Empirical Bayes via ERM and Rademacher complexities: the Poisson model

Soham Jana, Yury Polyanskiy, Anzo Teh and Yihong Wu*

July 6, 2023

Abstract

We consider the problem of empirical Bayes estimation for (multivariate) Poisson means. Existing solutions that have been shown theoretically optimal for minimizing the regret (excess risk over the Bayesian oracle that knows the prior) have several shortcomings. For example, the classical Robbins estimator does not retain the monotonicity property of the Bayes estimator and performs poorly under moderate sample size. Estimators based on the minimum distance and non-parametric maximum likelihood (NPMLE) methods correct these issues, but are computationally expensive with complexity growing exponentially with dimension. Extending the approach of [BZ22], in this work we construct monotone estimators based on empirical risk minimization (ERM) that retain similar theoretical guarantees and can be computed much more efficiently. Adapting the idea of offset Rademacher complexity [LRS15] to the non-standard loss and function class in empirical Bayes, we show that the shape-constrained ERM estimator attains the minimax regret within constant factors in one dimension and within logarithmic factors in multiple dimensions.

Contents

*S.J. is with the Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ, email: soham.jana@princeton.edu. Y.P. and A.T. are with the Department of EECS, MIT, Cambridge, MA, email: yp@mit.edu and anzoteh@mit.edu. Y.W. is with the Department of Statistics and Data Science, Yale University, New Haven, CT, email: yihong.wu@yale.edu.

1 Introduction

At the heart of modern large-scale inference [Efr12], empirical Bayes is a classical topic and powerful formalism in statistics and machine learning. Consider the Poisson model in one dimension as a concrete example. In a Bayesian setting, the latent parameter θ is drawn from a prior π and the observation X is then sampled from $\mathrm{Poi}(\theta)$, the Poisson distribution with mean θ. In other words, X is distributed according to the following Poisson mixture p_π with mixing distribution π:

$$p_\pi(x) = \int e^{-\theta} \frac{\theta^x}{x!} d\pi(\theta), \quad x \in \mathbb{Z}_+. \tag{1}$$

The Bayes estimator for θ that minimizes the squared error is the posterior mean, which can be expressed in terms of the mixture density as follows:

$$f^*(x) = (x+1) \frac{p_\pi(x+1)}{p_\pi(x)} \tag{2}$$

In the empirical Bayes setting, the prior π is unknown but we have access to a training sample $X_1, \dots, X_n$ drawn independently from the mixture p_π. The goal is to learn a data-driven rule that produces vanishing excess risk over the Bayes risk, known as the *regret*[1]

$$\mathsf{Regret}_\pi(\widehat{f}) \triangleq \mathbb{E}\left[(\widehat{f}(X) - \theta)^2\right] - \mathbb{E}\left[(f^*(X) - \theta)^2\right]. \tag{3}$$

The problem of interest in this context is thus:

> *Can we construct computationally efficient and practically sound estimators of f^* with optimal regret over a class of priors?*

Preliminary analyses of the Poisson empirical Bayes problem go back to [Rob51, Rob56], who proposed the following rule as an empirical approximation of (2):

$$\widehat{f}_{\mathsf{Rob}}(X) \triangleq \widehat{f}_{\mathsf{Rob}}(X; X_1, \dots, X_n) = (X+1) \frac{N_n(X+1)}{N_n(X) + 1} \tag{4}$$

where $N_n(x) = \sum_{i=1}^n \mathbf{1}_{\{X_i = x\}}$ is the empirical count for each $x \in \mathbb{Z}_+$ in the training sample. Such an approach is termed "f-modeling" that focuses on approximating the mixture density [Efr14]. Recent theoretical developments [BGR13, PW20] have established that the Robbins method achieves the optimal rate of regret when π has either bounded support or subexponential tails. On the other hand, in practice, it is well-recognized that the Robbins estimator suffers from multiple shortcomings such as numerical instability (cf. e.g. [Mar68, Section 1], [ML18, Section 1.9], [EH21, Section 6.1]) and lack of regularity properties, including, notably, the desired monotonicity property of the Bayes rule f^* (see [HS83]).

[1] In the literature there are multiple ways to formulate the regret in empirical Bayes estimation [Zha03]. As opposed to the formulation (known as the individual regret) in (3), where the data are split into the training set $X_1, \dots, X_n$ and the test set X, one can consider the total excess risk of estimating the latent parameters $\theta_1, \dots, \theta_n$ based on $X_1, \dots, X_n$ over the Bayes risk. This quantity, known as the total regret, in fact equals to n times the individual regret (3) (with n replaced by $n-1$) as shown in [PW21, Lemma 5].

In another approach to the empirical Bayes problem, known as "g-modeling" [Efr14], one tries to mimic the structure of the Bayes estimator by substituting the prior in the posterior mean with a suitable estimator. It has recently been shown that optimal regret can be attained by g-modeling estimators based on the minimum distance methodology that first finds the best approximation $p_{\widehat{\pi}}$ to the empirical distribution of the training data under suitable distances then applies the Bayes rule with the learned prior $\widehat{\pi}$. A prominent example is the nonparametric maximum likelihood estimator (NPMLE)

$$\widehat{\pi}_{\mathsf{NPMLE}} = \operatorname*{argmax}_{Q} \prod_{i=1}^{n} p_Q(X_i) \tag{5}$$

which minimizes the Kullback-Leibler divergence. Thanks to their Bayesian form, these estimators inherit the desired regularity of Bayes estimator (such as monotonicity) and lead to more stable, accurate, and interpretable estimates in practice. Recently, [JPW22] has shown that a suite of minimum-distance estimators, including the NPMLE, attain the optimal regret similar to the Robbins estimator for both bounded or subexponential priors. In addition, when π has heavier (polynomial) tails, the NPMLE achieves the corresponding optimal regret while Robbins estimator provably fails [SW22]. However, the downside of g-modeling is its much higher computational cost. For example, (5) entails solving an infinite-dimensional convex optimization. Although in one dimension faster algorithms akin to Frank-Wolfe have been proposed [Lin83, JPW22], for multiple dimensions existing solvers essentially all boil down to maximizing the weights over a discretized domain [KM14] which clearly does not scale with the dimension.

1.1 Empirical Bayes via Empirical Risk Minimization

In this paper we propose a new approach for Poisson empirical Bayes by incorporating a framework based on *empirical risk minimization* (ERM) and the needed technology from learning theory, notably, the *offset Rademacher complexity*, refined via localization, to establish the optimality of the achieved regret. In contrast to f-modeling and g-modelling that aim at approximating the mixture density and the prior respectively, the main idea is to directly approximate the Bayes rule by solving a suitable ERM subject to certain structural constraints satisfied by the Bayesian oracle. We note that a similar technique has been applied earlier in [BZ22] to the Gaussian model; however, the theoretical guarantees therein are highly suboptimal.

The benefits of the ERM-based methodology are manifold:

1. Unlike the Robbins method, the constrained ERM produces an estimator that enjoys the same regularity as that of the Bayes rule, at a small permillage of the computational cost of g-modeling methods such as the NPMLE and other minimum-distance estimators.

2. The ERM-based estimator is scalable to high dimensions and runs in time that is polynomial in both n and the dimension d. In contrast, all existing algorithms for NPMLE are essentially grid-based and scales poorly with the dimension as $n^{\Theta(d)}$.

3. The ERM approach invites powerful tools from empirical processes theory (such as Rademacher complexity and variants) to bear on its regret.

4. The flexibility of the ERM framework allows one to easily incorporate extra constraints or replace the function class by more powerful ones (such as neural nets) in order to tackle more challenging empirical Bayes problems in high dimensions for which there is no feasible proposal so far.

To summarize, the ERM can be seen as an alternative solution to the empirical Bayes problem, that excels over the Robbins method in terms of retaining the regularity properties of the Bayes estimator, and is computationally much efficient than the other existing non-parametric alternatives. We will also show that theoretically it achieves the optimal regret for certain light-tailed classes of priors. Whether these guarantees carry over to the heavy-tailed classes of prior, where the Robbins method is known to be suboptimal and NPMLE is known to be optimal [SW22], is beyond the scope of the current paper.

Next we describe the construction of the ERM-based empirical Bayes estimator in details. To derive the objective function for the ERM, note that using $f^*(X) = \mathbb{E}[\theta|X]$, we have

$$\begin{aligned} f^* = \operatorname*{argmin}_f \mathbb{E}[(f(X)-\theta)^2] &= \operatorname*{argmin}_f \mathbb{E}[(f(X))^2 - 2\theta f(X)] \\ &= \operatorname*{argmin}_f \mathbb{E}\left[f(X)^2 - 2Xf(X-1)\right], \end{aligned}$$

where we get the last step applying the identity $\mathbb{E}[\theta f(X)] = \mathbb{E}[Xf(X-1)]$ for $X \sim \mathrm{Poi}(\theta)$. Since f^* is monotone, this naturally leads to the ERM-based estimator

$$\widehat{f}_{\mathsf{erm}} \in \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{\mathbb{E}}[f(X)^2 - 2Xf(X-1)], \tag{6}$$

where $\widehat{\mathbb{E}}[h(X)] \triangleq \frac{1}{n}\sum_{i=1}^n h(X_i)$ denotes the empirical expectation of a function h based on the sample $X_1,\dots,X_n$, and the minimization (6) is over the class of monotone functions $\mathcal{F} = \{f : f(x) \le f(x+1), \forall x \ge 0\}$. We also note that the solution (6) is only uniquely specified on the set $S \triangleq \{X_1,\dots,X_n\} \cup \{X_1-1,\dots,X_n-1\}$, which can be easily computed by an algorithm akin to isotonic regression (see Lemma 1). We then extend this solution to the whole $\mathbb{Z}_+$ in a piecewise constant manner: for those $x < \min S$, set $\widehat{f}_{\mathsf{erm}}(x) = 0$; for those $x > \max S = X_{\max} \triangleq \max\{X_1,\dots,X_n\}$, set $f(x) = f(X_{\max})$; for the remaining $x \notin S$, set $\widehat{f}_{\mathsf{erm}}(x) = \widehat{f}_{\mathsf{erm}}(\max\{y \in S : y \le x\})$. This natural piecewise constant extension clearly retains monotonicity.

We note that the above construction of the ERM-based empirical Bayes estimator can be done in a principled way for other mixture models than Poisson (see Table 1). Indeed, [BZ22] was the first to apply this approach to the Gaussian mixture model. However, only the *slow rate* of $\frac{\mathsf{polylog}(n)}{\sqrt{n}}$ is obtained for the regret by applying standard empirical process theory. In addition, they use extra constraints, such as the ones based on bounded derivatives, bounds on the parameter space, etc. These constraints can be used to further improve upon the practical performances of the ERM estimator we use for the Poisson model; however the corresponding analysis is beyond the scope of the current paper. One of the major technical contributions of the present paper is to introduce a suitable version of the *offset Rademacher complexity* [LRS15] that leads to the *fast rate* of $\frac{\mathsf{polylog}(n)}{n}$ (even with the optimal logarithmic factors!)

Mixture	$p(X\|\theta)$	Bayes estimator	ERM Objective
$\mathrm{Geo}(\theta)$	$\theta^X(1-\theta)$	$1 - \frac{p_\pi(X+1)}{p_\pi(X)}$	$\widehat{\mathbb{E}}[f(X)^2 - 2f(X) + 2f(X-1)\mathbf{1}_{\{X>0\}}]$
$\mathrm{NB}(r,\theta)$	$\binom{k+r-1}{k}(1-\theta)^r\theta^k$	$\frac{X+1}{X+r}\frac{p_\pi(X+1)}{p_\pi(X)}$	$\widehat{\mathbb{E}}[f(X)^2 - 2\frac{X+1}{X+r}f(X-1)\mathbf{1}_{\{X>0\}}]$
$\mathcal{N}(\theta,1)$	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(X-\theta)^2}{2}\right)$	$X + \frac{p'_\pi(X)}{p_\pi(X)}$	$\widehat{\mathbb{E}}[f(X)^2 - 2Xf(X) + 2f'(X)]$
$\mathrm{Exp}(\theta)$	$\theta\exp(-\theta X)$	$-\frac{p'_\pi(X)}{p_\pi(X)}$	$\widehat{\mathbb{E}}[f(X)^2 - 2f'(X)]$

Table 1: ERM objectives for other mixture models: geometric, negative binomial, normal location, and exponential distributions.

1.2 Regret optimality

In addition to its conceptual simplicity and computational advantage, the ERM-based estimator comes with strong statistical guarantees which we now describe. Let $\mathcal{P}[0,h]$ denote the class of all priors supported on the interval $[0,h]$ and $\mathsf{SubE}(s)$ the set of all s-subexponential distributions on $\mathbb{R}_+$, namely $\mathsf{SubE}(s) = \left\{G : G([t,\infty)]) \le 2e^{-t/s}, \forall t > 0\right\}$. Our main result is as follows:

Theorem 1 (Regret optimality of ERM-based estimators)**.** *Let $\widehat{f}_{\mathsf{erm}}$ be defined in (6), with $\mathcal{F}$ the class of all monotone functions on $\mathbb{Z}_+$. Then there exist s a constant $C > 0$ such that for any $h, s > 0$,*

$$\sup_{\pi\in\mathcal{P}([0,h])} \mathsf{Regret}_\pi(\widehat{f}_{\mathsf{erm}}) \le \frac{C\max\{1,h\}^3}{n}\left(\frac{\log n}{\log\log n}\right)^2, \qquad \sup_{\pi\in\mathsf{SubE}(s)} \mathsf{Regret}_\pi(\widehat{f}_{\mathsf{erm}}) \le \frac{C\max\{1,s\}^3}{n}(\log n)^3.$$

The regret bounds in Theorem 1 match the minimax lower bounds in [PW21, Theorem 2] up to constant factors, thereby establish the strong optimality of the ERM-based empirical Bayes estimators. Finally, as a side remark, we mention that, one can show that a monotone projection of the Robbins estimator, given by $\widehat{f}_{\mathsf{mono\text{-}Rob}} = \operatorname{argmin}_{f\in\mathcal{F}} \widehat{\mathbb{E}}[(f(X) - \widehat{f}_{\mathsf{Rob}}(X))^2]$, also attains similar regret guarantees as in Theorem 1. This is outside the scope of the current paper.

1.3 Multiple dimensions

The ERM-based estimator (6) can be easily extended to the d-dimension Poisson model. For clarity, we use the bold fonts to denote a vector, e.g., $\boldsymbol{\theta} = (\theta_1,\dots,\theta_d), \boldsymbol{\theta}_i = (\theta_{i1},\dots,\theta_{id}), \boldsymbol{X} = (X_1,\dots,X_d), \boldsymbol{X}_i = (X_{i1},\dots,X_{id}), \boldsymbol{x} = (x_1,\dots,x_d)$, etc. Let π be a prior distribution on $\mathbb{R}_+^d$. Consider the following data-generating process

$$\boldsymbol{\theta}_i \overset{\text{iid}}{\sim} \pi \qquad X_{ij} \overset{\text{ind.}}{\sim} \mathrm{Poi}(\theta_{ij}). \tag{7}$$

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

$$p_\pi(\boldsymbol{x}) = \int_{\boldsymbol{\theta}} \prod_{i=1}^d e^{-\theta_i}\frac{\theta_i^{x_i}}{x_i!} d\pi(\boldsymbol{\theta}), \quad \boldsymbol{x}\in\mathbb{Z}_+^d.$$

Similar to (3), let us define the regret of a given estimator $\boldsymbol{f} : \mathbb{Z}_+^d \to \mathbb{R}_+^d$ as

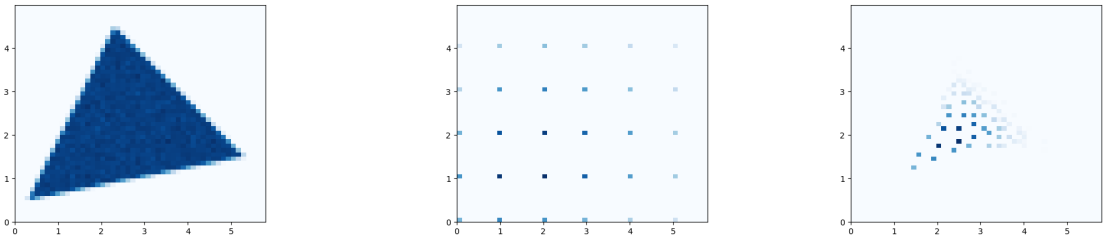
$$\mathsf{Regret}_\pi(\boldsymbol{f}) = \mathbb{E}\left[\|\boldsymbol{f}(\boldsymbol{X}) - \boldsymbol{\theta}\|^2\right] - \mathbb{E}\left[\|\boldsymbol{f}^*(\boldsymbol{X}) - \boldsymbol{\theta}\|^2\right], \tag{8}$$

where $\boldsymbol{X} \sim p_\pi$ is a test point independent from the training sample $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n \overset{\text{iid}}{\sim} p_\pi$. For each $\boldsymbol{f}$, let $\boldsymbol{f} = (f_1,\cdots,f_d)$ where $f_i : \mathbb{Z}_+^d \to \mathbb{R}_+$. Denote by $\boldsymbol{f}^*$ the Bayes estimator, whose i-th coordinate f_i^* is given by

$$f_i^*(\boldsymbol{x}) = \mathbb{E}[\theta_i|\boldsymbol{x}] = \frac{\int_{\boldsymbol{\theta}} \theta_i \prod_{j=1}^d e^{-\theta_i}\frac{\theta_i^{x_i}}{x_i!} d\pi(\boldsymbol{\theta})}{p_\pi(\boldsymbol{x})} = (x_i+1)\frac{p_\pi(\boldsymbol{x}+\boldsymbol{e}_i)}{p_\pi(\boldsymbol{x})}, \quad i = 1,\dots,d,$$

where $\boldsymbol{e}_i$ denote the i-th coordinate vector. Using Cauchy-Schwarz, one can show that the Bayes estimator for the i-th coordinate is increasing in the i-th coordinate of the input if all other coordinates are fixed, i.e.,

$$f_i^*(\boldsymbol{x}) \le f_i^*(\boldsymbol{x}+\boldsymbol{e}_i), \quad \forall i = 1,\dots,d, \quad \forall \boldsymbol{x}\in\mathbb{Z}_+^d \tag{9}$$



(a) Latent $\boldsymbol{\theta}_i$'s. (b) Observations $\boldsymbol{X}_i$'s. (c) Denoised $\widehat{\boldsymbol{f}}_{\mathsf{erm}}$.

