An Empirical Bayes Perspective on Heteroskedastic Mean Estimation

Yanjun Han YANJUNHAN@NYU.EDU
New York University
Abhishek Shetty SHETTY@MIT.EDU
Massachusetts Institute of Technology and Georgia Institute of Technology
Jacob Shkrob JAS10184@NYU.EDU
New York University

Editors: Steve Hanneke and Tor Lattimore

Abstract

Towards understanding the fundamental limits of estimation from data of varied quality, we study the problem of estimating a mean parameter from heteroskedastic Gaussian observations where the variances are unknown and may vary across observations. While, with known variances, a simple linear estimator attains the smallest mean squared error, estimation without this knowledge is challenging due to the large number of nuisance parameters. We propose a simple and principled approach based on empirical Bayes: model the observations as if they were i.i.d. from a normal scale mixture and compute the profile maximum likelihood estimator (MLE) for the mean, treating the nonparametric mixing distribution as nuisance. Our result shows that this estimator achieves near-optimal error bounds across various heteroskedastic models in the literature. In particular, for the subset-of-signals problem where an unknown subset of observations has small variance, our estimator adaptively achieves the minimax rate for all signal sizes, including the sharp phase transition, without any tuning parameters.

One of our key technical steps is a sharper metric entropy bound for normal scale mixtures, obtained via generalized moment matching and Chebyshev approximation. This approach yields an improved polylogarithmic, rather than polynomial, dependence on problem parameters, which could be of independent interest.

Keywords: Heteroskedastic mean estimation, empirical Bayes, nonparametric maximum likelihood, normal scale mixtures, metric entropy

1. Introduction

Estimation of a signal from heterogeneous data lies at the heart of statistics and machine learning. The heterogeneity can arise from variations in data quality, measurement precision, or sampling conditions and is ubiquitous in real-world applications. Thus, it is a fundamental statistical challenge to develop principled methods that allow one to estimate signals robustly from heterogeneous data with minimal knowledge of the data quality.

Perhaps the simplest abstraction of this general objective is estimation of a single one dimensional parameter from data of varying quality. In particular, given independent observations $X_1, \dots, X_n$ with $X_i \sim \mathcal{N}(\mu, \sigma_i^2)$, the learner's target is to estimate the mean parameter $\mu \in \mathbb{R}$. Even in this simple one dimensional setting, heterogeneity leads to statistical challenges. In the case when $(\sigma_1, \dots, \sigma_n)$ are known, the maximum likelihood estimator (MLE) for μ is

$$\widehat{\mu}_{\mathrm{known}-\sigma} = \frac{\sum_{i=1}^n \frac{X_i}{\sigma_i^2}}{\sum_{i=1}^n \frac{1}{\sigma_i^2}}, \tag{1}$$

which is also the uniformly minimum-variance unbiased estimator (UMVUE) and achieves the optimal error rate of $\mathbb{E}[|\widehat{\mu} - \mu|] = (\sum_{i=1}^n \frac{1}{\sigma_i^2})^{-1/2}$. The key aspect of this estimator is that this bound is *robust* to large variances. For example, if there are $n/2$ variances that are large while the other $n/2$ variances are small, then the error rate is still $O(1/\sqrt{n})$, as the large variances are effectively ignored in the weighted average. Compare this to the error of the sample average $\overline{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$, which has error rate $\mathbb{E}[(\overline{\mu} - \mu)^2] = \frac{1}{n^2}\sum_{i=1}^n \sigma_i^2$ and could be arbitrarily large even if a single variance is large. On the other hand, $\widehat{\mu}_{\text{known}-\sigma}$ critically relies on the knowledge of $(\sigma_1, \ldots, \sigma_n)$, which could not be estimated reliably simply because the number of unknown parameters $n+1$ is more than the number of observations n while the sample average $\overline{\mu}$ is oblivious to the knowledge of $(\sigma_1, \ldots, \sigma_n)$. The key question that we would like to address is:

> Are there general principled methods that allow one to estimate signals robustly from heterogeneous data with minimal knowledge of the data quality?

An important special case considered in the literature is the "subset of signals" problem (Liang and Yuan, 2020), where it is assumed that $|\{i \in [n] : \sigma_i \le 1\}| \ge m$, i.e., m out of n observations can be treated as "signals" (although the location of this subset of signals is still unknown). Even for this simple setting, the minimax rate has only been recently characterized (Compton and Valiant, 2024): there exists an estimator $\widehat{\mu}$ with

$$\mathbb{E}|\widehat{\mu} - \mu| = \begin{cases} \widetilde{O}\Big(\big(\frac{n}{m^4}\big)^{1/2}\Big) & \text{if } \log n \ll m \le n^{1/4}, \\ \widetilde{O}\Big(\big(\frac{n}{m^4}\big)^{1/6}\Big) & \text{if } n^{1/4} \le m \le n, \end{cases} \tag{2}$$

and this dependence on (m, n) (including the phase transition at $m \asymp n^{1/4}$) is not improvable in the worst case.[1] While the minimax rate is known, existing estimators (Compton and Valiant, 2024) achieving this minimax rate are often quite complicated, requiring multiple tunable parameters and are reliant on the particular formulation of the problem, and tend not to lead to unifying statistical principles for learning from heterogeneous data.

Towards building a unified theory of heteroskedasticity in estimation, we make a perhaps surprising connection to the study of *empirical Bayes* methods. Using this connection, we show that the simple principle of maximum likelihood estimation, in conjunction with an empirical Bayes framework, leads to a unified estimator that achieves the minimax rate adaptively for all ranges of m. In this framework, instead of modeling the heterogeneous variances $\sigma_1, \ldots, \sigma_n$ separately, we model them using their empirical distribution $G_n := \frac{1}{n}\sum_{i=1}^n \delta_{\sigma_i}$ and treat the observations $X_1, \ldots, X_n$ *as if* they were *i.i.d.* drawn from a normal scale mixture $f_{\mu,G_n}(x) := \mathbb{E}_{\sigma\sim G_n}[\frac{1}{\sigma}\varphi(\frac{x-\mu}{\sigma})]$, where φ denotes the standard normal density. Under this i.i.d. model, the joint MLE $(\widehat{\mu}, \widehat{G})$ is naturally defined as

$$(\widehat{\mu}, \widehat{G}) = \operatorname*{arg\,max}_{\mu\in\mathbb{R}, \operatorname{supp}(G)\subseteq[\sigma_{\min},\sigma_{\max}]} \frac{1}{n}\sum_{i=1}^n \log f_{\mu,G}(X_i), \tag{3}$$

and the final estimator is simply defined as $\widehat{\mu}^{\text{EB}} = \widehat{\mu}$ in (3). In other words, our estimator $\widehat{\mu}^{\text{EB}}$ is the profile MLE (Murphy and Van der Vaart, 2000) of μ, treating G as a nuisance parameter.

1. Here and throughout, $\widetilde{O}(\cdot)$ hides polylogarithmic factors in n.

Note that the nuisance estimator $\widehat{G}$ in (3) is a *nonparametric* MLE (NPMLE), since it may be any probability distribution supported on $[\sigma_{\min}, \sigma_{\max}]$. In particular, this formulation no longer enforces the structure of an empirical distribution, nor does it explicitly require the "subset-of-signals" structure $G([0,1]) \geq \frac{m}{n}$. Here $\sigma_{\min}, \sigma_{\max} > 0$ are hyperparameters used in the MLE and satisfy $\sigma_{\min} \leq \sigma_i \leq \sigma_{\max}$ for every $i \in [n]$; further discussion is provided later.

1.1. Main results

Our first result shows that, although the estimator $\widehat{\mu}^{\mathrm{EB}}$ is developed from an entirely different principle compared with existing estimators, it nevertheless has an instance-dependent error bound.

Theorem 1.1 *Let $\sigma_i \in [\sigma_{\min}, \sigma_{\max}]$ for all $i \in [n]$. With probability at least $1-\delta$, the estimator $\widehat{\mu}^{\mathrm{EB}}$ in (3) achieves (the exact logarithmic factor is displayed in Theorem 1.3)*

$$|\widehat{\mu}^{\mathrm{EB}} - \mu| \leq C\omega_{H^2,G_n}\Big(\frac{\mathrm{polylog}(n, \frac{\sigma_{\max}}{\sigma_{\min}}, \frac{1}{\delta})}{n}\Big),$$

where $\omega_{H^2,G_n}(t)$ is the Hellinger modulus of continuity in the location family:

$$\omega_{H^2,G_n}(t) = \sup\Big\{|\mu_1 - \mu_2| : \mu_1, \mu_2 \in \mathbb{R}, H^2(f_{\mu_1,G_n}, f_{\mu_2,G_n}) \leq t\Big\}.$$

We remark that Theorem 1.1 establishes an upper bound that is competitive with the best oracle estimator possessing knowledge of G_n, uniformly over all possible choices of G_n. Indeed, even in the mean estimation problem where $X_1, \ldots, X_n$ are i.i.d. drawn from f_{μ,G_n} with *known* G_n, Le Cam's two-point method (Le Cam, 1973, 1986) yields a minimax lower bound of $\omega_{H^2,G_n}(\frac{1}{n})$.[2] By comparison, the radius in the Hellinger modulus of continuity in our upper bound is $\widetilde{O}(\frac{1}{n})$, which essentially matches this lower bound even when G_n is unknown.

Specializing to several heterogeneous mean estimation models studied in the literature (Pensia et al., 2022; Devroye et al., 2023), Theorem 1.1 implies the following error bounds, all of which match the best known rates ((Pensia et al., 2022, Table 1) and (Devroye et al., 2023, Section 5.1)) up to logarithmic factors.

Corollary 1.1 *Let $L = \widetilde{O}(1)$ be the poly-logarithmic factor in Theorem 1.1. The following high-probability guarantee holds in specific examples:*

1. *Equal variance: when $\sigma_i \equiv 1$, then $|\widehat{\mu}^{\mathrm{EB}} - \mu| = O(\sqrt{\frac{L}{n}})$;*
2. *Quadratic variance: when $\sigma_i = i$, then $|\widehat{\mu}^{\mathrm{EB}} - \mu| = O(L \log n)$;*
3. *Two variances: when $\sigma_1 = \cdots = \sigma_m = 1$ and $\sigma_{m+1} = \cdots = \sigma_n = \sigma \geq 1$ with $m \leq \frac{n}{2}$, then $|\widehat{\mu}^{\mathrm{EB}} - \mu| = O(\sigma\sqrt{\frac{L}{n}})$. For specific ranges of m, sharper upper bounds can be obtained:*

$$|\widehat{\mu}^{\mathrm{EB}} - \mu| = \begin{cases} O(\frac{\sqrt{nL}}{m}) & \text{if } m \geq \sqrt{nL}, \\ O(1) & \text{if } \sqrt{\frac{nL}{\sigma}} + L \leq m < \sqrt{nL}. \end{cases}$$

2. This precise argument does not hold in the compound setting, but this intuition is conjectured to remain valid. See Section 5 for detailed discussion.

4. *α-mixture distribution: when $\sigma_1 = \cdots = \sigma_m = 1$ and $\sigma_{m+1} = \cdots = \sigma_n = n^\alpha$ with $m = cL$, then $|\widehat{\mu}^{\mathrm{EB}} - \mu| = O(n^{\alpha - 1/2}\sqrt{L})$ if $0 < \alpha < 1$ and $|\widehat{\mu}^{\mathrm{EB}} - \mu| = O(1)$ if $\alpha \geq 1$.*

Finally, for the subset-of-signals problem with m signals, the estimator $\widehat{\mu}^{\mathrm{EB}}$ achieves near-optimal error bounds, including the correct phase transition, across all ranges of m:

Theorem 1.2 *Let $\sigma_i \in [\sigma_{\min}, \sigma_{\max}]$ for all $i \in [n]$, and the number of signals be m in the subset-of-signals problem. With probability at least $1 - \delta$, estimator $\widehat{\mu}^{\mathrm{EB}}$ in (3) achieves*

$$|\widehat{\mu}^{\mathrm{EB}} - \mu| = \begin{cases} \widetilde{O}\left(\left(\frac{n}{m^4}\right)^{1/2}\right) & \text{if } \widetilde{O}(1) \ll m \leq n^{1/4}, \\ \widetilde{O}\left(\left(\frac{n}{m^4}\right)^{1/6}\right) & \text{if } n^{1/4} \leq m \leq n, \end{cases}$$

where $\widetilde{O}(\cdot)$ hides polylogarithmic factors in $(n, \frac{\sigma_{\max}}{\sigma_{\min}}, \frac{1}{\delta})$.

Compared with the known minimax risk (2) for the subset-of-signals problem, Theorem 1.2 shows that the estimator $\widehat{\mu}^{\mathrm{EB}}$ achieves the minimax rate up to logarithmic factors under the mild assumption $\log \frac{\sigma_{\max}}{\sigma_{\min}} = O(\mathrm{polylog}(n))$. However, like many empirical Bayes approaches, a particularly appealing advantage of our estimator is that it is *adaptive* and *parameter-free*: this estimator requires no tuning parameters (we set $\sigma_{\min} = 0$ and $\sigma_{\max} = \infty$ in our experiments; see Section 4), and adapts to the signal size m. These features make $\widehat{\mu}^{\mathrm{EB}}$ especially attractive in practice. In the practical scenario where the joint MLE (3) is computed only approximately, a similar guarantee remains valid for approximate MLEs; see Remark 1.1.

Compared with existing empirical Bayes approaches, while the idea of applying an NPMLE-based estimator is classical in the empirical Bayes literature (Kiefer and Wolfowitz, 1956; Robbins, 1956), our objective in (3) is conceptually different. In existing theoretical studies of empirical Bayes, one typically considers independent observations $X_i \sim P_{\theta_i}$ (the compound setting) and aims to estimate the parameter vector $(\theta_1, \ldots, \theta_n)$. The empirical Bayes approach proceeds by first estimating the empirical distribution $G_n := \frac{1}{n}\sum_{i=1}^n \delta_{\theta_i}$ (e.g., via the NPMLE), and then applying the resulting learned prior to estimate θ through the corresponding Bayes rule. By contrast, our procedure likewise replaces heterogeneity in the *nuisance parameter* $(\sigma_1, \ldots, \sigma_n)$ with a mixing distribution G_n and learns it from the data, but does not invoke explicit posterior inference. It is therefore an interesting theoretical observation that empirical Bayes methods are near-optimal in our setting, even though the procedure does not actually invoke a "Bayes" component.

We also provide some discussion on the hyperparameters $(\sigma_{\min}, \sigma_{\max})$ and computation. We acknowledge that our estimator requires additional knowledge of $(\sigma_{\min}, \sigma_{\max})$, which is not needed by previous estimators. This requirement is, however, intrinsic to our approach: without it, the solution to (3) would be degenerate, since one could assign a positive mass to $\widehat{G}(\{0\})$ and choose $\widehat{\mu} \in \{X_1, \ldots, X_n\}$, thereby making the likelihood unbounded. Fortunately, the error dependence on $\sigma_{\min}$ in Theorem 1.2 is only $\mathrm{polylog}(\frac{\sigma_{\max}}{\sigma_{\min}})$ and thus very mild, and our numerical experiments show that even setting $\sigma_{\min} = 0$ and $\sigma_{\max} = \infty$ still yields sound estimation performance. As for computation, solving the optimization program in (3) is, unfortunately, a nonconvex problem involving the infinite-dimensional parameter $\widehat{G}$. In Section 4, we adopt a successive maximization strategy. For fixed μ, (3) reduces to a convex optimization problem in G, for which it is standard to apply a fully corrective Frank–Wolfe algorithm. For fixed G, (3) becomes a one-dimensional maximization problem in μ, which can be efficiently solved via grid search. Although both Frank–Wolfe and successive maximization may converge to local maxima, our numerical results indicate that the resulting $\widehat{\mu}^{\mathrm{EB}}$ is empirically close to the ideal MLE obtained when G_n is known exactly.

1.2. Outline of the proof

Theorem 1.1 follows from a key result concerning the density estimation performance of the MLE, which accurately estimates the true mixing density f_{μ,G_n} in Hellinger distance.

