curve resembles a $\frac{\text{polylog}(n)}{n}$ decay. Furthermore, it appears that NPMLE performs the best in this scenario, then ERM, and then Robbins. This is slightly surprising since the upper bound on the ERM regret is asymptotically worse than the other two algorithms. However, it is possible that a better bound can be achieved. We have also only ran this simulation with a relatively small distribution, and it is possible that the constant factors overpower the power of the log term.

In the graphs with respect to k, the log regret curve looks linear, which agrees with our minimax regrets. It appears that in this case, the order of the algorithms' performances are reversed, with Robbins performing the best and NPMLE performing the worst. Again this is possible due to constant factors of the form c^k, but it is surprising that the ordering completely reverses for larger k.

Chapter 7

Conclusion and Future Directions

In this thesis, we have extended the classic problem of estimating θ on a Poisson mixture model to estimating θ^k. We have extended the f-modeling, g-modeling, and ERM based algorithms to estimate θ^k and proved regret bounds for each of these algorithms. In addition, we proved a lower bound for this estimation problem to show that the algorithms for f and g-modeling are both tight given the bounded and subexponential priors, and that ERM is tight given a subexponential prior. Lastly, we empirically evaluated the regrets of each of these algorithms by simulating them on various prior distributions and examining the relationship with n and k. In the remainder of this section, we will discuss different directions in which this research problem can be further studied.

7.1 Clipping the ERM Estimator

As mentioned in Remark 1, the ERM algorithm can be modified to take into account the bounds on the prior, and search only for monotone functions within these bounds, that is, we instead solve for

$$\widehat{f}_{\mathsf{erm},k,\mathsf{clipped}}(x) = \operatorname*{argmin}_{f\in\mathcal{F}_{\text{monotone}},f\leq h} \widehat{\mathbb{E}}\left[f(X)^2 - 2P(X,k)f(X-k)\right].$$

This can be useful since in practice, such a bound would usually be known. The first question is whether or not the solution satisfies

$$\widehat{f}_{\mathsf{erm},k,\mathsf{clipped}}(x) = \min(\widehat{f}_{\mathsf{erm},k}(x), h),$$

which may follow a similar proof to [4, Lemma 1]. If satisfied, we can easily modify the existing algorithm to solve the isotonic regression and then clip the function at h. In this case, the speed of the algorithm is not hindered. Otherwise, we will have to investigate other algorithms to solve this problem. We may also calculate new regret bounds by modifying the proof of Lemma 10 which may give rise to a tight bound on ERM in the bounded setting as well.

7.2 Smooth Functions

We have examined only monomials, but polynomials can easily be estimated as well by doing each term separately. Given that all coefficients are bounded, note that the regret from the leading term asymptotically overpowers all other regrets. This means the regret of a polynomial is asymptotically the same as the regret of the leading term. This gives rise to an interesting extension.

A useful property of polynomials is that they can be used to approximate any smooth function on a closed interval arbitrarily closely by the Stone-Weierstrass Theorem. Specifically, for any smooth function f on a range $[0, h]$ and error ϵ, there is some degree n polynomial for which $\|P_n(x) - f(x)\|_\infty \leq \epsilon$. Thus, we may be able to further generalize our analysis to apply to any smooth functions on an interval.

A natural approach is to approximate the function using a polynomial and then estimate each term individually. One issue to consider carefully would be that the coefficients may grow fast as we become more and more precise, and the scale of the coefficients will greatly affect the overall regret.

7.3 Heavy-tailed Priors

Another interesting extension is to examine heavier tailed distributions, such as ones with bounded p-th moment ($\mathbb{E}[|X|^p] < \infty$). In this thesis, we have only examined bounded and subexponential distributions, where encountering large values of θ is relatively unlikely. This makes the process of estimating functions of θ relatively more easy.

As discussed in Section 1.3.2, g-modeling has proven to be optimal in this setting whereas f-modeling has been proven not be [12]. The performance of the ERM algorithm on such priors is not yet known. It would be interesting to explore the following questions:

- What is the performance of the orignal ERM algorithm?
- What is the minimax lower bound?
- Can a generalized version of NPMLE or ERM can achieve optimality?

7.4 Other Mixture Models

In this thesis, we have focused on mixture models with a Poisson channel. However, real world examples are not limited to these models. For example, other common distributions may be Gaussian, exponential, or negative binomial.

There have been extensive research on the normal location model. Both the f[16] and g-modeling[17] approaches have been shown to obtain a nearly optimal fast rate of regret. However, the analogous ERM estimator has only been proven to achieve a slow rate for regret in [18], and it is still unknown whether a faster rate can be achieved. It may be possible to extend the framework for these upper bounds to θ^k. Furthermore, the method [15] used to prove a minimax lower bound for Gaussian mixture models is very similar to the Poisson

mixture model. It is likely that the way in which we extended the lower bound in Theorem 5 can also be applied to the Gaussian.

In addition, the ERM objective has also been extended to other distributions such as geometric, negative binomial, and exponential as done in [4], but regret bounds for these are still unknown. As these distributions are commonly used in the real world as well, proving bounds in these settings will also be useful.

Appendix A

Auxiliary Proofs for Modified Robbins

Proof of Lemma 2. For the first case, define $\bar{f}(x) = \frac{h^x e^{-h}}{x!} \geq p_\pi(x)$ just as in [15, Lemma 17]. Then

$$\begin{aligned}\sum_{x>x_0+k} \bar{f}(x)^2 P(x+k,k) &\leq \sum_{x>x_0+k} 2^k \bar{f}(x)^2 P(x,k) \\ &\overset{\text{(a)}}{\leq} (2h)^k \sum_{x>x_0+k} \bar{f}(x-k)\bar{f}(x) \\ &\overset{\text{(b)}}{\leq} (2h)^k \bar{f}(x_0) \sum_{x>x_0+k} \bar{f}(x) \\ &\overset{\text{(c)}}{\leq} 2(2h)^k \bar{f}(x_0)^2\end{aligned}$$

where (a) is by the identity $x\bar{f}(x) = h\bar{f}(x-1)$, (b) is by the monotonicity of $\bar{f}$ on the domain $[2h, \infty)$, and (c) is by [15, (Equation 133)]. Our choice of x_0 satisfies $\bar{f}(x_0) \leq \frac{1}{n}$, so we obtain (2.6).

By the proof of Lemma 17 in [15], we know $\widetilde{f}(x) \triangleq 2(1+\frac{1}{s})^{-x} \geq f(x)$, and then the properties of the geometric distribution give (2.7). □

Proof of Lemma 4. For the first bound, note that

$$\mathbb{E}_\pi[\theta^{4k}] = 4k \int_0^\infty x^{4k-1} \mathbb{P}_\pi(\theta > x) dx \leq 8k \int_0^\infty x^{4k-1} e^{-x/s} dx = 8k(4k-1)! s^{4k}.$$

For the second bound, using [11, (Equation 44)], we know $\mathbb{P}_{p_\pi}(X \geq K) \leq \frac{3}{2} e^{-K \log(1+\frac{1}{2s})}$.

Thus for any L,

$$\begin{aligned}
\mathbb{E}[X_{\max}^{\ell}] &= \ell \int_0^{\infty} x^{\ell-1}\mathbb{P}(X_{\max} > x)dx \\
&\leq 4L^{\ell} + n\int_L^{\infty} x^{\ell-1}\mathbb{P}(X > x)dx \\
&\leq 4L^{\ell} + \frac{3n}{2}\int_L^{\infty} x^{\ell-1}e^{-x\log(1+\frac{1}{2s})}dx \\
&\overset{(a)}{\leq} 4L^{\ell} + \frac{3n}{2(\log(1+\frac{1}{2s}))^{\ell}}\int_{L\log(1+\frac{1}{2s})}^{\infty} z^{\ell-1}e^{-z}dz \\
&\overset{(b)}{\leq} 4L^{\ell} + \frac{3n}{2(\log(1+\frac{1}{2s}))^{\ell}}\int_{L\log(1+\frac{1}{2s})}^{\infty} c(k)e^{-z/2}dz \\
&\leq 4L^{\ell} + \frac{3n}{(\log(1+\frac{1}{2s}))^{\ell}}c(k)e^{-L\log(1+\frac{1}{2s})/2}
\end{aligned}$$

where (a) follows from a change of variables with $z = x\log(1+\frac{1}{2s})$ and (b) follows since there exists some $c(k)$ such that $c(k)e^{z/2} \geq z^{\ell-1}$ for all positive z and $\ell \leq 4k$. Plugging in $L = \frac{2\log n}{\log(1+\frac{1}{2s})}$ gives the desired bound in (2.8). □