Figure 1: A two-dimensional experiment with $n = 10^6$: Left: $\boldsymbol{\theta}_i$'s are sampled uniformly from a triangle. Middle: the observations $\boldsymbol{X}_i$'s are drawn independently from $\mathrm{Poi}(\boldsymbol{\theta}_i)$, with their empirical distribution shown on the grid $\mathbb{Z}_+^2$ (notice that this is also the MLE estimator for $\boldsymbol{\theta}$, hence very different from the empirical Bayes solution). Right: the empirical Bayes denoised version obtained by applying $\widehat{f}_{\mathsf{erm}}$ in (10) to $\boldsymbol{X}_i$'s.

This leads to the following ERM procedure.

$$\widehat{\boldsymbol{f}}_{\mathsf{erm}} = \operatorname*{argmin}_{\boldsymbol{f}\in\boldsymbol{\mathcal{F}}} \quad \widehat{\mathbb{E}}\left[\|\boldsymbol{f}(\boldsymbol{X})\|^2 - 2\sum_{j=1}^{d} X_j f_j(\boldsymbol{X}-\boldsymbol{e}_i)\right],$$
$$\boldsymbol{\mathcal{F}} = \{\boldsymbol{f}: \mathbb{Z}_+^d \to \mathbb{R}_+^d : f_i(\boldsymbol{x}) \le f_i(\boldsymbol{x}+\boldsymbol{e}_i), \forall i = 1,\cdots,d, \forall \boldsymbol{x}\in\mathbb{Z}_+^d\}. \tag{10}$$

We again note that $\widehat{\boldsymbol{f}}_{\mathsf{erm}}$ is not uniquely defined for all $\boldsymbol{x} \in \mathbb{Z}_+^d$. To specify a minimizer, note that $(\widehat{f}_{\mathsf{erm}})_j$, the j-th coordinate of $\widehat{\boldsymbol{f}}_{\mathsf{erm}}$, is uniquely defined on $S \triangleq \{\boldsymbol{X}_i\} \cup \{\boldsymbol{X}_i - \boldsymbol{e}_j\}$. We may extend it to $\mathbb{Z}_+^d$ in the same manner as the one-dimensional case of (6) in a piecewise constant manner. That is, for each $\boldsymbol{x} \notin S$, if there exists $y \ge 0$ such that $\boldsymbol{x} - y\boldsymbol{e}_j \in S$, we set $(\widehat{f}_{\mathsf{erm}})_j(\boldsymbol{x}) = (\widehat{f}_{\mathsf{erm}})_j(\min_{\boldsymbol{x}-y\boldsymbol{e}_j\in S, y\ge 0} \boldsymbol{x} - y\boldsymbol{e}_j)$. Otherwise, set $(\widehat{f}_{\mathsf{erm}})_j(\boldsymbol{x}) = 0$. By convention, we also define $(\widehat{f}_{\mathsf{erm}})_j(-\boldsymbol{e}_j) = 0$.

Theorem 2. *The ERM estimator (10) satisfies the following regret bounds whenever $n \ge d$:*

1. *If π is supported on $[0,h]^d$, then $\mathsf{Regret}_\pi(\widehat{\boldsymbol{f}}_{\mathsf{erm}}) \le O(\frac{d}{n}\max\{c_1, c_2 h\}^{d+2}(\frac{\log(n)}{\log\log(n)})^{d+1})$;*
2. *If all marginals of π are s-subexponential for some $s > 0$, then $\mathsf{Regret}_\pi(\widehat{\boldsymbol{f}}_{\mathsf{erm}}) \le O(\frac{d}{n}(\max\{c_3, c_4 s\}\log(n))^{d+2})$,*

where $c_1, c_2, c_3, c_4 > 0$ are absolute constants.

We conjecture these regret bounds in Theorem 2 are nearly optimal and factors like $(\log n)^d$ are necessary. Indeed, for the Gaussian model in d dimensions, the minimax squared Hellinger risk for density estimation is shown to be at least $O((\log n)^d/n)$ for subgaussian mixing distributions and the minimax regret is typically even larger. A rigorous proof of matching lower bound for Theorem 2 will likely involve extending the regret lower bound based on Bessel kernels in [PW21] to multiple dimensions; this is left for future work.

Remark 1 (Time complexity)**.** *For the statistical rate of ERM in multiple dimensions to be meaningful, we require d to be significantly smaller than n. Nonetheless, even in the dimensions where*

the regret in Theorem 2 is vanishing, the ERM method is computationally much more scalable, compared with the conventional approach based on NPMLE or other minimum-distance estimators.

To elaborate on this, ERM is a linear program and has a dedicated solver due to its special form. NPMLE is an infinite-dimensional convex optimization, and the prevailing solver either discretizes the domain (at least $\sqrt{n}$ level in order to be statistically relevant, thus requires a grid of size $n^{\Theta(d)}$) or runs Frank-Wolfe style iteration, which is only known to converge slowly at $\frac{1}{t}$ rate [Lin83] and requires mode finding that is expensive in multiple dimensions. In contrast, the ERM approach scales much better with the dimension. To evaluate the d-dimensional ERM (10), as we will demonstrate in Remark 4, if ℓ is the number of distinct vector-valued observations $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$, our algorithm runs in $O(d\ell^2) \le O(dn^2)$ time (apart from reading the sample of size n). An almost linear time $O(d\ell \log \ell)$ algorithm (which is how we implemented in the simulations), exists but is beyond the scope of this paper. (We will describe the basic idea in Appendix B.)

On the empirical side, we demonstrate the multidimensional feasibility of ERM by running a simulation with $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n$ sampled uniformly from a triangle with $n = 10^6$ and compute the empirical Bayes denoiser $\widehat{\boldsymbol{f}}_{\mathsf{erm}}$ in (10) to $\boldsymbol{X}_i \overset{ind.}{\sim} \mathrm{Poi}(\boldsymbol{\theta}_i)$. Here, we see that $\widehat{\boldsymbol{f}}_{\mathsf{erm}}$ can recover the triangular structure of the prior, as in Fig. 1.3. To further compare the computational costs of ERM and minimum distance methods, we did a comparison in the statistical software R with the popular package "REBayes" [KG17] and the results are as follows. With the prior $\mathsf{Unif}(4, 30)$ and sample sizes $n = 50, 500, 5000, 50000$, we ran both REBayes and ERM 100 times and found that on average the ERM is respectively $21, 50, 212, 588$ times faster. This improvement is even more pronounced ($25, 58, 227.5, 2160$ times) if we supply the empirical distribution to the ERM instead of the full sample.

Remark 2 (Comparison with f-modelling)**.** *While both f-modelling (i.e. the Robbins estimator) and the ERM estimator $\widehat{f}_{\mathsf{erm}}$ are asymptotically optimal, we demonstrate more concretely the advantage of $\widehat{f}_{\mathsf{erm}}$ over Robbins. The shortcomings of the Robbins method have been widely observed in practice and discussed in the existing literature. Most recently, it has been demonstrated in [JPW22] extensively through both simulated and real data experiment. Expanding on Fig. 1(a), which compares the performance of the multidimensional Robbins method and $\widehat{f}_{\mathsf{erm}}$ under a uniform prior on the 2d triangle, for $n = 10^k, k = 4, 5, 6, 7$, we found that the Robbins method achieved a regret of $0.356, 0.0575, 0.00771, 0.00116$ and $\widehat{f}_{\mathsf{erm}}$ achieved a regret of $0.0748, 0.0161, 0.00276, 0.000463$, suggesting a much better performance. On another experiment, we also compared the methods in dimensions $1, 2, 3, 4$ using a product of $Exp(2)$ distributions as prior, fixing $n = 10000$. The Robbins method achieved regrets $0.0125, 0.0607, 0.185, 0.427$; $\widehat{f}_{\mathsf{erm}}$ achieved regrets $0.00422, 0.0208, 0.0660, 0.161$.*

1.4 Related work

Empirical Bayes estimation for the Poisson means incorporating shape constraint has a long research thread. However, the majority of the work relies on approximating the Robbins estimator using monotone functions. For example, [Mar66] used linear approximation to the Robbins estimator and [Mar69] represented the marginal distribution p_π based on a monotone ordinate fit to the Robbins and then used it to compute a maximum likelihood estimation of the ordinates. Both of these papers focus on numerical comparison of the corresponding error guarantees; see [ML18, Section 3.4.5] for a concise exposition. In recent work, [BGR13] discussed the numerical benefits of first performing a Rao-Blackwellization on the Robbins estimator and then using an isotonic regression to impose the monotonicity of the final estimator. An important theoretical contribution to the monotone smoothing of any given empirical Bayes estimator has been proposed in [VH77]. Using the monotone likelihood ratio property of the Poisson distribution, it is shown that any estimator

(e.g., the Robbins estimator) can be made monotone without increasing the regret. In contrast, our main estimator is computed directly via minimizing an empirical version of the regret. It might be possible to use the monotone smoothing of [VH77] to further improve the ERM-estimator which is not pursued in this work.

As mentioned in Section 1.1, the application of empirical risk minimization in empirical Bayes has been introduced in the one-dimensional normal mean model by [BZ22]. Using the monotonicity of the posterior mean, they construct an empirical Bayes estimator by solving the ERM under monotonicity constraint (see Table 1). However, the regret bound they establish is of the slow rate $\frac{\mathsf{polylog}(n)}{\sqrt{n}}$ which is highly suboptimal, compared with the nearly optimal rate of $O(\frac{(\log n)^5}{n})$ by [JZ09] (based on the g-modeling approach via NPMLE) and $O(\frac{(\log n)^8}{n})$ by [LGL05] (based on the f-modeling approach of polynomial kernel density estimates). As mentioned earlier, the NPMLE is computationally expensive, especially in multiple dimensions due to the reliance on grid-based approximation [KM14, SGS21]. In contrast, as mentioned before, ERM-based estimators algorithm can be easily constructed for multiple or high dimensions.

The rest of the paper is organized as follows. Section 2 provides a regret upper bound on the ERM-based estimator in one dimension in terms of the offset Rademacher complexities, and a proof sketch for Theorem 1. Section 3 contains the analysis for the multidimensional ERM-estimator and a proof sketch of Theorem 2. Omitted proofs are provided in the appendices.

2 Regret guarantees for the ERM estimator via Offset Rademacher complexity

2.1 The ERM algorithm

As mentioned in the last section, our proposed estimator is based on ERM framework. In many statistical problems, the statistician intends to find a function f that approximates a target statistic $s(X)$ in order to minimize the error $\mathbb{E}\left[\ell\left(s(X), f(X)\right)\right]$ for some suitable loss function ℓ. In the ERM framework, the population average is replaced by the empirical average $\widehat{\mathbb{E}}\left[\ell\left(s(X); f(X)\right)\right]$ over the training sample. There is a rich literature on using such methods to approximate nonparametric target functions. See, for example, [Nem85, VdG90] for regression problems, [Bar91, BC91, Bar94] for penalized empirical risk minimization, [BM93, LZ95] for consistency results of general nonparametric ERM-estimators, etc. In this paper, we aim to approximate the nonparametric target function f^* (the Bayes rule) by minimizing $\mathbb{E}\left[(f^*(X) - f(X))^2\right]$. As shown in Section 1.1, in the Poisson mixture model, this can be equivalently expressed as minimizing $\mathbb{E}\left[f(X)^2 - 2Xf(X-1)\right]$ and we minimize the corresponding empirical loss over the class of all monotone functions. Isotonic minimization of such quadratic loss is easy to compute; [BC90] showed that monotone projection can be done in linear time. In the following lemma we present one such minimization algorithm that we use in numerical analyses. The proof is deferred to Appendix B.

Lemma 1. *Let $a_1 < \cdots < a_n$ be a sequence of non-negative integers and $\{v_i\}_{i=1}^n, \{w_i\}_{i=1}^n$ be two non-negative sequences with $v_n > 0$ and $\max\{v_i, w_i\} > 0$ for all i. Consider the iterative b_i*

$$b_i = \begin{cases} 1 & i = 0 \\ 1 + \operatorname{argmin}_{b_{i-1} \le i^* \le n} \frac{\sum_{i=b_{i-1}}^{i^*} w_i}{\sum_{i=b_{i-1}}^{i^*} v_i} & i \ge 1 \end{cases}$$

where the fraction is $+\infty$ whenever the denominator is 0, and where tie exists at argmin*, choose*

biggest such i^*. We stop at $b_m = n+1$. Then the solution to

$$\widehat{f}_{\mathsf{erm}} = \operatorname*{argmin}_{f\in\mathcal{F}} \sum_{i=1}^{n} v_i f(a_i)^2 - 2w_i f(a_i)$$

is given as

$$\forall i = 1,\cdots,m, \forall x : b_m \le x < b_{m+1} : \widehat{f}_{\mathsf{erm}}(a_x) = \frac{\sum_{i=b_m}^{b_{m+1}-1} w_i}{\sum_{i=b_m}^{b_{m+1}-1} v_i}.$$

Remark 3. *Making the restriction $v_i \ge 0$ and $v_n > 0$ ensures that our solution will be well-formed. To apply this algorithm to estimate $\widehat{f}_{\mathsf{erm}}$, let $\{a_1,\cdots,a_k\} \subseteq \{1,\cdots,X_{\max}\}$ be such that either $N(a_i) > 0$ or $N(a_i+1) > 0$. Here, $v_i = N(a_i)$ and $w_i = (a_i+1)N(a_i+1)$. Our choice of a_i's for $i = 1,\ldots,k$ ensures that $\max\{v_i, w_i\} > 0$, and also $v_k > 0$.*

Remark 4. *Lemma 1 can be applied to compute the ERM estimator (10) for the multivariate case. Recall that the function class $\boldsymbol{\mathcal{F}}$ dictates the following form of monotonicity: for each vector $\boldsymbol{x}' = (x'_1,\cdots,x'_{j-1},x'_{j+1},\cdots,x_d)$ of length $d-1$, we define*

$$C_j(\boldsymbol{x}') \triangleq \{\boldsymbol{x} \in \mathbb{R}^d_+ : x_i = x'_i, \forall i \ne j\} \tag{11}$$

Here are several examples for $d = 3$:

$$C_0((0,0)) = \{(0,0,0),(1,0,0),(2,0,0),\cdots\} \quad C_1((0,0)) = \{(0,0,0),(0,1,0),(0,2,0),\cdots\}$$

$$C_2((0,0)) = \{(0,0,0),(0,0,1),(0,0,2),\cdots\}$$

Then $\boldsymbol{f} \in \boldsymbol{\mathcal{F}}$ if and only if for each $j \in [d]$, f_j restricted on each $C_j(\boldsymbol{x}')$ is monotone in the j-th coordinate of the argument. Since the objective function $\widehat{\mathbb{E}}[\|\boldsymbol{f}(\boldsymbol{X})\|^2 - 2\sum_{j=1}^d X_j f_j(\boldsymbol{X} - \boldsymbol{e}_j)]$ is separable, for each j we may determine $(\widehat{f}_{\mathsf{erm}})_j$ by partitioning the samples $\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n$ into classes of $C_j(\boldsymbol{x}')$, and then apply Lemma 1 to each class.

To bound the regret of such ERM-estimators, we used the technique of Rademacher complexities. The Rademacher analysis, popularized by [Kol01, Men02, BBL02], etc., uses a symmetrization argument to bound the error using the supremum of an empirical process of the form $\sup_{g\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i g(X_i)$, where $\epsilon_1,\ldots,\epsilon_n$ are iid Rademacher random variables, and $\mathcal{F}$ is some suitable function class. The complexity of such a function class is often characterized by the VC dimension or the covering numbers. An immediate bound on the complexity is produced by the uniform convergence bound when $\mathcal{F}$ is chosen to be the class of all possible candidate functions, however, this has been shown to guarantee only a slow rate of regret $(\frac{1}{\sqrt{n}})$, which is the case in the prior work [BZ22] that applies the ERM approach to the Gaussian model. An improvement on this is made by restricting $\mathcal{F}$ to be a smaller class, for example using the techniques of local Rademacher complexities [BBM05, KP04, LW04] which analyzes the complexity within a small ball around the target function, the empirical minimizer, etc. We employ a similar technique of using function classes with smaller complexity. Note that the empirical minimizer in (6) satisfies the following regularity property.

Lemma 2. *Let $\widehat{f}_{\mathsf{erm}}$ be the ERM-estimator defined in (6). Let $X_{\max} = \max\{X_1,\ldots,X_n\}$. Then $\max_{0\le x\le X_{\max}} \widehat{f}_{\mathsf{erm}}(x) \le X_{\max}$.*

Proof. Recall that $\widehat{f}_{\mathsf{erm}}$ is characterized by piecewise constancy, where for each maximal interval I on which $\widehat{f}_{\mathsf{erm}}$ is constant (maximal in the sense we cannot extend I further), we have

$$\forall x_0 \in I : \widehat{f}_{\mathsf{erm}}(x_0) = \frac{\sum_{x\in I}(x+1)N(x+1)}{\sum_{x\in I} xN(x)}$$

Now that we have defined $\widehat{f}_{\mathsf{erm}}(x) = \widehat{f}_{\mathsf{erm}}(X_{\max})$ for all $x > X_{\max}$, it suffices to show that $\widehat{f}_{\mathsf{erm}}(X_{\max}) \le X_{\max}$. Indeed, there exists an $i^* \le X_{\max}$ such that

$$\begin{aligned}\widehat{f}_{\mathsf{erm}}(k) &= \frac{\sum_{i=i^*}^{X_{\max}}(i+1)N(i+1)}{\sum_{i=i^*}^{X_{\max}} N(i)} \\ &\overset{\text{(a)}}{=} \frac{\sum_{i=i^*+1}^{X_{\max}} iN(i)}{\sum_{i=i^*}^{X_{\max}} N(i)} \le \frac{\sum_{i=i^*+1}^{X_{\max}} X_{\max}N(i)}{\sum_{i=i^*}^{X_{\max}} N(i)} = X_{\max}(1 - \frac{N(i^*)}{\sum_{i=i^*}^{k} N(i)}) \le X_{\max} \end{aligned} \tag{12}$$

where (a) is due to $N(X_{\max}+1) = 0$. □

When $X_1, \ldots, X_n$ are generated from the Poisson mixture with either a compactly supported or subexponential prior, the above result implies that the value of ERM-estimator is at most $\Theta(\mathsf{polylog}(n))$ with high probability. This, in essence, dictates the required complexity of the function class.

