

Empirical Bayes Selection for Value Maximization

Dominic Coey*

Kenneth Hung*

coey@meta.com

kenhung@meta.com

Central Applied Science

Meta

Menlo Park, California, USA

Abstract

We study the problem of selecting the best m units from a set of n as $m/n \rightarrow \alpha \in (0, 1)$, where noisy, heteroskedastic measurements of the units' true values are available and the decision-maker wishes to maximize the aggregate true value of the units selected. Given a parametric prior distribution, the empirical Bayes decision rule incurs $O_p(n^{-1})$ regret relative to the Bayesian oracle that knows the true prior. More generally, if the error in the estimated prior is of order $O_p(r_n)$, regret is $O_p(r_n^2)$. In this sense *selection* of the best units is fundamentally easier than *estimation* of their values. We show this regret bound is sharp in the parametric case, by giving an example in which it is attained. Using priors calibrated from a dataset of over four thousand internet experiments, we confirm that empirical Bayes methods perform well in detecting the best treatments with only a modest number of experiments.

CCS Concepts

• **Mathematics of computing** → *Density estimation*; • **General and reference** → *Experimentation*; • **Information systems** → *Decision support systems*.

Keywords

Empirical Bayes, compound decisions, ranking and selection

ACM Reference Format:

Dominic Coey and Kenneth Hung. 2025. Empirical Bayes Selection for Value Maximization. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25)*, August 3–7, 2025, Toronto, ON, Canada. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3711896.3736928>

KDD Availability Link:

The source code of the simulations that generate the plots has been made publicly available at <https://doi.org/10.5281/zenodo.15538137>.

1 Introduction

In many important scientific and economic applications, decision-makers are presented with data on the performance of n units, from which they must select a strict subset for further investigation or

treatment. Examples include identifying the best teachers, hospitals, or athletes (Brown [6], Chetty et al. [12], Dimick et al. [18]); genes associated with particular outcomes (Efron and Tibshirani [24]); drug candidates (Yu et al. [65]); or in the application of this paper, internet experiments. Each unit is associated with an unobserved true value, which is measured with heteroskedastic noise. The constraint that only $m < n$ units can be selected arises naturally when the decision-maker has limited resources to devote to the chosen units, and must restrict attention to the most promising candidates.

A desirable feature of a selection procedure is that the aggregate value of its selections is close to the maximum attainable value. Understanding how different selection procedures perform in this respect enables decision-makers to assess the quality of their decisions in the preceding applications. The empirical Bayes approach to this question involves estimating the unknown, prior distribution from which the true values are drawn, and selecting units with the highest estimated posterior means. We show that if the prior distribution is known to lie within some parametric class, empirical Bayes incurs regret of order $O_p(n^{-1})$ relative to the oracle Bayes decision rule in which the prior distribution is known.² This is faster than the usual $n^{-1/2}$ parametric rate of convergence from the central limit theorem. In this sense *selection* is fundamentally easier than *estimation*: picking a set of units with low regret is easier than pinning down the precise values of those units. This generalizes directly to the nonparametric case: regret converges to zero at the square of the rate that estimation error in the prior converges to zero.

The basic intuition for this result follows. First, mistakes, whether of inclusion or exclusion, are only likely to happen for those units whose true values are sufficiently close to a critical threshold. Units comfortably (or below) above that threshold will be correctly selected (or omitted) with high probability (i.e. with probability converging to 1, abbreviated as w.h.p.). Second, even those mistakes cannot be too costly, as units incorrectly included or excluded are likely to be marginal—almost good enough to be selected, or almost bad enough to be omitted. The regret, which is the product of two terms corresponding to these factors, will therefore be second-order small. We show that our $O_p(n^{-1})$ bound is sharp in the parametric case, by constructing an example in which regret is at least Cn^{-1} with non-vanishing probability for some positive constant C .

*Both authors contributed equally to this research.



This work is licensed under a Creative Commons Attribution 4.0 International License. *KDD '25, Toronto, ON, Canada*

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1454-2/2025/08

<https://doi.org/10.1145/3711896.3736928>

¹ $O_p(\cdot)$ is the stochastic O notation commonly used in statistics, as defined in [60]. We write $X_n = O_p(a_n)$ if X_n/a_n is bounded in probability.

²We refer to the oracle Bayes decision rule rather than simply the Bayes decision rule throughout, to emphasize that the prior is unknown to the decision-maker.

We illustrate this result with simulations based on internet experimentation data, where units correspond to experiments, and values to treatment effects. Heteroskedasticity arises because experiments vary in sample size. Technology companies may wish to identify a subset of best- or worst-performing experiments for further investigation, in the former case, as candidates to launch to production, or in the latter case, as candidates to stop early. When the follow-up investigation incurs some cost, it may only be feasible to select a strict subset of experiments for more analysis. In this application, as in others, the cost of mistakes depends on their magnitude—that is, on the difference between the aggregate value of the units selected and the units that should have been selected. We simulate true effects from a scale mixture of mean-zero Gaussians calibrated on this dataset, and evaluate the regret of the empirical Bayes approach for selecting the top 10% of experiments. Consistent with our theoretical results, we find that regret is $O_p(n^{-1})$. By comparison, identifying the set of the top 10% experiments with all misclassifications being equally penalized regardless of their magnitude, or estimating the treatment effects of the selected experiments, or estimating the prior distribution itself, are all structurally harder problems, each of which only exhibits convergence at the usual parametric rate.

1.1 Related Work

Our work builds on several large and active strands of the statistics and econometrics literature. Foundational work introducing and developing the empirical Bayes approach to statistics includes Efron and Morris [22], Kiefer and Wolfowitz [42], Robbins [53, 54]. Applications of the selection problem have proliferated, as the problem of discerning between units which perform well or poorly on the basis of noisy, heteroskedastic measurements describes many real-world settings of interest. Previous work has studied identifying the best teachers (Chetty et al. [11], Gilraine et al. [27], Harris and Sass [36], Jacob and Lefgren [39], Kane et al. [40]), the best medical facilities (Dimick et al. [18], Goldstein and Spiegelhalter [29], Hull [37], Thomas et al. [59]), the best baseball players (Brown [6], Efron and Morris [23]); differentially expressed genes (Efron and Tibshirani [24], Smyth [56]); promising drug candidates (Yu et al. [65]); geographic areas associated with the greatest intergenerational mobility (Bergman et al. [4]) or mortality (Marshall [47]), and employers exhibiting the most evidence of discrimination (Kline and Walters [43]). Internet experiments are particularly well-suited to empirical Bayes methods (Azevedo et al. [2, 3], Coey and Cunningham [13], Deng [16], Goldberg and Johndrow [28], Guo et al. [31]) as datasets are often large enough for accurate estimation of flexibly-specified priors, and the experiment-level sampling error is typically close to normally distributed. For these applications, the aggregate value of the selected units will often be an important component of the decision-maker’s utility function. Our results provide theoretical and empirical support for selection based on such methods.

The literature on post-selection inference, including Andrews et al. [1], Cohen and Sackrowitz [14], Dahiya [15], Fithian et al. [26], Guo and He [32], Gupta and Panchapakesan [34], Hung and Fithian [38], also studies selection problems, but differs from the present work in that its chief focus is estimating the values, differences

or ranks of the selected units, rather than analyzing the regret associated with the selection. Cohen and Sackrowitz [14], Dahiya [15] provide estimates for the value of a selection unit. Andrews et al. [1], Fithian et al. [26], Guo and He [32], Gupta and Panchapakesan [34], Hung and Fithian [38] largely aim at frequentist inferences. While the notion of regret we consider averages over draws from the distribution of units’ true values, an alternative line of inquiry beyond the scope of this paper would be to characterize admissible and minimax decision rules for the frequentist analog of the regret we define, considering the units’ values as fixed constants.

Gu and Koenker [30], Mogstad et al. [50] both study similar selection problems to the one we analyze. Gu and Koenker [30] take an empirical Bayes approach to selecting the best units while controlling the marginal false discovery rate; Mogstad et al. [50] assert frequentist control over the familywise error rate, which amounts to a zero-one loss based on the correctness of the ranks. Both consider loss functions different from ours. In their frameworks, mistakenly selecting or omitting any unit incurs a discrete cost, whereas in ours the cost of mistakenly selecting or omitting a marginal unit near the selection threshold is small. We view these as complementary perspectives. While in some decision problems mistakes may be undesirable *per se*, the aggregate performance of the selected units is typically still of interest. For teacher evaluations, for example, policy-makers may rightly be concerned with guarantees over the number of teachers who are incorrectly fired (Mogstad et al. [49]), but may also wish to understand how well their selection procedure is performing from the students’ perspective, in terms of aggregate teacher “value-added”. In other contexts, as in internet experimentation or drug discovery, the aggregate value of the selection is the primary concern, and it is harder to justify caring about the number of mistakes *per se*.

Closely related to our paper is [9], which notes the importance of empirical Bayes top- m selection to various social science applications, and derives regret bounds for the problem. Those rate results are more favorable than the ones we present in cases where we can recover the posterior mean but not the prior fast. However in other cases, e.g. the parametric case, [9] bound regret by a term converging slower than $n^{-1/2}$, while we prove n^{-1} convergence and show that rate cannot in general be improved upon. We summarize this comparison in Table 1. Furthermore, while the nonparametric rate of convergence of estimated priors to the truth is generally only logarithmic even for optimal procedures [7, 25], we may often observe a faster rate of convergence (e.g. see Figure 4) in practice, which our result translates to a tighter bound on the regret.

The bound in [9] goes through the mean squared error of the posterior means, which would be more pessimistic if the posterior mean of irrelevant items (e.g. those almost always or never selected) are hard to estimate. Meanwhile, our method relies on controlling the number of mistakes by estimating the distribution, which is likely more challenging than minimizing the mean squared error, leading to potential pessimism in a different way.

Our selection problem may remind readers of the multi-armed bandit literature that studies the problem of identifying the top m arms with a certain probability, e.g. Chen et al. [10], Shang et al. [55]. However to target the probability of correct selection is to consider discontinuous loss functions similar to ones in Gu and Koenker

Class of priors	Our bound	Bound from [9]
Parametric	$O_p(n^{-1})$	$\tilde{O}(n^{-1/2})$
Finite support of unknown cardinality [8]	$O_p(n^{-1/2})$	$\tilde{O}(n^{-1/2})$
Density function has k bounded derivative with bounded support [7]	$O_p((\log n)^{-k})$	$\tilde{O}(n^{-1/2})$

Table 1: Summary of regret convergence rates given by our bound and the bound from [9]. Note that \tilde{O} indicates some hidden log-factors, and also the difference that Chen [9] bounds the expectation of regret when we bound it stochastically.

[30], Mogstad et al. [50]. We also note that our selection problem is non-sequential which leads to new challenges, as poor choices of parameter in the prior cannot be overcome with additional samples in long run.