Appendix B

Auxiliary Proofs for NPMLE

Proof of Lemma 6. First note that $\mathsf{mmse}_k(\pi) \leq \mathbb{E}[\theta^{2k}] \leq \sqrt{\mathbb{E}[\theta^{4k}]} \leq \sqrt{M}$.
Let $\theta \sim \pi$ and $X|\theta \sim f_\theta$. Let f^* and f_h^* be the Bayes estimators for the priors π and π_h respectively. For $\widehat{\pi}$ independent of X,

$$\begin{aligned}
\mathbb{E}_\pi[(\widehat{f}(X)-\theta^k)^2] &= \mathbb{E}_\pi[(\widehat{f}(X)-\theta^k)^2\mathbf{1}_{\theta\leq h}] + \mathbb{E}_\pi[(\widehat{f}(X)-\theta^k)^2\mathbf{1}_{\theta>h}] \\
&\leq \mathbb{E}_\pi[(\widehat{f}(X)-\theta^k)^2|\theta\leq h] + \sqrt{\mathbb{E}_\pi[(\widehat{f}(X)-\theta^k)^4]\mathbb{P}_\pi(\theta>h)} \\
&\leq \mathbb{E}_{\pi_h}[(\widehat{f}(X)-\theta^k)^2] + \sqrt{8(\widehat{h}^{4k}+\mathbb{E}_\pi[\theta^{4k}])\mathbb{P}_\pi(\theta>h)} \\
&\leq \mathbb{E}_{\pi_h}[(\widehat{f}(X)-\theta^k)^2] + \sqrt{8(\widehat{h}^{4k}+M)\mathbb{P}_\pi(\theta>h)}
\end{aligned}$$

where the second line follows from Cauchy Schwarz and the third line follows from $(a+b)^4 \leq 8(a^4+b^4)$ for any real a, b. Then

$$\begin{aligned}
&\mathsf{Regret}_{\pi,k}(\widehat{f}) \\
=&\mathbb{E}_\pi[(\widehat{f}(X)-\theta^k)^2] - \mathsf{mmse}_k(\pi) \\
\leq&\mathbb{E}_{\pi_h}[(\widehat{f}(X)-\theta^k)^2] - \mathsf{mmse}_k(\pi_h) + \mathsf{mmse}_k(\pi_h) - \mathsf{mmse}_k(\pi) + \sqrt{8(\widehat{h}^{4k}+M)\mathbb{P}_\pi(\theta>h)} \\
\leq&\mathbb{E}_{\pi_h}[(\widehat{f}(X)-f_h^*(X))^2] + \left(\frac{1}{\mathbb{P}_\pi(\theta\leq h)}-1\right)\mathsf{mmse}_k(\pi) + \sqrt{8(\widehat{h}^{4k}+M)\mathbb{P}_\pi(\theta>h)} \\
\overset{(a)}{\leq}&\mathbb{E}_{\pi_h}[(\widehat{f}(X)-f_h^*(X))^2] + \frac{\mathbb{P}_\pi(\theta>h)}{\mathbb{P}_\pi(\theta\leq h)}\sqrt{M} + \sqrt{8(\widehat{h}^{4k}+M)\mathbb{P}_\pi(\theta>h)} \\
\leq&\mathbb{E}_{\pi_h}[(\widehat{f}(X)-f_h^*(X))^2] + \frac{(1+2\sqrt{2})\sqrt{(\widehat{h}^{4k}+M)\mathbb{P}_\pi(\theta>h)}}{\mathbb{P}_\pi(\theta\leq h)} \qquad \text{(B.1)}
\end{aligned}$$

where (a) follows from $\mathsf{mmse}_k(\pi) \le \sqrt{M}$. The second term is already in the right form, so we bound the first term. For any $K \ge 1$,

$$
\begin{aligned}
&\mathbb{E}_{\pi_h}[(\widehat{f}(X) - f_h^*(X))^2 \mathbf{1}_{X \le K-k}] \\
=&\sum_{x=0}^{K-k} P(x+k,k)^2 p_{\pi_h}(x) \left(\frac{p_{\widehat{\pi}}(x+k)}{p_{\widehat{\pi}}(x)} - \frac{p_{\pi_h}(x+k)}{p_{\pi_h}(x)} \right)^2 \\
\le&\sum_{x=0}^{K-k} P(x+k,k)^2 p_{\pi_h}(x) \Bigg(3 \left(\frac{p_{\widehat{\pi}}(x+k)}{p_{\widehat{\pi}}(x)} - \frac{2p_{\widehat{\pi}}(x+k)}{p_{\widehat{\pi}}(x) + p_{\pi_h}(x)} \right)^2 \\
&+ 3 \left(\frac{2p_{\pi_h}(x+k)}{p_{\widehat{\pi}}(x) + p_{\pi_h}(x)} - \frac{p_{\pi_h}(x+k)}{p_{\pi_h}(x)} \right)^2 + 3 \left(\frac{2p_{\widehat{\pi}}(x+k) - 2p_{\pi_h}(x+k)}{p_{\widehat{\pi}}(x) + p_{\pi_h}(x)} \right)^2 \Bigg) \\
=&3 \sum_{x=0}^{K-k} \Bigg(\left(\frac{P(x+k,k) p_{\widehat{\pi}}(x+k)}{p_{\widehat{\pi}}(x)} \right)^2 \frac{p_{\pi_h}(x)(p_{\pi_h}(x) - p_{\widehat{\pi}}(x))^2}{(p_{\widehat{\pi}}(x) + p_{\pi_h}(x))^2} \\
&+ \left(\frac{P(x+k,k) p_{\pi_h}(x+k)}{p_{\pi_h}(x)} \right)^2 \frac{p_{\pi_h}(x)(p_{\pi_h}(x) - p_{\widehat{\pi}}(x))^2}{(p_{\widehat{\pi}}(x) + p_{\pi_h}(x))^2} \\
&+ 4P(x+k,k)^2 \frac{p_{\pi_h}(x)(p_{\widehat{\pi}}(x+k) - p_{\pi_h}(x+k))^2}{(p_{\widehat{\pi}}(x) + p_{\pi_h}(x))^2} \Bigg)
\end{aligned}
$$

where we have used $(x+y+z)^2 \le 3(x^2+y^2+z^2)$. Since $p_{\pi_h}(x) < p_{\widehat{\pi}}(x) + p_{\pi_h}(x)$, we can further write

$$
\begin{aligned}
&\mathbb{E}_{\pi_h}[(\widehat{f}(X) - f_h^*(X))^2 \mathbf{1}_{X \le K-k}] \\
\le&3 \sum_{x=0}^{K-k} \Bigg(\left(\frac{P(x+k,k) p_{\widehat{\pi}}(x+k)}{p_{\widehat{\pi}}(x)} \right)^2 \frac{(p_{\pi_h}(x) - p_{\widehat{\pi}}(x))^2}{p_{\widehat{\pi}}(x) + p_{\pi_h}(x)} \\
&+ \left(\frac{P(x+k,k) p_{\pi_h}(x+k)}{p_{\pi_h}(x)} \right)^2 \frac{(p_{\pi_h}(x) - p_{\widehat{\pi}}(x))^2}{p_{\widehat{\pi}}(x) + p_{\pi_h}(x)} + 4P(x+k,k)^2 \frac{(p_{\widehat{\pi}}(x+k) - p_{\pi_h}(x+k))^2}{p_{\widehat{\pi}}(x) + p_{\pi_h}(x)} \Bigg) \\
=&3 \sum_{x=0}^{K-k} (\widehat{f}(x)^2 + f_h^*(x)^2) \frac{(p_{\pi_h}(x) - p_{\widehat{\pi}}(x))^2}{p_{\widehat{\pi}}(x) + p_{\pi_h}(x)} + 12 \sum_{x=0}^{K-k} P(x+k,k)^2 \frac{(p_{\widehat{\pi}}(x+k) - p_{\pi_h}(x+k))^2}{p_{\widehat{\pi}}(x) + p_{\pi_h}(x)} \\
\le&3(h^{2k} + \widehat{h}^{2k}) \sum_{x=0}^{K-k} \frac{(p_{\pi_h}(x) - p_{\widehat{\pi}}(x))^2}{p_{\widehat{\pi}}(x) + p_{\pi_h}(x)} + 12 \sum_{x=0}^{K-k} P(x+k,k)^2 \frac{(p_{\widehat{\pi}}(x+k) - p_{\pi_h}(x+k))^2}{p_{\widehat{\pi}}(x) + p_{\pi_h}(x)}.
\end{aligned}
$$

