Universal priors: solving empirical Bayes via Bayesian inference and pretraining

Nick Cannella NVC9912@NYU.EDU
Courant Institute, New York University

Anzo Teh ANZOTEH@MIT.EDU
EECS, MIT

Yanjun Han YANJUNHAN@NYU.EDU
Courant Institute, New York University

Yury Polyanskiy YP@MIT.EDU
EECS, MIT

Editors: Steve Hanneke and Tor Lattimore

Abstract

We theoretically justify the recent empirical finding of Teh et al. (2025) that a transformer pretrained on synthetically generated data achieves strong performance on empirical Bayes (EB) problems. We take an indirect approach to this question: rather than analyzing the model architecture or training dynamics, we ask why a pretrained Bayes estimator, trained under a *prespecified training distribution*, can adapt to *arbitrary test distributions*. Focusing on Poisson EB problems, we identify the existence of *universal priors* such that training under these priors yields a near-optimal regret bound of $\widetilde{O}(\frac{1}{n})$ uniformly over all test distributions. Our analysis leverages the classical phenomenon of *posterior contraction* in Bayesian statistics, showing that the pretrained Bayes estimator adapts to unknown test distributions precisely through posterior contraction. This perspective also explains the phenomenon of *length generalization*, in which the test sequence length exceeds the training length, as the model performs Bayesian inference using a fractional posterior.

Keywords: Empirical Bayes, regret bound, transformer, length generalization.

1. Introduction

Consider the following empirical Bayes (EB) task in the Poisson model: let $\theta_1, \dots, \theta_n$ be i.i.d. drawn from some unknown prior G_0 supported on $[0, A]$, and the observations X^n be conditionally independent with $X_i \sim \mathrm{Poi}(\theta_i)$ given θ^n. Here we assume knowledge of A, but do not impose any condition on the prior G_0 (not even continuity or smoothness). The target of empirical Bayes is to propose an estimator $\widehat{\theta}^n = \widehat{\theta}^n(X^n)$ that is nearly optimal *on every problem instance*, i.e., achieves competitive performance compared to the Bayes estimator with the oracle knowledge of G_0. The standard notion in empirical Bayes to quantify the estimator performance is the *regret*, defined as the excess MSE over the Bayes risk:

$$\begin{aligned}\mathsf{Regret}(\widehat{\theta}^n; G_0) &= \mathbb{E}_{G_0}\left[\frac{1}{n}\|\widehat{\theta}^n(X^n) - \theta^n\|_2^2 - \min_{\theta^\star(\cdot)} \frac{1}{n}\|\theta^\star(X^n) - \theta^n\|_2^2\right] \\ &= \mathbb{E}_{G_0}\left[\frac{1}{n}\|\widehat{\theta}^n - \theta_{G_0}(X^n)\|_2^2\right], \end{aligned} \tag{1}$$

where $\theta_{G_0}(X_i) = \mathbb{E}_{G_0}[\theta_i|X_i]$ is the Bayes estimator (posterior mean) with the knowledge of G_0, and $\theta_{G_0}(X^n) = (\theta_{G_0}(X_1), \dots, \theta_{G_0}(X_n))$. A small regret $\mathsf{Regret}(\widehat{\theta}^n) = o(1)$ means that the Bayes risk can be asymptotically attained by the legal estimator $\widehat{\theta}^n$. Compared with classical statistical estimators (like the MLE), empirical Bayes estimators usually enjoy a much better empirical performance due to instance-wise guarantees and implicit adaptations to the prior structure (Robbins, 1951, 1956; Jiang and Zhang, 2009; Han et al., 2025).

There exist several ways to solve empirical Bayes problems in the literature. Specializing to the Poisson empirical Bayes model, the earliest example is Robbins' estimator based on f-modeling (i.e., mimicking the form of the Bayes estimator $\theta_{G_0}(X)$). A numerically more stable approach is the g-modeling, which learns a prior from data and uses the Bayes estimator under the learned prior. A notable example for learning the prior is the nonparametric MLE (NPMLE), as well as a broader class of minimum distance estimators (Vandegar et al., 2021; Jana et al., 2025). Finally, a modern approach is to use empirical risk minimization (ERM), which minimizes a properly constructed loss function evaluated on X^n over a suitably chosen function class (Barbehenn and Zhao, 2022; Jana et al., 2023). With the exception of Robbins' estimator, all these estimators solve an optimization program at test time, meaning that these programs depend on X^n.

The recent work (Teh et al., 2025) proposes to solve empirical Bayes problems via a different strategy of using a *pretrained estimator*. Unlike the previous ERM approach which trains a separate model for each sample X^n, a pretrained model learns a function $\widehat{\theta}^n(\cdot)$ from a large pool of properly generated training data and enables extremely fast computation at test time (namely, applying $\widehat{\theta}^n(\cdot)$ directly to X^n). This strategy is inspired by recent successes of TabPFN (Hollmann et al., 2023, 2025), which similarly applies a single pretrained model across diverse categorical datasets. This therefore achieves *cost amortization* in a similar spirit with amortized inference (see, e.g. Zammit-Mangion et al. (2025)), where after pretraining with synthetic data, a transformer can perform quick inference on millions of batches of new observations. As shown via experimental results in Teh et al. (2025) for Poisson EB, a well-trained transformer indeed achieves a better performance than the state-of-the-art NPMLE-based estimator at only a tiny fraction of inference time.

In this paper, we study the statistical aspect of the approach in Teh et al. (2025). Specifically, we ask the following question:

> Why can a pretrained Bayes estimator trained under a *prespecified training distribution* adapt to *all possible* test distributions?

For Poisson EB, a pretrained estimator is constructed as follows: Given a large number of training batches $\{(\theta^{n,(m)}, X^{n,(m)})\}_{m=1}^M$ generated from a training prior $\theta^{n,(m)} \sim G_\Pi$ and Poisson model $X^{n,(m)}|\theta^{n,(m)} \sim \otimes_{i=1}^n \mathrm{Poi}(\theta_i^{(m)})$, the pretrained estimator is the following empirical risk minimizer:

$$\widehat{\theta}^n = \operatorname*{arg\,min}_{T:\mathbb{N}^n \to [0,A]^n} \frac{1}{M} \sum_{m=1}^M \|\theta^{n,(m)} - T(X^{n,(m)})\|_2^2. \tag{ERM}$$

At the population level $M \to \infty$, and assuming that the global minimizer of (ERM) is attained, this pretrained estimator $\widehat{\theta}^n$ is the posterior mean vector $\mathbb{E}_{G_\Pi}[\theta^n|X^n] = (\mathbb{E}_{G_\Pi}[\theta_1|X^n], \dots, \mathbb{E}_{G_\Pi}[\theta_n|X^n])$ under an n-dimensional prior G_Π for θ^n. We call such a training prior G_Π *universal* if this Bayes estimator achieves a vanishing regret over *all* test priors G_0, i.e.

$$\sup_{G_0 \in \mathcal{P}([0,A])} \mathsf{Regret}(\mathbb{E}_{G_\Pi}[\theta^n|X^n]; G_0) = o(1). \tag{2}$$

Algorithm 1: A pretrained empirical Bayes estimator via transformers

Input: Dimension parameter n, support parameter $A > 0$, number of training batches M.
Output: A pretrained estimator $\widehat{\theta}^n : \mathbb{N}^n \to [0, A]^n$.
Let $k \leftarrow \lceil c_0 \frac{\log n}{\log\log n} \rceil$; `// $c_0 > 0$ is a properly chosen hyperparameter`
for *batch* $m = 1, \cdots, M$ **do** `// Generate training data`
  Sample prior locations $\lambda_1, \ldots, \lambda_k \sim \mathrm{Unif}([0, A])$;
  Sample prior weights $(w_1, \ldots, w_k) \sim \mathsf{Dir}(1, \ldots, 1)$;
  Sample training outputs $\theta^{n,(m)} \sim (\sum_{j=1}^k w_j \delta_{\lambda_j})^{\otimes n}$;
  Sample training inputs $X^{n,(m)} \sim \otimes_{i=1}^n \mathrm{Poi}(\theta_i^{(m)})$;
end
Train a transformer T_ζ to minimize the training error, with parameter `// Training`

$$\widehat{\zeta} = \arg\min_{\zeta} \frac{1}{M} \sum_{m=1}^{M} \|\theta^{n,(m)} - \mathsf{T}_\zeta(X^{n,(m)})\|_2^2.$$

return *The trained transformer as the final estimator:* $\widehat{\theta}^n = \mathsf{T}_{\widehat{\zeta}}(X^n)$ *for any test data* X^n.

One might expect that universal priors must be carefully engineered to achieve this ambitious goal; interestingly, this turns out not to be the case. A natural way to construct G_Π is through a hierarchical prior or a *prior-on-prior (PoP)*: Let $\Pi \in \mathcal{P}(\mathcal{P}([0, A]))$[1] be a PoP (i.e. prior over probability distributions), we can generate $G \sim \Pi$ and $\theta^n | G \sim G^{\otimes n}$. In other words, the high-dimensional prior $G_\Pi = \mathbb{E}_\Pi[G^{\otimes n}]$ is an i.i.d. mixture induced by the PoP Π. The resulting Bayes estimator $\theta^n_\Pi(X^n) := \mathbb{E}_{G_\Pi}[\theta^n | X^n]$ will be called a *hierarchical Bayes estimator*, or a *nonparametric Bayes estimator* as the prior G_Π is usually nonparametric. We will also call a PoP Π *universal* if the corresponding prior G_Π is universal in the sense of (2). Under this hierarchical structure, it turns out that there exists a remarkably simple choice of a universal PoP Π.

Theorem 1.1 *Let the PoP* Π *be the distribution of a random prior* $G = \sum_{j=1}^k w_j \delta_{\lambda_j}$ *with* $\lambda_1, \ldots, \lambda_k \sim \mathrm{Unif}([0, A])$ *and* $(w_1, \ldots, w_k) \sim \mathsf{Dir}(1, \ldots, 1)$. *For* $k = \lceil c_0 \frac{\log n}{\log\log n} \rceil$ *with a large enough absolute constant* $c_0 > 0$*, it holds that*

$$\sup_{G_0 \in \mathcal{P}([0,A])} \mathsf{Regret}(\theta^n_\Pi; G_0) \le C \frac{\log^3 n}{n(\log\log n)^2},$$

where $C = C(A, c_0)$ *is an absolute constant depending only on* A *and* c_0.

The practical end-to-end implementation of θ^n_Π is displayed in Algorithm 1, where the class of transformers is used to find the sequence-to-sequence map T in (ERM). Theorem 1.1 shows that, if the approximation error of the pretrained transformer in Algorithm 1 to the Bayes estimator θ^n_Π turns out to be small, then this transformer achieves a vanishing regret for all test priors G_0.

1. Here and throughout, for a Polish space (i.e., a complete and separable metric space) (X, d), we write $\mathcal{P}(X)$ for the set of Borel probability measures on (X, d) equipped with the topology of weak convergence; under this topology, $\mathcal{P}(X)$ is itself a Polish space by Prokhorov's theorem.

We make some remarks on Theorem 1.1. First, the regret bound in Theorem 1.1 is near-optimal: For $A = \Theta(1)$, it was shown in Theorem 2 of Polyanskiy and Wu (2021) that the minimax regret is $\Theta(\frac{1}{n}(\frac{\log n}{\log\log n})^2)$. Therefore, although the estimator θ^n_Π is constructed in a Bayesian framework, it achieves near-optimal frequentist guarantees (off by a $\log n$ factor) even in the worst case.

Second, the universal prior G_Π, albeit remarkably simple, is a *high-dimensional* mixture of i.i.d. priors. Consequently, the posterior mean $\mathbb{E}_\Pi[\theta_1|X^n]$ depends on the entire X^n, not only X_1. In other words, the pretrained Bayes estimator is not separable, and this dependence turns out to be essential for Theorem 1.1 to hold. Since such expectations are difficult to compute classically, we employ transformers to approximate this sequence-to-sequence map.

Third, rather than analyzing the training dynamics or architecture-specific details of transformers, we adopt the key assumption that the transformers can approximate the Bayes estimator θ^n_Π, which we numerically verify in Section 4. We choose transformers (without positional encoding or masking) for their expressive power for sequence-to-sequence maps (Teh et al., 2025, Theorem 4.1, 4.2); (Furuya et al., 2025), ability to incorporate any sequence length, and permutation equivariance (see Section 3.1 for definition; θ^n_Π is also permutation-equivariant).

Finally, for approximate Bayes estimators close to θ^n_Π, Section 3.1 develops regret bounds that depend on the approximation error. In particular, such analysis shows that the same regret bound of Theorem 1.1 can be attained by the ERM with a finite number of batches M, though our current sufficient condition on M grows super-polynomially in n. This is consistent with the large data requirements observed in practical pretraining.

In the sequel, we present an in-depth analysis of the hierarchical Bayes estimator θ^n_Π and universal PoPs Π. First, we establish basic statistical results such as the existence of the least favorable PoP via the minimax theorem, and admissibility of hierarchical Bayes estimators while classical estimators (such as the NPMLE-based estimator) could be inadmissible. Second, based on the classical phenomenon of *posterior contraction* in Bayesian statistics, we will show the central result in our paper that *a broad class of prior-on-priors turns out to be universal*. Finally, drawing on the transformer structure, we explain why the pretrained transformer enjoys *length generalization*.

1.1. Least favorable PoP and admissibility

We establish basic properties of θ^n_Π in this section. Recall that θ^n_Π is the Bayes estimator (under the quadratic loss) in the following hierarchical Bayes model:

$$\begin{aligned} G &\sim \Pi, \\ \theta_i \mid G &\overset{\text{i.i.d.}}{\sim} G, \quad i = 1, \dots, n \\ X_i \mid \theta^n, G &\overset{\text{ind.}}{\sim} \mathrm{Poi}(\theta_i), \quad i = 1, \dots, n. \end{aligned} \tag{3}$$

Although the induced high-dimensional prior $G_\Pi = \mathbb{E}_\Pi[G^{\otimes n}]$ on θ^n is constrained to be an i.i.d. mixture, the following result shows that there exists a *least favorable PoP* $\Pi^\star$ such that the hierarchical Bayes estimator $\theta^n_{\Pi^\star}$ attains the minimax optimal regret.

Proposition 1.1 *For any $A > 0$ and $n \geq 1$, we have the following identity:*

$$\inf_{\widehat{\theta}^n} \sup_{G \in \mathcal{P}([0,A])} \mathsf{Regret}(\widehat{\theta}^n; G) = \sup_{\Pi \in \mathcal{P}(\mathcal{P}([0,A]))} \inf_{\widehat{\theta}^n} \mathbb{E}_{G\sim\Pi}[\mathsf{Regret}(\widehat{\theta}^n; G)].$$

Consequently, the saddle point gives a least favorable PoP $\Pi^\star$ such that the corresponding Bayes estimator $\theta^n_{\Pi^\star}$ achieves the minimax regret.

Proposition 1.1 is a minimax theorem; its proof verifies the compactness and continuity conditions, and is deferred to Section A.1. By the characterization of the minimax regret $\Theta(\frac{1}{n}(\frac{\log n}{\log\log n})^2)$ for Poisson EB (Polyanskiy and Wu, 2021), Proposition 1.1 immediately implies the existence of a universal PoP Π that attains (2). In addition, it justifies the i.i.d. mixture form of the high-dimensional prior G_Π used in pretraining, so the only task is to construct the PoP Π.

The next result concerns the admissibility of the hierarchical Bayes estimator θ^n_Π, with the classical notion of admissibility defined below for Poisson EB.

Definition 1 (Admissibility) *An estimator $\widehat{\theta}^n : \mathbb{N}^n \to [0,A]^n$ is called* inadmissible *if there exists another estimator $\widetilde{\theta}^n$ such that $\mathsf{Regret}(\widetilde{\theta}^n; G) \le \mathsf{Regret}(\widehat{\theta}^n; G)$ for all $G \in \mathcal{P}([0,A])$, and $\mathsf{Regret}(\widetilde{\theta}^n; G_0) < \mathsf{Regret}(\widehat{\theta}^n; G_0)$ for some $G_0 \in \mathcal{P}([0,A])$. If an estimator is not inadmissible, it is called* admissible.

We note that this definition remains equivalent if we replace the regret by the MSE $\mathbb{E}_{G_0}[\frac{1}{n}\|\widehat{\theta}^n - \theta^n\|_2^2]$, since the Bayes risk is independent of $\widehat{\theta}^n$. In the next result, we will compare the hierarchical Bayes estimator θ^n_Π with the NPMLE-based estimator

$$\widehat{\theta}^{\mathrm{NPMLE}}(X^n) = (\theta_{\widehat{G}}(X_1), \dots, \theta_{\widehat{G}}(X_n)), \quad \widehat{G} = \operatorname*{arg\,max}_{G\in\mathcal{P}([0,A])} \prod_{i=1}^n f_G(X_i),$$

where we recall that $\theta_G(x) = \mathbb{E}_G[\theta|X=x]$ is the 1-D Bayes estimator, and $f_G(x) = \mathbb{E}_{\theta\sim G}[\mathrm{Poi}(x;\theta)]$ is a shorthand for the marginal pmf of $X \sim \mathrm{Poi}(\theta)$ when $\theta \sim G$. Note that the support of the NPMLE $\widehat{G}$ is restricted to be $[0,A]$, since otherwise it is clearly inadmissible by truncating the output to $[0,A]^n$. The next result shows that while the hierarchical Bayes estimator θ^n_Π is admissible for most PoPs, the NPMLE-based estimator is not.

Proposition 1.2 *For all $\Pi \in \mathcal{P}(\mathcal{P}([0,A]))\backslash\{\delta_{\delta_0}\}$, the hierarchical Bayes estimator θ^n_Π is admissible. However, if $A \ge 3$ is an integer, then the NPMLE-based estimator $\widehat{\theta}^{\mathrm{NPMLE}}$ is inadmissible.*

1.2. Universal PoPs

While Proposition 1.1 justifies the pretrained approach in Algorithm 1, it still remains to identify PoPs that are approximately least favorable. This motivates the following definition of *thick* PoPs, which provide a sufficient condition for attaining a small worst-case regret.

Definition 2 (Thick PoP) *We call a PoP Π* thick *with rate function $E(\delta, r)$ if for every prior G_0 supported on $[0,A]$, there exists some G such that $\mathsf{TV}(f_{G_0}, f_G) \le \delta$,[2] and*

$$\Pi\left(\{G' : \chi^2(f_G\|f_{G'}) \le r^2\}\right) \ge e^{-E(\delta,r)}. \tag{4}$$

The term "thick" is motivated by the use of "thick priors" in Bickel and Kleijn (2012), referring to priors with continuous and strictly positive Lebesgue densities. Taking $G = G_0$, the main condition (4) states that the PoP Π puts a sufficient amount of mass near the true prior G_0. For technical reasons, "near" is measured through the χ^2 divergence between the marginal pmfs, and we include

2. Here $\mathsf{TV}(P,Q) = \frac{1}{2}\int|\mathrm{d}P - \mathrm{d}Q|$ and $\chi^2(P\|Q) = \int\frac{(\mathrm{d}P)^2}{\mathrm{d}Q} - 1$ denote the total variation distance and chi-squared divergence, respectively.

an additional prior G for an intermediate coupling. We note that the intuition behind (4) is standard in the literature of posterior contraction (Ghosal et al., 2000), commonly with χ^2 replaced by KL.

Our next result shows that thick PoPs (with proper rate function E) are universal, and establishes a regret bound on the hierarchical Bayes estimator under a thick PoP.

Theorem 1.2 *Let the PoP Π be thick with rate E, and $r_n > 0$ satisfy $E(n^{-2}, r_n) \le nr_n^2$. Then for an absolute constant $C = C(A) > 0$, the hierarchical Bayes estimator θ^n_Π satisfies*

$$\sup_{G_0 \in \mathcal{P}([0,A])} \mathsf{Regret}(\theta^n_\Pi; G_0) \le \frac{C \log n}{n \log\log n}\Big(\frac{\log n}{\log\log n} + nr_n^2\Big).$$

Theorem 1.1 is then a special case of Theorem 1.2 in conjunction with the following lemma.