2.2 Risk bounds for ERM via Rademacher complexities

Lemma 2 shows that $\widehat{f}_{\mathsf{erm}}$ coincides with the ERM over the following more restrictive class

$$\mathcal{F}_* \triangleq \{f : f \text{ is monotone}, f(X_{\max}) \le \max\{X_{\max}, f^*(X_{\max})\}\}. \tag{13}$$

Note that $\mathcal{F}_*$ is a (random) class that depends on the sample maximum. Furthermore, since it depends on the unknown ground truth f^*, it is not meant for data-driven optimization but only for theoretical analysis of the ERM (6). In addition, our work utilizes the quadratic structure of the empirical loss to obtain a stronger notion of the Rademacher complexity measure, which closely resembles and is motivated by the offset Rademacher complexity introduced in [LRS15].

Theorem 3. *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*

$$\mathsf{Regret}_\pi(\widehat{f}) \le \frac{3}{n}T_1(n) + \frac{4}{n}T_2(n) \tag{14}$$

where

$$T_1(n) = \mathbb{E}\left[\sup_{f\in\mathcal{F}_{p_n}\cup\mathcal{F}_{p_n'}} \sum_{i=1}^n (\epsilon_i - \frac{1}{6})(f(X_i) - f^*(X_i))^2\right], \tag{15}$$

$$\begin{aligned} T_2(n) = \mathbb{E}\Bigg[\sup_{f\in\mathcal{F}_{p_n}\cup\mathcal{F}_{p_n'}} \sum_{i=1}^n \Bigg\{ &2\epsilon_i(f^*(X_i)(f^*(X_i) - f(X_i)) - X_i(f^*(X_i - 1) - f(X_i - 1))) \\ &- \frac{1}{4}(f^*(X_i) - f(X_i))^2 \Bigg\}\Bigg], \end{aligned} \tag{16}$$

and $\mathcal{F}_{p_n'}$ is defined in the same way as $\mathcal{F}_{p_n}$ with respect to an independent copy of $X_1, \ldots, X_n$.

Proof. Define

$$\mathsf{R}(f) = \mathbb{E}\left[f(X)^2 - 2Xf(X-1)\right], \quad \widehat{\mathsf{R}}(f) = \widehat{\mathbb{E}}\left[f(X)^2 - 2Xf(X-1)\right]. \tag{17}$$

We first note that $\widehat{f}$ satisfies the following inequality, thanks to the convexity of $\mathcal{F}$:

$$\widehat{\mathsf{R}}(h) - \widehat{\mathsf{R}}(\widehat{f}) \geq \widehat{\mathbb{E}}[(h - \widehat{f})^2], \quad \forall h \in \mathcal{F}. \tag{18}$$

To show this claim, since $\mathcal{F}$ is convex,for any $\epsilon \in [0,1]$, $(1-\epsilon)\widehat{f} + \epsilon h$ is inside the class $\mathcal{F}$, so with $\widehat{\mathsf{R}}(\widehat{f}) \leq \widehat{\mathsf{R}}((1-\epsilon)\widehat{f} + \epsilon h)$ we have

$$\frac{\partial}{\partial \epsilon}\widehat{\mathsf{R}}((1-\epsilon)\widehat{f} + \epsilon h) = 2\widehat{\mathbb{E}}[(h(X) - \widehat{f}(X))((1-\epsilon)\widehat{f}(X) + \epsilon h(X)) - X(h(X-1) - \widehat{f}(X-1))]$$

By the ERM minimality of $\widehat{f}$, such derivative must be nonnegative when evaluated at 0. That is,

$$\widehat{\mathbb{E}}[(h(X) - \widehat{f}(X))\widehat{f}(X) - X(h(X-1) - \widehat{f}(X-1))] \geq 0 \tag{19}$$

Therefore, evaluating the difference gives us

$$\begin{aligned}
&\widehat{\mathsf{R}}(h) - \widehat{\mathsf{R}}(\widehat{f}) - \widehat{\mathbb{E}}[(h(X) - \widehat{f}(X))^2] \\
&= \widehat{\mathbb{E}}[(h(X)^2 - \widehat{f}(X)^2) - 2X(h(X-1) - f(X-1))] - \widehat{\mathbb{E}}[(h(X) - \widehat{f}(X))^2] \\
&= 2\widehat{\mathbb{E}}[h(X)\widehat{f}(X) - \widehat{f}(X)^2 - X(h(X-1) - \widehat{f}(X-1))] \geq 0
\end{aligned} \tag{20}$$

as desired. Then using $\mathsf{Regret}_\pi(\widehat{f}) = \mathsf{R}(\widehat{f}) - \mathsf{R}(f^*)$ we get

$$\begin{aligned}
&\mathsf{Regret}_\pi(\widehat{f}) \\
&\leq \mathbb{E}\left[\mathsf{R}(\widehat{f}) - \mathsf{R}(f^*) + \widehat{\mathsf{R}}(f^*) - \widehat{\mathsf{R}}(\widehat{f}) - \widehat{\mathbb{E}}(f^* - \widehat{f})^2\right] \\
&= \mathbb{E}\Big[(\mathsf{R}(\widehat{f}) - \mathsf{R}(f^*) - \mathbb{E}[(f^* - \widehat{f})^2]) + (\widehat{\mathsf{R}}(f^*) - \widehat{\mathsf{R}}(\widehat{f}) + \widehat{\mathbb{E}}[(f^* - \widehat{f})^2]) \\
&\quad + \mathbb{E}[(f^* - \widehat{f})^2] - 2\widehat{\mathbb{E}}[(f^* - \widehat{f})^2]\Big] \\
&= \mathbb{E}\Big[\widehat{\mathbb{E}}[2f^*(X)(f^*(X) - \widehat{f}(X)) - 2X(f^*(X-1) - \widehat{f}(X-1))] \\
&\quad -\mathbb{E}[2f^*(X)(f^*(X) - \widehat{f}(X)) - 2X(f^*(X-1) - \widehat{f}(X-1))] - \frac{1}{4}(\widehat{\mathbb{E}}[(f^* - \widehat{f})^2] + \mathbb{E}[(f^* - \widehat{f})^2])\Big] \qquad (21) \\
&\quad + \mathbb{E}\left[\frac{5}{4}\mathbb{E}[(f^*(X) - \widehat{f}(X))^2] - \frac{7}{4}\widehat{\mathbb{E}}[(f^*(X) - \widehat{f}(X))^2]\right]. \qquad (22)
\end{aligned}$$

We separately bound the two terms (21) and (22) in the above display in terms of the Rademacher complexities using the following symmetrization result.

Lemma 3. *Let $\epsilon_1, \cdots, \epsilon_n$ as independent Rademacher symbols. Let T, U be two operators mapping $f(x)$ to $Tf(x)$ and $Uf(x)$. Then*

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}_{p_n}} [\mathbb{E}[Tf(X)] - \widehat{\mathbb{E}}[Tf(X)] - (\mathbb{E}[Uf(X)] + \widehat{\mathbb{E}}[Uf(X)])]\right] \leq \frac{2}{n}\mathbb{E}\left[\sup_{f \in \mathcal{F}_{p_n} \cup \mathcal{F}_{p_n'}} \sum_{i=1}^n \epsilon_i Tf(X_i) - Uf(X_i)\right]$$

where p_n' is an independent copy of the empirical distribution p_n.

Proof. Here, we note that the symmetrization technique has been introduced in [LRS15, p.11-12]. However, given that we are taking a supremum over a data-dependent subclass of $\mathcal{F}$, some extra care needs to be taken.

$$
\begin{aligned}
&\mathbb{E}[\sup_{f\in\mathcal{F}_{p_n}\cup\mathcal{F}_{p'_n}} \widehat{\mathbb{E}}'[T(f(X))] - \widehat{\mathbb{E}}[T(f(X))] - (\widehat{\mathbb{E}}'[U(f(X))] + \widehat{\mathbb{E}}[U(f(X))])] \\
&\overset{(a)}{=} \frac{1}{2}\mathbb{E}[\sup_{f\in\mathcal{F}_{p_n}\cup\mathcal{F}_{p'_n}} \widehat{\mathbb{E}}'[T(f(X))] - \widehat{\mathbb{E}}[T(f(X))] - (\widehat{\mathbb{E}}'[U(f(X))] + \widehat{\mathbb{E}}[U(f(X))])] \\
&+ \frac{1}{2}\mathbb{E}[\sup_{f\in\mathcal{F}_{p_n}\cup\mathcal{F}_{p'_n}} \widehat{\mathbb{E}}[T(f(X))] - \widehat{\mathbb{E}}'[T(f(X))] - (\widehat{\mathbb{E}}'[U(f(X))] + \widehat{\mathbb{E}}[U(f(X))])] \\
&= \frac{1}{2n}\mathbb{E}[\sup_{f,g\in\mathcal{F}_{p_n}\cup\mathcal{F}_{p'_n}} \sum_{i=1}^{n} T(f)(X'_i) - T(f)(X_i) - U(f)(X_i) - U(f)(X'_i) \\
&+ T(g)(X_i) - T(g)(X'_i) - U(g)(X_i) - U(g)(X'_i)] \\
&\leq \frac{1}{2n}\mathbb{E}[\sup_{f_1,g_1\in\mathcal{F}_{p_n}\cup\mathcal{F}_{p'_n}} \sum_{i=1}^{n} T(g_1)(X_i) - T(f_1)(X_i) - U(f_1)(X_i) - U(g_1)(X_i)] \\
&+ \frac{1}{2n}\mathbb{E}[\sup_{f_2,g_2\in\mathcal{F}_{p_n}\cup\mathcal{F}_{p'_n}} \sum_{i=1}^{n} T(f_2)(X'_i) - T(g_2)(X'_i) - U(f_2)(X'_i) - U(g_2)(X'_i)] \\
&\overset{(b)}{=} \frac{1}{n}\mathbb{E}[\sup_{f,g\in\mathcal{F}_{p_n}\cup\mathcal{F}_{p'_n}} \sum_{i=1}^{n} T(g)(X_i) - T(f)(X_i) - U(f)(X_i) - U(g)(X_i)] \\
&\overset{(c)}{\leq} \frac{2}{n}\mathbb{E}[\sup_{f\in\mathcal{F}_{p_n}\cup\mathcal{F}_{p'_n}} \sum_{i=1}^{n} \epsilon_i T(f)(X_i) - U(f)(X_i)] \qquad (23)
\end{aligned}
$$

where (a), (b) are symmetry and (c) is Jensen's inequality. □

As $f \in \mathcal{F}_{p_n}$, applying the last lemma to previous display above, with the choice for the first expectation (21)

$$Tf(x) = -\left[2f^*(x)(f^*(x) - f(x)) - 2x(f^*(x-1) - f(x-1))\right], \quad Uf(x) = \frac{1}{4}(f^*(x) - f(x))^2,$$

and the choice for the second expectation (22) $Tf(x) = \frac{3}{2}(f^*(x) - f(x))^2, Uf(x) = \frac{1}{2}(f^*(x) - f(x))^2$, we get the desired result. □

2.3 Controlling the Rademacher complexities

To prove Theorem 1, we apply Theorem 3 with the function class $\mathcal{F}_{p_n} = \mathcal{F}_*$ defined in (13). Denote by $\mathcal{F}_{p'_n} = \mathcal{F}'_*$ its independent copy based on a fresh sample $X'_1, \ldots, X'_n$. Let us define the following

generalization of (15) and (16): For $b > 1$,

$$T_1(b,n) = \mathbb{E}\left[\sup_{f\in\mathcal{F}_*\cup\mathcal{F}'_*}\sum_{i=1}^{n}(\epsilon_i - \frac{1}{b})(f(X_i)-f^*(X_i))^2\right], \tag{24}$$

$$T_2(b,n) = \mathbb{E}\Bigg[\sup_{f\in\mathcal{F}_*\cup\mathcal{F}'_*}\sum_{i=1}^{n} 2\epsilon_i(f^*(X_i)(f^*(X_i)-f(X_i)) - X_i(f^*(X_i-1)-f(X_i-1))) - \frac{1}{b}(f^*(X_i)-f(X_i))^2\Bigg]. \tag{25}$$

Then we have the following bound on the complexities.

Lemma 4. *Let $\pi \in \mathcal{P}[0,h]$ with h being either a constant or $h = s\log n$ for some $s > 0$. Let $M := M(n,h) > h$ be such that*

- $\sup_{\pi\in\mathcal{P}([0,h])}\mathbb{P}_{X\sim p_\pi}[X > M] \le \frac{1}{n^7}$.
- *For $X_i \overset{iid}{\sim} p_\pi$, $\mathbb{E}\left[X_{\max}^k\right] \le c(k)M^k$ for $k = 1,\dots,4$ and absolute constant $c > 0$.*

Then there exists a constant $c_0(b) > 0$ such that

$$T_1(b,n), T_2(b,n) \le c_0(b)\left(\max\{1,h^2\}M + \max\{1,h\}M^2\right). \tag{26}$$

The first condition on the probability is an artifact of the proof. In general, any tail bounds on the random variable X that decay polynomially in n, such as the ones satisfied by bounded priors or priors with subexponential tails, are good enough for our proofs to go through.

Proof of Lemma 4. We consider the following notations.

$$N(x) = \sum_{i=1}^{n}\mathbf{1}_{\{X_i=x\}} \qquad \epsilon(x) = \sum_{i=1}^{n}\epsilon_i\mathbf{1}_{\{X_i=x\}} \tag{27}$$

where $\epsilon_1,\cdots,\epsilon_n$ are independent Rademacher symbols.

Bound on $T_2(b,n)$: Using $f(-1) = 0$ we note that

$$\begin{aligned}
&\sum_{i=1}^{n} 2\epsilon_i(f^*(X_i)(f^*(X_i)-f(X_i)) - X_i(f^*(X_i-1)-f(X_i-1))) - \frac{1}{b}(f^*(X_i)-f(X_i))^2\\
=&\sum_{x\ge 0} 2\epsilon(x)(f^*(x)(f^*(x)-f(x)) - x(f^*(x-1)-f(x-1))) - \frac{N(x)}{b}(f^*(x)-f(x))^2\\
=&\sum_{x\ge 0} 2(\epsilon(x)f^*(x) - (x+1)\epsilon(x+1))(f^*(x)-f(x)) - \frac{N(x)}{b}(f^*(x)-f(x))^2
\end{aligned} \tag{28}$$

In view of the above, we can bound $T_2(b,n)$ using the sum of the following two terms

$$t_1(n) \triangleq \mathbb{E}\{\sup_{f\in\mathcal{F}_*\cup\mathcal{F}'_*}[\sum_{x\ge 0} 2(\epsilon(x)f^*(x)-(x+1)\epsilon(x+1))(f^*(x)-f(x)) - \frac{N(x)}{b}(f^*(x)-f(x))^2]\mathbf{1}_{\{N(x)>0\}}\}$$

$$t_0(n) \triangleq \mathbb{E}\{\sup_{f\in\mathcal{F}_*\cup\mathcal{F}'_*}[\sum_{x\ge 0} -2(x+1)\epsilon(x+1)(f^*(x)-f(x))]\mathbf{1}_{\{N(x)=0\}}\}.$$

For analyzing the term $t_1(n)$, since $N(x) > 0$, using $2ax - bx^2 \le \frac{a^2}{b}$ for any a, x and $b > 0$ we get

$$t_1(n) \le b \cdot \mathbb{E}\left[\sum_{x\ge 0} \frac{(\epsilon(x)f^*(x) - (x+1)\epsilon(x+1))^2}{N(x)} \mathbf{1}_{\{N(x)>0\}}\right] \tag{29}$$

Using $\mathbb{E}\left\{\epsilon(x)|X_1,\dots,X_n\right\} = 0$ and $\mathbb{E}\left[(\epsilon(x))^2|X_1,\dots,X_n\right] = N(x)$ we get

$$\mathbb{E}\left[\frac{(f^*(x)\epsilon(x) - (x+1)\epsilon(x+1))^2}{N(x)} \mathbf{1}_{\{N(x)>0\}}\right] = \mathbb{E}\left[\left((f^*(x))^2 + \frac{(x+1)^2 N(x+1)}{N(x)}\right) \mathbf{1}_{\{N(x)>0\}}\right].$$

Using the results that

(P1) $N(x) \sim \text{Binom}(n, p_\pi(x))$ and for absolute constant $c' > 0$ [PW21, Lemma 16]

$$\mathbb{E}\left[\frac{\mathbf{1}_{\{N(x)>0\}}}{N(x)}\right] \le c' \min\left\{np_\pi(x), \frac{1}{np_\pi(x)}\right\},$$

(P2) conditioned on $N(x), N(x+1) \sim \text{Binom}(n - N(x), \frac{p_\pi(x+1)}{1-p_\pi(x)})$,

(P3) $f^*(x) = (x+1)\frac{p_\pi(x+1)}{p_\pi(x)} = \mathbb{E}\left[\theta|X = x\right] \le h$ for all $x \ge 0$,

(P4) Since for every $x > 0$, $\frac{x^y e^{-x}}{y!} \le \frac{y^y e^{-y}}{y!} \le \frac{1}{\sqrt{2\pi y}}$ (Stirling's), we have

$$p_\pi(y) < \frac{1}{\sqrt{2\pi y}}, \quad y \ge 1, \tag{30}$$

we continue (29) to get

$$\begin{aligned}
\frac{1}{b}t_1(n) &\le \mathbb{E}\left[\sum_{x\ge 0} f^*(x)^2 \mathbf{1}_{\{N(x)>0\}}\right] + \sum_{x\ge 0}(x+1)^2 \frac{np_\pi(x+1)}{1-p_\pi(x)} \mathbb{E}\left[\frac{\mathbf{1}_{\{N(x)>0\}}}{N(x)}\right] \\
&\le h^2 \mathbb{E}\left[1 + X_{\max}\right] + \frac{np_\pi(1)}{1-p_\pi(0)} \mathbb{E}\left[\frac{\mathbf{1}_{\{N(0)>0\}}}{N(0)}\right] + n\sum_{x\ge 1}(x+1)^2 p_\pi(x+1) \mathbb{E}\left[\frac{\mathbf{1}_{\{N(x)>0\}}}{N(x)}\right] \\
&\le h^2 \mathbb{E}\left[1 + X_{\max}\right] + \frac{c' p_\pi(1)}{(1-p_\pi(0))p_\pi(0)} + c' h \sum_{x\ge 1}(x+1) \min\left\{(np_\pi(x))^2, 1\right\}.
\end{aligned}$$

Let $M > h$ be as in the lemma statement. For the second term notice that $\frac{p_\pi(1)}{(1-p_\pi(0))p_\pi(0)} \le \max\left\{1, h\right\}$. For the third term, we use the bound

$$\begin{aligned}
h\sum_{x\ge 1}(x+1) \min\left\{(np_\pi(x))^2, 1\right\} &\le hM^2 + h\sum_{x\ge M}(x+1) \min\left\{(np_\pi(x))^2, 1\right\} \\
&\le hM^2 + 2n^2 h \sum_{x\ge M} x(p_\pi(x))^2 \overset{(a)}{\le} hM^2 + 2n^2 h^2 \mathbb{P}_{X\sim p_\pi}[X > M] \le 2(hM^2 + \frac{2h^2}{n^5}).
\end{aligned} \tag{31}$$

where (a) is due to that $xp_\pi(x) = f^*(x-1)p_\pi(x-1) \le h$ for all $x \ge 1$. We finally note that since h is either constant or in the form $O(s \log n)$ for some constant s, the term $\frac{h^2}{n^5}$ can be neglected.