Now, note that

$$
(\sqrt{p_{\widehat{\pi}}(x)} + \sqrt{p_{\pi_h}(x)})^2 \le 2(p_{\widehat{\pi}}(x) + p_{\pi_h}(x))
$$

so we have the bound

$$
(p_{\widehat{\pi}}(x) - p_{\pi_h}(x))^2 \le 2(p_{\widehat{\pi}}(x) + p_{\pi_h}(x)) \left(\sqrt{p_{\widehat{\pi}}(x)} - \sqrt{p_{\pi_h}(x)} \right)^2.
$$

We also see that $P(x+k,k) \le K^k$ for $x \le K-k$, we have

$$\begin{aligned}
&\mathbb{E}_{\pi_h}[(\widehat{f}(X) - f_h^*(X))^2 \mathbf{1}_{X \le K-k}] \\
\le& 6(h^{2k} + \widehat{h}^{2k}) \sum_{x=0}^{K-k} (\sqrt{p_{\widehat{\pi}}(x)} - \sqrt{p_{\pi_h}(x)})^2 \\
&+ 24K^k \max_{x \le K-k} \frac{P(x+k,k)p_{\widehat{\pi}}(x+k) + P(x+k,k)p_{\pi_h}(x+k)}{p_{\widehat{\pi}}(x) + p_{\pi_h}(x)} \sum_{x=0}^{K-k} \left(\sqrt{p_{\widehat{\pi}}(x+k)} - \sqrt{p_{\pi_h}(x+k)} \right)^2 \\
\le& \left(6(h^{2k} + \widehat{h}^{2k}) + 24K^k \max_{x \le K-k} (f_{\widehat{\pi}}(x) + f_{\pi_h}(x)) \right) H^2(p_{\widehat{\pi}}, p_{\pi_h}) \\
\le& \left(6(h^{2k} + \widehat{h}^{2k}) + 24K^k (h^k + \widehat{h}^k) \right) H^2(p_{\widehat{\pi}}, p_{\pi_h}). \qquad \text{(B.2)}
\end{aligned}$$

Note that by triangle inequality on Hellinger distance,

$$H^2(p_{\widehat{\pi}}, p_{\pi_h}) \le (H(p_{\widehat{\pi}}, p_\pi) + H(p_\pi, p_{\pi_h}))^2 \le 2H^2(p_{\widehat{\pi}}, p_\pi) + 2H(p_\pi, p_{\pi_h})^2.$$

But

$$H(p_\pi, p_{\pi_h})^2 \le 2\mathrm{TV}(p_\pi, p_{\pi_h}) \le 2\mathrm{TV}(\pi, \pi_h) = 4\mathbb{P}_\pi(\theta > h)$$

where TV is the total variation, the middle inequality is from the data processing inequality, and the last equality $\mathrm{TV}(\pi, \pi_h) = 2\mathbb{P}_\pi(\theta > h)$ is justified in [11, Appendix B]. Combining this with (B.2),

$$\mathbb{E}_{\pi_h}[(\widehat{f}(X) - f_h^*(X))^2 \mathbf{1}_{X \le K-k}] \le \left(12(h^{2k} + \widehat{h}^{2k}) + 48K^k (h^k + \widehat{h}^k) \right) \left(H^2(p_{\widehat{\pi}}, p_\pi) + 4\mathbb{P}_\pi(\theta > h) \right).$$

We can also bound

$$\mathbb{E}_{\pi_h}[(\widehat{f}(X) - f_h^*(X))^2 \mathbf{1}_{X > K-k}] \le (h^k + \widehat{h}^k)^2 \mathbb{P}_{f_{\pi_h}}(X > K-k) = (h^k + \widehat{h}^k)^2 \frac{\mathbb{P}_{f_\pi}(X > K-k)}{\mathbb{P}_\pi(\theta \le h)}.$$

Using $\mathbb{P}_\pi(\theta \le h) \ge \frac{1}{2}$, summing these two inequalities, and combining with (B.1) gives the desired bound. □

Appendix C

Auxiliary Proofs for ERM

Proof of Lemma 10. Recall $N(x)$ is the sample frequency and define the quantity

$$\epsilon(x) = \sum_{i=1}^{n} \epsilon_i \mathbf{1}_{X_i = x}.$$

We first prove the bound on $T_2(b, n)$. Defining $f(x) = f^*(x) = 0$ for $x < 0$, we have

$$\begin{aligned}
&\sum_{i=1}^{n} 2\epsilon_i (f^*(X_i)(f^*(X_i) - f(X_i)) - P(X_i, k)(f^*(X_i - k) - f(X_i - k))) - \frac{1}{b}(f^*(X_i) - f(X_i))^2 \\
=&\sum_{x \geq 0} 2\epsilon(x)(f^*(x)(f^*(x) - f(x)) - P(x, k)(f^*(x - k) - f(x - k))) - \frac{N(x)}{b}(f^*(x) - f(x))^2 \\
=&\sum_{x \geq 0} 2(\epsilon(x) f^*(x) - P(x + k, k)\epsilon(x + k))(f^*(x) - f(x)) - \frac{N(x)}{b}(f^*(x) - f(x))^2 \qquad \text{(C.1)}
\end{aligned}$$

We substitute (C.1) back into $T_2(b, n)$ and then split it into two terms

$$\begin{aligned}
t_1(n) = \mathbb{E}\Bigg\{ \sup_{f \in \mathcal{F}_* \cup \mathcal{F}'_*} \Bigg[\sum_{x \geq 0} \Bigg(&2(\epsilon(x) f^*(x) - P(x + k, k)\epsilon(x + k))(f^*(x) - f(x)) \\
&- \frac{N(x)}{b}(f^*(x) - f(x))^2 \Bigg) \mathbf{1}_{N(x) > 0} \Bigg] \Bigg\} \qquad \text{(C.2)}
\end{aligned}$$

$$t_0(n) = \mathbb{E}\left\{ \sup_{f \in \mathcal{F}_* \cup \mathcal{F}'_*} \left[\sum_{x \geq 0} -2P(x + k, k)\epsilon(x + k)(f^*(x) - f(x)) \mathbf{1}_{N(x) = 0} \right] \right\} \qquad \text{(C.3)}$$

We start with the $t_1(n)$ term. Since $N(x) > 0$ and $2ax - bx^2 \leq \frac{a^2}{b}$, (C.2) becomes

$$t_1(n) \leq b \cdot \mathbb{E}\left[\sum_{x \geq 0} \frac{(\epsilon(x) f^*(x) - P(x + k, k)\epsilon(x + k))^2}{N(x)} \mathbf{1}_{N(x) > 0} \right]. \qquad \text{(C.4)}$$

Plugging in $\mathbb{E}[\epsilon(x)|X_1,\dots,X_n]=0$ and $\mathbb{E}[(\epsilon(x))^2|X_1,\dots,X_n]=N(x)$ into (C.4), we get

$$\begin{aligned} t_1(n) \leq & b\cdot\mathbb{E}\left[\sum_{x\geq 0}\frac{(\epsilon(x)f^*(x)-P(x+k,k)\epsilon(x+k))^2}{N(x)}\mathbf{1}_{N(x)>0}\right] \\ = & b\cdot\mathbb{E}\left[\sum_{x\geq 0}\left((f^*(x))^2+\frac{P(x+k,k)^2N(x+k)}{N(x)}\right)\mathbf{1}_{N(x)>0}\right]. \end{aligned} \tag{C.5}$$

We now split up the summation in (C.5) to get

$$\begin{aligned} \frac{1}{b}t_1(n) \leq & \mathbb{E}\left[\sum_{x\geq 0}f^*(x)^2\mathbf{1}_{N(x)>0}\right]+\sum_{x\geq 0}P(x+k,k)^2\frac{np_\pi(x+k)}{1-p_\pi(x)}\mathbb{E}\left[\frac{\mathbf{1}_{N(x)>0}}{N(x)}\right] \\ \overset{(a)}{\leq} & h^{2k}\mathbb{E}[1+X_{max}]+\frac{n(k!)^2p_\pi(k)}{1-p_\pi(0)}\mathbb{E}\left[\frac{\mathbf{1}_{N(0)>0}}{N(0)}\right] \\ & +\frac{n}{1-\frac{1}{\sqrt{2\pi}}}\sum_{x\geq 1}P(x+k,k)^2p_\pi(x+k)\mathbb{E}\left[\frac{\mathbf{1}_{N(x)>0}}{N(x)}\right] \\ \overset{(b)}{\leq} & h^{2k}\mathbb{E}[1+X_{max}]+2c'\max\{h^k,1\}+c''h^k\sum_{x\geq 1}P(x+k,k)\min\{(np_\pi(x))^2,1\} \end{aligned} \tag{C.6}$$

where (a) follows since the first summation is 0 for anything over X_{max} and the summand is at most h^{2k} and (b) follows from (P1), (P3), and (2.5). For the third term of (C.6), we have