Lemma 1.1 *For a large enough hyperparameter $c_0 > 0$, the PoP used in Algorithm 1 is thick with $E(n^{-2}, r) = O(\frac{\log^2 n}{\log\log n})$ for $r^2 \ge \frac{1}{n}$. In particular, we can choose $r_n^2 = O(\frac{\log^2 n}{n \log\log n})$.*

We sketch the proof idea of Theorem 1.2 via *posterior contraction* in Bayesian statistics (Ghosal et al., 2000; Shen and Wasserman, 2001). By Lemma 2.3, the i-th coordinate of θ^n_Π equals $\theta_{G_i}(X_i)$, the one-dimensional Bayes estimator under the posterior mean $G_i = \mathbb{E}_{G \sim \Pi_{G|X_{\setminus i}}}[G]$. For the true data $X^n \sim f_{G_0}^{\otimes n}$, posterior contraction implies that G_i concentrates around the test distribution G_0, so that the pretrained estimator $\theta_{G_i}(X_i)$ adapts to the true Bayes estimator $\theta_{G_0}(X_i)$ under the unknown prior G_0. These intuitions will be made precise in Section 2.

1.3. Length generalization

The hierarchical Bayes estimator θ^n_Π is a map from $\mathbb{N}^n$ to $\mathbb{R}^n$, but a transformer can take variable-length inputs and gives a map from $\mathbb{N}^m$ to $\mathbb{R}^m$ for every $m \in \mathbb{N}$. Therefore, if we use a transformer to represent the hierarchical Bayes estimator θ^n_Π on length n, it can also take input sequence X^m of a longer length $m > n$. This is *length generalization* of transformers: for Poisson EB, the empirical evidence in Teh et al. (2025) shows that even if the transformer is solely trained on training length n, it achieves small regret on larger test lengths $n_{\mathsf{test}} > n$. Our perspective of Bayesian inference and posterior contraction framework provides a theoretical explanation for this phenomenon.

To this end, we need to generalize our definition of the hierarchical Bayes estimator $\theta_\Pi : \mathbb{N}^n \to \mathbb{R}^n$ into a map that can incorporate any input sequence length. We use the following observations on length generalization for transformers (Furuya et al., 2025, Section 2.2). If a transformer architecture only consists of softmax attention and token-wise operations[3] (e.g. input embedding, MLP, residual connection, and layer normalization), then it is *permutation equivariant*, i.e. $\pi \circ \mathsf{T}(X^n) = \mathsf{T}(\pi \circ X^n)$ for all $\pi \in S_n$. As a result of permutation equivariance, its i-th output takes the form $\mathsf{T}_i(X^n) = f_n(X_i, \mu_n)$ for some $f_n : \mathcal{X} \times \mathcal{P}(\mathcal{X}) \to \mathcal{Y}$ and the *empirical distribution* μ_n of all inputs X^n. Furthermore, under this length generalization model for transformers with softmax attention and token-wise operations, the induced map can be represented by a single function $f_n \equiv f$ independent of n. Therefore, a single function $f : \mathcal{X} \times \mathcal{P}(\mathcal{X}) \to \mathcal{Y}$ encodes the transformer map for general input length: given input $X^{n_{\mathsf{test}}}$, the i-th output is $\mathsf{T}_i(X^{n_{\mathsf{test}}}) = f(X_i, \mu_{n_{\mathsf{test}}})$. For example, this implies that if $\mathsf{T}(X^n) = Y^n$, then $\mathsf{T}(X^n, \dots, X^n) = (Y^n, \dots, Y^n)$.

Applying the above length generalization model to θ^n_Π, we have the following theorem.

3. Crucially, this transformer does not contain positional encoding, causal masking, or the final softmax layer to convert logits into a probability distribution.

Theorem 1.3 *For any PoP Π and training length n, the hierarchical Bayes estimator θ^n_Π is permutation equivariant, so that there exists a map $f_{\Pi,n} : \mathbb{N} \times \mathcal{P}(\mathbb{N}) \to \mathbb{R}_+$, determined solely by Π and n, such that $\theta_{\Pi,i}(X^n) = f_{\Pi,n}(X_i, \mu_n)$ for every input $X^n \in \mathbb{N}^n$, with $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$.*

Next, suppose that Π is thick with rate E. For any test sequence length $n_{\mathsf{test}} \geq n$, let $r > 0$ satisfy $E(n_{\mathsf{test}}^{-2}, r) \vee 1 \leq nr^2$. The following regret bound holds for the estimator $f^{n_{\mathsf{test}}}_{\Pi,n}(X^{n_{\mathsf{test}}}) := (f_{\Pi,n}(X_1, \mu_{n_{\mathsf{test}}}), \ldots, f_{\Pi,n}(X_{n_{\mathsf{test}}}, \mu_{n_{\mathsf{test}}}))$ with an absolute constant $C = C(A)$:

$$\sup_{G_0 \in \mathcal{P}([0,A])} \mathsf{Regret}(f^{n_{\mathsf{test}}}_{\Pi,n}; G_0) \leq \frac{C \log n_{\mathsf{test}}}{\log\log n_{\mathsf{test}}} \left(\frac{\log n_{\mathsf{test}}}{n_{\mathsf{test}} \log\log n_{\mathsf{test}}} + r^2 \right).$$

The first part of Theorem 1.3 shows that the hierarchical Bayes estimator θ^n_Π under every PoP can be represented by some function $f_{\Pi,n}$. The second part illustrates both strengths and weaknesses of length generalization for a pretrained transformer that learns the map $f_{\Pi,n}$: it still achieves a vanishing regret on longer lengths as the training length $n \to \infty$, but the regret remains $\widetilde{O}(r^2) = \widetilde{O}(\frac{1}{n})$ rather than $\widetilde{O}(\frac{1}{n_{\mathsf{test}}})$. This result conforms to the empirical finding in Teh et al. (2025) that the pretrained transformer enjoys a low regret for all $n_{\mathsf{test}} \geq n$, but the regret stops decreasing when $n_{\mathsf{test}} \gg n$; see Figure 1 for an illustration. Intuitively, this is because the map $f_{\Pi,n}$ depends on the training length n, and is thus not fully adaptive to a longer sequence length $n_{\mathsf{test}} \geq n$ at test time.

The proof of Theorem 1.3 again relies on posterior contraction, but now for *fractional/generalized posteriors*. For $n_{\mathsf{test}} \geq n$, the new estimator $f^{n_{\mathsf{test}}}_{\Pi,n}$ is no longer a hierarchical Bayes estimator, but it performs Bayesian inference using an *α-posterior*, i.e. with posterior update

$$\Pi^\alpha(\mathrm{d}G|X^{n_{\mathsf{test}}}) \propto \Pi(\mathrm{d}G)\Big(\prod_{i=1}^{n_{\mathsf{test}}} f_G(X_i)\Big)^\alpha, \quad \text{with } \alpha = \frac{n}{n_{\mathsf{test}}} \leq 1. \tag{5}$$

In Section 2, we develop posterior contraction results such that the α-posterior $\Pi^\alpha(\mathrm{d}G|X^{n_{\mathsf{test}}})$ still concentrates around G_0 for $X^{n_{\mathsf{test}}} \sim f_{G_0}^{\otimes n_{\mathsf{test}}}$, with a rate depending now on α. We note that such fractional posteriors have appeared previously in the Bayesian literature on model misspecification (Bhattacharya et al., 2019; Medina et al., 2022). Moreover, although (5) only holds for the ideal estimator $f^{n_{\mathsf{test}}}_{\Pi,n}$, for a pretrained transformer with training length n and test length n_{test}, our numerical experiments in Section 4.2 demonstrate that its output is indeed close to a hierarchical Bayes model performing $\frac{n}{n_{\mathsf{test}}}$-posterior inference.

1.4. Related work

Empirical Bayes. EB was introduced alongside compound decision theory (Robbins, 1951, 1956), motivated by the idea that estimating a sequence of parameters can achieve lower risk when each component is allowed to depend on the entire sequence, a phenomenon classically illustrated by the James–Stein estimator (Stein, 1956; James and Stein, 1961). For the Poisson model, known estimators are based either on the Tweedie's formula (Robbins, 1956), posterior density estimation (Kiefer and Wolfowitz, 1956; Lindsay, 1983; Shen and Wu, 2026; Jana et al., 2025; Han et al., 2025), or ERM (Jana et al., 2023). The former two approaches are also dubbed f-modeling and g-modeling, respectively (Efron, 2019). Neural approaches have also been explored to approximate the maximum likelihood via generative modeling (Wang et al., 2019; Vandegar et al., 2021); see also (Ghosh et al., 2025; Chen and Cui, 2025) for the normal means model. For the regret in the Poisson-EB problem, tight bounds $\Theta(\frac{1}{n}(\frac{\log n}{\log\log n})^2)$ and $\Theta(\frac{\log^3 n}{n})$ are established for compactly supported and

subexponential priors (Brown et al., 2013; Polyanskiy and Wu, 2021; Jana et al., 2023, 2025), with non-trivial extensions to unbounded supports (Shen and Wu, 2026). There is also a rich line of work on the optimal regret in the normal means model (Jiang and Zhang, 2009; Saha and Guntuboyina, 2020; Polyanskiy and Wu, 2021; Soloff et al., 2025; Chen and Wu, 2026). A recent work (Kang et al., 2026) also explores function estimation in the empirical Bayes setting, which we discuss in Theorem 3.3. We also refer to (Efron, 2024; Ignatiadis and Sen, 2025) for surveys on EB.

Meta-learning. Our use of a pretrained transformer, trained over a distribution of priors and applied to test sequences without per-instance optimization at test time, falls within the framework of *meta-learning* (Hospedales et al., 2021). A rich line of recent work studies in-context learning of transformers, a form of meta learning usually instantiated through autoregressive next-token prediction (Garg et al., 2022; Bai et al., 2023; Akyürek et al., 2023; Von Oswald et al., 2023). In particular, connections between in-context learning and Bayesian inference have been observed both empirically (Müller et al., 2022; Panwar et al., 2024; Aggarwal et al., 2025) and theoretically (Xie et al., 2022; Wakayama and Suzuki, 2025; Ma et al., 2025b). This connection inspired the seminal TabPFN framework (Hollmann et al., 2023, 2025) and several follow-up works on amortized inference for tabular data (Ma et al., 2025a; Reuter et al., 2025; Mittal et al., 2025a,b). Similar posterior-contraction ideas have appeared in autoregressive (Xie et al., 2022; Ma et al., 2025b) and diffusion settings (Jia et al., 2026); by contrast, our EB application views the transformer as a sequence-to-sequence map (Teh et al., 2025). This Bayesian view of this sequence-to-sequence map also gives a statistical explanation of length generalization (Anil et al., 2022; Peng et al., 2024; Zhou et al., 2024a,b) through fractional posteriors, rather than transformer-specific architectural or algorithmic mechanisms (Ahuja and Mansouri, 2024; Izzo et al., 2026; Huang et al., 2025).

"Bayes" empirical Bayes. The use of a hierarchical/nonparametric Bayes model (3) to solve EB falls into the framework of "Bayes empirical Bayes" by Deely and Lindley (1981), where a line of work (Antoniak, 1974; Gu and Koenker, 2017) chooses the Dirichlet Process (DP) (Ferguson, 1973) as the prior. Statistical guarantees for hierarchical Bayes estimators in EB date back to the asymptotic optimality results of Datta (1991), with related developments in Petrone et al. (2014); Rizzelli et al. (2024); non-asymptotic guarantees were only established in our work and the independent concurrent work of Ignatiadis and Kankanala (2026). Both works establish regret guarantees via posterior contraction, focusing on Poisson and Gaussian EB problems, respectively. The two papers differ conceptually in their motivations. Ignatiadis and Kankanala (2026) study the optimality of nonparametric Bayes estimators under a DP prior, with computation carried out by a Gibbs sampler (Neal, 2000). By contrast, our motivation is more ML-oriented: we focus on pretrained transformers, where computation is amortized through pretraining, and thereby establish length generalization results. Other minor differences include that we establish minimax theorems in Proposition 1.1, whereas their work extends posterior contraction arguments to compound decision problems; we also note that our admissibility results for Poisson EB in Proposition 1.2 are motivated by theirs.

2. Analysis via posterior contraction

2.1. Preliminaries

We first provide some preliminaries on the Poisson mixture model and the hierarchical Bayes model induced by a general training PoP Π. For $\theta \sim G$ and $X|\theta \sim \mathrm{Poi}(\theta)$, the Bayes estimator of θ under

the squared loss is the posterior mean, defined as

$$\theta_G(x) = \mathbb{E}_G[\theta|X = x] = (x+1)\frac{f_G(x+1)}{f_G(x)}, \tag{6}$$

where $f_G(x) = \int \mathrm{Poi}(x;\theta)G(\mathrm{d}\theta)$ is the marginal pmf of X. Here and throughout, we will use $\mathbb{E}_G$ to denote the expectation with respect to the prior G. The following result, taken from Jana et al. (2025, Lemma 4), is a regret-Hellinger inequality that relates the posterior mean difference $\theta_G - \theta_{G_0}$ to the Hellinger distance $H(f_G, f_{G_0})$.[4]

Lemma 2.1 *Let G_0, G be two priors supported on $[0, A]$, and $\varepsilon \in (0, e^{-e})$ be any real number. Then for an absolute constant $C = C(A) > 0$,*

$$\mathbb{E}_{X\sim f_{G_0}}\left[(\theta_G(X) - \theta_{G_0}(X))^2\right] \le C\Big(\frac{\log(1/\varepsilon)}{\log\log(1/\varepsilon)}H^2(f_G, f_{G_0}) + \varepsilon\Big).$$

For the class of Poisson mixtures $\mathcal{P} = \{f_G : \mathrm{supp}(G) \subseteq [0, A]\}$, we will use a metric entropy upper bound under the Hellinger metric. Let $N(\varepsilon, \mathcal{P}, H)$ denote the ε-covering number of $\mathcal{P}$ under the Hellinger metric (i.e. the minimum number of ε-balls to cover $\mathcal{P}$), we define the *local* covering number as $N_{\mathrm{loc}}(\varepsilon, \mathcal{P}, H) := \sup_{P_0\in\mathcal{P}} \sup_{\rho\ge\varepsilon} N(\rho, \mathcal{P} \cap \{P : H(P_0, P) \le 2\rho\}, H)$.

Lemma 2.2 *For $\mathcal{P} = \{f_G : \mathrm{supp}(G) \subseteq [0, A]\}$, there is an absolute constant $C = C(A) > 0$ with $\log N_{\mathrm{loc}}(\varepsilon, \mathcal{P}, H) \le C\frac{\log(1/\varepsilon)}{\log\log(1/\varepsilon)}$ for all $\varepsilon \in (0, e^{-e})$.*

The last lemma gives a simple leave-one-out structure of the hierarchical Bayes estimator θ^n_Π.

Lemma 2.3 *For $i \in [n]$, it holds that $\mathbb{E}_\Pi[\theta_i|X^n] = \mathbb{E}_{G\sim\Pi_{G|X^n}}[\theta_G(X_i)] = \theta_{G_i}(X_i)$. Here $G_i := \mathbb{E}_{G\sim\Pi_{G|X_{\setminus i}}}[G]$, and θ_G is the posterior mean defined in (6).*

2.2. Posterior contraction

The key technical step in proving Theorem 1.2 is the following posterior contraction lemma. Below we will state a general version not limited to Poisson EB which will later be applied to a Gaussian example in Section 3.3. Let the true data $X_1, \dots, X_n \sim f_{G_0}$ be i.i.d. with $G_0 \in \mathcal{G}$, where f_{G_0} is a general density which may or may not take the form of mixture distributions. In the definition of the Bayes estimator, let $G \sim \Pi$ and $X_1, \dots, X_n \mid G \sim f_G$, where Π is a prior distribution over $\mathcal{G}$ (a PoP in the hierarchical setting). Finally, we set $\mathcal{P} = \{f_G : G \in \mathcal{G}\}$, and note that the thickness definition in Definition 2 naturally extends to the prior Π over the density class $\mathcal{P}$.

Lemma 2.4 *Let Π be thick over $\mathcal{P}$ in the sense of Definition 2, with rate function E. Let $\varepsilon_n, r_n > 0$ satisfy $\log N_{\mathrm{loc}}(\varepsilon_n, \mathcal{P}, H) \le n\varepsilon_n^2$ and $E(n^{-2}, r_n) \le nr_n^2$. Then there exist absolute constants $c, C > 0$ such that for every $G_0 \in \mathcal{G}$ and every $\varepsilon^2 \ge \varepsilon_n^2 + r_n^2 + \frac{1}{n}$,*

$$f_{G_0}^{\otimes(n-1)}\Big(\Big\{H^2(f_{G_0}, \Pi_{X_n|X^{n-1}}) > C\varepsilon^2\Big\}\Big) \le n^{-1} + e^{-cn\varepsilon^2}.$$

Here the probability is with respect to $X^{n-1} \sim f_{G_0}^{\otimes(n-1)}$, and $\Pi_{X_n|X^{n-1}}$ is the conditional distribution of X_n given X^{n-1} under the Bayes model with prior Π.

4. The squared Hellinger distance between P and Q is defined as $H^2(P, Q) := \int(\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q})^2$.

The Poisson EB setting is a special case of Lemma 2.4 by marginalizing out θ^n from the joint distribution Π_{G,θ^n,X^n}. Posterior contraction asserts that, given test data $X_1,\dots,X_{n-1} \sim f_{G_0}$, the posterior distribution $\Pi_{G|X^{n-1}}$ "concentrates" around the true prior G_0 with high probability, and this concentration is characterized through the induced pushforward measure from G to X (i.e. $\Pi_{X_n|X^{n-1}}$ is statistically close to f_{G_0}). This is precisely the statement of Lemma 2.4. The proof of Lemma 2.4 mirrors the classical posterior contraction arguments (Ghosal et al., 2000, 2008; Ghosal and Van der Vaart, 2017), with a few adaptations to obtain a high-probability statement; the proof details are postponed to the appendix.

Proof of Theorem 1.2, assuming Lemma 2.4. By the regret definition in (1) and symmetry, we focus on the last coordinate and write

$$\begin{aligned}\mathsf{Regret}(\theta^n_\Pi; G_0) &= \mathbb{E}_{X^n\sim f_{G_0}^{\otimes n}}\left[(\theta_{\Pi,n}(X^n) - \theta_{G_0}(X_n))^2\right] \\ &\le C\Big(\frac{\log n}{\log\log n}\cdot \mathbb{E}_{X^{n-1}\sim f_{G_0}^{\otimes(n-1)}}\left[H^2(f_{G_n}, f_{G_0})\right] + \frac{1}{n}\Big).\end{aligned}$$

Here the last step uses $\theta_{\Pi,n}(X^n) = \theta_{G_n}(X_n)$ by Lemma 2.3, and Lemma 2.1 applied to $G = G_n$ and $\varepsilon = \frac{1}{n}$. Since $G_n = \mathbb{E}_{G\sim\Pi_{G|X^{n-1}}}[G]$, we have $f_{G_n}(x_n) = \mathbb{E}_{G\sim\Pi_{G|X^{n-1}}}[f_G(x_n)] = \Pi_{X_n=x_n|X^{n-1}}$. By Lemma 2.2, we can choose $\varepsilon_n^2 = O(\frac{\log n}{n\log\log n})$. Finally, Theorem 1.2 follows from Lemma 2.4 and integrating the tails up to the deterministic upper bound $H^2 \le 2$.

2.3. Posterior contraction for the α-posterior

Similar to Lemma 2.4, we have the following posterior contraction result for α-posteriors in (5).