Finally, our work is related to the compound decision framework introduced in Robbins [52], in which a simple decision is made for each unit and the overall loss is the sum of the loss from each individual decision. Convergence and rate results are available for empirical Bayes as applied to compound decision problems, e.g. Gupta and Li [33], Hannan and Van Ryzin [35], Polyanskiy and Wu [51], Van Ryzin and Susarla [61], Zhang [66], but as Weinstein [62] observes this framework is rather restrictive and does not encompass the value maximization problem studied here of selecting the best m of n units. Weinstein [62] generalizes further to a class of simultaneous decision problems that are permutation invariant, which encompasses our problem of selecting m units. However the optimal frequentist solution requires knowledge of the empirical c.d.f. (e.c.d.f.), or equivalently the order statistics, of the true effects μ_i . Instead of studying the performance under pathological choices of μ_i as would be required in a minimax analysis, we take a more Bayesian approach to enable an analysis of regret without knowledge of the order statistics of the μ_i 's.

1.2 Our Contribution

Our main contribution relative to this existing literature is to provide the first sharp regret bounds for parametric empirical Bayes selection, and to show how these same ideas extend to control regret in the nonparametric case. Our empirical work on internet experimentation complements this by verifying that regret is quantitatively modest in practice, given a reasonable number experiments of moderate precision. Together, our theoretical and empirical results suggest optimism for empirical Bayes approaches to selection when the decision-maker is primarily concerned with maximizing the aggregate value of the selected units, as opposed to correctly classifying the top units or estimating their values.

2 The Top- m Selection Problem

2.1 Setup

There are n units, each of which is associated with a unobserved true value $\mu_i \in \mathbb{R}$ and an observed noise standard deviation $\sigma_i > 0$.³ The μ_i and σ_i are distributed independently from each other and independently across experiments.⁴ Their unknown marginal

distributions are denoted G_0 and H_0 ,

$$(\mu_i, \sigma_i) \sim G_0 \times H_0.$$

We consider nondegenerate prior distributions G belong to a potentially nonparametric family \mathcal{M} , that forms a metric space with 1-Wasserstein distance $W_1(\cdot, \cdot)$. We assume the family is not misspecified, i.e. the family includes the truth G_0 . For each unit i , the decision-maker observes a measurement $X_i \in \mathbb{R}$, which is distributed as

$$X_i | \mu_i, \sigma_i \sim \mathcal{N}(\mu_i, \sigma_i^2).$$

The decision-maker must choose m units for some $m < n$. Their average utility given the index set of choices $J \subset \{1, 2, \dots, n\}$ is $U(J) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(i \in J) \mu_i$. Let

$$f_{G, \sigma_i}(X_i) = \frac{\int \mu \frac{1}{\sigma_i} \phi\left(\frac{X_i - \mu}{\sigma_i}\right) dG}{\int \frac{1}{\sigma_i} \phi\left(\frac{X_i - \mu}{\sigma_i}\right) dG} \quad (1)$$

denote the posterior mean of μ_i given X_i, σ_i , assuming that the prior distribution of μ_i is G , where $\phi(\cdot)$ is the probability density function (p.d.f.) of a standard Gaussian. The true posterior mean is $f_{G_0, \sigma_i}(X_i)$. An estimator $\hat{G} = \hat{G}(X_1, \dots, X_n, \sigma_1, \dots, \sigma_n)$ of G_0 is available, where \hat{G} converges to G_0 at some rate r_n in 1-Wasserstein distance. It is used as the empirical Bayes prior, and in constructing posterior mean estimates, $f_{\hat{G}, \sigma_i}(X_i)$. For simplicity we denote $f_{G_0, \sigma_i}(X_i)$ and $f_{\hat{G}, \sigma_i}(X_i)$, the oracle and empirical Bayes posterior means for unit i , as θ_i and $\hat{\theta}_i$ respectively. We can view θ_i as being drawn i.i.d. from a distribution, which we denote P .

Given the observed data, an oracle Bayesian decision-maker maximizes expected utility (where the expectation is with respect to the posterior distribution over the unknown true values) by selecting the m units with the highest values of θ_i , breaking ties randomly. The empirical Bayes decision-maker mimics this rule, by selecting the m units with the highest values of $\hat{\theta}_i$, breaking ties randomly. Letting J_{EB} and J_{Bayes} be the empirical Bayes and oracle Bayes choice sets, the regret from empirical relative to oracle Bayes is

$$\mathcal{R} = \mathbb{E}[U(J_{\text{Bayes}}) | X_1, \dots, X_n, \sigma_1, \dots, \sigma_n] - \mathbb{E}[U(J_{\text{EB}}) | X_1, \dots, X_n, \sigma_1, \dots, \sigma_n] \quad (2)$$

$$\begin{aligned} &= \frac{1}{n} \sum_{i=1}^n (\mathbb{1}(i \in J_{\text{Bayes}}) - \mathbb{1}(i \in J_{\text{EB}})) \mathbb{E}[\mu_i | X_i, \sigma_i] \\ &= \frac{1}{n} \sum_{i=1}^n (\mathbb{1}(i \in J_{\text{Bayes}}) - \mathbb{1}(i \in J_{\text{EB}})) \theta_i, \end{aligned} \quad (3)$$

³We assume σ_i to be known, in line with past applications of empirical Bayes methods, e.g. Deng et al. [17], Guo et al. [31], Weinstein et al. [63].

⁴Independence of μ_i and σ_i is a common maintained assumption in empirical Bayes methods, but may be unrealistic in some applications. [9] treats this topic in detail.

where $\mathbb{1}(\cdot)$ is the indicator function.⁵ We aim to characterize how quickly \mathcal{R} converges to zero as $n \rightarrow \infty$ and $m/n \rightarrow \alpha \in (0, 1)$.⁶ The proof that $\mathcal{R} = O_p(r_n^2)$ proceeds by bounding the regret \mathcal{R} by the product of two terms: the proportion of mistakes, and the maximum possible magnitude of the loss caused by a mistake. We show that each of these terms are of the same order as the estimation error in \hat{G} , and consequently regret must be second-order small, i.e. $O_p(r_n) \cdot O_p(r_n) = O_p(r_n^2)$.

Note that in the homoskedastic case when all variances are equal, $\sigma_1 = \dots = \sigma_n$, the posterior mean of μ_i is monotone in X_i for any choice of prior (Efron [20], Koenker and Mizera [44]). Hence the oracle Bayes selection rule amounts to selecting the top- m observations ordered by X_i , and this selection problem is trivial.

2.2 Establishing a Convergence Bound

To establish the convergence bound, we enlist Assumptions 1 and 2.

Assumption 1. $W_1(\hat{G}, G_0) = O_p(r_n)$ for some sequence $(r_n)_{n \in \mathbb{N}}$ with $r_n \geq n^{-1/2}$ for all n .

Assumption 2. The support of the distribution H_0 of σ_i is compact and bounded away from 0.

Assumption 1 will be satisfied under mild conditions by the maximum likelihood estimator when \mathcal{M} is parameterized by some finite-dimensional parameter η , with $r_n = n^{-1/2}$ (Keener [41, Theorem 9.14]).⁷ This includes the commonly used “normal-normal” model, in which the prior is $\mathcal{N}(m_g, v_g)$ for unknown m_g, v_g which are estimated by maximum likelihood. In the nonparametric case we will generally obtain slower rates of convergence, as allowed for by this assumption. For example, from the existing literature on convergence rates for deconvolution problems:

- if the prior takes on some finite but unknown number of values, Chen [8] shows that the best possible convergence rate for estimating the prior in the 1-Wasserstein metric⁸ is $n^{-1/4}$;
- if the prior has a density function with k bounded derivatives, Carroll and Hall [7] shows that the fastest rate of convergence of any estimator of the prior is $(\log n)^{-k/2}$ in the L_1 -norm of the p.d.f. Furthermore, if we assume the support of G is bounded, the same rate of convergence applies to the 1-Wasserstein metric.

⁵We can also consider the loss $U(J) - U(J_{\text{EB}})$ for some other choice of benchmark J . However, other natural choices of J may require oracle knowledge of the order statistics of the μ_i 's, e.g. when J is optimal among the class of permutation invariant choice sets (Weinstein [62]).

⁶For simplicity of exposition, we only consider fixed m . We expect our main results to extend to the case where m is allowed to be mildly data-driven, with $m/n \rightarrow \alpha$ in probability, as when selecting units with positive posterior means.

⁷The convergence rate holds for estimating η , but this translates to common families such as finite mixture families parameterized only by their weights and canonical exponential families with finite variance. For finite mixture families, note that the 1-Wasserstein metric can be bounded by the product of the L_1 -norm of the weight parameters and the maximum 1-Wasserstein metric between any two mixture components. For exponential families, see Lemma 5 in Appendix A.

⁸The 1-Wasserstein metric is a natural choice here. We need a statistical distance that reflects the metric on the observation space, as our regret is tied to that metric as well. We also do not require the distributions to have the same support, or more precisely, be absolutely continuous with respect to G_0 . Other common statistical distances such as total variation distance or Kullback–Leibler divergence do not meet these two desiderata.

Assumption 2 states that there are non-trivial upper and lower bounds on the precision with which the true values are measured, as would be the case in experiments with sample sizes bounded below and above.

Under Assumptions 1 and 2, our main result that $\mathcal{R} = O_p(r_n^2)$ follows. We give a brief overview of the proof strategy behind this theorem, before establishing supporting lemmas and giving the proof itself. Regret arises because our estimated posterior means, $\hat{\theta}_i$, are different from their oracle Bayes counterparts, θ_i , and consequently the top units ranked by the former may differ from the top units ranked by the latter. The difference between $\hat{\theta}_i$ and θ_i is the error relative to oracle Bayes shrinkage for observation i . We show that regret can be bounded above by the product of the maximum magnitude of this shrinkage error from mistakes (whether of inclusion or exclusion) and the proportion of such mistakes. It suffices to show that each of these terms is $O_p(r_n)$. Using the facts that the posterior mean function $f_{G,\sigma}(X_i)$ is sufficiently well-behaved around the G_0 when the observations X_i belong to a compact set (Lemma 2), and that the X_i 's associated with all mistakes lie within a compact set w.h.p. (Lemma 3), we can show that the maximum magnitude of the shrinkage error from mistakes is bounded above by a constant times $W_1(\hat{G}, G_0)$, and hence is $O_p(r_n)$ by Assumption 1. Next we argue that for any neighborhood around $P^{-1}(1 - \frac{m}{n})$ shrinking slower than r_n , the true values associated with mistakes will lie within that neighborhood w.h.p. This allows us to control the proportion of mistakes, and conclude that they are also $O_p(r_n)$.

The following lemma is a key preliminary result, establishing continuity of both the posterior mean function $f_{G,\sigma}(X)$ and its inverse, and will be used to establish Lemmas 2 and 3. The existence of the inverse follows immediately from a classic result by Efron [20].

Lemma 1. Under Assumption 2, the posterior mean function $f_{G,\sigma}(X)$ and its inverse $f_{G,\sigma}^{-1}(X)$ are both continuous in $(G, \sigma, X) \in \mathcal{M} \times \text{supp}(H_0) \times \mathbb{R}$.