Next, we evaluate $t_0(n)$. As $|\epsilon(x+1)| \le N(x+1)$ and $N(x+1) = 0$ for $x \ge X_{\max}$ we get

$$t_0(n) \le \mathbb{E}\left[\sum_{x\ge 0} 2(x+1)N(x+1) \sup_{f\in\mathcal{F}_*\cup\mathcal{F}'_*} |f^*(x) - f(x)|\, \mathbf{1}_{\{N(x)=0\}}\right]$$
$$\le \mathbb{E}\left[\sum_{x=0}^{X_{\max}-1} 2(x+1)\left(f^*(x) + X_{\max} + X'_{\max}\right) N(x+1)\mathbf{1}_{\{N(x)=0\}}\right]. \quad (32)$$

Let $M > 0$ be as in the lemma statement and $A = \{X_{\max} \le M, X'_{\max} \le M\}$. Then $\mathbb{P}[A^c] \le \frac{2}{n^6}$ via the union bound argument. Thus we have, for some absolute constant $c > 0$,

$$\mathbb{E}\left[\sum_{x=0}^{X_{\max}-1} 2(x+1)\left(f^*(x) + X_{\max} + X'_{\max}\right) N(x+1)\mathbf{1}_{\{N(x)=0\}} \cdot \mathbf{1}_{\{A^c\}}\right]$$
$$\le \mathbb{E}\left[X_{\max}(h + X_{\max} + X'_{\max}) \sum_{x=0}^{X_{\max}-1} N(x+1)\mathbf{1}_{\{N(x)=0\}}\mathbf{1}_{\{A^c\}}\right]$$
$$\overset{(a)}{\le} n\mathbb{E}\left[(X_{\max})\cdot\left(h + X_{\max} + X'_{\max}\right)\mathbf{1}_{\{A^c\}}\right] \overset{(b)}{\le} n\sqrt{\mathbb{E}\left[\left(h + X_{\max} + X'_{\max}\right)^4\right]}\sqrt{\mathbb{P}[A^c]} \le \frac{cM^2}{n^2}. \quad (33)$$

with (a) due to that $\sum_{x=0}^{X_{\max}-1} N(x+1) \le \sum_{x=0}^{\infty} N(x) = n$, and (b) the Cauchy-Schwarz inequality and $\mathbb{E}\left[X^4_{\max}\right] \lesssim 2M^4$.

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

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

where (a) uses $f^*(x) \le h$ for all x, and also $nw(1-w)^{n-1} \le (1 - \frac{1}{n})^{n-1} < 1$ for all $w \in [0, 1]$. We conclude our proof by combining the above with (33).

Bound on $T_1(b, n)$: Denote $m_b = b + 1$. Conditional on the sample $X_1, \ldots, X_n$ and $X_1, \ldots, X_n$, given any $f \in \mathcal{F}_* \cup \mathcal{F}'_*$ define

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

Then using the above definition we get for each $f \in \mathcal{F}_* \cup \mathcal{F}'_*$, conditional on the samples,

$$\sum_{i=1}^{n}(\epsilon_i - \frac{1}{b})(f(X_i) - f^*(X_i))^2 = \sum_{x:N(x)>0}(\epsilon(x) - \frac{1}{b}N(x))(f(x) - f^*(x))^2$$

$$= \left(\sum_{x=0}^{v(f)} + \sum_{x=v(f)+1}^{X_{\max}}\right)(\epsilon(x) - \frac{1}{b}N(x))(f(x) - f^*(x))^2$$

$$\leq m_b^2 h^2 \sum_{x=0}^{X_{\max}} \max\left\{\epsilon(x) - \frac{1}{b}N(x), 0\right\} \tag{34}$$

$$+ \sup_{v\geq 0}\left\{\sup_{m_b h \leq f \leq X_{\max}}\left\{\sum_{x>v}^{X_{\max}}(\epsilon(x) - \frac{1}{b}N(x))(f(x) - f^*(x))^2\right\}\right\}. \tag{35}$$

For the first term (34), we invoke the following lemma, to be proven in Appendix B.

Lemma 5. *For each x and $b > 1$, conditioned on X_1^n we have*

$$\mathbb{E}[\max\{\epsilon(x) - \frac{1}{b}N(x), 0\}] \leq \frac{1 - \frac{1}{b}}{e \cdot D(\frac{1+\frac{1}{b}}{2}||\frac{1}{2})}$$

For brevity, we denote $N_b \triangleq \frac{1-\frac{1}{b}}{e \cdot D(\frac{1+\frac{1}{b}}{2}||\frac{1}{2})}$. This gives us

$$\mathbb{E}\left[m_b^2 h^2 \sum_{x=0}^{X_{\max}} \max\left\{\epsilon(x) - \frac{1}{b}N(x), 0\right\}\middle| X_1^n\right] \leq N_b m_b^2 h^2 \mathbb{E}[(1 + X_{\max})]. \tag{36}$$

For the second term (35), we note that for any f with values in $[m_b h, X_{\max}]$, we have $\frac{m_b - 1}{m_b} f \leq f - f^* \leq f$ and hence

$$(\epsilon(x) - \frac{1}{b}N(x))(f(x) - f^*(x))^2 \leq \max\left\{\left(\epsilon(x) - \frac{1}{b}N(x)\right), \left(\frac{m_b - 1}{m_b}\right)^2\left(\epsilon(x) - \frac{1}{b}N(x)\right)\right\} f(x)^2. \tag{37}$$

Now given that $-N(x) \leq \epsilon(x) \leq N(x)$, define function $g : [-1, 1] \to \mathbb{R}$ given by

$$g(x) = \max\left(\left(x - \frac{1}{b}\right), \left(\frac{m_b - 1}{m_b}\right)^2\left(x - \frac{1}{b}\right)\right) \tag{38}$$

Since g is the maximum of two linear functions, it is convex, and therefore bounded by the line joining their endpoints, $(-1, -(\frac{1}{b} + 1) \cdot \left(\frac{m_b - 1}{m_b}\right)^2)$ and $(1, 1 - \frac{1}{b})$. Now define:

$$\alpha = \frac{1}{2}\left[\left(1 + \frac{1}{b}\right) \cdot \left(\frac{m_b - 1}{m_b}\right)^2 + \left(1 - \frac{1}{b}\right)\right]; \quad \beta = \frac{1}{2}\left[\left(1 + \frac{1}{b}\right) \cdot \left(\frac{m_b - 1}{m_b}\right)^2 - \left(1 - \frac{1}{b}\right)\right] = \frac{1}{2b(b+1)} \tag{39}$$

using the fact that $m_b = b+1$. Note that $0 < \beta < \alpha$. Then we have $g(x) \leq \alpha x - \beta$ for all $x \in [-1, 1]$. Hence, we have

$$(\epsilon(x) - \frac{1}{b}N(x))(f(x) - f^*(x))^2 \leq (\alpha\epsilon(x) - \beta N(x))f(x)^2 \tag{40}$$

Hence (35) can be bounded by, modulo a constant multiplicative factor $c_2(b)$ depending on b,

$$\sup_{v\geq 0}\left\{\sup_{m_b h\leq f\leq X_{\max}}\left\{\sum_{x>v}^{X_{\max}}(\epsilon(x)-\frac{1}{b}N(x))(f(x)-f^*(x))^2\right\}\right\}$$
$$\leq c_2(b)\left[\sup_{v\geq 0}\left\{\sup_{m_b h\leq f\leq X_{\max}}\left\{\sum_{x>v}^{X_{\max}}(\epsilon(x)-\frac{\beta}{\alpha}N(x))f(x)^2\right\}\right\}\right]. \tag{41}$$

Note that the above f-based maximization problem is a linear programming of the form

$$\sup_{a_1,\ldots,a_k}\sum_{i=1}^{k}v_i a_i,\quad (m_b h)^2\leq a_1\cdots\leq a_k\leq (X_{\max})^2,$$

with $k=X_{\max}+1$. The optimization happens on the corner points of the above convex set, that are given by $X_{\max}+1$ length vectors of the form

$$\left\{(m_b h)^2,\ldots,(m_b h)^2,(X_{\max})^2,\ldots,(X_{\max})^2\right\}.$$

This implies we can bound (41) by

$$(m_b h)^2\sum_{x=0}^{X_{\max}}\max\left\{\epsilon(x)-\frac{\beta}{\alpha}N(x),0\right\}+(X_{\max})^2\sup_{v\geq 0}\left\{\sum_{x>v}^{X_{\max}}(\epsilon(x)-\frac{\beta}{\alpha}N(x))\right\}. \tag{42}$$

The bound of the first term, conditional on the data, is given as per Lemma 5 as $m_b^2h^2N_b(1+X_{\max})$. For the second term, we first note the following result.

Lemma 6. *Let $c>0$ be given. For $\epsilon=(\epsilon_1,\cdots,\epsilon_n)$ n independent Rademacher symbols, denote*

$$L_c(\epsilon)=\max_{0\leq j\leq n}\left\{\sum_{i=1}^{j}\epsilon_i-cj\right\} \tag{43}$$

Then $\mathbb{E}[L_c(\epsilon)]\leq M_c$ where $M_c\triangleq 1+(1-\exp(-D(\frac{c+1}{2}||\frac{1}{2})))^{-2}$.

The proof of the above result is provided in Appendix B.

Therefore, using Lemma 6, we have

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

for some constant $c(b)>0$ via Lemma 6. Thus we get

$$\mathbb{E}\left[(X_{\max})^2\sup_{v\geq 0}\left\{\sum_{x>v}^{X_{\max}}(\epsilon(x)-\frac{\beta}{\alpha}N(x))\right\}\middle|X_1,\ldots,X_n\right]\leq c(b)(1+X_{\max})^2. \tag{44}$$

Combining (41), (42), and (44) we get

$$\mathbb{E}\left[\sup_{v\geq 0}\left\{\sup_{m_b h\leq f\leq X_{\max}}\left\{\sum_{x>v}^{X_{\max}}(\epsilon(x)-\frac{1}{b}N(x))(f(x)-f^*(x))^2\right\}\right\}\middle|X_1^n\right]\leq c_3(b)\left(h^2(1+X_{\max})+(1+X_{\max})^2\right) \tag{45}$$

for a constant $c_3(b)$ depending on b. Then taking expectation on both the sides and using the definition of M in the lemma statement we finish the proof.

□

2.4 Proof of Regret optimality (Theorem 1)

We use the above result to first prove the regret bound for bounded priors in $\mathcal{P}([0,h])$. Note that by Lemma 10 and Lemma 12, there are constants $c_1, c_2 > 0$ such that for any fixed $h > 0$ such that $M = \max\{c_2, c_1 h\} \cdot \frac{\log n}{\log\log n}$ satisfies both conditions in Lemma 4, and we get $O(\frac{\max\{1,h^3\}}{n}(\frac{\log n}{\log\log n})^2)$ bound on the regret, which is optimal up to constants that possibly depend on h.

Next we extend the above proof to the subexponential case. Given $\pi \in \mathsf{SubE}(s)$ define the truncated version $\pi_{c,n}[\theta \in \cdot] = \pi[\theta \in \cdot \mid \theta \le c\log n]$ for $c > 0$. Then we have the following reduction.

Lemma 7. *There exists constants $c_1, c_2, c_3 > 0$ such that*

$$\mathsf{Regret}_\pi(\widehat{f}_{\mathsf{erm}}) \le \mathsf{Regret}_{\pi_{c_1 s,n}}(\widehat{f}_{\mathsf{erm}}) + \frac{\max\{c_2, c_3 s\}}{n}.$$

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

$$\varepsilon = \mathbb{P}[\theta > c(s)\log n] \le \frac{1}{n^{10}}, \quad \theta \sim \pi \tag{46}$$

Denote, also, the event $E = \{\theta_i \le c(s)\log n, \forall i = 1, \cdots, n\}$; we have $\mathbb{P}[E^c] \le n^{-9}$. Let $\pi_{c(s),n}$ as the truncated prior $\pi_{c(s),n}[\theta \in \cdot] = \pi[\theta \in \cdot \mid \theta \le c(s)\log n]$. Define $\mathrm{mmse}(\pi) \triangleq \min_f \mathbb{E}_{\theta\sim\pi}[(f(X)-\theta)^2]$ (i.e. the error by the Bayes estimator). Then we may use [PW21, Equation 131] to obtain

$$\mathsf{Regret}_\pi(\widehat{f}_{\mathsf{erm}}) \le \mathsf{Regret}_{\pi_{c,n}}(\widehat{f}_{\mathsf{erm}}) + \mathrm{mmse}(\pi_{c,n}) - \mathrm{mmse}(\pi) + \mathbb{E}_\pi[(\widehat{f}_{\mathsf{erm}}(X)-\theta)^2 \mathbf{1}_{\{E^c\}}] \tag{47}$$

By [WV12, Lemma 2], $\mathrm{mmse}(\pi_{c,n}) - \mathrm{mmse}(\pi) \le \frac{\varepsilon}{1-\varepsilon}\mathrm{mmse}(\pi) \le 2\varepsilon$ whenever $\varepsilon \le \frac{1}{2}$. In addition, Lemma 2 entails that $\widehat{f}_{\mathsf{erm}}(X) \le X_{\max}$, which means that $\mathbb{E}[\widehat{f}^4_{\mathsf{erm}}(X)] \le \mathbb{E}[X^4_{\max}] \le O(\max\{1,s^4\}(\log n)^4)$ as per Lemma 13. Meanwhile, for all $\pi \in \mathsf{SubE}(s)$ we have $\mathbb{E}_\pi[\theta^4] \in O(s^4)$. This means $\mathbb{E}_\pi[(\widehat{f}_{\mathsf{erm}}-\theta)^4] \lesssim_s (\log n)^4$. Thus by Cauchy-Schwarz inequality

$$\mathbb{E}_\pi[(\widehat{f}_{\mathsf{erm}}(X)-\theta)^2\mathbf{1}_{\{E^c\}}] \le \sqrt{\mathbb{P}[E^c]\mathbb{E}_\pi[(\widehat{f}_{\mathsf{erm}}(X)-\theta)^4]} \le \sqrt{n^{-9}\mathbb{E}_\pi[(\widehat{f}_{\mathsf{erm}}(X)-\theta)^4]} \lesssim \frac{\max\{1,s^2\}}{n}.$$

□

Given this lemma, it suffices to bound $\mathsf{Regret}_{\pi_{c,n}}(\widehat{f}_{\mathsf{erm}})$. Then by Lemma 11 and Lemma 12 there exist constants $c_1, c_2 > 0$ such that $M = \max\{c_1, c_2 s\}\log n$ satisfies both the requirements in Lemma 4. Hence we get the desired regret bound of $O(\frac{\max\{1,s^3\}(\log n)^3}{n})$.