Lemma 2.5 *Let Π be thick over $\mathcal{P}$ in the sense of Definition 2, with rate function E, and $\alpha \in (0,1]$. Let $\varepsilon_n, r_{n,\alpha} > 0$ satisfy $\log N_{\mathrm{loc}}(\varepsilon_n, \mathcal{P}, H) \le n\varepsilon_n^2$ and $E(n^{-2}, r_{n,\alpha}) \le \alpha n r_{n,\alpha}^2$. Then there exist absolute constants $c, C > 0$ such that for every $G_0 \in \mathcal{G}$ and every $\varepsilon^2 \ge \varepsilon_n^2 + r_{n,\alpha}^2 + \frac{1}{\alpha n}$,*

$$f_{G_0}^{\otimes(n-1)}\left(\left\{\mathbb{E}_{G\sim\Pi^\alpha_{G|X^{n-1}}}[H^2(f_{G_0}, f_G)] > C\varepsilon^2\right\}\right) \le n^{-1} + e^{-c\alpha n\varepsilon^2},$$

where Π^α is the α-posterior defined in (5) with n_{test} replaced by n.

Compared with Lemma 2.4, the inequality $E(n^{-2}, r) \le \alpha n r^2$ involves an additional factor of α. When $E(n^{-2}, r) = \widetilde{O}(1)$, this gives $r_{n,\alpha}^2 = \widetilde{O}(\frac{1}{\alpha n})$, hence the squared Hellinger rate in posterior contraction increases from $\widetilde{O}(\frac{1}{n})$ to $\widetilde{O}(\frac{1}{\alpha n})$. For $(n,\alpha) = (n_{\mathsf{test}}, \frac{n}{n_{\mathsf{test}}})$, this rate is $\widetilde{O}(\frac{1}{n})$ rather than $\widetilde{O}(\frac{1}{n_{\mathsf{test}}})$. This additional factor of $\frac{1}{\alpha}$ is unsurprising, as it can be easily seen in the special example of normal location models with a normal prior. This change of scaling has also been observed in Bhattacharya et al. (2019, Theorem 3.2) on fractional posteriors with a different error probability.

Finally, using generalizations of Lemma 2.3 and Lemma 2.1 (cf. Lemma A.5 and Lemma A.6 in the appendix), we deduce the final regret bound from the Hellinger rate. This precisely explains why the regret upper bound for length generalization saturates at $\widetilde{O}(\frac{1}{n})$ even if $n_{\mathsf{test}} \gg n$.

3. Discussions

In this section, we provide discussions on pretraining on a finite number of batches, and extension beyond our Poisson EB setting with a bounded support of the prior. Deferred proofs are in Section C.

3.1. Finite number of batches

The population ERM solution in (ERM) coincides with the hierarchical Bayes estimator θ^n_Π only if $M \to \infty$. In practice, the number of batches M used in pretraining is finite and large. To this end, we consider a finite-sample ERM with a finite M and define

$$\widehat{\theta}^n = \arg\min_{f \in \mathcal{F}^{\mathrm{PE}}} \frac{1}{M} \sum_{m=1}^{M} \|\theta^{n,(m)} - f(X^{n,(m)})\|_2^2, \tag{7}$$

where $\mathcal{F}^{\mathrm{PE}}$ denotes the class of all permutation-equivariant functions $f(X^n)$, i.e. $\pi \circ f(X^n) = f(\pi \circ X^n)$ for all permutations $\pi \in S_n$. Note that most EB estimators (Robbins, NPMLE, ERM-monotone, etc.), as well as a transformer without positional encoding or masking, belong to $\mathcal{F}^{\mathrm{PE}}$; any estimator $\widehat{\theta}^n$ can also be symmetrized into a PE estimator $\widehat{\theta}^{n,\mathrm{PE}} = \frac{1}{n!}\sum_{\pi \in S_n} \pi^{-1} \circ \widehat{\theta}^n(\pi \circ X^n)$ without increasing the pointwise regret. The performance of this permutation-equivariant ERM in the Poisson model is summarized in the following result.

Lemma 3.1 *Let r_n be defined in Theorem 1.2. The estimator in (7) with $M \geq \exp(C(\frac{\log^2 n}{\log\log n} + nr_n^2))$ and an absolute constant $C = C(A)$ satisfies the same regret guarantee in Theorem 1.2.*

Lemma 3.1 shows that for the simple PoP in Algorithm 1 to attain the same regret of Theorem 1.1, $M = \exp(O(\frac{\log^2 n}{\log\log n}))$ batches suffices. This is superpolynomial in n, but it agrees with the practical intuition that pretraining typically requires a huge amount of training data. We leave it to future study if a better condition on M can be obtained by assuming different structures of $\widehat{\theta}^n$.

3.2. More discussions on Poisson EB

The compact support condition $G \in \mathcal{P}([0, A])$ can be generalized. First assume that G_0 is a subexponential prior, i.e. $G_0 \in \mathsf{SubE}(s) := \{G : \forall t \geq 0 : \mathbb{P}_G[\theta > t] < 2e^{-t/s}\}$. Since $G_0 \in \mathsf{SubE}(s)$ is effectively supported on $[0, O(\log n)]$, we modify Algorithm 1 to pick $k = \lceil c_0(s) \log n \rceil$ atoms for $G \in \mathsf{SubE}(s)$, chosen uniformly at random on $[0, c_1(s) \log n]$. The following regret bound holds for the hierarchical Bayes estimator θ^n_Π, which is near-optimal (Polyanskiy and Wu, 2021).

Theorem 3.1 *For large enough hyperparameters $c_0(s), c_1(s) > 0$, the hierarchical Bayes estimator θ^n_Π achieves a worst-case regret of $O_s(\frac{\log^4 n}{n})$ over all $G_0 \in \mathsf{SubE}(s)$.*

The next case is the regime where G_0 is supported on $[0, A]$ with $A = A_n \gg \log n$, where Shen and Wu (2026) showed that the optimal regret scales as $\widetilde{\Theta}(\frac{A^{1.5}}{n})$ for $A = O(n^2)$. This regret bound can be recovered by the hierarchical Bayes estimator with $k = \widetilde{\Theta}(\sqrt{A})$ atoms drawn uniformly at random on $[0, A]$, *provided that* one can establish $-\log \Pi_{X_n|X^{n-1}} = O(\mathsf{polylog}(n))$ with a sufficiently high probability over the randomness of the test data $X^n \sim f_{G_0}^{\otimes n}$. This condition is due to the regularization parameter $\rho = \exp(-O(\mathsf{polylog}(n)))$ used in the regret-Hellinger inequality (Shen and Wu, 2026, Theorem 3.3). We leave the verification of this condition as a conjecture.

Finally we comment on a potential route toward removing the extra $\log n$ factor in Theorem 1.1 relative to the optimal regret. Specializing to Poisson mixtures, a peeling argument into shells in Ghosal et al. (2000, Theorem 2.4) could improve the rate of posterior contraction, *provided that* one can show a local prior-doubling bound of the form $\log(\Pi(\{G : H^2(f_{G_0}, f_G) \leq Cr^2\})/\Pi(\{G :$

$\chi^2(f_{G_0}\|f_G) \le r^2\})) \le D = O(\frac{\log n}{\log\log n})$. Such a bound is plausible as f_G is effectively supported on $[0, D]$, but a rigorous proof faces two obstacles. First, unlike $H^2(f_{G_0}, f_G) \asymp \chi^2(f_{G_0}\|f_G)$ for Gaussian mixtures (Jia et al., 2023), this is generally false for Poisson mixtures due to the singularity of $\mathrm{Poi}(\theta)$ at $\theta = 0$. Second, although Gassiat and Van Handel (2014) establishes a local equivalence between $H(f_G, f_{G_0})$ and a weighted L_2 norm, a uniform non-asymptotic version of such a result is currently unavailable, which prevents us from carrying out the volume calculation.

3.3. Gaussian EB

In Gaussian EB, we have $\theta^n \sim G_0^{\otimes n}$ with some prior $G_0 \in \mathcal{P}([-A, A])$, and $X^n|\theta^n \sim \otimes_{i=1}^n \mathcal{N}(\theta_i, 1)$. Algorithm 1 is modified as follows: we still take $k = \lceil c_0 \frac{\log n}{\log\log n} \rceil$ for a large enough constant $c_0 > 0$, but now draw $\lambda_1, \dots, \lambda_k \sim \mathrm{Unif}([-A, A])$ and replace $\mathrm{Poi}(\lambda_j)$ by $\mathcal{N}(\lambda_j, 1)$. The following theorem summarizes the regret bound of the resulting hierarchical Bayes estimator θ^n_Π, which is near-optimal compared with the lower bound $\Omega(\frac{1}{n}(\frac{\log n}{\log\log n})^2)$ in Polyanskiy and Wu (2021).

Theorem 3.2 *For a large enough hyperparameter $c_0 > 0$, the hierarchical Bayes estimator θ^n_Π achieves a worst-case regret of $O_A(\frac{\log^3 n}{n(\log\log n)^2})$ over all $G_0 \in \mathcal{P}([-A, A])$.*

3.4. Function estimation

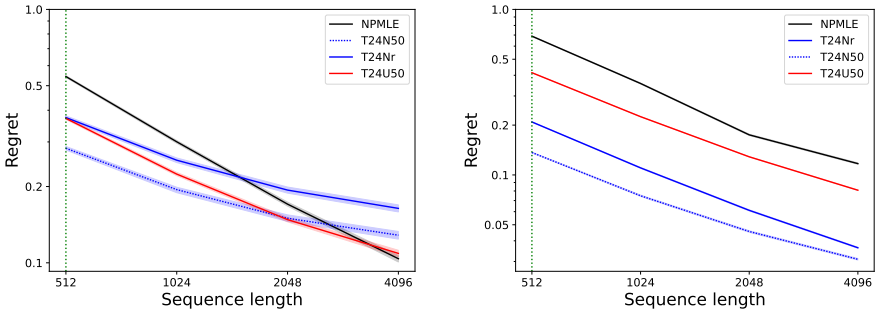
Motivated by the recent work (Kang et al., 2026), our result also extends to the setting where one is interested in estimating a function $g(\theta)$ of θ. In this case, the hierarchical Bayes estimator becomes $g^n_\Pi = \mathbb{E}_\Pi[g(\theta^n)|X^n]$, and the oracle estimator is $g_{G_0}(X) = \mathbb{E}_{G_0}[g(\theta)|X]$. For the special case of the polynomial function $g(\theta) = \theta^p$ with $p \in \mathbb{N}$, we establish the following regret bound which is near-optimal compared with the minimax regret of $\Theta(\frac{1}{n}(\frac{\log n}{\log\log n})^{p+1})$ in Kang et al. (2026).

Theorem 3.3 *Let the PoP Π be the same as Algorithm 1, with a large enough hyperparameter $c_0 > 0$. Then for $g(\theta) = \theta^p$, the hierarchical Bayes estimator g^n_Π achieves a worst-case regret of $O_{A,p}(\frac{1}{n}\frac{\log^{p+2} n}{(\log\log n)^{p+1}})$ over all $G_0 \in \mathcal{P}([0, A])$.*

4. Numerical experiments

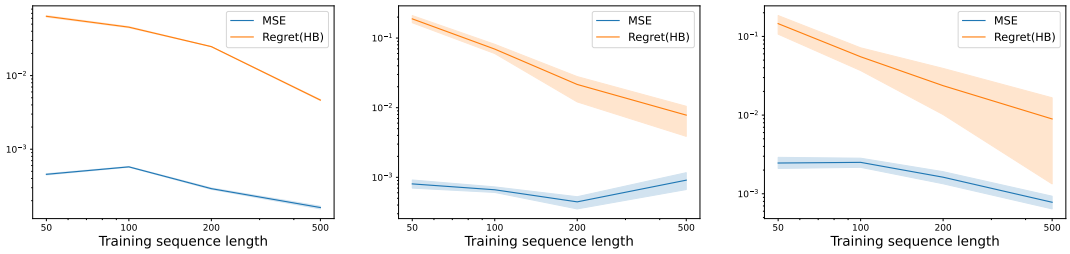
4.1. Regret performance and length generalization

Our first experiment validates the universality of our simple PoP in Algorithm 1 through transformers. Specifically, we pretrain a transformer T24U50 according to Algorithm 1 with $A = 50, k = 10$, and $n = 512$, where both the training procedure and transformer architecture are identical to those in Teh et al. (2025). The test priors are generated from the neural and multinomial PoPs also introduced in Teh et al. (2025), and we compare against the classical NPMLE-based estimator as well as two other pretrained transformers (T24N50 and T24Nr) from Teh et al. (2025). For each test sequence length n_{test}, Figure 1 displays the average regret of these estimators over 4096 runs, across different test priors. We make the following observations. First, when $n_{\mathsf{test}} = n$, all pretrained transformers outperform the NPMLE-based estimator, despite being a strong baseline, under both test priors. Second, as n_{test} increases, the regrets of all pretrained transformers continue to decrease, although this improvement may saturate for large n_{test}. These observations are consistent with both the theory of universal priors in Theorem 1.1 and the length generalization results in Theorem 1.3.



(*a*) Test prior from neural PoP (*b*) Test prior from multinomial PoP

Figure 1: Regrets of different estimators with different test lengths and test priors. For all transformers, the training length is fixed to be $n = 512$ (indicated by the vertical dotted green line).



(*a*) $m = 2$ (*b*) $m = 5$ (*c*) $m = 10$

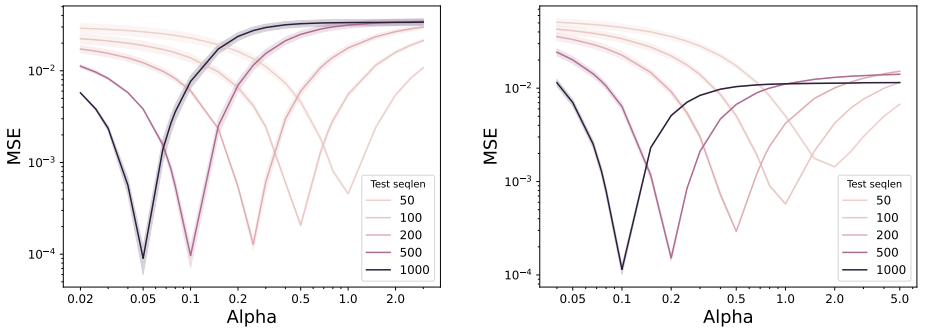
Figure 2: The regret of the hierarchical Bayes estimator, as well as its mean squared distance to the trained transformer, under simple training prior $G_{\Pi_m} = \frac{1}{m}\sum_{i=1}^{m} G_i^{\otimes n}$.

Finally, we emphasize that our experiments with the simple PoP are primarily intended as a proof of concept, and we do not claim its practical optimality; we believe that a more finely tuned PoP (as used by the other two transformers) can achieve better empirical performance.

4.2. Evidence on Bayesian inference

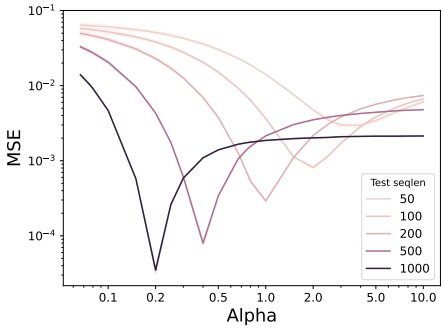
Our second set of experiments is designed to validate the assumptions underlying our main theorems, namely that (1) the pretrained transformer approximately implements the Bayes estimator under the training PoP, and (2) for length generalization, the pretrained transformer indeed performs Bayesian inference under α-posteriors.

For (1), we compare our pretrained transformer with the true hierarchical Bayes (HB) estimator θ^n_{Π}, in scenarios where finding θ^n_{Π} is computationally easy. Specifically, we choose a simple PoP $\Pi_m = \frac{1}{m}\sum_{i=1}^{m} \delta_{G_i}$ with small m, where $G_1, \ldots, G_m$ are randomly generated from our PoP in Algorithm 1, conditioned on the event that θ^n_{Π} does not perform uniformly well on $G_1, \ldots, G_m$. This conditioning ensures that the regret incurred by both θ^n_{Π} and the transformer is reasonably large and *much larger* than their difference; implementation details are deferred to the appendix. Figure 2 displays the regret of θ^n_{Π}, along with its mean squared distance to the trained transformer,

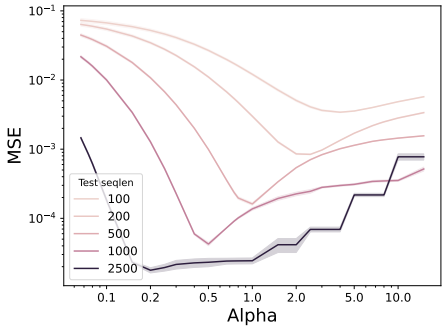


(*a*) $m = 2, n = 50$

(*b*) $m = 2, n = 100$



(*c*) $m = 2, n = 200$



(*d*) $m = 2, n = 500$

Figure 3: Plots of the mean squared distance between the pretrained transformer and the hierarchical Bayes estimator using various α-posteriors, with different training lengths n and test lengths n_{test}. This distance is indeed minimized at $\alpha \simeq \frac{n}{n_{\mathsf{test}}}$.

for $m \in \{2, 5, 10\}$ under the training PoP. We observe that, relative to the regret of θ^n_Π, the trained transformer is remarkably close to θ^n_Π, indicating that the pretrained transformer indeed performs approximate Bayesian inference under the training distribution.

For (2), we use the transformer trained on the above PoP Π_m, compute its outputs for various test lengths n_{test}, and compare them with the HB estimator based on α-posteriors for different values of α. Figure 3 presents the results for $m = 2$ across multiple training lengths n and test lengths n_{test}; results for larger m are deferred to the appendix. The quality of fit is striking: for each pair (n, n_{test}), the transformer output is extremely close to that of the HB estimator using α-posteriors, with $\alpha \simeq \frac{n}{n_{\mathsf{test}}}$. This demonstrates that our α-posterior conclusion in Theorem 1.3, although derived under a specific length generalization model, aligns well with the experimental findings.

Acknowledgement. We thank Nikolaos Ignatiadis for sharing the concurrent work (Ignatiadis and Kankanala, 2026), and anonymous reviewers for helpful comments on improving the organization of this paper. N.C. is supported by the Dean's Undergraduate Research Fund Grant at NYU. The authors acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing compute resources that have contributed to the experimental results reported in Section 4.

References

Naman Aggarwal, Siddhartha R Dalal, and Vishal Misra. The Bayesian geometry of transformer attention. *arXiv preprint arXiv:2512.22471*, 2025.

Kartik Ahuja and Amin Mansouri. On provable length and compositional generalization. *arXiv preprint arXiv:2402.04875*, 2024.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? Investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023.

Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. Exploring length generalization in large language models. *Advances in Neural Information Processing Systems*, 35: 38546–38556, 2022.

Charles E Antoniak. Mixtures of Dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics*, pages 1152–1174, 1974.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*, 36:57125–57211, 2023.

Alton Barbehenn and Sihai Dave Zhao. A nonparametric regression alternative to empirical Bayes approaches to simultaneous estimation. *arXiv preprint arXiv:2205.00336*, 2022.

Anirban Bhattacharya, Debdeep Pati, and Yun Yang. Bayesian Fractional Posteriors. *The Annals of Statistics*, 47(1):39–66, 2019.

PJ Bickel and BJK Kleijn. The semiparametric Bernstein-von Mises theorem. *The Annals of Statistics*, 40(1):206–237, 2012.

Lucien Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 65(2):181–237, 1983.

Lawrence D Brown, Eitan Greenshtein, and Ya'acov Ritov. The Poisson Compound Decision Problem Revisited. *Journal of the American Statistical Association*, pages 741–749, 2013.

Gongyu Chen and Ying Cui. Score-based diffusion modeling for nonparametric empirical Bayes in heteroscedastic Gaussian mixtures. In *Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Jiafeng Chen and Yihong Wu. Sharp regret-Hellinger bounds for Gaussian empirical Bayes via polynomial approximation. *arXiv preprint arXiv:2605.02070*, 2026.

Somnath Datta. Asymptotic optimality of bayes compound estimators in compact exponential families. *The Annals of Statistics*, 19(1):354–365, 1991.

JJ Deely and DV Lindley. Bayes empirical Bayes. *Journal of the American Statistical Association*, 76(376):833–841, 1981.

Richard M Dudley. *Real analysis and probability*. Chapman and Hall/CRC, 2018.

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

Bradley Efron. Empirical Bayes: Concepts and Methods. In *Handbook of Bayesian, Fiducial, and Frequentist Inference*, pages 8–34. Chapman and Hall/CRC, 2024.