PROOF. We first prove that $f_{G,\sigma}(X)$ is continuous in (G, σ, X) . From (1),

$$f_{G,\sigma}(X) = \frac{\int \mu \phi\left(\frac{X-\mu}{\sigma}\right) dG}{\int \phi\left(\frac{X-\mu}{\sigma}\right) dG} =: \frac{h_1(G, \sigma, X)}{h_0(G, \sigma, X)}, \quad (4)$$

where $\phi(\cdot)$ is the p.d.f. of a standard Gaussian. As $h_0 > 0$, it suffices to show that h_0 and h_1 are continuous.

Suppose we have a sequence $(G_k, \sigma_k, X_k) \rightarrow (G^*, \sigma^*, X^*)$ as $k \rightarrow \infty$. For h_1 , we wish to show that

$$\int \mu \phi\left(\frac{X_k - \mu}{\sigma_k}\right) dG_k \rightarrow \int \mu \phi\left(\frac{X^* - \mu}{\sigma^*}\right) dG^*.$$

Note that the function sequence $\mu \phi\left(\frac{X_k - \mu}{\sigma_k}\right)$ converges uniformly to $\mu \phi\left(\frac{X^* - \mu}{\sigma^*}\right)$.⁹ Hence for any $\varepsilon > 0$, when k is sufficiently large,

⁹For any $\varepsilon > 0$, there exists a compact interval C_ε such that $\mu \phi\left(\frac{X_k - \mu}{\sigma_k}\right) < \varepsilon$ on C_ε^c . The function sequence itself is equicontinuous and converges pointwise, so it also converges uniformly within C_ε . Hence for any $\varepsilon > 0$ there is sufficiently large k such that $\mu \phi\left(\frac{X_k - \mu}{\sigma_k}\right)$ is within ε of $\mu \phi\left(\frac{X^* - \mu}{\sigma^*}\right)$ pointwise.

we have

$$\begin{aligned} & \left| \int \mu \phi \left(\frac{X_k - \mu}{\sigma_k} \right) dG_k - \int \mu \phi \left(\frac{X^* - \mu}{\sigma^*} \right) dG_k \right| \\ & \leq \sup_{\mu \in \mathbb{R}} \left| \mu \phi \left(\frac{X_k - \mu}{\sigma_k} \right) - \int \mu \phi \left(\frac{X^* - \mu}{\sigma^*} \right) \right| \\ & < \varepsilon/2. \end{aligned} \quad (5)$$

Note also that $\mu \phi \left(\frac{X^* - \mu}{\sigma^*} \right)$ is Lipschitz. Therefore by Kantorovich–Rubinstein duality, when k is sufficiently large, $W_1(G_k, G^*)$ is sufficiently small and

$$\left| \int \mu \phi \left(\frac{X^* - \mu}{\sigma^*} \right) dG_k - \int \mu \phi \left(\frac{X^* - \mu}{\sigma^*} \right) dG^* \right| < \varepsilon/2. \quad (6)$$

Summing up (5) and (6) yields the convergence of the numerator. The proof for h_0 is almost identical, as $\phi \left(\frac{X_k - \mu}{\sigma_k} \right)$ converges uniformly to $\phi \left(\frac{X^* - \mu}{\sigma^*} \right)$ and $\phi \left(\frac{X^* - \mu}{\sigma^*} \right)$ is Lipschitz. The continuity of $f_{G,\sigma}^{-1}(X)$ follows from the continuity of $f_{G,\sigma}(X)$ by Lemmas 6 and 8. \square

The next lemma states that posterior mean function $f_{G,\sigma}(X)$ is locally Lipschitz around G_0 , uniformly in $(\sigma, X) \in \text{supp}(H_0) \times W$, for any compact W . This will be used in Theorem 4 to bound the shrinkage error by a constant times the estimation error in the prior parameter, \hat{G} .

Lemma 2. *Suppose Assumption 2 holds. Then for any compact $W \subset \mathbb{R}$, there exist positive constants K, δ such that for all $(G, \sigma, X) \in \mathcal{M} \times \text{supp}(H_0) \times W$, we have $|f_{G,\sigma}(X) - f_{G_0,\sigma}(X)| \leq KW_1(G, G_0)$ whenever $W_1(G, G_0) < \delta$.*

PROOF. Using the same definition for h_0 and h_1 as in (4), we have

$$\begin{aligned} & |f_{G,\sigma}(X) - f_{G_0,\sigma}(X)| \\ &= \left| \frac{h_1(G, \sigma, X)}{h_0(G, \sigma, X)} - \frac{h_1(G_0, \sigma, X)}{h_0(G_0, \sigma, X)} \right| \\ &\leq \frac{|h_1(G, \sigma, X) - h_1(G_0, \sigma, X)|}{h_0(G, \sigma, X)} \\ &\quad + \frac{|h_1(G_0, \sigma, X)| \cdot |h_0(G_0, \sigma, X) - h_0(G, \sigma, X)|}{h_0(G, \sigma, X)h_0(G_0, \sigma, X)}. \end{aligned}$$

It remains to show the following claims for when G is in a sufficiently small neighborhood of G_0 :

- h_0 is bounded away from 0: Since h_0 is continuous from the proof of Lemma 1, $h_0(G_0, \sigma, X)$ is strictly positive and $\text{supp}(H_0) \times W$ is compact, for sufficiently small δ , $h_0(G, \sigma, X)$ is bounded away from 0 whenever $W_1(G, G_0) < \delta$.
- h_1 is Lipschitz in G with a Lipschitz constant that does not depend on σ or X : The integrand in h_1 is $\mu \phi \left(\frac{X - \mu}{\sigma} \right)$, a Lipschitz function in μ . Since this Lipschitz constant is a continuous function in σ, X and $\text{supp}(H_0) \times W$ is compact, $\mu \phi \left(\frac{X - \mu}{\sigma} \right)$ is uniformly Lipschitz. The function h_1 is Lipschitz in G again by Kantorovich–Rubinstein duality.
- $h_1(G_0, \sigma, X)$ is bounded: From the proof of Lemma 1, h_1 and thus $h_1(G_0, \cdot, \cdot)$ are bounded. So $h_1(G_0, \sigma, X)$ is bounded since $\text{supp}(H_0) \times W$ is compact. \square

We will apply Lemma 2 on a specific W which contains all of the observations corresponding to mistakes made by empirical Bayes selection w.h.p. This is the subject of the following lemma. We use Δ to denote symmetric difference, so $J_{\text{Bayes}} \Delta J_{\text{EB}}$ is the index set of all mistakes.

Lemma 3. *If Assumptions 1 and 2 hold, there exists a compact set W such that $X_i \in W$ for all $i \in J_{\text{Bayes}} \Delta J_{\text{EB}}$ w.h.p.*

PROOF. Oracle Bayes selection essentially thresholds on θ^* , the m -th largest order statistic of the θ_i 's. For any $c > 0$, this threshold lands in $(P^{-1}(1 - \frac{m}{n}) - c, P^{-1}(1 - \frac{m}{n}) + c)$ w.h.p. and hence $(P^{-1}(1 - \alpha) - c, P^{-1}(1 - \alpha) + c)$ w.h.p., by van der Vaart [60, Corollary 21.5 and discussion thereof]. Let V be an open ball of G_0 in 1-Wasserstein whose radius is fixed but to be determined later. By Assumption 1, \hat{G} lies in V w.h.p. For any $i \notin J_{\text{EB}}$, under the high probability event A_n defined as $A_n = \{\theta^* \in (P^{-1}(1 - \alpha) - c, P^{-1}(1 - \alpha) + c)\} \cap \{\hat{G} \in V\}$,

$$\begin{aligned} \hat{\theta}_i &\leq \min_{i' \in J_{\text{EB}}} \hat{\theta}_{i'} \\ &= \min_{i' \in J_{\text{EB}}} f_{\hat{G}, \sigma_{i'}} \circ f_{G_0, \sigma_{i'}}^{-1}(\theta_{i'}) \\ &\leq \min_{i' \in J_{\text{EB}}} \max_{\sigma' \in \text{supp}(H_0)} f_{\hat{G}, \sigma'} \circ f_{G_0, \sigma'}^{-1}(\theta_{i'}) \end{aligned} \quad (7)$$

$$\begin{aligned} &= \max_{\sigma' \in \text{supp}(H_0)} f_{\hat{G}, \sigma'} \circ f_{G_0, \sigma'}^{-1} \left(\min_{i' \in J_{\text{EB}}} \theta_{i'} \right) \\ &\leq \max_{\sigma' \in \text{supp}(H_0)} f_{\hat{G}, \sigma'} \circ f_{G_0, \sigma'}^{-1} \left(\min_{i' \in J_{\text{Bayes}}} \theta_{i'} \right) \end{aligned} \quad (8)$$

$$\leq \max_{\sigma' \in \text{supp}(H_0)} f_{\hat{G}, \sigma'} \circ f_{G_0, \sigma'}^{-1}(P^{-1}(1 - \alpha) + c) \quad (9)$$

$$X_i \leq \max_{\sigma' \in \text{supp}(H_0)} f_{\hat{G}, \sigma_i}^{-1} \circ f_{\hat{G}, \sigma'} \circ f_{G_0, \sigma'}^{-1}(P^{-1}(1 - \alpha) + c) \quad (10)$$

$$\leq \max_{\sigma', \sigma'' \in \text{supp}(H_0)} f_{\hat{G}, \sigma''}^{-1} \circ f_{\hat{G}, \sigma'} \circ f_{G_0, \sigma'}^{-1}(P^{-1}(1 - \alpha) + c) \quad (11)$$

By Lemmas 6 and 8, the function $\max_{\sigma' \in \text{supp}(H_0)} f_{\hat{G}, \sigma'} \circ f_{G_0, \sigma'}^{-1}$ is a strictly increasing function, so we can move the minimum inside in (7). Also applying $f_{\hat{G}, \sigma_i}^{-1}$ to both sides of (9) yields (10). Note that the maximand in (11) is a composition of functions in $\hat{G}, \sigma', \sigma''$ that are continuous by Lemma 1, hence also a continuous function itself. By maximum theorem and the compactness of $\text{supp}(H_0)$, (11) is continuous in \hat{G} and locally bounded. In other words, for a sufficiently small open ball in 1-Wasserstein, V , centered at G_0 , (11) is bounded by some constant.

On the other hand, for any $i \in J_{\text{Bayes}}$, under the event A_n ,

$$\begin{aligned} X_i &= f_{G_0, \sigma_i}^{-1}(\theta_i) \\ &\geq f_{G_0, \sigma_i}^{-1}(P^{-1}(1 - \alpha) - c) \\ &\geq \min_{\sigma' \in \text{supp}(H_0)} f_{G_0, \sigma'}^{-1}(P^{-1}(1 - \alpha) - c). \end{aligned}$$

Together, there exists a constant bounded interval that contains all i in $J_{\text{Bayes}} \setminus J_{\text{EB}}$ under the event A_n . Likewise, there is also a constant bounded interval contains all i in $J_{\text{EB}} \setminus J_{\text{Bayes}}$ under the event A_n . Taking W to be the union of these two intervals completes the proof. \square

With these preliminaries we can prove our main result.