Theorem 1.3 *Let $\sigma_{\min} \le \sigma_i \le \sigma_{\max}$ for all $i \in [n]$, and $(\widehat{\mu}, \widehat{G})$ be the joint MLE in (3). Then with probability at least $1-\delta$,*

$$H^2(f_{\mu,G_n}, f_{\widehat{\mu},\widehat{G}}) \le \frac{C}{n}\left(\log^6\left(\frac{n\sigma_{\max}}{\sigma_{\min}}\right)\log\log\left(\frac{n\sigma_{\max}}{\sigma_{\min}}\right) + \log\frac{1}{\delta}\right).$$

Remark 1.1 *If $(\widehat{\mu}, \widehat{G})$ is an approximate MLE with $\prod_{i=1}^n f_{\widehat{\mu},\widehat{G}}(X_i) \ge \beta \cdot \max_{\mu,G}\prod_{i=1}^n f_{\mu,G}(X_i)$ with $\beta \in (0,1]$, by Lemma 3.1, the Hellinger bound in Theorem 1.3 has an additive factor $O(\frac{1}{n}\log\frac{1}{\beta})$.*

Theorem 1.3 establishes an upper bound in Hellinger distance for estimating the density f_{μ,G_n}, which is the true average distribution of $X_1, \dots, X_n$. Obtaining density estimation guarantees under the Hellinger metric is standard in the statistical literature for analyzing the MLE, thanks to its well-known connection with the metric entropy of the underlying density class (Wong and Shen, 1995; van de Geer, 2000). Theorem 1.1 is then a direct consequence of Theorem 1.3 and a symmetrization inequality for the Hellinger distance in Lemma A.1, presented in Section A.

Upper bound on the covering number In the proof of Theorem 1.3, a key technical step is to analyze the metric entropy of the family of normal scale mixtures $\{f_{0,G} : \mathrm{supp}(G) \subseteq [\sigma_{\min}, \sigma_{\max}]\}$ under the Hellinger metric. Although such normal scale mixtures have been studied in the literature (Ghosal and van der Vaart, 2001, 2007; Ignatiadis and Sen, 2025), the existing arguments based on (local) moment matching (Ghosal and van der Vaart, 2001; Zhang, 2009) yield metric entropy bounds with a dependence of $\mathrm{poly}(\frac{\sigma_{\max}}{\sigma_{\min}})$; by contrast, we critically need to improve this dependence to $\mathrm{polylog}(\frac{\sigma_{\max}}{\sigma_{\min}})$. To achieve this, we introduce two simple yet effective ideas.

- First, instead of applying classical moment matching based on polynomials, we perform a generalized moment matching by constructing a suitable basis $\{a_k(\sigma), g_k(x)\}$ to approximate the normal density $\frac{1}{\sigma}\varphi(\frac{x}{\sigma})$ by a separable expansion $\sum_{k=1}^L a_k(\sigma)g_k(x)$.
- Second, instead of using the usual Taylor approximating polynomial for a given analytic function, we employ Chebyshev polynomials, which more accurately capture growth behavior on a Bernstein ellipse rather than on a disk in the complex plane.

These arguments establish the $\mathrm{polylog}(\frac{\sigma_{\max}}{\sigma_{\min}})$ dependence, which we elaborate in Section 3.

Upper bounding Hellinger modulus of continuity To prove Theorem 1.2, we need to upper bound the Hellinger modulus of continuity $\omega_{H^2,G}(t)$. We prove such a bound in the case when $G([0,1]) \ge p$ for some $p > 0$, which captures instances such as the subset-of-signals problem.

Lemma 1.1 *Let $t \in [0,1]$, and G be a prior distribution over $[0,\infty)$ with $G([0,1]) \ge p > 0$. Then for a universal constant $C > 0$,*

$$\omega_{H^2,G}(t) \le C\begin{cases}(\frac{t^3}{p^4})^{1/6} & \text{if } t \le p^{4/3},\\ (\frac{t^3}{p^4})^{1/2} & \text{if } p^{4/3} < t \le \frac{p}{C}.\end{cases} \tag{4}$$

Lemma 1.1 is a functional inequality that exploits the structure of normal scale mixtures and will be proved in Section C. We note that this inequality is not improvable in general, as witnessed by the example $G = p\delta_1 + (1-p)\delta_{c\mu/\sqrt{t}}$, with $\mu = \mu(p,t)$ matching the upper bound in Lemma 1.1; we verify the algebra in Section C. The condition $t = O(p)$ is also necessary: for the above choice of G, the same Section C verifies that $\omega_{H^2,G}(t)$ can be unbounded for $t \geq 3p$. We note that Theorem 1.2 follows directly from Theorem 1.1 and Lemma 1.1, with the subset-of-signals assumption giving $G_n([0,1]) \geq \frac{m}{n} =: p$. Corollary 1.1 similarly follows from Theorem 1.1 and specific calculations of $\omega_{H^2,G_n}(t)$ for the respective choices of G_n, which we defer to Section C.3.

2. Related work

Heteroskedastic mean estimation. The problem of estimating a common location parameter from heterogeneous observations has attracted significant recent attention in both the statistics and theoretical computer science communities. Pensia, Jog, and Loh (Pensia et al., 2022) initiated the systematic study of this problem, characterizing minimax rates for location estimation under sample-heterogeneous distributions and proposing estimators based on iterative trimming and median-of-means techniques. Devroye, Lattanzi, Lugosi, and Zhivotovskiy (Devroye et al., 2023) further developed this line of work, obtaining sharp minimax bounds for heteroskedastic mean estimation under bounded variance assumptions and establishing the phase transition at $m \asymp n^{1/4}$ in the subset-of-signals regime. Both works reveal a fundamental tension in this setting: the sample mean is efficient when variances are known but highly sensitive to large variances, while robust alternatives such as the median (Huber, 1964; Hampel et al., 1986) are resilient but often suboptimal.

The heteroskedastic setting studied here differs from the classical Huber contamination model (Huber, 1964, 1981), where an ε-fraction of observations are adversarially corrupted. In the contamination model, robust estimators aim for error scaling with ε (Diakonikolas et al., 2019; Lai et al., 2016; Chen et al., 2018), whereas in our setting the noise levels can vary continuously across samples, leading to different minimax rates and requiring distinct estimation strategies.

Robust statistics and breakdown point. Classical robust statistics provides a rich toolkit for handling outliers and heavy-tailed distributions. Foundational work by Tukey (Tukey, 1960, 1975), Huber (Huber, 1964, 1981), and Hampel (Hampel, 1971; Hampel et al., 1986) established the concepts of breakdown point, influence function, and M-estimation that underpin modern robust methods. The median achieves the optimal breakdown point of $1/2$, meaning it remains bounded even when nearly half the observations are arbitrarily corrupted. Trimmed means (Bickel, 1965; Stigler, 1973) offer a tunable tradeoff between efficiency and robustness. However, these classical estimators do not achieve the optimal rates in the heteroskedastic setting characterized by (Pensia et al., 2022; Devroye et al., 2023), motivating the search for new approaches.

Empirical Bayes and compound decision theory. Empirical Bayes methodology, pioneered by Robbins (Robbins, 1951, 1956), provides a principled framework for estimation problems with many latent parameters. In the compound decision setting, one observes $X_i \sim P_{\theta_i}$ for $i = 1, \ldots, n$ and aims to estimate the parameter vector $(\theta_1, \ldots, \theta_n)$. The empirical Bayes approach treats the parameters as draws from an unknown prior G and estimates G from the marginal distribution of the observations. The nonparametric maximum likelihood estimator (NPMLE) for the mixing distribution, introduced by Kiefer and Wolfowitz (Kiefer and Wolfowitz, 1956), plays a central role in this framework. Laird (Laird, 1978) developed practical algorithms for computing the NPMLE, and

Lindsay (Lindsay, 1983) established that it is a discrete measure with at most n support points. More recently, Polyanskiy and Wu (Polyanskiy and Wu, 2020) discovered a remarkable self-regularization property of the NPMLE: even without explicit regularization, the NPMLE for subgaussian mixtures has $O(\log n)$ support points with high probability. The REBayes package (Koenker and Mizera, 2014; Koenker and Gu, 2017) provides efficient implementations of NPMLE-based procedures.

A major breakthrough in the theoretical study of empirical Bayes was made by Jiang and Zhang (Jiang and Zhang, 2009) in the Gaussian compound decision problem. This work inspired a flurry of subsequent studies in Gaussian (Brown and Greenshtein, 2009; Saha and Guntuboyina, 2020; Polyanskiy and Wu, 2021; Ghosh et al., 2025) and Poisson models (Brown et al., 2013; Polyanskiy and Wu, 2021; Shen and Wu, 2026; Jana et al., 2023, 2025; Han et al., 2025). We also refer to a recent book (Efron, 2024) for an overview of empirical Bayes approaches.

Our approach differs from classical empirical Bayes in an important way: while standard empirical Bayes treats the heterogeneous parameters $(\sigma_1, \dots, \sigma_n)$ as the primary estimation target and applies Bayes rules, we treat them as nuisance parameters and focus on estimating the common mean μ through a profile likelihood (Murphy and Van der Vaart, 2000; Severini, 2000). This perspective connects our work to the literature on profile likelihood inference with infinite-dimensional nuisance parameters (Murphy and Van der Vaart, 1997; Shen, 1997). Nevertheless, our approach is still based on an important idea of empirical Bayes, where we replace heterogeneity by a mixing distribution and learn it from data.

Density estimation and sieve MLE theory. The analysis of our estimator relies on density estimation theory for the MLE in mixture models. The foundational work of Wong and Shen (Wong and Shen, 1995) established probability inequalities for likelihood ratios and convergence rates for sieve MLEs, showing that the rate is governed by the Hellinger metric entropy of the density class. Van de Geer (van de Geer, 2000) developed a comprehensive theory of M-estimation with empirical processes. The similar idea is used in our entropic upper bound in Lemma 3.1.

Specializing to the MLE in mixture models, its density estimation performance under the Hellinger distance has been studied in normal location mixture (Ghosal and van der Vaart, 2001, 2007; Zhang, 2009), normal scale mixture (Ghosal and van der Vaart, 2001, 2007; Ignatiadis and Sen, 2025), and Poisson mixture (Shen and Wu, 2026; Jana et al., 2025) models. Existing approaches are mainly based on the idea of (local) moment matching and polynomial approximations (Birgé and Massart, 1998; Genovese and Wasserman, 2000), usually leading to $\mathsf{polylog}(\frac{1}{\varepsilon})$ dependence on the radius ε in the metric entropy bound. However, the existing dependence on $\frac{\sigma_{\max}}{\sigma_{\min}}$ in the normal scale mixture is not tight, and our analysis improves the dependence to logarithmic using generalized moment matching and Chebyshev polynomial approximations, which better capture the analytic structure of the Gaussian kernel. This improvement is crucial for obtaining minimax-optimal rates that are robust to the ratio $\frac{\sigma_{\max}}{\sigma_{\min}}$.

Modulus of continuity and minimax lower bounds. Our upper bounds are expressed in terms of the Hellinger modulus of continuity of the location family, following the geometric approach to minimax theory developed by Donoho and Liu (Donoho and Liu, 1991). Le Cam's two-point method (Le Cam, 1973, 1986) provides matching lower bounds: if $H^2(f_{\mu_1,G}, f_{\mu_2,G}) \le 1/n$, then one cannot reliably distinguish μ_1 from μ_2, yielding a lower bound of $\omega_{H^2,G}(1/n)$ for estimating μ. Therefore, Theorem 1.1 can be viewed as a duality result to Le Cam's lower bound; such duality has appeared in the literature for linear functionals (Donoho and Liu, 1991; Juditsky and Nemirovski, 2009; Polyanskiy and Wu, 2026) and more recently for mean estimation in location families (Comp-

ton and Valiant, 2025). There the mixing distribution G_n is known, and the estimator is of a different Birgé–Le Cam type. In comparison, our empirical Bayes perspective shows that the profile MLE attains a comparable bound even when G_n is unknown.

Computation. The optimization problem in Equation (3) is nonconvex due to the joint optimization over (μ, G), but for fixed μ reduces to a convex problem in G. Algorithms for computing the NPMLE in mixture models include the EM algorithm (Dempster et al., 1977; Laird, 1978), interior point methods (Lesperance and Kalbfleisch, 1992), and the Frank–Wolfe algorithm (Frank and Wolfe, 1956; Jaggi, 2013). Koenker and Mizera (Koenker and Mizera, 2014) reformulated the NPMLE computation as a convex optimization problem, enabling efficient solutions for large-scale problems. Our implementation uses a successive maximization strategy that alternates between Frank–Wolfe updates for G and line search for μ; similar block coordinate approaches are common in semiparametric estimation (Murphy et al., 1999; Van der Vaart, 1998).

3. Density Estimation: Proof of Theorem 1.3

In this section we prove the density estimation guarantee of the MLE in Theorem 1.3. We begin with a general entropic upper bound of the Hellinger distance in the compound (or i.n.i.d.) setting, and then present a metric entropy upper bound for the normal scale mixture with a fixed mean. The remaining details will be devoted to allowing μ to vary; we defer them to the appendix.

3.1. Entropic Upper Bound of the MLE in the Compound Setting

In this section we review and state the classical entropic upper bound for density estimation via the MLE, in a general i.n.i.d. (independent and *non-identically* distributed) setting. Let $X_1, \dots, X_n$ be independent, with $X_i \sim P_i$. Given a class $\mathcal{P}$ of distributions such that $\overline{P} := \frac{1}{n}\sum_{i=1}^n P_i \in \mathcal{P}$, let $\widehat{P}$ be a β-approximate MLE with $\prod_{i=1}^n \widehat{P}(X_i) \geq \beta \cdot \max_{P\in\mathcal{P}} \prod_{i=1}^n P(X_i)$. To bound the Hellinger distance $H(\widehat{P}, \overline{P})$, we need the following notation for the metric entropy. For $\delta > 0$, define the localized family $\mathcal{P}(\overline{P}, \delta) := \{P \in \mathcal{P} : H(P, \overline{P}) \leq \delta\}$.

Definition 3.1 (Bracketing Number) *The (δ-local) Hellinger bracketing number at scale ε around $\overline{P}$, denoted by $N_{[]}(\varepsilon, \mathcal{P}(\overline{P}, \delta), H)$, is the smallest integer N such that there exist N brackets of (positive) measures $[P_1^{\mathrm{L}}, P_1^{\mathrm{U}}], \dots, [P_N^{\mathrm{L}}, P_N^{\mathrm{U}}]$ such that $\max_{i\in[N]} H(P_i^{\mathrm{L}}, P_i^{\mathrm{U}}) \leq \varepsilon$, and for every $P \in \mathcal{P}(\overline{P}, \delta)$, there exists $i \in [N]$ such that $P_i^{\mathrm{L}} \leq P \leq P_i^{\mathrm{U}}$ (with $\leq$ being pointwise).*

In the above definition, we extend the definition of Hellinger distance between two probability measures in a natural way to general measures μ, ν by $H^2(\mu, \nu) := \int(\sqrt{\mathrm{d}\mu} - \sqrt{\mathrm{d}\nu})^2$. The following result states a general entropic upper bound on the Hellinger distance $H(\widehat{P}, \overline{P})$.

