Empirical Bayes When Estimation Precision Predicts Parameters

Jiafeng Chen
Department of Economics, Harvard University
Harvard Business School
jiafengchen@g.harvard.edu

Job Market Paper (latest)

Abstract. Empirical Bayes shrinkage methods usually maintain a *prior independence* assumption: The unknown parameters of interest are independent from the known standard errors of the estimates. This assumption is often theoretically questionable and empirically rejected. For one, the sample sizes associated with each estimate may select on or may influence the underlying parameters of interest, thereby making standard errors predictive of the unknown parameters. This paper instead models the conditional distribution of the parameter given the standard errors as a flexibly parametrized family of distributions, leading to a family of methods that we call close. This paper establishes that (i) close is rate-optimal for squared error Bayes regret, (ii) squared error regret control is sufficient for an important class of economic decision problems, and (iii) close is worst-case robust when our assumption on the conditional distribution is misspecified. Empirically, using close leads to sizable gains for selecting high-mobility Census tracts targeting a variety of economic mobility measures. Census tracts selected by close are substantially more mobile on average than those selected by the standard shrinkage method. This additional improvement is often multiple times the improvement of the standard shrinkage method over selection without shrinkage.

JEL codes. C10, C11, C44
Keywords. Empirical Bayes, g-modeling, regret, heteroskedasticity, nonparametric maximum likelihood, Opportunity Atlas, Creating Moves to Opportunity

---

Date: January 11, 2024. This paper is previously titled "Gaussian Heteroskedastic Empirical Bayes without Independence." I thank my advisors, Isaiah Andrews, Elie Tamer, Jesse Shapiro, and Edward Glaeser, for their guidance and generous support. I thank Harvey Barnhard, Raj Chetty, Dominic Coey, Aureo de Paula, Jiaying Gu, Nathaniel Hendren, Keisuke Hirano, Peter Hull, Kenneth Hung, Lawrence Katz, Patrick Kline, Scott Duke Kominers, Soonwoo Kwon, Lihua Lei, Andrew Lo, Michael Luca, Anna Mikusheva, Joris Pinkse, Mikkel Plagborg-Møller, Suproteem Sarkar, Ashesh Rambachan, David Ritzwoller, Brad Ross, Jonathan Roth, Neil Shephard, Rahul Singh, Asher Spector, Winnie van Dijk, Christopher Walker, Davide Viviano, and workshop and seminar participants at Brown, Harvard, Penn State, Philadelphia Fed, Rutgers, and Stanford. I am responsible for any and all errors. An R implementation of close is found at https://github.com/jiafengkevinchen/close.

1. Introduction

Applied economists often use empirical Bayes methods to shrink noisy parameter estimates, in hopes of accounting for the imprecision in the estimates and improving subsequent policy decisions.[1] The textbook empirical Bayes method assumes *prior independence*—that the precisions of the noisy estimates do not predict the underlying unknown parameters. However, prior independence is economically questionable and empirically rejected in many contexts. This is frequently because sample sizes associated with the estimates either *select on* or *affect* the underlying parameters, rendering the resulting standard errors highly predictive of the parameters.[2] Inappropriately imposing prior independence can harm empirical Bayes decisions, possibly even making them underperform decisions without using shrinkage. Motivated by these concerns, this paper introduces empirical Bayes methods that relax prior independence.

To be concrete, our primary empirical example (Bergman et al., 2023) computes empirical Bayes posterior means for economic mobility estimates of low-income children[3] published in the Opportunity Atlas (Chetty et al., 2020). Here, prior independence assumes that the

[1]Empirical Bayes methods are appropriate whenever many parameters for heterogeneous populations are estimated in tandem. For instance, value-added modeling, where the parameters are latent qualities for different service providers (e.g. teachers, schools, colleges, insurance providers, etc.), is a common thread in several literatures (Angrist et al., 2017; Mountjoy and Hickman, 2021; Chandra et al., 2016; Doyle et al., 2017; Hull, 2018; Einav et al., 2022; Abaluck et al., 2021; Dimick et al., 2010). Our application (Bergman et al., 2023) is in a literature on place-based effects, where the unknown parameters are latent features of places (Chyn and Katz, 2021; Finkelstein et al., 2021; Chetty et al., 2020; Chetty and Hendren, 2018; Diamond and Moretti, 2021; Baum-Snow and Han, 2019). Empirical Bayes methods are also applicable in studies of discrimination (Kline et al., 2022, 2023; Rambachan, 2021; Egan et al., 2022; Arnold et al., 2022; Montiel Olea et al., 2021), meta-analysis (Azevedo et al., 2020; Meager, 2022; Andrews and Kasy, 2019; Elliott et al., 2022; Wernerfelt et al., 2022; DellaVigna and Linos, 2022; Abadie et al., 2023), and correlated random effects in panel data (Chamberlain, 1984; Arellano and Bonhomme, 2009; Bonhomme et al., 2020; Bonhomme and Manresa, 2015; Liu et al., 2020; Giacomini et al., 2023).
In terms of policy decisions driven by empirical Bayes posterior means, Gilraine et al. (2020) report that by the end of 2017, 39 states require that teacher value-added measures—typically, empirical Bayes posterior means of teacher performance—be incorporated into the teacher evaluation process.

[2]To see this, take value-added modeling as an example. The precision of value-added estimates is usually a function of the number of customers associated with a service provider (e.g. number of students for a teacher). It is possible that customers select into higher quality providers. It is also possible that congestion effects render more popular service providers worse. These channels predict that the sample sizes for a provider are associated with latent value-added, and the direction of association depends on the interplay of the selection and congestion effects. Appendix A.5 outlines a formal discrete choice model to illustrate these effects. Potential failure of prior independence is noted by, among others, Bruhn et al. (2022), Kline et al. (2023), George et al. (2017), and Mehta (2019).

[3]Throughout this paper, measures of economic mobility are defined as certain average outcomes of children from low-income households. There are various definitions of economic mobility provided by Chetty et al. (2020), discussed later in the paper. They are all measures of economic outcomes for children from low-income households (households at the 25^{th} percentile of the national income distribution). One example is the probability that a Black person have incomes in the top 20 percentiles, whose parents have household incomes at the 25^{th} percentile. As another example, Bergman et al. (2023) measure economic mobility as the mean income rank of children growing up in households at the 25^{th} income percentile.

standard errors of these noisy mobility estimates do not predict true economic mobility. However, more upwardly mobile Census tracts tend to have fewer low-income children and hence noisier estimates of economic mobility. Consequently, the standard errors of the estimates and true economic mobility are positively correlated, violating prior independence.

Bergman et al. (2023) use empirical Bayes posterior means to select high-mobility Census tracts, choosing those with high estimated posterior means. Using a validation procedure that we develop, for a few measures of economic mobility where prior independence is severely violated, we find that screening on conventional empirical Bayes posterior means selects *less* economically mobile tracts, on average, than screening on the unshrunk estimates.[4] In contrast, screening on empirical Bayes posterior means computed by our method selects substantially more mobile tracts.

To describe our method, let Y_i be some noisy estimates for some parameters θ_i, with standard errors σ_i, over heterogeneous populations $i = 1, \ldots, n$. In our empirical application, (Y_i, σ_i) are published in the Opportunity Atlas for each Census tract i and are designed to measure true economic mobility θ_i. Motivated by the central limit theorem applied to the underlying micro-data, Y_i is approximately Gaussian:

$$Y_i \mid \theta_i, \sigma_i \sim \mathcal{N}(\theta_i, \sigma_i^2) \quad i = 1, \ldots, n. \tag{1.1}$$

If we knew the distribution of (θ_i, σ_i), then we can do no better than *oracle Bayes* decisions, based on the posterior distribution $\theta_i \mid \sigma_i, Y_i$. Empirical Bayes emulates such optimal decisions by estimating the oracle prior distribution of (θ_i, σ_i). Prior independence $\theta_i \perp\!\!\!\perp \sigma_i$ simplifies this estimation problem. However, empirical Bayes methods based on this assumption can have poor performance when it fails to hold.

We relax prior independence by modeling the prior distribution $\theta_i \mid \sigma_i$ flexibly, detailed in Section 2. We model $\theta_i \mid \sigma_i$ as a conditional location-scale family, controlled by σ_i-dependent location and scale hyperparameters and a σ_i-independent shape hyperparameter. Under this assumption, different values of the standard errors σ_i translate, compress, or dilate the distribution of the parameters $\theta_i \mid \sigma_i$, but the underlying shape of $\theta_i \mid \sigma_i$ does not vary. This model subsumes prior independence as the special case where the unknown location and scale parameters are constant functions of σ_i.

This conditional location-scale assumption leads naturally to a family of empirical Bayes methods that we call CLOSE. Since the unknown prior distribution $\theta_i \mid \sigma_i$ is fully described by its location, scale, and shape hyperparameters, CLOSE estimates these parameters flexibly

[4]Fortunately, for the measure of economic mobility (mean income rank pooling over all demographic groups whose parents are at the 25^{th} percentile of household income) used in Bergman et al. (2023), the violation of prior independence is sufficiently mild, so that screening on these empirical Bayes posterior means still outperforms screening on the raw estimates.

and plugs the estimated parameters into downstream decision rules. Among different estimation strategies for the hyperparameters, our preferred specification of CLOSE uses nonparametric maximum likelihood (NPMLE, Kiefer and Wolfowitz, 1956; Koenker and Mizera, 2014) to estimate the unknown shape of the prior distribution $\theta_i \mid \sigma_i$. We find that CLOSE-NPMLE inherits the favorable computational and theoretical properties of NPMLE documented in the literature (Soloff et al., 2021; Jiang, 2020; Polyanskiy and Wu, 2020).

Section 3 provides three statistical guarantees for CLOSE-NPMLE. First and foremost, CLOSE-NPMLE emulates the oracle as well as possible, at least in terms of squared error loss. Specifically, Corollary 1 and Theorem 2 establish that CLOSE-NPMLE is minimax rate-optimal—up to logarithmic factors and under the conditional location-scale assumptions—for *Bayes regret in squared error*, a standard performance metric (Jiang and Zhang, 2009). Bayes regret is the performance gap between CLOSE-NPMLE and oracle Bayes decisions made with knowledge of the distribution of (θ_i, σ_i).

Second, our guarantee for squared error regret also controls the Bayes regret for two ranking-related decision problems, including the problem of selecting high-mobility tracts encountered by Bergman et al. (2023). Theorem 3 shows that the Bayes regret in squared error dominates the Bayes regret for these decision problems. Thus, these ranking-related problems are easier than squared error estimation, and our squared error regret result implies upper bounds for the regrets of these problems.

Third, to assess robustness of CLOSE to the location-scale modeling assumption, Theorem 4 establishes that CLOSE-NPMLE is worst-case robust. Without imposing the location-scale assumptions, for a population version of CLOSE-NPMLE, we show that its worst-case mean-squared error is a bounded multiple of that of the minimax procedure. Since the minimax procedure optimizes its worst-case risk, this result shows that CLOSE-NPMLE does not perform exceedingly poorly even when the location-scale model is misspecified.

Since practitioners may want to assess how and whether CLOSE-NPMLE provides improvements in specific applications, Section 4.3 produces an out-of-sample validation procedure by extending the *coupled bootstrap* in Oliveira et al. (2021). If one had access to the micro-data, one could split the data into training and testing samples, use one to compute decisions, and use the other to evaluate them. Our validation procedure emulates this sample-splitting without needing access to the underlying micro-data. It provides unbiased loss estimates for any decision rules. In particular, this procedure allows practitioners to evaluate whether CLOSE provides improvements for their setting by comparing loss estimates for CLOSE and those for the standard shrinkage procedure.

To illustrate our method, Section 5 applies CLOSE to two empirical exercises, building on Chetty et al. (2020) and Bergman et al. (2023). The first exercise is a calibrated Monte Carlo simulation, in which we have access to the true distribution of (θ_i, σ_i). We find that CLOSE-NPMLE has mean-squared error (MSE) performance close to that of the oracle posterior, uniformly across the 15 measures of economic mobility that we include. For all 15 measures, CLOSE-NPMLE captures over 90% of possible MSE gains relative to no shrinkage, whereas conventional shrinkage captures only 70% on average and as little as 40% for some measures.

The second exercise evaluates the out-of-sample performance of various procedures for an economic policy problem. Bergman et al. (2023) use empirical Bayes procedures to select high-mobility Census tracts in Seattle. We consider a version of their exercise with different mobility measures, scaled up to the largest Commuting Zones in the United States. We find that CLOSE-NPMLE selects more economically mobile tracts than the conventional shrinkage method. These improvements are large relative to two benchmarks. First, they are on median 3.2 times the *value of basic empirical Bayes*—that is, the improvements the standard method delivers over screening on the raw estimates Y_i directly. Therefore, if one finds using the standard empirical Bayes method a worthwhile methodological investment, then the additional gain of using CLOSE is likewise meaningful. Second, for 6 out of 15 measures of mobility, CLOSE even improves over the standard method *by a larger amount* than the *value of data*—that is, the amount by which the standard method improves over selecting Census tracts completely at random. These improvements are substantial, since the value of data is likely economically significant if the mobility estimates are at all useful for the policy problem.

2. Model and proposed method

We observe estimates Y_i and their standard errors σ_i for parameters θ_i, over populations $i \in \{1, \ldots, n\}$. We maintain throughout that the estimates are conditionally Gaussian and independent across i:

$$Y_i \mid \theta_i, \sigma_i^2 \sim \mathcal{N}(\theta_i, \sigma_i^2) \qquad i = 1, \ldots, n. \tag{2.1}$$

The Normality in (2.1) is motivated by the central limit theorem applied to the underlying micro-data that generate the estimates Y_i. That is, let n_i denote the underlying sample size in the micro-data which generate (Y_i, σ_i). Standard large-sample approximation implies

$$\frac{Y_i - \theta_i}{\sigma_i} \xrightarrow{d} \mathcal{N}(0, 1) \tag{2.2}$$

as $n_i \to \infty$.[5]

We also assume that the population parameters (θ_i, σ_i) are sampled from some joint distribution. Throughout this paper, we condition on $\sigma_{1:n} = (\sigma_1, \ldots, \sigma_n)$ and treat them as

[5]Note that, under standard assumptions, the approximation (2.2) holds regardless of whether σ_i is an estimated standard error or its unknown population counterpart. This is because the estimation error in σ_i is typically of order $1/n_i$, which is smaller than that in Y_i, which is of order $1/\sqrt{n_i}$.

fixed. We assume that (θ_i, σ_i) are independently and identically drawn,[6] but the conditional distribution $\theta_i \mid \sigma_i$ may be different across σ_i:

$$\theta_i \mid \sigma_i \overset{\text{i.n.i.d.}}{\sim} G_{(i)}. \tag{2.3}$$

We use $G_{(i)}$ to denote the distribution of $\theta_i \mid \sigma_i$. We use P_0 to denote the distribution of $\theta_{1:n} \mid \sigma_{1:n}$, which is fully described by $(G_{(1)}, \ldots, G_{(n)})$. We refer to P_0 as the *oracle Bayes prior*.

These assumptions imply that the Bayes decision rule with respect to the oracle Bayes prior P_0 is optimal (Lehmann and Casella, 2006). Consider a loss function $L(\boldsymbol{\delta}, \theta_{1:n})$, which evaluates an action $\boldsymbol{\delta}$ at a vector of parameters $\theta_{1:n}$. For instance, in our empirical application, the loss function may measure how well we estimate true mobility $\theta_{1:n}$ or how well we select high mobility Census tracts.[7] At any realization of the data $(Y_{1:n}, \sigma_{1:n})$, the *oracle Bayes decision rule* $\boldsymbol{\delta}^\star$ picks an action that minimizes the posterior expected loss:

$$\boldsymbol{\delta}^\star(Y_{1:n}, \sigma_{1:n}; P_0) \in \arg\min_{\boldsymbol{\delta}} \mathbb{E}_{P_0}[L(\boldsymbol{\delta}, \theta_{1:n}) \mid Y_{1:n}, \sigma_{1:n}]. \tag{2.4}$$

Empirical Bayesians seek to approximate the oracle Bayes rule $\boldsymbol{\delta}^\star$ (Efron, 2014). With an estimate $\hat{P}$ for P_0, it is natural to plug $\hat{P}$ into (2.4):[8]

$$\boldsymbol{\delta}_{\text{EB}}(Y_{1:n}, \sigma_{1:n}; \hat{P}) \in \arg\min_{\boldsymbol{\delta}} \mathbf{E}_{\hat{P}}[L(\boldsymbol{\delta}, \theta_{1:n}) \mid Y_{1:n}, \sigma_{1:n}]. \tag{2.5}$$

Popular empirical Bayes methods impose more structure than (2.3) in order to simplify estimating P_0.[9] The standard parametric empirical Bayes method additionally models $G_{(i)}$ as identical across i and Gaussian: i.e., for all i, $G_{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(m_0, s_0^2)$ (Morris, 1983). Following the recipe (2.5), this approach estimates the prior parameters (m_0, s_0^2). Henceforth, we shall refer to this method as INDEPENDENT-GAUSS. On the other hand, state-of-the-art empirical

[6]Combined with the independence assumption of Y_i across i, we assume that $(\theta_i, \sigma_i, Y_i)$ are independently drawn unconditionally. The independence assumption for the estimates Y_i conditional on (θ_i, σ_i) holds when the underlying micro-data for different estimates Y_i are sampled independently. This assumption does not precisely hold for the Opportunity Atlas, but the correlation between Y_i and Y_j, which arises from individuals who move between tracts, is likely small. Papers imposing this assumption include Mogstad et al. (2020) and Andrews et al. (2023). Moreover, we discuss an interpretation of the procedure when we erroneously assume that Y_i and/or θ_i are independent across i in Appendix A.6.

[7]We formalize the sense of optimality and formalize three decision problems in Section 2.3.

[8]To emphasize the distinction between the true expectation with respect to the data-generating process (2.3) and a posterior mean taken with respect to some possibly estimated measure $\hat{P}$, we shall use $\mathbb{E}$ to refer to the former and $\mathbf{E}$ to refer to the latter. Subscripts typically make the distinction clear as well. Specifically,

$$\mathbf{E}_{\hat{P}}[L(\boldsymbol{\delta}, \theta_{1:n}) \mid Y_{1:n}, \sigma_{1:n}] = \frac{\int L\left(\boldsymbol{\delta}(Y_{1:n}, \sigma_{1:n}), \theta_{1:n}\right) \prod_{i=1}^n \varphi\left(\frac{y_i - \theta_i}{\sigma_i}\right) \hat{P}(d\theta_{1:n} \mid \sigma_{1:n})}{\int \prod_{i=1}^n \varphi\left(\frac{y_i - \theta_i}{\sigma_i}\right) \hat{P}(d\theta_{1:n} \mid \sigma_{1:n})},$$

where $\varphi(\cdot)$ is the probability density function of a standard Gaussian.

[9]The literature on empirical Bayes methods is vast. For theoretical and applied results of particular interest to economists, see the recent lecture by Gu and Walters (2022) and references therein. Efron (2019) and accompanying discussions are excellent introductions to the statistics literature.

Bayes methods (Jiang, 2020; Soloff et al., 2021; Jiang and Zhang, 2009; Koenker and Gu, 2019; Gilraine et al., 2020) assume that the marginal distributions are equal to some common, unknown distribution $G_{(0)}$, not necessarily Gaussian: i.e., for all i, $G_{(i)} \overset{\text{i.i.d.}}{\sim} G_{(0)}$. They estimate $G_{(0)}$ with nonparametric maximum likelihood and form decision rules according to (2.5). We refer to this method as INDEPENDENT-NPMLE. The "INDEPENDENT" here emphasizes that these methods assume *prior independence*: $\theta_i \perp\!\!\!\perp \sigma_i$ under the prior P_0.

We relax prior independence by instead modeling $\theta_i \mid \sigma_i$ as a location-scale family,[10] indexed by unknown hyperparameters $(m_0(\cdot), s_0(\cdot), G_0(\cdot))$: Specifically, we assume

$$\mathrm{P}\left(\theta_i \le t \mid \sigma_i\right) = G_0\left(\frac{t - m_0(\sigma_i)}{s_0(\sigma_i)}\right), \tag{2.6}$$

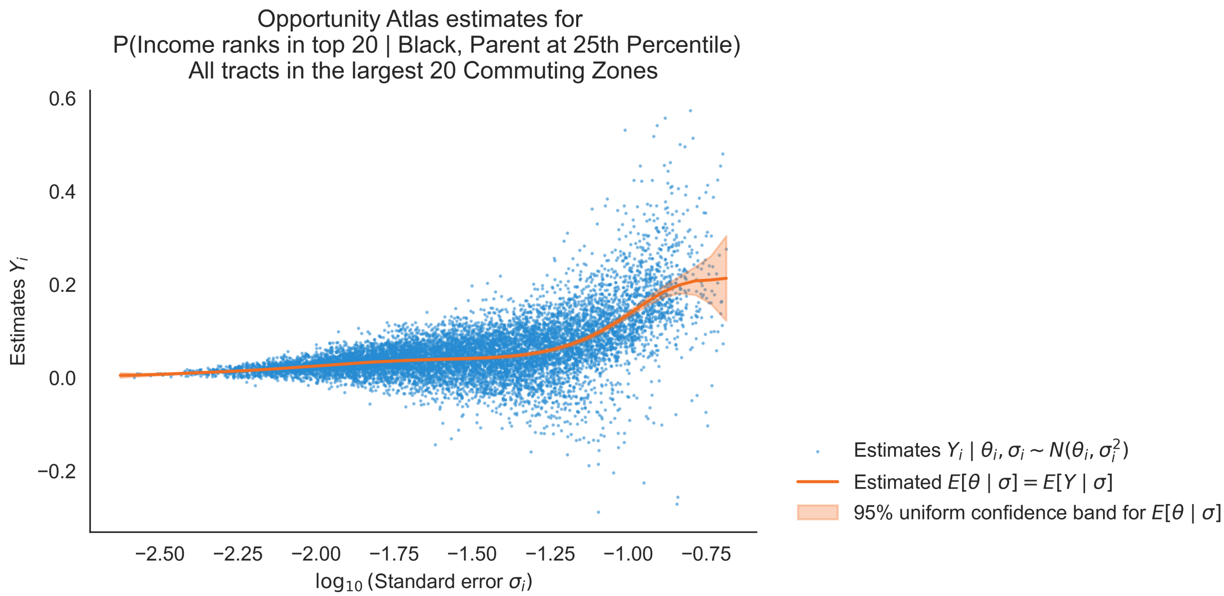
where the distribution G_0 is normalized to have zero mean and unit variance. Under (2.6), different values of σ may translate, compress, or dilate the conditional distribution of $\theta \mid \sigma$ via the location parameter $m_0(\cdot)$ and the scale parameter $s_0(\cdot)$, but the conditional distributions can be normalized to take the same shape $G_0(\cdot)$. Under this model, the oracle prior distribution P_0 is fully described by the hyperparameters $(m_0(\cdot), s_0(\cdot), G_0(\cdot))$. Our method, CLOSE, proposes to estimate P_0 with an estimate $\hat{P}$ derived from estimated hyperparameters $(\hat{m}(\cdot), \hat{s}(\cdot), \hat{G}_n)$. CLOSE then produces empirical Bayes decision rules with respect to the estimated prior $\hat{P}$, following the recipe (2.5).

Before specifying our procedure in detail in Section 2.2, we illustrate with an example where prior independence fails and show what happens to empirical Bayes decision rules that inappropriately impose prior independence.

2.1. **Plausibility of prior independence.** As a running example, let us define economic mobility θ_i as the probability of family income ranking in the top 20 percentiles of the national income distribution, for a Black individual growing up in tract i whose parents are at the 25$^{\text{th}}$ national income percentile. Note that the standard error σ_i for an estimate of θ_i is then related to the implicit sample size—the number of Black households at the 25$^{\text{th}}$ income percentile in tract i.

Prior independence is readily rejected for this measure of economic mobility. Figure 1 plots Y_i against $\log_{10}(\sigma_i)$ and imposes a nonparametric regression estimate of the conditional mean function $m_0(\sigma_i) \equiv \mathbb{E}[\theta_i \mid \sigma_i] = \mathbb{E}[Y_i \mid \sigma_i]$. If θ_i were independent of σ_i, then the true

[10]We explore alternatives to the location-scale model in Appendix A.7. We find that no alternative provides a free-lunch improvement over our assumptions.
More restrictive forms of this assumption also appear in the past and concurrent literature. For instance, Kline et al. (2023) model the dependence as a pure scale model $\theta \mid \sigma \sim s(\sigma) \cdot \tau$ for some $\tau \mid \sigma \overset{\text{i.i.d.}}{\sim} G$ (with additional parametric restrictions on $s(\cdot)$) and George et al. (2017) impose the location scale model (2.6) with $G_0 \sim \mathcal{N}(0, 1)$ (as well as additional parametric restrictions on $s_0(\cdot), m_0(\cdot)$).



Notes. All tracts within the largest 20 Commuting Zones (CZs) are shown. Due to the regression specification in Chetty et al. (2020), point estimates of $\theta_i \in [0, 1]$ do not always lie within $[0, 1]$. The orange line plots nonparametric regression estimates of the conditional mean $\mathbb{E}[Y \mid \sigma] = \mathbb{E}[\theta \mid \sigma] \equiv m_0(\sigma)$, estimated via local linear regression with automatic bandwidth selection implemented in Calonico et al. (2019). The orange shading shows a 95% uniform confidence band, constructed by the max-t confidence set over 50 equally spaced evaluation points. The confidence band excludes any constant function. See Appendix G for details on estimating conditional moments of θ_i given σ_i. □

FIGURE 1. Scatter plot of Y_i against $\log_{10}(\sigma_i)$ in the Opportunity Atlas

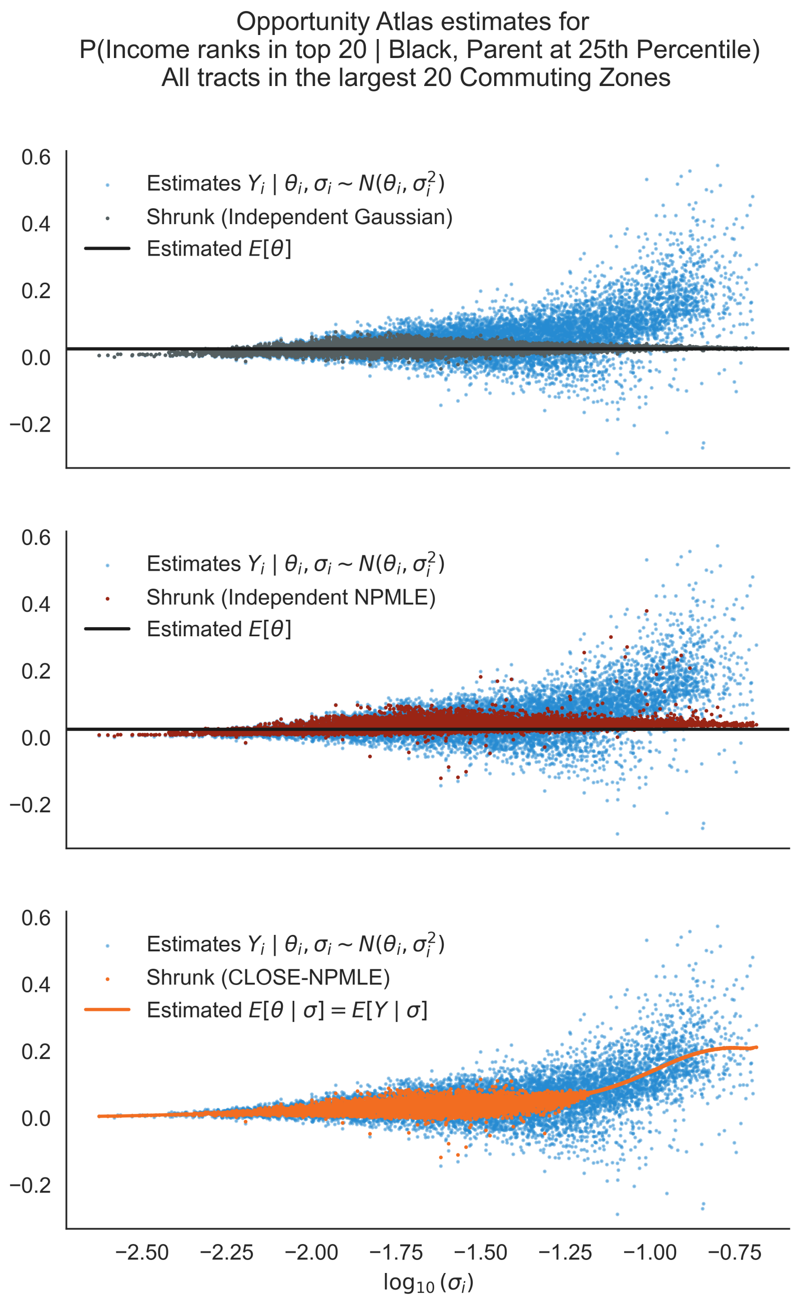
conditional mean function $m_0(\sigma_i)$ should be constant. Figure 1 shows the contrary—tracts with more imprecisely estimated Y_i tend to have higher economic mobility.[11]

This correlation is in part through the following channel. Since θ_i is an average outcome for children from poor Black families, tracts with more poor Black families tend to have more precise estimates of θ_i.[12] However, these tracts also tend to have lower economic mobility θ_i due to the pernicious effects of residential segregation.