Theorem 4. *If Assumptions 1 and 2 hold, then $\mathcal{R} = O_p(r_n^2)$, the square of the rate of convergence for estimating the prior.*

PROOF. We first decompose an upper bound for \mathcal{R} into two components.

$$\begin{aligned} \mathcal{R} &\leq \frac{1}{n} \sum_{i=1}^n (\mathbb{1}(i \in J_{\text{Bayes}}) - \mathbb{1}(i \in J_{\text{EB}})) \theta_i \\ &\quad - \frac{1}{n} \sum_{i=1}^n (\mathbb{1}(i \in J_{\text{Bayes}}) - \mathbb{1}(i \in J_{\text{EB}})) \widehat{\theta}_i \end{aligned} \quad (12)$$

$$\begin{aligned} &= \frac{1}{n} \sum_{i=1}^n (\mathbb{1}(i \in J_{\text{Bayes}}) - \mathbb{1}(i \in J_{\text{EB}})) (\theta_i - \widehat{\theta}_i) \\ &\leq \frac{1}{n} (\#(J_{\text{Bayes}} \setminus J_{\text{EB}}) + \#(J_{\text{EB}} \setminus J_{\text{Bayes}})) \\ &\quad \cdot \max_{i \in J_{\text{Bayes}} \Delta J_{\text{EB}}} |\theta_i - \widehat{\theta}_i| \end{aligned} \quad (13)$$

$$= 2 \cdot \underbrace{\frac{1}{n} \#(J_{\text{Bayes}} \setminus J_{\text{EB}})}_{\text{proportion of mistakes}} \cdot \underbrace{\max_{i \in J_{\text{Bayes}} \Delta J_{\text{EB}}} |\theta_i - \widehat{\theta}_i|}_{\text{max magnitude of shrinkage error}}, \quad (14)$$

where (12) follows from the fact that J_{EB} is the set of indices of the m largest $\widehat{\theta}_i$'s. In (13), since $\#J_{\text{Bayes}} = \#J_{\text{EB}}$, we have $\#(J_{\text{Bayes}} \setminus J_{\text{EB}}) = \#(J_{\text{EB}} \setminus J_{\text{Bayes}})$. From (14), it suffices to bound the proportion of mistakes and the maximum magnitude of shrinkage error.

We start by bounding the latter term. We denote the event that the observations associated with all mistakes belong in the set W from Lemma 3 and \widehat{G} belongs to a neighborhood V around G_0 by $A_n = \{X_i \in W \text{ for all } i \in J_{\text{Bayes}} \Delta J_{\text{EB}}\} \cap \{\widehat{G} \in V\}$. By Lemma 3, $\mathbb{P}(A_n) \rightarrow 1$, and under this high probability event, we have

$$\begin{aligned} \max_{i \in J_{\text{Bayes}} \Delta J_{\text{EB}}} |\theta_i - \widehat{\theta}_i| &= \max_{i \in J_{\text{Bayes}} \Delta J_{\text{EB}}} |f_{G_0, \sigma_i}(X_i) - f_{\widehat{G}, \sigma_i}(X_i)| \\ &\leq \max_{\substack{X \in W \\ \sigma \in \text{supp}(H_0)}} |f_{G_0, \sigma}(X) - f_{\widehat{G}, \sigma}(X)| \\ &\leq KW_1(G_0, \widehat{G}). \end{aligned} \quad (15)$$

(15) follows from Lemma 2, and is $O_p(r_n)$ by Assumption 1. Consequently

$$\max_{i \in J_{\text{Bayes}} \Delta J_{\text{EB}}} |\theta_i - \widehat{\theta}_i| = O_p(r_n). \quad (16)$$

Next we bound the proportion of mistakes. Let θ^* denote the m -th largest order statistic of the θ_i 's, and $\theta^{**} = P^{-1}(1 - m/n)$ the $(1 - m/n)^{\text{th}}$ quantile of P .

For any nondecreasing sequence $(b_n)_{n \in \mathbb{N}}$ with $\lim_{n \rightarrow \infty} b_n = \infty$, define the event B_n as

$$B_n = \left\{ \max_{i \in J_{\text{Bayes}} \setminus J_{\text{EB}}} |\theta_i - \theta^{**}| \leq b_n r_n \right\}.$$

Then

$$\begin{aligned} &\frac{1}{n} \#(J_{\text{Bayes}} \setminus J_{\text{EB}}) \\ &\leq \mathbb{1}(B_n^c) + \mathbb{1}(B_n) \frac{1}{n} \# \{i : |\theta_i - \theta^{**}| \leq b_n r_n\}. \end{aligned}$$

We first argue that $\mathbb{P}(B_n^c) \rightarrow 0$ and subsequently that $\frac{1}{n} \# \{i : |\theta_i - P^{-1}(1 - \frac{m}{n})| \leq b_n r_n\} = O_p(b_n r_n)$, giving $\frac{1}{n} \#(J_{\text{Bayes}} \setminus J_{\text{EB}}) = O_p(b_n r_n)$. Since $(b_n)_{n \in \mathbb{N}}$ was an arbitrary nondecreasing sequence converging to infinity, by Lemma 9 in Appendix A this implies $\frac{1}{n} \#(J_{\text{Bayes}} \setminus J_{\text{EB}}) = O_p(r_n)$.

For each i in $J_{\text{Bayes}} \setminus J_{\text{EB}}$, empirical Bayes selection must have excluded it because some other shrinkage estimate was larger, i.e. $\widehat{\theta}_i \leq \widehat{\theta}_{i'}$ for some $i' \in J_{\text{EB}} \setminus J_{\text{Bayes}}$. Hence for each $i \in J_{\text{Bayes}} \setminus J_{\text{EB}}$, there is a $i' \in J_{\text{EB}} \setminus J_{\text{Bayes}}$ such that $\theta^* \leq \theta_i \leq \theta_{i'} + 2 \max_{i \in J_{\text{Bayes}} \Delta J_{\text{EB}}} |\theta_i - \widehat{\theta}_i|$ and so

$$\max_{i \in J_{\text{Bayes}} \setminus J_{\text{EB}}} |\theta_i - \theta^*| \leq 2 \max_{i \in J_{\text{Bayes}} \Delta J_{\text{EB}}} |\theta_i - \widehat{\theta}_i|. \quad (17)$$

From the triangle inequality and union bound, we have

$$\begin{aligned} \mathbb{P}(B_n^c) &\leq \mathbb{P} \left(r_n^{-1} \max_{i \in J_{\text{Bayes}} \setminus J_{\text{EB}}} |\theta_i - \theta^*| > b_n/2 \right) + \\ &\quad \mathbb{P} (r_n^{-1} |\theta^* - \theta^{**}| > b_n/2). \end{aligned}$$

By (16) and (17), $r_n^{-1} \max_{i \in J_{\text{Bayes}} \setminus J_{\text{EB}}} |\theta_i - \theta^*| = O_p(1)$. By Lemma 7, given standard results on the convergence of sample quantiles (van der Vaart [60, Corollary 21.5]), we have $r_n^{-1} |\theta^* - P^{-1}(1 - \frac{m}{n})| = O_p(r_n^{-1} n^{-1/2})$. As $b_n \rightarrow \infty$, $\mathbb{P}(B_n^c) \rightarrow 0$.

The probability of θ_i that falls in $(P^{-1}(1 - \frac{m}{n}) - b_n r_n, P^{-1}(1 - \frac{m}{n}) + b_n r_n)$ is no greater than $P(\theta^{**} + b_n r_n) - P(\theta^{**} - b_n r_n) = O(b_n r_n)$ by the continuous differentiability of P from Lemma 7. So by Chebyshev's inequality, the proportion $\frac{1}{n} \# \{i : |\theta_i - P^{-1}(1 - \frac{m}{n})| < b_n r_n\}$ is $O_p(b_n r_n)$, and by arbitrariness of b_n , it must also be $O_p(r_n)$.

Because the first part and second parts of (14) are both $O_p(r_n)$, it follows that the regret \mathcal{R} is $O_p(r_n^2)$. \square

The two main estimation approaches for empirical Bayes are *f-modeling*, in which a model is specified for the observed outcomes, and *g-modeling*, in which a model is specified for the unobserved prior (Efron [21]). This theorem is consistent with either estimation approach. In the *f*-modelling case, if the estimated distribution for outcomes is consistent with some prior distribution for true effects, i.e. falls in the class characterized by Guo et al. [31], we can think of \widehat{G} as the prior implicitly specified by deconvolving the estimated observation distribution. For *g*-modelling, we can interpret \widehat{G} directly as the model specified for the unobserved prior.

The bound is also sharp when $r_n = n^{-1/2}$, as shown by our example in Section 3.

3 Sharpness of the Convergence Bound in the Parametric Case

We provide an example where the regret satisfies $\mathcal{R} \geq Cn^{-1}$ with non-vanishing probability for some positive constant C . Let the location family $G(\eta) = \mathcal{N}(\eta, 1)$ be the model for the prior, where the scalar location parameter η is estimated by maximum likelihood. In our example, the truth is $\eta_0 = 0$. We assume the standard deviation of the noise term is drawn i.i.d. from

$$\sigma_i = \begin{cases} 1 & \text{with probability } 1/2, \text{ and} \\ 2 & \text{with probability } 1/2; \end{cases}$$

and we will select $m = \lfloor \alpha n \rfloor$ units.

The maximum likelihood estimator $\widehat{\eta}$ converges to η_0 at rate $n^{-1/2}$. The oracle Bayes shrunk estimate of the posterior mean is $\theta_i = \frac{1}{\sigma_i^2 + 1} X_i$ and the empirical Bayes estimate is $\widehat{\theta}_i = \frac{1}{\sigma_i^2 + 1} X_i + \frac{\sigma_i^2}{\sigma_i^2 + 1} \widehat{\eta}$.