Lemma 3.1 *There exists a universal constant $C > 0$ such that the following holds. Let*

$$\Psi(\delta) \geq \int_{\delta^2/C}^{\delta} \sqrt{\log N_{[]}(u, \mathcal{P}(\overline{P}, \delta), H)}\mathrm{d}u \vee \delta$$

be any function such that $\Psi(\delta)/\delta^2$ is non-increasing in δ, and $\delta_n > 0$ satisfy $\sqrt{n}\delta_n^2 \geq C\Psi(\delta_n)$. Then for all $\delta \geq \delta_n$, any β-approximate MLE $\widehat{P}$ satisfies

$$\mathbb{P}\left(H^2(\widehat{P}, \overline{P}) \geq \delta^2 + \frac{C}{n}\log\frac{1}{\beta}\right) \leq C\exp\Big(-\frac{n\delta^2}{C^2}\Big).$$

In the i.i.d. case where $P_i \equiv P$, Lemma 3.1 reduces to the classical upper bounds on $H^2(\widehat{P}, P)$ in (Wong and Shen, 1995; van de Geer, 2000). The same proof technique extends to the i.n.i.d. setting, where the MLE $\widehat{P}$ is instead close to the average distribution $\overline{P}$. In heteroskedastic mean estimation, this average distribution is $\overline{P} = \frac{1}{n}\sum_{i=1}^n \mathcal{N}(\mu, \sigma_i^2) = f_{\mu, G_n}$, which motivates the form of Theorem 1.3. For completeness, we include the proof of Lemma 3.1 in the appendix, as the i.n.i.d. case in (van de Geer, 2000, Chapter 8.3) does not cover this precise setting.

3.2. Improved Covering Results for the Normal Scale Mixture

The central metric entropy bound in this section is the following:

Theorem 3.1 *Let $\mathcal{P} = \{f_{\mu,G} : \mu \in \mathbb{R}, \mathrm{supp}(G) \subseteq [\sigma_{\min}, \sigma_{\max}]\}$ such that $\overline{P} \in \mathcal{P}$. There exists a universal constant $C > 0$ such that for $0 < \varepsilon \le \frac{1}{8}$,*

$$\log N_{[]}(\varepsilon, \mathcal{P}(\overline{P}, \frac{1}{8}), H) \le C \log^6\left(\frac{\sigma_{\max}}{\varepsilon \sigma_{\min}}\right) \log\log\left(\frac{\sigma_{\max}}{\varepsilon\sigma_{\min}}\right).$$

In conjunction with Lemma 3.1, Theorem 3.1 gives the target upper bound of Theorem 1.3. The main technical challenge in the proof of Theorem 3.1 is to derive the bracketing entropy bound for the class of *normal scale mixtures*, i.e., the subset of $\mathcal{P}$ with $\mu = 0$ and a general mixture distribution G on $[\sigma_{\min}, \sigma_{\max}]$. Specifically, using a reparametrization $t = 1/\sigma$ and standard truncation arguments (deferred to Section B.2), Theorem 3.1 follows from the following technical lemma.

Lemma 3.2 *Let $0 < t_{\min} \le 1 \le t_{\max}$ and $0 < x_{\min} \le 1 \le x_{\max}$. Let $\mathcal{F}$ be the class of all functions $\mathbb{E}_{t\sim H}[t\exp(-t^2x^2/2)]$ with $H \in \mathcal{P}([t_{\min}, t_{\max}])$, and $L_\infty([x_{\min}, x_{\max}])$ denotes the sup norm on the interval $[x_{\min}, x_{\max}]$. Then for $\varepsilon \in (0, 1/2)$,*

$$\log N(\varepsilon, \mathcal{F}, L_\infty([x_{\min}, x_{\max}])) \le C \log^6\left(\frac{t_{\max}x_{\max}}{\varepsilon t_{\min}x_{\min}}\right) \log\log\left(\frac{t_{\max}x_{\max}}{\varepsilon t_{\min}x_{\min}}\right).$$

In Lemma 3.2, the polylog$(\frac{1}{\varepsilon})$ dependence on ε is perhaps unsurprising as $\mathcal{F}$ is effectively a parametric family by discretizing the support of H; the key challenge is to obtain a logarithmic dependence on $\frac{t_{\max}}{t_{\min}}$ and $\frac{x_{\max}}{x_{\min}}$. This is precisely where existing work based on polynomial approximation or Taylor series fail to capture, as detailed in the following sections.

3.2.1. GENERALIZED MOMENT MATCHING

In the literature for normal location or scale mixtures, the prevalent approach for establishing metric entropy bounds is through the idea of *moment matching*. Specifically, it is shown that if H and H' have the same first L moments, then the difference $\|\mathbb{E}_{t\sim H}[t\exp(-t^2x^2/2)] - \mathbb{E}_{t\sim H'}[t\exp(-t^2x^2/2)]\|_\infty$ decays exponentially in L. Since Carathéodory's theorem shows that matching first L moments can be realized by a discrete distribution with at most $O(L)$ atoms, covering such discrete distributions typically leads to a metric entropy bound of $\widetilde{O}(L)$. This idea has been used for approximating normal scale mixtures in (Ghosal and van der Vaart, 2001, 2007; Ignatiadis and Sen, 2025); an improved bound is also obtained via *local moment matching* in (Jiang and Zhang, 2009) for normal location mixtures, with extensions to Poisson mixtures in (Shen and Wu, 2026).

However, such moment matching approaches (even the localized version) in normal scale mixtures suffer from a polynomial dependence on the ratio $\frac{t_{\max}}{t_{\min}}$. The underlying mathematical reason is that, such moment matching arguments rely on the following polynomial approximation

$$\sup_{x\in[x_{\min},x_{\max}],t\in[t_{\min},t_{\max}]}\left|te^{-\frac{t^2x^2}{2}}-\sum_{k=1}^{L}g_k(x)t^k\right|\le\varepsilon,$$

and this polynomial is usually chosen to be a truncation of Taylor series. Based on this approximation, it is clear that $\mathbb{E}_{t\sim H}[t\exp(-t^2x^2/2)]$ approximately depends only on $\mathbb{E}_{t\sim H}[t^k], k=1,\dots,L$. However, even for a fixed x, approximation theory tells that the degree L must have a polynomial dependence on $(t_{\min},t_{\max})$. For example, (Aggarwal and Alman, 2022) shows that to approximate e^{-t} with a uniform approximation error ε on $t\in[0,B]$, the degree of the polynomial must be at least $\widetilde{\Omega}(\sqrt{B})$, exhibiting a polynomial dependence on B.

To address this issue, we make a simple yet important observation that we may apply a generalized moment matching with suitably chosen basis functions:

$$\sup_{x\in[x_{\min},x_{\max}],t\in[t_{\min},t_{\max}]}\left|te^{-\frac{t^2x^2}{2}}-\sum_{k=1}^{L}a_k(t)g_k(x)\right|\le\varepsilon. \tag{5}$$

Here, by choosing $a_k(t)$ beyond polynomials, we may find a better approximating basis of $t\exp(-t^2x^2/2)$ to achieve a smaller approximation error. The following lemma is an easy metric entropy bound if we can construct bounded functions $\{a_k(t),g_k(x)\}_{k=1}^L$ such that (5) holds.

Lemma 3.3 *Suppose* (5) *holds with* $|a_k(t)|+t|a_k'(t)|\le A$ *for all* $t\in[t_{\min},t_{\max}]$ *and* $|g_k(x)|\le G$ *for all* $x\in[x_{\min},x_{\max}]$. *Then* $\log N(2\varepsilon,\mathcal{F},L_\infty([x_{\min},x_{\max}]))\le CL\log\frac{AGL\log(t_{\max}/t_{\min})}{\varepsilon}$.

To choose the basis, we use the natural idea of writing $u=\log x+\log t$ and try to approximate $\exp(-t^2x^2/2)=\exp(-\frac{1}{2}e^{2u})\approx P_L(u)$, where P_L is a polynomial of degree at most L. Expanding $P_L(u)$ into a bivariate polynomial in $(\log x,\log t)$, we obtain the basis functions $(a_k(t),g_k(x))$ in (5) in the form of $t(\log t)^i$ and $(\log x)^j$. Therefore, it remains to find a uniform polynomial approximation for $\exp(-\frac{1}{2}e^{2u})$ on $u\in[\log(x_{\min}t_{\min}),\log(x_{\max}t_{\max})]$, or equivalently, for $\exp(-Ke^{\lambda u})$ on $u\in[-1,1]$ after translation and scaling, with $K=\frac{1}{2}x_{\min}x_{\max}t_{\min}t_{\max}>0$ and $\lambda=\log\frac{x_{\max}t_{\max}}{x_{\min}t_{\min}}$.

3.2.2. CHEBYSHEV APPROXIMATION

In this section we solve the polynomial approximation problem of approximating $\exp(-Ke^{\lambda u})$ on $[-1,1]$. As commonly used in the moment matching literature, a first natural idea is to apply the Taylor approximating polynomial of $\exp(-Ke^{\lambda u})$ at $u=0$. However, this leads to a poor approximation: after some algebra, we can show that the n-th Taylor coefficient is $a_n=\frac{\lambda^n}{n!}e^{-K}B_n(-K)$, where $B_n(x)=\sum_{k=0}^n S(n,k)x^k$ is the Bell/Touchard polynomial and $S(n,k)$ is the Stirling number of the second kind. Since $|B_n(-K)|$ grows at a speed faster than K^n, we see that $|a_n|\to0$ only if $n\gg\lambda K$. However, since K is polynomial in $(x_{\min},x_{\max},t_{\min},t_{\max})$, this means that the degree of the Taylor approximating polynomial must also be a polynomial in $(x_{\min},x_{\max},t_{\min},t_{\max})$. To improve over this dependence, we critically make use of the *Chebyshev approximation* to achieve an approximation error *independent of* K, as summarized in the following lemma.

Lemma 3.4 *Let $h(x) = \exp(-Ke^{\lambda x})$ on $[-1, 1]$, with $K > 0$ and $\lambda > 0$. Let P_L be the degree-L Chebyshev polynomial of h. Then $\|h - P_L\|_\infty \le \varepsilon$ for $L = O(\log(\frac{1}{\varepsilon}) + \lambda \log(\frac{\lambda}{\varepsilon}))$.*

Proof The proof bounds the Chebyshev approximation error using the Bernstein ellipse theorem.

Lemma 3.5 (Bernstein Ellipse Theorem, (Trefethen, 2019, Theorem 8.2)) *Let h be analytic in the open Bernstein ellipse $E_\rho = \{z \in \mathbb{C} : z = \frac{1}{2}(\rho e^{i\theta} + \rho^{-1} e^{-i\theta}), \theta \in [0, 2\pi]\}$ for some $\rho > 1$, and suppose $|h(z)| \le M$ on E_ρ. Let P_L be the degree-L Chebyshev truncation of h on $[-1, 1]$. Then $\|h - P_L\|_\infty \le \frac{2M}{\rho - 1}\rho^{-L}$.*

To conclude from Lemma 3.5, note that the Bernstein ellipse E_ρ is contained in the strip $|\Im z| \le \frac{1}{2}(\rho - \rho^{-1}) \le \rho - 1$ for $\rho \in [1, 2]$. In addition, $|h(z)| = \exp(-Ke^{\lambda x}\cos(\lambda y))$ for $z = x + iy$. Therefore, for $\rho = \min\{2, 1 + \frac{\pi}{2\lambda}\}$, we have $\cos(\lambda y) \ge 0$ for $|y| \le \rho - 1$ and thus $|h(z)| \le 1$ on E_ρ. Now Lemma 3.4 directly follows from Lemma 3.5. ■

Importantly, the final polynomial degree L in Lemma 3.4 is only polynomial in λ, and thus poly-logarithmic in $\frac{x_{\max} t_{\max}}{x_{\min} t_{\min}}$; this is the target feature of Lemma 3.2. It is interesting to note why Chebyshev approximation performs better than Taylor approximation here. In fact, Taylor approximation gives a convergence rate ρ^{-n} when $|h|$ is bounded on the disc $|z| = \rho$, whereas Chebyshev approximation operates on a different geometry E_ρ which more tightly envelops the desired interval. Since $|h(z)|$ could grow rapidly along the imaginary axis, here the Bernstein ellipse is much more suitable than the disc.

4. Numerical experiments and algorithms

In this section, we summarize important properties of the NPMLE that are well-known in the empirical Bayes literature (Lindsay, 1983, 1995), specialized to normal scale mixtures. We also detail our computational approach for finding the NPMLE based on the Frank–Wolfe procedure adapted from (Han et al., 2025), and numerically compare our approach to previous methods in the literature such as median and iterative truncation (Liang and Yuan, 2020; Pensia et al., 2022).

4.1. Basic properties of the NPMLE

The main difficulty in solving the MLE (3) is to solve for the NPMLE $\widehat{G}$ once $\widehat{\mu}$ is fixed. This section describes some basic properties of $\widehat{G}$, and WLOG we assume that $\widehat{\mu} = 0$. Since the log-likelihood is concave in G, it is well-known (Lindsay, 1983) that the KKT condition for $\widehat{G}$ is

$$D_{\widehat{G}}(\sigma) := D_{\widehat{G}}(\sigma; X^n) := \frac{1}{n}\sum_{i=1}^{n} \frac{\frac{1}{\sigma}\varphi(\frac{X_i}{\sigma})}{f_{0,\widehat{G}}(X_i)} \le 1, \quad \forall \sigma \ge 0, \tag{6}$$

and equality holds iff $\sigma \in \operatorname{supp}(\widehat{G})$. Based on this, we have the following structural result.

Lemma 4.1 *Suppose $X_i \neq 0$ for all $i \in [n]$. Then the NPMLE $\widehat{G}$ is a discrete distribution with $|\operatorname{supp}(\widehat{G})| \le |\{|X_1|, \dots, |X_n|\}|$, $\operatorname{supp}(\widehat{G}) \subseteq [\min_{i\in[n]} |X_i|, \max_{i\in[n]} |X_i|]$, and unique.*

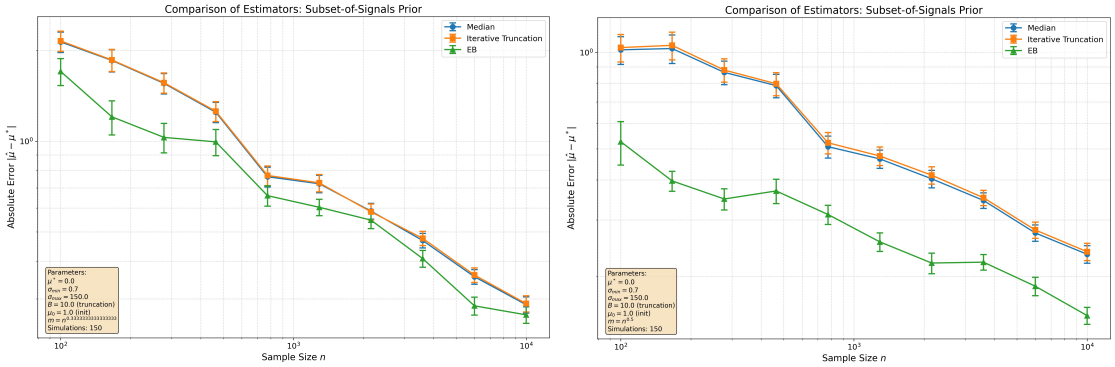
4.2. Successive maximization algorithm for the joint MLE

Motivated by the NPMLE properties in Lemma 4.1, especially the discrete nature of $\widehat{G}$, we use a fully-corrective Frank–Wolfe algorithm (Frank and Wolfe, 1956) (also known as the vertex-direction method in the mixture model literature) to compute $\widehat{G}$ with fixed $\widehat{\mu}$, in a similar manner to (Han et al., 2025). Specifically, given an initialization $\widehat{G}_0$, for $t = 1, 2, \ldots$, we repeat the following steps until convergence: (1) find a new atom $\sigma_t \in \arg\max_{\sigma>0} D_{\widehat{G}_{t-1}}(\sigma; X^n - \widehat{\mu})$ via a grid search, and then (2) add σ_t to the support of $\widehat{G}$ and optimize over the weights via a convex program:

$$\widehat{G}_t = \underset{\mathrm{supp}(G)\subseteq \mathrm{supp}(\widehat{G}_{t-1})\cup\{\sigma_t\}}{\arg\max} \frac{1}{n}\sum_{i=1}^{n} \log f_{\widehat{\mu},G}(X_i).$$

We note that in our implementation we do not enforce the constraints $\mathrm{supp}(\widehat{G}) \subseteq [\sigma_{\min}, \sigma_{\max}]$; this is because by Lemma 4.1, the data X^n have already constrained the support of $\widehat{G}$. Based on the Frank–Wolfe algorithm for fixed $\widehat{\mu}$, our final algorithm for joint optimization of $(\widehat{\mu}, \widehat{G})$ proceeds via successive maximization. Given an initialization $\widehat{\mu}_0$ (chosen as the median in our implementation), for $t = 1, 2, \ldots$, we iterate the following steps until convergence: (1) run the Frank–Wolfe algorithm with $\widehat{\mu}$ fixed at $\widehat{\mu}_{t-1}$ to obtain $\widehat{G}_t$; (2) perform a grid search to find $\widehat{\mu}_t$ that maximizes the total likelihood for fixed $\widehat{G}_t$. Since the likelihood improves at each iteration, we terminate when the increase in likelihood falls below a prescribed threshold. Although the overall algorithm is susceptible to local maxima, our experiments show that it converges well in practice.