What happens if we apply empirical Bayes methods that assume prior independence here? Figure 2 overlays empirical Bayes posterior means on the Y_i-against-$\log \sigma_i$ scatterplot. In the top panel, INDEPENDENT-GAUSS shrinks estimates Y_i towards a common estimated mean $\hat{m}_0$, depicted as the black line. INDEPENDENT-GAUSS shrinks noisier estimates more aggressively.

[11]Moreover, $\log \sigma_i$ remains predictive of Y_i even if we residualize Y_i against a vector of tract-level covariates (Figure B.9).
Prior independence is also readily rejected for the mobility measure used in Bergman et al. (2023), but its violation is not as severe once adjusted for tract-level covariates (see Section 5 and Figure B.8).

[12]Since θ_i is also the mean of a binary outcome, the asymptotic variance of its estimators also depend on mechanically on θ_i.



Notes. The top panel shows posterior mean estimates with INDEPENDENT-GAUSS shrinkage. The middle panel shows the same with INDEPENDENT-NPMLE shrinkage. The bottom panel displays posterior mean estimates from our preferred procedure, CLOSE-NPMLE. In the top panel, the estimates for m_0, s_0^2 are weighted by the precision $1/\sigma_i^2$ (as in Bergman et al., 2023). Under $\theta_i \perp\!\!\!\perp \sigma_i$, this weighting scheme improves efficiency of the (m_0, s_0)-estimates by underweighting noisier Y_i. $\square$

FIGURE 2. Posterior mean estimates under prior independence

When σ_i and θ_i are positively correlated—as is the case here—estimated posterior means under INDEPENDENT-GAUSS systematically undershoot θ_i for populations with imprecise estimates. Similarly, the middle panel of Figure 2 shows that INDEPENDENT-NPMLE suffers from the same undershooting, though less so. In contrast, the bottom panel of Figure 2 previews our preferred procedure, CLOSE-NPMLE, which shrinks towards the conditional mean $\mathbb{E}[\theta_i \mid \sigma_i]$, thus avoiding the undershooting.

This undershooting is particularly problematic if one would like to select high-mobility Census tracts. These high-mobility tracts are exactly those with high imprecision σ_i, owing to the positive correlation between θ_i and σ_i. By shrinking these tracts severely towards the estimated common mean, empirical Bayes under prior independence makes suboptimal selections that may even underperform screening directly based on Y_i.[13]

For a given empirical context, prior independence can always be checked empirically by plotting à la Figure 1. Nevertheless, we discuss the general plausibility of prior independence in the following remark.

Remark 1 (Plausibility of prior independence)**.** To describe the general channels underlying the potential failure of prior independence, let us write (2.2) in a different form

$$\sqrt{n_i}(Y_i - \theta_i) \xrightarrow{d} \mathcal{N}(0, \sigma_{0i}^2) \quad \text{where } \sigma_i \approx \frac{\sigma_{0i}}{\sqrt{n_i}}. \tag{2.7}$$

Expression (2.7) decomposes the (estimated) standard error into the underlying sample size n_i in the micro-data and the asymptotic variance σ_{0i}^2 of the (properly scaled) estimator. Both n_i and σ_{0i} may predict θ_i in a variety of empirical contexts.

Let us start with the implicit sample sizes n_i. It is possible that n_i is in part determined by θ_i, which we loosely term *selection*. In value-added modeling, n_i is the number of observations associated with a provider. It is possible that n_i selects on the latent quality θ_i of that provider. For instance, Chandra et al. (2016) find "higher quality hospitals have higher market shares and grow more over time." If market share and hospital size relate to the underlying sample size n_i (e.g. number of patient observations) for estimating hospital value-added, then this suggests non-independence between θ_i and σ_i (see George et al. (2017) for some empirical evidence). As another example, in meta-analysis, suppose θ_i represents the treatment effect of some intervention i. If researchers power studies based on informative priors for θ_i, then we should observe that interventions with larger conjectured effect sizes have smaller sample sizes n_i.

Another channel driving the correlation between n_i and θ_i can be loosely termed *congestion*, where n_i affects the latent feature θ_i. For our primary application, n_i represents the number of poor and minority households in a Census tract, and θ_i represents underlying economic or social mobility. Places with more poor and minority households experience white

[13]This latter point is similarly made in Mehta (2019), though for different loss functions.

flight and residential segregation (Cutler et al., 1999; Agan and Starr, 2020; Kain, 1968), develop oppressive institutions (Derenoncourt, 2022; Alesina et al., 2001), and provide worse public goods (Laliberté, 2021; Jackson and Mackevicius, 2021; Colmer et al., 2020). These factors contribute to lower economic mobility θ_i. Appendix A.5 contains more examples of violation of prior independence and outlines a model in which selection and congestion effects drive correlation between n_i and θ_i.

There are also channels for the asymptotic variance σ_{0i}^2 to correlate with θ_i. In the context of intergenerational mobility, a parallel literature on the *Great Gatsby curve* (Durlauf et al., 2022) seeks to explain a negative relationship between inequality—which contributes to σ_{0i}^2—and intergenerational mobility. For instance, Becker et al. (2018) posit that parental investment and parental human capital are complements for forming the skills of a child. As a result, parents with higher human capital—and more wealth—invest disproportionately more in their children's education than parents with lower human capital. This process then produces both inequality and low economic mobility. In other words, places that are more unequal (which may result in higher σ_{0i}^2) have lower mobility θ_i. ■

2.2. **Conditional location-scale relaxation of prior independence.** Having argued that (i) prior independence is theoretically suspect and empirically rejected and that (ii) inappropriately imposing it can harm empirical Bayes decision rules, we propose the conditional location-scale model (2.6) as a relaxation.[14] Here, we state the location-scale assumption (2.6) equivalently as the following representation with transformed parameters $\tau_i = \frac{\theta_i - m_0(\sigma_i)}{s_0(\sigma_i)}$:

$$\theta_i = m_0(\sigma_i) + s_0(\sigma_i)\tau_i \qquad \tau_i \mid \sigma_i \overset{\text{i.i.d.}}{\sim} G_0 \qquad \mathbb{E}_{G_0}[\tau_i] = 0 \qquad \mathrm{Var}_{G_0}(\tau_i) = 1. \tag{2.8}$$

To estimate P_0 under (2.8), it suffices to estimate the unknown hyperparameters (m_0, s_0, G_0). Expression (2.8) makes clear that, under the location-scale model, the transformed parameter $\tau_i \sim G_0$ is independent from σ_i. Analogously, let $Z_i = \frac{Y_i - m_0(\sigma_i)}{s_0(\sigma_i)}$ be the transformed estimates and $\nu_i = \frac{\sigma_i}{s_0(\sigma_i)}$ be their standard errors.

Crucially, (Z_i, τ_i, ν_i) obey an analogue of the Gaussian location model (2.1) in which prior independence holds:

$$Z_i \mid \nu_i, \tau_i \sim \mathcal{N}(\tau_i, \nu_i^2), \text{ independently across } i \text{ and } \tau_i \mid \sigma_i \overset{\text{i.i.d.}}{\sim} G_0.$$

Therefore, it is a natural to first transform (Y_i, σ_i) into (Z_i, ν_i) and then use empirical Bayes methods that assume prior independence on these transformed quantities to estimate G_0.

This strategy is still infeasible, since the transformation depends on unknown location and scale parameters $\eta_0 \equiv (m_0, s_0)$. Fortunately, $m_0(\cdot)$ and $s_0(\cdot)$ are readily estimable from the

[14] In the presence of covariates X_i—which do not predict the noise in Y_i, $Y_i \perp\!\!\!\perp X_i \mid \theta_i, \sigma_i$—the assumption (2.6) can be modified to accommodate additional covariates as well. We provide additional discussion of covariates in Appendix A.6.2.

data (Y_i, σ_i), as they only require conditional expectations and variances of Y given σ:

$$m_0(\sigma) = \mathbb{E}[\theta \mid \sigma] = \mathbb{E}[Y \mid \sigma] \quad \text{and} \quad s_0^2(\sigma) = \operatorname{Var}(\theta \mid \sigma) = \mathbb{E}[(Y - m_0(\sigma))^2 \mid \sigma] - \sigma^2. \tag{2.9}$$

Given estimates $\hat{m}$ and $\hat{s}$ of $m_0(\cdot)$ and $s_0(\cdot)$, we then form the estimated transformed data $\hat{Z}_i, \hat{\nu}_i$ as

$$\hat{Z}_i = \frac{Y_i - \hat{m}(\sigma_i)}{\hat{s}(\sigma_i)} \quad \text{and} \quad \hat{\nu}_i = \frac{\sigma_i}{\hat{s}(\sigma_i)}. \tag{2.10}$$

We then apply empirical Bayes methods assuming prior independence on $(\hat{Z}_i, \hat{\nu}_i)$. This leads to a family of empirical Bayes strategies that we refer to as conditional location-scale empirical Bayes, or CLOSE:[15]

CLOSE–STEP 1 Nonparametrically estimate $m_0(\sigma), s_0^2(\sigma)$ according to (2.9).

CLOSE–STEP 2 With the estimates $\hat{\eta} = (\hat{m}, \hat{s})$, transform the data according to (2.10). Apply empirical Bayes methods with prior independence to estimate G_0 with some $\hat{G}_n$ on the transformed data $(\hat{Z}_i, \hat{\nu}_i)$.

CLOSE–STEP 3 Having estimated $(\hat{\eta}, \hat{G}_n)$, which implies an estimate $\hat{P}$ of P_0, we then form empirical Bayes decision rules following (2.5).

This framework produces a family of empirical Bayes strategies, since **CLOSE–STEP 2** can take different forms. To leverage theoretical and computational advances, we will focus on—and recommend—using nonparametric maximum likelihood (NPMLE) to estimate G_0. That is, we maximize the log-likelihood of (an estimated version of) the transformed data Z_i, whose marginal distribution is the convolution $G_0 \star \mathcal{N}(0, \nu_i^2)$:[16]

$$\hat{G}_n \in \operatorname*{arg\,max}_{G \in \mathcal{P}(\mathbb{R})} \frac{1}{n} \sum_{i=1}^{n} \log \int_{-\infty}^{\infty} \varphi\left(\frac{\hat{Z}_i - \tau}{\hat{\nu}_i}\right) \frac{1}{\hat{\nu}_i} G(d\tau). \tag{2.11}$$

When the estimated moments $\hat{m}, \hat{s}$ are constant functions of σ, CLOSE-NPMLE estimates the same prior as INDEPENDENT-NPMLE. In the theoretical literature, under prior independence, INDEPENDENT-NPMLE is state-of-the-art in terms of computational ease and regret

[15] We give a more detailed walkthrough of these steps in Section 4. We also detail a local linear regression estimator in Appendix G for **CLOSE–STEP 1**.

[16] We use $(f \star g)(t) = \int_{-\infty}^{\infty} f(x) g(t-x)\, dx$ to denote convolution and $\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$ to denote the Gaussian probability density function. The maximization is over the set of all probability measures on $\mathbb{R}$, $\mathcal{P}(\mathbb{R})$.

properties.[17] Our subsequent results in Section 3 extend some of these favorable properties to CLOSE-NPMLE under the conditional location-scale model.

A simple alternative, which we call CLOSE-GAUSS and think of as a "lite" version of CLOSE-NPMLE, additionally models the shape G_0 as standard Gaussian. We briefly discuss its properties in the following remark.

Remark 2 (CLOSE-GAUSS)**.** Under $G_0 \sim \mathcal{N}(0,1)$, the oracle Bayes posterior means are simply

$$\theta^*_{i,\mathcal{N}(0,1),\eta_0} = \frac{\sigma_i^2}{s_0^2(\sigma_i)+\sigma_i^2} m_0(\sigma_i) + \frac{s_0^2(\sigma_i)}{s_0^2(\sigma_i)+\sigma_i^2} Y_i. \tag{2.12}$$

Equation (2.12) is the analogue of posterior means estimated by INDEPENDENT-GAUSS, where the unconditional mean m_0 and variance s_0^2 are replaced with their conditional counterparts $(m_0(\cdot), s_0^2(\cdot))$. As an empirical Bayes strategy, CLOSE-GAUSS then replaces the unknown conditional moments with their estimated counterparts.[18] Its properties depend on those of the oracle (2.12) it mimics, which we turn to now.

Despite being rationalized under the assumption $\theta_i \mid \sigma_i \sim \mathcal{N}(m_0(\sigma_i), s_0^2(\sigma_i))$, (2.12) enjoys strong robustness properties: It is optimal over a restricted class of decision rules and minimax over all decision rules—without imposing the location-scale assumption (2.6). First, (2.12) is the optimal decision rule for estimating θ_i when we restrict to the class of decision rules that are linear in Y_i (Weinstein et al., 2018). Second, (2.12) is minimax in the sense that it minimizes the worst-case mean squared error, where an adversary chooses $G_{(1)}, \ldots, G_{(n)}$, subjected to the constraint that $G_{(i)}$'s first two moments are $(m_0(\sigma_i), s_0^2(\sigma_i))$.[19]

[17]The nonparametric maximum likelihood has a long history in econometrics and statistics (Kiefer and Wolfowitz, 1956; Lindsay, 1995; Heckman and Singer, 1984). There is recent renewed interest. See, among others, Koenker and Gu (2019); Koenker and Mizera (2014); Jiang and Zhang (2009); Jiang (2020); Soloff et al. (2021); Saha and Guntuboyina (2020); Polyanskiy and Wu (2020); Shen and Wu (2022); Polyanskiy and Wu (2021). Empirical Bayes methods via NPMLE have computational and theoretical advantages, though much of the favorable theoretical results are proven in a homoskedastic setting. Its computational ease (Koenker and Mizera, 2014; Koenker and Gu, 2017) and lack of tuning parameters are advocated in Koenker and Gu (2019). Polyanskiy and Wu (2020) find that, with high probability, NPMLE recovers a distribution $\hat{G}_n$ with only $O(\log n)$ support points despite searching over the set of all distributions; they refer to this property as self-regularization. For regret control in the homoskedastic Gaussian model, Jiang and Zhang (2009)'s result is the best known and matches a lower bound up to log factors (Polyanskiy and Wu, 2021).

[18](2.12) is first proposed by Weinstein et al. (2018). Weinstein et al. (2018) propose estimating $m_0(\cdot), s_0(\cdot)$ in a particular manner to ensure the resulting empirical Bayes posterior means dominate the naive estimates Y_i uniformly over $\theta_{1:n}, \sigma_{1:n}$, which are conditioned upon.

[19]Formally,

$$\theta^*_{1:n,\mathcal{N}(0,1),\eta_0} \in \arg\min_{\delta_{1:n}} \sup_{G_{(1:n)}} \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{G_{(i)}}\left[(\delta_i(Y_{1:n}, \sigma_{1:n}) - \theta_i)^2\right],$$

where the supremum is taken over $G_{(i)}$ having moments $\eta_0(\sigma_i)$. To wit, note that the Bayes risk of (2.12) is the same regardless of choices of $G_{(1)}, \ldots, G_{(n)}$ under the moment constraint, and it is equal to the optimal Bayes risk when $G_{(i)} \sim \mathcal{N}(m_0(\sigma_i), s_0^2(\sigma_i))$. We therefore conclude that (2.12) is minimax by observing that the minimax Bayes risk is at least the risk of (2.12).

However, the Normality assumption does imply that (2.12), unlike CLOSE-NPMLE, fails to approximate the optimal decision (2.4) when the location-scale assumption (2.6) holds but $\theta_i \mid \sigma_i$ may not be Gaussian. Since we also show that CLOSE-NPMLE is worst-case robust—though with higher worst-case risk than CLOSE-GAUSS, we recommend CLOSE-NPMLE over CLOSE-GAUSS, unless the researcher is extremely concerned about the misspecification of the location-scale model. ■

2.3. **Decision problems.** To prepare for our theoretical results in the next section, we close this one by introducing decision theory notation and formalizing a few decision problems. Let $\boldsymbol{\delta}(Y_{1:n}, \sigma_{1:n})$ be a *decision rule* mapping the data $(Y_{1:n}, \sigma_{1:n})$ to *actions*. Let $L(\boldsymbol{\delta}, \theta_{1:n})$ denote a *loss function* mapping actions and parameters to a scalar. Let $R_{\mathrm{F}}(\boldsymbol{\delta}, \theta_{1:n}) = \mathbb{E}[L(\boldsymbol{\delta}, \theta_{1:n}) \mid \theta_{1:n}, \sigma_{1:n}]$ denote the *frequentist risk* associated with the loss function L, which integrates over the randomness in $Y_{1:n}$, keeping $\theta_{1:n}, \sigma_{1:n}$ fixed. Finally, let $R_{\mathrm{B}}(\boldsymbol{\delta}; P_0) = \mathbb{E}_{P_0}[R_{\mathrm{F}}(\boldsymbol{\delta}, \theta_{1:n}) \mid \sigma_{1:n}]$ be the *Bayes risk* of $\boldsymbol{\delta}$ under P_0, which additionally integrates over the conditional distribution $\theta_{1:n} \mid \sigma_{1:n}$.[20]

The oracle Bayes decision rule $\boldsymbol{\delta}^\star$ (2.4) is optimal in the sense that it minizes R_{B}. A natural metric of success for the empirical Bayesian (2.5) is thus the gap between the Bayes risks of $\boldsymbol{\delta}_{\mathrm{EB}}$ and $\boldsymbol{\delta}^\star$. We refer to this quantity as *Bayes regret*:

$$\mathrm{BayesRegret}_n(\boldsymbol{\delta}_{\mathrm{EB}}) = R_{\mathrm{B}}(\boldsymbol{\delta}_{\mathrm{EB}}; P_0) - R_{\mathrm{B}}(\boldsymbol{\delta}^\star; P_0) = \mathbb{E}[L(\boldsymbol{\delta}_{\mathrm{EB}}, \theta_{1:n}) - L(\boldsymbol{\delta}^\star, \theta_{1:n}) \mid \sigma_{1:n}] \quad (2.13)$$

where the right-hand side integrates over the randomness in $\theta_{1:n}, Y_{1:n}$, and, by extension, $\hat{P}$. If an empirical Bayes method achieves low Bayes regret, then it successfully imitates the decisions of the oracle Bayesian, and its decisions are thus approximately optimal. Our theoretical results focus on bounding Bayes regret for CLOSE.[21]

We introduce a few concrete decision problems by specifying the actions $\boldsymbol{\delta}$ and loss functions L and state the corresponding oracle Bayes and empirical Bayes decision rules.

Decision Problem 1 (Squared-error estimation of $\theta_{1:n}$)**.** The canonical statistical problem (Robbins, 1956) is estimating the parameters $\theta_{1:n}$ under mean-squared error (MSE). That is, the action $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$ collects estimates δ_i for parameters θ_i, evaluated with MSE:

$$L(\boldsymbol{\delta}, \theta_{1:n}) = \frac{1}{n} \sum_{i=1}^{n} (\delta_i - \theta_i)^2.$$

[20]Since $\sigma_{1:n}$ is kept fixed throughout, we suppress their appearances in $R_{\mathrm{B}}(\cdot), R_{\mathrm{F}}(\cdot)$.

[21]Bayes regret is likewise the focus of the literature in empirical Bayes that we build on (Jiang, 2020; Soloff et al., 2021). On the other hand, other optimality criteria are also considered. For instance, Kwon (2021), Xie et al. (2012), Abadie and Kasy (2019), and Jing et al. (2016) propose methods that use Stein's Unbiased Risk Estimate (SURE) to select hyperparameters for a class of shrinkage procedures. A common thread of these approaches is that they seek optimality in terms of the frequentist risk R_{F}—which is stronger than controlling the Bayes risk R_{B}—but limit attention to squared error and to a restricted class of methods.

The oracle Bayes decision rule $\boldsymbol{\delta}^\star = (\delta_1^\star, \ldots, \delta_n^\star)$ here is the posterior mean under P_0, denoted by $\theta_i^* = \theta_{i,P_0}^*$:

$$\delta_i^\star = \theta_{i,P_0}^* \equiv \mathbb{E}_{P_0}[\theta_i \mid Y_i, \sigma_i]$$

with empirical Bayesian counterpart $\hat{\theta}_{i,\hat{P}} = \mathbf{E}_{\hat{P}}[\theta_i \mid Y_i, \sigma_i]$. ■

Next, we describe two problems that are likely more relevant for policy-making, such as replacing low value-added teachers and recommending high economic mobility tracts (Gilraine et al., 2020; Bergman et al., 2023).[22]

Decision Problem 2 (UTILITY MAXIMIZATION BY SELECTION)**.** Suppose $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$, where $\delta_i \in \{0, 1\}$ is a selection decision for population i. For each population, selecting that population has benefit θ_i and known cost c_i. The decision maker wishes to maximize utility (i.e., negative loss):

$$-L(\boldsymbol{\delta}, \theta_{1:n}) = \frac{1}{n}\sum_{i=1}^n \delta_i \left(\theta_i - c_i\right).$$

The oracle Bayes rule selects all populations whose posterior mean benefit θ_{i,P_0}^* exceeds the selection cost c_i:

$$\delta_i^\star = \mathbb{1}\left(\theta_{i,P_0}^* \geq c_i\right).$$

One natural empirical Bayes decision rule replaces θ_{i,P_0}^* with $\theta_{i,\hat{P}}^*$, following (2.5).

In a context where the parameters are conditional average treatment effects for a particular covariate cell, $\theta_i = \text{CATE}(i) \equiv \mathbb{E}[Y(1) - Y(0) \mid X = i]$, and δ_i are treatment decisions, this problem is an instance of welfare maximization by treatment choice (Manski, 2004; Stoye, 2009; Kitagawa and Tetenov, 2018; Athey and Wager, 2021). In this setting, δ_i is a decision to treat individuals with covariate values in the i^{th} cell. The average benefit of treating these individuals is their conditional average treatment effect θ_i, and the cost of treatment is c_i.[23] ■

Decision Problem 3 (TOP-m SELECTION)**.** Similar to UTILITY MAXIMIZATION BY SELECTION, suppose $\boldsymbol{\delta}$ consists of binary selection decisions, with the additional constraint that exactly m populations are chosen: $\sum_i \delta_i = m$. The decision maker's utility is the average θ_i of the selected set:

$$-L(\boldsymbol{\delta}, \theta_{1:n}) = \frac{1}{m}\sum_{i=1}^n \delta_i\theta_i. \tag{2.14}$$

[22]We analyze these problems from a decision-theoretic perspective, under the sampling assumption (2.3). For a different and complementary perspective in terms of conditional-on-θ frequentist inference on ranks, see Mogstad et al. (2020, 2023). For additional ranking-related decision problems, see Gu and Koenker (2023).

[23]The literature on treatment choice uses a different notion of regret compared to this paper (based on R_{F} rather than R_{B}).

Oracle Bayes selects the populations corresponding to the m largest posterior means θ^*_{i,P_0} (breaking ties arbitrarily):

$$\delta_i^\star = \mathbb{1}\left(\theta^*_{i,P_0} \text{ is among the top-}m \text{ of } \theta^*_{1:n,P_0}\right).$$

Again, the empirical Bayes recipe (2.5) suggests replacing P_0 with the estimate $\hat{P}$.

The utility function (2.14) rationalizes the widespread practice of screening based on empirical Bayes posterior means. For instance, this objective may be reasonable for rewarding the top 5% of teachers or replacing the bottom 5%, according to value-added (Gilraine et al., 2020; Chetty et al., 2014; Kane and Staiger, 2008; Hanushek, 2011). In Bergman et al. (2023), where housing voucher holders are incentivized to move to Census tracts selected according to economic mobility, (2.14) represents the expected economic mobility of a mover if they move randomly to one of the selected tracts.[24] ■

3. Regret results for CLOSE-NPMLE

We observe $(Y_i, \sigma_i)_{i=1}^n$, where (θ_i, σ_i) satisfies the location-scale assumption (2.6) and $(Y_i, \theta_i, \sigma_i)$ obeys the Gaussian location model (2.1). Our recommended procedure, CLOSE-NPMLE, transforms the data (Y_i, σ_i) into $(\hat{Z}_i, \hat{\nu}_i)$, with estimated nuisance parameters $\hat{\eta} = (\hat{m}, \hat{s})$ for $\eta_0 = (m_0, s_0)$ in **CLOSE–STEP 1**. It then estimates the unknown shape parameter G_0 via NPMLE (2.11) on $(\hat{Z}_i, \hat{\nu}_i)_{i=1}^n$.

Our leading result shows that CLOSE-NPMLE mimics the oracle Bayesian as well as possible, for the problem of estimation under squared error loss, in the sense that its Bayes regret vanishes at the minimax optimal rate. Our second result connects squared error estimation to Decision Problems 2 and 3, by showing that if an empirical Bayesian has low regret in squared error loss, then they likewise have low regret for Decision Problems 2 and 3.

Since our main result assumes the location-scale model, one may be concerned about its potential misspecification. The last result in this section, Theorem 4, bounds the worst-case Bayes risk of an idealized version of CLOSE-NPMLE (i.e. with known η_0 and fixed but misspecified $\hat{G}_n$) as a multiple of a notion of minimax risk, without assuming (2.6). Thus, even under misspecification, CLOSE-NPMLE does not perform arbitrarily badly relative to the minimax procedure.

The rest of this section states and discusses these results formally. Practitioners who are less interested in the theoretical details are free to skip to Section 4, where we discuss a number of practical considerations.

Remark 3 (Notation)**.** In what follows, we use the symbol C to denote a generic positive and finite constant which does not depend on n. We use the symbol C_x to denote a generic positive

[24]Our theoretical results in Section 3.2 can accommodate a slightly more general decision problem, which allows for an expected mobility interpretation for movers who do not move uniformly randomly. See Remark 5.

and finite constant that depends only on x, some parameter(s) that describe the problem. Occurrences of the same symbol C, C_x may not refer to the same constants. Similarly, for $A_n, B_n \geq 0$, generally functions of n, we use $A_n \lesssim B_n$ to mean that some universal C exists such that $A_n \leq CB_n$ for all n, and we use $A \lesssim_x B$ to mean that some universal C_x exists such that $A_n \leq C_x B_n$ for all n. In logical statements, appearances of $\lesssim$ implicitly prepend "there exists a universal constant" to the statement.[25] Since all expectation or probability statements are with respect to the conditional distribution P_0 of $\theta_{1:n} \mid \sigma_{1:n}$, going forward, we treat $\sigma_{1:n}$ as fixed and simply write $\mathbb{E}[\cdot], \mathrm{P}(\cdot)$ to denote the expectation and probability over $\theta_{1:n} \mid \sigma_{1:n} \sim P_0$. We omit the P_0 subscript and the conditioning on $\sigma_{1:n}$. ■

3.1. **Regret rate in squared error.** Since we consider CLOSE-NPMLE in mean-squared error, we define

$$\mathrm{MSERegret}_n(G, \eta) \equiv \frac{1}{n}\sum_{i=1}^n (\hat{\theta}_{i,G,\eta} - \theta_i)^2 - \frac{1}{n}\sum_{i=1}^n (\theta_i^* - \theta_i)^2$$

$$\theta_i^* \equiv \theta_{i,P_0}^* = \mathbb{E}_{P_0}[\theta_i \mid Y_i, \sigma_i] \quad \hat{\theta}_{i,G,\eta} \equiv \mathbf{E}_{G,\eta}[\theta_i \mid Y_i, \sigma_i] \equiv \frac{\int \theta \varphi\left(\frac{Y_i - \theta}{\sigma_i}\right) \frac{1}{\sigma_i} dG\left(\frac{\theta - m(\sigma_i)}{s(\sigma_i)}\right)}{\int \varphi\left(\frac{Y_i - \theta}{\sigma_i}\right) \frac{1}{\sigma_i} dG\left(\frac{\theta - m(\sigma_i)}{s(\sigma_i)}\right)}$$

as the excess loss of the empirical Bayes posterior means—obtained by prior G and nuisance parameter estimate η for η_0—relative to that of the oracle Bayes posterior means. The Bayes regret for CLOSE-NPMLE in squared error is then the P_0-expectation of $\mathrm{MSERegret}_n$:

$$\mathrm{BayesRegret}_n = \mathbb{E}\left[\mathrm{MSERegret}_n(\hat{G}_n, \hat{\eta})\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\theta_i^* - \hat{\theta}_{i,\hat{G}_n,\hat{\eta}})^2\right]. \tag{3.1}$$

Equation (3.1) additionally notes that expected $\mathrm{MSERegret}_n$ is equal to the expected mean-squared difference between the empirical Bayesian posterior means $\hat{\theta}_{i,\hat{G}_n,\hat{\eta}}$ and the oracle Bayes posterior means.