$$\begin{aligned} & h^k\sum_{x\geq 1}P(x+k,k)\min\{(np_\pi(x))^2,1\} \leq h^kM^{k+1}+h^k\sum_{x\geq M}P(x+k,k)\min\{(np_\pi(x))^2,1\} \\ & \overset{(a)}{\leq} h^kM^{k+1}+2^kn^2h^k\sum_{x\geq M}P(x,k)(p_\pi(x))^2 \overset{(b)}{\leq} h^kM^{k+1}+2^kn^2h^{2k}\mathbb{P}_{X\sim p_\pi}[X>M] \\ & \leq 2^k\left(h^kM^{k+1}+\frac{h^{2k}}{n^5}\right) \end{aligned} \tag{C.7}$$

where (a) we used the crude inequality $P(x+k,k)\leq 2^kP(x,k)$ for $x\geq M\geq k$, and (b) is because

$$P(x,k)p_\pi(x)=f^*(x-k)p_\pi(x-k)\leq h^k.$$

The $\frac{h^{2k}}{n^5}$ term disappears asymptotically, so substituting (C.7) back into (C.6), we obtain

$$\frac{1}{b}t_1(n)\leq h^{2k}M+2\max\{h^k,1\}+2^kh^kM^{k+1}. \tag{C.8}$$

Now we bound $t_0(n)$. We know $|\epsilon(x+k)|\leq N(x+k)=0$ for $x\geq X_{\max}-k+1$. Thus

$$\begin{aligned} t_0(n) \leq & \mathbb{E}\left[\sum_{x\geq 0}2P(x+k,k)N(x+k)\sup_{f\in\mathcal{F}_*\cup\mathcal{F}_*'}|f^*(x)-f(x)|\mathbf{1}_{N(x)=0}\right] \\ \leq & \mathbb{E}\left[\sum_{x=0}^{X_{\max}-k}2P(x+k,k)(f^*(x)+X_{\max}^k+X_{\max}'^k)N(x+k)\mathbf{1}_{N(x)=0}\right] \end{aligned} \tag{C.9}$$

Let $A = \{X_{\max} \le M, X'_{\max} \le M\}$. Then $\mathbb{P}[A^C] \le \frac{2}{n^6}$ by union bounding. Thus for some absolute constant $c > 0$:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{x=0}^{X_{\max}-k} 2P(x+k,k)(f^*(x) + X_{\max}^k + X_{\max}'^k)N(x+k)\mathbf{1}_{N(x)=0}\mathbf{1}_{A^C}\right] \\
\le&\mathbb{E}\left[X_{\max}^k(h^k + X_{\max}^k + X_{\max}'^k)\sum_{x=0}^{X_{\max}-k} N(x+k)\mathbf{1}_{N(x)=0}\mathbf{1}_{A^C}\right] \\
\overset{(a)}{\le}&n\mathbb{E}\left[X_{\max}^k(h^k + X_{\max}^k + X_{\max}'^k)\mathbf{1}_{A^C}\right] \\
\overset{(b)}{\le}&n\sqrt{\mathbb{E}\left[X_{\max}^{2k}(h^k + X_{\max}^k + X_{\max}'^k)^2\right]}\sqrt{\mathbb{P}[A^C]} \le \frac{cM^{2k}}{n^2}
\end{aligned} \tag{C.10}$$

where (a) follows from $\sum_{x=0}^{X_{\max}-k} N(x+k) \le \sum_{x=0}^{\infty} N(x) = n$, and (b) follows from Cauchy-Schwarz.

For each $x \le M$, define $q_{\pi,M}(x) = \frac{p_\pi(x)}{\mathbb{P}_{X\sim p_\pi}[X \le M]}$. Note that $\mathbb{P}[N(x) = 0|A] = (1-q_{\pi,M}(x))^n$ and conditioned on A and $N(x) = 0$, the random variable $N(x+k) \sim \text{Binom}\left(n, \frac{q_{\pi,M}(x+k)}{1-q_{\pi,M}(x)}\right)$. Then

$$\begin{aligned}
&\mathbb{E}\left[\sum_{x=0}^{X_{\max}-k} 2P(x+k,k)(f^*(x) + X_{\max}^k + X_{\max}'^k)N(x+k)\mathbf{1}_{N(x)=0}\mathbf{1}_{A}\right] \\
\le&\mathbb{E}\left[\sum_{x=0}^{X_{\max}-k} 2P(x+k,k)(f^*(x) + X_{\max}^k + X_{\max}'^k)N(x+k)\mathbf{1}_{N(x)=0}\middle|A\right] \\
\le&\sum_{x=0}^{M-k} 2P(x+k,k)(h^k + 2M^k)\mathbb{E}\left[N(x+k)\mathbf{1}_{N(x)=0}|A\right] \\
=&\sum_{x=0}^{M-k} 2P(x+k,k)(h^k + 2M^k)\mathbb{E}\left[N(x+k)|N(x)=0,A\right]\mathbb{P}[N(x)=0|A] \\
\le&\sum_{x=0}^{M-k} 2P(x+k,k)(h^k + 2M^k)\frac{nq_{\pi,M}(x+k)}{1-q_{\pi,M}(x)}(1-q_{\pi,M}(x))^n \\
\overset{(a)}{=}&\sum_{x=0}^{M-k} 2(h^k + 2M^k)f^*(x)nq_{\pi,M}(x)(1-q_{\pi,M}(x))^{n-1} \\
\le&2Mh^k(h^k + 2M^k)
\end{aligned} \tag{C.11}$$

where (a) is because $f^*(x) \le h$ for all x and $nw(1-w)^{n-1} \le (1-\frac{1}{n})^{n-1} < 1$ for all $w \in [0,1]$. Summing (C.10) and (C.11) and continuing (C.9), we have

$$t_0(n) \le \frac{cM^{2k}}{n^2} + 2Mh^k(h^k + 2M^k)$$

and combining with (C.8), we obtain the desired bound

$$T_2(b,n) \le \max\{1,h^{2k}\}M + \max\{1,h^k\}M^{1+k} + \frac{M^{2k}}{n^2} = \max\{1,h^{2k}\}M + \max\{1,h^k\}M^{1+k} + o_{h,k}(1).$$

Next we bound $T_1(b,n)$. Let $m_b = b+1$. Given two sets of samples $X_1,\dots,X_n$, for any $f \in \mathcal{F}_* \cup \mathcal{F}'$, define

$$v(f) = \min\{\max\{x : f(x) \le m_b h^k\}, X_{\max}\}.$$

Then for each f, conditional on the samples,

$$\begin{aligned}
&\sum_{i=1}^{n}(\epsilon_i - \frac{1}{b})(f(X_i) - f^*(X_i))^2 \\
=&\sum_{x:N(x)>0}(\epsilon(x) - \frac{1}{b}N(x))(f(x) - f^*(x))^2 \\
=&\left(\sum_{x=0}^{v(f)} + \sum_{x=v(f)+1}^{X_{\max}}\right)(\epsilon(x) - \frac{1}{b}N(x))(f(x) - f^*(x))^2 \\
\le& m_b^2 h^{2k}\sum_{x=0}^{X_{\max}} \max\left\{\epsilon(x) - \frac{1}{b}N(x), 0\right\} \\
&+ \sup_{v\ge 0}\left\{\sup_{m_b h^k \le f \le X_{\max}}\left\{\sum_{x>v}^{X_{\max}}(\epsilon(x) - \frac{1}{b}N(x))(f(x) - f^*(x))^2.\right\}\right\} \qquad \text{(C.12)}
\end{aligned}$$

By [4, Lemma 5], the first term of (C.12) is bounded by

$$\mathbb{E}\left[m_b^2 h^{2k}\sum_{x=0}^{X_{\max}} \max\left\{\epsilon(x) - \frac{1}{b}N(x), 0\right\}\middle| X_1^n\right] \le N_b m_b^2 h^{2k}\mathbb{E}[1 + X_{\max}] \le N_b m_b^2 h^{2k}(1+M). \qquad \text{(C.13)}$$

where $N_b \triangleq \frac{1-\frac{1}{b}}{e\cdot D(\frac{1+\frac{1}{b}}{2}||\frac{1}{2})}$.