In particular, the magnitude of $\widehat{\theta}_i - \theta_i = \frac{\sigma_i^2}{\sigma_i^2 + 1} \widehat{\eta}$ increases with σ_i . In our setting θ_i and σ_i are measurable with respect to the Lebesgue

measure and the counting measure, respectively. The density of (θ_i, σ_i) with respect to the product measure is then $\frac{1}{2} \cdot \sqrt{2}\phi(\sqrt{2}\theta_i)$ for $\sigma_i = 1$ and $\frac{1}{2} \cdot \sqrt{5}\phi(\sqrt{5}\theta_i)$ for $\sigma_i = 2$. If we condition on θ^* and $\theta_i < \theta^*$, then θ_i are i.i.d. In fact $(\theta_i, \sigma_i) \mid \theta^*, \theta_i < \theta^*$ is i.i.d. with density

$$\begin{cases} \frac{\sqrt{2}\phi(\sqrt{2}\theta_i)}{\Phi(\sqrt{2}\theta^*) + \Phi(\sqrt{5}\theta^*)} \mathbb{1}(\theta_i < \theta^*) & \text{for } \sigma_i = 1, \\ \frac{\sqrt{5}\phi(\sqrt{5}\theta_i)}{\Phi(\sqrt{2}\theta^*) + \Phi(\sqrt{5}\theta^*)} \mathbb{1}(\theta_i < \theta^*) & \text{for } \sigma_i = 2, \end{cases} \quad (18)$$

with respect to the product measure, where $\Phi(\cdot)$ is the c.d.f. of a standard Gaussian. Likewise, $(\theta_i, \sigma_i) \mid \theta^*, \theta_i > \theta^*$ is i.i.d. with density

$$\begin{cases} \frac{\sqrt{2}\phi(\sqrt{2}\theta_i)}{1 - \Phi(\sqrt{2}\theta^*) + 1 - \Phi(\sqrt{5}\theta^*)} \mathbb{1}(\theta_i > \theta^*) & \text{for } \sigma_i = 1, \\ \frac{\sqrt{5}\phi(\sqrt{5}\theta_i)}{1 - \Phi(\sqrt{2}\theta^*) + 1 - \Phi(\sqrt{5}\theta^*)} \mathbb{1}(\theta_i > \theta^*) & \text{for } \sigma_i = 2. \end{cases} \quad (19)$$

Consider a compact interval $[\underline{a}, \bar{a}]$ that contains $P^{-1}(1 - \alpha)$ in its interior. Since θ^* converges to $P^{-1}(1 - \alpha)$ at rate $n^{-1/2}$, the event A_n where the interval $(\theta^* - cn^{-1/2}, \theta^* + cn^{-1/2})$ is a subset of $[\underline{a}, \bar{a}]$ happens w.h.p. for any positive constant $c > 0$. For any such θ^* , the density in (18) over $(\theta^* - cn^{-1/2}, \theta^*)$ and density in (19) over $(\theta^*, \theta^* + cn^{-1/2})$ are in some strictly positive bounded interval $[\underline{b}, \bar{b}]$ that does not depend on the value of θ^* .

There are three sets of units of interest:

- $K_n = \{i : \theta_i \in (\theta^*, \theta^* + cn^{-1/2}) \text{ and } \sigma_i = 1\}$,
- $L_n = \{i : \theta_i \in (\theta^* - dn^{-1/2}, \theta^*) \text{ and } \sigma_i = 2\}$, and
- $M_n = \{i : \theta_i \in (\theta^* - dn^{-1/2}, \theta^* - \frac{1}{2}dn^{-1/2}) \text{ and } \sigma_i = 2\}$,

where c, d are positive constants to be chosen. There are $\lfloor \alpha n \rfloor - 1$ realizations of θ_i greater than θ^* and $n - \lfloor \alpha n \rfloor$ realizations of θ_i smaller than θ^* . Conditional on θ^* , the cardinalities are consequently binomially distributed with

$$\#K_n \sim \text{Binomial}(\lfloor \alpha n \rfloor - 1, p_{K_n}(\theta^*)), \quad \text{where } p_{K_n}(\theta^*) > \underline{b}cn^{-1/2},$$

$$\#L_n \sim \text{Binomial}(n - \lfloor \alpha n \rfloor, p_{L_n}(\theta^*)),$$

$$\text{where } \underline{b}dn^{-1/2} < p_{L_n}(\theta^*) < \bar{b}dn^{-1/2},$$

$$\#M_n \sim \text{Binomial}(n - \lfloor \alpha n \rfloor, p_{M_n}(\theta^*)),$$

$$\text{where } p_{M_n}(\theta^*) > \frac{1}{2}\underline{b}dn^{-1/2}.$$

Marginalizing over the event A_n gives the same observation but removes the dependence on θ^* . Hence for some constants $c_K, c_L, c_M > 0$, we have w.h.p. $\#K_n \geq c_K n^{1/2}$, $\#L_n \geq c_L n^{1/2}$, and $\#M_n \geq c_M n^{1/2}$. Furthermore $d > 0$ can be chosen sufficiently small such that $\#K_n > \#L_n$ w.h.p.

The rest of the argument focuses on the event where $\hat{\eta} > \frac{10}{3}(c + d)n^{-1/2}$, which occurs with non-vanishing probability. Under this event, since $\hat{\eta} > 0$, empirical Bayes selection will only mistakenly select units with $\sigma_i = 2$ in place of other units with $\sigma_i = 1$. In particular, for any i in K_n and i' in L_n , we have $\theta_i > \theta_{i'}$ but

$$\hat{\theta}_i < \theta^* + cn^{-1/2} + \frac{1}{2}\hat{\eta} < \theta^* - dn^{-1/2} + \frac{4}{5}\hat{\eta} < \hat{\theta}_{i'}.$$

So w.h.p. at least $\min(\#K_n, \#L_n) = \#L_n$ mistakes were made. In fact since L_n consists of units immediately smaller than θ^* and the relative ordering of all units with $\sigma_i = 2$ does not change, all of $\#L_n$

will be mistakenly selected. This incurs a regret of at least

$$\frac{1}{n} \sum_{i \in L_n} (\theta^* - \theta_i) \geq \frac{1}{n} \sum_{i \in M_n} (\theta^* - \theta_i) \geq \frac{1}{n} \#M_n \cdot \frac{1}{2} dn^{-1/2} \geq \frac{1}{2} c_M dn^{-1},$$

with high probability.

4 Top- m Selection in Simulation

We illustrate Theorem 4 with a realistic simulation, based on the Upworthy dataset of internet experiments conducted between 2013 and 2015.¹⁰ The dataset contains a list of experiments, along with effect sizes and standard errors. For the prior G_0 , we fit a normal scale mixture with fixed components, parameterized only by the weights. The data and modelling details are described in Appendix B, and the notebook to reproduce the simulations and figures is available as an artifact¹¹.

We simulate a variety of signal-to-ratio regimes, and choices of family for the prior. In increasing order of flexibility, these are: (i) the family of normal priors, (ii) the family of scale mixtures of normals, and (iii) the family of all distributions. The priors for these cases are estimated using the `ebnm` R package (Willwerscheid and Stephens [64]). In particular, the normal scale mixture is estimated using adaptive shrinkage as in Stephens [58], and the fully nonparametric case is estimated by nonparametric maximum likelihood estimator (NPMLE) (Kiefer and Wolfowitz [42]). Henceforth we refer to these three estimators as EB-NN, EB-NSM, and EB-NPMLE. This enables comparison of the performance of empirical Bayes methods under misspecification (when the restrictive EB-NN estimator is used), under a parsimonious and well-specified model (EB-NSM), and under a highly flexibly and well-specified model (EB-NPMLE). For the distribution H_0 , we use the empirical distribution of standard errors in the dataset.

Top- m selection here corresponds to selecting a subset of experiments, given a constraint on the subset size. We pick $m = \lfloor 0.1n \rfloor$ and vary n , the number of simulated experiments, showing the distribution of regret for each choice of n . For each n we run 1000 iterations of the selection simulation. In each iteration, we

- (1) independently draw n true treatment effects $\mu_i \sim G_0$ and noise standard deviations $\sigma_i \sim H_0$;
- (2) generate the n observations X_i , where $X_i \mid \mu_i, \sigma_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$;
- (3) fit three models for the prior distribution of treatment effects μ_i : EB-NN, EB-NSM, and EB-NPMLE;
- (4) compute the choice sets J_{Bayes} , $J_{\text{EB-NN}}$, $J_{\text{EB-NSM}}$, $J_{\text{EB-NPMLE}}$, and J_{UN} corresponding to the oracle Bayes posterior mean estimators, the three empirical Bayes posterior mean estimators, and the unshrunk X_i ;
- (5) compute the regret relative to oracle Bayes selections, $\mathcal{R}_M = \frac{1}{n} \sum_{i=1}^n (\mathbb{1}(i \in J_{\text{Bayes}}) - \mathbb{1}(i \in J_M)) \theta_i$ for $M = \text{EB-NN, EB-NSM, EB-NPMLE, UN}$.

To assess performance in lower signal-to-noise regimes, we repeat this exercise with varying levels of sampling error. We use standard errors 1, 2 and 4 times greater than the baseline standard errors, corresponding to signal-to-noise ratios of roughly 1.3, 0.7 and 0.3.

¹⁰From the publicly accessible Upworthy Research Archive (Matias et al. [48]) which is downloadable at <https://osf.io/jd64p/>.

¹¹Also available on <https://github.com/facebookresearch/eb-selection>

The normal scale mixture is a parametric model once the number of components and the scale parameters are fixed. As ebnm fits by maximizing the likelihood, the remaining parameters—the weights—converge at $O_p(n^{-1/2})$. Hence by Theorem 4, $\mathcal{R}_{\text{EB-NSM}}$ is $O_p(n^{-1})$. We have no such guarantees for $\mathcal{R}_{\text{EB-NN}}$ or $\mathcal{R}_{\text{EB-UN}}$, corresponding to the misspecified normal prior and the “naive” choice which selects the units with the largest X_i ’s. EB-NPMLE is highly flexible and not misspecified, although its guaranteed convergence rate is very slow [57].

Figure 1 shows regret as a function of the number of experiments, for each selection method and each value of the noise multiplier. As n increases, the mean and 99th percentile across simulations of $\mathcal{R}_{\text{EB-NSM}}$ and $\mathcal{R}_{\text{EB-NPMLE}}$ both exhibit declines consistent with n^{-1} convergence, although the regret associated with the latter is larger, suggesting the NPMLE model incurs a cost from its greater flexibility. With just 1000 experiments, the regret of the EB-NSM approach can be as low as 10^{-4} times the standard error of the noise. The normal prior performs better than the unshrunk selection procedure, but neither has regret approaching zero. These patterns are consistent across different noise levels, although the regret are lower with less noise, as the oracle prior and estimated prior are closer.

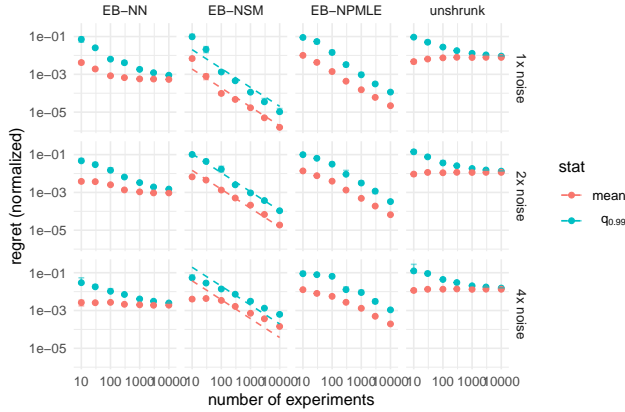


Figure 1: Regret \mathcal{R} as the number of experiments n increases, on a log scale and normalized by the standard deviation of the noise. The 95% confidence intervals are due to simulation uncertainty. For choice sets based on the correctly specified EB-NSM and EB-NPMLE models, both the 99th percentile and the mean show a trend of $O(n^{-1})$. Regret does not appear to converge to zero for choice sets based on EB-NN or the unshrunk estimates.