We assume that $P_0 \in \mathcal{P}_0$ belongs to some restricted class. Informally speaking, our first main result shows that for some constants $C, \beta > 0$ that depend solely on $\mathcal{P}_0$, the Bayes regret in squared error decays at the same rate as the maximum estimation error for η_0 squared:

$$\mathrm{BayesRegret}_n \leq C(\log n)^\beta \max\left(\mathbb{E}\|\hat{\eta} - \eta_0\|_\infty^2, \frac{1}{n}\right),$$

where we define $\|\eta\|_\infty = \max(\|m\|_\infty, \|s\|_\infty)$ for $\eta = (m, s)$. This result continues a recent statistics literature on empirical Bayes methods via NPMLE by characterizing the effect of an estimated nuisance parameter $\hat{\eta}$ in a first step.[26]

---

[25]For instance, statements like "under certain assumptions, $\mathrm{P}(A_n \lesssim B_n) \geq c_0$" should be read as "under certain assumptions, there exists a constant $C > 0$ such that for all n, $\mathrm{P}(A_n \leq CB_n) \geq c_0$."

[26]Our theory hews closely to—and extends—the results in Jiang (2020) and Soloff et al. (2021), which themselves are extensions of earlier results in the homoskedastic setting (Jiang and Zhang, 2009; Saha

Moreover, we show that controlling the Bayes regret is no easier than estimating m in $\|\cdot\|_2$, which is a corresponding lower bound on regret. There exists c such that for any estimator of θ_i, its worst-case regret is bounded below[27]

$$\sup_{P_0 \in \mathcal{P}_0} \text{BayesRegret}_n \geq c \inf_{\hat{m}} \sup_{m_0} \mathbb{E}\|\hat{m} - m_0\|_2^2.$$

Since the minimax estimation rates of $\|\hat{\eta} - \eta_0\|_\infty$ and of $\|\hat{\eta} - \eta_0\|_2$ are the same up to logarithmic factors, we conclude that our regret upper bound is rate-optimal up to logarithmic factors. We now introduce the assumptions on $P_0 \in \mathcal{P}_0$ needed for these results, state the upper and lower bounds, and provide a technical discussion.

3.1.1. *Assumptions for regret upper bound.* We first assume that $\hat{G}_n$ is an *approximate* maximizer of the log-likelihood on the transformed data $\hat{Z}_i$ and $\hat{\nu}_i$ satisfying some support restrictions. This is not a restrictive assumption, as the actual maximizers of the log-likelihood function satisfy it.[28]

Assumption 1. *Let* $\psi_i(Z_i, \hat{\eta}, G) \equiv \log\left(\int_{-\infty}^{\infty} \varphi\left(\frac{\hat{Z}_i - \tau}{\hat{\nu}_i}\right) G(d\tau)\right)$ *be the objective function in* (2.11), *ignoring a constant factor* $1/\hat{\nu}_i$. *We assume that* $\hat{G}_n$ *satisfies*

$$\frac{1}{n}\sum_{i=1}^n \psi_i(Z_i, \hat{\eta}, \hat{G}_n) \geq \sup_{H \in \mathcal{P}(\mathbb{R})} \frac{1}{n}\sum_{i=1}^n \psi_i(Z_i, \hat{\eta}, H) - \kappa_n \tag{3.2}$$

for tolerance κ_n

$$\kappa_n = \frac{2}{n}\log\left(\frac{n}{\sqrt{2\pi e}}\right). \tag{3.3}$$

Moreover, we require that $\hat{G}_n$ *has support points within* $[\min_i \hat{Z}_i, \max_i \hat{Z}_i]$. *To ensure that* κ_n *is positive, we assume that* $n \geq 7 = \lceil \sqrt{2\pi e} \rceil$.[29]

We now state further assumptions on the data-generating processes $\mathcal{P}_0$ beyond (2.6). First, we assume that G_0 is exponential-tailed with parameter α that controls the thickness of its tails. We state the restriction in an equivalent form of simultaneous moment control.[30]

---

and Guntuboyina, 2020). These results, under either homoskedasticity or prior independence, show that empirical Bayes derived from estimating the prior via NPMLE achieves fast regret rates. In particular, Soloff et al. (2021) show that the regret rate is of the form $C(\log n)^\beta \frac{1}{n}$ under prior independence and assumptions similar to ours.

[27] Our proof only exploits a lower bound for the performance of $\hat{m}$; doing so is without loss if m_0 and s_0 belong to the same smoothness class.

[28] In particular, the support restriction for $\hat{G}_n$ in Assumption 1 is satisfied by all maximizers of the likelihood function (see Corollary 3 in Soloff et al., 2021).

[29] The constants κ_n also feature in Jiang (2020) to ensure that the fitted likelihood is bounded away from zero. The particular constants in κ_n are chosen to simplify expressions and are not material to the result.

[30] An equivalent statement to Assumption 2 is that there exists $a_1, a_2 > 0$ such that $\mathrm{P}_{G_0}(|\tau| > t) \leq a_1 \exp(-a_2 t^\alpha)$ for all $t > 0$. Note that when $\alpha = 2$, G_0 is subgaussian, and when $\alpha = 1$, G_0 is subexponential (see the definitions in Vershynin, 2018), as commonly assumed in high-dimensional statistics. Assumption 2 is slightly stronger than requiring that all moments exist for G_0, and weaker than requiring

Assumption 2. *The distribution G_0 is has zero mean, unit variance, and admits simultaneous moment control with parameter $\alpha \in (0, 2]$: There exists a constant $A_0 > 0$ such that for all $p > 0$,*

$$(\mathbb{E}_{\tau \sim G_0}[|\tau|^p])^{1/p} \leq A_0 p^{1/\alpha}. \tag{3.4}$$

Next, Assumption 3 imposes that members of $\mathcal{P}_0$ have various variance parameters uniformly bounded away from zero and infinity. This is a standard assumption in the literature, maintained likewise by Jiang (2020) and Soloff et al. (2021).

Assumption 3. *The variances $(\sigma_{1:n}, s_0)$ admit lower and upper bounds:*

$$\sigma_\ell < \sigma_i < \sigma_u \text{ and } s_\ell < s_0(\cdot) < s_u,$$

where $0 < \sigma_\ell, \sigma_u, s_{0\ell}, s_{0u} < \infty$. This implies that $0 < \nu_\ell \leq \nu_i = \frac{\sigma_i}{s_0(\sigma_i)} \leq \nu_u < \infty$ for some ν_ℓ, ν_u.

Lastly, we require that m_0, s_0 satisfies some smoothness restrictions. We also require that $\hat{m}, \hat{s}$ satisfy some corresponding regularity conditions.

Assumption 4. *Let $C^p_{A_1}([\sigma_\ell, \sigma_u])$ be the Hölder class of order $p \geq 1$ with maximal Hölder norm $A_1 > 0$ supported on $[\sigma_\ell, \sigma_u]$.[31] We assume that*

(1) The true conditional moments are Hölder-smooth: $m_0, s_0 \in C^p_{A_1}([\sigma_\ell, \sigma_u])$.

Additionally, let $\beta_0 > 0$ be a constant. Let $\mathcal{V}$ be a set of bounded functions supported on $[\sigma_\ell, \sigma_u]$ that (i) admits the uniform bound $\sup_{f \in \mathcal{V}} \|f\|_\infty \leq C_{A_1}$ and (ii) admits the metric entropy bound

$$\log N(\epsilon, \mathcal{V}, \|\cdot\|_\infty) \leq C_{A_1, p, \sigma_\ell, \sigma_u} (1/\epsilon)^{1/p}.$$

We assume that the estimators for m_0 and s_0, $\hat{\eta} = (\hat{m}, \hat{s})$, satisfy the following assumptions.

(2) For any $\epsilon > 0$, there exists a sufficiently large $C = C(\epsilon)$, independently of n, such that for all n,

$$\mathrm{P}\left(\max\left(\|\hat{m} - m_0\|_\infty, \|\hat{s} - s_0\|_\infty\right) > C(\epsilon) n^{-\frac{p}{2p+1}} (\log n)^{\beta_0}\right) < \epsilon.$$

G_0 to have a moment-generating function. Similar tail assumptions feature in the theoretical literature on empirical Bayes (Soloff et al., 2021; Jiang and Zhang, 2009; Jiang, 2020).

[31] We recall the definition of a Hölder class from van der Vaart and Wellner (1996), Section 2.7.1. We specialize its definition to functions of one real variable. For an integer p, Hölder-p functions are $(p-1)$-times differentiable, with a Lipschitz continuous $(p-1)^{\text{st}}$ derivative.

Definition 1. For some set $\mathcal{X} \subset \mathbb{R}$ and constant $A > 0$, $p > 0$, let $C^p_A(\mathcal{X})$ be the set of continuous functions $f : \mathcal{X} \to \mathbb{R}$ with $\|f\|_{(p)} \leq A$. The norm $\|\cdot\|_{(p)}$ is defined as follows. Let $\underline{p}$ be the greatest integer strictly smaller than p. Define

$$\|f\|_{(p)} = \max_{k \leq \underline{p}} \sup_{x \in \mathcal{X}} \left|f^{(k)}(x)\right| + \sup_{x, y \in \mathcal{X}} \frac{\left|f^{(\underline{p})}(x) - f^{(\underline{p})}(y)\right|}{|x - y|^{p - \underline{p}}}.$$

We refer to $C^p_A(\mathcal{X})$ as a Hölder class of order p and $\|f\|_{(p)}$ as the Hölder norm.

(3) *The nuisance estimators take values in $\mathcal{V}$ almost surely:* $\mathrm{P}(\hat{m} \in \mathcal{V}, \hat{s} \in \mathcal{V}) = 1$.

(4) *The conditional variance estimator respects the conditional variance bounds in Assumption 3:* $\mathrm{P}\left(\frac{s_{0\ell}}{2} < \hat{s} < 2s_{0u}\right) = 1$.

Assumption 4 is a Hölder smoothness assumption on the nuisance parameters m_0 and s_0, which is a standard regularity condition in nonparametric regression; our subsequent minimax rate optimality statements are relative to this class. Moreover, it is also a high-level assumption on the quality of the estimation procedure for $(\hat{m}, \hat{s})$. Specifically, Assumption 4 expects that the nuisance parameter estimates $\hat{m}$ and $\hat{s}$ are rate-optimal up to logarithmic factors (Stone, 1980). Assumption 4 also expects that the nuisance parameter estimates belong to a class $\mathcal{V}$ with the same metric entropy behavior as the Hölder class $C^p_{A_1}([\sigma_\ell, \sigma_u])$.[32]

Assumptions 2 to 4 specify a class of distributions $\mathcal{P}_0$ and nuisance estimators $\hat{\eta}$ indexed by a set of hyperparameters $\mathcal{H} = (\sigma_\ell, \sigma_u, s_\ell, s_u, A_0, A_1, \alpha, \beta_0, p)$. Our subsequent theoretical results are finite sample, with implicit constants dependent on these hyperparameters $\mathcal{H}$. To review, $(\sigma_\ell, \sigma_u, s_\ell, s_u)$ are bounds on the variances $(\sigma_i^2, s^2(\sigma_i))$; (A_0, α) control the tails of G_0; and (A_1, p) control the smoothness of η_0; and β_0 is the power of the log factor in the $\|\cdot\|_\infty$ estimation rate for η_0.

3.1.2. *Regret results.* Consider the following "good event," indexed by $C > 0$,

$$\mathbf{A}_n(C) \equiv \left\{\|\hat{\eta} - \eta_0\|_\infty \le C n^{-\frac{p}{2p+1}} (\log n)^{\beta_0}\right\}. \tag{3.5}$$

$\mathbf{A}_n(C)$ indicates that the nuisance parameter estimates satisfy some rate in $\|\cdot\|_\infty$. Our main result derives a convergence rate for the expected MSE regret conditional on this good event $\mathbf{A}_n(C)$.

Theorem 1. *Assume Assumptions 1 to 4 hold. Then, for any $\delta \in (0, \frac{1}{2})$, there exists universal constants $C_{1,\mathcal{H},\delta} > 0$ and $C_{0,\mathcal{H},\delta} > 0$ such that (i) $\mathrm{P}(\mathbf{A}_n(C_{1,\mathcal{H},\delta})) \ge 1 - \delta$ and that (ii) the expected regret conditional on $\mathbf{A}_n(C_{1,\mathcal{H},\delta})$ is dominated by the rate function*

$$\mathbb{E}\left[\mathrm{MSERegret}_n(\hat{G}_n, \hat{\eta}) \mid \mathbf{A}_n(C_{1,\mathcal{H},\delta})\right] \le C_{0,\mathcal{H},\delta} n^{-\frac{2p}{2p+1}} (\log n)^{\frac{2+\alpha}{\alpha}+3+2\beta_0}. \tag{3.6}$$

If the event $\mathbf{A}_n(C)$ is sufficiently likely, we can control expected regret on the bad event $\mathbf{A}_n^{\mathrm{C}}$ as well. In Appendix G, we verify that local linear regression satisfies a weakening of these assumptions that are also sufficient for the conclusion of Corollary 1.

---

[32]Regarding Assumption 4(2), we note that kernel smoothing estimators attain the rates required for Hölder smooth functions m_0, s_0 (see Tsybakov (2008) and Appendix G). Regarding Assumption 4(3), if the nuisance parameters are p-Hölder smooth almost surely, we can simply take $\mathcal{V} = C^p_{A_1'}([\sigma_\ell, \sigma_u])$ for some potentially different A_1'. This can be achieved in practice by, say, projecting estimated nuisance parameters $\tilde{\eta}$ to $C_{A_1}([\sigma_\ell, \sigma_u])$ in $\|\cdot\|_\infty$. Finally, Assumption 4(4) also expects the nuisance parameter estimates to respect the boundedness constraints for s_0. This is mainly so that our results are easier to state; we discuss this assumption in Remark C.1.

Corollary 1. *Assume the same setting as Theorem 1. Suppose, additionally, for all sufficiently large $C_{1,\mathcal{H}} > 0$, $\mathrm{P}(\mathbf{A}_n(C_{1,\mathcal{H}})) \geq 1 - n^{-2}$. Then, there exists a constant $C_{0,\mathcal{H}} > 0$ such that the expected regret is dominated by the rate function*

$$\text{BayesRegret}_n = \mathbb{E}\left[\text{MSERegret}_n(\hat{G}_n, \hat{\eta})\right] \leq C_{0,\mathcal{H}} n^{-\frac{2p}{2p+1}} (\log n)^{\frac{2+\alpha}{\alpha}+3+2\beta_0}.$$

We can show a corresponding lower bound on the Bayes regret—i.e., a lower bound on the worst-case Bayes regret when an adversary picks G_0, η_0—by showing that any good posterior mean estimate $\hat{\theta}_i$ implies a good estimate $\hat{m}(\sigma_i)$ for m_0. Minimax lower bounds for estimation of m_0 then imply lower bounds for estimation of the oracle posterior means θ_i^*.[33]

Theorem 2. *Fix a set of valid hyperparameters $\mathcal{H} = (\sigma_\ell, \sigma_u, s_\ell, s_u, A_0, A_1, \alpha, \beta_0, p)$ for Assumptions 2 to 4. Let $\mathcal{P}(\mathcal{H}, \sigma_{1:n})$ be the set of distributions P_0 on support points $\sigma_{1:n}$ which satisfy (2.6) and Assumptions 2 to 4 corresponding to $\mathcal{H}$. For a given P_0, let $\theta_i^* = \mathbb{E}_{P_0}[\theta_i \mid Y_i, \sigma_i]$ denote the oracle posterior means. Then there exists a constant $c_{\mathcal{H}} > 0$ such that the worst-case Bayes regret of any estimator exceeds $c_{\mathcal{H}} n^{-\frac{2p}{2p+1}}$:*

$$\inf_{\hat{\theta}_{1:n}} \sup_{\substack{\sigma_{1:n} \in (\sigma_\ell, \sigma_u) \\ P_0 \in \mathcal{P}(\mathcal{H}, \sigma_{1:n})}} \mathbb{E}_{P_0}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2 - (\theta_i^* - \theta_i)^2\right] \geq c_{\mathcal{H}} n^{-\frac{2p}{2p+1}},$$

where the infimum is taken over all (possibly randomized) estimators of $\theta_{1:n}$.

As a result, the rate (3.6) is optimal up to logarithmic factors. The additional logarithmic factors are partly the price of having to estimate G_0 via NPMLE and partly deficiencies in the proof of Theorem 1. In any case, this cost is not substantial.

The regret upper bounds Theorem 1 and Corollary 1 are finite-sample statements. As a result, they hold uniformly over all distributions P_0 delineated by the problem parameters $\mathcal{H}$. However, the usefulness of Theorem 1 and Corollary 1 still lies in the convergence rate, as the constants implied by the proofs are not sharp.

These regret upper bounds readily extend to the case where covariates are present and the location-scale assumption is with respect to the additional covariates X_i:

$$\theta_i \mid \sigma_i, X_i \sim G_0\left(\frac{\theta_i - m_0(X_i, \sigma_i)}{s_0(X_i, \sigma_i)}\right),$$

under assumption on $m_0, s_0, \hat{m}, \hat{s}$ analogous to Assumption 4. Of course, the resulting convergence rate would suffer from the curse of dimensionality, and the term $n^{-\frac{2p}{2p+1}}$ would be replaced with $n^{-\frac{2p}{2p+1+d}}$, where d is the dimension of X.

Taken together, Corollary 1 and Theorem 2 are strong statistical optimality guarantees for CLOSE-NPMLE in the canonical problem of estimation with squared error loss. That is,

[33] A similar argument is considered in Ignatiadis and Wager (2019) for a related but distinct setting. See, also, Appendix A.6.2.

the worst-case performance gap of CLOSE-NPMLE relative to the oracle contracts at the best possible rate, meaning that CLOSE-NPMLE mimics the oracle as well as possible.

For interested readers, we provide an overview of the proof of our main result Theorem 1 in the following remark. A more detailed walkthrough is in Appendix C.3.

Remark 4 (Informal discussion of the proof for Theorem 1)**.** Regret results assuming prior independence are established by Soloff et al. (2021) and Jiang (2020), and we build on these results for Theorem 1. Applied to (Z_i, ν_i, τ_i), these results state that **(i)** approximate maximizers $\tilde{G}_n$ of the (infeasible) log-likelihood $\Psi_n(\eta_0, G) \equiv \frac{1}{n}\sum_i \psi_i(Z_i, \eta_0, G)$ are close to G_0 in terms of the *average Hellinger distance* of the induced densities of Z_i

$$\overline{h}^2(f_{\tilde{G}_n,\cdot}, f_{G_0,\cdot}) \equiv \frac{1}{n}\sum_{i=1}^n h^2\left(\mathcal{N}(0,\nu_i^2) \star \tilde{G}_n, \mathcal{N}(0,\nu_i^2) \star G_0\right), \quad h^2(f,g) \equiv 1 - \int_{-\infty}^{\infty} \sqrt{f(x)g(x)}\, dx$$

and **(ii)** if $\overline{h}^2(f_{\tilde{G}_n,\cdot}, f_{G_0,\cdot})$ is small, then posterior means for τ_i under $\tilde{G}_n$ are close to posterior means under G_0 in squared error.

Our results extend this argument by accommodating the fact that η_0 is unknown and must be estimated with $\hat{\eta}$.[34] To apply **(ii)** in the literature, we would like to show that **(i')** $\hat{G}_n$—an approximate maximizer of the feasible log-likelihood $\Psi_n(\hat{\eta}, G) = \frac{1}{n}\sum_i \psi_i(Z_i, \hat{\eta}, G)$—is close to G_0 in terms of $\overline{h}^2(\cdot,\cdot)$. This is not a straightforward task and is the most intricate part of our argument. To show **(i')**, we prove a lower bound for the likelihood $\Psi_n(\eta_0, \hat{G}_n)$ (Theorem D.1) and adapt the argument for **(i)** to accommodate our likelihood lower bound (Theorem E.1).

To lower bound $\Psi_n(\eta_0, \hat{G}_n)$, we relate the two likelihoods by linearization (formally, see (D.4)):

$$\Psi_n(\hat{\eta}, \hat{G}_n) - \Psi_n(\eta_0, \hat{G}_n) \approx \frac{1}{n}\sum_{i=1}^n \frac{\partial \psi_i(Z_i, \eta_0, \hat{G}_n)}{\partial \eta} \underbrace{(\hat{\eta}(\sigma_i) - \eta_0(\sigma_i))}_{\leq \|\hat{\eta}-\eta_0\|_\infty}.$$

Since $\hat{G}_n$ approximately maximizes the feasible likelihood $\Psi_n(\hat{\eta}, \cdot)$, $\Psi_n(\hat{\eta}, \hat{G}_n)$ is large by construction. Thus, if we can show that the right-hand side is small, then the infeasible likelihood $\Psi_n(\eta_0, \hat{G}_n)$ would be close to $\Psi_n(\hat{\eta}, \hat{G}_n)$ and hence would also be large. To obtain the rate (3.6), it is important to show that the right-hand side vanishes *strictly faster* than $\|\hat{\eta} - \eta_0\|_\infty$. To do so, we additionally need to show that the derivatives $\frac{1}{n}\sum_i \partial\psi_i(Z_i, \eta_0, \hat{G}_n)/\partial\eta$ are small. Without it, we would obtain a worse squared error regret rate of the form $n^{-\frac{p}{2p+1}}(\log n)^\beta$.

[34]We also translate the resulting regret guarantee on estimating τ_i to regret guarantees on estimating θ_i. In doing so, we identify an apparent gap in the arguments of Jiang (2020) and Soloff et al. (2021). We show a modified argument avoids the gap in our setting, which applies to the setting in Soloff et al. (2021) as well. See Remark F.1 for details.

In particular, we manage to relate the average derivative to the average Hellinger distance (see Lemmas D.1 and D.2)

$$\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\psi_i(Z_i,\eta_0,\hat{G}_n)}{\partial\eta}(\hat{\eta}(\sigma_i)-\eta_0(\sigma_i))\right| \lesssim (\log n)^{\gamma}\overline{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot})\|\hat{\eta}-\eta_0\|_\infty, \text{ for some } \gamma > 0.$$

Loosely, this is because the population score in η is mean-zero, $\mathbb{E}[\partial\psi_i(Z,\eta_0,G_0)/\partial\eta] = 0$. Thus if $\hat{G}_n$ is close to G_0, then the sample score evaluated at $\hat{G}_n$ should also be approximately zero. This is a key step in Appendix D.

This bound for $\Psi_n(\eta_0,\hat{G}_n)$ creates an additional complication when attempting to apply the claim **(i)**. The claim **(i)** upper bounds the Hellinger distance $\overline{h}(f_{\tilde{G}_n,\cdot}, f_{G_0,\cdot})$ using a lower bound for $\Psi_n(\eta_0,\tilde{G}_n)$. However, now our lower bound for the likelihood $\Psi_n(\eta_0,\hat{G}_n)$ itself depends on $\overline{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot})$, and so we cannot apply **(i)** directly. The proof for **(i')** additionally modifies the argument for **(i)** to accommodate our likelihood bound (Appendix E). ■

So far, our regret guarantees are only about estimation in squared error (Decision Problem 1). In the next subsection, we analyze regret for empirical Bayes decision rules targeted to the ranking-related problems (Decision Problems 2 and 3), and relate their performances to those for Decision Problem 1.

3.2. **Other decision objectives and relation to squared-error loss.** Notably, the oracle Bayes decision rules $\boldsymbol{\delta}^\star$ in Decision Problems 2 and 3 depend solely on the vector of oracle Bayes posterior means $\theta^*_{1:n}$.[35] Therefore, for these problems, the natural empirical Bayes decision rules simply replace oracle Bayes posterior means (θ_i^*) with empirical Bayes ones ($\hat{\theta}_i$) in the oracle decision rules.[36] For instance, if one is comfortable with the prior estimated by CLOSE-NPMLE, then the corresponding decision rules for Decision Problems 2 and 3 threshold based on estimated posterior means under CLOSE-NPMLE.

In these problems, BayesRegret_n (2.13) is equal to the expected risk gap between using the feasible decision rules $\hat{\boldsymbol{\delta}}$ and the oracle decision rules $\boldsymbol{\delta}^\star$. To specialize, we let UMRegret_n

[35]In principle, one could consider many other policy problems with a ranking flavor (Koenker and Gu, 2019; Kline et al., 2023). Among these problems, UTILITY MAXIMIZATION BY SELECTION and TOP-m SELECTION are special in that optimal decisions are simple functions of the posterior means. We caution that the worst-case regret rate for ranking-type problems without this property can be unfavorable—as Gu and Koenker (2023) put it, "inherently futile"—since their optimal decisions depend on functionals that are known to be difficult to estimate (i.e., they have logarithmic minimax rates of estimation, Pensky, 2017; Dedecker and Michel, 2013; Cai and Low, 2011), without stronger assumptions on the prior.
In general, the minimax squared error rate of estimating $\mathbb{E}[f(\theta)]$ is logarithmic, unless f is an analytic function, by an extension of the argument in Cai and Low (2011). Ranking-type problems often involve f of the form $f(\theta) = \mathbb{1}(\theta > c)$ or $f(\theta) = \max(\theta, c)$, which are not smooth. This observation suggests that these ranking-type problems may also suffer from logarithmic regret rates—though, this observation alone does not rigorously prove this, as difficulties in estimating $\mathbb{E}f(\theta)$ *in squared error* might not preclude a polynomial regret rate for these ranking-type problems.

[36]Theorem 3 applies to any estimators of the oracle Bayes posterior means—not necessarily derived through an empirical Bayes procedure—and does not impose the location-scale assumption. As a result, it may be of independent interest.

denote BayesRegret_n for Decision Problem 2 and we let $\text{TopRegret}_n^{(m)}$ denote BayesRegret_n for Decision Problem 3. The following result relates UMRegret_n and $\text{TopRegret}_n^{(m)}$ to MSERegret_n.

Theorem 3. *Suppose (2.3) holds, but (2.6) may or may not hold. Let $\hat{\delta}_i$ be the plug-in decisions with any vector of estimates $\hat{\theta}_i$, not necessarily from* CLOSE-NPMLE. *We have the following inequalities on the expected regret corresponding to the decision rules $\hat{\delta}_i$:*

(1) For UTILITY MAXIMIZATION BY SELECTION,

$$\mathbb{E}[\text{UMRegret}_n] \le \left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i^*)^2\right]\right)^{1/2}. \tag{3.7}$$

(2) For TOP-m SELECTION,

$$\mathbb{E}[\text{TopRegret}_n^{(m)}] \le 2\sqrt{\frac{n}{m}}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i^*)^2\right]\right)^{1/2}. \tag{3.8}$$

Theorem 3 shows that the two decision problems UTILITY MAXIMIZATION BY SELECTION and TOP-m SELECTION are easier than estimating the oracle Bayesian posterior means. As a result, our convergence rates from Theorem 1 and Corollary 1 also upper bound regret rates for these two decision problems, rendering the regret rates more immediately useful for policy problems. In particular, for $m/n \asymp 1$, both regret rates (3.7) and (3.8) are of the form $n^{-p/(2p+1)}(\log n)^c = o(1)$ under Corollary 1. Thus, the performance of the empirical Bayes decision rule approximates that of the oracle with at least the rate $O(n^{-p/(2p+1)})$ up to log factors.

Remark 5 (Mover interpretation of Theorem 3)**.** Recall that we can think of TOP-m SELECTION as the decision problem in Bergman et al. (2023). The utility function represents the expected mobility of a mover, assuming that the mover moves randomly into one of the high mobility Census tracts. Our proof of Theorem 3 in Appendix A.2 allows for a slightly more general decision problem. Suppose the decision now is to provide a full ranking of Census tracts for potential movers and maximize the expected mobility for a mover. Suppose that the probability that a mover moves to a tract depends decreasingly and solely on the tract's rank. To be more concrete, suppose the mover has probability π_1 of moving to the highest-ranked tract, π_2 to the second-highest, and so forth. Then, with the same argument, the corresponding regret is dominated by $2\sqrt{n\sum_{i=1}^n \pi_i^2} \cdot \left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n(\hat{\theta}_i - \theta_i^*)^2\right]\right)^{1/2}$, which generalizes (3.8). ■

Remark 6 (Tightness of Theorem 3)**.** We suspect that the actual performance of CLOSE-NPMLE for Decision Problems 2 and 3 may be better than predicted by Theorem 3. Take the bound for UMRegret_n, for instance. As would be clear from the proof, the bound (3.7) holds even when the c_i's are adversarially chosen[37] such that the empirical Bayesian makes every mistake: $\hat{\delta}_i \neq \delta_i^\star$ for every i. However, for a fixed vector c, we expect that $\hat{\delta}_i \neq \delta_i^\star$ only for a vanishing fraction of populations, and thus the actual performance of $\hat{\delta}_i$ may be better than the rate in Appendix A.2 implies.[38]

Though we conjecture that the rate in Theorem 3 does not match a lower bound, Theorem 3 is competitive with recent results. TOP-m SELECTION is recently studied by Coey and Hung (2022), who show that under prior independence, if $\hat{\theta}_{1:n}$ are posterior means for some estimate $\hat{G}$ of the prior $G_{(0)}$, then

$$\mathbb{E}[\text{TopRegret}_n^{(m)}] = O\left(W_1^2(G_{(0)}, \hat{G})\right)$$

where $W_1(P, Q)$ is the Wasserstein-1 distance between P, Q. Theorem 3 attains a worse rate in parametric settings, when the prior $G_{(0)}$ can be estimated at fast rates. However, in nonparametric settings, $G_{(0)}$ is often only estimable at logarithmic rates (Dedecker and Michel, 2013), and thus the rate in Theorem 3 is much more favorable in those settings. ■

3.3. **Robustness to the location-scale assumption** (2.6)**.** We prove our regret upper and lower bounds imposing the location-scale model (2.6). This is an optimistic assessment of the performance of CLOSE-NPMLE. While (2.6) nests prior independence, it may still be misspecified. We now consider the worst-case behavior of CLOSE-NPMLE without the location-scale assumption. Since without the location-scale assumption, CLOSE-NPMLE can no longer hope to emulate the oracle Bayes decisions, we focus on worst-case Bayes *risk* here, instead of on regret.