Note that for f with values in $[m_b h^k, X_{\max}^k]$, we have $\frac{m_b-1}{m_b} f \le f - f^* \le f$ so

$$(\epsilon(x) - \frac{1}{b}N(x))(f(x) - f^*(x))^2 \le \max\left\{\epsilon(x) - \frac{1}{b}N(x), \left(\frac{m_b-1}{m_b}\right)^2(\epsilon(x) - \frac{1}{b}N(x))\right\} f(x)^2.$$

Since $-N(x) \le \epsilon(x) \le N(x)$, dividing by $N(x)$, yields $\frac{\epsilon(x)}{N(x)} \in [-1,1]$. Now consider the function

$$g(x) = \max\left\{\left(x - \frac{1}{b}\right), \left(\frac{m_b-1}{m_b}\right)^2\left(x - \frac{1}{b}\right)\right\}.$$

Since it is the max of two linear functions, it is convex and thus upper bounded by the line connecting the two endpoints $(-1, -\frac{b}{b+1})$ and $(1, \frac{b-1}{b})$, which is $\frac{2b^2-1}{2b(b+1)}\left(x - \frac{1}{2b^2-1}\right)$. Thus, the second term of (C.12) satisfies

$$\begin{aligned}
&\sup_{v\ge 0}\left\{\sup_{m_b h^k \le f \le X_{\max}}\left\{\sum_{x>v}^{X_{\max}}(\epsilon(x) - \frac{1}{b}N(x))(f(x) - f^*(x))^2.\right\}\right\} \\
\le& c_2(b)\sup_{v\ge 0}\left\{\sup_{m_b h^k \le f \le X_{\max}}\left\{\sum_{x>v}^{X_{\max}}(\epsilon(x) - \frac{1}{2b^2-1}N(x))f(x)^2.\right\}\right\} \qquad \text{(C.14)}
\end{aligned}$$

This can be viewed as an $X_{\max}+1$ dimensional linear programming problem with unknowns being the values $f(0),\dots,f(X_{\max})$. The set of solutions $m_b^2 h^{2k} \le f(0)^2 \le \cdots \le f(X_{\max})^2 \le X_{\max}^{2k}$ is convex, so the optimum value must occur at one of the corners. Thus, we can further upper bound (C.14) by

$$m_b^2 h^{2k} \sum_{x=0}^{X_{\max}} \max\left\{\epsilon(x) - \frac{1}{2b^2-1}N(x), 0\right\} + (X_{\max})^{2k} \sup_{v\ge 0}\left\{\sum_{x>v}^{X_{\max}}\left(\epsilon(x) - \frac{1}{2b^2-1}N(x)\right)\right\}$$

Again by [4, Lemma 5], the first term is at most $N_b m_b^2 h^{2k}(1+X_{\max})$. Now by [4, Lemma 6], we have

$$\mathbb{E}\left[\sup_{v\ge 0}\left\{\sum_{x>v}^{X_{\max}}\left(\epsilon(x) - \frac{1}{2b^2-1}N(x)\right)\right\}\middle| X_1^n\right] \le \sup_{w:0\le w\le n}(\epsilon_{w+1}+\cdots+\epsilon_n) - \frac{1}{2b^2-1}(n-w) \le c(b)$$

for some constant $c(b)$. Thus

$$\mathbb{E}\left[(X_{\max})^{2k}\sup_{v\ge 0}\left\{\sum_{x>v}^{X_{\max}}\left(\epsilon(x) - \frac{1}{2b^2-1}N(x)\right)\right\}\middle| X_1^n\right] \le c(b)(1+X_{\max})^{2k}.$$

Substituting these bounds back into (C.14), we get

$$\begin{aligned}&\mathbb{E}\left[\sup_{v\ge 0}\left\{\sup_{m_b h\le f\le X_{\max}}\left\{\sum_{x>v}^{X_{\max}}(\epsilon(x) - \frac{1}{b}N(x))(f(x)-f^*(x))^2.\right\}\right\}\middle| X_1^n\right]\\ \le& c_3(b)(h^{2k}(1+X_{\max}) + (1+X_{\max})^{2k})\end{aligned}$$

and combining with (C.13) and (C.12), we have the desired bound

$$T_1(b,n) \le c_3(b)\left(h^{2k}(1+M) + M^{2k}\right).$$

□

Appendix D

Auxiliary Proofs for the Lower Bound

Proof of Lemma 12. As per [15] we introduce the kernel $K(x, y)$, defined as (and also satisfies)

$$K(x,y) = \frac{G_0(x)}{f_0(y)} e^{-x} \frac{x^y}{y!} \text{ and } Kr(y) = \int_{\mathbb{R}^+} dx r(x) K(x,y)$$

Motivated by the paper, we first consider the following identity that holds true for all integers $N \geq m$ and real c:

$$\frac{\partial^m}{\partial x^m} x^N e^{cx} = e^{cx} \sum_{j=0}^{m} \binom{m}{j} c^{m-j} P(N,j) x^{N-j}.$$

To show this, we use induction on m: base case $m = 0$ is clear; for induction step,

$$\begin{aligned}
\frac{\partial^{m+1}}{\partial x^{m+1}} x^N e^{cx} &= \frac{\partial}{\partial x} \left(e^{cx} \sum_{j=0}^{m} \binom{m}{j} c^{m-j} P(N,j) x^{N-j} \right) \\
&= e^{cx} \left(c \sum_{j=0}^{m} \binom{m}{j} c^{m-j} P(N,j) x^{N-j} + \sum_{j=0}^{m} \binom{m}{j} c^{m-j} (N-j) P(N,j) x^{N-j-1} \right) \\
&= e^{cx} \sum_{j=0}^{m+1} \left(\binom{m}{j} + \binom{m}{j-1} \right) c^{m+1-j} P(N,j) x^{N-j} \\
&= e^{cx} \sum_{j=0}^{m+1} \binom{m+1}{j} c^{m+1-j} P(N,j) x^{N-j}
\end{aligned}$$

note the use of $(N-j)P(N,j) = P(N,j+1)$ for $j < N$, and that $\binom{m}{j} + \binom{m}{j-1} = \binom{m}{j+1}$.

Now going back to our computation, we first consider the LHS, $K_k(r)$. First, we can easily see that

$$K(x^k r)(y) = \frac{f_0(y+k)}{f_0(y)} P(y+k,k) Kr(y+k).$$

Thus

$$K_k r = \frac{f_0(y+k)}{f_0(y)} P(y+k,k)(Kr(y+k) - Kr(y))$$
$$= \frac{f_0(y+k)}{f_0(y)} P(y+k,k) \int_{\mathbb{R}^+} dx r(x)(K(x,y+k) - K(x,y)). \tag{D.1}$$

In addition, by the definition of $K(x,y)$, we have $\frac{K(x,y+k)}{K(x,y)} = \frac{f_0(y+k)}{f_0(y)} \frac{x^k}{P(y+k,k)}$, giving the aforementioned term as

$$\left(x^k - \frac{f_0(y+k)}{f_0(y)} P(y+k,k) \right) \int_{\mathbb{R}^+} dx r(x) K(x,y)$$

If $G_0 = \text{Gamma}(\alpha, \beta)$, then f_0 is Negative binomial, and satisfies (c.f. [15, (Equation 54)])

$$f_0(y) = \binom{y+\alpha-1}{y} \left(\frac{\beta}{1+\beta} \right)^\alpha \left(\frac{1}{1+\beta} \right)^y$$

Thus (D.1) now becomes

$$\left(x^k - P(y+\alpha+k-1,k) \left(\frac{1}{1+\beta} \right)^k \right) \int_{\mathbb{R}^+} dx r(x) \frac{\beta^\alpha}{\Gamma(\alpha) f_0(y)} e^{-(1+\beta)x} \frac{x^{y+\alpha-1}}{y!}.$$

$$= \left(x^k - P(y+\alpha+k-1,k) \left(\frac{1}{1+\beta} \right)^k \right) \int_{\mathbb{R}^+} dx K(x,y) r(x)$$

Now onwards to the RHS. We first consider the general form of $K(x^m r^{(j)})$. Again, applying [15, (Equation 58)] iteratively for m times, we get for any function g we have

$$K(x^m g)(y) = \frac{f_0(y+m)}{f_0(y)} (y+1) \cdots (y+m) K g(y+m)$$

and for $j < m$, we see that $(\partial x)^j K(0, y+m) = 0$ for all $y \geq 0$. We know $r^{(j)}$ is bounded for $j < m$, and for the Gamma distribution, we also know $(\partial x)^j K(\infty, y+m) = 0$, giving rise the following (applying integration by parts as per [15] for j times)