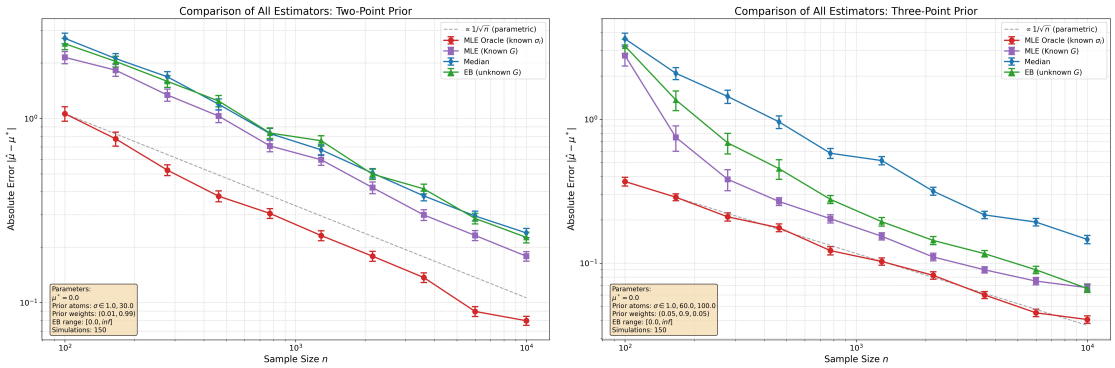
4.3. Experiments



(a) $m = n^{1/3}$ (b) $m = \sqrt{n}$

Figure 1: Average absolute estimation errors for $\widehat{\mu}^{\mathrm{EB}}$, sample median, and iterative truncation over 150 simulations, under the subset-of-signals prior $G_n = \frac{m}{n}\mathrm{Unif}([0.7, 1]) + \frac{n-m}{n}\mathrm{Unif}([1, 150])$, with different choices of (m, n).

In this section, we illustrate the empirical performance of our estimator $\widehat{\mu}^{\mathrm{EB}}$ on a variety of empirical distributions (priors) G_n of $(\sigma_1, \ldots, \sigma_n)$, including the subset-of-signals prior $G_n = \frac{m}{n}\mathrm{Unif}([\sigma_{\min}, 1]) + \frac{n-m}{n}\mathrm{Unif}([1, \sigma_{\max}])$, two-point and three-point scale mixture priors. We also consider the equal and quadratic variance models, but defer these results to Section D.2 for space considerations. On the subset-of-signals prior, we compare our estimator with two other estimators



(a) $G_n = \frac{1}{100}\delta_1 + \frac{99}{100}\delta_{30}$ (b) $G_n = \frac{1}{20}\delta_1 + \frac{9}{10}\delta_{60} + \frac{1}{20}\delta_{100}$

Figure 2: Average absolute estimation errors for $\widehat{\mu}^{\mathrm{EB}}$, sample median, and two oracle MLEs over 150 simulations, under two-point and three-point scale mixtures priors.

designed for the subset-of-signals problem: the sample median and the iterative truncation estimator in (Liang and Yuan, 2020). On the two-point and three-point priors, in addition to the median estimator, we compare with two *oracle* estimators: the linear estimator $\widehat{\mu}_{\mathrm{known}-\sigma}$ in (1) which we call the oracle MLE, and the MLE of μ in (3) with the true knowledge of G_n. We emphasize that apart from the grid search, number of starting atoms, and stopping criterion, our algorithm is completely free of tuning parameters.

Figure 1 and 2 display at the log scale the average absolute estimation error of μ over 150 simulations for both sets of experiments over a range of scales for n. In Figure 1, we set $m \in \{n^{1/3}, n^{1/2}\}$, and the hyperparameters $(B, \mu_0) = (10, 1)$ for the iterative truncation estimator. We observe that $\widehat{\mu}^{\mathrm{EB}}$ outperforms the other two estimators in both settings. In Figure 2, the scale mixture is a two-point prior $G_n = \frac{1}{100}\delta_1 + \frac{99}{100}\delta_{30}$ or three-point prior $G_n = \frac{1}{20}\delta_1 + \frac{9}{10}\delta_{60} + \frac{1}{20}\delta_{100}$. We observe that the performance of $\widehat{\mu}^{\mathrm{EB}}$, obtained without knowledge of G_n, is close to that of the MLE computed with knowledge of G_n. This indicates that the proposed successive maximization algorithm, in particular the Frank–Wolfe procedure used to estimate $\widehat{G}$, exhibits good empirical convergence, despite the potential for convergence to local maxima.

5. Conclusion and Open Problems

In this work, we demonstrate that a simple empirical Bayes approach based on maximum likelihood estimation achieves near-optimal performance for heteroskedastic mean estimation across a wide range of instances. Compared with traditional empirical Bayes approaches, our procedure replaces heterogeneity with a mixing distribution (prior) and learns it from the data, but does not invoke explicit posterior inference. Instead, the learned prior is treated as a nuisance estimate and is used in constructing our final mean estimator, which takes the form of a profile MLE. It is thus an interesting finding that this empirical Bayes approach still achieves near-optimal statistical performance in heteroskedastic mean estimation. Several interesting questions remain open.

Efficient computation. From a computational perspective, the optimization problem in (3) is non-convex (in μ) and infinite-dimensional (in G). While our numerical experiments suggest that suc-

cessive maximization with Frank–Wolfe performs well in practice, it would be valuable to develop provably efficient algorithms and/or to understand the landscape of this optimization problem.

Hyperparameters $(\sigma_{\min}, \sigma_{\max})$**.** Our theoretical guarantee involves a logarithmic factor in $\frac{\sigma_{\max}}{\sigma_{\min}}$, which primarily arises from our proof strategy for establishing the density estimation guarantee in Theorem 1.3. Indeed, when $\sigma_{\min} = 0$, the likelihood in (3) can be made unbounded by assigning positive mass to $\widehat{G}(\{0\})$ and choosing $\widehat{\mu} \in \{X_1, \dots, X_n\}$, in which case $H^2(f_{\widehat{\mu},\widehat{G}}, f_{\mu,G})$ is no longer small. However, this does *not* imply that the resulting estimator $\widehat{\mu}$ fails to estimate μ accurately. For example, in our numerical experiments, we always set $\sigma_{\min} = 0$ and $\sigma_{\max} = \infty$, yet the resulting estimator continues to perform well. As another illustration, the Le Cam–Birgé-type mean estimator studied in (Compton and Valiant, 2025) likewise satisfies $\widehat{\mu} \in \{X_1, \dots, X_n\}$, suggesting that estimators of this form can still achieve good performance. It is therefore an interesting question to remove the assumptions on $(\sigma_{\min}, \sigma_{\max})$.

Le Cam's lower bound for compound problems. Although $\omega_{H^2,G_n}(\frac{1}{n})$ is a minimax lower bound for the mean estimation problem where $X_1, \dots, X_n$ are i.i.d. drawn from f_{μ,G_n} with known G_n, the same claim is not rigorous in the compound setting. In the compound setting, Le Cam's method only gives the following lower bound: let $\mathbb{P}_{\mu,G_n} := \mathbb{E}_{\pi\sim\mathrm{Unif}(S_n)}[\bigotimes_{i=1}^n \mathcal{N}(\mu, \sigma^2_{\pi(i)})]$ be the n-dimensional "permutation mixture", then a minimax lower bound for mean estimation is

$$\sup\Big\{|\mu_1 - \mu_2| : \mu_1, \mu_2 \in \mathbb{R}, H^2(\mathbb{P}_{\mu_1,G_n}, \mathbb{P}_{\mu_2,G_n}) \le 2 - \Omega(1)\Big\}.$$

Assuming a mean-field approximation $\mathbb{P}_{\mu,G_n} \approx f^{\otimes n}_{\mu,G_n}$ for permutation mixtures, the above quantity would essentially reduce to $\omega_{H^2,G_n}(\frac{1}{n})$. However, although recent work (Han and Niles-Weed, 2024; Liang and Han, 2025) shows that such a mean-field approximation holds quantitatively and independently of the dimension n, these results only guarantee $\mathsf{TV}(\mathbb{P}_{\mu,G_n}, f^{\otimes n}_{\mu,G_n}) \le 1 - c$ for a small constant $c > 0$. Consequently, a direct application of the triangle inequality yields a vacuous bound (incurring a cost of $2 \cdot \mathsf{TV}$) and is therefore insufficient to establish the desired equivalence.

Multivariate settings. Extending our framework to multivariate settings, where observations $X_i \in \mathbb{R}^d$ have heteroskedastic covariance matrices, presents both statistical and computational challenges. On the statistical side, the density estimation-based proof strategy becomes unsuitable, since the Hellinger error typically scales exponentially in d. On the computational side, computing the NPMLE in high dimensions is challenging, and standard algorithms such as Frank–Wolfe become less effective.

Acknowledgments

Yanjun Han would like to thank Cun-Hui Zhang for helpful discussions at an early stage of this project. Abhishek Shetty would like to thank Shyam Narayanan, Ilias Diakonikolas, and Daniel Kane for helpful discussions. The numerical experiments were supported by NYU High Performance Computing (HPC) resources. We thank the reviewers for helpful comments.

References

Amol Aggarwal and Josh Alman. Optimal-degree polynomial approximations for exponentials and Gaussian kernel density estimation. In *37th Computational Complexity Conference*, pages 22:1–23, 2022.

Peter J Bickel. On some robust estimates of location. *The Annals of Mathematical Statistics*, 36(3): 847–858, 1965.

Lucien Birgé and Pascal Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

Lawrence D Brown and Eitan Greenshtein. Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics*, pages 1685–1704, 2009.

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

Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance and scatter matrix estimation under Huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960, 2018.

Spencer Compton and Gregory Valiant. Near-optimal mean estimation with unknown, heteroskedastic variances. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 194–200, 2024.

Spencer Compton and Gregory Valiant. Attainability of two-point testing rates for finite-sample location estimation. *arXiv preprint arXiv:2502.05730*, 2025.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–38, 1977.

Luc Devroye, Silvio Lattanzi, Gábor Lugosi, and Nikita Zhivotovskiy. On mean estimation for heteroscedastic random variables. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, volume 59, pages 1–20. Institut Henri Poincaré, 2023.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

David L Donoho and Richard C Liu. Geometrizing rates of convergence, II. *The Annals of Statistics*, 19(2):633–667, 1991.

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

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

Christopher R Genovese and Larry Wasserman. Rates of convergence for the Gaussian mixture sieve. *The Annals of Statistics*, 28(4):1105–1127, 2000.

S Ghosal and AW van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *Annals of Statistics*, 35(2):697–723, 2007.

Subhashis Ghosal and Aad W van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics*, 29(5): 1233–1263, 2001.

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.

Frank R Hampel. A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, 42(6):1887–1896, 1971.

Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, 1986.

Yanjun Han and Jonathan Niles-Weed. Approximate independence of permutation mixtures. *arXiv preprint arXiv:2408.09341*, 2024.

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.

Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Peter J Huber. *Robust Statistics*. John Wiley & Sons, 1981.

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

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435. PMLR, 2013.

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.

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

Anatoli B Juditsky and Arkadi S Nemirovski. Nonparametric estimation via convex programming. *Annals of statistics*, 37(5):2278–2300, 2009.

Jack Kiefer and Jacob Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 27(4):887–906, 1956.

Roger Koenker and Jiaying Gu. REBayes: an R package for empirical Bayes mixture methods. *Journal of Statistical Software*, 82(8):1–26, 2017.

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

Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674. IEEE, 2016.

Nan Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811, 1978.

Lucien Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53, 1973.

Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, 1986.

Mary L Lesperance and John D Kalbfleisch. An algorithm for computing the nonparametric MLE of a mixing distribution. *Journal of the American Statistical Association*, 87(417):120–126, 1992.

Yiguo Liang and Yanjun Han. Sharp mean-field analysis of permutation mixtures and permutation-invariant decisions. *arXiv preprint arXiv:2509.12584*, 2025.

Yingyu Liang and Hui Yuan. Learning entangled single-sample Gaussians in the subset-of-signals model. In *Conference on Learning Theory*, pages 2712–2737. PMLR, 2020.

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

Bruce G Lindsay. Mixture models: theory, geometry, and applications. Ims, 1995.

Susan A Murphy and Aad W Van der Vaart. Semiparametric likelihood ratio inference. *The Annals of Statistics*, 25(4):1471–1509, 1997.

Susan A Murphy and Aad W Van der Vaart. On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465, 2000.

Susan A Murphy, Aad W Van der Vaart, James Robins, and Joke PJ Slaets. Current status data with two competing risks. *The Annals of Statistics*, pages 1751–1770, 1999.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Estimating location parameters in sample-heterogeneous distributions. *Information and Inference: A Journal of the IMA*, 11(3):959–1036, 2022.

George Pólya and Gabor Szegö. Aufgaben und lehrsätze aus der analysis: Zweiter band: Funktionentheorie· nullstellen polynome· determinanten zahlentheorie. 1925.

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

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

Yury Polyanskiy and Yihong Wu. Dualizing Le Cam's method for functional estimation I: General theory. *The Annals of Statistics*, 54(1):1–24, 2026.

Herbert Robbins. Asymptotically subminimax solutions of compound statistical decision problems. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 131–148, 1951.

Herbert Robbins. An empirical Bayes approach to statistics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1:157–163, 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.

Thomas A Severini. Likelihood methods in statistics. 2000.

Xiaotong Shen. Methods of sieves and penalization. *The Annals of Statistics*, 25(6):2555–2591, 1997.

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

Stephen M Stigler. Simon Newcomb, Percy Daniell, and the history of robust estimation 1885–1920. *Journal of the American Statistical Association*, 68(344):872–879, 1973.

Lloyd N Trefethen. *Approximation theory and approximation practice, extended edition*. SIAM, 2019.

John W Tukey. A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics*, pages 448–485, 1960.

John W Tukey. Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians*, 2:523–531, 1975.

Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Aad W Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Wing Hung Wong and Xiaotong Shen. Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *The Annals of Statistics*, pages 339–362, 1995.

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

Appendix A. Proof of Theorem 1.1

To complete the proof of Theorem 1.1, we prove a symmetrization inequality which reduces possibly different scale mixtures to a single mixture.

Lemma A.1 *Let $\mu_1, \mu_2 \in \mathbb{R}$, and G_1, G_2 be two prior distributions over $[0, \infty)$. Then*

$$H^2(f_{\mu_1,G_1}, f_{\mu_2,G_2}) \geq \frac{1}{4} H^2(f_{|\mu_1-\mu_2|,G_1}, f_{-|\mu_1-\mu_2|,G_1}).$$

The proof of Theorem 1.1 follows by $(\mu_1, G_1) = (\mu, G_n)$ and $(\mu_2, G_2) = (\widehat{\mu}, \widehat{G})$ in Lemma A.1 and Theorem 1.3. Crucially, thanks to Lemma A.1, the prior G_2, which corresponds to the estimated prior $\widehat{G}$, does not appear on the right-hand side and is therefore not subject to any structural requirements such as those imposed on G_1. This asymmetry is the key technical reason why our estimator adapts to the signal size m in the subset-of-signals problem.