We will do so by considering an idealized version of the procedure. So long as $\theta_i \mid \sigma_i$ has two moments, $\eta_0(\cdot) = (m_0(\cdot), s_0(\cdot))$ are well-defined as conditional moments of $\theta_i \mid \sigma_i$ without imposing the location-scale assumption. We will assume that m_0, s_0 are known. Without the location-scale model, G_0 is ill-defined, but we will assume that we obtain some pseudo-true value G_0^* that has zero mean and unit variance. This is a reasonable condition to impose,

[37] That said, if the c_i's are indeed adversarially chosen given knowledge of $(Y_{1:n}, \sigma_{1:n}, P_0)$, then Theorem 3 does match a corresponding lower bound, derived by choosing $c_i = (\hat{\theta}_i + \theta_i^\star)/2$.

[38] Upper and lower bounds are derived in related but distinct settings by Audibert and Tsybakov (2007); Bonvini et al. (2023); Liang (2000); some upper bounds, under possibly stronger assumptions, appear better than implied by Appendix A.2.

For UTILITY MAXIMIZATION BY SELECTION, suppose we impose a margin condition of the form

$$\text{For all } i,\ \mathrm{P}(|\theta_i^* - c_i| \le t) \lesssim t^\xi \quad \xi \in (0, \infty), t \in (0, c_0]$$

where if θ_i^* has (uniform-in-i) bounded density around c_i, then ξ can be taken to be 1. Proposition 2 in Bonvini et al. (2023) then yields the sharper result that

$$\text{UMRegret}_n \lesssim_\xi \frac{1}{n}\sum_{i=1}^n \mathbb{E}[(\hat{\theta}_i - \theta_i)^2] \le \left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2\right]\right)^{\frac{1}{2}+\frac{1}{2}\frac{\xi}{2+\xi}}.$$

Further applications of Audibert and Tsybakov (2007) and Bonvini et al. (2023) to the Gaussian sequence setting remain open.

since every conditional prior distribution $\tau_i \mid \sigma_i$ obeys this moment constraint.[39] Thus, for estimating $\tau_i = \frac{\theta_i - m_0(\sigma_i)}{s_0(\sigma_i)}$, whose true prior is $\tau_i \mid \sigma_i \sim G_i$, this idealized procedure uses some misspecified prior $G_0^* \neq G_i$, which does have the correct first two moments.

Using results we develop in a related note (Chen, 2023), we show that this idealized procedure has maximum risk within a constant factor of the minimax risk, uniformly over η_0. The minimax risk here is defined with respect to a game where the analyst knows m_0, s_0 and an adversary chooses the shape of the distribution $\tau_i \mid \sigma_i$ for every i.

Theorem 4. *Under (2.3) but not (2.6), assume the conditional distribution $\theta_i \mid \sigma_i$ has mean $m_0(\sigma_i)$ and variance $s_0^2(\sigma_i)$. Denote the set of distributions of $\theta_{1:n} \mid \sigma_{1:n}$ which obey these restrictions as $\mathcal{P}(m_0, s_0)$. Let $\hat{\theta}_{i,G_0^*,\eta_0}$ denote the posterior mean estimates with some prior P^* under the location-scale model $P^*(\theta_i \le t \mid \sigma_i) = G_0^*\left(\frac{t-m_0(\sigma_i)}{s_0(\sigma_i)}\right)$, for some fixed G_0^* with zero mean and unit variance. Let $\overline{\rho} = \max_i s_0^2(\sigma_i)/\sigma_i^2 < \infty$ be the maximal conditional signal-to-noise ratio and assume that it is bounded. Then, for some $C_{\overline{\rho}} < \infty$ that solely depends on $\overline{\rho}$,*

$$\sup_{P_0 \in \mathcal{P}(m_0,s_0)} \mathbb{E}_{P_0}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_{i,G_0^*,\eta_0} - \theta_i)^2\right] \le C_{\overline{\rho}} \cdot \inf_{\hat{\theta}_{1:n}} \sup_{P_0 \in \mathcal{P}(m_0,s_0)} \mathbb{E}_{P_0}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2\right]. \quad (3.9)$$

where the infimum on the right-hand side is over all (possibly randomized) estimators of θ_i given $(Y_i, \sigma_i)_{i=1}^n$ and $\eta_0(\cdot)$.

Theorem 4 shows that the worst-case behavior of an idealized version of CLOSE-NPMLE must come within a factor of the minimax risk and hence is not arbitrarily unreasonable, even under misspecification. We caution that (3.9) is a fairly weak guarantee, in that the decision rule that simply outputs the prior conditional mean ($\delta_i = \hat{\theta}_{i,\boldsymbol{\delta}_0,\eta_0} = m_0(\sigma_i)$) also satisfies it. Nevertheless, even so, (3.9) *does not* hold for the idealized version of INDEPENDENT-GAUSS, plugging in known unconditional moments $m_0 = \frac{1}{n}\sum_{i=1}^n m_0(\sigma_i)$ and $s_0^2 = \frac{1}{n}\sum_{i=1}^n (m_0(\sigma_i) - m_0)^2 + s_0^2(\sigma_i)$.[40] To provide additional reassurance for CLOSE-NPMLE under misspecification, Appendix A.7.3 discusses an interpretation of CLOSE-NPMLE under misspecification of (2.6), and the validation procedure developed in Section 4.3 provides unbiased evaluation without relying on the location-scale model.

4. **Practical considerations**

4.1. **A detailed recipe.** We now describe the implementation of CLOSE-NPMLE in more detail, following our previous outline in **CLOSE–STEP 1** to **CLOSE–STEP 3**.

[39]We do not know if the maximizer G of the population analogue to (2.11) respects the moment constraints. In any case, imposing these moment constraints computationally in NPMLE is feasible, as they are simply linear constraints over the optimizing variables. Projecting the estimated $\hat{G}_n$ to these moment constraints makes little difference in our empirical exercise (Appendix B.2).

[40]To wit, take $s_0(\sigma_i) \approx 0$. Then, the minimax risk as a function of $(s_0(\cdot), m_0(\cdot))$ is approximately zero, but $m_0(\cdot)$ can be chosen such that the risk of INDEPENDENT-GAUSS is bounded away from zero.

The first step $\boxed{\textbf{CLOSE–STEP 1}}$ estimates the conditional moments $\eta_0 = (m_0, s_0)$ nonparametrically. Since the two conditional moments can be written as conditional expectations

$$\begin{aligned} m_0(\sigma) &= \mathbb{E}[\theta \mid \sigma] = \mathbb{E}[Y \mid \sigma] \\ s_0^2(\sigma) &= \mathrm{Var}(\theta \mid \sigma) = \mathbb{E}[(Y - m_0(\sigma))^2 \mid \sigma] - \sigma^2, \end{aligned} \tag{4.1}$$

we can estimate them accordingly with off-the-shelf methods (e.g., local polynomial kernel smoothing methods implemented by Calonico et al., 2019). Specifically, estimating m_0 with $\hat{m}$ is directly a nonparametric regression of Y_i on σ_i.[41] Estimating $s_0^2(\cdot)$ can be operationalized by first nonparametrically regressing $(Y_i - \hat{m}(\sigma_i))^2$ on σ_i, and then subtracting off σ_i^2. This is a plug-in estimator for s_0^2, as it replaces quantities in (4.1) with their empirical counterparts.[42]

A wrinkle is that the plug-in estimate $\hat{s}$ may be negative.[43] Truncating $\hat{s}$ at zero results in observations whose estimated prior variances $\hat{s}^2(\sigma_i) = 0$. These observations also have implied $\hat{\nu}_i = \infty$. For these observations, an empirical Bayesian taking $\hat{s}^2(\sigma_i) = 0$ at face value has degenerate priors at $\hat{m}(\sigma_i)$. Since observations with $\nu_i = \infty$ do not contribute to the likelihood objective for NPMLE, excluding them from the NPMLE computation does not alter the estimated $\hat{G}_n$. Thus, we can continue to use $(\hat{m}, \hat{s}^2, \hat{G}_n)$ as the estimated posterior—an observation with $\hat{s}^2(\sigma_i) = 0$ would have a point mass at $\hat{m}(\sigma_i)$ as its estimated posterior. In our experience, this simple approach does not appear to affect performance. Nevertheless, in Appendix G, we propose a heuristic but data-driven truncation rule, borrowing from a statistics literature on estimating non-centrality parameters for non-central χ^2 distributions (Kubokawa et al., 1993). Appendix G also discusses tuning parameter selection for estimating (m_0, s_0) and verifies that our local linear regression estimators satisfy the regularity conditions in Section 3.

Next, in the second step $\boxed{\textbf{CLOSE–STEP 2}}$, we form the transformed estimates $\hat{Z}_i = \frac{Y_i - \hat{m}(\sigma_i)}{\hat{s}(\sigma_i)}$ and the transformed standard errors $\hat{\nu}_i = \sigma_i / \hat{s}(\sigma_i)$. We then estimate the NPMLE on the data $(\hat{Z}_i, \hat{\nu}_i)$ by maximizing (2.11). In practice, the infinite-dimensional optimization problem (2.11) is approximated with a finite-dimensional one by discretizing distributions on a grid. To be precise, let $\min_i \hat{Z}_i = \tau_{(1)} \le \cdots \le \tau_{(J)} = \max_i \hat{Z}_i$ be a pre-specified grid of points, not

[41] We take $\log(\sigma_i)$ in our empirical implementation since the distribution of σ_i tends to be right-skewed, and thus we suspect regressing on $\log(\sigma_i)$ has a better fit.

[42] Since (4.1) can be written in different forms, there are other reasonable plug-in estimators for s_0. We investigate one such alternative estimator in Appendix B.2 and find very similar performance in our empirical exercise.

[43] The negative estimated variance phenomenon similarly may occur with estimating the prior variance with INDEPENDENT-GAUSS and with conditional variance estimation in Armstrong et al. (2022). This is in part caused by estimation noise in $\mathrm{Var}(Y_i \mid \sigma_i)$. However, there is some evidence that observations with large estimated σ_i's are underdispersed for the measures of economic mobility in the Opportunity Atlas (see Appendix B.1.)

necessarily equally spaced, and denote it by $\boldsymbol{\tau}$.[44] The feasible version of (2.11) maximizes the concave program $\pi^\star \equiv \max_{\pi \in \mathbb{R}^J_{\geq 0}, \pi'1=1} \sum_{i=1}^n \log \left(\sum_{j=1}^J \pi_j \varphi \left(\frac{\hat{Z}_i - \tau_{(j)}}{\hat{\nu}_i} \right) \right)$. The estimated NPMLE $\hat{G}_n$ is a discrete distribution with support points $\tau_{(j)}$ and corresponding masses $\pi_j^\star$.

Finally, given the estimate $\hat{G}_n = (\boldsymbol{\tau}, \pi^\star)$, we can compute empirical Bayes decision rules and implement CLOSE–STEP 3 by minimizing posterior expected loss. Since $\hat{G}_n$ is a discrete distribution, the posterior for τ_i is given by the probability mass function

$$\mathrm{P}_{\hat{G}_n}(\tau_i = \tau_{(j)} \mid \hat{Z}_i = z, \hat{\nu}_i = \nu) \propto \pi_j^\star \exp\left(-\frac{1}{2\nu^2}(z - \tau_{(j)})^2 \right),$$

normalized so that the probabilities sum to 1. This probability mass function can be plugged into (2.5) to compute the empirical Bayes decision rule for any loss function L.[45]

4.2. **When does relaxing prior independence matter?** When prior independence holds, CLOSE-NPMLE is the same as INDEPENDENT-NPMLE, up to the estimation of the constant conditional moments $(m_0(\cdot), s_0(\cdot))$. Since CLOSE-NPMLE has to estimate the conditional moments, we expect it to underperform INDEPENDENT-NPMLE, though not by much in large samples.

When prior independence does not hold, but when the conditional location-scale model (2.6) approximately holds, we expect CLOSE to outperform methods that assume prior independence. Qualitatively speaking, we expect the improvement of CLOSE-methods to be large when the conditional expectation accounts for large portions of the unconditional signal variance $\mathrm{Var}(\theta_i)$. Since we can decompose $\mathrm{Var}(\theta_i) = \mathbb{E}[s_0^2(\sigma_i)] + \mathrm{Var}(m_0(\sigma_i))$, we expect the improvement of CLOSE-methods to be large when the variance of the conditional expectation $\mathrm{Var}(m_0(\sigma_i))$ is large compared to $\mathbb{E}[s_0^2(\sigma_i)]$. Intuitively, this is the case when σ_i is highly predictive of θ_i. Whether this is the case can be easily checked by plotting Y_i against σ_i, as in Figure 1, and inspecting the estimated conditional moments.

Finally, when the conditional distributions $\theta_i \mid \sigma_i$ are non-Gaussian, and in particular when they are discrete, skewed, or thick-tailed, we expect CLOSE-NPMLE to additionally outperform INDEPENDENT-GAUSS due to not assuming Normality of θ_i. When the conditional priors are Gaussian, estimating it via the NPMLE pays a modest statistical price. Admittedly, it is often difficult to diagnose whether the underlying conditional distributions $\theta_i \mid \sigma_i$ have these

[44]Since the gridding is a computational approximation to the infinite dimensional optimization problem, the sole downside of a finer grid is computational burden (cf. bias-variance tradeoffs in typical tuning parameter selection problems). Ideally, adjacent grid points should have a sufficiently small and economically insignificant gap between them. Since the true prior G_0 for τ_i have zero mean and unit variance, we find that a fine grid within $[-6, 6]$ (e.g., 400 equally spaced grid points), with a coarse grid on $[\min_i \hat{Z}_i, \max_i \hat{Z}_i] \backslash [-6, 6]$ (e.g., 100 equally spaced grid points), performs well. Also see recommendations in Koenker and Gu (2017) and Koenker and Mizera (2014).

[45]In the leading use-case, the posterior means for θ_i are simply $\hat{m}(\sigma_i) + \hat{s}(\sigma_i)\mathbf{E}_{\hat{G}_n, \hat{\nu}_i}[\tau_i \mid \hat{Z}_i, \hat{\nu}_i]$. In practice, REBayes::GLmix (Koenker and Gu, 2017) in R implements estimation of the NPMLE and computation of the posterior means $\mathbf{E}_{\hat{G}_n, \hat{\nu}_i}[\tau_i \mid \hat{Z}_i, \hat{\nu}_i]$.

properties, since we only observe (Y_i, σ_i). Likewise, so far the discussion in this subsection is heuristic. To be more certain of the extent of improvement of CLOSE-NPMLE over other methods, it is helpful to have out-of-sample validation. The next subsection proposes a minor extension of Oliveira et al. (2021), which allows for an unbiased estimate of loss and serves as a validation procedure.

4.3. **A formal validation procedure via coupled bootstrap.** Consider (Y_i, σ_i) where $Y_i \mid \sigma_i, \theta_i \sim \mathcal{N}(\theta_i, \sigma_i^2)$. For some $\omega > 0$ and an independent Gaussian noise $W_i \sim \mathcal{N}(0,1)$, consider adding to Y_i and subtracting from Y_i some scaled version of W_i:

$$Y_i^{(1)} = Y_i + \sqrt{\omega}\sigma_i W_i \quad Y_i^{(2)} = Y_i - \frac{1}{\sqrt{\omega}}\sigma_i W_i.$$

Oliveira et al. (2021) call $(Y_i^{(1)}, Y_i^{(2)})$ the *coupled bootstrap* draws. Observe that the two draws are conditionally independent:

$$\begin{bmatrix} Y_i^{(1)} \\ Y_i^{(2)} \end{bmatrix} \mid \theta_i, \sigma_i^2 \sim \mathcal{N}\left(\begin{bmatrix} \theta_i \\ \theta_i \end{bmatrix}, \begin{bmatrix} (1+\omega)\sigma_i^2 & 0 \\ 0 & (1+\omega^{-1})\sigma_i^2 \end{bmatrix} \right). \tag{4.2}$$

The conditional independence allows us to use $Y_i^{(2)}$ as an out-of-sample validation for decision rules computed based on $Y_i^{(1)}$. We denote their variances by $\sigma_{i,(1)}^2$ and $\sigma_{i,(2)}^2$.

It is helpful to think of $Y_i^{(1)}$ as training data and $Y_i^{(2)}$ as testing data. In fact, the coupled bootstrap precisely emulates sample-splitting on the micro-data. To see that, suppose $Y_i = \frac{1}{n_i}\sum_{j=1}^n Y_{ij}$ is a sample mean of i.i.d. micro-data $\{Y_{ij} : j = 1, \ldots, n_i\}$. Suppose we split the micro-data $\{Y_{ij} : j = 1, \ldots, n_i\}$ into a training set and a testing set, with proportions $\frac{1}{\omega+1}$ and $\frac{\omega}{\omega+1}$, respectively. Let $Y_i^{(1)}$ and $Y_i^{(2)}$ be the training and testing set sample means, respectively. Then the central limit theorem implies that, approximately,

$$Y_i^{(1)} \mid \theta_i, \sigma_i^2 \sim \mathcal{N}\left(\theta_i, (1+\omega)\sigma_i^2\right) \quad Y_i^{(2)} \mid \theta_i, \sigma_i^2 \sim \mathcal{N}\left(\theta_i, (1+\omega^{-1})\sigma_i^2\right) \tag{4.3}$$

independently. Note that the two representations (4.2) and (4.3) are equivalent, and hence coupled bootstrap emulates sample-splitting. For instance, coupled bootstrap with a value of $\omega = 1/9$ is statistically equivalent to splitting the micro-data with a 90-10 train-test split.

Just as we can perform out-of-sample validation with sample-splitting on the micro-data, we can also do so with the coupled bootstrap emulation of sample-splitting. The following proposition formalizes this and states unbiased estimators for the loss of these decision rules, as well as their accompanying standard errors.[46]

[46] Oliveira et al. (2021) state the unbiased estimation result for the mean-squared error estimation problem. They develop the result further by connecting the coupled bootstrap estimator to Stein's unbiased risk estimate. Our analogous calculation for other loss functions and for the standard errors is a minor extension of their results. Proposition 1 can also be easily generalized to other loss functions that admit unbiased estimators (Effectively, the loss is a function of a Gaussian location θ_i. For unbiased estimation of functions of Gaussian parameters, see Table A1 in Voinov and Nikulin, 2012).

Table 1. Unbiased estimators for loss of decision rules and associated conditional variance expressions (Proposition 1)

Problem	Unbiased estimator of loss, $T\left(Y_{1:n}^{(2)}, \boldsymbol{\delta}\right)$	$\operatorname{Var}\left(T\left(Y_{1:n}^{(2)}, \boldsymbol{\delta}\right) \mid \mathcal{F}\right)$
Decision Problem 1	$\frac{1}{n}\sum_{i=1}^n \left(Y_i^{(2)} - \delta_i(Y_{1:n}^{(1)})\right)^2 - \sigma_{i,(2)}^2$	$\frac{1}{n^2}\sum_{i=1}^n \operatorname{Var}\left((Y_i^{(2)} - \delta_i(Y_{1:n}^{(1)}))^2 \mid \mathcal{F}\right)$
Decision Problem 2	$-\frac{1}{n}\sum_{i=1}^n \delta_i(Y_{1:n}^{(1)})(Y_i^{(2)} - c_i)$	$\frac{1}{n^2}\sum_{i=1}^n \delta_i(Y_{1:n}^{(1)})\sigma_{i,(2)}^2$
Decision Problem 3	$-\frac{1}{m}\sum_{i=1}^n \delta_i(Y_{1:n}^{(1)})Y_i^{(2)}$	$\frac{1}{m^2}\sum_{i=1}^n \delta_i(Y_{1:n}^{(1)})\sigma_{i,(2)}^2$

Proposition 1. *Suppose (Y_i, σ_i) obey the Gaussian heteroskedastic location model, assumed to be independent across i (2.3). Fix some $\omega > 0$ and let $Y_{1:n}^{(1)}, Y_{1:n}^{(2)}$ be the coupled bootstrap draws. For some decision problem, let $\boldsymbol{\delta}(Y_{1:n}^{(1)})$ be some decision rule using only data $\left(Y_i^{(1)}, \sigma_{i,(1)}^2\right)_{i=1}^n$. Let $\mathcal{F} = \left(\theta_{1:n}, Y_{1:n}^{(1)}, \sigma_{1:n,(1)}, \sigma_{1:n,(2)}\right)$, for Decision Problems 1 to 3, the estimators $T(Y_{1:n}^{(2)}, \boldsymbol{\delta})$ displayed in Table 1 are unbiased for the corresponding loss:*

$$\mathbb{E}\left[T(Y_{1:n}^{(2)}, \boldsymbol{\delta}(Y_{1:n}^{(1)})) \mid \mathcal{F}\right] = L\left(\boldsymbol{\delta}(Y_{1:n}^{(1)}), \theta_{1:n}\right).$$

Moreover, their conditional variances are equal to those expressions displayed in the third column of Table 1.

Proposition 1 allows for an out-of-sample evaluation of decision rules, as well as uncertainty quantification around the estimate of loss, solely imposing the heteroskedastic Gaussian model. This is a useful property in practice for comparing different empirical Bayes methods. The alternative is to take some estimated prior—say the one learned by CLOSE-NPMLE—as the true prior, and evaluate performance of competing methods. Doing so likely tips the scale in favor of a particular method, and we advocate for the coupled bootstrap instead.

5. Empirical illustration

How does CLOSE-NPMLE perform in the field? We now consider two empirical exercises related to the Opportunity Atlas (Chetty et al., 2020) and Creating Moves to Opportunity (Bergman et al., 2023). We first summarize these papers.

5.1. The Opportunity Atlas and Creating Moves to Opportunity.

Chetty et al. (2020) and Bergman et al. (2023) are motivated by a growing literature in neighborhood effects on upward mobility. There is a large body of quasiexperimental evidence that the neighborhood a child grows up in has substantial causal effects on upward mobility (Chetty and Hendren, 2018; Chetty et al., 2016; Laliberté, 2021; Chyn and Katz, 2021). Consequently, social programs that encourage low-income families to move to better neighborhoods can potentially benefit upward mobility.

Such programs hinge on two economic questions and one econometric question. First, how do we measure neighborhood mobility? Second, are low-income families currently living in low-opportunity neighborhoods because they *prefer* some unobserved quality of these neighborhoods, or is it due to certain economic and informational barriers? Third, econometrically, given noisy measures of mobility, how do we identify high-mobility neighborhoods?

Motivated by the first question, Chetty et al. (2020) provide Census tract-level estimates of poor children's outcomes in adulthood and argue that these observational measures of mobility predict neighborhoods' causal effects. Motivated by the second question, Bergman et al. (2023) show that financial assistance and informational support do induce low-income families to move to neighborhoods that researchers recommend, indicating that these families indeed face barriers to moving to opportunity. The third question is naturally answered by empirical Bayes methods.

Specifically, using longitudinal Census micro-data, Chetty et al. (2020) estimate tract-level children's outcomes in adulthood and publish the estimates in a collection of datasets called the Opportunity Atlas. Each dataset contains estimates and standard errors for some particular definition of the economic parameter of interest, at the Census tract i level. Taking these estimates from the Opportunity Atlas, Bergman et al. (2023) conducted a program in Seattle called Creating Moves to Opportunity. They provided assistance to treated low-income individuals[47] to move to "Opportunity Areas"—Census tracts with empirical Bayes posterior means in the top third.[48] We view Bergman et al.'s (2023) objectives as TOP-m SELECTION (Decision Problem 3), for m equal to one third of the number of tracts in King County, Washington (Seattle).

The Opportunity Atlas also includes tract-level covariates, a complication that we have so far abstracted away from. In the ensuing empirical exercises—as well as in Bergman et al. (2023)—the estimates and parameters are residualized against the covariates as a preprocessing step. We now let $\tilde{Y}_i$ denote the raw Opportunity Atlas estimates for a pre-residualized parameter ϑ_i and let (Y_i, θ_i) be their residualized counterparts against a vector of tract-level covariates X_i, with regression coefficient β.[49] We can apply the empirical Bayes procedures in this paper to (Y_i, σ_i^2) and obtain an estimated posterior for θ_i. This estimated posterior for the residualized parameter θ_i then implies an estimated posterior for

[47]They are families with a child below age 15 who are issued Section 8 vouchers between April 2018 and April 2019, with median household income of \$19,000. About half of the sampled households are Black and about a quarter are white (Table 1, Bergman et al., 2023).

[48]There are also adjustments to make the selected tracts geographically contiguous. See Bergman et al. (2023) for details.

[49]Precisely speaking, let X_i be a vector of tract-level covariates. Let $\tilde{Y}_i$ be the raw Opportunity Atlas estimates of a parameter ϑ_i, with accompanying standard errors σ_i. Let β be some vector of coefficients, typically estimated by weighted least-squares of Y_i on X_i. Let $Y_i = \tilde{Y}_i - X_i'\beta$ and $\theta_i = \vartheta_i - X_i'\beta$ be the residuals. We assume that the tract-level covariates do not predict the estimation noise in $\tilde{Y}_i$: i.e., $X_i \perp\!\!\!\perp \tilde{Y}_i \mid \theta_i, \sigma_i^2$. Since β is precisely estimated, we ignore its estimation noise. Then, the residualized objects (Y_i, θ_i) obey the Gaussian location model $Y_i \mid \theta_i, \sigma_i \sim \mathcal{N}(\theta_i, \sigma_i^2)$. See additional discussion on covariates in Appendix A.6.2. Figure B.6 contains empirical results without residualizing against covariates.

the original parameter $\vartheta_i = \theta_i + X_i'\beta$, by adding back the fitted values $X_i'\beta$ (Fay and Herriot, 1979). When there are no covariates, $\vartheta_i = \theta_i$ and $Y_i = \tilde{Y}_i$.

We consider several measures of economic mobility ϑ_i. For our purposes, these definitions of ϑ_i take the following form: ϑ_i is the population mean of *some* outcome for individuals of *some* demographic subgroup growing up in tract i, whose parents are at the 25$^{\text{th}}$ income percentile. We will consider three types of outcomes:

(1) Percentile rank of adult income
(2) An indicator for whether the individual has incomes in the top 20 percentiles
(3) An indicator for whether the individual is incarcerated

for the following demographic subgroups:[50] (1) all individuals (POOLED), (2) white individuals, (3) white men, (4) Black individuals, and (5) Black men. As shorthands, we refer to the three types of outcomes as MEAN RANK, TOP-20 PROBABILITY, and INCARCERATION, respectively. The outcome we use in Section 2 corresponds to TOP-20 PROBABILITY for Black individuals, while Bergman et al. (2023) consider MEAN RANK POOLED.[51]

The remainder of this section compares several empirical Bayes approaches on two exercises. The first exercise is a calibrated simulation. In the simulation, we compare MSE performance of various methods to the that of the oracle posterior. We find that CLOSE-NPMLE has near-oracle performance in terms of MSE, and substantially outperforms INDEPENDENT-GAUSS. The second exercise is an empirical application to a scale-up of the exercise in Bergman et al. (2023). It uses the coupled bootstrap to evaluate whether CLOSE-NPMLE selects more economically mobile tracts than INDEPENDENT-GAUSS. We find that it does, and the magnitude of improvement is substantial compared to two benchmarks, which we refer to as the value of basic empirical Bayes methods and the value of data.

5.2. **Calibrated simulation.** Our first empirical exercise is a calibrated simulation. To devise a data-generating process that does not impose the location-scale assumption, we partition σ into vingtiles, fit a location-scale model within each vingtile, and draw from the estimated model (see Appendix B.3 for details). Since the location-scale model is only imposed within each vingtile, this data-generating process does not impose (2.6) on the entire dataset. Figure 3 shows an overlay of real and simulated data for one of the variables we consider. Visually, at least, the simulated data resemble the real estimates.

[50]We focus on men as a subgroup since incarceration rates for women are extremely low.

[51]In each Opportunity Atlas dataset, the estimates $\tilde{Y}_i, \sigma_i$ are computed from the fitted value of a semiparametric regression procedure on the Census micro-data. The regression procedure implicitly pools observation with similar parent income ranks and is not fully nonparametric. As a result of this extrapolation, the estimates Y_i need not respect support conditions for Bernoulli means. For instance, some estimates for TOP-20 PROBABILITY and for INCARCERATION are negative. Similarly, the standard errors for estimates for binarized ϑ_i are typically not precisely of the form $\sqrt{\vartheta_i(1-\vartheta_i)/n_i}$. We refer interested readers to Chetty et al. (2020) for details of their regression specification.