We compute other quantities of interest from (14), such as the proportion of mistakes in Figure 2 and the maximum magnitude of shrinkage error in Figure 3, as well as the 1-Wasserstein distance between the true prior and the estimated prior in Figure 4. As expected, we see that the proportion of mistakes, their magnitude, and the 1-Wasserstein distance between the true and estimated prior in the correctly specified EB-NSM model all converge to zero at $n^{-1/2}$. The misspecified EB-NN model and the unshrunk procedure perform poorly in comparison, with the proportion and magnitude

of mistakes not converging to zero, or even increasing, with the number of experiments. The most flexible model, EB-NPMLE, performs worse along every dimension than the more parsimonious EB-NSM, although the proportion and magnitude of its mistakes both converge to zero.

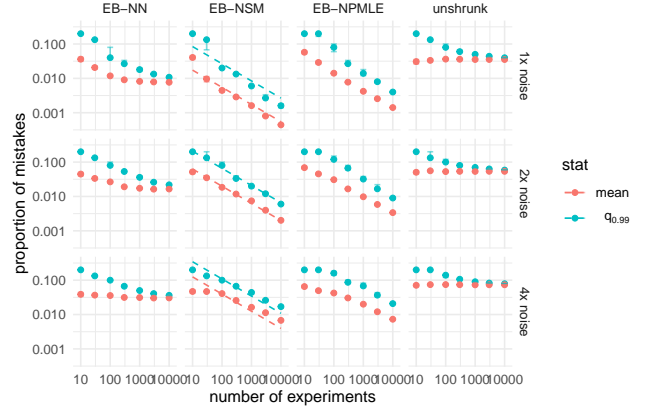


Figure 2: Proportion of mistakes as n increases, on a log scale. The 95% confidence intervals are due to simulation uncertainty. For the correctly specified $J_{\text{EB-NSM}}$, both the 99th percentile and the mean show a trend of $O(n^{-1/2})$. The highly flexible NPMLE shows a similar trend but generally makes more mistakes. The proportion does not appear to decrease towards zero for $J_{\text{EB-NN}}$ or J_{UN} .

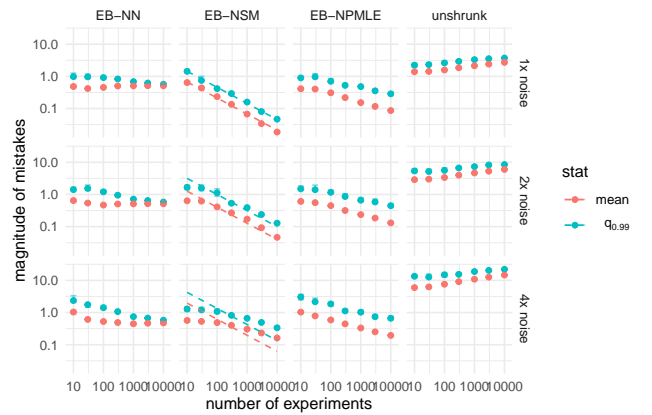


Figure 3: Maximum magnitude of shrinkage error as n increases, on a log scale and normalized by the standard deviation of the noise. The 95% confidence intervals are due to simulation uncertainty. For the correctly specified $J_{\text{EB-NSM}}$, both the 99th percentile and the mean show a trend of $O(n^{-1/2})$ in the lower noise settings. The maximum magnitude does not appear to decrease indefinitely for $J_{\text{EB-NN}}$ or J_{UN} .

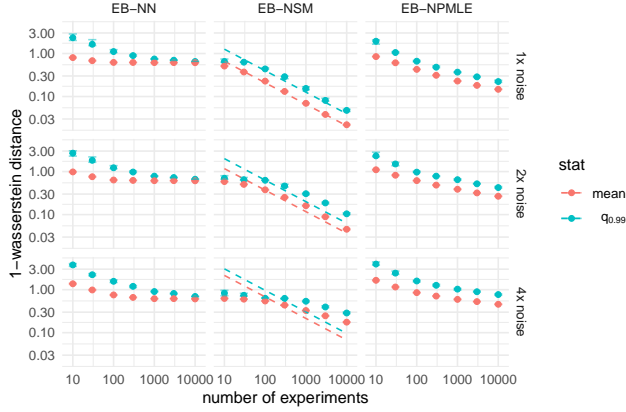


Figure 4: The 1-Wasserstein distance between the true prior and the estimated prior, as n increases, on a log scale and normalized by the standard deviation of the noise. The 95% confidence intervals are due to simulation uncertainty. As n gets large, both the 99th percentile and the mean show a trend of $O(n^{-1/2})$ for EB-NSM. The distance levels off away from zero for EB-NN because of misspecification. The distance for EB-NPMLE decreases at a slower rate than $O(n^{-1/2})$.

4.1 Estimated standard error

The simulations above assume the known standard error to be known, which is reasonable for large-scale online experiments where each experiments have million of units. We complement the simulations above to demonstrate how the noise in the estimated standard error will affect the performance of empirical Bayes methods, showing the regret as the number of units increase and the estimation for standard error improves in Figure 5.

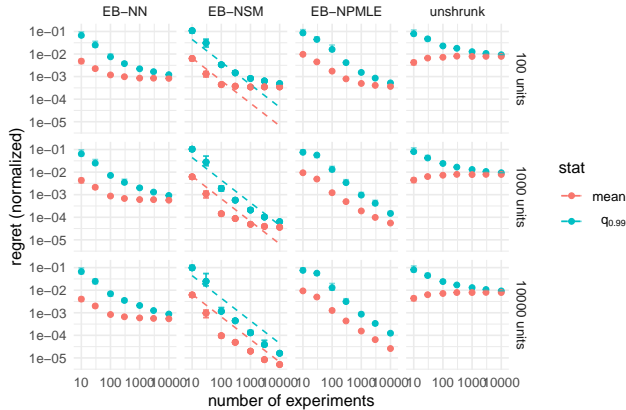


Figure 5: Regret \mathcal{R} as the number of experiments n increases, on a log scale and normalized by the standard deviation of the noise. Error bars around each point are the 95% confidence intervals from uncertainty due to simulation error. Our result holds better as the number of units increase.

5 Conclusion

Our results show that empirical Bayes methods perform well in maximizing the aggregate value of the selected units, in the sense that the regret they incur converges to zero faster than the estimation error in the values themselves. This stands in contrast to prior work emphasizing the difficulty of accurately selecting the best units when the decision-maker incurs a discrete loss from each misclassification (e.g. Gu and Koenker [30], Lin et al. [45], Lockwood et al. [46]). This underscores that rather than selection being an inherently difficult problem, it depends on whether misclassification errors should be weighted by their severity in the utility function. Finally, we note that many extensions and variations on this setting are yet to be fully explored, including characterizing the performance of decision rules for the frequentist analog of the Bayesian regret we study, treating the true values of units as non-stochastic,¹² improving performance by incorporating unit-specific covariates into the analysis; and extending to an empirical Bayes knapsack problem where the selected units incur heterogeneous costs.

As discussed in Section 1, the frequentist optimal solution requires unavailable oracle knowledge of the order statistics of μ_i 's. This implies that the empirical Bayes solution is not optimal in a frequentist sense, with mainly two gaps: (i) the optimal solution in Weinstein [62, Theorem 1] is the Bayesian solution with a uniform prior on the permutations of μ_i 's, while empirical Bayes uses \hat{G} instead; (ii) Weinstein [62] focused on the loss for a specific set of μ_i 's, while our analysis averages this over G_0 .

We suspect these gaps are small. For the first gap, Weinstein [62, Section 6] conjectures that the Bayesian solution using the e.c.d.f. of μ_i 's is asymptotically optimally. We believe this can be reasonably recovered as \hat{G} when the class of priors \mathcal{M} is sufficiently large. For the second gap, Weinstein [62, Theorem 1] showed that minimizing the loss is equivalently to minimizing the loss averaged over a uniform permutation of μ_i 's. Asymptotically this should be close to the loss averaged over G_0 , our regret \mathcal{R} . Putting this together, both the solution and loss function are similar between the frequentist and the empirical Bayes settings, hinting at some loose frequentist optimality of the empirical Bayes approach.

References

- [1] Isaiah Andrews, Toru Kitagawa, and Adam McCloskey. 2019. *Inference on winners*. Technical Report. National Bureau of Economic Research.
- [2] Eduardo M Azevedo, Alex Deng, José Luis Montiel Olea, Justin Rao, and E Glen Weyl. 2020. A/B testing with fat tails. *Journal of Political Economy* 128, 12 (2020), 4614–000.
- [3] Eduardo M Azevedo, Alex Deng, José L Montiel Olea, and E Glen Weyl. 2019. Empirical bayes estimation of treatment effects with many A/B tests: An overview. In *AEA Papers and Proceedings*, Vol. 109. 43–47.
- [4] Peter Bergman, Raj Chetty, Stefanie DeLuca, Nathaniel Hendren, Lawrence F Katz, and Christopher Palmer. 2019. *Creating moves to opportunity: Experimental evidence on barriers to neighborhood choice*. Technical Report. National Bureau of Economic Research.
- [5] Lawrence D. Brown. 1986. *Fundamentals of statistical exponential families with applications in statistical decision theory*. Institute of Mathematical Statistics.
- [6] Lawrence D Brown. 2008. In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *The Annals of Applied Statistics* 2, 1 (2008), 113–152.
- [7] Raymond J Carroll and Peter Hall. 1988. Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.* 83, 404 (1988), 1184–1186.

¹²Specifically, studying the utility $U(J_{EB})$ under permutation invariance as outlined in Weinstein [62].