Proof [Proof of Lemma A.1] By translation and reflection invariance, we may assume that $\mu_1 = \mu \geq 0$, and $\mu_2 = 0$. The proof will rely on an easy symmetric relationship:

$$f_{\mu,G}(x) = \mathbb{E}_{\sigma\sim G}\left[\frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)\right] = \mathbb{E}_{\sigma\sim G}\left[\frac{1}{\sigma}\varphi\left(\frac{-x+\mu}{\sigma}\right)\right] = f_{-\mu,G}(-x). \tag{7}$$

We repeatedly apply this symmetry to get

$$\begin{aligned}
H^2(f_{\mu,G_1}, f_{0,G_2}) &= \int_{-\infty}^{\infty}\left(\sqrt{f_{\mu,G_1}(x)} - \sqrt{f_{0,G_2}(x)}\right)^2 \mathrm{d}x \\
&= \frac{1}{2}\int_{-\infty}^{\infty}\left[\left(\sqrt{f_{\mu,G_1}(x)} - \sqrt{f_{0,G_2}(x)}\right)^2 + \left(\sqrt{f_{\mu,G_1}(-x)} - \sqrt{f_{0,G_2}(-x)}\right)^2\right]\mathrm{d}x \\
&\overset{(7)}{=} \frac{1}{2}\int_{-\infty}^{\infty}\left[\left(\sqrt{f_{\mu,G_1}(x)} - \sqrt{f_{0,G_2}(x)}\right)^2 + \left(\sqrt{f_{\mu,G_1}(-x)} - \sqrt{f_{0,G_2}(x)}\right)^2\right]\mathrm{d}x \\
&\geq \frac{1}{4}\int_{-\infty}^{\infty}\left(\sqrt{f_{\mu,G_1}(x)} - \sqrt{f_{\mu,G_1}(-x)}\right)^2 \mathrm{d}x \\
&\overset{(7)}{=} \frac{1}{4}\int_{-\infty}^{\infty}\left(\sqrt{f_{\mu,G_1}(x)} - \sqrt{f_{-\mu,G_1}(x)}\right)^2 \mathrm{d}x \\
&= \frac{1}{4}H^2(f_{\mu,G_1}, f_{-\mu,G_1}),
\end{aligned}$$

and the inequality step uses $a^2 + b^2 \geq \frac{(a-b)^2}{2}$. This is the desired result. ■

Appendix B. Density Estimation: Proof of Theorem 1.3

In this section, we provide the proofs of the key lemmas used to establish Theorem 1.3.

B.1. Proof of Lemma 3.1

Our proof mostly follows the arguments in (van de Geer, 2000). We first mimic the arguments of (van de Geer, 2000, Lemma 4.1). Let P_n be the empirical distribution of $X_1, \ldots, X_n$, then

$$\begin{aligned}
-\frac{1}{n}\log\frac{1}{\beta} &\overset{(a)}{\leq} \int \log\frac{\widehat{P}}{\overline{P}}\mathrm{d}P_n \overset{(b)}{\leq} \frac{1}{2}\int \log\frac{\widehat{P}+\overline{P}}{2\overline{P}}\mathrm{d}P_n \\
&= \frac{1}{2}\int \log\frac{\widehat{P}+\overline{P}}{2\overline{P}}\mathrm{d}(P_n - \overline{P}) - \frac{1}{2}\mathsf{KL}\left(\overline{P}\middle\|\frac{\widehat{P}+\overline{P}}{2}\right) \\
&\overset{(c)}{\leq} \frac{1}{2}\int \log\frac{\widehat{P}+\overline{P}}{2\overline{P}}\mathrm{d}(P_n - \overline{P}) - \frac{1}{2}H^2\left(\overline{P}, \frac{\widehat{P}+\overline{P}}{2}\right),
\end{aligned}$$

where (a) follows from the definition of $\widehat{P}$ and $\overline{P} \in \mathcal{P}$, (b) is due to the concavity of $x \mapsto \log x$, and (c) uses the inequality $\mathsf{KL} \geq H^2$. Therefore,

$$H^2(\widehat{P}, \overline{P}) \overset{\text{(d)}}{\leq} 16H^2\Big(\overline{P}, \frac{\widehat{P}+\overline{P}}{2}\Big) \leq 16 \int \log \frac{\widehat{P}+\overline{P}}{2\overline{P}} \mathrm{d}(P_n - \overline{P}) + \frac{32}{n} \log \frac{1}{\beta}, \tag{8}$$

where (d) follows from (van de Geer, 2000, Lemma 4.2). This is called the "basic inequality" in (van de Geer, 2000).

The next step is to prove a high-probability upper bound on the integral in (8), using empirical processes. For $P \in \mathcal{P}$, let $Z_i(P) = \frac{1}{2} \log \frac{P+\overline{P}}{2\overline{P}}(X_i)$. For the subexponential norm $\rho^2(g) := 2\mathbb{E}[e^{|g|} - 1 - |g|]$, it holds that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \rho^2(Z_i(P)) &\overset{\text{(e)}}{\leq} \frac{1}{n}\sum_{i=1}^n 8\mathbb{E}\Big[\Big(\Big(\frac{P+\overline{P}}{2\overline{P}}(X_i)\Big)^{1/2} - 1\Big)^2\Big] \\ &= 8\mathbb{E}_{\overline{P}}\Big[\Big(\Big(\frac{P+\overline{P}}{2\overline{P}}(X_i)\Big)^{1/2} - 1\Big)^2\Big] = 8H^2\left(\frac{P+\overline{P}}{2}, \overline{P}\right) \leq 4H^2(P, \overline{P}),\end{aligned}$$

where (e) uses (van de Geer, 2000, Lemma 7.1) and that $Z_i(P) \geq -\frac{1}{2}\log 2$. In addition, for any bracket $[P^{\mathrm{L}}, P^{\mathrm{U}}]$, we have $Z_i(P^{\mathrm{L}}) \leq Z_i(P^{\mathrm{U}})$, and the similar arguments to (van de Geer, 2000, Lemma 7.3) yield

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \rho^2(Z_i(P^{\mathrm{U}}) - Z_i(P^{\mathrm{L}})) &\leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\Big(\Big(\frac{P^{\mathrm{U}}+\overline{P}}{P^{\mathrm{L}}+\overline{P}}(X_i)\Big)^{1/2} - 1\Big)^2\Big] \\ &= \mathbb{E}_{\overline{P}}\Big[\Big(\Big(\frac{P^{\mathrm{U}}+\overline{P}}{P^{\mathrm{L}}+\overline{P}}(X_i)\Big)^{1/2} - 1\Big)^2\Big] \\ &\leq 2H^2\left(\frac{P^{\mathrm{U}}+\overline{P}}{2}, \frac{P^{\mathrm{L}}+\overline{P}}{2}\right) \leq H^2(P^{\mathrm{U}}, P^{\mathrm{L}}).\end{aligned}$$

Therefore, compared with the Hellinger bracketing number $N_{[]}(\varepsilon, \mathcal{P}(\delta), H)$, the conditions of (van de Geer, 2000, Definition 8.1) are fulfilled with $(\delta, R) = (\varepsilon, 2\delta)$ and $F^c = \varnothing$. Since

$$\frac{1}{n}\sum_{i=1}^n Z_i(P) - \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n Z_i(P)\Big] = \frac{1}{2}\int \log \frac{P+\overline{P}}{2\overline{P}} \mathrm{d}(P_n - \overline{P})$$

coincides with the integral in the basic inequality (8), now (van de Geer, 2000, Theorem 8.13) (with $R = 2^{s+1}\delta$ and $a = c2^{2s}\delta^2\sqrt{n}$) yields

$$\mathbb{P}\Big(\exists P \in \mathcal{P} : \Big|\int \log \frac{P+\overline{P}}{2\overline{P}} \mathrm{d}(P_n - \overline{P})\Big| \geq c2^{2s}\delta^2 \wedge H(P, \overline{P}) \leq 2^{s+1}\delta\Big) \leq C\exp\Big(-\frac{n(2^s\delta)^2}{C^2}\Big)$$

for every $s \geq 0$ with $2^s\delta \leq 1$. Here the conditions in (van de Geer, 2000, Eqn. (8.43) and (8.44)) are clearly fulfilled for a small enough $c > 0$, and the condition

$$c2^{2s}\delta^2\sqrt{n} \geq C_0\Big(\int_{c2^{2s-6}\delta^2}^{2^{s+1}\delta} \sqrt{\log N_{[]}(u, \mathcal{P}(\overline{P}, 2^{s+1}\delta), H)}\mathrm{d}u \vee 2^{s+1}\delta\Big)$$

required in (van de Geer, 2000, Eqn. (8.45)) is also satisfied by the definition of δ_n, the assumption $\delta \geq \delta_n$, and adjusting constants.

Finally, for $\delta_0^2 := \delta^2 + \frac{64}{n}\log\frac{1}{\beta}$, by summing over $s = 0, 1, \ldots, S$ where $S = \min\{s : 2^s\delta_0 > 1\}$, a standard peeling argument similar to (van de Geer, 2000, Theorem 7.4) gives

$$\begin{aligned}
\mathbb{P}\Big(H^2(\widehat{P}, \overline{P}) \geq \delta_0^2\Big) &\leq \sum_{s=0}^{S} \mathbb{P}\Big(2^s\delta_0 \leq H(\widehat{P}, \overline{P}) \leq 2^{s+1}\delta_0\Big) \\
&\overset{(8)}{\leq} \sum_{s=0}^{S} \mathbb{P}\Big(\int \log\frac{\widehat{P}+\overline{P}}{2\overline{P}}\mathrm{d}(P_n - \overline{P}) \geq \frac{1}{16}\Big(2^{2s}\delta_0^2 - \frac{32}{n}\log\frac{1}{\beta}\Big) \wedge H(\widehat{P}, \overline{P}) \leq 2^{s+1}\delta_0\Big) \\
&\leq \sum_{s=0}^{S} \mathbb{P}\Big(\exists P \in \mathcal{P} : \Big|\int \log\frac{P+\overline{P}}{2\overline{P}}\mathrm{d}(P_n - \overline{P})\Big| \geq \frac{1}{32}2^{2s}\delta_0^2 \wedge H(P, \overline{P}) \leq 2^{s+1}\delta_0\Big) \\
&\leq \sum_{s=0}^{\infty} C\exp(-\frac{n(2^s\delta_0)^2}{C^2}) \leq C_0\exp(-\frac{n\delta_0^2}{C_0^2}),
\end{aligned}$$

for a suitably chosen universal constant $C_0 > 0$.

B.2. Proof of Theorem 3.1

In this section, we prove Theorem 3.1 based on Lemma 3.2. The proof decomposes into three steps.

B.2.1. L_∞ COVERING FOR SCALE MIXTURES

Our first step is to establish an upper bound on the covering number of *pure* scale mixtures

$$\mathcal{P}_0 = \{f_{0,G} : \mathrm{supp}(G) \subseteq [\sigma_{\min}, \sigma_{\max}]\}$$

under the L_∞ norm.

Lemma B.1 *For $\varepsilon \in (0, 1/2)$ and $\sigma_{\min} \leq 1$,*

$$\log N(\varepsilon, \mathcal{P}_0, L_\infty) \leq C\log^6\left(\frac{1}{\varepsilon\sigma_{\min}}\right)\log\log\left(\frac{1}{\varepsilon\sigma_{\min}}\right).$$

Proof The proof relies on Lemma 3.2, with a suitable reparametrization and truncation. Let $t = 1/\sigma$ and H be the pushforward measure of G under $\sigma \mapsto 1/\sigma$, then H is supported on $[0, \sigma_{\min}^{-1}]$, and

$$f_{0,G}(x) = \mathbb{E}_{\sigma\sim G}\Big[\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{x^2}{2\sigma^2})\Big] = \mathbb{E}_{t\sim H}\Big[\frac{t}{\sqrt{2\pi}}\exp(-\frac{t^2x^2}{2})\Big] =: \widetilde{f}_H(x).$$

To construct an ε-cover of $\{\widetilde{f}_H : \mathrm{supp}(H) \subseteq [0, \sigma_{\min}^{-1}]\}$, we invoke Lemma 3.2 with parameters

$$t_{\min} = \frac{\varepsilon}{4}, \quad t_{\max} = \frac{1}{\sigma_{\min}}, \quad x_{\min} = c_0(\varepsilon\sigma_{\min}^3)^{1/2}, \quad x_{\max} = C_0\varepsilon^{-1}\sqrt{\log(1/\varepsilon)}, \tag{9}$$

where $c_0, C_0 > 0$ are appropriate universal constants. By Lemma 3.2, there exists a cover $\mathcal{H}$ with $|\mathcal{H}| = \exp(O(\log^3\frac{1}{\varepsilon\sigma_{\min}}\log^4\log\frac{1}{\varepsilon\sigma_{\min}}))$ such that for every H supported on $[t_{\min}, t_{\max}]$, there exists $H_0 \in \mathcal{H}$ such that

$$\sup_{x\in[x_{\min}, x_{\max}]}\Big|\widetilde{f}_H(x) - \widetilde{f}_{H_0}(x)\Big| \leq \frac{\varepsilon}{4}.$$

We show that a similar inequality still holds even if $x \notin [x_{\min}, x_{\max}]$:

1. If $x \geq x_{\max}$, the function $t \mapsto te^{-t^2x^2/2}$ is decreasing for $x \geq 1/t_{\min}$, hence

$$0 \leq \widetilde{f}_H(x) \leq \frac{t_{\min}}{\sqrt{2\pi}} \exp\Big(-\frac{t_{\min}^2 x^2}{2}\Big) \leq \frac{\varepsilon}{4}$$

for a large constant C_0. The same bound also holds for $\widetilde{f}_{H_0}$, so that $|\widetilde{f}_H(x) - \widetilde{f}_{H_0}(x)| \leq \frac{\varepsilon}{4}$.

2. If $0 \leq x \leq x_{\min}$, note that

$$|\widetilde{f}_H'(x)| = x\mathbb{E}_{t\sim H}\Big[\frac{t^3}{\sqrt{2\pi}} \exp(-\frac{t^2x^2}{2})\Big] \leq \frac{x_{\min}}{\sqrt{2\pi}\sigma_{\min}^3}.$$

Therefore, for a small constant c_0,

$$\begin{aligned}
&|\widetilde{f}_H(x) - \widetilde{f}_{H_0}(x)| \\
&\leq |\widetilde{f}_H(x) - \widetilde{f}_H(x_{\min})| + |\widetilde{f}_H(x_{\min}) - \widetilde{f}_{H_0}(x_{\min})| + |\widetilde{f}_{H_0}(x_{\min}) - \widetilde{f}_{H_0}(x)| \\
&\leq 2 \cdot \frac{x_{\min}^2}{\sqrt{2\pi}\sigma_{\min}^3} + \frac{\varepsilon}{4} \leq \frac{\varepsilon}{2}.
\end{aligned}$$

3. If $x \leq 0$, since $\widetilde{f}_H(x)$ is an even function, the statement follows from symmetry.

Therefore, we have established that if $\mathrm{supp}(H) \subseteq [t_{\min}, t_{\max}]$, then $\|\widetilde{f}_H - \widetilde{f}_{H_0}\|_\infty \leq \frac{\varepsilon}{2}$.