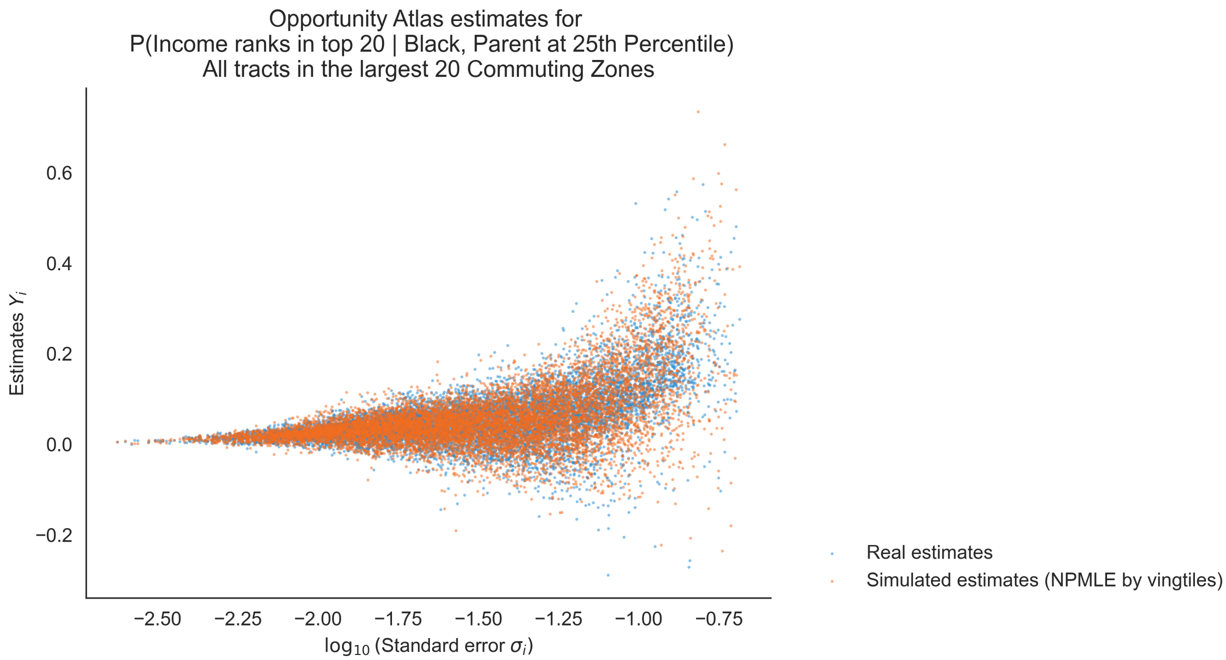


FIGURE 3. A draw of real vs. simulated data for estimates of TOP-20 PROBABILITY for Black individuals

On the simulated data, we then put various empirical Bayes strategies to test. We consider the feasible procedures NAIVE, INDEPENDENT-GAUSS, INDEPENDENT-NPMLE, CLOSE-GAUSS, and CLOSE-NPMLE, where NAIVE sets $\hat{\theta}_i = Y_i$.[52] Because we have the ground truth data-generating process, we additionally have two infeasible benchmarks:

- ORACLE: A Bayesian who has access to the distribution of (θ_i, σ_i) and uses the true posterior means for θ_i.[53]
- ORACLE-GAUSS: A Bayesian who knows (m_0, s_0) and uses (2.12).

For this exercise, we focus on estimating the parameters ϑ_i in MSE (Decision Problem 1).

Figure 4 plots the main results from this calibrated simulation. For each method and each target variable, we display a relative measure of gain in terms of mean-squared error. For each method, we calculate its squared error gain over NAIVE, as a percentage of the squared error gain of ORACLE over NAIVE. If we think of the ORACLE–NAIVE difference as the total size of the "statistical pie," then Figure 4 shows how much of this pie each method captures.

[52]We note that none of the feasible procedures (NAIVE, INDEPENDENT-GAUSS, INDEPENDENT-NPMLE, CLOSE-GAUSS, and CLOSE-NPMLE) have access to the true projection coefficient β of $\tilde{Y}_i$ onto X_i, which they must estimate by residualizing against covariates on the data. Additionally, we weight the estimation of m_0 and s_0 in INDEPENDENT-GAUSS by the precision $1/\sigma_i^2$, following Bergman et al. (2023).

[53]These posterior means are computed by approximating the true prior with the empirical distribution of a large sample drawn from the true prior.

What % of Naive-to-Oracle MSE gain do we capture?

	Indep-Gauss (No residualization)	Indep-NPMLE (No residualization)	CLOSE-Gauss (No residualization)	CLOSE-NPMLE (No residualization)	Indep-Gauss	Indep-NPMLE	CLOSE-Gauss	Oracle-Gauss	CLOSE-NPMLE
Mean income rank	-4	25	49	50	85	88	91	91	91
Mean income rank [white]	55	60	66	66	87	90	94	95	95
Mean income rank [Black]	30	61	87	87	82	88	93	94	93
Mean income rank [white male]	63	69	74	75	89	92	93	94	95
Mean income rank [Black male]	32	54	86	87	83	86	93	93	94
P(Income ranks in top 20)	-160	9	67	67	57	81	91	93	93
P(Income ranks in top 20 | white)	31	51	65	65	75	80	94	97	95
P(Income ranks in top 20 | Black)	-6	24	93	95	46	53	95	97	97
P(Income ranks in top 20 | white male)	23	46	71	72	70	76	90	94	94
P(Income ranks in top 20 | Black male)	-8	21	94	96	37	45	95	97	97
Incarceration	-5	32	68	68	51	59	88	95	91
Incarceration [white]	61	72	90	96	74	81	91	93	97
Incarceration [Black]	42	51	94	95	48	52	96	98	97
Incarceration [white male]	43	53	92	96	60	64	93	95	98
Incarceration [Black male]	25	42	90	90	42	49	96	99	96
Column median	30	51	86	87	70	80	93	95	95

Notes. Each column is an empirical Bayes strategy that we consider, and each row is a different definition of ϑ_i. The table shows relative performance, defined as the squared error improvement over NAIVE, normalized as a percentage of the improvement of ORACLE over NAIVE. That is, if we think of going to ORACLE from NAIVE as the total extent of risk gains via empirical Bayes methods, this relative performance denotes how much of those gains each method captures. The last row shows the column median. Since we rely on Monte Carlo approximations of ORACLE, the resulting Monte Carlo error causes CLOSE-NPMLE to outperform ORACLE in the top right. Results are averaged over 1,000 Monte Carlo draws. For absolute, un-normalized performance of INDEPENDENT-GAUSS, INDEPENDENT-NPMLE, CLOSE-NPMLE, and ORACLE, see Figure B.10. □

FIGURE 4. Table of relative squared error Bayes risk for various empirical Bayes approaches

A value of 70 in Figure 4, for instance, indicates that a particular method captures 70% of the possible extent of risk gains for a particular parameter definition.

The first four columns show the relative mean-squared error performance *without* residualizing against covariates, applying empirical Bayes methods directly on $(\tilde{Y}_i, \sigma_i)$. We see that methods which assume prior independence—INDEPENDENT-GAUSS and INDEPENDENT-NPMLE—perform worse than methods based on CLOSE.[54] Across the 15 variables, the median

[54]It may be surprising that INDEPENDENT-GAUSS can perform worse than NAIVE on MSE, since Gaussian empirical Bayes typically has a connection to the James–Stein estimator, which dominates the MLE. We note that, as in Bergman et al. (2023), when we estimate the prior mean and prior variance, we *weight* the data with precision weights proportional to $1/\sigma_i^2$. When the independence between θ and σ holds, these

proportion of possible gains captured by INDEPENDENT-GAUSS is only 30%. This value is 51% for INDEPENDENT-NPMLE, and 87% for CLOSE-NPMLE. Individually for each variable, among the first four columns, CLOSE-NPMLE uniformly dominates all three other methods. This is because the standard error σ_i contains much of the predictive power of the covariates, and using that information can be very helpful when the researcher does not have rich covariate information.

The next five columns show performance when the methods do have access to covariate information. Compared to their no-covariates counterparts, the methods that assume prior independence do substantially better, since the covariates absorb some dependence between ϑ_i and σ_i. For MEAN RANK, after covariate residualization, there appears to be little dependence between θ_i and σ_i. INDEPENDENT-NPMLE and CLOSE-NPMLE perform similarly, capturing almost all of the available gains. Both methods slightly outperform the Gaussian methods for MEAN RANK.[55]

For the other two outcome variables, TOP-20 PROBABILITY and INCARCERATION, the dependence between θ_i and σ_i is stronger, and CLOSE-based methods display substantial improvements over INDEPENDENT-GAUSS and INDEPENDENT-NPMLE. CLOSE-NPMLE achieves near-oracle performance across the different definitions of θ_i (capturing a median of 95% of the ORACLE-NAIVE gap), and uniformly dominates all other feasible methods.

So far, we have tested the methods in a synthetic environment set up to imitate the real data. Next, we turn to an empirical application that uses the coupled bootstrap (Section 4.3) estimator of performance.

5.3. **Validation exercise via coupled bootstrap.** Our second empirical exercise uses the coupled bootstrap described in Section 4.3 for the ranking policy problem in Bergman et al. (2023). Throughout, we choose ω to emulate a 90-10 train-test split on the micro-data.

Bergman et al. (2023) use empirical Bayes methods to select the top third Census tracts in Seattle, based on economic mobility—which we view as a TOP-m SELECTION problem (Decision Problem 3). Can CLOSE-NPMLE make better selections—can it select tracts with higher ϑ_i on average? Specifically, we imagine scaling up Bergman et al. (2023)'s exercise and perform INDEPENDENT-GAUSS and CLOSE-NPMLE for all Census tracts in the largest twenty Commuting Zones. We then select the top third of tracts *within* each Commuting Zone, according to empirical Bayesian posterior means for ϑ_i. Additionally, to faithfully mimic

precision weights typically improve efficiency. However, the weighting does break the connection between Gaussian empirical Bayes and James–Stein, and the resulting posterior mean does not always dominate the MLE (i.e., NAIVE). To take an extreme example, if a particular observation has $\sigma_i \approx 0$, then that observation is highly influential for the prior mean estimate. If $\mathbb{E}[\theta_i \mid \sigma_i]$ is very different for that observation than the other observations, then the estimated prior mean is a bad target to shrink towards.

[55] Appendix B.4 contains an alternative data-generating process in which the true prior is Weibull, which has thicker tails and higher skewness. Under such a scenario, NPMLE-based methods substantially outperform methods assuming Gaussian priors.

Bergman et al. (2023), here we perform all empirical Bayes procedures *within Commuting Zone*. That is, for each of the twenty Commuting Zones that we consider, we execute all empirical Bayes methods—including the residualization by covariates—with only $\tilde{Y}_i, \sigma_i$ of tracts within the Commuting Zone.[56]

Figure 5(a) shows the estimated performance gap between a given empirical Bayes method and NAIVE as the x-position of the dots. The estimated performance of each method,[57] defined as the average ϑ_i among those selected (2.14), is shown in the annotated figures. According to these estimates, CLOSE-NPMLE generally improves over INDEPENDENT-GAUSS.[58]

For the MEAN RANK variables, using CLOSE-NPMLE generates substantial gains for mobility measures for Black individuals (0.8 percentile ranks for Black men and 0.5 percentile ranks for Black individuals). To put these gains in dollar terms, the Housing Choice Voucher holders in Bergman et al. (2023) have incomes around \$19,000, and for these individuals, an incremental percentile rank amounts to about \$1,000. Thus, the estimated gain in terms of mean income rank is roughly \$500–800. For the other two outcomes, TOP-20 PROBABILITY and INCARCERATION,[59] the gains are even more sizable, especially for Black individuals. These gains are as high as 2–3 percentage points on average in terms of these two variables.

Bergman et al. (2023) select tracts based on MEAN RANK POOLED. For this measure, there is little additional gain from using CLOSE-NPMLE, at least when residualized against sufficiently rich covariates. Nevertheless, since about half of the trial participants are Black in Bergman et al.'s (2023) setting, one might consider providing more personalized recommendations by targeting measures of economic mobility for finer demographic subgroups. If we select tracts based on these demographic-specific measures of economic mobility, CLOSE-NPMLE then provides economically significant improvements.[60]

We can think of the performance gap between INDEPENDENT-GAUSS and NAIVE as the *value of basic empirical Bayes*. If practitioners find using the standard empirical Bayes method to be a worthwhile investment over screening on the raw estimates directly, perhaps they reveal that the value of basic empirical Bayes is economically significant. Across the

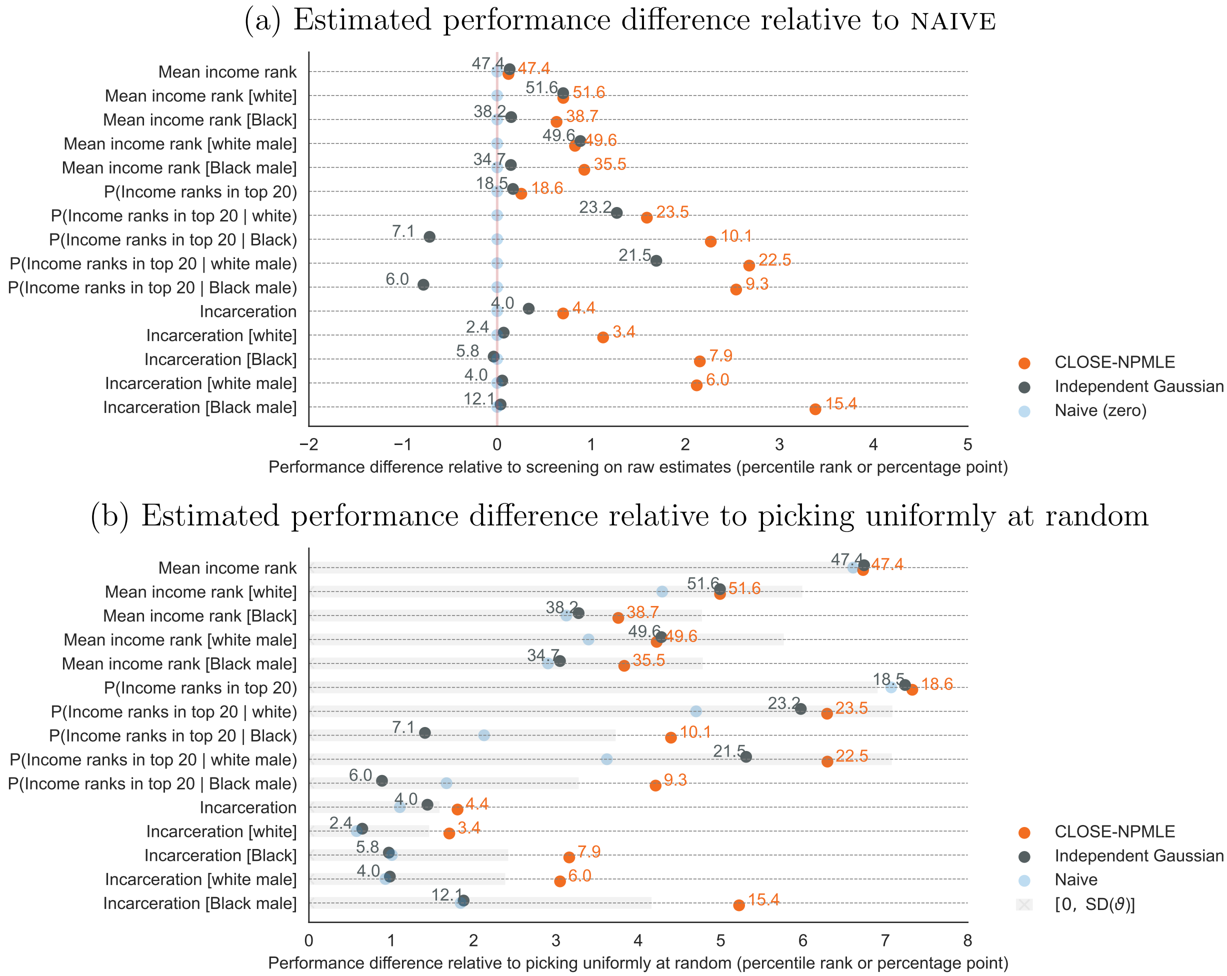
[56] Appendix B.6 contains results where we perform empirical Bayes pooling over all Commuting Zones and select the top third within each Communting Zone. We obtain very similar results. Appendix B.6 also contains results without residualizing against covariates, and INDEPENDENT-GAUSS performs very poorly in that setting. Appendix B.5 contains results on estimating ϑ_i in MSE (Decision Problem 1) in this context.

[57] By virtue of Proposition 1, these estimated performances are unbiased for the true (negative) Bayes risk. Despite being averaged over 1,000 coupled bootstrap draws, these estimates are not free of sampling error, since, for one, the stochastic components in Y_i are not redrawn.

[58] For MEAN RANK POOLED, CLOSE-NPMLE is worse by 0.012 percentile ranks, and CLOSE-NPMLE is worse by 0.058 percentile ranks for MEAN RANK for white males. In either case, the estimated disimprovement is small.

[59] For incarceration, we consider a policy objective of encouraging people to move *out* of high-incarceration areas.

[60] Appendix B.7 shows that screening with mobility measures for Black individuals outperforms screening mobility for Black individuals with the POOLED estimate.



Notes. These figures show the estimated performance of various decision rules over 1,000 draws of coupled bootstrap. Empirical Bayes methods, including residualization with respect to the covariates, are applied *within* each Commuting Zone. Performance is measured as the mean ϑ_i among selected Census tracts. All decision rules select the top third of Census tracts within each Commuting Zone. Figure (a) plots the estimated performance *gap* relative to NAIVE, where we annotate with the estimated performance for CLOSE-NPMLE and INDEPENDENT-GAUSS. Figure (b) plots the estimated performance gap relative to picking uniformly at random; we continue to annotate with the estimated performance. The shaded regions in Figure (b) have lengths equal to the unconditional standard deviation of the underlying parameter ϑ. □

FIGURE 5. Performance of decision rules in top-m selection exercise

15 measures, the improvement of CLOSE-NPMLE over INDEPENDENT-GAUSS is on median 320% of the value of basic empirical Bayes, where the median is attained by MEAN RANK for Black individuals. Thus, the additional gain of CLOSE-NPMLE over INDEPENDENT-GAUSS is substantial compared to the value of basic empirical Bayes. If the latter is economically significant, then it is similarly worthwhile to use CLOSE-NPMLE instead.

For 3 of the 15 measures, including our running example, INDEPENDENT-GAUSS in fact underperforms NAIVE, rendering the estimated value of basic empirical Bayes negative. As a result, we consider a different normalization in Figure 5(b). Figure 5(b) plots the difference between a given method's performance and the estimated mean ϑ_i for a given measure. Analogous to the value of basic empirical Bayes, we think of the difference between INDEPENDENT-GAUSS's performance and the estimated mean ϑ_i as the *value of data*, since choosing the tracts randomly in the absence of data has expected performance equal to mean ϑ_i. If the mobility estimates are at all useful for decision-making, the value of data must be economically significant.

Across the 15 measures considered, the gain of CLOSE-NPMLE is on median 25% of the value of data. For six of the 15 measures, the gain of CLOSE-NPMLE exceeds the value of data. For MEAN RANK for Black individuals, the incremental value of CLOSE-NPMLE over INDEPENDENT-GAUSS is about 15% of the value of data, which is already sizable. These relative gains are more substantial for the binarized outcome variables TOP-20 PROBABILITY and INCARCERATION. For our running example (TOP-20 PROBABILITY for Black individuals), this incremental gain of CLOSE-NPMLE is 210% the value of data. That is, relative to choosing randomly, CLOSE-NPMLE delivers *gains 3.1 times that of* INDEPENDENT-GAUSS.

6. **Conclusion**

This paper studies empirical Bayes methods in the heteroskedastic Gaussian location model. We argue that prior independence—the assumption that the precision of estimates does not predict the true parameter—is theoretically questionable and often empirically rejected. Empirical Bayes shrinkage methods that rely on prior independence can generate worse posterior mean estimates, and screening decisions based on these estimates can suffer as a result. They may even be worse than the selection decisions made with the unshrunk estimates directly.

Instead of treating θ_i as independent from σ_i, we model its conditional distribution as a location-scale family. This assumption leads naturally to a family of empirical Bayes strategies that we call CLOSE. We prove that CLOSE-NPMLE attains minimax-optimal rates in Bayes regret, extending previous theoretical results. That is, it approximates infeasible oracle Bayes posterior means as competently as statistically possible. Our main theoretical results are in terms of squared error, which we further connect to ranking-type decision problems in Bergman et al. (2023). Additionally, we show that an idealized version of CLOSE-NPMLE is robust, with finite worst-case Bayes risk. Lastly, we introduce a simple validation procedure based on coupled bootstrap (Oliveira et al., 2021) and highlight its utility for practitioners choosing among empirical Bayes shrinkage methods.

Simulation and validation exercises demonstrate that CLOSE-NPMLE generates sizable gains relative to the standard parametric empirical Bayes shrinkage method. Across calibrated simulations, CLOSE-NPMLE attains close-to-oracle mean-squared error performance. In a hypothetical, scaled-up version of Bergman et al. (2023), across a wide range of economic mobility measures, CLOSE-NPMLE consistently selects more mobile tracts than does the standard empirical Bayes method. The gains in the average economic mobility among selected tracts, relative to the standard empirical Bayes procedure, are often comparable to—or even multiples of—the value of basic empirical Bayes. These gains are even comparable to the benefit of using standard empirical Bayes procedures over ignoring the data.

We close by highlighting some future directions. In Section 5, we use kernel smoothing methods to estimate the unknown conditional moments $\eta_0 = (m_0, s_0)$. These methods presume a given level of smoothness and do not adapt to the true smoothness of η_0. We can imagine replacing the conditional moment estimation with adaptive methods (e.g., van der Vaart and van Zanten, 2009). With cross-fitting, the regret result should similarly adapt to the $\|\cdot\|_\infty$ rate of the estimators. Additionally, for the purpose of frequentist inference, the procedure of Armstrong et al. (2022) apply in our setting as well and provide confidence sets for the vector of parameters $\theta_{1:n}$ with average coverage guarantees. For frequentist inference on the oracle posterior mean $\mathbb{E}_{P_0}[\theta_i \mid Y_i, \sigma_i]$, we conjecture that a version of Ignatiadis and Wager's (2022) procedure—which so far only applies in the homoskedastic Gaussian case—is valid under the location-scale model (2.6).

References

Abadie, Alberto and Maximilian Kasy, "Choosing among regularized estimators in empirical economics: The risk of machine learning," *Review of Economics and Statistics*, 2019, *101* (5), 743–762. 14

———, **Anish Agarwal, Guido Imbens, Siwei Jia, James McQueen, and Serguei Stepaniants**, "Estimating the Value of Evidence-Based Decision Making," *arXiv preprint arXiv:2306.13681*, 2023. 2

Abaluck, Jason, Mauricio Caceres Bravo, Peter Hull, and Amanda Starc, "Mortality effects and choice across private health insurance plans," *The quarterly journal of economics*, 2021, *136* (3), 1557–1610. 2

Agan, Amanda Y and Sonja B Starr, "Employer neighborhoods and racial discrimination," Technical Report, National Bureau of Economic Research 2020. 11

Alesina, Alberto F, Edward L Glaeser, and Bruce Sacerdote, "Why doesn't the US have a European-style welfare system?," 2001. 11

Andrews, Isaiah and Maximilian Kasy, "Identification of and correction for publication bias," *American Economic Review*, 2019, *109* (8), 2766–94. 2

———, **Toru Kitagawa, and Adam McCloskey**, "Inference on Winners," 2023. 6

Angrist, Joshua D, Peter D Hull, Parag A Pathak, and Christopher R Walters, "Leveraging lotteries for school value-added: Testing and estimation," *The Quarterly Journal of Economics*, 2017, *132* (2), 871–919. 2

Arellano, Manuel and Stéphane Bonhomme, "Robust priors in nonlinear panel data models," *Econometrica*, 2009, *77* (2), 489–536. 2

Armstrong, Timothy B, Michal Kolesár, and Mikkel Plagborg-Møller, "Robust empirical bayes confidence intervals," *Econometrica*, 2022, *90* (6), 2567–2602. 27, 39

Arnold, David, Will Dobbie, and Peter Hull, "Measuring racial discrimination in bail decisions," *American Economic Review*, 2022, *112* (9), 2992–3038. 2

Athey, Susan and Stefan Wager, "Policy learning with observational data," *Econometrica*, 2021, *89* (1), 133–161. 15

Audibert, Jean-Yves and Alexandre B Tsybakov, "Fast learning rates for plug-in classifiers," *Annals of Statistics*, 2007. 25

Azevedo, Eduardo M, Alex Deng, José Luis Montiel Olea, Justin Rao, and E Glen Weyl, "A/b testing with fat tails," *Journal of Political Economy*, 2020, *128* (12), 4614–000. 2

Barrett, Nathan and Eugenia F Toma, "Reward or punishment? Class size and teacher quality," *Economics of education review*, 2013, *35*, 41–52. 53

Baum-Snow, Nathaniel and Lu Han, "The microgeography of housing supply," *Work in progress, University of Toronto*, 2019. 2

Becker, Gary S, Scott Duke Kominers, Kevin M Murphy, and Jörg L Spenkuch, "A theory of intergenerational mobility," *Journal of Political Economy*, 2018, *126* (S1), S7–S25. 11

Bergman, Peter, Raj Chetty, Stefanie DeLuca, Nathaniel Hendren, Lawrence F Katz, and Christopher Palmer, "Creating moves to opportunity: Experimental evidence on barriers to neighborhood choice," Technical Report, National Bureau of Economic Research 2023. 2, 3, 4, 5, 8, 9, 15, 16, 24, 30, 31, 32, 33, 34, 35, 36, 38, 39, 63, 67, 70

Bonhomme, Stéphane and Elena Manresa, "Grouped patterns of heterogeneity in panel data," *Econometrica*, 2015, *83* (3), 1147–1184. 2

———, **Kerstin Holzheu, Thibaut Lamadon, Elena Manresa, Magne Mogstad, and Bradley Setzler**, "How much should we trust estimates of firm effects and worker sorting?," 2020. 2

Bonvini, Matteo, Edward H Kennedy, and Luke J Keele, "Minimax optimal subgroup identification," *arXiv preprint arXiv:2306.17464*, 2023. 25

Boucheron, Stéphane, Gábor Lugosi, and Pascal Massart, *Concentration inequalities: A nonasymptotic theory of independence*, Oxford university press, 2013. 114, 117

Brown, Lawrence D, "In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies," 2008. 57

——— **and Eitan Greenshtein**, "Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means," *The Annals of Statistics*, 2009, pp. 1685–1704. 58

Bruhn, Jesse, Scott Imberman, and Marcus Winters, "Regulatory arbitrage in teacher hiring and retention: Evidence from Massachusetts charter schools," *Journal of Public Economics*, 2022, *215*, 104750. 2, 54

Cai, T Tony and Mark G Low, "Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional," *The Annals of Statistics*, 2011, pp. 1012–1041. 23

Calonico, Sebastian, Matias D Cattaneo, and Max H Farrell, "nprobust: Nonparametric kernel-based estimation and robust bias-corrected inference," *arXiv preprint arXiv:1906.00198*, 2019. 8, 27, 118

Chamberlain, Gary, "Panel data," *Handbook of econometrics*, 1984, *2*, 1247–1318. 2

Chandra, Amitabh, Amy Finkelstein, Adam Sacarny, and Chad Syverson, "Health care exceptionalism?