Ky Fan. Minimax theorems. *Proceedings of the National Academy of Sciences*, 39(1):42–47, 1953.

Thomas S Ferguson. A Bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

Takashi Furuya, Maarten V de Hoop, and Gabriel Peyré. Transformers are Universal In-context Learners. In *The Thirteenth International Conference on Learning Representations*, 2025.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? A case study of simple function classes. *Advances in neural information processing systems*, 35:30583–30598, 2022.

Elisabeth Gassiat and Ramon Van Handel. The local geometry of finite mixtures. *Transactions of the American Mathematical Society*, 366(2):1047–1072, 2014.

Subhashis Ghosal and Aad W Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

Subhashis Ghosal, Jayanta K Ghosh, and Aad W Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, pages 500–531, 2000.

Subhashis Ghosal, Jüri Lember, and Aad van der Vaart. Nonparametric Bayesian model selection and averaging. *Electronic Journal of Statistics*, 2:63–89, 2008.

Sulagna Ghosh, Nikolaos Ignatiadis, Frederic Koehler, and Amber Lee. Stein's unbiased risk estimate and Hyvärinen's score matching. *arXiv preprint arXiv:2502.20123*, 2025.

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

Yanjun Han, Jonathan Niles-Weed, Yandi Shen, and Yihong Wu. Besting Good–Turing: Optimality of Non-Parametric Maximum Likelihood for Distribution Estimation. *arXiv preprint arXiv:2509.07355*, 2025.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. In *The Eleventh International Conference on Learning Representations*, 2023.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 5149–5169, 2021.

Xinting Huang, Andy Yang, Satwik Bhattamishra, Yash Sarrof, Andreas Krebs, Hattie Zhou, Preetum Nakkiran, and Michael Hahn. A formal framework for understanding length generalization in transformers. In *The Thirteenth International Conference on Learning Representations*, 2025.

Nikolaos Ignatiadis and Sid Kankanala. Compound decisions and empirical Bayes via Bayesian nonparametrics. *arXiv preprint arXiv:2602.20115*, 2026.

Nikolaos Ignatiadis and Bodhisattva Sen. Empirical Bayes. *Lecture notes*, 2025.

Zachary Izzo, Eshaan Nichani, and Jason D Lee. Quantitative bounds for length generalization in transformers. In *International Conference on Learning Representations*, 2026.

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

Soham Jana, Yury Polyanskiy, Anzo Z Teh, and Yihong Wu. Empirical Bayes via ERM and Rademacher complexities: the poisson model. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5199–5235. PMLR, 2023.

Soham Jana, Yury Polyanskiy, and Yihong Wu. Optimal empirical Bayes estimation for the Poisson model via minimum-distance methods. *Information and Inference: A Journal of the IMA*, 14(4): iaaf027, 2025.

Jing Jia, Wei Yuan, Sifan Liu, Liyue Shen, and Guanyang Wang. Weak diffusion priors can still achieve strong inverse-problem performance. *arXiv preprint arXiv:2601.22443*, 2026.

Zeyu Jia, Yury Polyanskiy, and Yihong Wu. Entropic characterization of optimal rates for learning gaussian mixtures. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4296–4335. PMLR, 2023.

Wenhua Jiang and Cun-Hui Zhang. General Maximum Likelihood Empirical Bayes Estimation of Normal Means. *The Annals of Statistics*, pages 1647–1684, 2009.

Benjamin Kang, Yury Polyanskiy, and Anzo Teh. Function estimation in the empirical Bayes setting. *arXiv preprint arXiv:2601.18689*, 2026.

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

Lucien Le Cam. Asymptotic Methods in Statistical Decision Theory. *Springer Series in Statistics*, 1986.

Erich Leo Lehmann and George Casella. *Theory of point estimation*. Springer, 1998.

Bruce G Lindsay. The geometry of mixture likelihoods: a general theory. *The Annals of Statistics*, pages 86–94, 1983.

Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Alex Labach, Jesse Cresswell, Keyvan Golestan, Guangwei Yu, Anthony L Caterini, and Maks Volkovs. TabDPT: Scaling tabular foundation models on real data. *Advances in Neural Information Processing Systems*, 38:172692–172722, 2025a.

Tianyi Ma, Tengyao Wang, and Richard J Samworth. Provable test-time adaptivity and distributional robustness of in-context learning. *arXiv preprint arXiv:2510.23254*, 2025b.

Marco Avella Medina, José Luis Montiel Olea, Cynthia Rush, and Amilcar Velez. On the Robustness to Misspecification of α-posteriors and Their Variational Approximations. *Journal of Machine Learning Research*, 23(147):1–51, 2022.

Sarthak Mittal, Yoshua Bengio, Nikolay Malkin, and Guillaume Lajoie. In-context parametric inference: Point or distribution estimators? *arXiv preprint arXiv:2502.11617*, 2025a.

Sarthak Mittal, Niels Leif Bracher, Guillaume Lajoie, Priyank Jaini, and Marcus Brubaker. Amortized in-context bayesian posterior estimation. *arXiv preprint arXiv:2502.06601*, 2025b.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do Bayesian inference. In *International Conference on Learning Representations*, 2022.

Radford M Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

Madhur Panwar, Kabir Ahuja, and Navin Goyal. In-Context Learning through the Bayesian Prism. In *The Twelfth International Conference on Learning Representations*, 2024.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Sonia Petrone, Judith Rousseau, and Catia Scricciolo. Bayes and empirical Bayes: do they merge? *Biometrika*, pages 285–302, 2014.

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

Arik Reuter, Tim GJ Rudner, Vincent Fortuin, and David Rügamer. Can transformers learn full Bayesian inference in context? In *Forty-second International Conference on Machine Learning*, 2025.

Stefano Rizzelli, Judith Rousseau, and Sonia Petrone. Empirical Bayes in Bayesian learning: understanding a common practice. *arXiv preprint arXiv:2402.19036*, 2024.

Herbert Robbins. Asymptotically subminimax solutions of compound statistical decision problems. *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, abs/1904.10040:131–149, 1951.

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.

Sujayam Saha and Adityanand Guntuboyina. On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising. *The Annals of Statistics*, 48(2):738–762, 2020.

Xiaotong Shen and Larry Wasserman. Rates of convergence of posterior distributions. *The Annals of Statistics*, 29(3):687–714, 2001.

Yandi Shen and Yihong Wu. Empirical Bayes estimation: When does g-modeling beat f-modeling in theory (and in practice)? *The Annals of Statistics*, 54(1):146–175, 2026.

Jake A Soloff, Adityanand Guntuboyina, and Bodhisattva Sen. Multivariate, heteroscedastic empirical Bayes via nonparametric maximum likelihood. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(1):1–32, 2025.

Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, volume 1, pages 197–206, 1956.

Anzo Teh, Mark Jabbour, and Yury Polyanskiy. Solving empirical Bayes via transformers. *arXiv preprint arXiv:2502.09844*, 2025.

Aad van der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media, 1996.

Maxime Vandegar, Michael Kagan, Antoine Wehenkel, and Gilles Louppe. Neural empirical Bayes: Source distribution estimation and its applications to simulation-based inference. In *International Conference on Artificial Intelligence and Statistics*, pages 2107–2115. PMLR, 2021.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.

Tomoya Wakayama and Taiji Suzuki. In-context learning is provably Bayesian inference: a generalization theory for meta-learning. *arXiv preprint arXiv:2510.10981*, 2025.

Yixin Wang, Andrew C Miller, and David M Blei. Comment: Variational autoencoders as empirical Bayes. 2019.

Yihong Wu and Sergio Verdú. Functional properties of minimum mean-square error and mutual information. *IEEE Transactions on Information Theory*, 58(3):1289–1301, 2011.

Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

Yihong Wu and Pengkun Yang. Optimal estimation of gaussian mixtures via denoised method of moments. *The Annals of Statistics*, 48(4):1981–2007, 2020.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An Explanation of In-context Learning as Implicit Bayesian Inference. In *International Conference on Learning Representations*, 2022.

Andrew Zammit-Mangion, Matthew Sainsbury-Dale, and Raphaël Huser. Neural methods for amortized inference. *Annual Review of Statistics and Its Application*, 12(1):311–335, 2025.

Hattie Zhou, Arwen Bradley, Etai Littwin, Noam Razin, Omid Saremi, Josh Susskind, Samy Bengio, and Preetum Nakkiran. What algorithms can transformers learn? A study in length generalization. In *The Twelfth International Conference on Learning Representations*, 2024a.

Yongchao Zhou, Uri Alon, Xinyun Chen, Xuezhi Wang, Rishabh Agarwal, and Denny Zhou. Transformers can achieve length generalization but not robustly. *arXiv preprint arXiv:2402.09371*, 2024b.

Appendix A. Deferred Proofs in Section 1

A.1. Proof of Proposition 1.1

We first utilize the continuity of regret as follows.

Lemma A.1 *For all estimators $\widehat{\theta}^n : \mathbb{N}^n \to [0, A]^n$ and $G \in \mathcal{P}([0, A])$, $\mathsf{Regret}(\widehat{\theta}^n; G)$ is continuous in G under the metric of weak convergence.*

The proof can be found in Appendix A.1.1. Next, note that since $[0, A]$ with the Euclidean metric is a compact Hausdorff space, by Prokhorov's Theorem, $\mathcal{P}([0, A])$ is compact under the metric of weak convergence (see, for example, Dudley (2018, Chapter 11.5)). This means the supremum can always be attained, i.e. for fixed estimator $\widehat{\theta}^n$,

$$\sup_{G\in\mathcal{P}([0,A])} \mathsf{Regret}(\widehat{\theta}^n; G) = \max_{G\in\mathcal{P}([0,A])} \mathsf{Regret}(\widehat{\theta}^n; G).$$

Next, we note that the LHS has the following equivalent formulation:

$$\inf_{\widehat{\theta}^n} \max_{G\in\mathcal{P}([0,A])} \mathsf{Regret}(\widehat{\theta}^n; G) = \inf_{\widehat{\theta}^n} \max_{\Pi\in\mathcal{P}(\mathcal{P}([0,A]))} \mathbb{E}_{G\sim\Pi}[\mathsf{Regret}(\widehat{\theta}^n; G)]$$

Clearly, $\mathbb{E}_{G\sim\Pi}[\mathsf{Regret}(\widehat{\theta}^n; G)]$ is convex in $\widehat{\theta}^n$ for fixed Π and affine in Π for fixed $\widehat{\theta}^n$. Therefore, combined with Lemma A.1, we use the Ky Fan minimax theorem (Fan, 1953, Theorem 2) to establish that

$$\inf_{\widehat{\theta}^n} \max_{\Pi\in\mathcal{P}(\mathcal{P}([0,A]))} \mathbb{E}_{G\sim\Pi}[\mathsf{Regret}(\widehat{\theta}^n; G)] = \max_{\Pi\in\mathcal{P}(\mathcal{P}([0,A]))} \inf_{\widehat{\theta}^n} \mathbb{E}_{G\sim\Pi}[\mathsf{Regret}(\widehat{\theta}^n; G)].$$

In addition, the final $\inf$ can be replaced by $\min$ as given Π, the minimizer is the hierarchical Bayes estimator $\theta^n_\Pi(x^n) = \mathbb{E}_\Pi[\theta^n | X^n = x^n]$.

A.1.1. PROOF OF LEMMA A.1

We note the following expansion of regret:

$$\begin{aligned}\mathsf{Regret}(\widehat{\theta}^n; G) &= \frac{1}{n}\mathbb{E}_{X^n\sim f_G^{\otimes n}}[\|\widehat{\theta}^n(X^n) - \theta_G(X^n)\|_2^2] \\ &= \frac{1}{n}\sum_{x^n\in\mathbb{N}^n}\left(\prod_{i=1}^n f_G(x_i)\right)(\|\widehat{\theta}^n\|_2^2 + \|\theta_G\|_2^2 - 2\widehat{\theta}^n\cdot\theta_G).\end{aligned}$$

Since $f_G(x)$ decays uniformly in $G \in \mathcal{P}([0, A])$ for $x \gg A$, by dominated convergence, it suffices to show that for each $x^n \in \mathbb{N}^n$, each of the following three terms is continuous in G under the metric of weak convergence:

$$L_G(x^n)\|\widehat{\theta}^n(x^n)\|_2^2, \quad L_G(x^n)\cdot\widehat{\theta}^n(x^n)\cdot\theta_G(x^n), \quad \text{and } L_G(x^n)\|\theta_G(x^n)\|_2^2,$$

where $L_G(x^n) = \prod_{i=1}^n f_G(x_i)$ is the likelihood of x^n under f_G.

Now for each $x \geq 0$, the Poisson PMF $\mathrm{Poi}(x; \theta)$ is bounded and continuous in θ. Therefore, the marginal PMF $f_G(x) = \mathbb{E}_G[\mathrm{Poi}(x; \theta)]$ is continuous in G under the metric of weak convergence, and so is the product $L_G(x^n)$. This gives the continuity of the first term $L_G(x^n)\|\widehat{\theta}^n(x^n)\|_2^2$.

For the other two terms, we use the following identity of the Bayes estimator $\theta_G(x) = (x+1)\frac{f_G(x+1)}{f_G(x)}$. In addition, it suffices to consider each coordinate separately; w.l.o.g. we will take the first coordinate. The cross term now becomes

$$\begin{aligned} L_G(x^n) \cdot (\widehat{\theta}^n(x^n))_1 \cdot \theta_G(x_1) &= \left(\prod_{i=1}^n f_G(x_i)\right) (\widehat{\theta}^n(x^n))_1 \theta_G(x_1) \\ &= (x_1+1)\left(\prod_{i=2}^n f_G(x_i)\right) (\widehat{\theta}^n(x^n))_1 f_G(x_1+1), \end{aligned}$$

which is also continuous in G. The final term, meanwhile, becomes

$$\left(\prod_{i=1}^n f_G(x_i)\right) \theta_G(x_1)^2 = (x_1+1)^2 \left(\prod_{i=2}^n f_G(x_i)\right) \frac{f_G(x_1+1)^2}{f_G(x_1)},$$

where the last term is defined to be 0 if $f_G(x_1) = 0$. If $f_G(x_1) > 0$, its continuity follows from the continuity of f_G in G. As for the continuity at $f_G(x_1) = 0$, note that

$$0 \le \frac{f_G(x_1+1)^2}{f_G(x_1)} = \frac{(\mathbb{E}_G[e^{-\theta}\frac{\theta^{x_1+1}}{(x_1+1)!}])^2}{\mathbb{E}_G[e^{-\theta}\frac{\theta^{x_1}}{x_1!}]} \le \frac{x_1!}{[(x_1+1)!]^2}\mathbb{E}_G\left[e^{-\theta}\theta^{x_1+2}\right] \le \frac{A^2}{(x_1+1)^2} f_G(x_1)$$

by Cauchy–Schwarz and $G \in \mathcal{P}([0,A])$. Taking $f_G(x_1) \to 0$ indeed gives the desired continuity.

A.2. Proof of Proposition 1.2

For the first statement, suppose that θ^n_Π is inadmissible and dominated by another estimator $\widehat{\theta}^n$. By inadmissibility,

$$\begin{aligned} 0 &\le \mathbb{E}_{G\sim\Pi}[\mathsf{Regret}(\theta_\Pi; G)] - \mathbb{E}_{G\sim\Pi}[\mathsf{Regret}(\widehat{\theta}^n; G)] \\ &= \mathbb{E}_{G\sim\Pi}\mathbb{E}_G\left[\frac{1}{n}\|\theta^n_\Pi(X^n) - \theta^n\|_2^2 - \frac{1}{n}\|\widehat{\theta}^n(X^n) - \theta^n\|_2^2\right] \\ &= \mathbb{E}_{G\sim\Pi}\mathbb{E}_G\left[-\frac{2}{n}\langle\widehat{\theta}^n(X^n) - \theta^n_\Pi(X^n), \theta^n_\Pi(X^n) - \theta^n\rangle - \frac{1}{n}\|\widehat{\theta}^n(X^n) - \theta^n_\Pi(X^n)\|_2^2\right] \\ &= -\frac{2}{n}\mathbb{E}_{X^n}\langle\widehat{\theta}^n(X^n) - \theta^n_\Pi(X^n), \theta^n_\Pi(X^n) - \mathbb{E}_\Pi[\theta^n|X^n]\rangle - \frac{1}{n}\mathbb{E}_{G\sim\Pi}\mathbb{E}_G\|\widehat{\theta}^n(X^n) - \theta^n_\Pi(X^n)\|_2^2 \\ &= -\frac{1}{n}\mathbb{E}_{G\sim\Pi}\mathbb{E}_G\|\widehat{\theta}^n(X^n) - \theta^n_\Pi(X^n)\|_2^2 \\ &= -\frac{1}{n}\sum_{x^n\in\mathbb{N}^n}\|\widehat{\theta}^n(x^n) - \theta^n_\Pi(x^n)\|_2^2 \cdot \mathbb{E}_{G\sim\Pi}\left[\prod_{i=1}^n f_G(x_i)\right]. \end{aligned}$$

Since $f_G(x) > 0$ for all $G \neq \delta_0$ and $x \in \mathbb{N}$, the assumption $\Pi \neq \delta_{\delta_0}$ implies that $\mathbb{E}_{G\sim\Pi}[\prod_{i=1}^n f_G(x_i)] > 0$ for all $x^n \in \mathbb{N}^n$. Therefore, $\widehat{\theta}^n(x^n) = \theta^n_\Pi(x^n)$ for all $x^n \in \mathbb{N}^n$, so $\widehat{\theta}^n$ cannot achieve a strictly smaller regret at some $G_0 \in \mathcal{P}([0,A])$. This is the desired contradiction, and θ^n_Π is admissible.

For the second statement, we need the following property of the NPMLE-based estimator: for every integer $x \in [0,A]$,

$$\widehat{\theta}^{\mathrm{NPMLE}}(x,\ldots,x) = (x,\ldots,x). \tag{8}$$

To see this, recall that Jana et al. (2025, Theorem 1) shows that the NPMLE $\widehat{G}$ is always supported on $[\min_i X_i, \max_i X_i]$. Therefore, when $X^n = (x, \dots, x)$, the NPMLE is $\widehat{G} = \delta_x$, so the Bayes estimator is $\theta_{\widehat{G}}(x) = x$. We claim that, for integer $A \geq 3$, any admissible estimator cannot satisfy (8) for all $x \in \{A-2, A-1, A\}$.

To prove the claim, assume by contradiction that there exists an admissible $\widehat{\theta} : \mathbb{Z}_+^n \to \mathbb{R}_+^n$ that satisfies (8) for $x \in \{A-2, A-1, A\}$. By the complete class theorem (e.g. Lehmann and Casella (1998, Theorem 7.15) with restriction $G \neq \delta_0$), this $\widehat{\theta}$ must be a pointwise limit of Bayes estimators, i.e. there exists a sequence of PoPs $\Pi_m \in \mathcal{P}(\mathcal{P}([0, A]))$ such that

$$\lim_{m\to\infty} \theta_{\Pi_m}(x, \dots, x) = \widehat{\theta}(x, \dots, x) = (x, \dots, x), \quad \forall x \in \{A-2, A-1, A\}.$$

Writing $m_x(G) := \mathbb{E}_{\theta\sim G}[e^{-\theta}\theta^x]$, the above display implies that

$$\lim_{m\to\infty} \frac{\mathbb{E}_{\Pi_m}[m_x(G)^{n-1} m_{x+1}(G)]}{\mathbb{E}_{\Pi_m}[m_x(G)^n]} = x, \quad \forall x \in \{A-2, A-1, A\}. \tag{9}$$

Since $m_x(G)$ is a sequence of moments, the map $x \mapsto \frac{m_{x+1}(G)}{m_x(G)} = \theta_G(x)$ is non-decreasing; in addition, since $G \in \mathcal{P}([0, A])$, it holds that $\frac{m_{x+1}(G)}{m_x(G)} \in [0, A]$ for all $x \in \mathbb{N}$. Let $\widetilde{\Pi}_m(\mathrm{d}G) \propto \Pi_m(\mathrm{d}G) m_{A-2}(G)^n$ be a tilt of Π_m, and μ_m be the pushforward measure of $\widetilde{\Pi}_m$ of the map

$$h : G \mapsto \left(\frac{m_{A-1}(G)}{m_{A-2}(G)}, \frac{m_A(G)}{m_{A-1}(G)}, \frac{m_{A+1}(G)}{m_A(G)} \right). \tag{10}$$

Then μ_m is a probability measure over $\Omega = \{(r, s, t) \in [0, A]^3 : r \leq s \leq t\}$, and (9) can be rewritten as

$$\lim_{m\to\infty} \int r \mathrm{d}\mu_m = A-2, \quad \lim_{m\to\infty} \int r^n (A-1-s) \mathrm{d}\mu_m = 0, \quad \lim_{m\to\infty} \int r^n s^n (A-t) \mathrm{d}\mu_m = 0.$$

Since Ω is compact, a proper subsequence of $\{\mu_m\}$ admits a weak limit μ, with

$$\int r \mathrm{d}\mu = A-2, \quad \int r^n (A-1-s) \mathrm{d}\mu = 0, \quad \int r^n s^n (A-t) \mathrm{d}\mu = 0. \tag{11}$$

To proceed, we need a technical lemma on the image set of h in (10).