Next we extend to the case where $\mathrm{supp}(H)$ could be $[0, t_{\max}]$. Note that if $\mathrm{supp}(H) \subseteq [0, t_{\min}]$, then $\widetilde{f}_H(x) \leq t_{\min} = \frac{\varepsilon}{4}$. Therefore, we can modify the cover $\mathcal{H}$ by a new cover

$$\mathcal{H}' = \Big\{(1-w)\delta_0 + wH_0 : H_0 \in \mathcal{H}, w \in \{0, \delta, 2\delta, \ldots, 1\}\Big\} \tag{10}$$

with $\delta = \frac{\varepsilon}{4t_{\max}}$, so that $|\mathcal{H}'| = O(|\mathcal{H}|/\delta)$. Now for any H supported on $[0, t_{\max}]$, we write $H = (1-w)H_1 + wH_2$ with $\mathrm{supp}(H_1) \subseteq [0, t_{\min}), \mathrm{supp}(H_2) \subseteq [t_{\min}, t_{\max}]$, and $w \geq 0$. Then for $H' = (1-w')\delta_0 + w'H_0 \in \mathcal{H}'$ with $|w - w'| \leq \delta$ and H_2 approximated by $H_0 \in \mathcal{H}$, we get

$$\begin{aligned}
\|\widetilde{f}_H - \widetilde{f}_{H'}\|_\infty &= \|(1-w)\widetilde{f}_{H_1} + w\widetilde{f}_{H_2} - w'\widetilde{f}_{H_0}\|_\infty \\
&\leq \|\widetilde{f}_{H_1}\|_\infty + \|\widetilde{f}_{H_2} - \widetilde{f}_{H_0}\| + |w - w'| \cdot \|\widetilde{f}_{H_0}\|_\infty \\
&\leq \frac{\varepsilon}{4} + \frac{\varepsilon}{2} + \delta\frac{t_{\max}}{\sqrt{2\pi}} \leq \varepsilon.
\end{aligned}$$

This shows that $\{\widetilde{f}_{H'} : H' \in \mathcal{H}'\}$ is an ε-cover of $\{\widetilde{f}_H : \mathrm{supp}(H) \subseteq [0, \sigma_{\min}^{-1}]\}$ under L_∞, with

$$\log|\mathcal{H}'| = \log|\mathcal{H}| + O\Big(\log\frac{1}{\delta}\Big) = O\Big(\log^3\frac{1}{\varepsilon\sigma_{\min}} \log^4 \log\frac{1}{\varepsilon\sigma_{\min}}\Big).$$

■

B.2.2. HELLINGER BRACKETING FOR SCALE MIXTURES

For the class of scale mixtures $\mathcal{P}_0$, by the standard approach in (Ghosal and van der Vaart, 2001), we can transform an L_∞-cover into a Hellinger bracket.

Lemma B.2 *For $\varepsilon \in (0, 1/2)$,*

$$\log N_{[]}(\varepsilon, \mathcal{P}_0, H) \le C \log^6\left(\frac{\sigma_{\max}}{\varepsilon \sigma_{\min}}\right) \log\log\left(\frac{\sigma_{\max}}{\varepsilon \sigma_{\min}}\right).$$

Proof Since the bracketing number is invariant by scaling $(\sigma_{\min}, \sigma_{\max})$ simultaneously, WLOG we may assume that $\sigma_{\min} \le 1$ so that Lemma B.1 can be applied. First, note that for (not necessarily probability) measures μ and ν,

$$H^2(\mu, \nu) = \int (\sqrt{\mathrm{d}\mu} - \sqrt{\mathrm{d}\nu})^2 \le \int |\mathrm{d}\mu - \mathrm{d}\nu| = \|\mu - \nu\|_1.$$

Therefore, $\log N_{[]}(\varepsilon, \mathcal{P}_0, H) \le \log N_{[]}(\varepsilon^2, \mathcal{P}_0, L_1)$. To construct a L_1-bracket, let $p_1, \dots, p_N$ be an η-cover of $\mathcal{P}_0$ in L_∞, with $\eta > 0$ to be specified later. We then construct the brackets $\{[P_i^L, P_i^U] : i \in [N]\}$ by $P_i^{\mathrm{L}} := \max\{p_i - \eta, 0\}, P_i^{\mathrm{U}} := \min\{p_i + \eta, R\}$, with $R(x) = \frac{1}{\sigma_{\min}} \exp(-\frac{x^2}{2\sigma_{\max}^2})$. Since every $p \in \mathcal{P}_0$ satisfies

$$p(x) = \mathbb{E}_{\sigma \sim G}\Big[\frac{1}{\sigma} \exp(-\frac{x^2}{2\sigma^2})\Big] \le \frac{1}{\sigma_{\min}} \exp(-\frac{x^2}{2\sigma_{\max}^2}) = R(x),$$

the inequality $\|p - p_i\|_\infty \le \eta$ indeed implies $P_i^{\mathrm{L}} \le p \le P_i^{\mathrm{U}}$. To upper bound the L_1 norm, for every $B \ge 0$ we have

$$\int_{\mathbb{R}} |P_i^{\mathrm{L}}(x) - P_i^{\mathrm{U}}(x)| \mathrm{d}x \le \int_{|x| \ge B} R(x) \mathrm{d}x + 2\eta B \le \frac{2\sigma_{\max}^2}{B\sigma_{\min}} \exp(-\frac{B^2}{2\sigma_{\max}^2}) + 2\eta B,$$

where the last inequality uses the Mills' ratio upper bound. Therefore, by choosing

$$B = C_0 \sigma_{\max} \sqrt{\log(\frac{\sigma_{\max}}{\sigma_{\min} \varepsilon^2})}, \qquad \eta = \frac{\varepsilon^2}{C_0 B}$$

with a large universal constant $C_0 > 0$, we ensure that $\|P_i^{\mathrm{U}} - P_i^{\mathrm{L}}\|_1 \le \varepsilon^2$. Finally, by Lemma B.1, we can take $\log N = O(\log^3(\frac{1}{\sigma_{\min}\eta}) \log^4 \log(\frac{1}{\sigma_{\min}\eta})) = O(\log^3(\frac{\sigma_{\max}}{\sigma_{\min}\varepsilon}) \log^4 \log(\frac{\sigma_{\max}}{\sigma_{\min}\varepsilon}))$. ■

B.2.3. LOCAL HELLINGER BRACKETING FOR LOCATION-SCALE MIXTURES

Finally, we extend the Hellinger bracketing for the smaller family $\mathcal{P}_0$ in Lemma B.2 to a *local* Hellinger bracketing for the entire family $\mathcal{P} = \{f_{\mu,G} : \mu \in \mathbb{R}, \mathrm{supp}(G) \subseteq [\sigma_{\min}, \sigma_{\max}]\}$ involving the location parameter in Theorem 3.1.

Proof [Proof of Theorem 3.1] Let $\overline{P} = f_{\mu_0, G_0} \in \mathcal{P}$ be the center of the local ball; by translation invariance we may assume that $\mu_0 = 0$. We first establish an upper bound on $|\mu|$ for every $f_{\mu,G} \in \mathcal{P}(\overline{P}, \delta)$. In fact, by the symmetrization inequality in Lemma A.1, we have

$$\delta \ge H(f_{0,G_0}, f_{\mu,G}) \ge \frac{1}{2} H(f_{\mu,G_0}, f_{-\mu,G_0}).$$

By the variational lower bound of the Hellinger distance in (18), we get

$$H^2(f_{\mu,G_0}, f_{-\mu,G_0}) \geq \frac{1}{4}(f_{\mu,G_0}([0,\infty)) - f_{-\mu,G_0}([0,\infty)))^2 \geq \frac{1}{4}\Big(2\Phi(\frac{|\mu|}{\sigma_{\max}}) - 1\Big)^2,$$

where Φ is the CDF of $\mathcal{N}(0,1)$. For $\delta \leq 1/8$, solving $|\mu|$ from the above two inequalities yields $|\mu| \leq C\sigma_{\max}$ for a universal constant $C > 0$, and therefore

$$\mathcal{P}(\overline{P}, \delta) \subseteq \big\{f_{\mu,G} : |\mu| \leq C\sigma_{\max}, \mathrm{supp}(G) \subseteq [\sigma_{\min}, \sigma_{\max}]\big\} =: \mathcal{P}_{\mathrm{loc}}.$$

To construct a Hellinger bracket for $\mathcal{P}_{\mathrm{loc}}$, we start from the Hellinger bracket $\{[P_i^{\mathrm{L}}, P_i^{\mathrm{U}}] : i \in [N]\}$ for $\mathcal{P}_0$ constructed in Lemma B.2, with $\|P_i^{\mathrm{U}} - P_i^{\mathrm{L}}\|_1 \leq \frac{\varepsilon^2}{4}$. Let $\eta > 0$ be a parameter to be specified later, and $\{\mu_1, \ldots, \mu_M\}$ be a uniform discretization of $[-C\sigma_{\max}, C\sigma_{\max}]$ with spacing 2η. Here $M = O(\frac{\sigma_{\max}}{\eta})$, and for $i \in [N], j \in [M]$, define

$$P_{i,j}^{\mathrm{L}}(x) = \inf_{|z|\leq\eta} P_i^{\mathrm{L}}(x - \mu_j - z), \quad P_{i,j}^{\mathrm{U}}(x) = \sup_{|z|\leq\eta} P_i^{\mathrm{U}}(x - \mu_j - z).$$

We claim that $\{[P_{i,j}^{\mathrm{L}}, P_{i,j}^{\mathrm{U}}] : i \in [N], j \in [M]\}$ is a bracket for $\mathcal{P}_{\mathrm{loc}}$. In fact, for any $f_{\mu,G} \in \mathcal{P}_{\mathrm{loc}}$, assume $P_i^{\mathrm{L}} \leq f_{0,G} \leq P_i^{\mathrm{U}}$ and $|\mu - \mu_j| \leq \eta$. For this pair (i,j), we have

$$f_{\mu,G}(x) = f_{0,G}(x - \mu) \geq \inf_{|z|\leq\eta} f_{0,G}(x - \mu_j - z) \geq \inf_{|z|\leq\eta} P_i^{\mathrm{L}}(x - \mu_j - z) = P_{i,j}^{\mathrm{L}}(x),$$

and similarly $f_{\mu,G} \leq P_{i,j}^{\mathrm{U}}$. To upper bound the L_1 norm of this bracket, recall that $P_i^{\mathrm{L}} = \max\{p_i - \eta', 0\}$ for some $p_i \in \mathcal{P}_0$ and $\eta' > 0$ in the proof of Lemma B.2. Therefore,

$$\begin{aligned}\sup_{|z|\leq\eta} P_i^{\mathrm{L}}(x - \mu_j - z) - \inf_{|z|\leq\eta} P_i^{\mathrm{L}}(x - \mu_j - z) &\leq 2\eta \cdot \sup_{|z|\leq\eta} |p'(x - \mu_j - z)| \\ &\leq \frac{2\eta}{\sqrt{2\pi}\sigma_{\min}^2} \exp(-\frac{(|x - \mu_j| - \eta)_+^2}{2\sigma_{\max}^2}),\end{aligned}$$

where the last step follows from simple algebra, with $x_+ = \max\{x, 0\}$. Consequently,

$$\begin{aligned}\Big\| \sup_{|z|\leq\eta} P_i^{\mathrm{L}}(\cdot - \mu_j - z) - \inf_{|z|\leq\eta} P_i^{\mathrm{L}}(\cdot - \mu_j - z)\Big\|_1 &\leq \frac{2\eta}{\sqrt{2\pi}\sigma_{\min}^2} \int_{\mathbb{R}} \exp(-\frac{(|x - \mu_j| - \eta)_+^2}{2\sigma_{\max}^2})\mathrm{d}x \\ &\leq \frac{C\eta}{\sigma_{\min}^2}\Big(\eta + \sigma_{\max}\Big). \end{aligned} \tag{11}$$

On the other hand, since both functions P_i^{U} and P_i^{L} constructed in the proof of Lemma B.2 are even and non-increasing on $[0, \infty)$, we have

$$\sup_{|z|\leq\eta} P_i^{\mathrm{U}}(x - \mu_j - z) = P_i^{\mathrm{U}}((|x - \mu_j| - \eta)_+),$$

and similarly for P_i^{L}. Therefore, integrating separately over two regimes $|x - \mu_j| \leq \eta$ and $|x - \mu_j| > \eta$ yields

$$\begin{aligned}&\Big\| \sup_{|z|\leq\eta} P_i^{\mathrm{U}}(\cdot - \mu_j - z) - \sup_{|z|\leq\eta} P_i^{\mathrm{L}}(\cdot - \mu_j - z)\Big\|_1 \\ &= 2\eta|P_i^{\mathrm{U}}(0) - P_i^{\mathrm{L}}(0)| + \|P_i^{\mathrm{U}} - P_i^{\mathrm{L}}\|_1 \leq 4\eta\eta' + \frac{\varepsilon^2}{4} \leq \frac{\varepsilon^2}{2}, \end{aligned} \tag{12}$$

where the first inequality follows from $\|P_i^{\mathrm{U}} - P_i^{\mathrm{L}}\|_\infty \le 2\eta'$ and $\|P_i^{\mathrm{U}} - P_i^{\mathrm{L}}\|_1 \le \frac{\varepsilon^2}{4}$, and the second inequality follows from the choice of $\eta' \le \frac{\varepsilon^2}{4}$ in the proof of Lemma B.2 and $\eta \le \frac{1}{4}$. Now by the triangle inequality and (11), (12), we obtain $\|P_{i,j}^{\mathrm{U}} - P_{i,j}^{\mathrm{L}}\|_1 \le \varepsilon^2$ as long as

$$\eta = c_0\Big(\varepsilon\sigma_{\min} \wedge \frac{\varepsilon^2\sigma_{\min}^2}{\sigma_{\max}}\Big)$$

for a small universal constant $c_0 > 0$. Since $H^2(\mu,\nu) \le \|\mu - \nu\|_1$, this implies that $\{[P_{i,j}^{\mathrm{L}}, P_{i,j}^{\mathrm{U}}] : i \in [N], j \in [M]\}$ is an ε-Hellinger bracket with size at most

$$\exp\left(O\Big(\log^6(\frac{\sigma_{\max}}{\varepsilon\sigma_{\min}})\log\log(\frac{\sigma_{\max}}{\varepsilon\sigma_{\min}}) + \log\frac{\sigma_{\max}}{\eta}\Big)\right) = \exp\left(O\Big(\log^6(\frac{\sigma_{\max}}{\varepsilon\sigma_{\min}})\log\log(\frac{\sigma_{\max}}{\varepsilon\sigma_{\min}})\Big)\right).$$

This completes the proof. ∎

B.3. Proof of Lemma 3.2

From Lemma 3.3, it suffices to find functions $a_1(t), \dots, a_L(t)$ and $g_1(x), \dots, g_L(x)$ such that $|a_k(t)| + t|a_k'(t)| \le A$ for all $t \in [t_{\min}, t_{\max}]$, $|g_k(x)| \le G$ for all $x \in [x_{\min}, x_{\max}]$, and

$$\sup_{x\in[x_{\min},x_{\max}],t\in[t_{\min},t_{\max}]} \Big|te^{-\frac{t^2x^2}{2}} - \sum_{k=1}^{L} a_k(t)g_k(x)\Big| \le \varepsilon. \tag{13}$$

Recall from Lemma 3.4, for the function $h(v) = e^{-Ke^{\lambda v}}$ on $v \in [-1,1]$, we have the approximation

$$\big|h(v) - P_L(v)\big| \le \varepsilon' \tag{14}$$

for all $v \in [-1,1]$ where $L = O(\log(1/\varepsilon') + \lambda\log(\lambda/\varepsilon'))$ and P_L is a polynomial of degree L in v. Recall that the degree-L Chebyshev approximation of h corresponds to the polynomial $P_L(v) = \sum_{j=0}^{L} c_j T_j(v)$, where $T_j(x)$ is the degree-j Chebyshev polynomial with $T_j(\cos(\theta)) = \cos(j\theta)$, and c_j is the Chebyshev coefficient of h defined as

$$c_0 = \frac{1}{\pi}\int_0^\pi h(\cos(\theta))\mathrm{d}\theta, \quad c_j = \frac{2}{\pi}\int_0^\pi h(\cos(\theta))\cos(j\theta)\mathrm{d}\theta, \quad j \ge 1. \tag{15}$$

Since T_j are polynomials in v, each of degree at most L, we can write $P_L(v) = \sum_{j=0}^{L} p_j v^j$. Now set $u := \log t + \log x$, so that $u \in [u_{\min}, u_{\max}]$ where $u_{\min} = \log(t_{\min}x_{\min})$ and $u_{\max} = \log(t_{\max}x_{\max})$. Define the affine rescaling

$$v(u) := \frac{2u - (u_{\min} + u_{\max})}{u_{\max} - u_{\min}} \in [-1,1].$$

Then

$$e^{-\frac{t^2x^2}{2}} = \exp\Big(-\frac{1}{2}e^{2u}\Big) = \exp\Big(-Ke^{\lambda v(u)}\Big),$$

where $K = \frac{1}{2}e^{u_{\min}+u_{\max}} = \frac{1}{2}t_{\min}t_{\max}x_{\min}x_{\max}$ and $\lambda = u_{\max} - u_{\min} = \log\frac{t_{\max}x_{\max}}{t_{\min}x_{\min}}$. In particular,

$$L = O\Big(\log\frac{1}{\varepsilon'} + \lambda\log\frac{\lambda}{\varepsilon'}\Big) = O\Big(\log^2\frac{t_{\max}x_{\max}}{t_{\min}x_{\min}\varepsilon}\Big).$$