Performance and allocation in the US health care sector," *American Economic Review*, 2016, *106* (8), 2110–44. 2, 10, 53

Chen, Hong-Bin and Jonathan Niles-Weed, "Asymptotics of smoothed Wasserstein distances," *Potential Analysis*, 2022, *56* (4), 571–595. 60

Chen, Jiafeng, "Mean-variance constrained priors have finite maximum Bayes risk in the normal location model," *arXiv preprint arXiv:2303.08653*, 2023. 26, 51

Chetty, Raj and Nathaniel Hendren, "The impacts of neighborhoods on intergenerational mobility I: Childhood exposure effects," *The Quarterly Journal of Economics*, 2018, *133* (3), 1107–1162. 2, 30

———, John N Friedman, and Jonah E Rockoff, "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates," *American economic review*, 2014, *104* (9), 2593–2632. 16

———, ———, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter, "The opportunity atlas: Mapping the childhood roots of social mobility," Technical Report 2020. 2, 4, 8, 30, 31, 32, 63, 67

———, Nathaniel Hendren, and Lawrence F Katz, "The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment," *American Economic Review*, 2016, *106* (4), 855–902. 30

Chyn, Eric and Lawrence F Katz, "Neighborhoods matter: Assessing the evidence for place effects," *Journal of Economic Perspectives*, 2021, *35* (4), 197–222. 2, 30

Coey, Dominic and Kenneth Hung, "Empirical Bayesian Selection for Value Maximization," *arXiv preprint arXiv:2210.03905*, 2022. 25

Colmer, Jonathan, Ian Hardman, Jay Shimshack, and John Voorheis, "Disparities in PM2. 5 air pollution in the United States," *Science*, 2020, *369* (6503), 575–578. 11

Cutler, David M, Edward L Glaeser, and Jacob L Vigdor, "The rise and decline of the American ghetto," *Journal of political economy*, 1999, *107* (3), 455–506. 11

Dedecker, Jérôme and Bertrand Michel, "Minimax rates of convergence for Wasserstein deconvolution with supersmooth errors in any dimension," *Journal of Multivariate Analysis*, 2013, *122*, 278–291. 23, 25

DellaVigna, Stefano and Elizabeth Linos, "RCTs to scale: Comprehensive evidence from two nudge units," *Econometrica*, 2022, *90* (1), 81–116. 2

Derenoncourt, Ellora, "Can you move to opportunity? Evidence from the Great Migration," *American Economic Review*, 2022, *112* (2), 369–408. 11

Diamond, Rebecca and Enrico Moretti, "Where is standard of living the highest? Local prices and the geography of consumption," Technical Report, National Bureau of Economic Research 2021. 2

Dimick, Justin B, Douglas O Staiger, and John D Birkmeyer, "Ranking hospitals on surgical mortality: the importance of reliability adjustment," *Health services research*, 2010, *45* (6p1), 1614–1629. 2

Doyle, Joseph J, John A Graves, Jonathan Gruber et al., "Evaluating measures of hospital quality," Technical Report, National Bureau of Economic Research 2017. 2

Durlauf, Steven N, Andros Kourtellos, and Chih Ming Tan, "The Great Gatsby Curve," *Annual Review of Economics*, 2022, *14*, 571–605. 11

Efron, Bradley, "Two modeling strategies for empirical Bayes estimation," *Statistical science: a review journal of the Institute of Mathematical Statistics*, 2014, *29* (2), 285. 6, 58

———, "Empirical Bayes deconvolution estimates," *Biometrika*, 2016, *103* (1), 1–20. 59

———, "Bayes, oracle Bayes and empirical Bayes," *Statistical science*, 2019, *34* (2), 177–201. 6

——— and Carl Morris, "Data analysis using Stein's estimator and its generalizations," *Journal of the American Statistical Association*, 1975, *70* (350), 311–319. 57

Egan, Mark, Gregor Matvos, and Amit Seru, "When Harry fired Sally: The double standard in punishing misconduct," *Journal of Political Economy*, 2022, *130* (5), 1184–1248. 2

Einav, Liran, Amy Finkelstein, and Neale Mahoney, "Producing Health: Measuring Value Added of Nursing Homes," Technical Report, National Bureau of Economic Research 2022. 2, 54

Elliott, Graham, Nikolay Kudrin, and Kaspar Wüthrich, "Detecting p-hacking," *Econometrica*, 2022, *90* (2), 887–906. 2

Fay, Robert E and Roger A Herriot, "Estimates of income for small places: an application of James-Stein procedures to census data," *Journal of the American Statistical Association*, 1979, *74* (366a), 269–277. 32, 56

Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams, "Place-based drivers of mortality: Evidence from migration," *American Economic Review*, 2021, *111* (8), 2697–2735. 2

Fu, Luella J, Gareth M James, and Wenguang Sun, "Nonparametric empirical bayes estimation on heterogeneous data," *arXiv preprint arXiv:2002.12586*, 2020. 57

George, Edward I, Veronika Ročková, Paul R Rosenbaum, Ville A Satopää, and Jeffrey H Silber, "Mortality rate estimation and standardization for public reporting: Medicare's hospital compare," *Journal of the American Statistical Association*, 2017, *112* (519), 933–947. 2, 7, 10

Giacomini, Raffaella, Sokbae Lee, and Silvia Sarpietro, "A Robust Method for Microforecasting and Estimation of Random Effects," 2023. 2

Gilraine, Michael, Jiaying Gu, and Robert McMillan, "A new method for estimating teacher value-added," Technical Report, National Bureau of Economic Research 2020. 2, 7, 15, 16

Gu, Jiaying and Christopher Walters, "SI 2022 Methods Lectures - Empirical Bayes Methods, Theory and Application," https://www.nber.org/conferences/si-2022-methods-lectures-empirical-bayes-methods-theory-and-application 7 2022. 6

______ and Roger Koenker, "Empirical Bayesball remixed: Empirical Bayes methods for longitudinal data," *Journal of Applied Econometrics*, 2017, *32* (3), 575–599. 57

______ and ______, "Invidious comparisons: Ranking and selection as compound decisions," *Econometrica*, 2023, *91* (1), 1–41. 15, 23

Hanushek, Eric A, "The economic value of higher teacher quality," *Economics of Education review*, 2011, *30* (3), 466–479. 16

Heckman, James and Burton Singer, "A method for minimizing the impact of distributional assumptions in econometric models for duration data," *Econometrica: Journal of the Econometric Society*, 1984, pp. 271–320. 13

Hull, Peter, "Estimating hospital quality with quasi-experimental data," *Available at SSRN 3118358*, 2018. 2

Ignatiadis, Nikolaos and Stefan Wager, "Covariate-powered empirical Bayes estimation," *Advances in Neural Information Processing Systems*, 2019, *32*. 21, 55

______ and ______, "Confidence intervals for nonparametric empirical Bayes analysis," *Journal of the American Statistical Association*, 2022, *117* (539), 1149–1166. 39

Jackson, C Kirabo and Claire Mackevicius, "The distribution of school spending impacts," Technical Report, National Bureau of Economic Research 2021. 11

Jiang, Wenhua, "On general maximum likelihood empirical Bayes estimation of heteroscedastic IID normal means," *Electronic Journal of Statistics*, 2020, *14* (1), 2272–2297. 4, 7, 13, 14, 17, 18, 19, 22, 60, 73, 77, 81, 92, 93, 96, 98, 99, 109, 112, 114

______ and Cun-Hui Zhang, "General maximum likelihood empirical Bayes estimation of normal means," *The Annals of Statistics*, 2009, *37* (4), 1647–1684. 4, 7, 13, 17, 19, 58, 60, 112, 114, 117

______ and ______, "Empirical Bayes in-season prediction of baseball batting averages," in "Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown," Vol. 6, Institute of Mathematical Statistics, 2010, pp. 263–274. 56

Jing, Bing-Yi, Zhouping Li, Guangming Pan, and Wang Zhou, "On sure-type double shrinkage estimation," *Journal of the American Statistical Association*, 2016, *111* (516), 1696–1704. 14

Kain, John F, "Housing segregation, negro employment, and metropolitan decentralization," *The quarterly journal of economics*, 1968, *82* (2), 175–197. 11

Kane, Thomas J and Douglas O Staiger, "Estimating teacher impacts on student achievement: An experimental evaluation," Technical Report, National Bureau of Economic Research 2008. 16

Kiefer, Jack and Jacob Wolfowitz, "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters," *The Annals of Mathematical Statistics*, 1956, pp. 887–906. 4, 13

Kitagawa, Toru and Aleksey Tetenov, "Who should be treated? empirical welfare maximization methods for treatment choice," *Econometrica*, 2018, *86* (2), 591–616. 15

Kline, Patrick, Evan K Rose, and Christopher R Walters, "Systemic discrimination among large US employers," *The Quarterly Journal of Economics*, 2022, *137* (4), 1963–2036. 2

______, ______, and Christopher Walters, "A Discrimination Report Card," *arXiv preprint arXiv:2306.13005*, 2023. 2, 7, 23

Koenker, Roger and Ivan Mizera, "Convex optimization, shape constraints, compound decisions, and empirical Bayes rules," *Journal of the American Statistical Association*, 2014, *109* (506), 674–685. 4, 13, 28

______ and Jiaying Gu, "REBayes: an R package for empirical Bayes mixture methods," *Journal of Statistical Software*, 2017, *82*, 1–26. 13, 28

______ and ______, "Comment: Minimalist g-modeling," 2019. 7, 13, 23

Kubokawa, T, CP Robert, and AK Md E Saleh, "Estimation of noncentrality parameters," *Canadian Journal of Statistics*, 1993, *21* (1), 45–57. 27, 119

Kwon, Soonwoo, "Optimal Shrinkage Estimation of Fixed Effects in Linear Panel Data Models," *EliScholar–A Digital Platform for Scholarly Publishing at Yal e*, 2021, p. 1. 14

Laliberté, Jean-William, "Long-term contextual effects in education: Schools and neighborhoods," *American Economic Journal: Economic Policy*, 2021, *13* (2), 336–377. 11, 30

Lazear, Edward P, "Educational production," *The Quarterly Journal of Economics*, 2001, *116* (3), 777–803. 53

Lehmann, Erich L and George Casella, *Theory of point estimation*, Springer Science & Business Media, 2006. 6

Liang, TaChen, "On an empirical Bayes test for a normal mean," *Annals of statistics*, 2000, pp. 648–655. 25

Lindsay, Bruce G, "Mixture models: theory, geometry, and applications," in "in" Ims 1995. 13

Liu, Laura, Hyungsik Roger Moon, and Frank Schorfheide, "Forecasting with dynamic panel data models," *Econometrica*, 2020, *88* (1), 171–201. 2, 58

Luo, Jiajun, Trambak Banerjee, Gourab Mukherjee, and Wenguang Sun, "Empirical Bayes Estimation with Side Information: A Nonparametric Integrative Tweedie Approach," 2023. 58

Manski, Charles F, "Statistical treatment rules for heterogeneous populations," *Econometrica*, 2004, *72* (4), 1221–1246. 15

Meager, Rachael, "Aggregating distributional treatment effects: A Bayesian hierarchical analysis of the microcredit literature," *American Economic Review*, 2022, *112* (6), 1818–47. 2

Mehta, Nirav, "Measuring quality for use in incentive schemes: The case of "shrinkage" estimators," *Quantitative Economics*, 2019, *10* (4), 1537–1577. 2, 10, 53

Mogstad, Magne, Joseph P Romano, Azeem M Shaikh, and Daniel Wilhelm, "A Comment on:"Invidious Comparisons: Ranking and Selection as Compound Decisions" by Jiaying Gu and Roger Koenker," *Econometrica*, 2023, *91* (1), 53–60. 15

———, ———, Azeem Shaikh, and Daniel Wilhelm, "Inference for ranks with applications to mobility across neighborhoods and academic achievement across countries," Technical Report, National Bureau of Economic Research 2020. 6, 15

Morris, Carl N, "Parametric empirical Bayes inference: theory and applications," *Journal of the American statistical Association*, 1983, *78* (381), 47–55. 6

Mountjoy, Jack and Brent R Hickman, "The returns to college (s): Relative value-added and match effects in higher education," Technical Report, National Bureau of Economic Research 2021. 2

Müller, Ulrich K and Mark W Watson, "Spatial Unit Roots," *manuscript, Princeton University*, 2022. 55

Olea, José Luis Montiel, Brendan O'Flaherty, and Rajiv Sethi, "Empirical Bayes Counterfactuals in Poisson Regression with an Application to Police Use of Deadly Force," *Available at SSRN 3857213*, 2021. 2

Oliveira, Natalia L, Jing Lei, and Ryan J Tibshirani, "Unbiased risk estimation in the normal means problem via coupled bootstrap techniques," *arXiv preprint arXiv:2111.09447*, 2021. 4, 29, 38

Olley, G. Steven and Ariel Pakes, "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 1996, *64* (6), 1263–1297. 54

Pensky, Marianna, "Minimax theory of estimation of linear functionals of the deconvolution density with or without sparsity," 2017. 23

Polyanskiy, Yury and Yihong Wu, "Self-regularizing property of nonparametric maximum likelihood estimator in mixture models," *arXiv preprint arXiv:2008.08244*, 2020. 4, 13

——— and ———, "Sharp regret bounds for empirical Bayes and compound decision problems," *arXiv preprint arXiv:2109.03943*, 2021. 13

Rambachan, Ashesh, "Identifying prediction mistakes in observational data," 2021. 2

Robbins, Herbert, "An empirical Bayes approach to statistics," 1956. 14

Saha, Sujayam and Adityanand Guntuboyina, "On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising," *The Annals of Statistics*, 2020, *48* (2), 738–762. 13, 17, 114

Sen, Bodhisattva, "A gentle introduction to empirical process theory and applications," *Lecture Notes, Columbia University*, 2018, *11*, 28–29. 121, 128

Shen, Yandi and Yihong Wu, "Empirical Bayes estimation: When does g-modeling beat f-modeling in theory (and in practice)?," *arXiv preprint arXiv:2211.12692*, 2022. 13

Soloff, Jake A, Adityanand Guntuboyina, and Bodhisattva Sen, "Multivariate, heteroscedastic empirical bayes via nonparametric maximum likelihood," *arXiv preprint arXiv:2109.03466*, 2021. 4, 7, 13, 14, 17, 18, 19, 22, 75, 76, 112, 114

Stone, Charles J, "Optimal rates of convergence for nonparametric estimators," *The annals of Statistics*, 1980, pp. 1348–1360. 20

Stoye, Jörg, "Minimax regret treatment choice with finite samples," *Journal of Econometrics*, 2009, *151* (1), 70–81. 15

Tsybakov, A.B., *Introduction to Nonparametric Estimation* Springer Series in Statistics, Springer New York, 2008. 20, 48, 120, 121, 123, 125

van der Vaart, Aad W and J Harry van Zanten, "Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth," 2009. 39

van der Vaart, Aad W and Jon A Wellner, *Weak convergence*, Springer, 1996. 19, 73, 87, 126, 128

Vershynin, Roman, *High-dimensional probability: An introduction with applications in data science*, Vol. 47, Cambridge university press, 2018. 18, 81, 86, 87, 91, 102

Voinov, Vasiliĭ Grigorévich and Mikhail Stepanovich Nikulin, *Unbiased estimators and their applications: volume 1: univariate case*, Vol. 263, Springer Science & Business Media, 2012. 29

Weinstein, Asaf, Zhuang Ma, Lawrence D Brown, and Cun-Hui Zhang, "Group-linear empirical Bayes estimates for a heteroscedastic normal mean," *Journal of the American Statistical Association*, 2018, *113* (522), 698–710. 13

Wernerfelt, Nils, Anna Tuchman, Bradley Shapiro, and Robert Moakler, "Estimating the Value of Offsite Data to Advertisers on Meta," *University of Chicago, Becker Friedman Institute for Economics Working Paper*, 2022, (114). 2

Xie, Xianchao, SC Kou, and Lawrence D Brown, "SURE estimates for a heteroscedastic hierarchical model," *Journal of the American Statistical Association*, 2012, *107* (500), 1465–1479. 14

Zhang, Cun-Hui, "Empirical Bayes and compound estimation of normal means," *Statistica Sinica*, 1997, *7* (1), 181–193. 100

Appendix for "Empirical Bayes When Estimation Precision Predicts Parameters"

Jiafeng Chen

January 11, 2024

Contents

Part 1. Proofs and discussions of results except the regret upper bound 47

Appendix A. Proofs and discussions of results except the regret upper bound 47

A.1. A simple regret rate lower bound: proof of Theorem 2 47

A.2. Relating other decision objects to squared-error loss 48

A.3. Worst-case risk 50

A.4. Unbiased loss estimation 51

A.5. A discrete choice model 52

A.6. Interpretation of empirical Bayes sampling model 54

A.7. Alternatives to CLOSE 56

Part 2. Additional empirical results 61

Appendix B. Additional empirical exercises 61

B.1. Positivity of $s_0(\cdot)$ in the Opportunity Atlas data 61

B.2. Robustness checks for the calibration exercise in Section 5 62

B.3. Simulation exercise setup 62

B.4. Different Monte Carlo setup 64

B.5. MSE in validation exercise with coupled bootstrap 64

B.6. Empirical Bayes pooling over all Commuting Zones in validation exercise 64

B.7. The tradeoff between accurate targeting and estimation precision 67

Part 3. Regret control proofs 69

Appendix C. Setup, assumptions, and notation 69

C.1. Assumptions 72

C.2. Regret control: result statement 74

C.3. Regret control: proof ideas 75

Appendix D. An oracle inequality for the likelihood 78

D.1. Proof of Corollary D.1 79

D.2. Proof of Theorem D.1 79

D.3. Bounding U_{1m} 82

D.4. Bounding U_{1s} 83

D.5. Bounding U_{2m} 84

D.6. Bounding U_{2s} 88

D.7. Bounding R_1 90

D.8. Bounding U_{3m}, U_{3s} 90

D.9. Bounding R_2 91

D.10. Derivative computations 92

D.11. Metric entropy of $\mathcal{P}(\mathbb{R})$ under moment-based distance 92
D.12. Auxiliary lemmas 98
Appendix E. A large-deviation inequality for the average Hellinger distance 103
E.1. Proof of Theorem E.1 105
E.2. Auxiliary lemmas 109
Appendix F. An oracle inequality for the Bayes squared-error risk 109
F.1. Controlling $\mathrm{MSERegret}_n$ on A_n^{C} 109
F.2. Controlling $\mathrm{MSERegret}_n$ on A_n 110
F.3. Controlling ξ_1 111
F.4. Controlling ξ_3 111
F.5. Controlling ξ_4 115
F.6. Auxiliary lemmas 115
Appendix G. Estimating η_0 by local linear regression 117
G.1. Auxiliary lemmas 124

Part 1. Proofs and discussions of results except the regret upper bound

Appendix A. Proofs and discussions of results except the regret upper bound

A.1. **A simple regret rate lower bound: proof of Theorem 2.** In this section, we prove Theorem 2, restated below.

Theorem 2. *Fix a set of valid hyperparameters $\mathcal{H} = (\sigma_\ell, \sigma_u, s_\ell, s_u, A_0, A_1, \alpha, \beta_0, p)$ for Assumptions 2 to 4. Let $\mathcal{P}(\mathcal{H}, \sigma_{1:n})$ be the set of distributions P_0 on support points $\sigma_{1:n}$ which satisfy (2.6) and Assumptions 2 to 4 corresponding to $\mathcal{H}$. For a given P_0, let $\theta_i^* = \mathbb{E}_{P_0}[\theta_i \mid Y_i, \sigma_i]$ denote the oracle posterior means. Then there exists a constant $c_{\mathcal{H}} > 0$ such that the worst-case Bayes regret of any estimator exceeds $c_{\mathcal{H}} n^{-\frac{2p}{2p+1}}$:*

$$\inf_{\hat{\theta}_{1:n}} \sup_{\substack{\sigma_{1:n} \in (\sigma_\ell, \sigma_u) \\ P_0 \in \mathcal{P}(\mathcal{H}, \sigma_{1:n})}} \mathbb{E}_{P_0}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2 - (\theta_i^* - \theta_i)^2\right] \geq c_{\mathcal{H}} n^{-\frac{2p}{2p+1}},$$

where the infimum is taken over all (possibly randomized) estimators of $\theta_{1:n}$.

Proof. We consider a specific choice of $G_0, \sigma_{1:n}$, and s_0. Namely, suppose $G_0 \sim \mathcal{N}(0, 1)$, $\sigma_{1:n}$ are equally spaced in $[\sigma_\ell, \sigma_u]$, and $s_0(\sigma) = (s_\ell + s_u)/2 \equiv s_0$ is constant. Note that we can represent

$$Y_i = \underbrace{\theta_i + \sigma_\ell W_i}_{V_i} + (\sigma_i^2 - \sigma_\ell)^{1/2} U_i.$$

for independent Gaussians $W_i, U_i \sim \mathcal{N}(0, 1)$. Suppose we are additionally given V_i, σ_ℓ. The expanded class of estimators $\tilde{\theta}_{1:n}$ that may depend on V_i, σ_ℓ is larger than the estimators $\hat{\theta}_{1:n}$. Moreover, since $((V_i, \sigma_i)_{i=1}^n, \sigma_\ell)$ is sufficient for $\theta_{1:n}$, we may restrict attention to $\tilde{\theta}_{1:n}$ that depend solely on $V_{1:n}, \sigma_{1:n}, \sigma_\ell$.

Under our assumptions, the oracle posterior means θ_i^* are equal to

$$\theta_i^* = \frac{s_0^2}{s_0^2 + \sigma_i^2} Y_i + \frac{\sigma_i^2}{s_0^2 + \sigma_i^2} m_0(\sigma_i)$$

For a given vector of estimates $\tilde{\theta}_{1:n}$, we can form

$$\hat{m}(\sigma_i) = \frac{s_0^2 + \sigma_i^2}{\sigma_i^2}\left(\tilde{\theta}_i - \frac{s_0^2}{s_0^2 + \sigma_i^2} Y_i\right)$$

Then

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\tilde{\theta}_i - \theta_i^*)^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left(\frac{\sigma_i^2}{s_0^2 + \sigma_i^2}\right)^2 (\hat{m}(\sigma_i) - m_0(\sigma_i))^2\right] \gtrsim \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\hat{m}(\sigma_i) - m_0(\sigma_i))^2\right].$$

We have just shown that

$$\inf_{\hat{\theta}_{1:n}} \sup_{\sigma_{1:n}, P_0} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2 - (\theta_i^* - \theta_i)^2\right] \gtrsim \inf_{\hat{m}} \sup_{m_0} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\hat{m}(\sigma_i) - m_0(\sigma_i))^2\right]$$

where the supremum is over m_0 satisfying Assumption 4, and the infimum is over all randomized estimators of $m_0(\sigma_1), \ldots, m_0(\sigma_n)$ with data (V_i, σ_i). Note that the squared error loss on the right-hand side takes expectation over the fixed design points $\sigma_1, \ldots, \sigma_n$.

Lastly, we connect the squared loss on the design points to the L_2 loss of estimating $m_0(\cdot)$ with homoskedastic data $V_i \sim \mathcal{N}(m_0(\sigma_i), \sigma_\ell^2 + s_0^2)$. Since we are simply confronted with a nonparametric regression problem, note that we may translate and rescale so that the design points $\sigma_{1:n}$ are equally spaced in $[0, 1]$ and the variance of V_i is 1—potentially changing the constant A_1 for the Hölder smoothness condition. The remaining task is to connect the average ℓ_2 loss on a set of equally spaced grid points to the L_2 loss over the interval.

Observe that for any $\hat{m}(\sigma_1), \ldots, \hat{m}(\sigma_n)$, there is a function $\tilde{m} : [0, 1] \to \mathbb{R}$ such that its average value on $[1 + (i-1)/n, 1 + i/n]$ is $\hat{m}(\sigma_i)$:

$$n \int_{[1+(i-1)/n, 1+i/n]} \tilde{m}(\sigma)\, d\sigma = \hat{m}(\sigma_i).$$

Now, note that

$$\begin{aligned}
\int_0^1 (\tilde{m}(x) - m_0(x))^2\, dx &= \sum_{i=1}^n \int_{[(i-1)/n, i/n]} (\tilde{m}(x) - m_0(x))^2\, dx \\
&\le 2 \sum_{i=1}^n \int_{[(i-1)/n, i/n]} (\tilde{m}(x) - m_0(\sigma_i))^2 + (m_0(\sigma_i) - m_0(x))^2\, dx \\
&\hspace{20em} \text{(Triangle inequality)} \\
&\le 2 \sum_{i=1}^n \left[\frac{1}{n} (\hat{m}_i - m_0(\sigma_i))^2 + \frac{L^2}{n^3} \right] \\
&= \frac{2}{n} \sum_{i=1}^n (\hat{m}_i - m_0(\sigma_i))^2 + \frac{2L^2}{n^2}.
\end{aligned}$$

The third line follows by observing (i) $\int_I (\tilde{m}(x) - m_0(\sigma_i))^2\, dx = \left(n \int_I \tilde{m}(x)\, dx - m_0(\sigma_i)\right)^2 \frac{1}{n}$ and (ii) $m_0(\cdot)$ is Lipschitz for some constant L since $p \ge 1$ in Assumption 4.

Therefore,

$$\inf_{\hat{m}} \sup_{m_0} \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n (\hat{m}(\sigma_i) - m_0(\sigma_i))^2 \right] \ge \frac{1}{2} \inf_{\tilde{m}} \sup_{m_0} \left\{ \mathbb{E} \left[\int_0^1 (\tilde{m}(x) - m_0(x))^2\, dx \right] - \frac{2L^2}{n^2} \right\} \gtrsim_{\mathcal{H}} n^{-\frac{2p}{2p+1}},$$

where the last inequality follows from the well-known result of L_2 minimax regression rate for Hölder classes. See, for instance, Corollary 2.3 in Tsybakov (2008). □

Remark A.1. For ease of interpretation, Theorem 2 is stated in the expected regret version, which is slightly disconnected from the upper bound Theorem 1, which conditions on a high-probability event. Observe that Theorem 1 immediately implies the in-probability upper bound on MSERegret_n:

$$\text{MSERegret}_n(\hat{G}_n, \hat{\eta}) = O_P\left(n^{-\frac{2p}{2p+1}} (\log n)^{\frac{2+\alpha}{\alpha} + 3 + 2\beta_0} \right).$$

Using the in-probability version of the minimax lower bound for nonparametric regression in Theorem 2 then implies an analogous lower bound (See, for instance, Theorems 2.4 and 2.5 in Tsybakov, 2008). ■

A.2. **Relating other decision objects to squared-error loss.**

Theorem 3. *Suppose (2.3) holds, but (2.6) may or may not hold. Let $\hat{\delta}_i$ be the plug-in decisions with any vector of estimates $\hat{\theta}_i$, not necessarily from* CLOSE-NPMLE. *We have the following inequalities on the expected regret corresponding to the decision rules $\hat{\delta}_i$:*

(1) For UTILITY MAXIMIZATION BY SELECTION,

$$\mathbb{E}[\text{UMRegret}_n] \le \left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_i - \theta_i^*)^2\right]\right)^{1/2}. \tag{3.7}$$

(2) For TOP-m SELECTION,

$$\mathbb{E}[\text{TopRegret}_n^{(m)}] \le 2\sqrt{\frac{n}{m}}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_i - \theta_i^*)^2\right]\right)^{1/2}. \tag{3.8}$$

Proof. (1) We compute

$$\begin{aligned}\text{UMRegret}_n &= \frac{1}{n}\sum_{i=1}^n \mathbb{1}(\theta_i^* \ge c_i)(\theta_i - c_i) - \frac{1}{n}\sum_{i=1}^n \mathbb{1}(\hat{\theta}_i \ge c_i)(\theta_i - c_i) \\ &= \frac{1}{n}\sum_{i=1}^n \left\{\mathbb{1}(\theta_i^* \ge c_i) - \mathbb{1}(\hat{\theta}_i \ge c_i)\right\}(\theta_i - c_i)\end{aligned}$$

By law of iterated expectations, since $\hat{\theta}_i, \theta_i^*$ are both measurable with respect to the data,[61]

$$\mathbb{E}[\text{UMRegret}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left\{\mathbb{1}(\theta_i^* \ge c_i) - \mathbb{1}(\hat{\theta}_i \ge c_i)\right\}(\theta_i^* - c_i)\right]$$

Note that, for $\mathbb{1}(\theta_i^* \ge c_i) - \mathbb{1}(\hat{\theta}_i \ge c_i)$ to be nonzero, c_i is between $\hat{\theta}_i$ and θ_i^*. Hence, $|\theta_i^* - c_i| \le |\theta_i^* - \theta_i|$ and thus

$$\mathbb{E}[\text{UMRegret}_n] \le \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n |\theta_i^* - \theta_i|\right] \le \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\theta_i^* - \theta_i)^2\right]^{1/2}. \qquad \text{(Jensen's inequality)}$$

(2) Let $\mathcal{J}^*$ collect the indices of the top-m entries of θ_i^* and let $\hat{\mathcal{J}}$ collect the indices of the top-m entries of $\hat{\theta}_i$. Then,

$$\frac{m}{n}\text{TopRegret}_n^{(m)} = \frac{1}{n}\sum_{i=1}^n \left\{\mathbb{1}(i \in \mathcal{J}^*) - \mathbb{1}(i \in \hat{\mathcal{J}})\right\}\theta_i$$

and hence, by law of iterated expectations,

$$\frac{m}{n}\mathbb{E}[\text{TopRegret}_n^{(m)}] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\{\mathbb{1}(i \in \mathcal{J}^*) - \mathbb{1}(i \in \hat{\mathcal{J}})\right\}\theta_i^*\right].$$

[61]For a randomized decision rule $\hat{\theta}_i$ that is additionally measurable with respect to some U independent of $(\theta_i, Y_i, \sigma_i)_{i=1}^n$, this step continues to hold since $\mathbb{E}[\theta_i \mid U, Y_i, \sigma_i] = \theta_i^*$.

Observe that this can be controlled by applying Proposition A.1, where $w_i = 0$ for all $i \le n - m$ and $w_i = 1$ for all $i > n - m$. In this case, $\|w\| = \sqrt{m}$. Hence,

$$\frac{m}{n}\mathbb{E}[\mathrm{TopRegret}_n^{(m)}] \le 2\sqrt{\frac{m}{n}}\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n(\hat\theta_i - \theta_i^*)^2\right)^{1/2}\right] \le 2\sqrt{\frac{m}{n}}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n(\hat\theta_i - \theta_i^*)^2\right]\right)^{1/2}.$$

Divide through by m/n to obtain the result.