- [8] Jiahua Chen. 1995. Optimal rate of convergence for finite mixture models. *The Annals of Statistics* (1995), 221–233.
- [9] Jiafeng Chen. 2022. Empirical Bayes when estimation precision predicts parameters. *arXiv preprint* (2022).
- [10] Lijie Chen, Jian Li, and Mingda Qiao. 2017. Nearly Instance Optimal Sample Complexity Bounds for Top-k Arm Selection. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 54)*, Aarti Singh and Jerry Zhu (Eds.). PMLR, 101–110.
- [11] Raj Chetty, John N Friedman, and Jonah E Rockoff. 2014. Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American economic review* 104, 9 (2014), 2593–2632.
- [12] Raj Chetty, John N Friedman, and Jonah E Rockoff. 2014. Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American economic review* 104, 9 (2014), 2633–79.
- [13] Dominic Coey and Tom Cunningham. 2019. Improving treatment effect estimators through experiment splitting. In *The World Wide Web Conference*. 285–295.
- [14] Arthur Cohen and Harold B. Sackrowitz. 1989. Two stage conditionally unbiased estimators of the selected mean. *Statistics & Probability Letters* 8, 3 (1989), 273–278. doi:10.1016/0167-7152(89)90133-8
- [15] Ram C. Dahiya. 1974. Estimation of the Mean of the Selected Population. *J. Amer. Statist. Assoc.* 69, 345 (1974), 226–230. doi:10.1080/01621459.1974.10480159
- [16] Alex Deng. 2015. Objective bayesian two sample hypothesis testing for online controlled experiments. In *Proceedings of the 24th International Conference on World Wide Web*. 923–928.
- [17] Alex Deng, Yicheng Li, Jiannan Lu, and Vivek Ramamurthy. 2021. On Post-Selection Inference in A/B Testing. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (Virtual Event, Singapore) (KDD '21)*. Association for Computing Machinery, New York, NY, USA, 2743–2752. doi:10.1145/3447548.3467129
- [18] Justin B Dimick, Douglas O Staiger, and John D Birkmeyer. 2010. Ranking hospitals on surgical mortality: the importance of reliability adjustment. *Health services research* 45, 6p1 (2010), 1614–1629.
- [19] Rick Durrett. 2009. *Probability: Theory and examples*. Cambridge University Press.
- [20] Bradley Efron. 2011. Tweedie's formula and selection bias. *J. Amer. Statist. Assoc.* 106, 496 (2011), 1602–1614.
- [21] Bradley Efron. 2014. Two modeling strategies for empirical Bayes estimation. *Statistical science: a review journal of the Institute of Mathematical Statistics* 29, 2 (2014), 285.
- [22] Bradley Efron and Carl Morris. 1973. Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Amer. Statist. Assoc.* 68, 341 (1973), 117–130.
- [23] Bradley Efron and Carl Morris. 1977. Stein's paradox in statistics. *Scientific American* 236, 5 (1977), 119–127.
- [24] Bradley Efron and Robert Tibshirani. 2002. Empirical Bayes methods and false discovery rates for microarrays. *Genetic epidemiology* 23, 1 (2002), 70–86.
- [25] Jianqing Fan. 1991. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics* (1991), 1257–1272.
- [26] William Fithian, Dennis Sun, and Jonathan Taylor. 2014. Optimal inference after model selection. *arXiv preprint* (2014).
- [27] Michael Gilraine, Jiaying Gu, and Robert McMillan. 2020. *A new method for estimating teacher value-added*. Technical Report. National Bureau of Economic Research.
- [28] David Goldberg and James E Johndrow. 2017. A decision theoretic approach to A/B testing. *arXiv preprint* (2017).
- [29] Harvey Goldstein and David J Spiegelhalter. 1996. League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 159, 3 (1996), 385–409.
- [30] Jiaying Gu and Roger Koenker. 2020. Invidious comparisons: Ranking and selection as compound decisions. *arXiv preprint* (2020).
- [31] F Richard Guo, James McQueen, and Thomas S Richardson. 2020. Empirical Bayes for Large-scale Randomized Experiments: a Spectral Approach. *arXiv preprint* (2020).
- [32] Xinzhou Guo and Xuming He. 2021. Inference on selected subgroups in clinical trials. *J. Amer. Statist. Assoc.* 116, 535 (2021), 1498–1506.
- [33] Shanti S Gupta and Jianjun Li. 2005. On empirical Bayes procedures for selecting good populations in a positive exponential family. *Journal of Statistical planning and Inference* 129, 1-2 (2005), 3–18.
- [34] Shanti S Gupta and Subramanian Panchapakesan. 2002. *Multiple decision procedures: theory and methodology of selecting and ranking populations*. SIAM.
- [35] James F Hannan and JR Van Ryzin. 1965. Rate of convergence in the compound decision problem for two completely specified distributions. *The Annals of Mathematical Statistics* (1965), 1743–1752.
- [36] Douglas N Harris and Tim R Sass. 2014. Skills, productivity and the evaluation of teacher performance. *Economics of Education Review* 40 (2014), 183–204.
- [37] Peter Hull. 2018. Estimating hospital quality with quasi-experimental data. Available at SSRN 3118358 (2018).
- [38] Kenneth Hung and William Fithian. 2019. Rank verification for exponential families. *The Annals of Statistics* 47, 2 (2019), 758 – 782. doi:10.1214/17-AOS1634
- [39] Brian A Jacob and Lars Lefgren. 2008. Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of labor Economics* 26, 1 (2008), 101–136.
- [40] Thomas J Kane, Jonah E Rockoff, and Douglas O Staiger. 2008. What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education review* 27, 6 (2008), 615–631.
- [41] Robert W Keener. 2010. *Theoretical statistics: Topics for a core course*. Springer.
- [42] Jack Kiefer and Jacob Wolfowitz. 1956. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics* (1956), 887–906.
- [43] Patrick Kline and Christopher Walters. 2021. Reasonable Doubt: Experimental Detection of Job-Level Employment Discrimination. *Econometrica* 89, 2 (2021), 765–792.
- [44] Roger Koenker and Ivan Mizera. 2014. Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Amer. Statist. Assoc.* 109, 506 (2014), 674–685.
- [45] Rongheng Lin, Thomas A Louis, Susan M Paddock, and Greg Ridgeway. 2006. Loss function based ranking in two-stage, hierarchical models. *Bayesian Analysis* 1, 4 (2006), 915.
- [46] JR Lockwood, Thomas A Louis, and Daniel F McCaffrey. 2002. Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of educational and behavioral statistics* 27, 3 (2002), 255–270.
- [47] Roger J Marshall. 1991. Mapping disease and mortality rates using empirical Bayes estimators. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 40, 2 (1991), 283–294.
- [48] J. Nathan Matias, Kevin Munger, Marianne Aubin Le Quere, and Charles Ebersole. 2021. The Upworthy Research Archive, a time series of 32,487 experiments in U.S. media. *Scientific Data* 8, 1 (2021), 195. doi:10.1038/s41597-021-00934-7
- [49] M Mogstad, J Romano, A Shaikh, and D Wilhelm. 2022. Comment on "Invidious Comparisons: Ranking and Selection as Compound Decisions". *Econometrica* (2022).
- [50] Magne Mogstad, Joseph P Romano, Azeem M Shaikh, and Daniel Wilhelm. 2024. Inference for ranks with applications to mobility across neighbourhoods and academic achievement across countries. *Review of Economic Studies* 91, 1 (2024), 476–518.
- [51] Yury Polyanskiy and Yihong Wu. 2021. Sharp regret bounds for empirical Bayes and compound decision problems. *arXiv preprint* (2021).
- [52] Herbert Robbins. 1951. Asymptotically Subminimax Solutions of Compound Statistical Decision Problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability: Volume 2*. University of California Press, 131–149.
- [53] Herbert Robbins. 1956. An Empirical Bayes Approach to Statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. University of California Press, 157–163.
- [54] Herbert Robbins. 1964. The empirical Bayes approach to statistical decision problems. *The Annals of Mathematical Statistics* 35, 1 (1964), 1–20.
- [55] Xuedong Shang, Rianne de Heide, Pierre Menard, Emilie Kaufmann, and Michal Valko. 2020. Fixed-confidence guarantees for Bayesian best-arm identification. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 108)*, Silvia Chiappa and Roberto Calandra (Eds.). PMLR, 1823–1832.
- [56] Gordon K Smyth. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology* 3, 1 (2004).
- [57] Jake A Soloff, Adityanand Guntuboyina, and Bodhisattva Sen. 2024. Multivariate, heteroscedastic empirical Bayes via nonparametric maximum likelihood. *Journal of the Royal Statistical Society Series B: Statistical Methodology* (05 2024), qkae040. doi:10.1093/jrsssb/qkae040
- [58] Matthew Stephens. 2016. False discovery rates: a new deal. *Biostatistics* 18, 2 (10 2016), 275–294. doi:10.1093/biostatistics/kxw041
- [59] Neal Thomas, Nicholas T Longford, and John E Rolph. 1994. Empirical Bayes methods for estimating hospital-specific mortality rates. *Statistics in medicine* 13, 9 (1994), 889–903.
- [60] Aad W van der Vaart. 2000. *Asymptotic statistics*. Vol. 3. Cambridge university press.
- [61] John Van Ryzin and Vyaghreshwarudu Susarla. 1977. On the empirical Bayes approach to multiple decision problems. *The Annals of Statistics* 5, 1 (1977), 172–181.
- [62] Asaf Weinstein. 2021. On Permutation Invariant Problems in Large-Scale Inference. *arXiv preprint* (2021).
- [63] Asaf Weinstein, Zhuang Ma, Lawrence D. Brown, and Cun-Hui Zhang. 2018. Group-Linear Empirical Bayes Estimates for a Heteroscedastic Normal Mean. *J. Amer. Statist. Assoc.* 113, 522 (2018), 698–710. doi:10.1080/01621459.2017.1280406
- [64] Jason Willwerscheid and Matthew Stephens. 2021. ebnm: An R Package for Solving the Empirical Bayes Normal Means Problem Using a Variety of Prior Families. *arXiv preprint* (2021).
- [65] Peng Yu, Spencer S Ericksen, Anthony Gitter, and Michael A Newton. 2020. Bayes Optimal Informer Sets for Early-Stage Drug Discovery. *arXiv preprint* (2020).

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

Acknowledgments

We thank Eytan Bakshy, Kevin Chen, Matt Goldman, Daniel Jiang, Jelena Markovic, Sepehr Akhavan Masouleh, Jingang Miao, Adam Obeng, Alex Peysakovich, Okke Schrijvers, Yanyi Song, Daniel Ting and Mark Tygert for their comments and suggestions.

A Supporting Proofs

Lemma 5. *For any exponential family $G(\eta)$ with finite variance, the mapping $\eta \mapsto G(\eta)$ is locally Lipschitz with respect to the L_1 -norm (or equivalently any L_p -norm) in the domain and the 1-Wasserstein distance in the codomain.*

PROOF. Suppose the exponential family is given by

$$\exp(\eta' T(x) - A(\eta)) h(x) d\mu(x)$$

with X as the variable and $T = T(X)$ as the canonical statistic. We wish to show that $|\int f(x) dG(\eta_1) - \int f(x) dG(\eta_0)| / \|\eta_1 - \eta_0\|_1$ is bounded for $\eta_0 \neq \eta_1$, locally in η and uniformly over all functions f with Lipschitz constant 1. By mean value theorem,

$$\begin{aligned} & \left| \int f(x) dG(\eta_1) - \int f(x) dG(\eta_0) \right| \\ &= \left| (\eta_1 - \eta_0)' \nabla_\eta \int f(x) dG(\eta) \Big|_{\eta=\eta'} \right| \\ &\leq \|\eta_1 - \eta_0\|_1 \left\| \nabla_\eta \int f(x) dG(\eta) \Big|_{\eta=\eta'} \right\|_\infty. \end{aligned}$$

for some η' that is also local. By Keener [41, Theorem 2.4],

$$\begin{aligned} \nabla_\eta \int f(x) dG(\eta) &= \int f(x) (T(x) - A'(\eta)) dG(\eta) \\ &= \text{cov}_\eta(f(X), T(X)). \end{aligned}$$