Therefore, applying the Chebyshev approximation of $h(v) = e^{-Ke^{\lambda v}}$ and substituting $v = v(u)$ yields

$$\left|e^{-\frac{t^2x^2}{2}} - \sum_{i=0}^{L} p_i\, v(u)^i\right| \le \varepsilon' \tag{16}$$

for all $t \in [t_{\min}, t_{\max}]$ and $x \in [x_{\min}, x_{\max}]$. Since $v(u)$ is affine in u, we can expand $\sum_{i=0}^{L} p_i v(u)^i$ into a degree-L polynomial in u, i.e.,

$$\sum_{i=0}^{L} p_i v(u)^i = \sum_{i=0}^{L} \widetilde{p}_i u^i,$$

and then expanding the term $(\log t + \log x)^i$ using the binomial theorem, we have

$$\left|te^{-\frac{t^2x^2}{2}} - \sum_{i=0}^{L}\sum_{j=0}^{i} \widetilde{p}_i\binom{i}{j}(\log t)^j(\log x)^{i-j}t\right| \le t_{\max}\varepsilon'. \tag{17}$$

We are left to bound the size of the coefficients $\widetilde{p}_i\binom{i}{j}$. Since $v(u)$ is an affine change of variables, with coefficients $2/\lambda$ and $\log(2K)/\lambda$, we have using the binomial theorem $|\widetilde{p}_i| \le \max_i|p_i|\, L \cdot 2^{2L}((\log(2K)/\lambda)^L \wedge 1)$. To bound p_i, first note that since $h(u) = e^{-Ke^{\lambda u}}$ is a bounded function on $[-1, 1]$, by (15), all Chebyshev coefficients have magnitude at most 2. Further, we have that the coefficients of the Chebyshev polynomial $T_k(u)$ are bounded by 2^{k-1} in absolute value (see e.g., (Trefethen, 2019)). We also trivially have $\binom{i}{j} \le 2^i \le 2^L$. Finally, for the functions in the approximation, we have $\left|t(\log t)^j\right| \le \max\{\log^L(1/t_{\min}), t_{\max}\log^L(t_{\max})\}$ and $|\log x|^{i-j} \le \max\{\log^L(x_{\max}), \log^L(1/x_{\min})\}$ for all $t \in [t_{\min}, t_{\max}]$ and $x \in [x_{\min}, x_{\max}]$. Combining these three bounds, we have that an approximation of the form in (13) holds with bound on the size of the terms given by $A \cdot G \le L2^{4L}t_{\max}(\log(t_{\max}/t_{\min}))^{2L}(\log(x_{\max}/x_{\min}))^{2L}$. Further, there are at most L^2 such terms in the approximation. Setting $\varepsilon' = \varepsilon/t_{\max}$, and plugging into Lemma 3.3, we conclude that the L_∞ covering number of $\mathcal{F}$ is bounded by

$$\begin{aligned}
&\log N(\varepsilon, \mathcal{F}, L_\infty([x_{\min}, x_{\max}])) \\
&\le O\Big(L^2\log\frac{AGL^2\log(t_{\max}/t_{\min})}{\varepsilon'}\Big) \\
&= O\Big(L^2\log\Big(\frac{2^{4L}t_{\max}^2(\log(t_{\max}/t_{\min}))^{2L+1}(\log(x_{\max}/x_{\min}))^{2L}L^3}{\varepsilon}\Big)\Big) \\
&= O\Big(\log^6\Big(\frac{x_{\max}t_{\max}}{x_{\min}t_{\min}\varepsilon}\Big)\cdot\log\log\Big(\frac{x_{\max}t_{\max}}{x_{\min}t_{\min}\varepsilon}\Big)\Big)
\end{aligned}$$

as required.

B.4. Proof of Lemma 3.3

By Carathéodory's theorem, for any $H \in \mathcal{P}([t_{\min}, t_{\max}])$, there exists a distribution $H' \in \mathcal{P}([t_{\min}, t_{\max}])$ with at most $O(L)$ atoms such that $\mathbb{E}_{t\sim H}[a_k(t)] = \mathbb{E}_{t\sim H'}[a_k(t)]$ for all $k \in [L]$. By discretizing the support of H' into a geometric grid $\{t_{\min}, e^{\delta}t_{\min}, \dots, t_{\max}\}$, and the weights of H' into a δ-net on the simplex under $\|\cdot\|_1$, with $\delta = \frac{\varepsilon}{2AGL}$, this forms a finite set $\mathcal{H}$ of size $\exp(O(L \log \frac{AGL \log(t_{\max}/t_{\min})}{\varepsilon}))$. In addition, if $H' = \sum_{i=1}^m w_i \delta_{t_i}$ and $H'' = \sum_{i=1}^m w_i' \delta_{t_i'}$ with $\max_{i\in[m]} |\log t_i - \log t_i'| \le \delta$ and $\|w - w'\|_1 \le \delta$, we have

$$\begin{aligned}|\mathbb{E}_{t\sim H'}[a_k(t)] - \mathbb{E}_{t\sim H''}[a_k(t)]| &\le \Big|\sum_{i=1}^m w_i(a_k(t_i) - a_k(t_i'))\Big| + \Big|\sum_{i=1}^m (w_i - w_i')a_k(t_i')\Big| \\ &\le \sum_{i=1}^m w_i A|\log t_i - \log t_i'| + \|w - w'\|_1 \cdot A \le 2\delta A.\end{aligned}$$

Since $\delta = \frac{\varepsilon}{2AGL}$ and $|g_k(x)| \le G$, we conclude that any $H' \in \mathcal{P}([t_{\min}, t_{\max}])$ with at most $O(L)$ atoms can be approximated by some $H'' \in \mathcal{H}$ such that

$$\sup_{x\in[x_{\min},x_{\max}], t\in[t_{\min},t_{\max}]} \Big|\sum_{k=1}^L \mathbb{E}_{t\sim H'}[a_k(t)]g_k(x) - \sum_{k=1}^L \mathbb{E}_{t\sim H''}[a_k(t)]g_k(x)\Big| \le \varepsilon.$$

By triangle inequality, this finite set $\mathcal{H}$ induces a 2ε-covering of $\mathcal{F}$.

Appendix C. Bounding Modulus of Continuity

In this section, we establish upper bounds of the Hellinger modulus of continuity in (4), with various choices of G_n. Specifically, we prove the upper bound in Lemma 1.1 for the subset-of-signals problem, and upper bounds in Corollary 1.1 for explicit examples of G_n.

C.1. Tightness of Lemma 1.1

Fix any $t \in (0, 1)$ and $p \in (0, \frac{1}{2})$. To verify the tightness of Lemma 1.1, we show that there exist universal constants $c_1, c_2 > 0$ such that if

$$G = p\delta_1 + (1-p)\delta_\sigma, \quad \mu = c_1 \begin{cases} (\frac{t^3}{p^4})^{1/6} & \text{if } t \le p^{4/3} \\ (\frac{t^3}{p^4})^{1/2} & \text{if } p^{4/3} \le t \le p \end{cases}, \quad \sigma = \frac{c_2\mu}{\sqrt{t}},$$

then $H^2(f_{\mu,G}, f_{-\mu,G}) \le t$. Indeed, by simple algebra, we have

$$\begin{aligned}
H^2(f_{\mu,G}, f_{-\mu,G}) &\le \int_{-\infty}^{\infty} \frac{(f_{\mu,G} - f_{-\mu,G})^2}{f_{\mu,G} + f_{-\mu,G}} \mathrm{d}x \\
&\le \frac{1}{1-p} \int_{-\infty}^{\infty} \frac{(f_{\mu,G}(x) - f_{-\mu,G}(x))^2}{\varphi_\sigma(x-\mu) + \varphi_\sigma(x+\mu)} \mathrm{d}x \\
&\le \frac{e^{\mu^2/\sigma^2}}{2(1-p)} \int_{-\infty}^{\infty} \frac{(f_{\mu,G}(x) - f_{-\mu,G}(x))^2}{\varphi_\sigma(x)} \mathrm{d}x \\
&\le \frac{e^{\mu^2/\sigma^2} p^2}{1-p} \int_{-\infty}^{\infty} \frac{(\varphi(x-\mu) - \varphi(x+\mu))^2}{\varphi_\sigma(x)} \mathrm{d}x \\
&\quad + e^{\mu^2/\sigma^2}(1-p) \int_{-\infty}^{\infty} \frac{(\varphi_\sigma(x-\mu) - \varphi_\sigma(x+\mu))^2}{\varphi_\sigma(x)} \mathrm{d}x.
\end{aligned}$$

Here we use φ and φ_σ to denote the densities of $\mathcal{N}(0,1)$ and $\mathcal{N}(0,\sigma^2)$, respectively. To proceed, since $\sigma = c_2\mu/\sqrt{t} \ge c_2\mu$, we have $\mu^2/\sigma^2 \le 1$ whenever $c_2 \ge 1$. For the remaining integrals, as long as $\tau^2 < 2\sigma^2$, simple algebra gives

$$\int_{-\infty}^{\infty} \frac{(\varphi_\tau(x-\mu) - \varphi_\tau(x+\mu))^2}{\varphi_\sigma(x)} \mathrm{d}x = \frac{2\sigma^2}{\tau\sqrt{2\sigma^2 - \tau^2}} \Big[\exp\Big(\frac{\mu^2}{2\sigma^2 - \tau^2}\Big) - \exp\Big(-\frac{\mu^2}{\tau^2}\Big)\Big].$$

Plugging in $\tau \in \{1, \sigma\}$, and using that $\sigma \ge 1$ in both regimes, we get

$$H^2(f_{\mu,G}, f_{-\mu,G}) = O\Big(p^2\sigma(\mu^2 \wedge 1) + \frac{\mu^2}{\sigma^2}\Big) = O\Big(\frac{c_2 p^2}{\sqrt{t}}(\mu^3 \wedge \mu) + \frac{t}{c_2^2}\Big).$$

Choosing $c_2 > 0$ large enough and $c_1 > 0$ small enough, we have $H^2(f_{\mu,G}, f_{-\mu,G}) \le t$, as desired.

Finally, if $t \ge 3p$, by convexity of squared Hellinger distance, for every $\mu > 0$ it holds that

$$H^2(f_{-\mu,G}, f_{\mu,G}) \le 2p + H^2(\mathcal{N}(-\mu, \tfrac{c_2^2\mu^2}{t}), \mathcal{N}(\mu, \tfrac{c_2^2\mu^2}{t})) \le 2p + \frac{t}{3} \le t,$$

as long as $c_2 > 0$ is large enough. Therefore, $\omega_{H^2,G}(t) \ge 2\mu$ in this case, and letting $\mu \to \infty$ shows that $\omega_{H^2,G}(t)$ does not admit a uniform upper bound.

C.2. Modulus of Continuity for SoS Priors: Proof of Lemma 1.1

The proof of Lemma 1.1 relies on the following variational lower bound of squared Hellinger distance:

$$H^2(P,Q) = \int \frac{(P-Q)^2}{(\sqrt{P} + \sqrt{Q})^2} \ge \frac{1}{2} \int \frac{(P-Q)^2}{P+Q} = \frac{1}{2} \sup_T \frac{(\mathbb{E}_P[T] - \mathbb{E}_Q[T])^2}{\mathbb{E}_P[T^2] + \mathbb{E}_Q[T^2]}, \tag{18}$$

where the supremum is over all test functions $T : \mathcal{X} \to \mathbb{R}$, and the last step is Cauchy–Schwarz. In the definition of Hellinger modulus of continuity, suppose the choices $P = f_{\mu,G}$ and $Q = f_{0,G}$ satisfy $H^2(P,Q) \le t$. Now choosing test function $T_\Delta(x) = 1_{[\mu,\mu+\Delta]}(x)$, the variational form gives

$$\frac{\Big[\mathbb{E}_G\Big(\Phi\Big(\frac{\Delta}{\sigma}\Big) - \Phi(0) - \Phi(\frac{\Delta+\mu}{\sigma}) + \Phi(\frac{\mu}{\sigma})\Big)\Big]^2}{\mathbb{E}_G\Big(\Phi(\frac{\Delta}{\sigma}) - \Phi(0) + \Phi(\frac{\Delta+\mu}{\sigma}) - \Phi(\frac{\mu}{\sigma})\Big)} \le 2t. \tag{19}$$

Here $\Phi(x) = \int_{-\infty}^{x} \varphi(t)dt$ denote the standard Gaussian CDF. Taking $\Delta = \infty$, and noting that the denominator always lies in $[0, 1]$, we obtain

$$\sqrt{2t} \geq \mathbb{E}_G\Big(\Phi(\frac{\mu}{\sigma}) - \Phi(0)\Big) \geq \Phi'(1) \cdot \mathbb{E}_G\Big[\frac{\mu \mathbf{1}_{\{\sigma \geq \mu\}}}{\sigma}\Big]$$

due to $\Phi(x) - \Phi(0) \geq \Phi'(1)x$ for $x \in [0, 1]$. This shows that for a universal $c_1 > 0$,

$$\mathbb{E}_G\Big[\frac{\mathbf{1}_{\{\sigma \geq \mu\}}}{\sigma}\Big] \leq \frac{c_1\sqrt{t}}{\mu}. \tag{20}$$

Next we take $\Delta = 1$ in (19). For the denominator, we have

$$\begin{aligned}\mathbb{E}_G\Big(\Phi(\frac{1}{\sigma}) - \Phi(0) + \Phi(\frac{1+\mu}{\sigma}) - \Phi(\frac{\mu}{\sigma})\Big) &\leq 2\mathbb{E}_G\Big(\Phi(\frac{1}{\sigma}) - \Phi(0)\Big) \\ &\leq 2\left(\frac{1}{2}G([0,1]) + \Phi'(0) \cdot \mathbb{E}_G\left[\frac{\mathbf{1}_{\{\sigma \geq 1\}}}{\sigma}\right]\right) \\ &\leq G([0,1]) + \mathbb{E}_G\Big[\frac{\mathbf{1}_{\{\sigma \geq 1\}}}{\sigma}\Big]. \end{aligned} \tag{21}$$

For the numerator, we distinguish into two cases.

Case I: $\mu \geq 1$**.** Let

$$\Xi_{\mu,\sigma} := \Phi(\frac{1}{\sigma}) - \Phi(0) - \Phi(\frac{1+\mu}{\sigma}) + \Phi(\frac{\mu}{\sigma}).$$

When $\mu \geq 1$, simple algebra yields that for a universal constant $c_2 > 0$,

$$\Xi_{\mu,\sigma} \geq c_2 \begin{cases} 1 & \text{if } \sigma < 1, \\ \frac{1}{\sigma} & \text{if } 1 \leq \sigma < \mu, \\ 0 & \text{if } \sigma \geq \mu. \end{cases} \tag{22}$$

Therefore,

$$\mathbb{E}_G[\Xi_{\mu,\sigma}] \geq c_2\Big(G([0,1]) + \mathbb{E}_G\Big[\frac{\mathbf{1}_{\{1 \leq \sigma \leq \mu\}}}{\sigma}\Big]\Big), \tag{23}$$

and

$$\begin{aligned}\Big(G([0,1]) + \mathbb{E}_G\Big[\frac{\mathbf{1}_{\{1 \leq \sigma \leq \mu\}}}{\sigma}\Big]\Big)^2 &\overset{\text{(a)}}{\leq} c_3 t\Big(G([0,1]) + \mathbb{E}_G\Big[\frac{\mathbf{1}_{\{\sigma \geq 1\}}}{\sigma}\Big]\Big) \\ &\overset{\text{(b)}}{\leq} c_3 t\Big(G([0,1]) + \mathbb{E}_G\Big[\frac{\mathbf{1}_{\{1 \leq \sigma \leq \mu\}}}{\sigma}\Big] + \frac{c_1\sqrt{t}}{\mu}\Big).\end{aligned}$$

Here (a) follows from (19) and (21), and (b) uses (20). Solving this quadratic inequality gives

$$G([0,1]) + \mathbb{E}_G\Big[\frac{\mathbf{1}_{\{1 \leq \sigma \leq \mu\}}}{\sigma}\Big] \leq c_4\Big(t + \frac{t^{3/4}}{\mu^{1/2}}\Big).$$

As $G([0,1]) \geq p \geq Ct$ for $C \geq 2c_4$, this inequality gives that $\mu = O(\frac{t^{3/2}}{p^2})$.