Lemma A.2 *Let K be the closure of the image of h in (10). If $(r, s, t) \in K$ with $r > 0$ and $t = A$, then $s = A$.*

We postpone the proof of Lemma A.2 to the end of this section. Since each μ_m is supported on K, so is the weak limit μ. Since $A \geq 3$, the first identity of (11) implies that $\mu(\{r > 0\}) > 0$. Since $s \geq r$, on the set $\{r > 0\}$ we must also have $s > 0$. Finally, since $t \leq A$, the last identity of (11) implies that $t = A$ μ-a.s. on the set $\{r > 0\}$. By Lemma A.2, this also implies $s = A$ μ-a.s. on the set $\{r > 0\}$. Then the middle identity of (11) implies

$$0 = \int r^n (A-1-s) \mathrm{d}\mu = \int_{r>0} r^n (A-1-s) \mathrm{d}\mu = -\int_{r>0} r^n \mathrm{d}\mu,$$

which forces $\mu(\{r > 0\}) = 0$, a contradiction! This contradiction shows that no admissible estimator can satisfy (8) at all three points $x = A-2, A-1, A$. Since the NPMLE-based estimator does satisfy (8) for all integers x, it is inadmissible.

A.2.1. PROOF OF LEMMA A.2

Suppose $G_m \in \mathcal{P}([0,A])$ is a sequence of priors such that $h(G_m) = (r_m, s_m, t_m) \to (r,s,t)$ as $m \to \infty$, with $r > 0$ and $t = A$. Since $Am_x(G) - m_{x+1}(G) = \mathbb{E}_{\theta\sim G}[(A-\theta)e^{-\theta}\theta^x]$, for $G \in \mathcal{P}([0,A])$, Cauchy–Schwarz gives

$$(Am_A(G) - m_{A+1}(G))(Am_{A-2}(G) - m_{A-1}(G)) \geq (Am_{A-1}(G) - m_A(G))^2.$$

Applying the above inequality to $G = G_m$ yields

$$r_m s_m(A - t_m) \cdot (A - r_m) \geq [r_m(A - s_m)]^2 \Longrightarrow r_m(A - s_m)^2 \leq (A - r_m)s_m(A - t_m).$$

As $t_m \to t = A$ and $r_m \to r > 0$, passing to the limit clearly gives $s_m \to A$, as claimed.

A.3. Proof of Lemma 1.1

Since $r \mapsto E(\delta, r)$ can be made a non-increasing function, we verify Definition 2 for $\delta = n^{-2}$ and $r^2 = n^{-1}$. Our first step is to first identify an intermediate prior G with $\mathsf{TV}(f_{G_0}, f_G) \leq n^{-2}$. To this end, we show that for $L \asymp_A \frac{\log n}{\log\log n}$ and any $G_0 \in \mathcal{P}([0,A])$, there exists a $G \in \mathcal{P}([0,A])$ such that:

- $G = \sum_{j=1}^L w_j \delta_{\lambda_j}$ is supported on L atoms $\lambda_1, \ldots, \lambda_L$ such that $\min_{j\leq L}\{\lambda_j\} \geq \frac{1}{4n^2}$ and $\min_{j\leq L}\{w_j\} \geq \frac{1}{8n^2 L}$.
- $\mathsf{TV}(f_G, f_{G_0}) \leq n^{-2}$.

To begin, we use the following result from Wu and Yang (2016, Lemma 3).

Lemma A.3 *Let V and V' be random variables on $[0,A]$. If $\mathbb{E}[V^j] = \mathbb{E}[V'^j]$ for $j = 0, 1, \ldots, L$ with $L > 2eA$, then*

$$\mathsf{TV}(\mathbb{E}[\mathrm{Poi}(V)], \mathbb{E}[\mathrm{Poi}(V')]) \leq \left(\frac{2eA}{L}\right)^L.$$

To apply this lemma, choose $L \asymp_A \frac{\log n}{\log\log n}$ such that the RHS equals $\frac{1}{2n^2}$. Next, let $V \sim G_0$, and $V' \sim G_1$ be any random variable on $[0,A]$ that matches all first L moments of V. By Carathéodory's theorem, G_1 can be chosen as a discrete distribution with support size at most L. By Lemma A.3, we have $\mathsf{TV}(f_{G_0}, f_{G_1}) \leq \frac{1}{2n^2}$.

Next, we consider the following two identities, which we defer the proof to Section A.3.1.

Lemma A.4 *Let $L > 0$ be an integer. Denote $G_1 = \sum_{j=1}^L w_j \delta_{\lambda_j}$ and $G_2 = \sum_{j=1}^L w'_j \delta_{\lambda'_j}$. Consider the Poisson mixtures f_{G_1} and f_{G_2}. Then the following holds true:*

- $\mathsf{TV}(f_{G_1}, f_{G_2}) \leq \sum_{j=1}^L |w_j - w'_j| + \max_{j=1}^L \mathsf{TV}(\mathrm{Poi}(\lambda_j), \mathrm{Poi}(\lambda'_j))$.
- $\chi^2(f_{G_1} \| f_{G_2}) \leq (1 + \chi^2(w\|w'))(1 + \max_{j=1}^L \chi^2(\mathrm{Poi}(\lambda_j)\|\mathrm{Poi}(\lambda'_j))) - 1$.

To apply this lemma, write $G_1 = \sum_{j=1}^L w_{j,1}\delta_{\lambda_{j,1}}$. Our goal is to construct $G := \sum_{j=1}^L w_j \delta_{\lambda_j}$ such that:

- $\min_{j=1}^{L}\{w_j\} \geq \frac{1}{8n^2L}$, $\sum_{j=1}^{L} w_j = 1$, $\min_{j=1}^{L}\{\lambda_j\} \geq \frac{1}{4n^2}$.
- $\sum_{j=1}^{L} |w_j - w_{j,1}| \leq \frac{1}{4n^2}$, $\max \mathsf{TV}(\mathrm{Poi}(\lambda_j), \mathrm{Poi}(\lambda_{j,1})) \leq \frac{1}{4n^2}$.

Then Lemma A.4 would ensure $\mathsf{TV}(f_{G_1}, f_G) \leq \frac{1}{2n^2}$ and thus $\mathsf{TV}(f_{G_0}, f_G) \leq \frac{1}{n^2}$.

We first claim that $\mathsf{TV}(\mathrm{Poi}(\mu), \mathrm{Poi}(\nu)) \leq 1 - e^{-|\mu-\nu|} \leq |\mu - \nu|$ for $\mu, \nu > 0$. Indeed, let $\mu > \nu$, $X \sim \mathrm{Poi}(\nu)$, $Z \sim \mathrm{Poi}(\mu - \nu)$, and $Y = X + Z$. By coupling, $\mathsf{TV}(\mathrm{Poi}(\mu), \mathrm{Poi}(\nu)) \leq \mathbb{P}[Z > 0] = 1 - e^{-(\mu-\nu)}$, as claimed. In this sequel, the condition $\mathsf{TV}(\mathrm{Poi}(\lambda_j), \mathrm{Poi}(\lambda_{j,1})) \leq \frac{1}{4n^2}$ may be replaced with $|\lambda_j - \lambda_{j,1}| \leq \frac{1}{4n^2}$, so that a sufficient construction is $\lambda_j = \lambda_{j,1} \vee \frac{1}{4n^2}$. To construct w_j, let $m = \arg\max_j w_{j,1}$, and

$$w_j = \begin{cases} w_{j,1} \vee \frac{1}{8n^2L} & \text{if } j \neq m, \\ 1 - \sum_{j \neq m} w_j & \text{if } j = m. \end{cases}$$

By simple algebra, $\min_j w_j \geq \frac{1}{8n^2L}$ for large n and $\sum_j |w_j - w_{j,1}| \leq \frac{1}{4n^2}$. Thus $G = \sum_{j=1}^{L} w_j \delta_{\lambda_j}$ satisfies the above conditions, and we take it to be the intermediate prior in Definition 2.

Next we prove the upper bound on E. Recall that $k = \lceil c_0 \frac{\log n}{\log\log n} \rceil$ is the number of atoms in Algorithm 1; choose $c_0 > 0$ large enough such that $k \geq L$. In this case, we write $G = \sum_{j=1}^{k} w_j \delta_{\lambda_j}$ by allowing repetitions in the atoms. Since

$$\begin{aligned} \chi^2\Big(\sum_{j=1}^{k} w_j \mathrm{Poi}(\lambda_j) \Big\| \sum_{j=1}^{k} w'_j \mathrm{Poi}(\lambda'_j)\Big) + 1 &\leq \big(\chi^2(w\|w') + 1\big)\left(\max_{j\in[k]} \chi^2(\mathrm{Poi}(\lambda_j)\|\mathrm{Poi}(\lambda'_j)) + 1\right) \\ &= \big(\chi^2(w\|w') + 1\big)\left(\max_{j\in[k]} \exp\left(\frac{(\lambda_j - \lambda'_j)^2}{\lambda'_j}\right)\right) \end{aligned}$$

a sufficient condition for the χ^2-divergence to be at most n^{-1} is $\chi^2(w\|w') \leq \frac{1}{3n}$ and

$$\max_{j\in[k]} \frac{(\lambda_j - \lambda'_j)^2}{\lambda'_j} \leq \frac{1}{3n}.$$

Since $w_j, \lambda_j = \Omega(n^{-3})$, a further sufficient condition is to make $\|w - w'\|_\infty \leq cn^{-3}$ and $\|\lambda - \lambda'\|_\infty \leq cn^{-3}$. Since w, w' are both on the simplex Δ^{k-1} and $\lambda, \lambda' \in [0, A]^k$, a standard volume argument establishes that the simple PoP in Algorithm 1 chooses such (λ', w') pair with probability at least

$$\Omega(\frac{1}{n^3})^{O(k)} = \exp(-O(k\log n)) = \exp\left(-O\Big(\frac{\log^2 n}{\log\log n}\Big)\right).$$

This verifies (4) with the claimed $E(n^{-2}, r) = O(\frac{\log^2 n}{\log\log n})$.

A.3.1. PROOF OF LEMMA A.4

For TV, we have

$$\begin{aligned}
\mathsf{TV}(f_{G_1}, f_{G_2}) &= \mathsf{TV}\Big(\sum_{j=1}^{L} w_j \mathrm{Poi}(\lambda_j), \sum_{j=1}^{L} w'_j \mathrm{Poi}(\lambda'_j)\Big) \\
&\overset{\text{(a)}}{\le} \frac{1}{2}\sum_{x\ge 0}\sum_{j=1}^{L} |w_j - w'_j|\mathrm{Poi}(x;\lambda_j) + w'_j|\mathrm{Poi}(x;\lambda_j) - \mathrm{Poi}(x;\lambda'_j)| \\
&= \frac{1}{2}\sum_{j=1}^{L}|w_j - w'_j| + \sum_{j=1}^{L} w'_j \mathsf{TV}(\mathrm{Poi}(\lambda_j), \mathrm{Poi}(\lambda'_j)) \\
&\overset{\text{(b)}}{\le} \sum_{j=1}^{L}|w_j - w'_j| + \max_{j\le L}\{\mathsf{TV}(\mathrm{Poi}(\lambda_j), \mathrm{Poi}(\lambda'_j))\}
\end{aligned}$$

where (a) is obtained by expanding the PMFs of the Poisson mixture at each x, and using the identity $|ab - a'b'| \le |b - b'|a + b'|a - a'|$, and (b) uses $w'_j \ge 0$ and $\sum_{j=1}^{L} w'_j = 1$. For χ^2, we have

$$\begin{aligned}
1 + \chi^2(f_{G_1}\|f_{G_2}) &= \sum_{x\ge 0}\frac{f_{G_1}(x)^2}{f_{G_2}(x)} \\
&= \sum_{x\ge 0}\frac{(\sum_{j=1}^{L} w_j \mathrm{Poi}(x;\lambda_j))^2}{\sum_{j=1}^{L} w'_j \mathrm{Poi}(x;\lambda'_j)} \\
&\overset{\text{(c)}}{\le} \sum_{x\ge 0}\sum_{j=1}^{L}\frac{(w_j\mathrm{Poi}(x;\lambda_j))^2}{w'_j\mathrm{Poi}(x;\lambda'_j)} \\
&= \sum_{j=1}^{L}\frac{w_j^2}{w'_j}(1 + \chi^2(\mathrm{Poi}(\lambda_j)\|\mathrm{Poi}(\lambda'_j))) \\
&\le (1 + \chi^2(w\|w'))(1 + \max_{j\le L}\chi^2(\mathrm{Poi}(\lambda_j)\|\mathrm{Poi}(\lambda'_j))),
\end{aligned}$$

where (c) follows from Cauchy–Schwarz.

A.4. Proof of Theorem 1.3

The first statement directly follows from the following lemma.

Lemma A.5 *For $i \in [n]$, the posterior mean $\mathbb{E}_\Pi[\theta_i|X^n]$ can be expressed as $\mathbb{E}_\Pi[\theta_i|X^n] = f_{\Pi,n}(X_i, \mu_n)$, where $\mu_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$ is the empirical distribution of X^n, and*

$$f_{\Pi,n}(x,\mu) = \frac{\int \Pi(\mathrm{d}G)\exp(n\int \log f_G(x')\mu(\mathrm{d}x'))\cdot\theta_G(x)}{\int \Pi(\mathrm{d}G)\exp(n\int \log f_G(x')\mu(\mathrm{d}x'))}.$$

In particular, for $\mu_m = \frac{1}{m}\sum_{i=1}^{m}\delta_{X_i}$ with a different length m, we have

$$f_{\Pi,n}(X_i,\mu_m) = \mathbb{E}_{G\sim\Pi^\alpha_{G|X^m}}[\theta_G(X_i)] = \frac{\mathbb{E}_{G\sim\Pi^\alpha_{G|X_{\setminus i}}}[\theta_G(X_i)f_G(X_i)^\alpha]}{\mathbb{E}_{G\sim\Pi^\alpha_{G|X_{\setminus i}}}[f_G(X_i)^\alpha]},$$

where $\alpha := \frac{n}{m}$*, and* Π^α *denotes the* α*-posterior:*

$$\Pi^\alpha(\mathrm{d}G|X^m) := \frac{\Pi(\mathrm{d}G)(\prod_{i=1}^m f_G(X_i))^\alpha}{\int \Pi(\mathrm{d}G')(\prod_{i=1}^m f_{G'}(X_i))^\alpha}.$$

For the second statement, we need the following generalization of Lemma 2.1.

Lemma A.6 *Let* G_0 *be a fixed prior supported on* $[0, A]$*, and* G *be a random prior supported on* $[0, A]$ *almost surely. Fix any* $\varepsilon \in (0, e^{-e})$*, and* $\alpha \in (0, 1]$*. Then for an absolute constant* $C = C(A) > 0$*,*

$$\mathbb{E}_{X\sim f_{G_0}}\Big[\Big(\frac{\mathbb{E}[\theta_G(X) f_G(X)^\alpha]}{\mathbb{E}[f_G(X)^\alpha]} - \theta_{G_0}(X)\Big)^2\Big] \le C\Big(\frac{\log(1/\varepsilon)}{\log\log(1/\varepsilon)}\mathbb{E}\left[H^2(f_G, f_{G_0})\right] + \varepsilon\Big).$$

To prove Theorem 1.3, we first integrate the tails in Lemma 2.5 to get

$$\begin{aligned}\mathbb{E}_{X_n}\mathbb{E}_{G\sim\Pi^\alpha_{G|X^{m-1}}}\left[H^2(f_G, f_{G_0})\right] &\le C\left(\varepsilon_m^2 + r_{m,\alpha}^2 + \frac{1}{\alpha m}\right) + \int_{C(\varepsilon_m^2 + r_{m,\alpha}^2 + \frac{1}{\alpha m})}^2 \left(m^{-1} + e^{-c\alpha m t}\right)\mathrm{d}t \\ &\lesssim \varepsilon_m^2 + r_{m,\alpha}^2 + \frac{1}{\alpha m}.\end{aligned}$$

Next we apply Lemma A.5 and A.6 to $m = n_{\mathsf{test}}$, the random prior $G \sim \Pi^\alpha_{G|X^{n_{\mathsf{test}}-1}}$ with $\alpha = \frac{n}{n_{\mathsf{test}}}$, an independent $X = X_n \sim f_{G_0}$, and a small parameter $\varepsilon = n_{\mathsf{test}}^{-1}$:

$$\sup_{G_0\in\mathcal{P}([0,A])} \mathsf{Regret}(f^{n_{\mathsf{test}}}_{\Pi,n}; G_0) \le \frac{C\log n_{\mathsf{test}}}{\log\log n_{\mathsf{test}}}\left(\varepsilon^2_{n_{\mathsf{test}}} + r^2_{n_{\mathsf{test}},\alpha} + \frac{1}{n}\right).$$

Finally, Lemma 2.2 gives $\varepsilon^2_{n_{\mathsf{test}}} = O(\frac{\log n_{\mathsf{test}}}{n_{\mathsf{test}}\log\log n_{\mathsf{test}}})$, and the inequality $E(n_{\mathsf{test}}^{-2}, r) \vee 1 \le nr^2$ implies that $r^2 = \Omega(r^2_{n_{\mathsf{test}},\alpha} + n^{-1})$. This proves Theorem 1.3.