3 Regret bounds in multiple dimensions

To prove the regret bound for the multidimensional estimator $\widehat{\boldsymbol{f}} = (\widehat{f}_1, \ldots, \widehat{f}_d)$ we use the approximation error for the different coordinates. In particular, similar to (17) we define

$$\mathsf{R}(\boldsymbol{f}) \triangleq \mathbb{E}\left[\|\boldsymbol{f}(\boldsymbol{X})\|^2 - 2\sum_{i=1}^d X_i f_i(\boldsymbol{X}-\boldsymbol{e}_i)\right], \quad \widehat{\mathsf{R}}(\boldsymbol{f}) \triangleq \widehat{\mathbb{E}}\left[\|\boldsymbol{f}(\boldsymbol{X})\|^2 - 2\sum_{i=1}^d X_i f_i(\boldsymbol{X}-\boldsymbol{e}_i)\right] \tag{48}$$

and note that

$$\mathsf{Regret}_\pi(\widehat{\boldsymbol{f}}_{\mathsf{erm}}) = \mathbb{E}\left[\mathsf{R}(\widehat{\boldsymbol{f}}_{\mathsf{erm}}) - \mathsf{R}(\boldsymbol{f}^*)\right] \tag{49}$$

As mentioned before, in the multidimensional setup our estimator is produced by optimizing over the class of coordinate-wise monotone functions $\boldsymbol{\mathcal{F}}$ in (10) and $\boldsymbol{f}^* \in \boldsymbol{\mathcal{F}}$ as well. Using the quadratic structure of the regret and the convexity of $\boldsymbol{\mathcal{F}}$, we can mimic the proof of (18) to get

$$\widehat{\mathsf{R}}(\boldsymbol{f}) - \widehat{\mathsf{R}}(\widehat{\boldsymbol{f}}) \geq \widehat{\mathbb{E}}\left[\|\boldsymbol{f} - \widehat{\boldsymbol{f}}\|^2\right], \quad \boldsymbol{f} \in \boldsymbol{\mathcal{F}}. \tag{50}$$

Then following a similar argument as in (21), (22), using (49) we have

$$\begin{aligned}
\mathsf{Regret}_\pi(\widehat{\boldsymbol{f}}_{\mathsf{erm}}) &\leq \mathbb{E}\left[\mathsf{R}(\widehat{\boldsymbol{f}}) - \mathsf{R}(\boldsymbol{f}^*) + \widehat{\mathsf{R}}(\boldsymbol{f}^*) - \widehat{\mathsf{R}}(\widehat{\boldsymbol{f}}) - \widehat{\mathbb{E}}\left\|\boldsymbol{f}^* - \widehat{\boldsymbol{f}}\right\|^2\right] \\
&= \mathbb{E}\Bigg[\widehat{\mathbb{E}}\Bigg[\sum_{j=1}^d 2f_j^*(\boldsymbol{X})(f_j^*(\boldsymbol{X}) - \widehat{f}_j(\boldsymbol{X})) - 2\boldsymbol{X}_j(f_j^*(\boldsymbol{X} - \boldsymbol{e}_j) - \widehat{\boldsymbol{f}}_j(\boldsymbol{X} - \boldsymbol{e}_j))\Bigg] \\
&\quad - \mathbb{E}\Bigg[\sum_{j=1}^d 2f_j^*(\boldsymbol{X})(f_j^*(\boldsymbol{X}) - \widehat{\boldsymbol{f}}_j(\boldsymbol{X})) - 2X_j(\boldsymbol{f}_j^*(\boldsymbol{X} - \boldsymbol{e}_j) - \widehat{\boldsymbol{f}}_j(\boldsymbol{X} - \boldsymbol{e}_j))\Bigg] \\
&\quad - \frac{1}{4}(\widehat{\mathbb{E}}\left[\|\boldsymbol{f}^* - \widehat{\boldsymbol{f}}\|^2\right] + \mathbb{E}[\|\boldsymbol{f}^*(\boldsymbol{X}) - \widehat{\boldsymbol{f}}(\boldsymbol{X})\|^2])\Bigg] \qquad (51) \\
&\quad + \mathbb{E}\left[\frac{5}{4}\mathbb{E}[\|\boldsymbol{f}^*(\boldsymbol{X}) - \widehat{\boldsymbol{f}}(\boldsymbol{X})\|^2] - \frac{7}{4}\widehat{\mathbb{E}}[\|\boldsymbol{f}^*(\boldsymbol{X}) - \widehat{\boldsymbol{f}}(\boldsymbol{X})\|^2]\right]. \qquad (52)
\end{aligned}$$

As Lemma 3 is still directly applicable in the multidimensional setting, applying it with

$$T(\boldsymbol{f}(\boldsymbol{x})) = -\sum_{j=1}^d [2\boldsymbol{f}_j^*(\boldsymbol{x})(f_j^*(\boldsymbol{x}) - f_j(\boldsymbol{x})) - 2x_j(f_j^*(\boldsymbol{x} - \boldsymbol{e}_j) - f_j(\boldsymbol{x} - \boldsymbol{e}_j))], \quad U(\boldsymbol{f}(\boldsymbol{x})) = \frac{1}{4}\|\boldsymbol{f}^*(x) - \boldsymbol{f}(x)\|^2$$

to bound (51) and with $T(\boldsymbol{f}(\boldsymbol{x})) = \frac{3}{2}\|\boldsymbol{f}^*(x) - \boldsymbol{f}(x)\|^2$, $U(\boldsymbol{f}(\boldsymbol{x})) = \frac{1}{2}\|\boldsymbol{f}^*(x) - \boldsymbol{f}(x)\|^2$ to bound (52) we get: for any function class $\boldsymbol{\mathcal{F}}_{p_n}$ depending on the empirical distribution p_n of the sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ that includes $\widehat{\boldsymbol{f}}_{\mathsf{erm}}$ and $\boldsymbol{f}^*$ and its independent copy $\boldsymbol{\mathcal{F}}_{p'_n}$ based on an independent sample $\boldsymbol{X}'_1, \ldots, \boldsymbol{X}_n$

$$\begin{aligned}
\mathsf{Regret}_\pi(\widehat{\boldsymbol{f}}_{\mathsf{erm}}) \leq\ & \frac{3}{n}\mathbb{E}\left[\sup_{\boldsymbol{f} \in \boldsymbol{\mathcal{F}}_{p_n} \cup \boldsymbol{\mathcal{F}}_{p'_n}} \sum_{i=1}^n (\epsilon_i - \frac{1}{6})(f_j(\boldsymbol{X}_i) - f_j^*(\boldsymbol{X}_i))^2\right] \\
&+ \frac{2}{n}\mathbb{E}\Bigg[\sup_{\boldsymbol{f} \in \boldsymbol{\mathcal{F}}_{p_n} \cup \boldsymbol{\mathcal{F}}_{p'_n}} \sum_{i=1}^n 2\epsilon_i(f_j^*(\boldsymbol{X}_i)(f_j^*(\boldsymbol{X}_i) - f_j(\boldsymbol{X}_i)) - X_{ij}(f_j^*(\boldsymbol{X}_i - \boldsymbol{e}_j) \\
&\qquad - f_j(\boldsymbol{X}_i - \boldsymbol{e}_j))) - \frac{1}{4}(f_j^*(\boldsymbol{X}_i) - f_j(\boldsymbol{X}_i))^2\Bigg]
\end{aligned} \tag{53}$$

To achieve the best possible bound we choose $\boldsymbol{\mathcal{F}}_{p_n}$ with low complexity. Note that the objective function R defined in (48) is separable into sum of individual loss functions. Thus, given the definition of $\boldsymbol{\mathcal{F}}$ in (10), for each coordinate j and each class $C_j(\boldsymbol{x}')$ defined in (11), we have

$$(\widehat{f}_{\mathsf{erm}})_j|_{C_j(\boldsymbol{x}')} = \operatorname*{argmin}_{f \in \mathcal{F}_1} \widehat{\mathbb{E}}\left[f_j(\boldsymbol{X}) - 2X_j f_j(\boldsymbol{X} - \boldsymbol{e}_j) | \boldsymbol{X} \in C_j(\boldsymbol{x}')\right], \qquad \forall \boldsymbol{x}' \in \mathbb{R}_+^{d-1}.$$

where $\mathcal{F}_1$ is the class of all one-dimensional monotone function from $\mathbb{Z}_+ \to \mathbb{R}_+$. Considering this for all classes $C_j(\boldsymbol{x}')$ and from Lemma 2, we have

$$(\widehat{f}_{\mathsf{erm}})_j(\boldsymbol{X}_i) \leq X_{j,\max}, \ X_{j,\max} \triangleq \max_{i=1}^n X_{ij}, \quad j = 1, \ldots, d. \tag{54}$$

Given the sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ define the sample based function class

$$\boldsymbol{\mathcal{F}}_* \triangleq \left\{ \boldsymbol{f} \in \boldsymbol{\mathcal{F}} : f_j(\boldsymbol{X}_i) \le \max\left\{ f_j^*(\boldsymbol{X}_i), X_{j,\max} \right\}, \quad j = 1, \ldots, d, i = 1, \ldots, n \right\}. \tag{55}$$

Let $\boldsymbol{\mathcal{F}}'_*$ be an independent copy of $\boldsymbol{\mathcal{F}}_*$. Then simplifying (53) with $\boldsymbol{\mathcal{F}}_{p_n} = \boldsymbol{\mathcal{F}}_*, \boldsymbol{\mathcal{F}}_{p_n} = \boldsymbol{\mathcal{F}}'_*$ we get

$$\begin{aligned}
\mathsf{Regret}_\pi(\widehat{\boldsymbol{f}}_{\mathsf{erm}}) &\le \frac{1}{n} \sum_{j=1}^{d} \left(3U_1(j,n) + 4U_2(j,n)\right) \\
U_1(j,n) &\triangleq \mathbb{E}\left[\sup_{\boldsymbol{f} \in \boldsymbol{\mathcal{F}}_* \cup \boldsymbol{\mathcal{F}}'_*} \sum_{i=1}^{n} (\epsilon_i - \frac{1}{6})(f_j(\boldsymbol{X}_i) - f_j^*(\boldsymbol{X}_i))^2 \right] \\
U_2(j,n) &\triangleq \mathbb{E}\left[\sup_{\boldsymbol{f} \in \boldsymbol{\mathcal{F}}_* \cup \boldsymbol{\mathcal{F}}'_*} \sum_{i=1}^{n} \epsilon_i (f_j^*(\boldsymbol{X}_i)(f_j^*(\boldsymbol{X}_i) - f_j(\boldsymbol{X}_i)) - X_{ij}(f_j^*(\boldsymbol{X}_i - \boldsymbol{e}_j) \right. \\
&\qquad \left. - f_j(\boldsymbol{X}_i - \boldsymbol{e}_j))) - \frac{1}{8}(f_j^*(\boldsymbol{X}_i) - f_j(\boldsymbol{X}_i))^2 \right].
\end{aligned} \tag{56}$$

We bound these $2d$ Rademacher complexities to arrive at the results. Note that as we want to analyze the supremum over all possible prior distributions whose marginals are subject to the same tail assumption (either supported on $[0, h]$ or s-subexponential), by the inherent symmetry on the d coordinates, it suffices to consider only a single coordinate, say, the j-th, when bounding the offset Rademacher complexity. The final regret bound then includes an extra factor of d over this single instance of Rademacher complexity. Note that in our problem the function class $\boldsymbol{\mathcal{F}}_*$ is supported over the hypercube $\prod_{j=1}^{d}[0, X_{j,\max}]$. The high-level idea for our analysis is that the effective size of this hypercube, corresponding to different classes of priors, controls the Rademacher complexity and hence the regret upper bound.

3.1 Bounding Rademacher Complexity for Bounded Prior

Here we first prove a bound for the generalization of the Rademacher complexities in (56) for $b > 1$:

$$\begin{aligned}
U_1(b,j,n) &\triangleq \mathbb{E}\left[\sup_{\boldsymbol{f} \in \boldsymbol{\mathcal{F}}_* \cup \boldsymbol{\mathcal{F}}'_*} \sum_{i=1}^{n} (\epsilon_i - \frac{1}{b})(f_j(\boldsymbol{X}_i) - f_j^*(\boldsymbol{X}_i))^2 \right] \\
U_2(b,j,n) &\triangleq \mathbb{E}\left[\sup_{\boldsymbol{f} \in \boldsymbol{\mathcal{F}}_* \cup \boldsymbol{\mathcal{F}}'_*} \sum_{i=1}^{n} 2\epsilon_i (f_j^*(\boldsymbol{X}_i)(f_j^*(\boldsymbol{X}_i) - f_j(\boldsymbol{X}_i)) - X_{ij}(f_j^*(\boldsymbol{X}_i - \boldsymbol{e}_j) \right. \\
&\qquad \left. - f_j(\boldsymbol{X}_i - \boldsymbol{e}_j))) - \frac{1}{b}(f_j^*(\boldsymbol{X}_i) - f_j(\boldsymbol{X}_i))^2 \right]
\end{aligned} \tag{57}$$

We have the following result similar to Lemma 4.

Lemma 8. *Let $\pi \in \mathcal{P}[0, h]$ with h being either a constant or $h = s \log n$ for some $s > 0$. Given $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be iid observations from p_π, let $M := M(n, h) > h$ be such that*

- *For each coordinate $j = 1, \cdots, d$, we have the j-th coordinate X_j of $\boldsymbol{X}$ satisfying*

$$\sup_{\pi \in \mathcal{P}([0,h])^d} \mathbb{P}_{\boldsymbol{X} \sim p_\pi}\left[X_j > M \right] \le \frac{1}{n^7}.$$

- *For $\beta = 1, 2, 3, 4$, constants $c_1(\beta)$ depending on β and absolute constant $c > 0$*

$$\mathbb{E}\left[(X_{j,\max})^4\right] \le cM^4, \quad \mathbb{E}\left[(1 + X_{j,\max})^\beta \prod_{\substack{k=1 \\ k\neq j}}^{d} (1 + X_{k,\max})\right] \le c_1(\beta) M^{d-1+\beta}.$$

Then there exists a constant $r(b) > 0$ such that for all $n \ge d$,

$$U_1(b, j, n), U_2(b, j, n) \le r(b) \left\{\max\{1, h^2\} + \max\{1, h\}M\right\} (1 + M)^d. \tag{58}$$

Proof. At a high level, using the monotonicity of $\mathcal{F}$, for a target coordinate j we partition the samples $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$ such that samples in the same class differ by (possibly) only the j-th coordinate. Then for each class, using monotonicity, we mimic the proof for the one-dimensional case. Before proceeding with the proof we define the following notations for all $j = 1, \ldots, d$ and $x' \in \mathbb{Z}_+^{d-1}$

$$\begin{aligned} C_j(\boldsymbol{x}') &\triangleq \{\boldsymbol{x} \in \mathbb{Z}_+^d : x_i = x'_i \ \forall i \le j-1 \text{ and } x_i = x'_{i-1} \ \forall i \ge j+1\}, \\ N_j(\boldsymbol{x}') &= \sum_{\boldsymbol{x} \in \mathbb{Z}_+^d} N(\boldsymbol{x}) \mathbf{1}_{\{\boldsymbol{x} \in C_j(\boldsymbol{x}')\}}. \end{aligned} \tag{59}$$

In addition, we will use multiple times that by union bound we have

$$\sup_{\pi \in \mathcal{P}([0,h])^d} \mathbb{P}_{\boldsymbol{X} \sim p_\pi}\left[\boldsymbol{X} \notin [0, M]^d\right] \le \sum_{i=1}^{d} \sup_{\pi \in \mathcal{P}([0,h])^d} \mathbb{P}_{\boldsymbol{X} \sim p_\pi}\left[X_j > M\right] \le \frac{d}{n^7}$$

Bound on $U_1(b, j, n)$. Denote $m_b = 1 + b$ and note that for each $\boldsymbol{f} \in \boldsymbol{\mathcal{F}}$, and for each class $C_j(\boldsymbol{x}')$, as f_j is monotone over the j-th coordinate of all $\boldsymbol{x}$-s in $C_j(\boldsymbol{x}')$, there exists $v \triangleq v(f_j, \boldsymbol{x}')$ such that for all $\boldsymbol{x} \in C_j(\boldsymbol{x}')$, $f_j(\boldsymbol{x}) \le m_b h$ if and only if $x_j \le v$. Using the above we can write

$$\begin{aligned}
&\sup_{\boldsymbol{f} \in \boldsymbol{\mathcal{F}}_* \cup \boldsymbol{\mathcal{F}}'_*} \sum_{i=1}^{n} \left(\epsilon_i - \frac{1}{b}\right) (f_j^*(\boldsymbol{X}_i) - f_j(\boldsymbol{X}_i))^2 = \sup_{\boldsymbol{f} \in \boldsymbol{\mathcal{F}}_* \cup \boldsymbol{\mathcal{F}}'_*} \sum_{\boldsymbol{x}: N(\boldsymbol{x}) > 0} (\epsilon(\boldsymbol{x}) - \frac{1}{b} N(\boldsymbol{x}))(f_j(\boldsymbol{x}) - f_j^*(\boldsymbol{x}))^2 \\
&= \sup_{\boldsymbol{f} \in \boldsymbol{\mathcal{F}}_* \cup \boldsymbol{\mathcal{F}}'_*} \sum_{\boldsymbol{x}': N_j(\boldsymbol{x}') > 0} \sum_{\boldsymbol{x} \in C_j(\boldsymbol{x}')} (\epsilon(\boldsymbol{x}) - \frac{1}{b} N(\boldsymbol{x}))(f_j(\boldsymbol{x}) - f_j^*(\boldsymbol{x}))^2 \\
&= \sup_{\boldsymbol{f} \in \boldsymbol{\mathcal{F}}_* \cup \boldsymbol{\mathcal{F}}'_*} \sum_{\boldsymbol{x}': N_j(\boldsymbol{x}') > 0} \left(\sum_{\boldsymbol{x} \in C_j(\boldsymbol{x}'), x_j \le v} + \sum_{\boldsymbol{x} \in C_j(\boldsymbol{x}'), x_j > v} \right) (\epsilon(\boldsymbol{x}) - \frac{1}{b} N(\boldsymbol{x}))(f_j(\boldsymbol{x}) - f_j^*(\boldsymbol{x}))^2 \\
&\le \sup_{\boldsymbol{f} \in \boldsymbol{\mathcal{F}}_* \cup \boldsymbol{\mathcal{F}}'_*} \sum_{\boldsymbol{x}': N_j(\boldsymbol{x}') > 0} \left(m_b^2 h^2 \sum_{\substack{\boldsymbol{x} \in C_j(\boldsymbol{x}'), \\ x_j \le v}} \max\{0, \epsilon(\boldsymbol{x}) - \frac{1}{b} N(\boldsymbol{x})\} + \sum_{\substack{\boldsymbol{x} \in C_j(\boldsymbol{x}'), \\ x_j > v}} (\epsilon(\boldsymbol{x}) - \frac{1}{b} N(\boldsymbol{x}))(f_j(\boldsymbol{x}) - f_j^*(\boldsymbol{x}))^2 \right) \\
&\le m_b^2 h^2 \sum_{N(\boldsymbol{x}) > 0} \max\{0, \epsilon(\boldsymbol{x}) - \frac{1}{b} N(\boldsymbol{x})\} \\
&+ \left\{ \sum_{\boldsymbol{x}': N_j(\boldsymbol{x}') > 0} \sup_{\substack{f \in \boldsymbol{\mathcal{F}}_* \cup \boldsymbol{\mathcal{F}}'_*, \\ N_c h \le f_j \le X_{j,\max}}} \left\{ \sup_{v(\boldsymbol{x}') \ge 0} \sum_{\substack{\boldsymbol{x} \in C_j(\boldsymbol{x}'), \\ x_j > v(\boldsymbol{x}')}} (\epsilon(\boldsymbol{x}) - \frac{1}{b} N(\boldsymbol{x}))(f_j(\boldsymbol{x}) - f_j^*(\boldsymbol{x}))^2 \right\} \right\}
\end{aligned} \tag{60}$$

As there are at most $\prod_{j=1}^{d}(1+X_{j,\max})$ vectors $\boldsymbol{x}$ with $N(\boldsymbol{x})>0$, we apply Lemma 5 to bound the expectation of the first term in the above display as

$$\begin{aligned}
&m_b^2h^2\mathbb{E}[\sum_{N(\boldsymbol{x})>0}\max\{0,\epsilon(\boldsymbol{x})-\frac{1}{b}N(\boldsymbol{x})\}|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n]\\
&\leq m_b^2h^2\mathbb{E}[\sum_{N(\boldsymbol{x})>0}1|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n]\overset{(a)}{\leq} r_1(b)m_b^2h^2\prod_{j=1}^{d}(1+X_{j,\max}). \qquad (61)
\end{aligned}$$

where (a) followed from Lemma 5 with $r_1(b)=\frac{1-1/b}{e\cdot D(\frac{1+1/b}{2}\|\frac{1}{2})}$.