Case II: $0 \le \mu < 1$**.** In this case, one can verify the following lower bounds for $\Xi_{\mu,\sigma}$:

$$\Xi_{\mu,\sigma} \ge c_2 \begin{cases} 1 & \text{if } \sigma < \mu, \\ \frac{\mu}{\sigma} & \text{if } \mu \le \sigma < 1, \\ 0 & \text{if } \sigma \ge 1. \end{cases} \tag{24}$$

Therefore, the numerator is lower bounded as

$$\mathbb{E}_G[\Xi_{\mu,\sigma}] \ge c_2 \Big(G([0,\mu]) + \mathbb{E}_G\Big[\frac{\mu \mathbf{1}_{\{\mu \le \sigma \le 1\}}}{\sigma}\Big]\Big) \ge c_2 \mu \cdot G([0,1]). \tag{25}$$

Consequently, by (19), (20), and (21), we get

$$\mu^2 G([0,1])^2 \le c_3 t \Big(G([0,1]) + \mathbb{E}_G\Big[\frac{\mathbf{1}_{\{\sigma \ge 1\}}}{\sigma}\Big]\Big) \le c_3 t \Big(G([0,1]) + \frac{c_1\sqrt{t}}{\mu}\Big),$$

and further solve this quadratic inequality to find

$$G([0,1]) \le c_4 \Big(\frac{t}{\mu^2} + \frac{t^{3/4}}{\mu^{3/2}} \Big).$$

Finally, as $G([0,1]) \ge p$, we obtain $\mu = O(\frac{t^{1/2}}{p^{1/2}} + \frac{t^{1/2}}{p^{2/3}}) = O(\frac{t^{1/2}}{p^{2/3}})$ as $p \le 1$.

To reach the desired conclusion, note that when $p^{4/3} < t \le \frac{p}{C}$, we have either $\mu = O(\frac{t^{3/2}}{p^2})$ (Case I), or $\mu \le 1$ (assumption of Case II). Since the upper bound in Case I is larger, we conclude that $\mu = O(\frac{t^{3/2}}{p^2})$. Similarly, when $t \le p^{4/3}$, we have either $\mu = O(\frac{t^{3/2}}{p^2})$ (Case I), or $\mu = O(\frac{t^{1/2}}{p^{2/3}})$ (Case II). Here the upper bound in Case II dominates, so that $\mu = O(\frac{t^{1/2}}{p^{2/3}})$.

C.3. Modulus of Continuity for Selected Priors: Proof of Corollary 1.1

To prove Corollary 1.1, we invoke Theorem 1.1 and establish upper bounds of the Hellinger modulus of continuity for various choices of G_n.

Equal variance. When $\sigma_i \equiv 1$, we have $G_n = \delta_1$. In this case,

$$H^2(f_{\mu_1,G_n}, f_{\mu_2,G_n}) = 2 - 2\exp\left(-\frac{(\mu_1-\mu_2)^2}{8}\right),$$

so that $\omega_{H^2,G_n}(t) = O(\sqrt{t})$. Now plugging in $t = O(\frac{L}{n})$ gives the result.

Quadratic variance. Consider any $\mu > 0$ such that (19) holds for all $\Delta \ge 0$, with $G = G_n$. Next we choose $\Delta = \mu$ in (19) and proceed in the same way as (21) to obtain

$$G_n([0,\mu])^2 \le c_1 t \left(G_n([0,\mu]) + \mu \mathbb{E}_{G_n}\left[\frac{\mathbf{1}_{\{\sigma \ge \mu\}}}{\sigma}\right]\right).$$

For $t \le c_2$ with a small enough constant $c_2 > 0$, the above inequality shows that $\mu > n$ is impossible. For $1 \le \mu \le n$, since $\mathbb{E}_{G_n}[\frac{\mathbf{1}_{\{\sigma \ge \mu\}}}{\sigma}] \le \mathbb{E}_{G_n}[\frac{1}{\sigma}] = \frac{1}{n}\sum_{i=1}^n \frac{1}{i} = O(\frac{\log n}{n})$ and $G_n([0,\mu]) \asymp \frac{\mu}{n}$, we can solve that $\mu = O(nt\log n)$. Combining with the remaining case $\mu \le 1$ leads to the final upper bound

$$\omega_{H^2,G_n}(t) = O\left(nt\log n \vee 1\right), \quad \text{if } t \le c_2.$$

Plugging in $t = O(\frac{L}{n})$ proves the target result.

Two variances. Again, consider some $\mu > 0$ such that (19) holds for every $\Delta \geq 0$, with $G = G_n = \frac{m}{n}\delta_1 + (1-\frac{m}{n})\delta_\sigma$. We consider two choices of Δ. First, choosing $\Delta = \sigma$ and upper bounding the denominator by 2 in (19), we get

$$\frac{n-m}{n}\left(\Phi(1) - \Phi(0) - \Phi(1+\frac{\mu}{\sigma}) + \Phi(\frac{\mu}{\sigma})\right) \leq c_1\sqrt{t}.$$

Clearly, for $t \leq c_2$ with a small enough constant $c_2 > 0$, this inequality implies that $\mu \leq \sigma$. In this case, the left-hand side further scales as $\Omega(\frac{\mu}{\sigma})$, so that we get

$$\omega_{H^2,G_n}(t) = O(\sigma\sqrt{t}), \quad \text{if } t \leq c_2.$$

For $t = O(\frac{L}{n})$, this gives the general $O(\sigma\sqrt{\frac{L}{n}})$ upper bound.

The second choice of Δ is $\Delta = 1$. If we upper bound the denominator of (19) by 2, then

$$\frac{m}{n}\left(\Phi(1) - \Phi(0) - \Phi(1+\mu) + \Phi(\mu)\right) \leq c_1\sqrt{t}.$$

Therefore, if $\frac{n\sqrt{t}}{m} \leq c_2$ for a small enough constant $c_2 > 0$, we solve that $\mu = O(\frac{n\sqrt{t}}{m})$. For $t = O(\frac{L}{n})$, this proves the $O(\frac{\sqrt{nL}}{m})$ upper bound for $m \geq \sqrt{nL}$. Instead, if we upper bound the denominator of (19) via (21), we obtain

$$\frac{m}{n}\left(\Phi(1) - \Phi(0) - \Phi(1+\mu) + \Phi(\mu)\right) \leq c_1\sqrt{t\left(\frac{m}{n} + \frac{1}{\sigma}\right)}.$$

For $t = O(\frac{L}{n})$ and $m \geq C(\sqrt{\frac{nL}{\sigma}} + L)$, it holds that

$$\frac{m}{n} \geq 10c_1\sqrt{t\left(\frac{m}{n} + \frac{1}{\sigma}\right)}.$$

Solving the inequality gives the target upper bound $\mu = O(1)$.

α-mixture distributions. We use the upper bound in the two variances example. For $\alpha \in (0,1)$, we use the upper bound $O(\sigma\sqrt{\frac{L}{n}}) = \widetilde{O}(n^{\alpha-1/2}\sqrt{L})$. For $\alpha \geq 1$, we note that for $m = cL$ with a large enough constant $c > 0$, it holds that

$$C\left(\sqrt{\frac{nL}{n^\alpha}} + L\right) \leq m < \sqrt{nL}.$$

Therefore, we can apply the $O(1)$ upper bound in this case.

Appendix D. Additional Details of Section 4

D.1. Proof of Lemma 4.1

Since $X_i \neq 0$ for all $i \in [n]$, the function $G \mapsto f_{0,G}(X_i)$ has a finite upper bound depending only on X_i, so the log-likelihood $\sum_{i=1}^n \log f_{0,G}(X_i)$ cannot reach $+\infty$. Then standard compactness

argument shows the existence of the NPMLE $\widehat{G}$, and by the strict concavity of $x \mapsto \log x$, the vector of densities $(f_{0,\widehat{G}}(X_1), \ldots, f_{0,\widehat{G}}(X_n))$ is unique. Denote it by $(f_1, \ldots, f_n)$.

By the KKT condition (6), $\operatorname{supp}(\widehat{G})$ is a subset of the set of global maximizers of $\sigma \mapsto D_{\widehat{G}}(\sigma)$. By differentiation, the set of critical points of $\sigma \mapsto D_{\widehat{G}}(\sigma)$ is

$$C := \Big\{\sigma > 0 : \frac{1}{n}\sum_{i=1}^{n} \frac{f_i}{\sqrt{2\pi}} e^{-X_i^2/2\sigma^2}\Big(\frac{X_i^2}{\sigma^2} - 1\Big) = 0\Big\}. \tag{26}$$

Clearly, $C \subseteq [\min_i |X_i|, \max_i |X_i|]$, so the same holds for $\operatorname{supp}(\widehat{G})$. By combining repeated appearances of $|X_i|$, WLOG we assume that $|X_1|, \ldots, |X_n|$ are distinct. Next we show that C is a finite set, with $|C| \le 2n - 1$. Therefore, for some distinct constants $c_1, c_2, \ldots, c_n$ and a non-zero vector $(d_1, \ldots, d_{2n})$,

$$\Big\{\frac{1}{\sigma^2} : \sigma \in C\Big\} \subseteq \Big\{x > 0 : \sum_{i=1}^{n} (d_i + d_{n+i}x)e^{c_i x} = 0\Big\}.$$

To proceed, we recall the following result taken from (Pólya and Szegö, 1925, Page 48).

Lemma D.1 *Let $c_1, \ldots, c_n \in \mathbb{R}$ be distinct, and $P_i(x)$ be a polynomial in x with degree $m_i - 1$. Then the equation $\sum_{i=1}^{n} P_i(x)e^{c_i x} = 0$ has at most $(\sum_{i=1}^{n} m_i) - 1$ real solutions.*

By Lemma D.1, the above equation has at most $2n-1$ solutions. This shows that $|C| \le 2n-1$. Finally, since the continuous map $\sigma \mapsto D_{\widehat{G}}(\sigma)$ must have at least one critical point between two global maximizers, we conclude from the upper bound on $|C|$ that $|\operatorname{supp}(\widehat{G})| \le n$.

Finally we show the uniqueness of $\widehat{G}$. Since the map $\sigma \mapsto D_{\widehat{G}}(\sigma)$ only depends on $\widehat{G}$ through the likelihood vector $(f_1, \ldots, f_n)$, the set D of its global maximizers is a fixed set for all versions of $\widehat{G}$. In addition, $|D| \le n$, so we can write $D = \{\sigma_1, \ldots, \sigma_m\}$ with $m \le n$, and any NPMLE $\widehat{G}$ takes the form $\sum_{j=1}^{m} w_j \delta_{\sigma_j}$. It remains to show the uniqueness of $(w_1, \ldots, w_m)$. Now by the uniqueness of $(f_1, \ldots, f_n)$, $(w_1, \ldots, w_m)$ is a solution to the linear system

$$\sum_{j=1}^{m} w_j \frac{1}{\sigma_j}\varphi\Big(\frac{X_i}{\sigma_j}\Big) = f_{0,\widehat{G}}(X_i) = f_i, \quad \forall i \in [n].$$

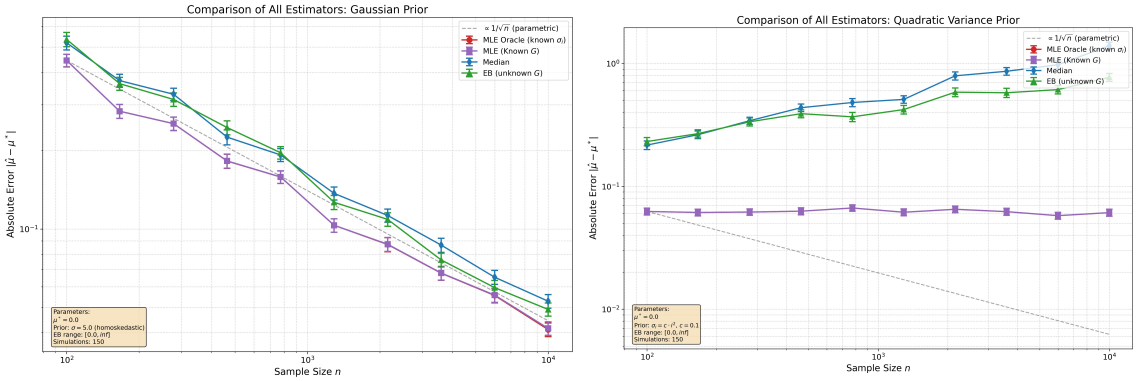
To prove uniqueness, it suffices to show that the matrix $A = (A_{ij})_{i\in[n], j\in[m]}$ with $A_{ij} = \frac{1}{\sigma_j}\varphi(\frac{X_i}{\sigma_j})$ has full column rank. By scaling the columns and taking $m \times m$ submatrices, it further suffices to prove that $B = (B_{ij}) \in \mathbb{R}^{m\times m}$ with $B_{ij} = \varphi(\frac{X_i}{\sigma_j})$ has full rank. Assuming the contrary, then $Bz = 0$ would have a non-zero solution $z \in \mathbb{R}^m$, and the map

$$t \mapsto \sum_{j=1}^{m} \exp\Big(-\frac{t}{\sigma_j^2}\Big) z_j$$

would have at least m distinct zeros. Since $\sigma_1, \ldots, \sigma_m > 0$ are distinct, this is a contradiction to the statement of Lemma D.1. This concludes the uniqueness of $\widehat{G}$.

D.2. Additional Experimental Details in Section 4

We provide additional experimental details related to our implementation, as well as figures for the equal and quadratic variance cases of heteroskedastic mean estimation. For all experiments, we set $\sigma_{\min} = 0$ and $\sigma_{\max} = \infty$, but in general one can initialize with prespecified values for $(\sigma_{\min}, \sigma_{\max})$. For every instance of the Frank–Wolfe algorithm (i.e., estimating $\widehat{G}$ with a fixed μ), we use Lemma 4.1 to set a data-driven support $I = [\min_i |X_i - \mu|, \max_i |X_i - \mu|]$ for the new atoms σ_t, and apply a grid search over $N = 5,000$ points in I to find the location of the new atom σ_t. In addition, we initialize the atoms of $\widehat{G}_0$ to be 5 evenly distributed points in I. When finding the MLE estimator $\widehat{\mu}$ given the current $\widehat{G}$, we also apply a grid search over $N = 5,000$ points in the interval $[\min_{i\in[n]} X_i, \max_{i\in[n]} X_i]$. Therefore, aside from typical tolerance thresholds, our algorithm is completely tuning-parameter free. Code for our implementation and simulation results can be found in the repository https://github.com/AshettyV/NPMLE, and we also include the experimental results for the cases of equal and quadratic variance in Corollary 1.1.



(a) $P_1 = \cdots = P_n = \mathcal{N}(0, 5)$ (b) $P_i = \mathcal{N}(0, \frac{i^2}{10}), \quad 1 \le i \le n$

Figure 3: Average absolute estimation error for $\widehat{\mu}^{\text{EB}}$, sample median, and two oracle MLEs over $N = 150$ simulations, under the equal variance model where $P_1 = \cdots = P_n = \mathcal{N}(0, 5)$ and the quadratic variance model $P_i = \mathcal{N}(0, \frac{i^2}{10})$ for $i \in [n]$.