A.4.1. PROOF OF LEMMA A.5

For the first statement, note that Lemma 2.3 yields $\widehat{\theta}_i(X^n) = \mathbb{E}_{G\sim\Pi_{G|X^n}}[\theta_G(X_i)]$, with posterior

$$\Pi(\mathrm{d}G|X^n) = \frac{\Pi(\mathrm{d}G)\prod_{j=1}^n f_G(X_j)}{\int\Pi(\mathrm{d}G')\prod_{j=1}^n f_{G'}(X_j)} = \frac{\Pi(\mathrm{d}G)\exp(n\int\log f_G(x)\mu_n(\mathrm{d}x))}{\int\Pi(\mathrm{d}G')\exp(n\int\log f_{G'}(x)\mu_n(\mathrm{d}x))}.$$

Therefore, we conclude that

$$\widehat{\theta}_i(X^n) = \int\Pi(\mathrm{d}G|X^n)\theta_G(X_i) = \frac{\int\Pi(\mathrm{d}G)\exp(n\int\log f_G(x)\mu_n(\mathrm{d}x))\cdot\theta_G(X_i)}{\int\Pi(\mathrm{d}G)\exp(n\int\log f_G(x)\mu_n(\mathrm{d}x))} =: f_{\Pi,n}(X_i, \mu_n).$$

This completes the first claim. For the second claim, note that on a new sequence X^m,

$$\begin{aligned}f_{\Pi,n}(X_i, \mu_m) &= \frac{\int\Pi(\mathrm{d}G)\exp(n\int\log f_G(x)\mu_m(\mathrm{d}x))\cdot\theta_G(X_i)}{\int\Pi(\mathrm{d}G)\exp(n\int\log f_G(x)\mu_m(\mathrm{d}x))} \\ &= \frac{\int\Pi(\mathrm{d}G)(\prod_{j=1}^m f_G(X_j))^\alpha\cdot\theta_G(X_i)}{\int\Pi(\mathrm{d}G)(\prod_{j=1}^m f_G(X_j))^\alpha} =: \mathbb{E}_{G\sim\Pi^\alpha_{G|X^m}}[\theta_G(X_i)],\end{aligned}$$

where $\alpha = \frac{n}{m}$, and we use the notation $\Pi^\alpha(\mathrm{d}G|X^m)$ to denote the α-posterior:

$$\Pi^\alpha(\mathrm{d}G|X^m) := \frac{\Pi(\mathrm{d}G)(\prod_{i=1}^m f_G(X_i))^\alpha}{\int \Pi(\mathrm{d}G')(\prod_{i=1}^m f_{G'}(X_i))^\alpha}.$$

Similar to the proof of Lemma 2.3, we also have

$$f_{\Pi,n}(X_i, \mu_m) = \mathbb{E}_{G\sim\Pi^\alpha_{G|X^m}}[\theta_G(X_i)] = \frac{\mathbb{E}_{G\sim\Pi^\alpha_{G|X_{\setminus i}}}[\theta_G(X_i)f_G(X_i)^\alpha]}{\mathbb{E}_{G\sim\Pi^\alpha_{G|X_{\setminus i}}}[f_G(X_i)^\alpha]}.$$

A.4.2. PROOF OF LEMMA A.6

Let $\mathcal{E} := \{x \in \mathbb{N} : \mathbb{E}[f_G(x)^\alpha] \geq \frac{1}{2} f_{G_0}(x)^\alpha\}$. By Markov's inequality, for $x \in \mathcal{E}^c$,

$$\mathbb{P}\left(f_G(x) \leq \frac{3}{4} f_{G_0}(x)\right) \geq 1 - \frac{\mathbb{E}[f_G(x)^\alpha]}{\left(\frac{3}{4} f_{G_0}(x)\right)^\alpha} \geq 1 - \frac{1}{2}\left(\frac{4}{3}\right)^\alpha \geq \frac{1}{3}.$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[H^2(f_{G_0}, f_G)\right] &= \sum_{x\in\mathbb{N}} \mathbb{E}\left(\sqrt{f_{G_0}(x)} - \sqrt{f_G(x)}\right)^2 \geq \sum_{x\in\mathcal{E}^c} \mathbb{E}\left(\sqrt{f_{G_0}(x)} - \sqrt{f_G(x)}\right)^2 \\
&\geq \sum_{x\in\mathcal{E}^c} \frac{1}{3}\left(\sqrt{f_{G_0}(x)} - \sqrt{\frac{3}{4} f_{G_0}(x)}\right)^2 \geq c_0 \sum_{x\in\mathcal{E}^c} f_{G_0}(x) = c_0 f_{G_0}(\mathcal{E}^c),
\end{aligned}$$

for some universal constant $c_0 > 0$. Consequently, we have arrived at

$$f_{G_0}(\mathcal{E}^c) \leq C_0 \cdot \mathbb{E}\left[H^2(f_{G_0}, f_G)\right].$$

Therefore, as $\theta_G(X), \theta_{G_0}(X) \in [0, A]$ almost surely,

$$\mathbb{E}_{X\sim f_{G_0}}\left[\left(\frac{\mathbb{E}[\theta_G(X)f_G(X)^\alpha]}{\mathbb{E}[f_G(X)^\alpha]} - \theta_{G_0}(X)\right)^2 \mathbf{1}_{\{X\in\mathcal{E}^c\}}\right] \leq A^2 f_{G_0}(\mathcal{E}^c) \leq C_1 \cdot \mathbb{E}\left[H^2(f_{G_0}, f_G)\right]. \tag{12}$$

Next, for $X \in \mathcal{E}$, we have

$$\begin{aligned}
&\mathbb{E}_{X\sim f_{G_0}}\left[\left(\frac{\mathbb{E}[\theta_G(X)f_G(X)^\alpha]}{\mathbb{E}[f_G(X)^\alpha]} - \theta_{G_0}(X)\right)^2 \mathbf{1}_{\{X\in\mathcal{E}\}}\right] \\
&= \mathbb{E}_{X\sim f_{G_0}}\left[\left(\frac{\mathbb{E}[(\theta_G(X) - \theta_{G_0}(X))f_G(X)^\alpha]}{\mathbb{E}[f_G(X)^\alpha]}\right)^2 \mathbf{1}_{\{X\in\mathcal{E}\}}\right] \\
&\overset{(a)}{\leq} \mathbb{E}_{X\sim f_{G_0}}\left[\frac{\mathbb{E}[(\theta_G(X) - \theta_{G_0}(X))^2 f_G(X)^\alpha]}{\mathbb{E}[f_G(X)^\alpha]} \mathbf{1}_{\{X\in\mathcal{E}\}}\right] \\
&\overset{(b)}{\leq} 2 \cdot \mathbb{E}_{X\sim f_{G_0}}\left[\frac{\mathbb{E}[(\theta_G(X) - \theta_{G_0}(X))^2 f_G(X)^\alpha]}{f_{G_0}(X)^\alpha}\right] \\
&= 2 \cdot \mathbb{E}\left[\sum_{x\in\mathbb{N}} (\theta_G(x) - \theta_{G_0}(x))^2 f_G(x)^\alpha f_{G_0}(x)^{1-\alpha}\right] \\
&\overset{(c)}{\leq} 2\alpha \cdot \mathbb{E}\left[\mathbb{E}_{X\sim f_G}\left((\theta_G(X) - \theta_{G_0}(X))^2\right)\right] + 2(1-\alpha) \cdot \mathbb{E}\left[\mathbb{E}_{X\sim f_{G_0}}\left((\theta_G(X) - \theta_{G_0}(X))^2\right)\right] \\
&\overset{(d)}{\leq} C_2\left(\frac{\log(1/\varepsilon)}{\log\log(1/\varepsilon)} \mathbb{E}\left[H^2(f_G, f_{G_0})\right] + \varepsilon\right).
\end{aligned}$$

Here (a) is due to Cauchy–Schwarz, (b) uses the definition of $\mathcal{E}$, (c) follows from Young's inequality $x^\alpha y^{1-\alpha} \le \alpha x + (1-\alpha)y$ for $\alpha \in (0,1]$ and $x, y \ge 0$, and (d) applies Lemma 2.1 twice to (G, G_0) and (G_0, G). Combining this inequality with (12) completes the proof of Lemma A.6.

Appendix B. Deferred Proofs in Section 2

B.1. Proof of Lemma 2.1

Fix $\varepsilon \in (0, e^{-e})$. The following inequality was shown in Jana et al. (2025, Lemma 4): for every integer $K \ge 0$,

$$\mathbb{E}_{X\sim f_{G_0}}\left[(\theta_G(X) - \theta_{G_0}(X))^2\right] \le C\left(KH^2(f_G, f_{G_0}) + \varepsilon_K(G_0)\right),$$

with $\varepsilon_K(G_0) := \sum_{y\ge K} f_{G_0}(y)$. Since $\mathrm{supp}(G_0) \subseteq [0, A]$, standard Poisson tails yield $\varepsilon_K(G_0) \le \varepsilon$ for $K = O_A(\frac{\log(1/\varepsilon)}{\log\log(1/\varepsilon)})$. Plugging this choice of K gives the claimed bound.

B.2. Proof of Lemma 2.2

Fix $\varepsilon \in (0, e^{-e})$, any $\rho \ge \varepsilon$, and $G_0 \in \mathcal{P}([0, A])$. Since the Hellinger distance is a metric, it suffices to construct an *improper* covering of $\{f_G : H(f_G, f_{G_0}) \le 2\rho\}$ using Hellinger balls of radius $\rho/2$. Let $L = \lceil C\frac{\log(1/\varepsilon)}{\log\log(1/\varepsilon)}\rceil$; by Poisson tail bounds, there is a large absolute constant $C = C(A) > 0$ such that

$$\sup_{G\in\mathcal{P}([0,A])} \sum_{x>L} f_G(x) \le \frac{\varepsilon^2}{16} \le \frac{\rho^2}{16}.$$

Let $H_{\le L}(P, Q) := (\sum_{i=0}^{L}(\sqrt{P(i)} - \sqrt{Q(i)})^2)^{1/2}$ be the truncated Hellinger distance, we have

$$H^2(f_{G_1}, f_{G_2}) \le H^2_{\le L}(f_{G_1}, f_{G_2}) + f_{G_1}(X > L) + f_{G_2}(X > L).$$

Therefore, for $G_1, G_2 \in \mathcal{P}([0, A])$, the inequality $H_{\le L}(f_{G_1}, f_{G_2}) \le \frac{\rho}{4}$ implies $H(f_{G_1}, f_{G_2}) \le \frac{\rho}{2}$. As a result, by truncating all Poisson mixtures onto a smaller support $\{0, 1, \dots, L\}$, it suffices to restrict to the space $\mathcal{S}$ of discrete sub-distributions

$$\mathcal{S} = \Big\{P = (p_0, \dots, p_L) \in \mathbb{R}_+^{L+1} : \sum_{i=0}^{L} p_i \le 1\Big\},$$

and cover the set $\{P \in \mathcal{S} : H(P, P_0) \le 2\rho\}$ using Hellinger balls of radius $\rho/4$. This covering problem is easily solved via a volume argument: parametrizing $q_i = \sqrt{p_i}$, the set $\mathcal{S}$ becomes the intersection of the unit ℓ_2 ball and the non-negative orthant for $Q = (q_0, \dots, q_L) \in \mathbb{R}^{L+1}$, and the Hellinger distance becomes the ℓ_2 metric: $H(P_1, P_2) = \|Q_1 - Q_2\|_2$. Therefore, the covering problem is equivalent to covering an ℓ_2 ball of radius 2ρ by smaller ℓ_2 balls of radius $\rho/4$ in $\mathbb{R}^{L+1}$; a simple volume argument shows that the log covering number is $\log N \le C(L+1)$. Therefore,

$$\log N_{\mathrm{loc}}(\varepsilon, \mathcal{P}, H) = O(L) = O\Big(\frac{\log(1/\varepsilon)}{\log\log(1/\varepsilon)}\Big).$$

B.3. Proof of Lemma 2.3

By repeated applications of Bayes' rule, the conditional density of θ_i given X^n is

$$\frac{\mathbb{E}_{G\sim\Pi}\left[G(\theta_i)\mathrm{Poi}(X_i;\theta_i)\prod_{j\neq i}f_G(X_j)\right]}{\mathbb{E}_{G\sim\Pi}\left[\prod_{j=1}^n f_G(X_j)\right]} = \mathbb{E}_{G\sim\Pi_{G|X^n}}\left[\frac{G(\theta_i)\mathrm{Poi}(X_i;\theta_i)}{f_G(X_i)}\right] = \mathbb{E}_{G\sim\Pi_{G|X^n}}\left[\mathbb{P}_G\left(\theta_i|X_i\right)\right].$$

Here $\mathbb{P}_G$ denotes the joint distribution $\theta_i\sim G$ and $X_i|\theta_i\sim\mathrm{Poi}(\theta_i)$. Therefore,

$$\mathbb{E}_\Pi[\theta_i|X^n] = \mathbb{E}_{G\sim\Pi_{G|X^n}}\left[\int\theta_i\mathbb{P}_G\left(\theta_i|X_i\right)\mathrm{d}\theta_i\right] = \mathbb{E}_{G\sim\Pi_{G|X^n}}\left[\theta_G(X_i)\right],$$

by definition of θ_G in (6). This establishes the first identity. For the second identity, we further use the Bayes' rule to write

$$\mathbb{E}_{G\sim\Pi_{G|X^n}}\left[\theta_G(X_i)\right] = \frac{\mathbb{E}_{G\sim\Pi_{G|X_{\setminus i}}}[\theta_G(X_i)f_G(X_i)]}{\mathbb{E}_{G\sim\Pi_{G|X_{\setminus i}}}[f_G(X_i)]} \overset{(6)}{=} (X_i+1)\cdot\frac{\mathbb{E}_{G\sim\Pi_{G|X_{\setminus i}}}[f_G(X_i+1)]}{\mathbb{E}_{G\sim\Pi_{G|X_{\setminus i}}}[f_G(X_i)]}.$$

Since $G_i = \mathbb{E}_{G\sim\Pi_{G|X_{\setminus i}}}[G]$, we have $f_{G_i}(x) = \mathbb{E}_{G\sim\Pi_{G|X_{\setminus i}}}[f_G(x)]$ for every $x\in\mathbb{N}$. Therefore, we can continue the above expression and obtain

$$\mathbb{E}_{G\sim\Pi_{G|X^n}}\left[\theta_G(X_i)\right] = (X_i+1)\frac{f_{G_i}(X_i+1)}{f_{G_i}(X_i)} \overset{(6)}{=} \theta_{G_i}(X_i).$$

B.4. Proof of Lemma 2.4

In this section we prove Lemma 2.4. We essentially apply the same classical posterior contraction arguments in Ghosal et al. (2000), with a few adaptations to obtain a high-probability statement. Let G be given in Definition 2 such that $\mathsf{TV}(f_G, f_{G_0})\le n^{-2}$. We claim that it suffices to consider the case $G_0 = G$. In fact, once we establish Lemma 2.4 with G in place of G_0, we can move to G_0 by noting that $\mathsf{TV}(f_{G_0}^{\otimes(n-1)}, f_G^{\otimes(n-1)})\le(n-1)\mathsf{TV}(f_{G_0},f_G)\le n^{-1}$ and

$$\begin{aligned}H^2(f_{G_0},\Pi_{X_n|X^{n-1}}) &\lesssim H^2(f_G,\Pi_{X_n|X^{n-1}}) + H^2(f_G,f_{G_0})\\ &\le H^2(f_G,\Pi_{X_n|X^{n-1}}) + 2\mathsf{TV}(f_G,f_{G_0})\\ &\le H^2(f_G,\Pi_{X_n|X^{n-1}}) + O\left(n^{-2}\right).\end{aligned}$$

Hence, this step only replaces H^2 by $O(H^2+n^{-2})$ and amplifies the error probability by an additive factor of n^{-1}; in the sequel we assume that $G = G_0$ and establish the exponential error probability.

We apply the arguments in Ghosal et al. (2000, Theorem 8.1). Let $N_{\mathrm{loc}}(\varepsilon) := N_{\mathrm{loc}}(\varepsilon,\mathcal{P},H)$ be the local covering number of $\mathcal{P}$, and $\varepsilon_n>0$ satisfy $\log N_{\mathrm{loc}}(\varepsilon_n)\le n\varepsilon_n^2$. By Ghosal et al. (2000, Theorem 7.1) (or the classical results in Birgé (1983); Le Cam (1986)), for $\varepsilon\ge C\varepsilon_n$ with a large absolute constant $C\ge 2$, there exists a test $\phi_n = \phi_n(X^{n-1})\in\{0,1\}$ such that

$$\mathbb{E}_{f_{G_0}^{\otimes(n-1)}}[\phi_n] \le N_{\mathrm{loc}}(\varepsilon_n)\exp\left(-cn\varepsilon^2\right), \tag{13}$$

$$\sup_{f_G\in\mathcal{P}:H(f_G,f_{G_0})\ge\varepsilon}\mathbb{E}_{f_G^{\otimes(n-1)}}[1-\phi_n] \le \exp\left(-cn\varepsilon^2\right), \tag{14}$$

where $c > 0$ is a universal constant. By (13),

$$f_{G_0}^{\otimes(n-1)}\left(\{\phi_n = 1\}\right) \le N_{\mathrm{loc}}(\varepsilon_n)\exp\left(-cn\varepsilon^2\right) \le \exp\left(-\frac{c}{2}n\varepsilon^2\right), \tag{15}$$

by $\log N_{\mathrm{loc}}(\varepsilon_n) \le n\varepsilon_n^2$ and $\varepsilon \ge C\varepsilon_n$ for large enough $C > 0$. On the other hand, let $\Pi_{G|X^{n-1}}$ be the posterior distribution of G given X^{n-1}, then

$$\begin{aligned}\Pi_{G|X^{n-1}}\left(H^2(f_{G_0}, f_G) > \varepsilon^2\right) &= \frac{\int_{G:H(f_{G_0},f_G)>\varepsilon}\Pi(\mathrm{d}G)\prod_{i=1}^{n-1} f_G(X_i)}{\int \Pi(\mathrm{d}G)\prod_{i=1}^{n-1} f_G(X_i)} \\ &= \frac{\int_{G:H(f_{G_0},f_G)>\varepsilon}\Pi(\mathrm{d}G)\prod_{i=1}^{n-1}\frac{f_G(X_i)}{f_{G_0}(X_i)}}{\int \Pi(\mathrm{d}G)\prod_{i=1}^{n-1}\frac{f_G(X_i)}{f_{G_0}(X_i)}}\end{aligned}$$

by the Bayes' rule. We bound the numerator and denominator separately.

1. Numerator: Adding the event $\phi_n = 0$ and taking expectation with respect to $X^{n-1} \sim f_{G_0}^{\otimes(n-1)}$ yields

$$\begin{aligned}&\mathbb{E}_{f_{G_0}^{\otimes(n-1)}}\left[(1-\phi_n)\int_{G:H(f_{G_0},f_G)>\varepsilon}\Pi(\mathrm{d}G)\prod_{i=1}^{n-1}\frac{f_G(X_i)}{f_{G_0}(X_i)}\right] \\ &= \int_{G:H(f_{G_0},f_G)>\varepsilon}\Pi(\mathrm{d}G)\cdot\mathbb{E}_{f_G^{\otimes(n-1)}}[1-\phi_n] \overset{(14)}{\le} e^{-cn\varepsilon^2}.\end{aligned}$$

   Therefore, by Markov's inequality,

$$f_{G_0}^{\otimes(n-1)}\left(\left\{\int_{G:H(f_{G_0},f_G)>\varepsilon}\Pi(\mathrm{d}G)\prod_{i=1}^{n-1}\frac{f_G(X_i)}{f_{G_0}(X_i)} > e^{-\frac{c}{2}n\varepsilon^2}\right\}\cap\{\phi_n = 0\}\right) \le \exp\left(-\frac{c}{2}n\varepsilon^2\right). \tag{16}$$

2. Denominator: Let $r_n > 0$ satisfy $E(n^{-2}, r_n) \le nr_n^2$, and $U = \left\{G' : \chi^2(f_{G_0}\|f_{G'}) \le r_n^2\right\}$ be a neighborhood of G_0. Since we have assumed $G = G_0$ without loss of generality, assumption (4) ensures $\Pi(U) \ge e^{-E_n}$, with $E_n := E(n^{-2}, r_n)$. By Jensen's inequality,

$$\int\Pi(\mathrm{d}G)\prod_{i=1}^{n-1}\frac{f_G(X_i)}{f_{G_0}(X_i)} \ge \int_U\Pi(\mathrm{d}G)\prod_{i=1}^{n-1}\frac{f_G(X_i)}{f_{G_0}(X_i)} \ge \Pi(U)\exp\left(\int\Pi_{|U}(\mathrm{d}G)\sum_{i=1}^{n-1}\log\frac{f_G(X_i)}{f_{G_0}(X_i)}\right),$$

   where $\Pi_{|U}$ denotes the restriction of Π to U. By Lemma B.1 below (which itself is a refinement of Ghosal et al. (2000, Lemma 8.3)) with $P = f_G, Q = f_{G_0}$, the random variable

$$Y_i = \int\Pi_{|U}(\mathrm{d}G)\log\frac{f_G(X_i)}{f_{G_0}(X_i)}$$

   satisfies $\mathbb{E}[e^{|Y_i|} - 1 - |Y_i|] \le \int\Pi_{|U}(\mathrm{d}G)\chi^2(f_{G_0}\|f_G) \le r_n^2$ by convexity of $x \mapsto e^{|x|} - 1 - |x|$ and definition of U. Therefore, by the discussion below Bernstein's inequality in van der Vaart and Wellner (1996, Lemma 2.2.11), we have

$$f_{G_0}^{\otimes(n-1)}\left(\sum_{i=1}^{n-1}(Y_i - \mathbb{E}[Y_i]) \ge -\frac{c}{8}n\varepsilon^2\right) \ge 1 - \exp\left(-\frac{c'(cn\varepsilon^2/8)^2}{nr_n^2 + cn\varepsilon^2/8}\right) \ge 1 - \exp\left(-c''n\varepsilon^2\right),$$

for universal constants $c, c' > 0$, under an additional assumption that $\varepsilon \geq Cr_n$. Since

$$\mathbb{E}[Y_i] = -\int \Pi_{|U}(\mathrm{d}G)\mathsf{KL}(f_{G_0}\|f_G) \geq -\int \Pi_{|U}(\mathrm{d}G)\chi^2(f_{G_0}\|f_G) \geq -r_n^2,$$

we conclude that

$$f_{G_0}^{\otimes(n-1)}\left(\int \Pi(\mathrm{d}G)\prod_{i=1}^{n-1}\frac{f_G(X_i)}{f_{G_0}(X_i)} \geq \exp\left(-E_n - \frac{c}{8}n\varepsilon^2 - nr_n^2\right)\right) \geq 1 - \exp\left(-c''n\varepsilon^2\right).$$