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

We now focus on the case where the prior G_0 is Gamma function. For $y \geq j$, we have

$$(\partial_x)^j K(x,y) = \frac{1}{f_0(y) y!} (\partial_x)^j (G_0(x) e^{-x} x^y) = \frac{\beta^\alpha}{\Gamma(\alpha) f_0(y) y!} (\partial_x)^j (e^{-(\beta+1)x} x^{\alpha-1+y})$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha) f_0(y) y!} e^{-(\beta+1)x} \sum_{i=0}^{j} \binom{j}{i} (-(\beta+1))^{j-i} P(\alpha-1+y, i) x^{(\alpha-1+y)-i}$$

$$= K(x,y)\sum_{i=0}^{j}\binom{j}{i}(-(\beta+1))^{j-i}P(\alpha-1+y,i)x^{-i}$$

i.e. for all $m \geq j$ we have

$$(\partial_x)^j K(x,y+m) = K(x,y+m)\sum_{i=0}^{j}\binom{j}{i}(-(\beta+1))^{j-i}P(\alpha-1+y+m,i)x^{-i}$$

$$= K(x,y)\frac{f_0(y)}{(y+1)\cdots(y+m)f_0(y+m)}\sum_{i=0}^{j}\binom{j}{i}(-(\beta+1))^{j-i}P(\alpha-1+y+m,i)x^{m-i}$$

Thus to summarize:

$$\begin{aligned}
K(x^m r^{(j)})(y) &= \frac{f_0(y+m)}{f_0(y)}(y+1)\cdots(y+m)Kr^{(j)}(y+m) \\
&= \frac{f_0(y+m)}{f_0(y)}(y+1)\cdots(y+m)(-1)^j\int_{\mathbb{R}^+}dxr(x)(\partial_x)^j K(x,y+m) \\
&= (-1)^j\int_{\mathbb{R}^+}dxr(x)K(x,y)\sum_{i=0}^{j}\binom{j}{i}(-(\beta+1))^{j-i}P(\alpha-1+y+m,i)x^{m-i}
\end{aligned}$$

Note that when $m = k$, $\sum_{j=1}^{k}\frac{(-1)^{j+1}}{(1+\beta)^j}K(x^k r^{(j)})(y)$ has the following as the coefficient of $r(x)K(x,y)dx$:

$$\begin{aligned}
&\sum_{j=1}^{k}\frac{(-1)^{j+1}}{(1+\beta)^j}\binom{k}{j}\left((-1)^j\sum_{i=0}^{j}\binom{j}{i}(-(1+\beta))^{j-i}P(\alpha-1+y+k,i)x^{k-i}\right) \\
=&\sum_{i=0}^{k}(-1)^{i+1}(1+\beta)^{-i}x^{k-i}P(\alpha-1+y+k,i)\sum_{j=\max(1,i)}^{k}(-1)^j\binom{k}{j}\binom{j}{i} \qquad \text{(D.2)}
\end{aligned}$$

When $i = k$, the inner summation is just $(-1)^k$. When $i < k$, we can rearrange $\binom{k}{j}\binom{j}{i} = \binom{k}{i}\binom{k-i}{j-i}$. Using this, $k > i \geq 1$ implies $\max(1,i) = i$ so

$$\sum_{j=i}^{k}(-1)^j\binom{k}{j}\binom{j}{i} = \binom{k}{i}\sum_{j=i}^{k}(-1)^j\binom{k-i}{j-i} = \binom{k}{i}(-1)^i\sum_{j=0}^{k-i}(-1)^j\binom{k-i}{j} = 0.$$

When $i = 0$, we instead have

$$\sum_{j=1}^{k}(-1)^j\binom{k}{j} = -1.$$

Substituting back into (D.2), we obtain

$$x^k - (1+\beta)^{-k}P(\alpha-1+y+k,k)$$

which agrees with the form $K_k(r)$. This implies that

$$K_k r = \sum_{j=1}^{k} \frac{(-1)^{j+1}}{(1+\beta)^j} \binom{k}{j} K(x^k r^{(j)})$$

□

We now construct the functions to be used in Lemma 11. Like [15, Appendix B] we define $S \triangleq K^* K$ and $S_k \triangleq K_k^* K_k$, satisfying the following as per [15, (Equation 89)]:

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

Next we define the same set of functions as [15, (Equation 100)] using the generalized Laguerre polynomials L_n^ν:

$$\Gamma_n(x) = e^{-\gamma_1 x} L_n^\nu(\gamma_2 x), \quad z = (\sqrt{1+\beta} - \sqrt{\beta})^2, \quad \gamma_2 = 2\sqrt{\beta(1+\sqrt{\beta})} = 2\gamma_1, \quad \nu = \alpha - 1 \tag{D.3}$$

which satisfy

$$(S\Gamma_n, \Gamma_m) = b_n \mathbf{1}_{n=m}, \quad b_n = C_2(\alpha, \beta) z^n \frac{\Gamma(n+\alpha)}{n!}. \tag{D.4}$$

We first develop some properties of these generalized Laguerre polynomials. These polynomials grow exponentially [19, p. 22.14.13], specifically

$$|L_n^\nu(x)| \le e^{x/2} \binom{n+\nu}{n}. \tag{D.5}$$

They also follow two recurrence relations.

Lemma 16. *For all $i \ge 0$,*

$$x^i \frac{d^i}{dx^i} L_n^\nu(x) = \sum_{\ell=0}^{i} (-1)^\ell \binom{i}{\ell} P(n-\ell, i-\ell) P(n+\nu, \ell) L_{n-\ell}^\nu(x).$$

Proof. We proceed with induction. First, $i = 0$ is trivially true and $i = 1$ is just the recurrence relation from [19, p. 22.8]. Now suppose the statement is true for i. Taking the derivative on both sides and multiplying by x, we obtain

$$\begin{aligned}
&x^{i+1} \frac{d^{i+1}}{dx^{i+1}} L_n^\nu(x) + i x^i \frac{d^i}{dx^i} L_n^\nu(x) \\
&= \sum_{\ell=0}^{i} (-1)^\ell \binom{i}{\ell} P(n-\ell, i-\ell) P(n+\nu, \ell) x \frac{d}{dx} L_{n-\ell}^\nu(x) \\
&= \sum_{\ell=0}^{i} (-1)^\ell \binom{i}{\ell} P(n-\ell, i-\ell) P(n+\nu, \ell) \left((n-\ell) L_{n-\ell}^\nu(x) - (n-\ell+\nu) L_{n-\ell-1}^\nu(x) \right)
\end{aligned} \tag{D.6}$$

Plugging in $x^i \frac{d^i}{dx^i} L_n^\nu(x)$ and subtracting, (D.6) becomes

$$\begin{aligned}
&x^{i+1}\frac{d^{i+1}}{dx^{i+1}}L_n^\nu(x) \\
&= \sum_{\ell=0}^{i}(-1)^\ell \binom{i}{\ell} P(n-\ell, i-\ell)P(n+\nu,\ell)\left((n-\ell-i)L_{n-\ell}^\nu(x) - (n-\ell+\nu)L_{n-\ell-1}^\nu(x)\right) \\
&\overset{(a)}{=} \sum_{\ell=0}^{i}(-1)^\ell \binom{i}{\ell} P(n-\ell, i-\ell)P(n+\nu,\ell)(n-\ell-i)L_{n-\ell}^\nu(x) \\
&\quad - \sum_{\ell=1}^{i+1}(-1)^{\ell-1} \binom{i}{\ell-1} P(n-\ell+1, i-\ell+1)P(n+\nu,\ell-1)(n-\ell+\nu+1)L_{n-\ell}^\nu(x) \\
&= \sum_{\ell=0}^{i+1}(-1)^\ell P(n-\ell, i-\ell)P(n+\nu,\ell)L_{n-\ell}^\nu(x)\left(\binom{i}{\ell}(n-\ell-i) + \binom{i}{\ell-1}(n-\ell+1)\right) \\
&\overset{(b)}{=} \sum_{\ell=0}^{i+1}(-1)^\ell P(n-\ell, i-\ell)P(n+\nu,\ell)L_{n-\ell}^\nu(x)\binom{i+1}{\ell}(n-i) \\
&= \sum_{\ell=0}^{i+1}(-1)^\ell \binom{i+1}{\ell} P(n-\ell, i+1-\ell)P(n+\nu,\ell)L_{n-\ell}^\nu(x) \qquad \text{(D.7)}
\end{aligned}$$

where (a) comes from splitting up the summation and shifting ℓ by 1, and (b) follows from properties of binomial coefficients. More specifically:

$$\begin{aligned}
\binom{i}{\ell}(n-\ell-i) + \binom{i}{\ell-1}(n-\ell+1) &= (n-\ell-i)(\binom{i}{\ell} + \binom{i}{\ell-1}) + \binom{i}{\ell-1}(i+1) \\
&= (n-\ell-i)\binom{i+1}{\ell} + \binom{i}{\ell-1}\ell = (n-i)\binom{i+1}{\ell}.
\end{aligned}$$

This proves the inductive step since (D.7) matches the desired statement with $i+1$. □

The recurrence relation of $x^i L_n^\nu(x)$ is trickier. Nevertheless, we will see later that we only need the coefficient of $L_{n-i}^\nu(x)$.