For the second term in (60), note that for the vectors in the set $C_j(\boldsymbol{x}')$, the only coordinate that takes different values is the j-th coordinate, and the function f_j is monotone when we condition on the coordinates $\{1,\ldots,j-1,j+1,\ldots,d\}$. It follows that conditional on $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$, for this class $C_j(\boldsymbol{x}')$, we can mimic the proof for (45) in one dimensional case of $T_1(b,n)$ to bound the innermost term as

$$\begin{aligned}
&\mathbb{E}\left[\sup_{v}\sup_{\boldsymbol{f}\in\boldsymbol{\mathcal{F}}_*\cup\boldsymbol{\mathcal{F}}'_*}\left\{\sum_{\substack{\boldsymbol{x}\in C_j(\boldsymbol{x}'),\\ x_j>v}}\max\{0,(\epsilon(\boldsymbol{x})-\frac{1}{b}N(\boldsymbol{x}))(f_j(\boldsymbol{x})-f_j^*(\boldsymbol{x}))^2\}\right\}\Bigg|\boldsymbol{X}_1^n\right]\\
&\leq r_2(b)\left(h^2(1+X_{j,\max})+(1+X_{j,\max})^2\right)
\end{aligned}$$

for a constant $c(b)$ depending on b. Finally, the number of such classes with $N_j(\boldsymbol{x}')>0$ is bounded above by $\prod_{\substack{k=1\\k\neq j}}^{d}(1+X_{k,\max})$. Therefore, summing over all classes and taking the expectation, and including (61), we get the bound

$$\begin{aligned}
U_1(b,j,n)&=\mathbb{E}\left[\sup_{\boldsymbol{f}\in\boldsymbol{\mathcal{F}}_*\cup\boldsymbol{\mathcal{F}}'_*}\sum_{i=1}^{n}\left(\epsilon_i-\frac{1}{b}\right)(f_j^*(\boldsymbol{X}_i)-f_j(\boldsymbol{X}_i))^2\right]\\
&\leq r_1(b)m_b^2h^2\mathbb{E}\left[\prod_{j=1}^{d}(1+X_{j,\max})\right]+r_2(b)\mathbb{E}\left[\prod_{\substack{k=1\\k\neq j}}^{d}(1+X_{k,\max})\cdot(h^2X_{j,\max}+X_{j,\max}^2)\right]\\
&\leq (r_1(b)+r_2(b))(c_1(1)h^2+c_1(2)M)M^d. \qquad (62)
\end{aligned}$$

Bounding $U_2(b,j,n)$. As per the one dimensional case, we bound the Rademacher complexity term $U_2(b,j,n)$ with $t_0(n)+t_1(n)$, where

$$\begin{aligned}
t_1(n)\triangleq\mathbb{E}\Bigg[&\sup_{\boldsymbol{f}\in\boldsymbol{\mathcal{F}}_*\cup\boldsymbol{\mathcal{F}}'_*}\sum_{\boldsymbol{x}}(2(\epsilon(\boldsymbol{x})f_j^*(\boldsymbol{x})-(x_j+1)\epsilon(\boldsymbol{x}+\boldsymbol{e}_j))(f_j^*(\boldsymbol{x})-f_j(\boldsymbol{x}))\\
&-\frac{N(\boldsymbol{x})}{b}(f_j^*(\boldsymbol{x})-f_j(\boldsymbol{x}))^2\mathbf{1}_{\{N(\boldsymbol{x})>0\}}\Bigg] \qquad (63)\\
t_0(n)\triangleq\mathbb{E}\Bigg[&\sup_{\boldsymbol{f}\in\boldsymbol{\mathcal{F}}_*\cup\boldsymbol{\mathcal{F}}'_*}\sum_{\boldsymbol{x}}-2(x_j+1)\epsilon(\boldsymbol{x}+\boldsymbol{e}_j)(f_j^*(\boldsymbol{x})-f_j(\boldsymbol{x}))\mathbf{1}_{\{N(\boldsymbol{x})=0\}}\Bigg] \qquad (64)
\end{aligned}$$

We first analyze $t_1(n)$. Using the inequality $2ax-bx^2\leq\frac{a^2}{b}$ for any $b>0$ we have

$$\frac{1}{b}t_1(n)\leq\mathbb{E}\left[\sum_{\boldsymbol{x}}\frac{(\epsilon(\boldsymbol{x})f_j^*(\boldsymbol{x})-(x_j+1)\epsilon(\boldsymbol{x}+\boldsymbol{e}_j))^2}{N(\boldsymbol{x})}\mathbf{1}_{\{N(\boldsymbol{x})>0\}}\right] \qquad (65)$$

Using the facts

- $\mathbb{E}[\epsilon(\boldsymbol{x})|\boldsymbol{X}_1,\dots,\boldsymbol{X}_n]=0,\ \mathbb{E}[\epsilon(\boldsymbol{x})\epsilon(\boldsymbol{x}+\boldsymbol{e}_j)|\boldsymbol{X}_1,\dots,\boldsymbol{X}_n]=0$
- $\mathbb{E}[\epsilon(\boldsymbol{x})^2|\boldsymbol{X}_1,\dots,\boldsymbol{X}_n]=N(\boldsymbol{x})$, and,
- $\mathbb{E}[N(\boldsymbol{x}+\boldsymbol{e}_j)\mid N(\boldsymbol{x})]=\frac{(n-N(\boldsymbol{x}))p_\pi(\boldsymbol{x}+\boldsymbol{e}_j)}{1-p_\pi(\boldsymbol{x})}\le\frac{np_\pi(\boldsymbol{x}+\boldsymbol{e}_j)}{1-p_\pi(\boldsymbol{x})}$

we continue the last display to get

$$\begin{aligned}
\frac{1}{b}t_1(n)&\le\mathbb{E}[\sum_{\boldsymbol{x}}\left(f_j^*(\boldsymbol{x})^2+\frac{(x_j+1)^2N(\boldsymbol{x}+\boldsymbol{e}_j)}{N(\boldsymbol{x})}\right)\mathbf{1}_{\{N(\boldsymbol{x})>0\}}]\\
&\le\mathbb{E}[\sum_{\boldsymbol{x}}h^2\mathbf{1}_{\{N(\boldsymbol{x})>0\}}]+\mathbb{E}[\sum_{\boldsymbol{x}}\frac{(x_j+1)^2np_\pi(\boldsymbol{x}+\boldsymbol{e}_j)}{1-p_\pi(\boldsymbol{x})}\cdot\frac{\mathbf{1}_{\{N(\boldsymbol{x})>0\}}}{N(\boldsymbol{x})}]\\
&\overset{(a)}{\le}\mathbb{E}[\sum_{\boldsymbol{x}}h^2\mathbf{1}_{\{N(\boldsymbol{x})>0\}}]+c'\cdot\sum_{\boldsymbol{x}}\frac{(x_j+1)^2np_\pi(\boldsymbol{x}+\boldsymbol{e}_j)}{1-p_\pi(\boldsymbol{x})}\cdot\min\{np_\pi(\boldsymbol{x}),\frac{1}{np_\pi(\boldsymbol{x})}\}\\
&\overset{(b)}{=}\mathbb{E}[\sum_{\boldsymbol{x}}h^2\mathbf{1}_{\{N(\boldsymbol{x})>0\}}]+c'\cdot\sum_{\boldsymbol{x}}\frac{(x_j+1)f_j^*(\boldsymbol{x})}{1-p_\pi(\boldsymbol{x})}\cdot\min\{1,(np_\pi(\boldsymbol{x}))^2\}\\
&\overset{(c)}{\le}h^2\mathbb{E}[\prod_{j=1}^d(1+X_{j,\max})]+\frac{c'f_j^*(\mathbf{0})}{1-p_\pi(\mathbf{0})}+c'\sum_{\boldsymbol{x}\neq\mathbf{0}}(x_j+1)f_j^*(\boldsymbol{x})\cdot\min\{1,(np_\pi(\boldsymbol{x}))^2\}
\end{aligned}\tag{66}$$

(here c' is an absolute constant), where:

- (a) is due to Property (P1) in the analysis of $T_2(b,n)$;
- (b) is using $f_j^*(\boldsymbol{x})=(x_j+1)\frac{p_\pi(\boldsymbol{x}+\boldsymbol{e}_j)}{p_\pi(\boldsymbol{x})}=\mathbb{E}\left[\theta_j|X=\boldsymbol{x}\right]\le h$;
- (c): for the first term, we use the fact that the number of vectors $\boldsymbol{x}$ with $N(\boldsymbol{x})>0$ is bounded by $\prod_{j=1}^d(1+X_{j,\max})$; for the third term, for each $\boldsymbol{x}\neq\mathbf{0}$ we may choose a coordinate k with $x_k>0$. Thus setting p_{π_k} as the marginal distribution of x_k we have by Stirling's inequality, again,

$$p_\pi(\boldsymbol{x})\le p_{\pi_k}(x_k)\le\sup_{\theta\ge0}\mathbb{P}_{X\sim\mathrm{Poi}(\theta)}[X=x_k]=\sup_{\theta\ge0}\frac{\theta^{x_k}e^{-\theta}}{x_k!}=\frac{x_k^{x_k}e^{-x_k}}{x_k!}\le\frac{1}{\sqrt{2\pi x_k}}\le\frac{1}{\sqrt{2\pi}}$$

and therefore $\frac{1}{1-p_\pi(\boldsymbol{x})}\le\frac{1}{1-\frac{1}{\sqrt{2\pi}}}\le O(1)$.

Now, the first term in (66) is bounded by $h^2c(1)M^d$. For the second term, using $p_\pi(\boldsymbol{e}_j)\le1-p_\pi(\mathbf{0})$ we have $\frac{f_j^*(\mathbf{0})}{1-p_\pi(\mathbf{0})}\le\frac{f_j^*(\mathbf{0})}{p_\pi(\boldsymbol{e}_j)}=\frac{1}{p_\pi(\mathbf{0})}$, so

$$\frac{f_j^*(\mathbf{0})}{1-p_\pi(\mathbf{0})}\le\min\left\{\frac{f_j^*(\mathbf{0})}{1-p_\pi(\mathbf{0})},\frac{1}{p_\pi(\mathbf{0})}\right\}\le2\max\{f_j^*(\mathbf{0}),1\}\le2\max\{h,1\}\tag{67}$$

given that $\boldsymbol{f}^*$ is bounded by h in each coordinates. Finally, the third term in (66) has the following bound:

$$
\begin{aligned}
&\sum_{\boldsymbol{x}\neq\mathbf{0}}(x_j+1)f_j^*(\boldsymbol{x})\cdot\min\{1,(np_\pi(\boldsymbol{x}))^2\}\\
&\leq h\sum_{\boldsymbol{x}\in[0,M]^d}(x_j+1)+n^2h\sum_{\boldsymbol{x}\notin[0,M]^d}(x_j+1)\cdot p_\pi(\boldsymbol{x})^2\\
&\overset{\text{(a)}}{\leq} h(1+M)^{d+1}+n^2h\mathbb{P}_{\boldsymbol{X}\sim p_\pi}\left[\boldsymbol{X}\notin[0,M]^d\right]\mathbb{E}_{\boldsymbol{X}\sim p_\pi}[X_j+1]\overset{\text{(b)}}{\leq} h(1+M)^{d+1}+hdn^{-4}(1+c_1(4)^{1/4}M)
\end{aligned}
\tag{68}
$$

where (a) followed as there are $(1+M)^d$ elements in $[0,M]^d$, and (b) is due to the assumptions in Lemma 8 and $\mathbb{E}[X_{j,\max}+1]\leq\left\{\mathbb{E}[(X_{j,\max}+1)^4]\right\}^{1/4}$. Thus, summarizing (66),(67),(68), we have

$$
\begin{aligned}
t_1(n)&\leq c''\cdot b\left(h^2c_1(1)M^d+\max\{h,1\}+h(1+M)^{d+1}+hdn^{-4}M\right)\\
&\leq 2c''b\left(\max\{1,h\}(1+M)^{d+1}+\max\{1,h^2\}c_1(1)(1+M)^d+hdn^{-4}M\right)
\end{aligned}
$$

for an absolute constant c'', as desired. Since $d\leq n$, $hdn^{-4}M\leq hn^{-3}M<h(1+M)^d$, and can therefore be negleected.

Next we analyze $t_0(n)$. Since we have $|\epsilon(\boldsymbol{x}+\boldsymbol{e}_j)|\leq N(\boldsymbol{x}+\boldsymbol{e}_j)$ and $N(\boldsymbol{x}+\boldsymbol{e}_j)=0$ for all $\boldsymbol{x}$ with $\boldsymbol{x}+\boldsymbol{e}_j\notin\prod_{k=1}^d[0,X_{k,\max}]$, we get

$$
\begin{aligned}
t_0(n)&=\mathbb{E}\left[\sup_{\boldsymbol{f}\in\boldsymbol{\mathcal{F}}_*\cup\boldsymbol{\mathcal{F}}'_*}\sum_{\boldsymbol{x}}[-2(x_j+1)\epsilon(\boldsymbol{x}+\boldsymbol{e}_j)(f_j^*(\boldsymbol{x})-f_j(\boldsymbol{x}))\mathbf{1}_{\{N(\boldsymbol{x})=0\}}]\right]\\
&\leq\mathbb{E}\left[\sum_{\boldsymbol{x}+\boldsymbol{e}_j\in\prod_{k=1}^d[0,X_{k,\max}]}2(x_j+1)N(\boldsymbol{x}+\boldsymbol{e}_j)\sup_{f\in\mathcal{F}_*\cup\mathcal{F}'_*}\left|f_j^*(\boldsymbol{x})-f_j(\boldsymbol{x})\right|\mathbf{1}_{\{N(\boldsymbol{x})=0\}}\right]\\
&\leq\mathbb{E}\left[\sum_{\boldsymbol{x}+\boldsymbol{e}_j\in\prod_{k=1}^d[0,X_{k,\max}]}2(x_j+1)\left(f_j^*(\boldsymbol{x})+X_{j,\max}+X'_{j,\max}\right)N(\boldsymbol{x}+\boldsymbol{e}_j)\mathbf{1}_{\{N(\boldsymbol{x})=0\}}\right]
\end{aligned}
\tag{69}
$$

where $X'_{j,\max}$ is the maximum of j-th coordinate on n samples independent of $\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n$.

Define $A=\left\{\boldsymbol{X}_i,\boldsymbol{X}_{i'}\in[0,M]^d,\forall i=1,\cdots,n\right\}$. We have $\mathbb{P}\left[A^c\right]\leq\frac{2d}{n^6}$ via union bound. Then we have for an absolute constant $c_1'>0$

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{\boldsymbol{x}+\boldsymbol{e}_j\in\prod_{k=1}^d[0,X_{k,\max}]}2(x_j+1)\left(f_j^*(\boldsymbol{x})+X_{j,\max}+X'_{j,\max}\right)N(\boldsymbol{x}+\boldsymbol{e}_j)\mathbf{1}_{\{N(\boldsymbol{x})=0\}}\cdot\mathbf{1}_{\{A^c\}}\right]\\
&\leq\mathbb{E}\left[2(X_{j,\max}+1)\left(h+X_{j,\max}+X'_{j,\max}\right)\sum_{\boldsymbol{x}+\boldsymbol{e}_j\in\prod_{k=1}^d[0,X_{k,\max}]}N(\boldsymbol{x}+\boldsymbol{e}_j)\mathbf{1}_{\{N(\boldsymbol{x})=0\}}\cdot\mathbf{1}_{\{A^c\}}\right]\\
&\overset{\text{(a)}}{\leq} n\mathbb{E}\left[(X_{j,\max}+1)\left(h+X_{j,\max}+X'_{j,\max}\right)\mathbf{1}_{\{A^c\}}\right]\\
&\overset{\text{(b)}}{\leq} n\sqrt{\mathbb{E}\left[\left(h+X_{j,\max}+X'_{j,\max}\right)^2(X_{j,\max}+1)^2\right]}\sqrt{\mathbb{P}\left[A^c\right]}\leq\ c_1'\frac{hd^{1/2}M^2}{n^2}\overset{\text{(c)}}{\leq}\frac{c_1'hM^2}{n},
\end{aligned}
\tag{70}
$$

where (a) is using $\sum_{\boldsymbol{x}+\boldsymbol{e}_j \in \prod_{k=1}^d [0, X_{k,\max}]} N(\boldsymbol{x}+\boldsymbol{e}_j)\mathbf{1}_{\{N(\boldsymbol{x})=0\}} \le \sum_{\boldsymbol{x}} N(\boldsymbol{x}) = n$, (b) is via Cauchy-Schwarz inequality and $\mathbb{E}[(X_{j,\max})^4], \mathbb{E}[(X'_{j,\max})^4] \le cM^4$, and (c) is because $d \le n$ by our assumption.