Finally, since $E_n \leq nr_n^2$, as long as $\varepsilon \geq Cr_n$ with a large enough constant $C > 0$, this implies

$$f_{G_0}^{\otimes(n-1)}\left(\int \Pi(\mathrm{d}G)\prod_{i=1}^{n-1}\frac{f_G(X_i)}{f_{G_0}(X_i)} \geq \exp\left(-\frac{c}{4}n\varepsilon^2\right)\right) \geq 1 - \exp\left(-c''n\varepsilon^2\right). \tag{17}$$

By (15), (16), and (17), we conclude that for every $\varepsilon^2 \geq C(\varepsilon_n^2 + r_n^2 + \frac{1}{n})$, with probability at least $1 - e^{-c_1 n\varepsilon^2}$ over the randomness of $X^{n-1} \sim f_{G_0}^{\otimes(n-1)}$, it holds that

$$\Pi_{G|X^{n-1}}\left(H^2(f_{G_0}, f_G) > \varepsilon^2\right) \leq e^{-c_2 n\varepsilon^2}, \quad \text{with } c_2 = \frac{c}{4}.$$

This argument applies with ε^2 replaced by $j\varepsilon^2$ for every integer $j \geq 1$. Taking a union bound over $1 \leq j \leq \lceil 2/\varepsilon^2 \rceil$, we obtain that, with probability at least $1 - \sum_{j=1}^{\lceil 2/\varepsilon^2 \rceil} e^{-c_1 nj\varepsilon^2}$, the following bound holds simultaneously for all such j:

$$\Pi_{G|X^{n-1}}\left(H^2(f_{G_0}, f_G) > j\varepsilon^2\right) \leq e^{-c_2 nj\varepsilon^2}.$$

On this event, since $H^2 \leq 2$, the layer-cake formula gives

$$\begin{aligned}\mathbb{E}_{G\sim\Pi_{G|X^{n-1}}}\left[H^2(f_{G_0}, f_G)\right] &\leq \varepsilon^2 \sum_{j=0}^{\lceil 2/\varepsilon^2 \rceil} \Pi_{G|X^{n-1}}\left(H^2(f_{G_0}, f_G) > j\varepsilon^2\right) \\ &\leq \varepsilon^2\left(1 + \sum_{j\geq 1} e^{-c_2 nj\varepsilon^2}\right) \leq C\left(\varepsilon^2 + \frac{1}{n}\right) \leq C'\varepsilon^2,\end{aligned}$$

where the last inequality uses $\varepsilon^2 \geq \frac{C}{n}$. The exceptional probability is bounded by

$$\sum_{j\geq 1} e^{-c_1 nj\varepsilon^2} \leq Ce^{-c_1 n\varepsilon^2} \leq e^{-c_1' n\varepsilon^2},$$

after adjusting constants. Finally, as $\Pi_{X_n|X^{n-1}} = \mathbb{E}_{\Pi_{G|X^{n-1}}}[f_G]$, on the above event, the convexity of the squared Hellinger distance gives

$$H^2(f_{G_0}, \Pi_{X_n|X^{n-1}}) \leq \mathbb{E}_{\Pi_{G|X^{n-1}}}\left[H^2(f_{G_0}, f_G)\right] \leq C'\varepsilon^2.$$

This is the claimed result.

Lemma B.1 *For probability distributions P and Q,*

$$\mathbb{E}_Q\Big[\exp\Big(\Big|\log\frac{P}{Q}\Big|\Big)-1-\Big|\log\frac{P}{Q}\Big|\Big]\le\chi^2(Q\|P).$$

Proof Using $\log x\ge 1-\frac{1}{x}$, when $P\ge Q$, this integral is

$$\int_{P\ge Q}\Big(P-Q-Q\log\frac{P}{Q}\Big)\le\int_{P\ge Q}\Big(P-Q-Q\Big(1-\frac{Q}{P}\Big)\Big)=\int_{P\ge Q}\frac{(Q-P)^2}{P}.$$

When $P<Q$, this integral becomes

$$\int_{P<Q}\Big(\frac{Q^2}{P}-Q-Q\log\frac{Q}{P}\Big)\le\int_{P<Q}\Big(\frac{Q^2}{P}-Q-Q\Big(1-\frac{P}{Q}\Big)\Big)=\int_{P<Q}\frac{(Q-P)^2}{P}.$$

Summing up gives the target upper bound $\chi^2(Q\|P)$. ■

B.5. Proof of Lemma 2.5

The proof precisely mimics the arguments of Lemma 2.4 and uses the same test construction ϕ_n satisfying (13) and (14). The main difference starts from

$$\Pi^{\alpha}_{G|X^{n-1}}\left(H^2(f_{G_0},f_G)>\varepsilon^2\right)=\frac{\int_{G:H(f_{G_0},f_G)>\varepsilon}\Pi(\mathrm{d}G)\Big(\prod_{i=1}^{n-1}\frac{f_G(X_i)}{f_{G_0}(X_i)}\Big)^{\alpha}}{\int\Pi(\mathrm{d}G)\Big(\prod_{i=1}^{n-1}\frac{f_G(X_i)}{f_{G_0}(X_i)}\Big)^{\alpha}}.$$

Again, we bound the numerator and denominator separately.

1. Numerator: Since $0\le\alpha\le 1$, by Hölder's inequality,

$$\begin{aligned}
&\mathbb{E}_{f_{G_0}^{\otimes(n-1)}}\left[(1-\phi_n)\int_{G:H(f_{G_0},f_G)>\varepsilon}\Pi(\mathrm{d}G)\Big(\prod_{i=1}^{n-1}\frac{f_G(X_i)}{f_{G_0}(X_i)}\Big)^{\alpha}\right]\\
&=\int_{G:H(f_{G_0},f_G)>\varepsilon}\Pi(\mathrm{d}G)\cdot\int_{\phi_n=0}f_G^{\otimes(n-1)}(\mathrm{d}x^{n-1})^{\alpha}f_{G_0}^{\otimes(n-1)}(\mathrm{d}x^{n-1})^{1-\alpha}\\
&\le\int_{G:H(f_{G_0},f_G)>\varepsilon}\Pi(\mathrm{d}G)\cdot f_G^{\otimes(n-1)}(\{\phi_n=0\})^{\alpha}f_{G_0}^{\otimes(n-1)}(\{\phi_n=0\})^{1-\alpha}\overset{(14)}{\le}e^{-c\alpha n\varepsilon^2}.
\end{aligned}$$

Therefore, by Markov's inequality,

$$\begin{aligned}
&f_{G_0}^{\otimes(n-1)}\left(\left\{\int_{G:H(f_{G_0},f_G)>\varepsilon}\Pi(\mathrm{d}G)\Big(\prod_{i=1}^{n-1}\frac{f_G(X_i)}{f_{G_0}(X_i)}\Big)^{\alpha}>e^{-\frac{c}{2}\alpha n\varepsilon^2}\right\}\cap\{\phi_n=0\}\right)\\
&\le\exp\Big(-\frac{c}{2}\alpha n\varepsilon^2\Big).
\end{aligned}\tag{18}$$

2. Denominator: Again, let $U = \{G' : \chi^2(f_{G_0} \| f_{G'}) \le r_{n,\alpha}^2\}$ be a neighborhood of G_0, and $\Pi_{|U}$ be the restriction of Π to U. By Jensen's inequality,

$$\int \Pi(\mathrm{d}G)\Big(\prod_{i=1}^{n-1}\frac{f_G(X_i)}{f_{G_0}(X_i)}\Big)^\alpha \ge \int_U \Pi(\mathrm{d}G)\Big(\prod_{i=1}^{n-1}\frac{f_G(X_i)}{f_{G_0}(X_i)}\Big)^\alpha \ge \Pi(U)\exp\left(\alpha\int \Pi_{|U}(\mathrm{d}G)\sum_{i=1}^{n-1}\log\frac{f_G(X_i)}{f_{G_0}(X_i)}\right).$$

Using $\Pi(U) \ge e^{-E(n^{-2}, r_{n,\alpha})} \ge e^{-\alpha n r_{n,\alpha}^2}$ and the same Bernstein concentration for the exponent, as long as $\varepsilon \ge C r_{n,\alpha}$, we have

$$f_{G_0}^{\otimes(n-1)}\left(\int \Pi(\mathrm{d}G)\Big(\prod_{i=1}^{n-1}\frac{f_G(X_i)}{f_{G_0}(X_i)}\Big)^\alpha \ge \exp\Big(-\frac{c}{4}\alpha n\varepsilon^2\Big)\right) \ge 1 - \exp\left(-c'' n\varepsilon^2\right). \quad (19)$$

By (13), (18), and (19), the same arguments in the proof of Lemma 2.4 lead to

$$\begin{aligned}
&f_{G_0}^{\otimes(n-1)}\left(\left\{\Pi^\alpha_{G|X^{n-1}}\left(H^2(f_{G_0}, f_G) > \varepsilon^2\right) > e^{-c\alpha n\varepsilon^2/4}\right\}\right)\\
&\le f_{G_0}^{\otimes(n-1)}\left(\left\{\Pi^\alpha_{G|X^{n-1}}\left(H^2(f_{G_0}, f_G) > \varepsilon^2\right) > e^{-c\alpha n\varepsilon^2/4}\right\} \cap \{\phi_n = 0\}\right) + f_{G_0}^{\otimes(n-1)}\{\phi_n = 1\}\\
&\le e^{-c_1\alpha n\varepsilon^2},
\end{aligned}$$

as long as $\varepsilon^2 \ge C(\varepsilon_n^2 + r_{n,\alpha}^2 + \frac{1}{\alpha n})$. The same layer-cake argument in the proof of Lemma 2.4 gives the target upper bound of $\mathbb{E}_{G\sim\Pi^\alpha_{G|X^{n-1}}}[H^2(f_{G_0}, f_G)]$.

Appendix C. Deferred proofs in Section 3

C.1. Proof of Lemma 3.1

We first derive the expression of $\widehat{\theta}^n$ in (7). For each $X^n \in \mathbb{N}^n$, let $\mathcal{S}(X^n) = \{m : \exists \pi \in S^n, \pi \circ X^{n,(m)} = X^n\}$ be the collection of data batches $X^{n,(m)}$ which are permutations of X^n. In addition, for all $m \in \mathcal{S}(X^n)$, we use π_m to record the m-th permutation, i.e., $\pi_m \circ X^{n,(m)} = X^n$. Note that $\mathcal{S}(X^n)$ only depends on the *type* (or the sorted version) of X^n, denoted by $T(X^n)$; in addition, for a permutation-equivariant estimator $\widehat{\theta}^n$, we only need to specify its output for each input type. Therefore, the ERM objective in (7) takes a separable form on types, and for each type T, $\widehat{\theta}(X^n)$ (with input type T) is the minimizer of

$$\begin{aligned}
\sum_{m\in\mathcal{S}(T)} \|\theta^{n,(m)} - \widehat{\theta}(X^{n,(m)})\|_2^2 &= \sum_{m\in\mathcal{S}(T)} \|\theta^{n,(m)} - \pi_m^{-1}\circ\widehat{\theta}(X^n)\|_2^2\\
&= \sum_{m\in\mathcal{S}(T)} \|\pi_m\circ\theta^{n,(m)} - \widehat{\theta}(X^n)\|_2^2.
\end{aligned}$$

This gives that

$$\widehat{\theta}^n(X^n) = \frac{1}{|\mathcal{S}(X^n)|}\sum_{m\in\mathcal{S}(X^n)} \pi_m\circ\theta^{n,(m)}$$

whenever $\mathcal{S}(X^n) \neq \emptyset$; when $\mathcal{S}(X^n) = \emptyset$, we set $\widehat{\theta}^n(X^n)$ to an arbitrary vector in $[0, A]^n$.

To continue, we claim the following that connects the approximation error and regret bound uniform over all $G_0 \in \mathcal{P}([0, A])$, which we will defer the proof to Section C.1.1.

Lemma C.1 *Let r_n be defined in Theorem 1.2. Let $\delta > 0$ and a set $\mathcal{G} \subseteq \mathbb{N}^n$ such that for all priors $G_0 \in \mathcal{P}([0, A])$, we have $f_{G_0}^{\otimes n}(\mathcal{G}^c) \leq \delta$. Suppose an estimator $\widehat{\theta}^n \in [0, A]^n$ is an approximate population risk minimizer on $\mathcal{G}$, satisfying*

$$\mathbb{E}_\Pi \left[\frac{1}{n} \|\widehat{\theta}^n(X^n) - \mathbb{E}_\Pi[\theta^n | X^n]\|_2^2 \mathbf{1}_{\{X^n \in \mathcal{G}\}} \right] \leq \varepsilon^2. \tag{20}$$

Then for an absolute constant $C = C(A)$,

$$\sup_{G_0} \mathsf{Regret}(\widehat{\theta}^n; G_0) \leq \frac{C \log n}{n \log \log n} \left(\frac{\log n}{\log \log n} + n r_n^2 \right) + C \left(\varepsilon e^{n r_n^2} + \delta \right).$$

We take $\mathcal{G} = [0, L]^n$ with $L = C(A) \frac{\log n}{\log \log n}$, so that $\delta = O(\frac{1}{n})$. We show that $\varepsilon \leq \frac{1}{n} \exp(-n r_n^2)$ for the estimator $\widehat{\theta}^n$ in (7) and $M = \exp(O_A(\frac{\log^2 n}{\log \log n} + n r_n^2))$. Let $\mathbf{X} = (X^{n,(m)})_{m \in [M]}$ denote the collection of all X sequences in training data. We note that whenever $|\mathcal{S}(x^n)| \geq 1$,

$$\mathbb{E}[\widehat{\theta}^n(x^n) | \mathbf{X}] = \frac{1}{|\mathcal{S}(x^n)|} \sum_{m \in \mathcal{S}(x^n)} \mathbb{E}[\pi_m \circ \theta^{n,(m)} | \mathbf{X}] = \mathbb{E}_\Pi[\theta^n | X^n = x^n],$$

where (θ^n, X^n) is a single generic batch generated from Π. By independence among batches, we conclude from the above unbiasedness that

$$\begin{aligned} \mathbb{E}\Big[\|\widehat{\theta}^n(x^n) - \mathbb{E}_\Pi[\theta^n | X^n = x^n]\|_2^2 \mid \mathbf{X} \Big] &= \frac{1}{|\mathcal{S}(x^n)|} \mathbb{E}\Big[\|\theta^n - \mathbb{E}_\Pi[\theta^n | X^n = x^n]\|_2^2 \mid X^n = x^n \Big] \\ &\leq \frac{n A^2}{|\mathcal{S}(x^n)|}. \end{aligned}$$

For $|\mathcal{S}(x^n)| = 0$, we simply replace the denominator by 1. Finally, taking expectation with respect to $\mathbf{X}$ and x^n gives

$$\mathbb{E} \left[\frac{1}{n} \|\widehat{\theta}^n(X^n) - \mathbb{E}_\Pi[\theta^n | X^n]\|_2^2 \mathbf{1}_{\{X^n \in [0,L]^n\}} \right] \leq \mathbb{E} \left[\frac{A^2}{|\mathcal{S}(X^n)| \vee 1} \mathbf{1}_{\{X^n \in [0,L]^n\}} \right].$$

To upper bound the final quantity, we sum over all possible types T obtained from $[0, L]^n$. Let $p(T)$ be the probability that a fresh draw $X^n \sim \Pi_{X^n}$ has type T, then $|\mathcal{S}(X^n)| \sim \mathsf{B}(M, p(T))$. Consequently,

$$\begin{aligned} \mathbb{E} \left[\frac{A^2}{|\mathcal{S}(X^n)| \vee 1} \Big| T(X^n) = T \right] &= A^2 \cdot (1 - p(T))^M + A^2 \sum_{k=1}^{M} \frac{1}{k} \binom{M}{k} p(T)^k (1 - p(T))^{M-k} \\ &\overset{(a)}{\lesssim} (1 - p(T))^M + \min\{1, \frac{1}{M p(T)}\} \\ &\overset{(b)}{\lesssim} \min\{1, \frac{1}{M p(T)}\} \end{aligned}$$

for some absolute constant $C > 0$. Here (a) is due to (Polyanskiy and Wu, 2021, Lemma 16), and (b) is due to the inequality $(1-x)^M \le \min\{1, \frac{1}{Mx}\}$ for all $x \in [0,1]$. Let N be the number of types for $X^n \in [0,L]^n$, with

$$N = \binom{L+n}{L} = O((n+L)^L) = O\left(\exp(\frac{\log^2 n}{\log\log n})\right).$$

Denoting $T_L^{(n)}$ as the set of all types in $[0,L]^n$, we may continue the previous display to get

$$\varepsilon^2 \lesssim \mathbb{E}\left[\min\{1, \frac{1}{Mp(T)}\}\mathbf{1}_{\left\{T \in T_L^{(n)}\right\}}\right] \le \sum_{T \in T_L^{(n)}} \frac{p(T)}{Mp(T)} = \frac{N}{M}.$$

Therefore, by taking $M = n^2 N \exp(2nr_n^2) \le \exp\left(O\left(\frac{\log^2 n}{\log\log n} + nr_n^2\right)\right)$, we obtain $\varepsilon \le \frac{1}{n} e^{-nr_n^2}$, as desired.