□

Proposition A.1. *Suppose $\sigma(\cdot)$ is a permutation such that $\hat\theta_{\sigma(n)} \ge \cdots \ge \hat\theta_{\sigma(1)}$. Then*

$$\frac{1}{n}\sum_{i=1}^n w_i\theta^*_{(i)} - \frac{1}{n}\sum_{i=1}^n w_i\theta^*_{\sigma(i)} \le \frac{2\|w\|}{\sqrt{n}}\sqrt{\frac{1}{n}\sum_{i=1}^n(\hat\theta_i - \theta_i^*)^2},$$

where $\|w\| = \sqrt{\sum_i w_i^2}$.

Proof. We compute

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n w_i\theta^*_{(i)} - \frac{1}{n}\sum_{i=1}^n w_i\theta^*_{\sigma(i)} &\le \left|\frac{1}{n}\sum_{i=1}^n w_i\theta^*_{(i)} - \frac{1}{n}\sum_{i=1}^n w_i\hat\theta_{\sigma(i)}\right| + \left|\frac{1}{n}\sum_{i=1}^n w_i(\hat\theta_{\sigma(i)} - \theta^*_{\sigma(i)})\right| \\ &\le \frac{\|w\|_2}{\sqrt{n}}\cdot\left(\frac{1}{n}\sum_{i=1}^n(\theta^*_{(i)} - \hat\theta_{\sigma(i)})^2\right)^{1/2} + \frac{\|w\|_2}{\sqrt{n}}\sqrt{\frac{1}{n}\sum_{i=1}^n(\hat\theta_i - \theta_i^*)^2} \\ &\le 2\frac{\|w\|_2}{\sqrt{n}}\sqrt{\frac{1}{n}\sum_{i=1}^n(\hat\theta_i - \theta_i^*)^2}.\end{aligned}$$

The last step follows from the observation that

$$\sum_{i=1}^n(\theta^*_{(i)} - \hat\theta_{\sigma(i)})^2 \le \sum_{i=1}^n(\hat\theta_i - \theta_i^*)^2.$$

The left-hand side is the sorted difference between θ_i^* and $\hat\theta_i$. This is smaller than the unsorted difference by an application of the rearrangement inequality.[62] □

A.3. **Worst-case risk.**

Theorem 4. *Under (2.3) but not (2.6), assume the conditional distribution $\theta_i \mid \sigma_i$ has mean $m_0(\sigma_i)$ and variance $s_0^2(\sigma_i)$. Denote the set of distributions of $\theta_{1:n} \mid \sigma_{1:n}$ which obey these restrictions as $\mathcal{P}(m_0, s_0)$. Let $\hat\theta_{i,G_0^*,\eta_0}$ denote the posterior mean estimates with some prior P^* under the location-scale model $P^*(\theta_i \le t \mid \sigma_i) = G_0^*\left(\frac{t-m_0(\sigma_i)}{s_0(\sigma_i)}\right)$, for some fixed G_0^* with zero mean and unit variance. Let $\overline{\rho} = \max_i s_0^2(\sigma_i)/\sigma_i^2 < \infty$ be the maximal conditional signal-to-noise ratio and assume that it is bounded. Then, for some $C_{\overline{\rho}} < \infty$ that solely depends on $\overline{\rho}$,*

$$\sup_{P_0\in\mathcal{P}(m_0,s_0)}\mathbb{E}_{P_0}\left[\frac{1}{n}\sum_{i=1}^n(\hat\theta_{i,G_0^*,\eta_0} - \theta_i)^2\right] \le C_{\overline{\rho}}\cdot\inf_{\hat\theta_{1:n}}\sup_{P_0\in\mathcal{P}(m_0,s_0)}\mathbb{E}_{P_0}\left[\frac{1}{n}\sum_{i=1}^n(\hat\theta_i - \theta_i)^2\right]. \tag{3.9}$$

[62] That is, for all real numbers $x_1 \le \cdots \le x_n, y_1 \le \cdots \le y_n$, $\sum_i x_iy_{\pi(i)} \le \sum_i x_iy_i$ for any permutation $\pi(\cdot)$.

where the infimum on the right-hand side is over all (possibly randomized) estimators of θ_i given $(Y_i, \sigma_i)_{i=1}^n$ and $\eta_0(\cdot)$.

Proof. Note that

$$\hat{\theta}_{i,G_0^*,\eta_0} = s_0(\sigma_i)\hat{\tau}_{i,G_0^*,\eta_0} + m_0(\sigma_i)$$

and

$$\theta_i = s_0(\sigma_i)\tau_i + m_0(\sigma_i).$$

Thus,

$$\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2 = \frac{1}{n}\sum_{i=1}^n s_0^2(\sigma_i)(\hat{\tau}_{i,G_0^*,\eta_0} - \tau_i)^2.$$

Chen (2023) shows that

$$\overline{R}_B \equiv \sup\left\{\mathbb{E}_{\tau_i\sim G_{(i)}, Z_i|\tau_i\sim\mathcal{N}(\tau_i,\nu_i^2)}[(\hat{\tau}_{i,G_0^*,\eta_0} - \tau_i)^2] : \nu_i > 0, G_{(i)}, G_0^* \text{ has zero mean and unit variance}\right\}$$

is finite. Taking the expected value with respect to $P_0 \in \mathcal{P}(m_0, s_0)$ and apply the bound $\overline{R_{\mathrm{B}}}$, we have that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2\right] \le \overline{R_{\mathrm{B}}}\frac{1}{n}\sum_{i=1}^n s_0^2(\sigma_i).$$

Note that when P_0 is such that $\theta_i \mid \sigma_i \sim \mathcal{N}(m_0(\sigma_i), s_0^2(\sigma_i))$, the risk of any procedure exceeds the Bayes risk (achieved by (2.12)). Hence, the Bayes risk under this P_0 lower bounds the minimax risk

$$\frac{1}{n}\sum_{i=1}^n \frac{\sigma_i^2}{\sigma_i^2 + s_0^2(\sigma_i)} s_0^2(\sigma_i) \le \inf_{\hat{\theta}_{1:n}} \sup_{P_0\in\mathcal{P}(m_0,s_0)} \mathbb{E}_{P_0}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2\right].$$

Note that, for some $c_{\sigma_\ell, s_u} > 0$,

$$\frac{1}{n}\sum_{i=1}^n \frac{\sigma_i^2}{\sigma_i^2 + s_0^2(\sigma_i)} s_0^2(\sigma_i) = \frac{1}{n}\sum_{i=1}^n \frac{1}{1 + s_0^2(\sigma_i)/\sigma_i^2} s_0^2(\sigma_i) \ge c_{\overline{\rho}}\frac{1}{n}\sum_{i=1}^n s_0^2(\sigma_i).$$

Hence

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2\right] \le \frac{\overline{R}_B}{c_{\overline{\rho}}}\frac{1}{n}\sum_{i=1}^n \frac{\sigma_i^2}{\sigma_i^2 + s_0^2(\sigma_i)} s_0^2(\sigma_i) \le C_{\overline{\rho}} \inf_{\hat{\theta}_{1:n}} \sup_{P_0\in\mathcal{P}(m_0,s_0)} \mathbb{E}_{P_0}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2\right].$$

□

A.4. **Unbiased loss estimation.**

Proposition 1. *Suppose (Y_i, σ_i) obey the Gaussian heteroskedastic location model, assumed to be independent across i (2.3). Fix some $\omega > 0$ and let $Y_{1:n}^{(1)}, Y_{1:n}^{(2)}$ be the coupled bootstrap draws. For some decision problem, let $\boldsymbol{\delta}(Y_{1:n}^{(1)})$ be some decision rule using only data $\left(Y_i^{(1)}, \sigma_{i,(1)}^2\right)_{i=1}^n$. Let $\mathcal{F} = \left(\theta_{1:n}, Y_{1:n}^{(1)}, \sigma_{1:n,(1)}, \sigma_{1:n,(2)}\right)$, for Decision Problems 1 to 3, the estimators $T(Y_{1:n}^{(2)}, \boldsymbol{\delta})$ displayed in Table 1 are unbiased for the corresponding loss:*

$$\mathbb{E}\left[T(Y_{1:n}^{(2)}, \boldsymbol{\delta}(Y_{1:n}^{(1)})) \mid \mathcal{F}\right] = L\left(\boldsymbol{\delta}(Y_{1:n}^{(1)}), \theta_{1:n}\right).$$

Moreover, their conditional variances are equal to those expressions displayed in the third column of Table 1.

Proof. These are straightforward calculations of the expectation. Since every expectation and variance is conditional on $\theta_{1:n}, Y_{1:n}^{(1)}, \sigma_{1:n,(1)}, \sigma_{1:n,(2)}$, we write $\mathbb{E}[\cdot \mid \mathcal{F}]$ and $\mathrm{Var}(\cdot \mid \mathcal{F})$ without ambiguity.

(1) (Decision Problem 1) The unbiased estimation follows directly from the calculation

$$\mathbb{E}\left[(Y_i^{(2)} - \delta_i(Y_{1:n}^{(1)}))^2 \mid \mathcal{F}\right] = (\theta_i^{(2)} - \delta_i(Y_{1:n}^{(1)}))^2 + \sigma_{i,(2)}^2$$

The conditional variance statement holds by definition.

(2) (Decision Problem 2) The unbiased estimation follows directly from the calculation

$$\mathbb{E}\left[\delta_i(Y_{1:n}^{(1)})(Y_i^{(2)} - c_i) \mid \mathcal{F}\right] = \delta_i(Y_{1:n}^{(1)})(\theta_i - c_i).$$

The conditional variance statement follows from

$$\mathrm{Var}\left[\delta_i(Y_{1:n}^{(1)})(Y_i^{(2)} - c_i) \mid \mathcal{F}\right] = \delta_i(Y_{1:n}^{(1)})\sigma_{1:n,(2)}^2.$$

(3) (Decision Problem 3) The loss function for Decision Problem 3 is the same as that for Decision Problem 2 with $c_i = 0$. Since we condition on $Y_{1:n}^{(1)}$, the argument is thus analogous.

$\square$

A.5. **A discrete choice model.** There are n options facing N consumers, where each consumer chooses one option. Each option is characterized by idiosyncratic quality β_j and inherent quality α_j. The latent quality of an option is $\theta_j = \alpha_j + \rho\frac{N_j}{\mathbb{E}[N]}$, where $N_j \le N$ is the number of consumers using option j, generated in equilibrium from a discrete choice model. The term ρN_j reflects externalities generated by the users of an option (congestion). We assume that $\alpha_j, \beta_j \overset{\text{i.i.d.}}{\sim} F$ where μ denotes $\mathbb{E}[\alpha_j + \beta_j]$ and $\sigma_\alpha^2, \sigma_\beta^2, \sigma_{\alpha\beta}$ denotes the variances and covariance of α and β.

To connect this model to our setting, we can imagine that the data analyst has estimates Y_j for θ_j, whose standard errors are a function of N_j. The discrete choice model specifies how N_j selects on the quality component α_j, and ρ determines how θ_j is affected by N_j. We characterize $\mathrm{Cov}(\theta_j, N_j)$ as a function of the primitives $\rho, \mu, \sigma_\alpha, \sigma_\beta, \sigma_{\alpha\beta}$.

Each individual i is endowed with a private type $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iJ})$ of i.i.d Type-1 extreme value random utilities. This prior for ϵ_i is common knowledge and well-specified. $\alpha_{1:n}, \beta_{1:n}, N$ are common knowledge as well. Each individual i is an expected utility maximizer, where the utility of item j is

$$V_j = \left(\alpha_j + \beta_j + \rho\frac{N_{j,-i}}{N-1}\right)\exp\left(\epsilon_{ij}\right)$$

where $N_{j,-i}$ is the number of other individuals choosing item j. Since individuals other than i are symmetric to i, the expected utility (conditional on what i observes) is[63]

$$\mathbb{E}_i V_j = \left(\alpha_j + \beta_j + \rho\pi_{-ij}\right)\exp\left(\epsilon_{ij}\right),$$

[63]Note that the externality that enters the utility is different from the externality in θ. This is for analytical tractability purposes.
To prevent the utility component from becoming negative, we additionally assume that $\alpha_j + \beta_j > -\rho$ almost surely, which imposes that $\rho > -\mu$.

where π_{-ij} is i's prior expectation of $N_{j,-i}/(N-1)$. A Bayes-Nash equilibrium is one in which individual i chooses the option with the highest $\mathbb{E}_i V_j$ and his beliefs about other individuals, π_{-ij}, are correct.

Since individuals are ex-ante symmetric, we assume that

$$\pi_{-ij} = \pi_j = \mathrm{P}(\mathbb{E}_i V_j \geq \mathbb{E}_i V_k \quad \forall k).$$

In such a symmetric equilibrium, π solves the system of equations

$$\frac{\alpha_j + \beta_j + \rho\pi_j(N-1)}{\sum_j \alpha_j + \beta_j + \rho\pi_j(N-1)} = \pi_j \implies \pi_j = \frac{\alpha_j + \beta_j}{\sum_j \alpha_j + \beta_j}.$$

Finally, we assume that the total number of consumers is ex ante random

$$N \mid (\alpha_{1:n}, \beta_{1:n}) \sim \mathrm{Pois}\left(\lambda \cdot \left(\sum_{j=1}^n \alpha_j + \beta_j\right)\right).$$

Assume that the data-generating process draws α, β, N, and individuals play the Bayes–Nash equilibrium under symmetric beliefs π. By the thinning property of Poisson processes, we have that

Lemma A.1. $N_j \mid (\alpha_{1:n}, \beta_{1:n}) \sim \mathrm{Pois}\left(\lambda(\alpha_j + \beta_j)\right)$ *independently across* j.

Now, under this process, we can compute the covariance between the latent quality θ_j and the sample size N_j in closed form:

$$\begin{aligned}
\mathrm{Cov}\left(\theta_j, N_j\right) &= \underbrace{\mathrm{Cov}(\alpha_j, N_j)}_{\text{selection}} + \underbrace{\frac{\rho}{\lambda n \mu}\mathrm{Var}(N_j)}_{\text{congestion}} \\
&= \lambda(\sigma_\alpha^2 + \sigma_{\alpha\beta}) + \frac{\rho}{\lambda n \mu}\left[\lambda\mu + \lambda^2(\sigma_\alpha^2 + \sigma_\beta^2 + 2\sigma_{\alpha\beta})\right]
\end{aligned}$$

This is positive—meaning that the latent quality is positively associated with precision—iff

$$\frac{\rho}{\lambda n \mu} > -\frac{\mathrm{Cov}(\alpha_j, N_j)}{\mathrm{Var}(N_j)} = -\frac{\sigma_\alpha^2 + \sigma_{\alpha\beta}}{\mu + \lambda(\sigma_\alpha^2 + \sigma_\beta^2 + 2\sigma_{\alpha\beta})}.$$

When the selection effect is positive ($\mathrm{Cov}(\alpha_j, N_j)$), the above display requires the externality ρ to not be too negative so as to dominate the selection effect. Note that the sign of the selection contribution depends on the covariance between α and β, and thus could be negative. Moreover, if α instead were an undesirable trait to consumers, then the selection effect may also be negative. The congestion effect similarly does not have to be negative. We allow for positive spillovers by $\rho > 0$.

We can interpret various empirical observations through this model:

- For hospital value-added (Chandra et al., 2016), N_j positively selects on hospital quality α_j. This is likely true for most value-added settings.
- For teacher value-added, it is possible (Lazear, 2001; Barrett and Toma, 2013; Mehta, 2019) that teachers may prefer smaller classes, and school administrators may reward good

teachers by letting them teach smaller classes. In the lens of this model, N_j negatively selects on quality.[64]

- In integenerational mobility, N_j is the number of poor minority households. Higher N_j leads to oppressive institutions and residential segregation. We can interpret these pernicious effects as a negative ρ.

However, this model does not capture all channels through which θ_j can be correlated with σ_j. For instance, the following is difficult to map to the discrete choice model.

- In unbalanced panel data settings, the length of the observed period for a unit—which relates to the precision of the unit's estimated fixed effect—may be correlated with the underlying fixed effect. This observation dates at least to Olley and Pakes (1996), who note that in a firm panel, those firms with shorter observed period are probably less productive and have to shut down sooner. For value-added modeling of nursing homes, Einav et al. (2022) note that patients with shorter stays at nursing homes typically experience an adverse health event, including death. Such events are presumably more likely for worse nursing homes, again inducing a correlation between nursing home qualities and the sample sizes used to estimate them. Similarly, for teacher value-added, Bruhn et al. (2022) find that teachers who have shorter observed spells in administrative datasets tend to be worse and have noisier value added estimates.

A.6. **Interpretation of empirical Bayes sampling model.** When the empirical Bayes sampling model fails to hold, empirical Bayes methods do not precisely mimic an oracle Bayesian's decision. However, in many cases, we can still interpret the empirical Bayes decision rules. In most such cases, the interpretation is in terms of emulating an oracle Bayesian who is *constrained*. The oracles are constrained either by removing its access to certain information or by restricting its decisions to a particular class. We will consider two scenarios when such an interpration is natural.

A.6.1. *Interpretation when independence of units fails.* We consider the interpretation of the sampling model (2.3) when it is misspecified. Recall that we assume $(Y_i, \theta_i, \sigma_i)$ are sampled independently across i, with $Y_i \mid \theta_i, \sigma_i \sim \mathcal{N}(\theta_i, \sigma_i)$. This sampling model can fail in two ways. First, it is possible that $Y_{1:n} \mid \theta_{1:n}, \sigma_{1:n}$ are correlated but still multivariate Gaussian. Second, it is possible that (θ_i, σ_i) are correlated across i. Here, we limit our discussion to Decision Problem 1.

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$. Let us assume instead that

$$\boldsymbol{Y} \mid \boldsymbol{\theta}, \Sigma \sim \mathcal{N}(\boldsymbol{\theta}, \Sigma)$$

where $\operatorname{diag}(\Sigma) = [\sigma_1^2, \ldots, \sigma_n^2]$ and the variance-covariance matrix Σ is known. Let Q_0 be the joint distribution of $\boldsymbol{\theta} \mid \Sigma$. Now, the oracle Bayesian—who knows Q_0—would use $\mathbb{E}_{Q_0}[\theta_i \mid \boldsymbol{Y}, \Sigma]$ as their decision rule. The empirical Bayesian can similarly emulate that oracle Bayes decision rule by estimating Q_0. If the empirical Bayesian is willing to assume that the location-scale assumption

[64]Though the channel is not through student-level discrete choice of teachers.

(2.6) describes Q_0:

$$(\theta_i \mid \Sigma) \sim (\theta_i \mid \sigma_{1:n}) \sim G_0\left(\frac{\cdot - m_0(\sigma_i)}{s_0(\sigma_i)}\right),$$

then the empirical Bayesian can similarly implement CLOSE, and output estimates of $\mathbb{E}_{Q_0}[\theta_i \mid \boldsymbol{Y}, \Sigma]$. We should caveat that the NPMLE step no longer maximizes the full likelihood of $\boldsymbol{Y}$ with respect to G_0, but a quasi-likelihood that averages over the log-likelihood of each Y_i separately, ignoring their joint distribution.

Now, let us consider what interpretation our method has when we erroneously assume either the independence of Y_i across i or that $\boldsymbol{\theta}_i \mid \Sigma$ are independent across i. The latter independence may fail, for instance, when the populations index places, and the θ_i's are thought to be spatially correlated (e.g., in Müller and Watson, 2022). Consider the class of *separable* decision rules, where the forecast for θ_i can depend solely on Y_i, σ_i:

$$\delta_i(\boldsymbol{Y}, \sigma_{1:n}) = \delta_i(Y_i, \sigma_i).$$

Consider a constrained oracle Bayesian who is forced to use a separable decision rule. They would use $\mathbb{E}_{Q_0}[\theta_i \mid Y_i, \sigma_i]$. Note that this constrained decision rule depends on Q_0 only through the distribution $\theta_i \mid \sigma_i$ (and not $\theta_i \mid \Sigma$). Thus, under the location-scale assumption

$$(\theta_i \mid \sigma_{1:n}) \sim G_0\left(\frac{\cdot - m_0(\sigma_i)}{s_0(\sigma_i)}\right),$$

CLOSE-based methods emulate this oracle Bayesian constrained to separable decision rules. Of course, the resulting empirical Bayesian decision rule is *not* separable (since $\hat{G}_n$ presumably depends on all the data), but it seeks to emulate the best possible separable rule. This interpretation in terms of emulating a constrained oracle Bayesian holds regardless of the joint distribution of $\boldsymbol{Y}$ or of $\boldsymbol{\theta}$, so long as our specification of the marginal distribution holds. Of course, our regret results do not immediately carry over to this setting.

A.6.2. *Interpretation with additional covariates X_i.* Additionally, we may also have population-level covariates X_i. Let us maintain that X_i does not predict the noise in Y_i:

$$Y_i \perp\!\!\!\perp X_i \mid \theta_i, \sigma_i.$$

Here, we will discuss two questions. First, how do we handle covariates? Second, what is the difference between using X_i and σ_i—is the standard error simply a covariate?[65]

On the first question, there are two ways of incorporating covariates, under similar but distinct assumptions. First, CLOSE-methods can be extended to incorporate covariates by augmenting (2.6)

[65]Covariates are considered in Ignatiadis and Wager (2019). They assume a homoskedastic setting where the prior depends on some covariates X_i: i.e., in our notation, $\theta_i \mid X_i \sim \mathcal{N}(m(X_i), s_0^2)$ and $Y_i \mid \theta_i \sim \mathcal{N}(\theta_i, \sigma^2)$. Starting from our setting (2.6), to obtain theirs, one would (i) restrict to homoskedasticity $\sigma_i = \sigma$, (ii) consider some covariates X_i that predict θ_i, and model $\theta_i \mid X_i$ as a conditional location—but not scale—family, and (iii) restrict $G_0 \sim \mathcal{N}(0, 1)$.
Their minimax lower bound on the regret uses essentially the same argument as we do in Theorem 2.

to incorporate covariates. That is, we can instead assume that

$$P_0(\theta_i \le t \mid X_i, \sigma_i) \sim G_0\left(\frac{t - m_0(\sigma_i, X_i)}{s_0(\sigma_i, X_i)}\right) \tag{A.1}$$

and estimate m_0, s_0 nonparametrically. Instead of being one-dimensional nonparametric regression problems, they are now $(d+1)$-dimensional nonparametric problems. Under the same Hölder-type smoothness conditions, the corresponding regret rate replaces $n^{-\frac{2p}{2p+1}}$ with $n^{-\frac{2p}{2p+1+d}}$. Second, as we do in the empirical exercises, one could consider a strategy of residualizing against X_i in some arbitrary way, performing empirical Bayes, and undoing the residualization. This strategy dates back to Fay and Herriot (1979). That is, with raw data $\tilde{Y}_i$ for parameter ϑ_i, we can consider forming the residuals $Y_i = \tilde{Y}_i - b(X_i)$ and $\theta_i = Y_i - b(X_i)$, and perform empirical Bayes methods on $(Y_i, \theta_i, \sigma_i)$. At a high level, we can rationalize this strategy as mimicking a constrained oracle Bayesian who solely has access to Y_i, σ_i, who knows the joint distribution of (θ_i, σ_i), but who does not have access to X_i. Note that this interpretation is coherent regardless of the transformation $b(X_i)$, allowing us to be more blasé about modeling X_i than the previous approach. In particular, choosing $b(X_i) = 0$ ignores the covariate entirely; the resulting empirical Bayes procedure mimics an oracle that does not have access to X_i. Of course, when we impose the location-scale assumption (2.6) on (θ_i, σ_i), different $b(X_i)$ gives rise to different—and possibly mutually exclusive—underlying models on $(\vartheta_i, \sigma_i, X_i)$.

On the second question, in an operational sense, σ_i is simply another covariate. σ_i is not particularly special in the assumption (A.1), and one interpretation of CLOSE is treating σ_i precisely as a covariate to be regressed out. However, σ_i does occupy a special place in the statistical structure of the problem. The *likelihood* of the data, $Y_i \mid \theta_i, \sigma_i$, depends on σ_i but not X_i. This special role of σ_i means that we must treat it with more care so that the resulting procedure has a coherent interpretation. If we wanted to ignore covariates X_i, we can imagine an oracle Bayesian who does not have access to X_i, and the resulting empirical procedure simply mimics that constrained oracle. This line of reasoning does not work with σ_i, since any oracle Bayesian—constrained or otherwise—must have access to σ_i. As a result, we cannot avoid the problem of modeling $\theta_i \mid \sigma_i$ as easily as we could have avoided modeling $\theta_i \mid X_i, \sigma_i$ by changing the goalpost.

A.7. **Alternatives to CLOSE.**

A.7.1. *Alternative methods.* Let us turn to a few specific alternative methods that consider failure of prior independence. We argue that they do not provide a free-lunch improvement over our assumptions. At a glance, these alternative methods have properties summarized in Table 2.

Alternative 1 (Working with t-ratios)**.** We may consider normalizing σ_i away by working with t-ratios $T_i \equiv \frac{Y_i}{\sigma_i} \mid (\sigma_i, \theta_i) \sim \mathcal{N}(\theta_i/\sigma_i, 1)$. The resulting problem is homoskedastic by construction. It is natural to consider performing empirical Bayes shrinkage assuming that $\frac{\theta_i}{\sigma_i} \overset{\text{i.i.d.}}{\sim} H_0$, and use, say, $\sigma_i \mathbf{E}_{\hat{H}_n}\left[\frac{\theta_i}{\sigma_i} \mid T_i\right]$ as an estimator for the posterior mean of θ_i (Jiang and Zhang, 2010). However, such an approach approximates the optimal decision rule within a restricted class on a different objective.

Table 2. Properties of alternative methods

	t-ratios	Var. stab. transforms	Random $\hat{\sigma}_i$	SURE
Restrict to a class of procedures	✗			✗
Change the loss function	✗	✗		
Require access to micro-data			✗	
Assume θ_i is independent from some other known nuisance parameter, e.g. n_i		✗	✗	
Parametric restrictions on the micro-data		✗	✗	

Let us restrict decision rules to those of the form $\delta_{i,\text{t-stat}}(Y_i, \sigma_i) = \sigma_i h(Y_i/\sigma_i)$. The oracle Bayes choice of h is $h^\star(T_i) = \frac{\mathbb{E}[\sigma_i\theta_i|T_i]}{\mathbb{E}[\sigma_i^2|T_i]}$. However, $h^\star$ is not the posterior mean of θ_i/σ_i given the t-ratio T_i, unless $\sigma_i^2 \perp\!\!\!\perp \theta_i/\sigma_i$. On the other hand, the loss function that does rationalize the posterior mean $h(T_i) = \mathbb{E}[\theta_i/\sigma_i \mid T_i]$ is the precision-weighted compound loss $L(\boldsymbol{\delta}, \theta_{1:n}) = \frac{1}{n}\sum_{i=1}^n \sigma_i^{-2}(\delta_i - \theta_i)^2$. Thus, rescaling posterior means on t-ratios achieves optimality for a weighted objective among a restricted class of decision rules $\delta_{i,\text{t-stat}}$. ∎

Alternative 2 (Variance-stabilizing transforms)**.** Second, we may consider a variance-stabilizing transform when the underlying micro-data are Bernoulli and θ_i is a Bernoulli mean (Efron and Morris, 1975; Brown, 2008). Specifically, we rely on the asymptotic approximation

$$\sqrt{n_i}(Y_i - \theta_i) \xrightarrow[n_i\to\infty]{d} \mathcal{N}(0, \theta_i(1-\theta_i)).$$

A variance-stabilizing transform can disentangle the dependence: Let $W_i = 2\arcsin(\sqrt{Y_i})$ and $\omega_i = 2\arcsin(\sqrt{\theta_i})$, and, by the delta method,

$$\sqrt{n_i}\,(W_i - \omega_i) \xrightarrow[n_i\to\infty]{d} \mathcal{N}(0,1). \quad \text{Thus, approximately, } W_i \mid \omega_i, n_i \sim \mathcal{N}\left(\omega_i, \frac{1}{n_i}\right).$$

One might consider an empirical Bayes approach on the resulting W_i. Note that W_i may still violate prior independence, since ω_i may not be independent of n_i. Moreover, squared error loss on estimating $\omega_i = 2\arcsin(\sqrt{\theta_i})$ is different from squared error loss on estimating θ_i. We do not know of any guarantees for the loss function on θ_i, $\frac{1}{n}\sum_{i=1}^n(\delta_i - \sin^2(\omega_i/2))^2$, when we perform empirical Bayes analysis on ω_i. ∎

Alternative 3 (Treating the standard error as estimated)**.** Lastly, if the researcher has access to micro-data, Gu and Koenker (2017) and Fu et al. (2020) propose empirical Bayes strategies that treat σ_i as noisy as well, in which we know the likelihood of (Y_i, σ_i). This approach allows for dependence between θ_i and σ_i but assumes independence between (θ_i, σ_i) and some other known nuisance parameter. To describe their model, we introduce more notation. Let $Y_{ij}, j = 1, \ldots, n_i$, denote the micro-data for population i, where, for each i, we are interested in the mean of Y_{ij}. Let Y_i denote their sample mean and S_i^2 denote their sample variance, where $\sigma_i^2 = S_i^2/n_i$. Let σ_{i0}^2 denote the true variance of observations from population i.