Lemma 17. *For all $i \geq 0$,*

$$x^i L_n^\nu(x) = \sum_{p=-i}^{i} c(p,i,n,\nu) L_{n+p}^\nu(x)$$

for some function c that satisfies $c(-i,i,n,\nu) = (-1)^i P(n+\nu,i)$.

Proof. We again proceed with induction. First, $i=0$ is trivially true and $i=1$ follows from the recurrence relations from [19, p. 22.7]. Now suppose the statement is true for i.

Multiplying by x on both sides, we obtain

$$\begin{aligned}
&x^{i+1}L_n^\nu(x)\\
&=\sum_{p=-i}^{i} c(p,i,n,\nu)xL_{n+p}^\nu(x)\\
&=\sum_{p=-i}^{i} c(p,i,n,\nu)\left((2n+2p+\nu+1)L_{n+p}^\nu(x)-(n+p+1)L_{n+p+1}^\nu(x)-(n+p+\nu)L_{n+p-1}\right)\\
&=\sum_{p=-i-1}^{i+1} c(p,i+1,n,\nu)xL_{n+p}^\nu(x) \qquad \text{(D.8)}
\end{aligned}$$

where the last equality is because the only L terms in (D.8) are $L_{n-i-1}^\nu,\dots,L_{n+i+1}^\nu$ and we can define $c(p,i+1,n,\nu)$ as needed. Furthermore, the only L_{n-i-1}^ν term happens when $p=-i$, so substituting the value of $c(-i,i,p,\nu)$ gives

$$c(-i-1,i+1,n,\nu)=(-1)^{-i}P(n+\nu,i)(-(n-i+\nu))=(-1)^{-i-1}P(n+\nu,i+1)$$

as desired. □

For the rest of this paper, we will use $\gamma=\gamma_1=\frac{\gamma_2}{2}$, so $\Gamma_q(x)=e^{-\gamma x}L_q^\nu(2\gamma x)$. Now we prove a generalization of [15, Lemma 15]. This will allow us to construct a set of functions that are orthogonal and have bounded magnitude.

Lemma 18. *The functions $\Gamma_q(x)$ satisfy*

$$\|\Gamma_q(x)\|_\infty \le \binom{q+\alpha}{q} \qquad \text{(D.9)}$$

$$(S_k\Gamma_q,\Gamma_q)\ge \frac{b_q}{2^{2k}\beta^{k-1}(1+\beta)^{k+1}}P(q+\nu,k)P(q,k)z^{-k} \qquad \text{(D.10)}$$

$$(S_k\Gamma_{q_1},\Gamma_{q_2})=0 \quad \forall|q_1-q_2|\ge 2k+1 \qquad \text{(D.11)}$$

Proof. First, (D.9) follows directly from (D.3) and (D.5), so it remains to show the identities with S_k. The functions Γ all have bounded jth derivatives for $j<k$, so by Lemma 12, we have

$$\begin{aligned}
(S_k\Gamma_{q_1},\Gamma_{q_2})&=\left(\sum_{j=1}^{k}\frac{(-1)^{j+1}}{(1+\beta)^j}\binom{k}{j}K(x^k\Gamma_{q_1}^{(j)}),\sum_{j=1}^{k}\frac{(-1)^{j+1}}{(1+\beta)^j}\binom{k}{j}K(x^k\Gamma_{q_2}^{(j)})\right)\\
&=\left(\sum_{j=1}^{k}\frac{(-1)^{j+1}}{(1+\beta)^j}\binom{k}{j}S(x^k\Gamma_{q_1}^{(j)}),\sum_{j=1}^{k}\frac{(-1)^{j+1}}{(1+\beta)^j}\binom{k}{j}x^k\Gamma_{q_2}^{(j)}\right) \qquad \text{(D.12)}
\end{aligned}$$

To evaluate $x^k\Gamma_q^{(j)}$, we apply the recurrence relations Lemma 16 and Lemma 17.

$$
\begin{aligned}
&x^k\Gamma_q^{(j)}\\
=&x^k\sum_{i=0}^{j}\binom{j}{i}(-\gamma)^{j-i}e^{-\gamma x}(2\gamma)^i(L_q^\nu)^{(i)}(2\gamma x)\\
=&e^{-\gamma x}(2\gamma)^{-k}\left\{x^k\sum_{i=0}^{j}\binom{j}{i}(-\gamma)^{j-i}(2\gamma)^i(L_q^\nu)^{(i)}\right\}(2\gamma x)\\
=&e^{-\gamma x}(2\gamma)^{-k+j}\left\{\sum_{i=0}^{j}\binom{j}{i}\left(-\frac{1}{2}\right)^{j-i}x^{k-i}x^i(L_q^\nu)^{(i)}\right\}(2\gamma x)\\
=&e^{-\gamma x}(2\gamma)^{-k+j}\left\{\sum_{i=0}^{j}\binom{j}{i}\left(-\frac{1}{2}\right)^{j-i}x^{k-i}\sum_{\ell=0}^{i}(-1)^\ell\binom{i}{\ell}P(q-\ell,i-\ell)P(q+\nu,\ell)L_{q-\ell}^\nu(x)\right\}(2\gamma x)\\
=&e^{-\gamma x}(2\gamma)^{-k+j}\Bigg\{\sum_{i=0}^{j}\sum_{\ell=0}^{i}\binom{j}{i}\left(-\frac{1}{2}\right)^{j-i}(-1)^\ell\binom{i}{\ell}P(q-\ell,i-\ell)P(q+\nu,\ell)\\
&\qquad\sum_{p=-(k-i)}^{k-i}c(p,k-i,q-\ell,\nu)L_{q-\ell+p}^\nu(x)\Bigg\}(2\gamma x).
\end{aligned}
\tag{D.13}
$$

Note that $q-\ell+p$ can only be in the range $[q-k,q+k]$. Furthermore, for each i, $L_{q-k}^\nu(x)$ is only achieved when $\ell=i$ and $p=-(k-i)$, so the coefficient of it in the summation is

$$
\begin{aligned}
&\sum_{i=0}^{j}\binom{j}{i}\left(-\frac{1}{2}\right)^{j-i}(-1)^iP(q+\nu,i)(-1)^{k-i}P(q-i+\nu,k-i)\\
=&(-1)^kP(q+\nu,k)\sum_{i=0}^{j}\binom{j}{i}\left(-\frac{1}{2}\right)^{j-i}\\
=&(-1)^kP(q+\nu,k)2^{-j}.
\end{aligned}
$$

Plugging this back into (D.13), for some function c', we can write

$$
\begin{aligned}
x^k\Gamma_q^{(j)}&=e^{-\gamma x}\left\{\sum_{i=q-k}^{q+k}c'(i,j,k,q,\nu)L_i^\nu(x)\right\}(2\gamma x)\\
&=\sum_{i=q-k}^{q+k}c'(i,j,k,q,\nu)\Gamma_i(x)
\end{aligned}
\tag{D.14}
$$

where $c'(q-k,j,k,q,\nu)=(2\gamma)^{-k+j}(-1)^kP(q+\nu,k)2^{-j}=(-2\gamma)^{-k}\gamma^jP(q+\nu,k)$. Clearly when $|q_1-q_2|\geq 2k+1$, none of the Γ terms in the expansion will intersect, so the orthogonality of Γ makes $(S_k\Gamma_{q_1},\Gamma_{q_2})=0$, satisfying (D.11). Now using (D.14), for some c'', we have

$$
\sum_{j=1}^{k}\frac{(-1)^{j+1}}{(1+\beta)^j}\binom{k}{j}x^k\Gamma_q^{(j)}=\sum_{i=q-k}^{q+k}c''(i,k,q,\nu)\Gamma_i(x)
\tag{D.15}
$$

where

$$\begin{aligned} c''(q-k,k,q,\nu) &= \sum_{j=1}^{k} \frac{(-1)^{j+1}}{(1+\beta)^j}\binom{k}{j}(-2\gamma)^{-k}\gamma^j P(q+\nu,k) \\ &= -(-2\gamma)^{-k}P(q+\nu,k)\sum_{j=1}^{k}\binom{k}{j}\left(-\frac{\gamma}{1+\beta}\right)^j \\ &= (-2\gamma)^{-k}P(q+\nu,k)\left(1-\left(1-\frac{\gamma}{1+\beta}\right)^k\right) \end{aligned}$$

Therefore, plugging (D.15) back into (D.12) gives

$$\begin{aligned} (S_k\Gamma_q,\Gamma_q) &= \left(\sum_{i=q-k}^{q+k} c''(i,k,q,\nu)S\Gamma_i(x), \sum_{i=q-k}^{q+k} c''(i,k,q,\nu)\Gamma_i(x)\right) \\ &\overset{(a)}{=} \sum_{i=q-k}^{q+k} c''(i,k,q,\nu)^2 b_i \\ &\overset{(b)}{\geq} (-2\gamma)^{-2k}P(q+\nu,k)^2\left(1-\left(1-\frac{\gamma}{1+\beta}\right)^k\right)^2 b_{q-k} \\ &\overset{(c)}{=} (2\gamma)^{-2k}P(q+\nu,k)P(q,k)z^{-k}\left(1-\left(1-\frac{\gamma}{1+\beta}\right)^k\right)^2 b_q \\ &\overset{(d)}{\geq} \frac{b_q}{2^{2k}\beta^{k-1}(1+\beta)^{k+1}}P(q+\mu,k)P(q,k)z^{-k} \end{aligned}$$

where (a) follows from the orthogonality of $K\Gamma_k$, (b) follows by consiering only the case $i = q-k$, (c) uses the closed form of b_q from (D.4) and the (d) follows from plugging in $\gamma = \sqrt{\beta(1+\beta)}$ and using the fact that $1-\left(1-\frac{\gamma}{1+\beta}\right)^k \geq \frac{\gamma}{1+\beta}$ (note also $\gamma < 1+\beta$ given our choice of γ) when $k \geq 1$. Thus, we satisfy (D.10). □

With Lemma 18, we are able to prove Lemma 13 and Lemma 14.