Next, we condition on the event A. Similar to the proof of bound on $T_2(b,n)$ in the one-dimensional setup, we define $q_{\pi,M}(\boldsymbol{x}) \triangleq \frac{p_\pi(\boldsymbol{x})}{\mathbb{P}_{\boldsymbol{X}\sim p_\pi}[\boldsymbol{X}\in[0,M]^d]}$. We have $\mathbb{P}[N(\boldsymbol{x})=0|A] = (1-q_{\pi,M}(\boldsymbol{x}))^n$, and conditioned on the set A and $\{N(\boldsymbol{x})=0\}$, $N(\boldsymbol{x}+\boldsymbol{e}_j) \sim \text{Binom}\left(n, \frac{q_{\pi,M}(\boldsymbol{x}+\boldsymbol{e}_j)}{1-q_{\pi,M}(\boldsymbol{x})}\right)$. Therefore:

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{\boldsymbol{x}+\boldsymbol{e}_j \in \prod_{k=1}^d [0, X_{k,\max}]} 2(x_j+1)\left(f_j^*(\boldsymbol{x}) + X_{j,\max} + X'_{j,\max}\right) N(\boldsymbol{x}+\boldsymbol{e}_j)\mathbf{1}_{\{N(\boldsymbol{x})=0\}} \,\middle|\, A\right] \\
&\le \sum_{\boldsymbol{x}\in\prod_{k=1}^d [0,M]^d} 2(x_j+1)(h+2M)\mathbb{E}\left[N(\boldsymbol{x}+\boldsymbol{e}_j)|\{N(\boldsymbol{x})=0\}, A\right]\mathbb{P}\left[N(\boldsymbol{x})=0|A\right] \\
&\le \sum_{\boldsymbol{x}\in\prod_{k=1}^d [0,M]^d} 2(x_j+1)(h+2M)\frac{nq_{\pi,M}(\boldsymbol{x}+\boldsymbol{e}_j)}{1-q_{\pi,M}(\boldsymbol{x})}\left(1-q_{\pi,M}(\boldsymbol{x})\right)^n \\
&\overset{(a)}{=} \sum_{\boldsymbol{x}\in\prod_{k=1}^d [0,M]^d} 2(h+2M)f_j^*(\boldsymbol{x})nq_{\pi,M}(\boldsymbol{x})\left(1-q_{\pi,M}(\boldsymbol{x})\right)^{n-1} \le 2(M+1)^d h(h+2M).
\end{aligned}
$$

where (a) followed using $f_j^*(\boldsymbol{x}) = (x_j+1)\frac{p_\pi(\boldsymbol{x}+\boldsymbol{e}_j)}{p_\pi(\boldsymbol{x})}$ and the definition of $q_{\pi,M}(x+\boldsymbol{e}_j)$, and for the last inequality, we used the fact that $nx(1-x)^{n-1} \le (1-\frac{1}{n})^{n-1} < 1$ for all x with $0 < x < 1$ and $f_j^*(\boldsymbol{x}) \le h$. Collecting terms and using $M > h$, we therefore have

$$
t_0(n) \le c_1' \frac{hd^{1/2}M^2}{n^2} + h(M+1)^{d+1} \le c_2' h(M+1)^{d+1} \tag{71}
$$

for absolute constants c_1', c_2' as required. □

3.2 Proof of Regret bound in the multidimensional setup (Theorem 2)

We start by describing the bounds on $\mathbb{E}[\prod_{j=1}^d (1+X_{j,\max})^{k_j}]$ in this multidimensional setting, which we claim the following.

Lemma 9. *Given any $s, h > 0$ and integer $\beta \ge 0$ there exist constants $c(\beta), c_1, c_2, c_3, c_4 > 0$ such that*

1. *For all $\pi \in \mathcal{P}([0,h]^d)$, $\mathbb{E}\left[(1+X_{j,\max})^\beta \prod_{\substack{k=1\\k\neq j}}^d (1+X_{k,\max})\right] \le c(\beta)\left(\max\{c_1, c_2 h\}\frac{\log(n)}{\log\log(n)}\right)^{d-1+\beta}$;*
2. *For all $\pi \in \mathcal{P}([0, s\log n]^d)$, $\mathbb{E}\left[(1+X_{j,\max})^\beta \prod_{\substack{k=1\\k\neq j}}^d (1+X_{k,\max})\right] \le c(\beta)(\max\{c_3, c_4 s\}\log(n))^{d-1+\beta}$.*

We will defer the proof to Appendix A.

For $\pi \in \mathcal{P}([0,h])^d$, by Lemma 9, there exist constants c_1, c_2 such that we may take $M = \max\{c_1, c_2 h\}\frac{\log(n)}{\log\log(n)}$ into Lemma 8. Note that This gives the overall regret bound as $\frac{d}{n}\max\{c_1, c_2 h\}^{d+2}(\frac{\log(n)}{\log\log(n)})^{d+1}$.

Now assume that each marginals of π_j are of $\mathsf{SubE}(s)$ for some $s > 0$. We now show that the multidimensional version of Lemma 7 applies here.

Here, we choose $c = c(s) \triangleq 11s$ such that for each $j = 1, \cdots, d$, we have $\mathbb{P}[X_j > c(s)\log(n)] \le \frac{1}{n^{10}}$. This means that we now have

$$\varepsilon = \mathbb{P}[\boldsymbol{X} \notin [0, c(s)\log(n)]^d] \le \sum_{j=1}^{d} \mathbb{P}[X_j > c(s)\log n] \le \frac{d}{n^{10}} \tag{72}$$

the middle inequality via union bound on each coordinate.

Define the event $E = \{\boldsymbol{X}_i \in [0, c(s)\log(n)]^d, \forall i = 1, \cdots, n\}$, and we have $\mathbb{P}[E^c] \le dn^{-9}$. Again we define the truncated prior $\pi_{c,n}[\boldsymbol{X} \in \cdot] = \pi[\boldsymbol{X} \in \cdot \mid \boldsymbol{X} \in [0, c(s)\log(n)]^d]$. Then, similar to (47) in the one-dimensional case, the following equation applies:

$$\mathsf{Regret}_{\pi}(\widehat{\boldsymbol{f}}_{\mathsf{erm}}) \le \mathsf{Regret}_{\pi_{c,n}}(\widehat{\boldsymbol{f}}_{\mathsf{erm}}) + \mathrm{mmse}(\pi_{c,n}) - \mathrm{mmse}(\pi) + \mathbb{E}_{\pi,c}[\|\widehat{\boldsymbol{f}}_{\mathsf{erm}}(\boldsymbol{X}) - \boldsymbol{\theta}\|^2 \mathbf{1}_{\{E^c\}}] \tag{73}$$

Given that $\widehat{f}_j(\cdot) \le X_{j,\max}$, we have $\mathbb{E}[(\widehat{f}_j)^4] \le \mathbb{E}[X_{j,\max}^4] \le O(s^4(\log n)^4)$ by Lemma 11, and $\mathbb{E}_\pi[\theta_j^4] \le O(s^4 \log^4 n)$ from the properties of subexponential priors. The logic $\mathbb{E}_\pi[(f_j^* - \theta_j)^4] \le O((s\log n)^4)$ and

$$\mathbb{E}_\pi[(f_{\mathsf{erm},j}(\boldsymbol{X}) - \theta_j)^2 \mathbf{1}_{\{E^c\}}] \le \sqrt{\mathbb{P}[E^c]\mathbb{E}_\pi[(f_{\mathsf{erm},j}(\boldsymbol{X}) - \theta_j)^4]} \lesssim \frac{s^2 d^{1/2}}{n^2}, \qquad \forall j = 1, 2, \cdots, d$$

then follows from there. This gives $\mathbb{E}_{\pi,c}[\|\widehat{\boldsymbol{f}}_{\mathsf{erm}}(\boldsymbol{X}) - \boldsymbol{\theta}\|^2 \mathbf{1}_{\{E^c\}}] \le \frac{d^{3/2}}{n^4}$ by considering all the d coordinates.

The identity $\mathrm{mmse}(\pi_c) - \mathrm{mmse}(\pi) \le \frac{\varepsilon}{1-\varepsilon}\mathrm{mmse}(\pi) \le 2d\varepsilon \le \frac{2d^2}{n^2}$ still applies here in the following sense. Let $\boldsymbol{f}^*$ be the Bayes estimator corresponding to π. Then denoting $M \triangleq c(s)\log(n)$ here we have

$$\begin{aligned}
\mathrm{mmse}(\pi) &= \mathbb{E}[\|\boldsymbol{f}^*(\boldsymbol{X}) - \boldsymbol{\theta}\|^2] \\
&= \mathbb{E}_{\boldsymbol{\theta}\sim\pi}[\mathbb{E}_{\boldsymbol{X}\sim\mathrm{Poi}(\boldsymbol{\theta})}[\|\boldsymbol{f}^*(\boldsymbol{X}) - \boldsymbol{\theta}\|^2]|\boldsymbol{\theta}] \\
&\ge \mathbb{E}_{\boldsymbol{\theta}\sim\pi}[\mathbb{E}_{\boldsymbol{X}\sim\mathrm{Poi}(\boldsymbol{\theta})}[\|\boldsymbol{f}^*(\boldsymbol{X}) - \boldsymbol{\theta}\|^2]\mathbf{1}_{\{\boldsymbol{\theta}\in[0,M]^d\}}|\boldsymbol{\theta}] \\
&= \mathbb{P}[\boldsymbol{\theta} \in [0, M]^d]\mathbb{E}_{\boldsymbol{\theta}\sim\pi}[\mathbb{E}_{\boldsymbol{X}\sim\mathrm{Poi}(\boldsymbol{\theta})}[\|\boldsymbol{f}^*(\boldsymbol{X}) - \boldsymbol{\theta}\|^2]\mathbf{1}_{\{\boldsymbol{\theta}\in[0,M]^d\}}|\boldsymbol{\theta}] \\
&\ge (1-\epsilon)\mathrm{mmse}(\pi_{c,n})
\end{aligned} \tag{74}$$

and that $\mathrm{mmse}(\pi) \le d$ given that the naive estimation of $\boldsymbol{f}_{\mathsf{id}}(\boldsymbol{x}) = \boldsymbol{x}$ achieves an expected loss of d (i.e. 1 for each coordinate). This shows that we also have $\mathsf{Regret}_\pi(\boldsymbol{f}^*) \le \mathsf{Regret}_{\pi_{c,n}}(\boldsymbol{f}^*) + O(\frac{d^2 s^2}{n^2}) \le \mathsf{Regret}_{\pi_{c,n}}(\boldsymbol{f}^*) + O(\frac{ds^2}{n})$ in this multidimensional case (given that $d \le n$). Thus, it suffices to work on prior $\pi_{c,n}$ supported on $[0, c\log(n)]^d$ for some $c \triangleq c(s)$.

Now under this truncated prior, by Lemma 9 there exist absolute constants c_3, c_4 such that we may take $M = \max\{c_3, c_4 s\}\log n$ and substitute into Lemma 8. This gives an overall regret bound of $\frac{d}{n}(\max\{c_3, c_4 s\}\log(n))^{d+2}$.

References

[Bar91] Andrew R Barron. Complexity regularization with application to artificial neural networks. In *Nonparametric functional estimation and related topics*, pages 561–576. Springer, 1991.

[Bar94] Andrew R Barron. Approximation and estimation bounds for artificial neural networks. *Machine learning*, 14(1):115–133, 1994.

[BBL02] Peter L Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48(1):85–113, 2002.

[BBM05] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4), August 2005. arXiv:math/0508275.

[BC90] Michael J. Best and Nilotpal Chakravarti. Active set algorithms for isotonic regression; A unifying framework. *Mathematical Programming*, 47(1-3):425–439, May 1990.

[BC91] Andrew R Barron and Thomas M Cover. Minimum complexity density estimation. *IEEE transactions on information theory*, 37(4):1034–1054, 1991.

[BGR13] Lawrence D Brown, Eitan Greenshtein, and Ya'acov Ritov. The poisson compound decision problem revisited. *Journal of the American Statistical Association*, 108(502):741–749, 2013.

[BM93] Lucien Birgé and Pascal Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97(1):113–150, 1993.

[BZ22] Alton Barbehenn and Sihai Dave Zhao. A nonparametric regression approach to asymptotically optimal estimation of normal means. *arXiv preprint arXiv:2205.00336*, 2022.

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

[Efr14] Bradley Efron. Two modeling strategies for empirical bayes estimation. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(2):285, 2014.

[EH21] Bradley Efron and Trevor Hastie. *Computer Age Statistical Inference, Student Edition: Algorithms, Evidence, and Data Science*, volume 6. Cambridge University Press, 2021.

[HS83] JC van Houwelingen and Th Stijnen. Monotone empirical Bayes estimators for the continuous one-parameter exponential family. *Statistica Neerlandica*, 37(1):29–43, 1983.

[JPW22] Soham Jana, Yury Polyanskiy, and Yihong Wu. Optimal empirical bayes estimation for the poisson model via minimum-distance methods. *arXiv preprint arXiv:2209.01328*, 2022.

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

[KG17] Roger Koenker and Jiaying Gu. Rebayes: an r package for empirical bayes mixture methods. *Journal of Statistical Software*, 82:1–26, 2017.

[KM14] Roger Koenker and Ivan Mizera. Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association*, 109(506):674–685, 2014.

[Kol01] Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

[KP04] Vladimir Koltchinskii and Dmitry Panchenko. Rademacher processes and bounding the risk of function learning, May 2004. arXiv:math/0405338.

[LGL05] Jianjun Li, Shanti S Gupta, and Friedrich Liese. Convergence rates of empirical Bayes estimation in exponential family. *Journal of statistical planning and inference*, 131(1):101–115, 2005.

[Lin83] Bruce G Lindsay. The geometry of mixture likelihoods: a general theory. *The annals of statistics*, pages 86–94, 1983.

[LRS15] Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pages 1260–1285. PMLR, 2015.

[LW04] Gábor Lugosi and Marten Wegkamp. Complexity regularization via localized random penalties. *The Annals of Statistics*, 32(4):1679–1697, 2004.

[LZ95] Gábor Lugosi and Kenneth Zeger. Nonparametric estimation via empirical risk minimization. *IEEE Transactions on information theory*, 41(3):677–687, 1995.

[Mar66] JS Maritz. Smooth empirical bayes estimation for one-parameter discrete distributions. *Biometrika*, 53(3-4):417–429, 1966.

[Mar68] JS Maritz. On the smooth empirical Bayes approach to testing of hypotheses and the compound decision problem. *Biometrika*, 55(1):83–100, 1968.

[Mar69] JS Maritz. Empirical bayes estimation for the Poisson distribution. *Biometrika*, 56(2):349–359, 1969.

[Men02] Shahar Mendelson. Rademacher averages and phase transitions in glivenko-cantelli classes. *IEEE transactions on Information Theory*, 48(1):251–263, 2002.

[ML18] Johannes S Maritz and T Lwin. *Empirical bayes methods*. Chapman and Hall/CRC, 2018.

[MU05] Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.

[Nem85] Arkadii Nemirovskii. Nonparametric estimation of smooth regression functions. *Soviet Journal of Computer and Systems Sciences*, 23(6):1–11, 1985.

[PW20] Yury Polyanskiy and Yihong Wu. Self-regularizing property of nonparametric maximum likelihood estimator in mixture models. *arXiv preprint arXiv:2008.08244*, 2020.

[PW21] Yury Polyanskiy and Yihong Wu. Sharp regret bounds for empirical bayes and compound decision problems. *arXiv preprint arXiv:2109.03943*, 2021.

[PW22] Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2022+.

[Rob51] Herbert Robbins. Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, pages 131–149. University of California Press, 1951.

[Rob56] Herbert Robbins. An Empirical Bayes Approach to Statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1956.

[SGS21] Jake A. Soloff, Adityanand Guntuboyina, and Bodhisattva Sen. Multivariate, Heteroscedastic Empirical Bayes via Nonparametric Maximum Likelihood. Technical Report arXiv:2109.03466, arXiv, September 2021. arXiv:2109.03466 [math, stat] type: article.

[SW22] Yandi Shen and Yihong Wu. Empirical bayes estimation: When does g-modeling beat f-modeling in theory (and in practice)? *arXiv preprint arXiv:2211.12692*, 2022.

[VdG90] Sara Van de Geer. Estimating a regression function. *The Annals of Statistics*, pages 907–924, 1990.

[VH77] JC Van Houwelingen. Monotonizing empirical bayes estimators for a class of discrete distributions with monotone likelihood ratio. *Statistica Neerlandica*, 31(3):95–104, 1977.

[WV12] Yihong Wu and Sergio Verdu. Functional Properties of Minimum Mean-Square Error and Mutual Information. *IEEE Transactions on Information Theory*, 58(3):1289–1301, March 2012.

[Zha03] Cun-Hui Zhang. Compound decision theory and empirical Bayes methods. *Annals of Statistics*, pages 379–390, 2003.

A Properties of Poisson mixtures

Lemma 10. *There exist constants c_1, c_2 such that for all $h > 0, k \geq 1$ and $\pi \in \mathcal{P}([0,h])$, $X_{\max}$ on $n \geq 3$ samples have the following bound:*

$$\mathbb{P}[1 + \max X_i \geq \max\{c_2, c_1 h\} \cdot k\frac{\log n}{\log\log n}] \leq n^{-k}$$

Proof. Consider $\lambda \in [0,h]$. Then for $x \geq h$ we have the following approximation for $X \sim \mathrm{Poi}(\lambda)$ via Chernoff's bound [MU05, p.97-98]:

$$\mathbb{P}[X \geq x] \leq \frac{(e\lambda)^x e^{-\lambda}}{x^x} \leq \frac{(eh)^x e^{-h}}{x^x} \tag{75}$$

Therefore for $X \sim p_\pi$ and $x \geq h$ we have $\mathbb{P}(X \geq x) \leq \frac{(eh)^x e^{-h}}{x^x}$.

Now choose c_0 such that $c_0 \geq \max\{4, h\}$, and for all $n \geq 3$,

$$\log\log n + \log c_0 - \log\log\log n - \log h - 1 \geq \frac{1}{2}\log\log n$$

That is, denoting $L = \sup_{n\geq 3}\left\{\log\log\log n - \frac{1}{2}\log\log n\right\}$, we take $\log c_0 \geq \log h + 1 + L$. Notice that this mean we may take $c_0 = \max\{4, \max\{1, \exp(1+L)\} \cdot h\}$. Then for all $k \geq 1$, $c_0 k\frac{\log n}{\log\log n} \geq c_0\frac{\log n}{\log\log n} \geq c_0 \geq h$ given that $n > \log n$ for all $n > 1$, so the tail bound in (75) can be applied. Setting $x = c_0 k\frac{\log n}{\log\log n}$, we have

$$\begin{aligned}
\log(\frac{(eh)^x e^{-h}}{x^x}) &= -h + c_0 k\frac{\log n}{\log\log n}(1 + \log h - \log c_0 - \log k - \log\log n + \log\log\log n) \\
&\leq -h + 4k\frac{\log n}{\log\log n}(-\frac{1}{2}\log\log n) \\
&< 2k\log n\,,
\end{aligned} \tag{76}$$

which implies that $\mathbb{P}[X \geq c_0 k \frac{\log n}{\log\log n}] \leq n^{-2k}$. Finally, taking $c = 2c_0 = \max\{8, \max\{2, 2\exp(1+L)\} \cdot h\}$, we have

$$\mathbb{P}[1 + X_{\max} \geq ck\frac{\log n}{\log\log n}] \overset{(a)}{\leq} n\mathbb{P}[1 + X \geq ck\frac{\log n}{\log\log n}] \overset{(b)}{\leq} n\mathbb{P}[X \geq c_0 k\frac{\log n}{\log\log n}] \overset{(c)}{\leq} n^{-k}$$

where (a) is union bound on $X_1, \cdots, X_n$, (b) is using $\frac{\log n}{\log\log n} > 1$ for all $n \geq 3$ and $\frac{\log n}{\log\log n} k(c - c_0) \geq c_0 k \geq c_0 > 1$ for all $k \geq 1$, and (c) is $2k - 1 \geq k$ for all $k \geq 1$. □

Lemma 11. *There exist constants $c_1, c_2 > 0$ such that for all $s > 0, k \geq 1$ and $\pi \in \mathcal{P}([0, s\log n])$, $X_{\max}$ on $n \geq 2$ samples has the following bound:*

$$\mathbb{P}[X_{\max} \geq \max\{c_2, c_1 s\} k \log n] \leq n^{-k}$$

Proof. Again, consider the following argument via Chernoff's bound [MU05, p.97-98]: for $x \geq s\log n$ and $X \sim p_\pi$ we have

$$\mathbb{P}[X \geq x] \leq \sup_{0 \leq \lambda \leq s\log n} \frac{(e\lambda)^x e^{-\lambda}}{x^x} \leq \frac{(es\log n)^x e^{-s\log n}}{x^x} = \exp(-s\log n + x(1 + \log(s\log n) - \log x))$$

Now, choose $c_0 = \max\{2 + s, e^2 s\}$. Then for $k \geq 1$ and $x = kc_0 \log n$ we have

$$\begin{aligned}
&- s\log n + (kc_0\log n)(1 + \log(s\log n) - \log(kc_0\log n)) \\
&= (\log n)(-s + kc_0(1 + \log s - \log k - \log c_0)) \\
&= (\log n)(-s + kc_0(1 - \log k - 2)) \\
&\leq (\log n)(-s - k(2+s)) \leq \ (\log n)(-2k) \leq \ (\log n)(-(k+1))
\end{aligned} \tag{77}$$

Therefore $\mathbb{P}[X \geq c_0 k \log n] \leq n^{-(k+1)}$.