C.1.1. PROOF OF LEMMA C.1

By Theorem 1.2 and the triangle inequality $(a+b)^2 \le 2(a^2+b^2)$, we only need to show that for every G_0 supported on $[0,A]$,

$$\mathbb{E}_{G_0}\left[\frac{1}{n}\|\widehat{\theta}^n(X^n) - \mathbb{E}_\Pi[\theta^n|X^n]\|_2^2\right] \le C\left(\varepsilon e^{nr_n^2} + \delta\right).$$

We first note that

$$\mathbb{E}_{G_0}\left[\frac{1}{n}\|\widehat{\theta}^n(X^n) - \mathbb{E}_\Pi[\theta^n|X^n]\|_2^2 \mathbf{1}_{\{X^n \in \mathcal{G}^c\}}\right] \le A^2 \mathbb{P}_{G_0}[X^n \in \mathcal{G}^c] \le A^2\delta$$

hence it suffices to show that

$$\mathbb{E}_{G_0}\left[\frac{1}{n}\|\widehat{\theta}^n(X^n) - \mathbb{E}_\Pi[\theta^n|X^n]\|_2^2 \mathbf{1}_{\{X^n \in \mathcal{G}\}}\right] \le C e^{nr_n^2}\varepsilon$$

To this end, first note that the data distribution $f_{G_0}^{\otimes n}$ can be changed into $f_G^{\otimes n}$ for some G, in view of Definition 2, by incurring a cost at most $\mathsf{TV}(f_G^{\otimes n}, f_{G_0}^{\otimes n}) \le n\mathsf{TV}(f_G, f_{G_0}) = O(\frac{1}{n})$. Therefore, WLOG we may assume that $G = G_0$ and write

$$\begin{aligned}
\chi^2(f_{G_0}^{\otimes n}\|\Pi_{X^n}) + 1 &= \sum_{x^n \in \mathbb{N}^n} \frac{f_{G_0}^{\otimes n}(x^n)^2}{\Pi_{X^n}(x^n)} = \sum_{x^n \in \mathbb{N}^n} \frac{f_{G_0}^{\otimes n}(x^n)^2}{\mathbb{E}_{G\sim\Pi}[f_G^{\otimes n}(x^n)]} \\
&\overset{\text{(a)}}{\le} \frac{1}{\Pi(U)} \sum_{x^n \in \mathbb{N}^n} \frac{f_{G_0}^{\otimes n}(x^n)^2}{\mathbb{E}_{G\sim\Pi_{|U}}[f_G^{\otimes n}(x^n)]} \\
&\overset{\text{(b)}}{\le} \frac{1}{\Pi(U)} \mathbb{E}_{G\sim\Pi_{|U}}\left[\chi^2(f_{G_0}^{\otimes n}\|f_G^{\otimes n}) + 1\right] \\
&= \frac{1}{\Pi(U)} \mathbb{E}_{G\sim\Pi_{|U}}\left[\left(\chi^2(f_{G_0}\|f_G) + 1\right)^n\right] \\
&\overset{\text{(c)}}{\le} e^{E(n^{-2}, r_n) + nr_n^2} \overset{\text{(d)}}{\le} e^{2nr_n^2},
\end{aligned}$$

where (a) defines $U := \left\{G : \chi^2(f_{G_0} \| f_G) \le r_n^2\right\}$ and $\Pi_{|U}$ as the restriction of Π to U, (b) follows from convexity of $x \mapsto \frac{1}{x}$, (c) uses the definition of U and Definition 2, and (d) uses the definition of r_n that $E(n^{-2}, r_n) \le nr_n^2$. Next, we invoke the following form of the Cauchy-Schwarz inequality on arbitrary distributions P and Q, an event E, and arbitrary nonnegative function g:

$$(\mathbb{E}_P[g\mathbf{1}_{\{E\}}])^2 \le \mathbb{E}_Q[g^2\mathbf{1}_{\{E\}}](\chi^2(P\|Q) + 1).$$

Using $P = f_{G_0}^{\otimes n}$, $Q = \Pi_{X^n}$, $g = \frac{1}{n}\|\widehat{\theta}^n(X^n) - \mathbb{E}_\Pi[\theta^n|X^n]\|_2^2$, and E denotes the event $X^n \in \mathcal{G}$, we conclude that

$$\begin{aligned}\mathbb{E}_{G_0}\left[g\mathbf{1}_{\{X^n\in\mathcal{G}\}}\right] &\le \sqrt{e^{2nr_n^2}\cdot \mathbb{E}_\Pi\left[g^2\mathbf{1}_{\{X^n\in\mathcal{G}\}}\right]} \\ &\le \sqrt{e^{2nr_n^2}\cdot A^2\mathbb{E}_\Pi\left[g\mathbf{1}_{\{X^n\in\mathcal{G}\}}\right]} \le Ce^{nr_n^2}\varepsilon.\end{aligned}$$

This is the desired claim.

C.2. Proof of Theorem 3.1

To generalize our approach to subexponential priors, we consider the prior truncation techniques (as generalized from Jana et al. (2023, Lemma 12)). In other words, we show that for any reasonable estimator $\widehat{\theta}$ and a prior $G \in \mathsf{SubE}(s)$, we can find G' supported on $[0, c(s)\log n]$ such that the extra regret incurred by this prior truncation is at most $O(\frac{1}{n})$.

Lemma C.2 *Let $G \in \mathsf{SubE}(s)$, and consider any estimator $\widehat{\theta}^n \in [0, M]^n$. Denote G' as the prior truncated at $c\log n$ where $c := c(s)$ is such that $G(\theta > c\log n) < \frac{1}{n^{10}}$, and $G'(\theta \in \cdot) = G(\theta \in \cdot | \theta \le c\log n)$. Then*

$$\mathsf{Regret}(\widehat{\theta}^n; G) \le \mathsf{Regret}(\widehat{\theta}^n; G') + O_s\left(\frac{1+M^2}{n^4}\right).$$

The proof is deferred to Appendix C.2.1.

For the hierarchical Bayes estimator θ_Π^n, since $\Pi \in \mathcal{P}(\mathcal{P}([0, c_0\log n]))$, we have $M = O(\log n)$ in Lemma C.2. Therefore, the overhead in Lemma C.2 is negligible, and we may assume that G_0 is supported on $[0, c_0\log n]$.

Next we solve for ε_n and r_n in the posterior contraction lemma (Lemma 2.4). First, for $\mathcal{P} = \{f_G : \mathrm{supp}(G) \subseteq [0, c_0\log n]\}$, the same truncation argument (with the truncation threshold at $\Theta(\log n)$ rather than $\Theta(\frac{\log n}{\log\log n})$) in Lemma 2.2 gives $\log N_{\mathrm{loc}}(\varepsilon, \mathcal{P}, H) = O(\log n)$ for all $\varepsilon^2 \ge \frac{1}{n}$. Therefore, we can take $\varepsilon_n^2 = O(\frac{\log n}{n})$ for the inequality $\log N_{\mathrm{loc}}(\varepsilon_n, \mathcal{P}, H) \le n\varepsilon_n^2$. Next, to upper bound $E(n^{-2}, r)$, Lemma A.3 implies that every $f_G \in \mathcal{P}$ can be approximated by a finite Poisson mixture with $L = O(\log n)$ atoms. Therefore, by the same arguments as in the proof of Lemma 1.1, we get $E(n^{-2}, r) = O(\log^2 n)$ for $r^2 \ge \frac{1}{n}$, so that we can take $r_n^2 = O(\frac{\log^2 n}{n})$ from the inequality $E(n^{-2}, r_n) \le nr_n^2$. Therefore, by Lemma 2.4 and integrating the tails, we obtain

$$\mathbb{E}_{X^{n-1}\sim f_{G_0}^{\otimes(n-1)}}[H^2(f_{G_0}, f_{G_n})] = O\left(\frac{\log^2 n}{n}\right),$$

with $G_n = \mathbb{E}_{G\sim\Pi_{G|X^{n-1}}}[G]$ defined in Lemma 2.3.

We now conclude by connecting Hellinger distance regret bound by using the following tools (Jana et al., 2025, Lemma 4): for $G, G_0 \in \mathcal{P}([0, A])$ and any $K \in \mathbb{N}$,

$$\mathsf{Regret}(\theta_G; G_0) \le C(AKH^2(f_G, f_{G_0}) + \varepsilon_K(G_0)),$$

with $\varepsilon_K(G_0) := \sum_{y\ge K} f_{G_0}(y)$. Standard Poisson tails yield that $\varepsilon_K(G_0) \le \frac{1}{n}$ for some $K = O(\log n)$. Choosing $G = G_n$, taking expectation over G_n, and noting that $A = c_0 \log n$, the regret bound is now $O(\frac{\log^4 n}{n})$, as desired.

C.2.1. PROOF OF LEMMA C.2

Denote the event $E = \{\max_{i=1}^n \theta_i \le c\log n\}$; we have $\mathbb{P}[E^c] \le n^{-9}$. Let $\mathsf{mmse}(G) := \min_{\widehat\theta}\mathbb{E}_{\theta\sim G}[(\widehat\theta(X) - \theta)^2]$, i.e. the MSE achieved by the Bayes estimator. Then by Polyanskiy and Wu (2021, Eqn. (131)),

$$\mathsf{Regret}(\widehat{\theta}^n; G) \le \mathsf{Regret}(\widehat{\theta}^n; G') + \mathsf{mmse}(G') - \mathsf{mmse}(G) + \mathbb{E}_G\Big[\frac{1}{n}\|\widehat\theta - \theta\|_2^2 \mathbf{1}_{\{E^c\}}\Big]$$

By Wu and Verdú (2011, Lemma 2), $\mathsf{mmse}(G') - \mathsf{mmse}(G) \le \frac{\varepsilon}{1-\varepsilon}\mathsf{mmse}(G) = O_s(\varepsilon)$, with $\varepsilon = G(\theta > c\log n) \le n^{-10}$. It remains to bound $\mathbb{E}_\pi[(\widehat\theta - \theta)^2 \mathbf{1}_{\{E^c\}}]$: by Cauchy-Schwarz,

$$\mathbb{E}_G\Big[\frac{1}{n}\|\widehat\theta - \theta\|_2^2 \mathbf{1}_{\{E^c\}}\Big] \le \sqrt{\mathbb{P}[E^c]\mathbb{E}_G\Big[\frac{1}{n^2}\|\widehat\theta - \theta\|_2^4\Big]} \le n^{-4}(M^2 + O_s(1)) = O_s\left(\frac{1+M^2}{n^4}\right),$$

where we have used that $\mathbb{E}_G[\theta^4] = O_s(1)$ for all $G \in \mathsf{SubE}(s)$. This completes the proof.

C.3. Proof of Theorem 3.2

We first establish the rate of posterior contraction in the Gaussian case, by working out ε_n and r_n in Lemma 2.4. To upper bound the local entropy $\log N_{\mathrm{loc}}(\varepsilon, \mathcal{P}, H)$ for the class of Gaussian mixtures $\mathcal{P} = \{f_G(x) = \mathbb{E}_{\theta\sim G}[\varphi(x-\theta)] : G \in \mathcal{P}([-A, A])\}$, we will overbound it by the global entropy $\log N(\varepsilon, \mathcal{P}, H)$. We quote the following relationship between the TV distance of Gaussian mixtures and moment matching (Wu and Yang, 2020, Lemma 9):

$$\mathsf{TV}(f_{G_0}, f_{G_1}) \le \frac{1}{2}\left[\sum_{m=0}^\infty \frac{|\mathbb{E}_{U\sim G_1}[U^m] - \mathbb{E}_{V\sim G_2}[V^m]|^2}{m!}\right]^{1/2}.$$

Since $H^2 \le 2\mathsf{TV}$, by simple algebra and Carathéodory's theorem, every $f_{G_0} \in \mathcal{P}$ is $O(\frac{1}{n^2})$-close to a finite Gaussian mixture with $L = O_A(\frac{\log n}{\log\log n})$ atoms. Therefore, by quantizing the atom locations and weights of an L-component Gaussian mixture, we get

$$\log N(n^{-1/2}, \mathcal{P}, H) = O\left(\frac{\log^2 n}{\log\log n}\right).$$

We can therefore choose $\varepsilon_n^2 = O(\frac{\log^2 n}{n\log\log n})$ in Lemma 2.4.

The above approximation by an L-component Gaussian mixture also gives an upper bound of $E(n^{-2}, r)$. Based on the finite Gaussian mixture, the same argument in the proof of Lemma 1.1 yields

$$E(n^{-2}, r) = O(L \log n) = O\left(\frac{\log^2 n}{\log\log n}\right), \quad r^2 \geq \frac{1}{n}.$$

Therefore, we can take $r_n^2 = O(\frac{\log^2 n}{n \log\log n})$ in Lemma 2.4, which gives

$$f_{G_0}^{\otimes(n-1)}\left(H^2(f_{G_0}, f_{G_n}) \geq C\varepsilon^2\right) \leq \frac{1}{n} + e^{-cn\varepsilon^2}, \quad \text{for } \varepsilon^2 \gtrsim \frac{\log^2 n}{n \log\log n},$$

where $G_n = \mathbb{E}_{G \sim \Pi_{G|X^{n-1}}}[G]$ is defined in Lemma 2.3.

Finally we quote a state-of-the-art regret-Hellinger inequality in the recent work (Chen and Wu, 2026, Theorem 1):

$$\mathsf{Regret}(\theta_{G_n}; G_0) \leq C H^2(f_{G_n}, f_{G_0}) \frac{\log \frac{1}{H^2(f_{G_n}, f_{G_0})}}{\log\log \frac{1}{H^2(f_{G_n}, f_{G_0})}}.$$

By the Hellinger guarantee above and tail integration, we obtain a regret of $O(\frac{\log^3 n}{n(\log\log n)^2})$ for θ_Π^n, as desired.

C.4. Proof of Theorem 3.3

The recent work (Kang et al., 2026, Lemma 16) tells that, for any G, G_0 supported on $[0, A]$ and any $K > p$, we have

$$\mathbb{E}_{G_0}\left[(g_G - g_{G_0})^2\right] \leq C(A^{2p} + A^p K^p) H^2(f_{G_0}, f_G) + A^{2p} f_{G_0}(X > K - p).$$

Taking $K = C(A)\frac{\log n}{\log\log n}$ such that $f_{G_0}(X > K - p) \leq \frac{1}{n^2}$, the target regret bound is then a direct consequence of the same Hellinger guarantee in Theorem 1.2 and Lemma 1.1:

$$\mathbb{E}_{X^{n-1} \sim f_{G_0}^{\otimes(n-1)}}[H^2(f_{G_0}, f_{G_n})] = O\left(\frac{\log^2 n}{n \log\log n}\right),$$

with $G_n = \mathbb{E}_{G \sim \Pi_{G|X^{n-1}}}[G]$, where Lemma 2.3 implies that $g_{\Pi,n} = g_{G_n}$.

Appendix D. Additional experimental details

D.1. Details of Experiments in Section 4.1

In this section, we mention the details of how the test priors are generated: they are the neural and the multinomial prior-on-priors, which are different from the PoP we used in Algorithm 1 that we use to train our transformer. Code release is available at https://github.com/Anzoteh96/eb-transformers.

Neural-generated prior-on-priors. This is described in Teh et al. (2025, Appendix A.2), reproduced here for clarity. We sample $\theta_{\mathsf{base}} \in [0, A]$ via the following: first, let $\mathcal{M}$ be the set of priors determined by some two-layer perceptron with a non-linear activation. This is defined as follows:

$$\mathcal{M} = \{\pi : \pi = \varphi_\sharp^{W_1, W_2, \sigma} \mathsf{Unif}[0, A]\}$$

where $\varphi^{W_1,W_2,\sigma}(x) = \mathsf{Sigmoid}(10 W_2 \sigma(W_1 x))$, W_1, W_2 are linear operators, and σ is an activation function chosen randomly from

$$\{\mathrm{GELU}, \mathrm{ReLU}, \mathrm{SELU}, \mathrm{CELU}, \mathrm{SiLU}, \mathrm{Tanh}, \mathrm{TanhShrink}\}.$$

The test Poisson means θ_{base} are then produced by sampling from a mixture of 4 priors in $\mathcal{M}$.

Multinomial prior-on-prior. Here, the prior-on-prior Π_{test} consists of priors in the form $\sum_{j=1}^{10A} w_j \delta_{\frac{j}{10}}$ with $(w_1, \cdots, w_{10A}) \sim \mathrm{Dir}(1, 1, \ldots, 1)$. In other words, the test prior is a discrete distribution uniformly chosen from the simplex over a fixed grid.

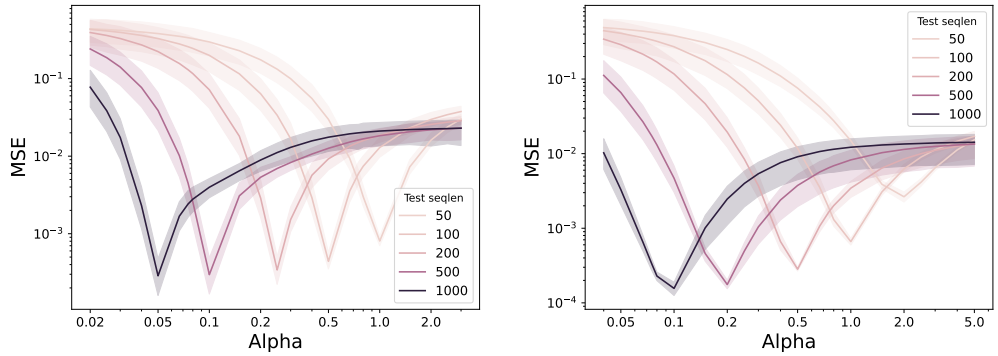
D.2. Details of Experiments in Section 4.2

Here, we describe how the experiments in Section 4.2 are set up. For each $m = 2, 5, 10$ we consider $N = 2,000$ runs of sampling m priors $G_1, \cdots, G_m$, take the simple PoP as $\Pi_m = \frac{1}{m} \sum_{i=1}^m \delta_{G_i}$, and choose the ones (among the N runs) such that the hierarchical Bayes of Π_m has the highest average regret across $G_1, \cdots, G_m$.

Next, we train a transformer that trains exclusively on $G_1, \ldots, G_m$. For all $m = 2, 5, 10$, we take $M = 200,000$ steps, and for each step i we sample 200 batches, each in the form of (θ^n, X^n) pairs sampled from G_j where $j = (i-1)\%m + 1$. The transformer architecture and training details are identical to Teh et al. (2025).

D.3. Additional Results in Section 4.2

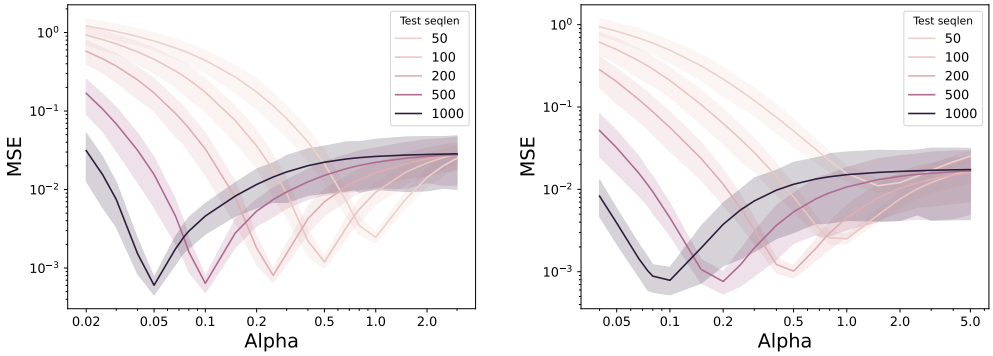
In this section we demonstrate some of the numerical results deferred from the main parts. Precisely, we tabulate the results of α-posterior on $m = 5$ and $m = 10$, as tabulated in Figure 4 and Figure 5. We see that the hypothesis that the transformers are doing α-posterior for $\alpha \simeq \frac{n}{n_{\mathsf{test}}}$ continues to hold, despite larger error bars.



(*a*) $m = 5, n = 50$ (*b*) $m = 5, n = 100$

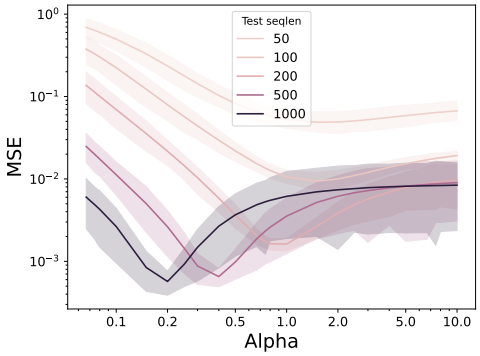
(*c*) $m = 5, n = 200$ (*d*) $m = 5, n = 500$

Figure 4: Mean squared distance between transformer output and the hierarchical Bayes estimator using various α-posteriors, trained on $m = 5$ priors.

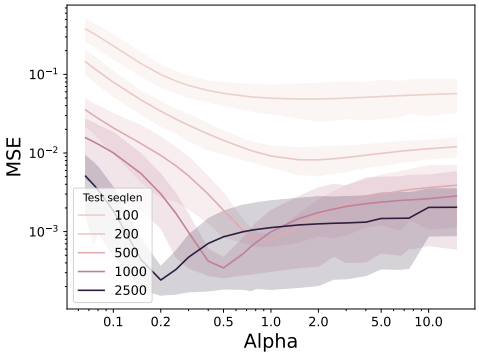


(a) $m = 10, n = 50$

(b) $m = 10, n = 100$



(c) $m = 10, n = 200$



(d) $m = 10, n = 500$

Figure 5: Mean squared distance between transformer output and the hierarchical Bayes estimator using various α-posteriors, with $m = 10$ priors.