For i -th component of the covariance vector, we have

$$\begin{aligned} |\text{cov}_\eta(f(X), T(X))_i| &= |\text{cov}_\eta(f(X), T_i(X))| \\ &\leq \sqrt{\text{var}_\eta f(X) \text{var}_\eta T_i(X)}. \end{aligned}$$

$\text{var}_\eta T_i(X)$ is given by $A''(\eta)_{ii}$, which is continuous by Brown [5, Theorem 2.2] and thus locally bounded in η . For $\text{var}_\eta f(X)$, suppose X' is an i.i.d. copy of X , then

$$\begin{aligned} \text{var}_\eta f(X) &= \frac{1}{2} (\text{var}_\eta f(X) + \text{var}_\eta f(X')) \\ &= \frac{1}{2} \text{var}_\eta [f(X) - \mathbb{E}_\eta f(X) + \mathbb{E}_\eta f(X') - f(X')] \\ &= \text{var}_\eta [f(X) - f(X')] \\ &\leq \text{var}_\eta (X - X') \\ &\leq \text{var}_\eta X, \end{aligned}$$

which is also continuous and thus locally bounded in η . \square

Lemma 6. *Under Assumption 2, the posterior mean function $f_{G,\sigma}$ is strictly increasing and differentiable in X , and thus admits an inverse $f_{G,\sigma}^{-1}$ over its image.*

PROOF. From Efron [20], under $\mu \sim G$ and $X | \mu \sim \mathcal{N}(\mu, \sigma^2)$, we have $\nabla_x f_{G,\sigma}(x) = \sigma^{-2} \text{var}(\mu | X = x) > 0$. \square

Lemma 7. *With Assumption 2, the c.d.f. P of θ_i is continuously differentiable with positive derivative, or equivalently, θ_i has positive continuous density.*

PROOF. The characteristic function of $X | \sigma$ is given by $\varphi_G(t) \exp(-\sigma^2 t^2/2)$, where φ_G is the characteristic function of G . Since $|\varphi_G(t) \exp(-\sigma^2 t^2/2)|$ is bounded by $\exp(-\sigma^2 t^2/2)$ which is integrable, $X | \sigma$ has bounded continuous density (Durrett [19, Theorem 3.3.14]). In fact the density is given by

$$p(X | \sigma) = \frac{1}{2\pi} \int e^{-itX} \varphi_G(t) \exp(-\sigma^2 t^2/2) dt.$$

Since σ is bounded away from 0, dominated convergence theorem implies joint continuity of the density above in (σ, X) . In fact, by dominated convergence theorem and the fact that $t^k \exp(-\sigma^2 t^2/2)$ is integrable for all integer $k \geq 0$, we can see that all higher derivatives of the density with respect to X are continuous in (σ, X) .

Consider the mapping $X \mapsto \theta = f_{G_0,\sigma}(X)$. By Lemma 6 it has a strictly positive derivative. Furthermore, since the derivative can be written in terms of the derivatives of $p(X | \sigma)$ (Efron [20]), it is also continuous in (σ, X) . In other words, the density $p(\theta | \sigma)$ is continuous in (σ, X) .

Assumption 2 assumes the support of σ is compact, so the density $p(\theta | \sigma)$ is naturally pointwise equicontinuous when viewed as a family of functions indexed by σ . Now for any θ and any $\varepsilon > 0$, we can select $\delta > 0$ such that for all σ and all θ' with $|\theta' - \theta| < \delta$, we have $|p(\theta' | \sigma) - p(\theta | \sigma)| < \varepsilon$ and so

$$\begin{aligned} & \left| \int p(\theta' | \sigma) dH_0 - \int p(\theta | \sigma) dH_0 \right| \\ &\leq \int |p(\theta' | \sigma) - p(\theta | \sigma)| dH_0 < \varepsilon, \end{aligned}$$

and the marginal density of θ is continuous. \square

Lemma 8. *Let A be a metric space. Suppose $f(a, x)$ as a function from $A \times \mathbb{R}$ to \mathbb{R} is continuous and has an inverse with respect to x , i.e. for all a there exists $f_a^{-1}(\cdot)$ such that $f_a^{-1} \circ f(a, \cdot) = \text{id}_{\mathbb{R}}$. Then $(a, y) \mapsto f_a^{-1}(y)$ is also continuous.*

PROOF. Let $(a_n, y_n) \rightarrow (a^*, y^*)$. It suffices to show that

$$f_{a_n}^{-1}(y_n) \rightarrow f_{a^*}^{-1}(y^*).$$

We first show that the sequence $x_n = f_{a_n}^{-1}(y_n)$ is bounded. The sequence y_n is bounded, so it is contained in some interval $(c + \varepsilon, d - \varepsilon)$ for some fixed $\varepsilon > 0$ and c, d . By continuity of f , for a sufficiently close to a^* , we have $f(a, f_{a^*}^{-1}(c))$ must be within ε of $f(a^*, f_{a^*}^{-1}(c)) = c$. Similarly, $f(a, f_{a^*}^{-1}(d))$ can be within ε of d . Since $f_{a^*}^{-1}$ is the inverse of a continuous function, it is monotonic. For a sufficiently close to a^* ,

$$\{y_n\} \text{ is bounded by } c + \varepsilon \text{ and } d - \varepsilon$$

$$\Leftrightarrow \{f(a, f_{a_n}^{-1}(y_n))\} \text{ is bounded by } c + \varepsilon \text{ and } d - \varepsilon$$

$$\Rightarrow \{f(a, f_{a_n}^{-1}(y_n))\} \text{ is bounded by } f(a, f_{a^*}^{-1}(c)) \text{ and } f(a, f_{a^*}^{-1}(d))$$

$$\Leftrightarrow \{f_{a_n}^{-1}(y_n)\} \text{ is bounded by } f_{a^*}^{-1}(c) \text{ and } f_{a^*}^{-1}(d)$$

So for sufficiently large n , a_n is sufficiently close to a^* , and $f_{a_n}^{-1}(y_n)$ is bounded.

Since the sequence $x_n = f_{a_n}^{-1}(y_n)$ is bounded, it must have a convergent subsequence. Consider any of such convergent subsequence indexed by n_k . We have

$$\begin{aligned} f(a^*, \lim_{k \rightarrow \infty} x_{n_k}) &= f(\lim_{k \rightarrow \infty} a_{n_k}, x_{n_k}) \\ &= \lim_{k \rightarrow \infty} f(a_{n_k}, x_{n_k}) \\ &= \lim_{k \rightarrow \infty} f(a_{n_k}, f_{a_n}^{-1}(y_n)) \\ &= y^* \\ &= f(a^*, f_{a^*}^{-1}(y^*)), \end{aligned}$$

and thus $\lim_{k \rightarrow \infty} x_{n_k} = f_{a^*}^{-1}(y^*)$. Since $f_{a_n}^{-1}(y_n)$ is bounded and any of its convergent subsequence converges to the same limit $f_{a^*}^{-1}(y^*)$, it also converges to the same limit, completing the proof of continuity. \square

Lemma 9. *If for any non-decreasing divergent sequence of real numbers $(a_n)_{n \in \mathbb{N}}$ the sequence of random variables $(X_n)_{n \in \mathbb{N}}$ is $O_p(a_n)$, then it is also $O_p(1)$.*

PROOF. Suppose $X_n \neq O_p(1)$. Then there exists $\varepsilon > 0$ such that for all $M > 0$, there are infinitely many n such that

$$\mathbb{P}(|X_n| > M) \geq \varepsilon. \quad (20)$$

Take n_1 to be the smallest n satisfying (20) with $M = 1$. For $i > 1$, take n_i to be the smallest $n > n_{i-1}$ satisfying (20) with $M = i^2$. Specifically, $(n_i)_{i \in \mathbb{N}}$ is a strictly increasing sequence such that

$$\mathbb{P}(|X_{n_i}| > i^2) \geq \varepsilon \quad \text{for all } i.$$

Now we are ready to set up a sequence that grows sufficiently slowly to cause a contradiction. For any n , take $b_n = i$ where $n \in [n_i, n_{i+1})$. Since $(n_i)_{i \in \mathbb{N}}$ is strictly increasing and only takes values in integers, $(b_n)_{n \in \mathbb{N}}$ is a non-decreasing sequence with $\lim_{n \rightarrow \infty} b_n = \infty$. So $X_n = O_p(b_n)$ and there exists M', N' such that

$$\mathbb{P}(|X_n|/b_n > M') < \varepsilon \quad \text{for all } n > N'.$$

So for sufficiently large $i > M'$,

$$\varepsilon > \mathbb{P}(|X_{n_i}|/b_{n_i} > M') = \mathbb{P}(|X_{n_i}| > M'i) \geq \mathbb{P}(|X_{n_i}| > i^2) \geq \varepsilon,$$

leading to a contradiction. \square

B Simulation details

Each experiment in the Upworthy Research Archive dataset involves two or more treatments corresponding to various combinations of headlines and image “packages” associated with an article.

The number of impressions and clicks are recorded for each package. The metric of interest is the click-through rate, defined as the ratio of clicks to impressions. We filter out article-package pairs with fewer than 1000 impressions or 100 clicks, to ensure normality approximations are reasonable. For the 4677 articles with at least two remaining packages, we arbitrarily consider the one with the most impressions to be the control group and the one with the second-most to be the treatment group, omitting any other packages for that article in the data. From these data, we compute the effect size estimate and the standard error for each experiment. The top- m selection problem is hence selecting the subset of articles, subject to a constraint on the number of articles that can be treated.

For the prior, we applied EB-NN, EB-NSM and EB-NPMLE to the real data. Figure 6 shows the density of the unshrunk treatment effects, as well as the observation densities implied by three estimated prior distributions corresponding to three different prior families. EB-NN is clearly misspecified and has thinner tails than the observations, indicating that the distribution of prior effects is not well approximated by a normal distribution. Both EB-NSM and EB-NPMLE result in close fits to the observed data and more realistic tail behavior. As a result, we base our simulation on the more parsimonious model of the two, EB-NSM.

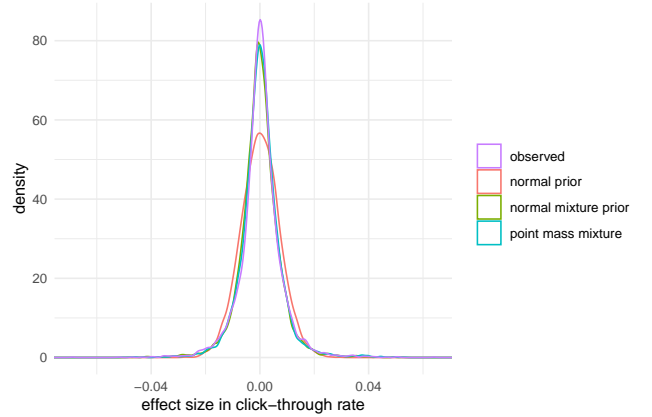


Figure 6: The density of the observed X , compared to the densities of observed X as generated by an estimated normal-normal model (EB-NN), an estimated normal scale mixture model (EB-NSM) and a model estimated by NPMLE (EB-NPMLE).