Proof of Lemma 13. Fix m and let

$$r_q = \frac{\Gamma_q}{\sqrt{(S_k\Gamma_q,\Gamma_q)}}, \quad q \in \mathcal{Q} = \{m, m+2k+1, \ldots, (2k+2)m\}.$$

Note that this definition guarantees (5.3) and (5.4). Since $z = \frac{1}{(\sqrt{1+\beta}+\sqrt{\beta})^2}$ and $\beta \geq 2$, we have $\frac{1}{6\beta} \leq z \leq \frac{1}{4\beta} \leq \frac{1}{8}$. Then

$$\|Kr_q\|^2_{L_2(f_0)} = \frac{(S\Gamma_q,\Gamma_q)}{(S_k\Gamma_q,\Gamma_q)} \leq \frac{2^{2k}\beta^{k-1}(1+\beta)^{k+1}z^k}{P(q+\nu,k)P(q,k)} \in O_k(\frac{\beta^k}{\alpha^k m^k})$$

where the last line follows since $z^k = \Theta_k(\beta^{-k})$, $q = \Theta_k(m)$, and $q + \nu = \Omega(\alpha)$. This proves (5.5).
From the proof of [15, Lemma 11], we know $\binom{q+\alpha}{q}^2 b_q^{-1} \leq \exp\{C'(\alpha + m\log\beta)\}$ for some absolute constant C'. Using (D.9),

$$\max_{q\in\mathcal{Q}} \|r_q\|_\infty \leq \sqrt{\frac{\beta^k}{\alpha^k q^k b_q}}\binom{q+\alpha}{q} \leq \sqrt{\frac{\beta^k}{\alpha^k}} e^{C(\alpha+m\log\beta)}$$

thus proving (5.6). □

Proof of Lemma 14. We choose the same r_q as in the previous proof. We know $\nu = \alpha - 1 = 0$ and β is a constant, so

$$\|Kr_q\|^2_{L_2(f_0)} = \frac{(S\Gamma_q, \Gamma_q)}{(S_k\Gamma_q, \Gamma_q)} \leq \frac{2^{2k}\beta^{k-1}(1+\beta)^{k+1}z^k}{P(q,k)^2} = O_{\beta,k}\left(\frac{1}{m^{2k}}\right).$$

From the proof of [15, Lemma 12], we know $b_q \asymp z^q$, so using (D.9) with $\alpha = 1$, we also have for some constant $c = c(k)$ and $C = C(k)$,

$$\|r_q\|_\infty = \frac{\|\Gamma_q\|_\infty}{\sqrt{(S_k\Gamma_q, \Gamma_q)}} \leq \frac{cm}{\sqrt{m^{2k}b_q}} \leq cm^{1-k}z^{-(2k+2)m} = m^{1-k}e^{Cm}.$$

since $z < 1$. □

References

[1] H. Robbins, "Asymptotically subminimax solutions of compound statistical decision problems," *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, vol. abs/1904.10040, pp. 131–149, 1951. URL: https://digitalassets.lib.berkeley.edu/math/ucb/text/math_s2_article-10.pdf.

[2] H. E. Robbins, "An empirical bayes approach to statistics," 1956. URL: https://api.semanticscholar.org/CorpusID:26161481.

[3] J. Kiefer and J. Wolfowitz, "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," *The Annals of Mathematical Statistics*, vol. 27, no. 4, pp. 887–906, 1956. DOI: 10.1214/aoms/1177728066. URL: https://doi.org/10.1214/aoms/1177728066.

[4] S. Jana, Y. Polyanskiy, A. Teh, and Y. Wu, *Empirical bayes via erm and rademacher complexities: The poisson model*, 2023. arXiv: 2307.02070 [math.ST].

[5] B. Efron, "Two modeling strategies for empirical bayes estimation," *Statistical sci- ence: a review journal of the Institute of Mathematical Statistics*, vol. 29, no. 2, pp. 285–301, 2014. URL: http://www.jstor.org/stable/43288481.

[6] L. D. Brown, E. Greenshtein, and Y. Ritov, "The poisson compound decision problem revisited," *Journal of the American Statistical Association*, vol. 108, no. 502, pp. 741–749, 2013. URL: https://www.jstor.org/stable/24246478.

[7] Y. Polyanskiy and Y. Wu, *Self-regularizing property of nonparametric maximum likelihood estimator in mixture models*, 2020. arXiv: 2008.08244 [math.ST].

[8] J. Maritz, "On the smooth empirical bayes approach to testing of hypotheses and the compound decision problem," *Biometrika*, vol. 53, no. 1, pp. 83–100, 1968.

[9] J. van Houwelingen and T. Stijnen, "Monotone empirical bayes estimators for the con- tinuous one-parameter exponential family," *Statistica Neerlandica*, vol. 37, no. 1, pp. 29–43, 1983. DOI: https://doi.org/10.1111/j.1467-9574.1983.tb00796.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9574.1983.tb00796.x.

[10] J. van Houwelingen, "Monotonizing empirical bayes estimators for a class of discrete distributions with monotone likelihood ratio," *Statistica Neerlandica*, vol. 31, no. 3, pp. 95–104, 1977. DOI: https://doi.org/10.1111/j.1467-9574.1977.tb00756.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9574.1977.tb00756.x.

[11] S. Jana, Y. Polyanskiy, and Y. Wu, *Optimal empirical bayes estimation for the poisson model via minimum-distance methods*, 2024. arXiv: 2209.01328 [math.ST].

[12] Y. Shen and Y. Wu, *Empirical bayes estimation: When does g-modeling beat f-modeling in theory (and in practice)?* 2022. arXiv: 2211.12692 [math.ST].

[13] V. Vapnik, "Principles of risk minimization for learning theory," in *Advances in Neural Information Processing Systems*, J. Moody, S. Hanson, and R. Lippmann, Eds., vol. 4, Morgan-Kaufmann, 1991. URL: https://proceedings.neurips.cc/paper_files/paper/1991/file/ff4d5fbbafdf976cfdc032e3bde78de5-Paper.pdf.

[14] B. G. Lindsay, "The Geometry of Mixture Likelihoods: A General Theory," *The Annals of Statistics*, vol. 11, no. 1, pp. 86–94, 1983. DOI: 10.1214/aos/1176346059. URL: https://doi.org/10.1214/aos/1176346059.

[15] Y. Polyanskiy and Y. Wu, "Sharp regret bounds for empirical bayes and compound decision problems," *arXiv preprint arXiv:2109.03943*, 2021.

[16] J. L. S. S. Gupta and F. Liese, "Convergence rates of empirical bayes es- timation in exponential family," *Journal of Statistical Planning and Inference*, vol. 131, no. 1, pp. 101–115, 2005. DOI: https://doi.org/10.1016/j.jspi.2003.12.017. URL: https://www.sciencedirect.com/science/article/abs/pii/S0378375804000643.

[17] W. Jiang and C.-H. Zhang, "General maximum likelihood empirical bayes esti- mation of normal means," *The Annals of Statistics*, vol. 37, no. 4, pp. 1647–1684, 2009. DOI: https://doi.org/10.1214/08-AOS638. URL: https://www.jstor.org/stable/30243683.

[18] A. Barbehenn and S. D. Zhao, *A nonparametric regression alternative to empirical bayes approaches to simultaneous estimation*, 2023. arXiv: 2205.00336 [math.ST].

[19] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Handbook of mathematical functions: with formulas, graphs, and mathematical tables, 1964.