Take $c_3 = c_0(1 + \frac{1}{\log 2})$, we have $1 + c_0 k \log n \leq c_3 k \log n$ for all $k \geq 1$. Therefore, union bound gives $\mathbb{P}[1 + X_{\max} \geq c_3 k \log n] \leq n\mathbb{P}[1 + X \geq c_3 k \log n] \leq n\mathbb{P}[X \geq c_0 k \log n] \leq n^{-k}$. It then follows that we can take $c_1 = e^2(1 + \frac{1}{\log 2})$ and $c_2 = 6(1 + \frac{1}{\log 2})$. □

Lemma 12. *Consider a random variable W. If there exists a function $p(n)$ such that for all integers $c \geq 1$, $\mathbb{P}(W \geq cp(n)) \leq n^{-c}$, then for each integer $m \geq 1$ there exists a constant $c(m)$ such that for all $n \geq 2$,*

$$\mathbb{E}[W^m \mathbf{1}_{\{W \geq p(n)\}}] \leq \left(2^m + \frac{3^m m!}{(\log n)^{m+1}}\right) \frac{p(n)^m}{n}$$

Proof of Lemma 12. Denote the event $E_k = \{kp(n) \leq W \leq (k+1)p(n)\}$, then for all $n \geq 2$, we consider the expansion of $P(m, n)$ as per the claim to get

$$\mathbb{E}[W^m \mathbf{1}_{\{W \geq p(n)\}}] = \sum_{k=1}^{\infty} \mathbb{E}[W^m \mathbf{1}_{\{E_k\}}] \leq (p(n))^m \sum_{k=1}^{\infty} \frac{(k+1)^m}{n^k} \leq \frac{(p(n))^m}{n} \left(2^m + 3^m \sum_{k=2}^{\infty} \frac{(k-1)^m}{n^{k-1}}\right) \tag{78}$$

Using the Gamma integration we bound the last term in the above display using

$$\sum_{k=2}^{\infty} \frac{(k-1)^m}{n^{k-1}} \leq \int_0^{\infty} x^m n^{-x} dx = \int_0^{\infty} x^m e^{-x\log n} dx = \frac{m!}{(\log n)^{m+1}}.$$

Plugging this bound back in (78) finishes the proof. □

Lemma 13. *Given $X_1, \cdots, X_n \overset{iid}{\sim} p_\pi \triangleq \mathrm{Poi} \circ \pi$. Let $k \geq 1$ be an integer. Then there exist constant $c_0(k), c_1, c_2, c_3, c_4$ such that:*

- $\mathbb{E}[(1 + X_{\max})^k] \leq c_0(k)(\max\{c_1, c_2 h\} \frac{\log n}{\log \log n})^k$ *for all* $\pi \in \mathcal{P}([0, h])$.
- $\mathbb{E}[(1 + X_{\max})^k] \leq c_0(k)(\max\{c_3, c_4 s\} \log n)^k$ *for all* $\pi \in \mathcal{P}([0, s \log n])$.

Proof. For $\pi \in \mathcal{P}([0, h])$, choose c_1, c_2 according to Lemma 10 and use Lemma 12 to obtain the constant $c_0(k) \triangleq (2^k + 2^k k!)$ with $p(n) \triangleq \max\{c_1, c_2 h\} \frac{\log n}{\log \log n}$ and $W = 1 + X_{\max}$. For $\pi \in \mathcal{P}([0, s \log n])$, choose c_3, c_4 according to Lemma 11 and use Lemma 12 with $p(n) \triangleq \max\{c_3, c_4 s\} \log n$ and $W = 1 + X_{\max}$. □

Proof of Lemma 9. We note that conditioned on $\theta_1, \cdots, \theta_d$, the coordinates $X_1, \cdots, X_d$ are independent (distributed as $X_i \sim \mathrm{Poi}(\theta_i)$). It then follows that

$$\mathbb{E}\left[(1 + X_{j,\max})^\beta \prod_{\substack{k=1 \\ k \neq j}}^{d} (1 + X_{k,\max}) \mid \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_n\right] = \prod_{i=1}^{d} \mathbb{E}\left[(1 + X_{i,\max})^{\beta_i} | \theta_{1i}, \cdots, \theta_{ni}\right]$$

where here β_i is β if $i = j$ and 1 otherwise.

For the bounded prior case, i.e. $\pi \in \mathcal{P}([0, h])^d$ for some $h > 0$, we may mimic the proof of Lemma 10 to obtain, for some absolute constant $c(h) \triangleq \max\{c_1, c_2 h\}$, $\mathbb{P}[1 + X_{i,\max} \geq k c(h) \frac{\log n}{\log \log n} \mid \theta_{1i}, \cdots, \theta_{ni}] \leq n^{-k}$ (given that $\theta \leq h$). Thus we may then adapt Lemma 12 to yield $\mathbb{E}[(1 + X_{i,\max})^{\beta_i} \mid \theta_{1i}, \cdots, \theta_{ni}] \leq c_0(\beta_i)(c(h) \frac{\log n}{\log \log n})^{\beta_i}$ for some absolute constant $c_0(\beta_i)$ that depends only on the exponents β_i. Since this inequality holds regardless of $\theta_{1i}, \cdots, \theta_{ni}$ (so long as they are in the range $[0, h]$), the desired bound now becomes

$$\mathbb{E}\left[(1 + X_{j,\max})^\beta \prod_{\substack{k=1 \\ k \neq j}}^{d} (1 + X_{k,\max})\right] \leq c_0(\beta) c_0(1)^{d-1} \left(c(h) \frac{\log n}{\log \log n}\right)^{d-1+\beta}$$

$$\leq c_0(\beta) \left(c(h) \max\{1, c_0(1)\} \frac{\log n}{\log \log n}\right)^{d-1+\beta}$$

Likewise, for the case $\pi \in ([0, s \log n]^d)$, we may mimic the proof of Lemma 11 to obtain, for some absolute constant $c'(s) \triangleq \max\{c_3, c_4 h\}$, $\mathbb{P}[1 + X_{i,\max} \geq k c(s) \log n \mid \theta_{1i}, \cdots, \theta_{ni}] \leq n^{-k}$. Using Lemma 12 again, $\mathbb{E}[(1 + X_{i,\max})^{\beta_i} \mid \theta_{1i}, \cdots, \theta_{ni}] \leq c_0(\beta_i)(c'(s) \log n)^{\beta_i}$. Considering all $\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_n$ we then get

$$\mathbb{E}\left[(1 + X_{j,\max})^\beta \prod_{\substack{k=1 \\ k \neq j}}^{d} (1 + X_{k,\max})\right] \leq c_0(\beta) \left(c'(s) \max\{1, c_0(1)\} \log n\right)^{d-1+\beta}$$

□

B Proof of technical results

Proof of Lemma 1

Throughout the solution, for $s \le t$ we denote $m(s,t) \triangleq \frac{\sum_{i=s}^{t} w_i}{\sum_{i=s}^{t} v_i}$, where $m(s,t) = \infty$ if $v_i = 0$ for $s \le i \le t$. Denote, also, the cost function $G(f) \triangleq \sum_{i=1}^{n} v_i f(a_i)^2 - 2w_i f(a_i)$. We restrict our attention to establishing $\widehat{f}_{\mathsf{erm}}(a_1)$; the rest follows similarly. Let i_2 be the maximum index such that $\widehat{f}_{\mathsf{erm}}(a_1) = \cdots = \widehat{f}_{\mathsf{erm}}(a_{i_2})$ for some $i_2 \ge 1$.

We first claim that $\widehat{f}_{\mathsf{erm}}(a_1) = m(1, a_{i_2})$. Indeed, for each real t, and integer $j = 1, \cdots, k$, we define the following function $f_{j,t}(a_i) \triangleq \begin{cases} \widehat{f}_{\mathsf{erm}}(a_i) + t & 1 \le i \le j \\ \widehat{f}_{\mathsf{erm}}(a_i) & \text{otherwise} \end{cases}$. Then by the maximality of i_2, for some small $\epsilon > 0$, $f_{i_2,t}$ is still monotone for some $t \in (-\epsilon, \epsilon)$. In addition,

$$\frac{\partial G(f_{j,t})}{\partial t} = \sum_{i=1}^{j} 2(v_i(\widehat{f}_{\mathsf{erm}}(a_i) + t) - w_i)\,. \tag{79}$$

Since $\widehat{f}_{\mathsf{erm}} = \operatorname{argmin} G(f)$, $\frac{\partial G(f_{i_2,t})}{\partial t}|_{t=0} = 0$. Therefore,

$$\widehat{f}_{\mathsf{erm}}(a_1) \sum_{i=1}^{i_2} v_i = \sum_{i=1}^{i_2} \widehat{f}_{\mathsf{erm}}(a_i) v_i = \sum_{i=1}^{i_2} w_i\,. \tag{80}$$

Since $\max\{v_i, w_i\} > 0$ and each v_i, w_i is nonnegative, we cannot have $\sum_{i=1}^{i_2} v_i = \sum_{i=1}^{i_2} w_i = 0$. It then follows that $\widehat{f}_{\mathsf{erm}}(a_1) = \frac{\sum_{i=1}^{i_2} w_i}{\sum_{i=1}^{i_2} v_i} = m(1, i_2)$.

It now remains to show that $m(1, i_2) \le m(1, j)$ for all $j = 1, \cdots, k$, and the inequality is strict for $j > i_2$. Now for any j with $1 \le j \le k$, for some small $\epsilon > 0$, $f_{j,t}$ is still monotone for some $t \in (-\epsilon, 0]$. Given also $\widehat{f}_{\mathsf{erm}} = \operatorname{argmin} G(f)$, $\frac{\partial G(f_{j,t})}{\partial t}|_{t=0} \le 0$. Since $\widehat{f}_{\mathsf{erm}}(a_i) \ge \widehat{f}_{\mathsf{erm}}(a_1)$ for all i, we have

$$\widehat{f}_{\mathsf{erm}}(a_1) \sum_{1 \le i \le j} v_i \le \sum_{1 \le i \le j} \widehat{f}_{\mathsf{erm}}(a_i) v_i \le \sum_{1 \le i \le j} w_i\,, \tag{81}$$

which implies that $m(1, j) \ge \widehat{f}_{\mathsf{erm}}(a_1) = m(1, i_2)$. To show that $m(1, j) > m(1, i_2)$ for all $j > i_2$, suppose otherwise that $m(1, j) = m(1, i_2)$ for some $j > i_2$. This means the inequality in (81) is an equality for this j. In particular,

$$\widehat{f}_{\mathsf{erm}}(a_1) \sum_{i=1}^{j} v_i = \sum_{i=1}^{j} \widehat{f}_{\mathsf{erm}}(a_i) v_i \tag{82}$$

In view of (80), from $\sum_{i=1}^{j} \widehat{f}_{\mathsf{erm}}(a_i) v_i = \sum_{i=1}^{j} w_i$ we have

$$\sum_{i=i_2+1}^{j} \widehat{f}_{\mathsf{erm}}(a_i) v_i = \sum_{i=i_2+1}^{j} w_i\,. \tag{83}$$

By the maximality of i_2, we have $\widehat{f}_{\mathsf{erm}}(a_i) > \widehat{f}_{\mathsf{erm}}(a_1)$ for all $i > i_2$. Given that $v_i \ge 0$ for all i, (82) then implies $v_i = 0$ for $i = i_2 + 1, \cdots, j$. This would imply that $\sum_{i=i_2+1}^{j} w_i = 0$, i.e. $w_i = 0$ for all $i = i_2 + 1, \cdots, j$. This contradicts $\max\{v_i, w_i\} > 0$ for each $i = 1, \cdots, n$.

Proof of Lemma 5

Recall that conditioned on X_1^n, $\epsilon(x) \sim 2 \cdot Binom(N(x), \frac{1}{2}) - N(x)$. Since $b > 1$, it then follows that

$$\begin{aligned}\mathbb{E}[\max\{\epsilon(x) - \frac{1}{b}N(x), 0\}] &= \mathbb{E}[(\epsilon(x) - \frac{1}{b}N(x))\mathbf{1}_{\{\epsilon(x) > \frac{1}{b}N(x)\}}] \\ &\le (1 - \frac{1}{b})N(x)\mathbb{P}[\epsilon(x) > \frac{1}{b}N(x)] \\ &\overset{(a)}{\le} (1 - \frac{1}{b})N(x)\exp(-N(x)D(\frac{1+\frac{1}{b}}{2}||\frac{1}{2})) \overset{(b)}{\le} \frac{1 - \frac{1}{b}}{e \cdot D(\frac{1+\frac{1}{b}}{2}||\frac{1}{2})}\end{aligned}$$

where (a) is from [PW22, Example 15.1, p.254] and (b) is using the fact that for all $a > 0$ and $y \ge 0$, $y\exp(-ay) \le \frac{1}{ae}$.

$O(X_{\max} \log X_{\max})$ Time Complexity Optimization

We now describe an algorithm based on stack that reduces the computation in Lemma 1 from $O(X_{\max}^2)$ to $O(X_{\max} \log X_{\max})$, with this log factor only used in sorting $\{(X, N(X))\}$ for $X = 0, 1, \cdots, X_{\max}$.

Let $W_1 < \cdots < W_k$ be the distinct elements in $\{X_1, \cdots, X_n\} \cup \{X_1 - 1, \cdots, X_n - 1\}$. We consider a stack S, initialized as $\emptyset$, with each element being the triple (I, w, t) where I denotes the interval of piecewise constancy, $w = \sum_{k \in I} N(W_k)$ and $t = \sum_{j \in I}(W_k + 1)N(W_k + 1)$. The invariant we are maintaining here is that the ratio $\frac{t}{w}$ is nondecreasing (this ratio is considered as $+\infty$ if $w = 0$).

At each step $t = 1, \cdots, k$ we do the following:

- Initialize $a \triangleq ([t, t], N(W_t), (W_t + 1)N(W_t + 1))$, the active element;
- Suppose, now, $a = (I, w, t)$. While the stack is nonempty and the top (most recent) element $a' = (I, w, t)$ $w't \le wt'$ (in particular, when $w, w' > 0$ we have the ratio $\frac{t}{w} \le \frac{t'}{w'}$), we pop a' from the stack, and set $a = (I \cup I', w + w', t + t')$.
- Push a onto the stack.

Then for each element in the form $([a, b], w, t)$ we have $\widehat{f}_{\mathsf{erm}}(x) = \frac{t}{w}$ for all $x = W_a, \cdots, W_b$. Notice that the largest element, W_k, has $N(W_k) > 0$, so the solution will always be well-formed.

To justify the time complexity, we see that there are at most k pushes into the stack. Each pop decreases the stack size by 1, so that cannot appear more than k times either. Assuming that each elementary computation (e.g. calculating $w't$ and wt') is $O(1)$, this stack operation takes $O(k)$. Since $k \le X_{\max}$, the claim follows.

Proof of Lemma 6

We will bound $\mathbb{P}[L_c(\epsilon) \ge k]$ for each integer $k \in [0, n]$. First, we see that $\sum_{i=1}^j \epsilon_i - cj \le (1-c)j$ (i.e. we'll only consider $j \ge k$) and for this sum to be positive we need $\sum_{i=1}^j \epsilon_i > cj$. If $X_j \sim Binom(j, \frac{1}{2})$ we have

$$\mathbb{P}[\sum_{i=1}^{j} \epsilon_i > cj] = \mathbb{P}[X_j > j(\frac{c+1}{2})] \le \exp(-jD(\frac{c+1}{2}||\frac{1}{2}))$$

by (i.e. Lemma 5). Now denoting $D(\frac{c+1}{2}||\frac{1}{2}) = c_1 > 0$, we have

$$\begin{aligned}\mathbb{P}[L_c(\epsilon) \geq k] &= \mathbb{P}[\exists j \geq k : \sum_{i=1}^{j} \epsilon_i - cj \geq k] \\ &\leq \sum_{j=k}^{n} \mathbb{P}[\sum_{i=1}^{j} \epsilon_i - cj \geq k] \leq \sum_{j=k}^{n} \exp(-jc_1) \leq \frac{\exp(-c_1 k)}{1-\exp(-c_1)} \end{aligned} \tag{84}$$

Therefore we have

$$\mathbb{E}[L_c(\epsilon)] \leq 1 + \sum_{k=0}^{n} \mathbb{P}[L_c(\epsilon) \geq k] \leq 1 + \sum_{k=0}^{n} \frac{\exp(-c_1 k)}{1-\exp(-c_1)} \leq 1 + \frac{1}{(1-\exp(-c_1))^2}.$$

as desired.