Both papers work under Gaussian assumptions on the micro-data. This parametric assumption[66] on the micro-data—which is stronger than we require—implies that $Y_i \perp\!\!\!\perp S_i^2 \mid (\sigma_{i0}, \theta_i, n_i)$ with marginal distributions:

$$Y_i \mid \sigma_{i0}, \theta_i, n_i \sim \mathcal{N}\left(\theta_i, \frac{\sigma_{i0}^2}{n_i}\right) \qquad S_i^2 \mid \sigma_{i0}, \theta_i, n_i \sim \text{Gamma}\left(\frac{n_i - 1}{2}, \frac{1}{2\sigma_{i0}^2}\right).$$

They then propose empirical Bayes methods treating $\mathbf{Y}_i \equiv (Y_i, S_i^2)$ as noisy estimates for parameters $\boldsymbol{\theta}_i \equiv (\theta_i, \sigma_{i0}^2)$. This formulation allows $\boldsymbol{\theta}_i$ to have a flexible distribution, and thus allows for dependence between θ_i and σ_{i0}^2. However, since the known sample size n_i enters the likelihood of $\mathbf{Y}_i$, this approach still assumes that $n_i \perp\!\!\!\perp \boldsymbol{\theta}_i$. ■

This discussion is not to say that CLOSE is necessarily preferable to these alternatives. It highlights that the possible dependence between θ_i and σ_i cannot be easily resolved. As summarized in Table 2, existing alternatives compromise on optimality, use a different loss function, or implicitly assume θ_i is independent from components of σ_i^2 (e.g., n_i). Of course, depending on the empirical context, these may well be reasonable features.

In contrast, our approach models $\theta_i \mid \sigma_i$ directly via the location-scale assumption (2.6). A natural question is whether other types of modeling may be superior—which we turn to next. We argue that the location-scale model uniquely capitalizes on the appealing properties of the NPMLE-based empirical Bayes approaches.

A.7.2. *Alternative models for $\theta_i \mid \sigma_i$.* One alternative is simply treating the joint distribution of (θ_i, σ_i) fully nonparametrically. For instance, an f-modeling approach with Tweedie's formula[67] implies that an estimate of the conditional distribution $Y_i \mid \sigma_i$ is all one needs for computing the posterior means (Brown and Greenshtein, 2009; Liu et al., 2020; Luo et al., 2023). However, conditional density estimation is a challenging problem, and most available methods do not exploit the restriction that $Y_i \mid \sigma_i$ is a Gaussian convolution. Similarly, one could consider flexible parametric

[66] The parametric restriction on the micro-data Y_{ij} can be relaxed by appealing to the asymptotic distribution of (Y_i, S_i^2)—resulting in the Gaussian likelihood $(Y_i, S_i^2) \mid \boldsymbol{\theta}_i, \Sigma_i \sim \mathcal{N}(\boldsymbol{\theta}_i, \Sigma_i)$. In general, however, Σ_i also depends on n_i and higher moments of Y_{ij}, which again may not be independent of $\boldsymbol{\theta}_i$.

[67] That is, the posterior mean can be written as a functional of the density of Y:

$$\mathbb{E}[\theta_i \mid Y_i, \sigma_i] = Y_i + \sigma_i^2 \frac{d}{dy} \log f(y \mid \sigma_i)\bigg|_{y=Y_i},$$

where $f(y \mid \sigma)$ is the conditional density of $Y \mid \sigma$. Empirical Bayes approaches exploiting this formula is known as *f-modeling* (Efron, 2014), since f usually denotes the marginal distribution of Y. This is in contrast to *g-modeling*, which seeks to estimate the prior distribution of θ_i.

Brown and Greenshtein (2009) develop an f-modeling approach with a kernel smoothing density estimator in the homoskedastic setting. Liu et al. (2020) extend this approach to a homoskedastic, balanced dynamic panel setting, where the initial outcome for each unit acts as a known nuisance parameter, much like σ_i in our case. Brown and Greenshtein (2009) and Liu et al. (2020) show that the squared error Bayes regret converges to zero faster than the oracle Bayes risk. These guarantees do not imply regret rate characterizations similar to those that we obtain. See Jiang and Zhang (2009) for additional discussion about the strengths of the theoretical results in Brown and Greenshtein (2009) compared to NPMLE-based g-modeling approaches.

g-modeling of $\theta_i \mid \sigma_i$ in the vein of the log-spline sieve of Efron (2016).[68] This has the advantage of estimating a smooth prior at the cost of having tuning parameters. We are not aware of regret results for this approach.

If we commit to making some substantive restriction on the joint distribution of (θ_i, σ_i), it is fair to ask why the conditional location-scale restriction (2.6) is necessarily preferable. However, if we wish to capitalize on the theoretical and computational advantages of NPMLE, it is natural to consider a class of procedures that transform the data in some way and use the NPMLE on the resulting transformed data to estimate the prior distribution (Appendix A.7.3 gives a heuristic justification for this strategy). If we wish to preserve the Gaussian location model structure on the transformed data, then effectively we can only consider affine transformations (i.e., $Z = a(\sigma) + b(\sigma)Y$) (shown in Lemma A.2 below). If we further wish that Z obeys a Gaussian location model in which prior independence holds (i.e., $\tau \equiv a(\sigma) + b(\sigma)\theta$ is independent from $\nu \equiv b(\sigma)\sigma$)—so that we can apply NPMLE-based approaches assuming prior independence—then we have no other choice but to assume (2.6). Thus, the conditional location-scale assumption is uniquely well-suited to capitalize on the favorable properties of NPMLE already established in the literature, which we extend via Theorem 1.

Lemma A.2. *Let $Y \sim \mathcal{N}(\theta, \sigma^2)$ with known σ^2. Consider a strictly increasing and differentiable function $g(\cdot)$. Let $Z = h(Y)$. Then the corresponding family of distributions of Z is a natural exponential family if and only if $h(Y) = a + bY$.*

Proof. The "if" part ($\Longleftarrow$) is immediate. We focus on the "only if" ($\Longrightarrow$) part. Writing the distribution of Y as an exponential family,

$$p_Y(y) \propto \exp\left(y\frac{\theta}{\sigma^2} + g(y,\sigma) + A(\theta,\sigma)\right)$$

for some $g(y,\sigma)$ and $A(\theta,\sigma)$. Note that we have

$$p_Z(z) = p_Y(y)\left|\frac{dy}{dz}\right| = p_Y(h^{-1}(z))\frac{dh^{-1}(z)}{dz}$$

Thus, writing in exponential family form, for some $\tilde{g}$, we have that

$$p_Z(z) \propto \exp\left(h^{-1}(z)\frac{\theta}{\sigma^2} + \tilde{g}(z,\sigma) + A(\theta,\sigma)\right)$$

Suppose Z follows a natural exponential family with natural parameter $q(\theta;\sigma)$. Then we can write

$$h^{-1}(z)\frac{\theta}{\sigma^2} = zq(\theta;\sigma) + v(\theta,\sigma) + w(z).$$

Since h is strictly monotone and differentiable, so is h^{-1}. Taking the z-derivative of both sides:

$$\frac{dh^{-1}}{dz} = \frac{\sigma^2}{\theta}q(\theta;\sigma) + w'(z)\frac{\sigma^2}{\theta}.$$

[68] Generalizing Efron (2016), we may model $g(\theta \mid \sigma) \propto \exp(\sum_{j=1}^J a_j(\sigma;\alpha_j)p_j(\theta))$ where $p_1,\ldots,p_J$ are flexible sieve expansions (e.g. spline basis functions) and $a_j(\sigma;\alpha_j)$ are flexible functions indexed by finite-dimensional parameters α_j. The parameters $\alpha_1,\ldots,\alpha_J$ can be estimated by maximizing the penalized likelihood of $Y_{1:n}$.

Since the left-hand side does not depend on θ, it follows that

$$\frac{q(\theta;\sigma)+w'(z)}{\theta}$$

is free of θ for all z. Suppose $w'(z)$ is not constant, then for $z_1 \neq z_2$ and $w'(z_1) \neq w'(z_2)$, the difference is θ-dependent

$$\frac{q(\theta;\sigma)+w'(z_1)}{\theta} - \frac{q(\theta;\sigma)+w'(z_2)}{\theta} = \frac{w'(z_1)-w'(z_2)}{\theta}.$$

Hence $w'(z)$ is a constant. As a result, $\frac{dh^{-1}}{dz}$ does not depend on z, and hence $h(z) = a + bz$. ☐

A.7.3. *Model-free interpretation of* CLOSE-NPMLE. When the location-scale model fails to hold, it remains sensible to consider estimating the NPMLE on an affine transformation of the data, as in CLOSE-NPMLE.

Let us first consider a given affine transformation of the data—not necessarily $\tau = \frac{Z-m_0(\sigma)}{s_0(\sigma)}$—into (Z_i, τ_i, ν_i) for which $\tau_i \mid \nu_i \sim H_{(i)}$, and ask why NPMLE is reasonable. In population, NPMLE seeks to minimize the average Kullback–Leibler (KL) divergence between the distribution of the estimates Z_i and the distribution implied by the convolution $H \star \mathcal{N}(0, \nu_i^2)$:

$$\max_H \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{Z_i \sim f_{H_{(i)},\nu_i}} \left[\log f_{H,\nu_i}(Z_i)\right], \text{ equivalent to } \min_H \frac{1}{n}\sum_{i=1}^n \mathrm{KL}\left(f_{H_{(i)},\nu_i} \,\|\, f_{H,\nu_i}\right),$$

where $f_{H,\nu}$ is the density of the convolution $H \star \mathcal{N}(0,\nu^2)$. As shown by Jiang and Zhang (2009) and Jiang (2020) (see Appendix C.3), the regret in mean-squared error under a misspecified prior $\tau_i \sim H$ is upper bounded by the average squared Hellinger distance between the distribution of the data and the distribution implied by H. The average Hellinger distance is further upper bounded by the average KL divergence:

$$\frac{1}{n}\sum_{i=1}^n h^2\left(f_{H_{(i)},\nu_i}, f_{H,\nu_i}\right) \leq \frac{1}{n}\sum_{i=1}^n \mathrm{KL}\left(f_{H_{(i)},\nu_i} \,\|\, f_{H,\nu_i}\right).$$

In this sense, even under misspecification ($H_{(i)} \neq H_{(j)}$), NPMLE chooses a common distribution H that minimizes an upper bound of regret.

Now that we have a justification for the NPMLE, let us consider the transformation we would like to choose. It is reasonable, then, to choose the affine transform $(a(\sigma), b(\sigma))$ so that the resulting conditional distributions $H_{(i)}$ of the transformed parameter $\tau_i \mid \sigma_i$ are similar—under some distance measure. Doing so does not recover prior independence on the transformed data but limits the extent of non-independence. Choosing $a(\sigma), b(\sigma)$ to ensure that $\tau_i \mid \sigma_i$ has the same first two moments is intuitively reasonable, and actually has a formal interpretation in terms of information-theoretic divergences and optimal transport metrics, at least in a large-σ regime (Chen and Niles-Weed, 2022).

Part 2. Additional empirical results

Appendix B. Additional empirical exercises

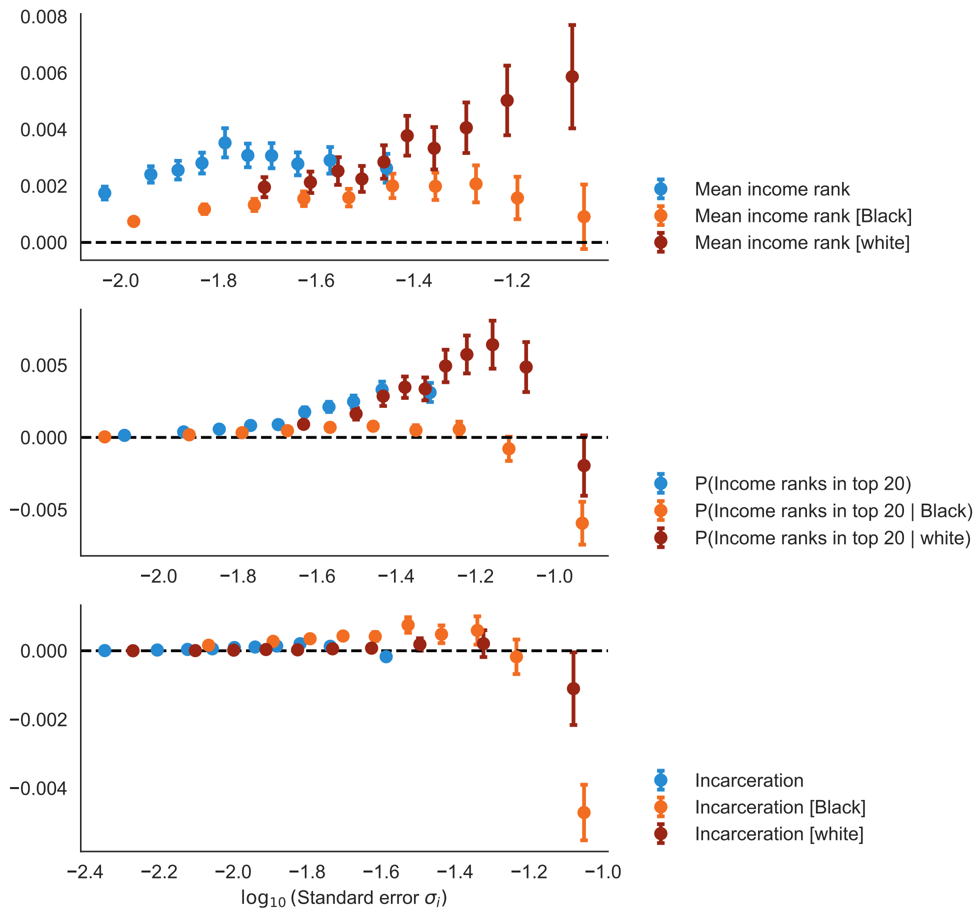


Figure B.1. Estimated conditional variance $s_0^2(\sigma)$, binned into deciles, with 95% uniform confidence intervals shown.

B.1. **Positivity of $s_0(\cdot)$ in the Opportunity Atlas data.** In the Opportunity Atlas data, we often observe that the estimated conditional variance is negative: $\hat{s}_0^2 < 0$. To test if this is due to sampling variation or underdispersion of the Opportunity Atlas estimates relative to the estimated standard error, we consider the following upward-biased estimator of $s_0^2(\sigma_i)$. Without loss, let us sort the Y_i, σ_i by σ_i, where $\sigma_1 \le \cdots \le \sigma_n$. Let $S_i = \frac{1}{2}\left[(Y_{i+1} - Y_i)^2 - (\sigma_i^2 + \sigma_{i+1}^2)\right]$. Note that

$$\mathbb{E}[S_i \mid \sigma_{1:n}] = \frac{1}{2}\mathbb{E}[(\theta_{i+1} - \theta_i)^2 \mid \sigma_{1:n}] = \frac{s_0^2(\sigma_{i+1}) + s_0^2(\sigma_i)}{2} + \frac{1}{2}(m_0(\sigma_{i+1}) - m_0(\sigma_i))^2 \ge \frac{s_0^2(\sigma_{i+1}) + s_0^2(\sigma_i)}{2}.$$

Hence S_i is an overestimate of the successive averages of $s_0(\sigma)$. Figure B.1 plot the estimated conditional expectation of S_i given σ_i, using a sample of $(S_1, S_3, S_5, \ldots)$ so that the S_i's used are mutually independent. We see that for many measures of economic mobility, we can reject $\mathbb{E}[S_i \mid \sigma_i] \ge 0$, indicating some overdispersion in the data.

What % of Naive-to-Oracle MSE gain do we capture?

	Indep-Gauss	Indep-NPMLE	CLOSE-Gauss	CLOSE-NPMLE (with $\hat{E}[(Y-\hat{m})^2-\sigma^2 \mid \sigma]$)	CLOSE-NPMLE	CLOSE-NPMLE (Estimated prior standardized)
Mean income rank	85.0	88.4	91.4	91.7	91.8	91.7
Mean income rank [white]	87.0	90.3	94.2	95.0	95.1	94.9
Mean income rank [Black]	81.9	88.5	93.2	93.4	93.5	92.9
Mean income rank [white male]	89.4	92.3	93.5	94.9	94.9	94.7
Mean income rank [Black male]	82.9	85.9	92.6	93.6	93.7	93.6
P(Income ranks in top 20)	57.7	80.8	91.4	92.8	92.9	92.9
P(Income ranks in top 20 | white)	74.6	80.3	93.8	94.9	94.9	94.8
P(Income ranks in top 20 | Black)	46.0	53.0	95.4	97.8	97.5	97.2
P(Income ranks in top 20 | white male)	69.6	75.7	90.2	93.5	93.6	93.4
P(Income ranks in top 20 | Black male)	36.8	44.8	94.4	97.5	97.0	96.6
Incarceration	50.6	58.9	88.2	91.2	91.0	90.7
Incarceration [white]	73.9	80.7	91.2	96.3	96.8	95.1
Incarceration [Black]	47.8	52.4	96.4	97.9	97.4	97.2
Incarceration [white male]	59.6	64.0	93.2	97.4	97.6	96.8
Incarceration [Black male]	41.7	49.3	96.0	96.6	96.3	96.2
Column median	69.6	80.3	93.2	94.9	94.9	94.8

FIGURE B.2. Additional CLOSE-NPMLE variants for the calibrated simulation in Section 5. Here the results average over 100 replications.

B.2. **Robustness checks for the calibration exercise in Section 5.** In Figure B.2, we evaluate two variants of CLOSE-NPMLE. The first variant (column 4) uses an estimator for $s_0(\cdot)$ that smoothes the difference $(Y - \hat{m}(\sigma))^2 - \sigma^2$, rather than smoothing $(Y - \hat{m}(\sigma))^2$ and then subtracting σ^2. Since local linear regression suffers from bias coming from the convexity of the underlying unknown function, smoothing the difference can perform better, as the convexity bias differences out. The second variant (column 6) projects the estimated NPMLE $\hat{G}_n$ to the space of mean zero and variance one distributions, by normalizing by its estimated first and second moments. Neither variant performs appreciably differently from the main version of CLOSE-NPMLE (column 5) that we demonstrate in the main text.

B.3. **Simulation exercise setup.** This section describes the details of the simulation exercise in Section 5. We restrict to the 10,109 tracts within the twenty largest Commuting Zones. Tracts with missing information are dropped for each measure of mobility. Specifically, the simulated data-generating process is as follows:

(Sim-1) Residualize $\tilde{Y}_i$ against some covariates X_i to obtain β and residuals Y_i. Estimate the conditional moments m_0, s_0 on (Y_i, σ_i) via local linear regression, described in Appendix G.

(Sim-2) Partition σ into vingtiles. Within each vingtile j, estimate an NPMLE G_j over the data $\left(\frac{Y_i - m_0(\sigma_i)}{s_0(\sigma_i)}, \frac{\sigma_i}{s_0(\sigma_i)}\right)$ and normalize G_j to have zero mean and unit variance. Sample $\tau_i^* \mid \sigma_i \sim G_j$ if observation i falls within vingtile j.

(Sim-3) Let $\vartheta_i^* = s_0(\sigma_i)\tau_i^* + m_0(\sigma_i) + \beta' X_i$ and let $\tilde{Y}_i^* \mid \theta_i^*, \sigma_i \sim \mathcal{N}(\theta_i^*, \sigma_i^2)$.

The estimated β, m_0, s_0 will serve as the basis for the true data-generating process in the simulation, and as a result we do not denote it with hats.

The covariates used are poverty rate in 2010, share of Black individuals in 2010, mean household income in 2000, log wage growth for high school graduates, mean family income rank of parents, mean family income rank of Black parents, the fraction with college or post-graduate degrees in 2010, and the number of children—and the number of Black children—under 18 living in the given tract with parents whose household income was below the national median. These covariates are included in Chetty et al.'s (2020) publicly available data, and these descriptions are from their codebook. This set of covariates is not precisely the same as what is used in Bergman et al. (2023). Bergman et al. (2023) additionally use economic mobility estimates for a later birth cohort, which are not included in the publicly released version of the Opportunity Atlas. The "number of children" variables are used by (Chetty et al., 2020) as a population weighting variable; they contain some information on the implicit micro-data sample sizes n_i.

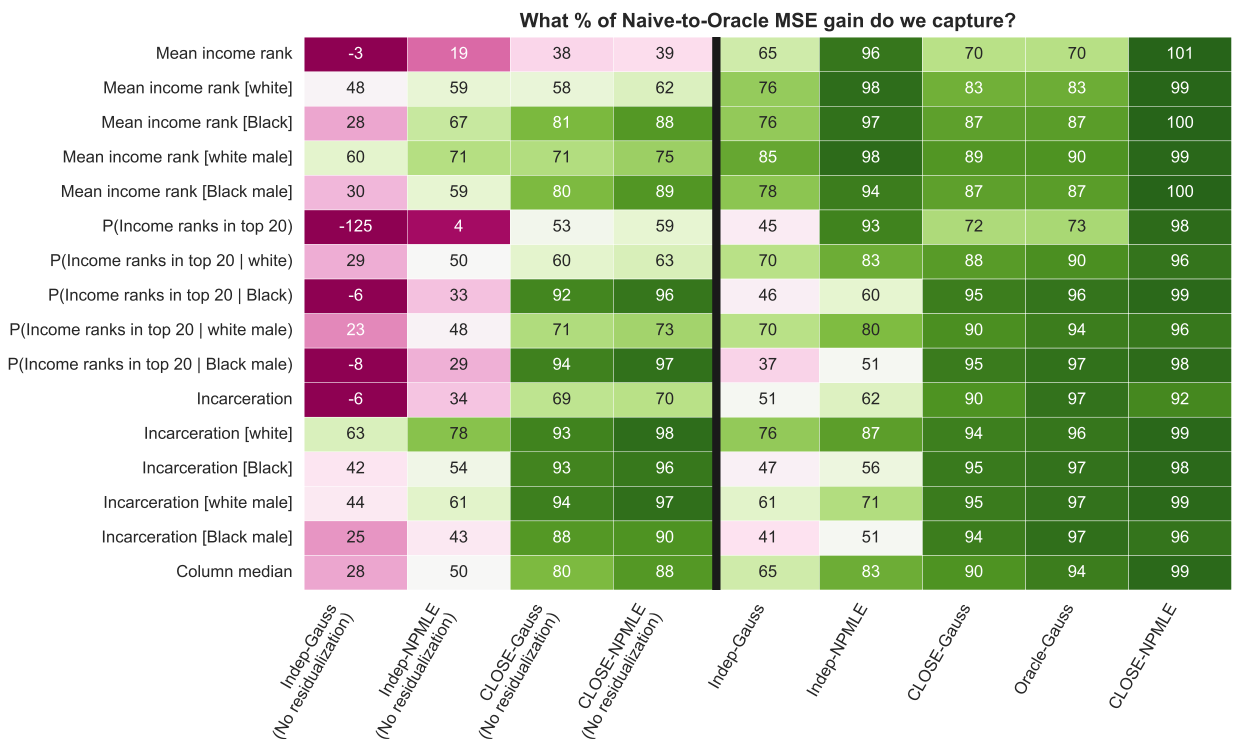


FIGURE B.3. Analogue of Figure 4 for the data-generating process in Appendix B.4. Here the results average over 100 replications.

B.4. **Different Monte Carlo setup.** We have also conducted a Monte Carlo exercise where we replace (Sim-2) with the following step:

- For each σ_i, let
$$\alpha_i = \frac{1}{2} + \frac{1}{2}\frac{m_0(\sigma_i) - \min_i m_0(\sigma_i)}{\max_i m_0(\sigma_i) - \min_i(\sigma_i)} \in [1/2, 1]$$
We sample $\tau_i^* \mid \sigma_i$ as a scaled and shifted Weibull distribution with shape α_i. The scaling and translation ensures that $\tau_i \mid \sigma_i$ has mean zero and variance one. Because we choose the Weibull distribution, the shape parameter α_i corresponds exactly to α in Assumption 2. Our choices of α_i implies that $\tau_i \mid \sigma_i$ has thicker tails than exponential and does not have a moment-generating function.

The Weibull distribution has thicker tails and is skewed, and as a result, NPMLE-based methods tend to greatly outperform methods based on assuming Gaussian priors. Figure B.3 show the analogue of Figure 4 for this data-generating process. Indeed, we see that INDEPENDENT-NPMLE improves over INDEPENDENT-GAUSS considerably, and similarly for CLOSE-NPMLE and ORACLE-GAUSS.

B.5. **MSE in validation exercise with coupled bootstrap.** We compare empirical Bayes procedures for the squared error estimation problem (Decision Problem 1), in the setting of the validation exercise in Section 5. Since this is an empirical application on real, rather than synthetic, data, we no longer have access to oracle estimators. As a result, for the relative MSE performance, we normalize by a different benchmark. We can think of the performance gain of INDEPENDENT-GAUSS over NAIVE as the value of doing basic, standard empirical Bayes shrinkage. We normalize each method's estimated MSE improvement against NAIVE as a multiple of this "value of basic empirical Bayes." Figure B.4(a) shows the resulting relative performance. Since our notion of relative performance has changed, we use a different color scheme. A value of 1 means that a method does exactly as well as INDEPENDENT-GAUSS, and a value of 2 means that, relative to NAIVE, a method doubles the gain of basic empirical Bayes. Performance on a non-relative scale is shown in Figure B.4(b).

We find that our empirical patterns from the calibrated simulation Figure 4 mostly persists on real data. In particular, INDEPENDENT-NPMLE offers small improvements over INDEPENDENT-GAUSS. Nevertheless, CLOSE-NPMLE continues to dominate other methods. Across the definitions of ϑ_i, CLOSE-NPMLE generates a median of 180% the value of basic empirical Bayes. That is, on mean-squared error, moving from INDEPENDENT-GAUSS to CLOSE-NPMLE is about half as valuable as moving from NAIVE to INDEPENDENT-GAUSS. For our running example (TOP-20 PROBABILITY for Black individuals), moving from INDEPENDENT-GAUSS to CLOSE-NPMLE is more valuable than moving from NAIVE to INDEPENDENT-GAUSS. If practitioners find using the standard empirical Bayes method to be a worthwhile investment over using the raw estimates directly, then they may find using CLOSE-NPMLE over INDEPENDENT-GAUSS to be a similarly worthwhile investment.

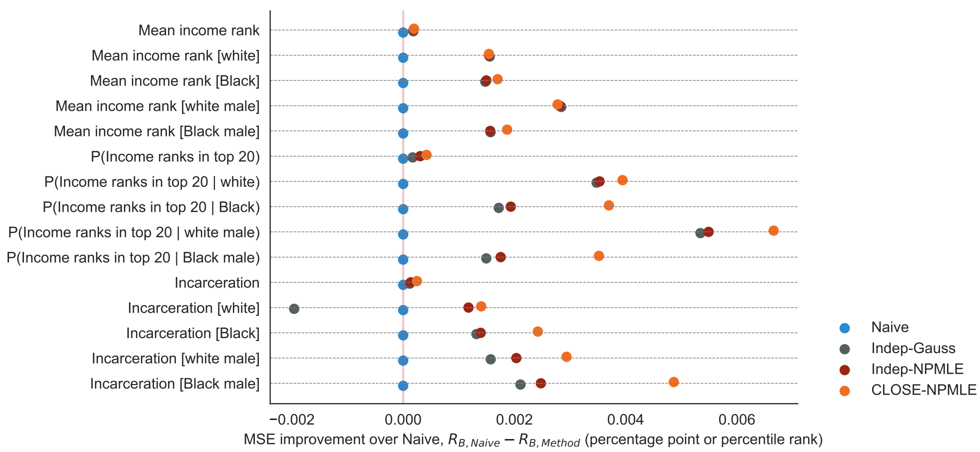
B.6. **Empirical Bayes pooling over all Commuting Zones in validation exercise.** Here, we repeat the exercise in Figure 5, but we now estimate empirical Bayes methods pooling over all

(a) Normalized performance

On MSE, how much do we gain over Naive as a multiple of Indep-Gauss's gain over Naive?

	Naive	Indep-Gauss	Indep-NPMLE	CLOSE-NPMLE
Mean income rank	0.0	1.0	1.0	1.1
Mean income rank [white]	0.0	1.0	1.0	1.0
Mean income rank [Black]	0.0	1.0	1.0	1.2
Mean income rank [white male]	0.0	1.0	1.0	1.0
Mean income rank [Black male]	0.0	1.0	1.0	1.2
P(Income ranks in top 20)	0.0	1.0	1.8	2.5
P(Income ranks in top 20 | white)	0.0	1.0	1.0	1.1
P(Income ranks in top 20 | Black)	0.0	1.0	1.1	2.2
P(Income ranks in top 20 | white male)	0.0	1.0	1.0	1.2
P(Income ranks in top 20 | Black male)	0.0	1.0	1.2	2.4
Incarceration	0.0	1.0	1.1	2.1
Incarceration [white]				
Incarceration [Black]	0.0	1.0	1.1	1.8
Incarceration [white male]	0.0	1.0	1.3	1.9
Incarceration [Black male]	0.0	1.0	1.2	2.3
Column median	0.0	1.0	1.1	1.8

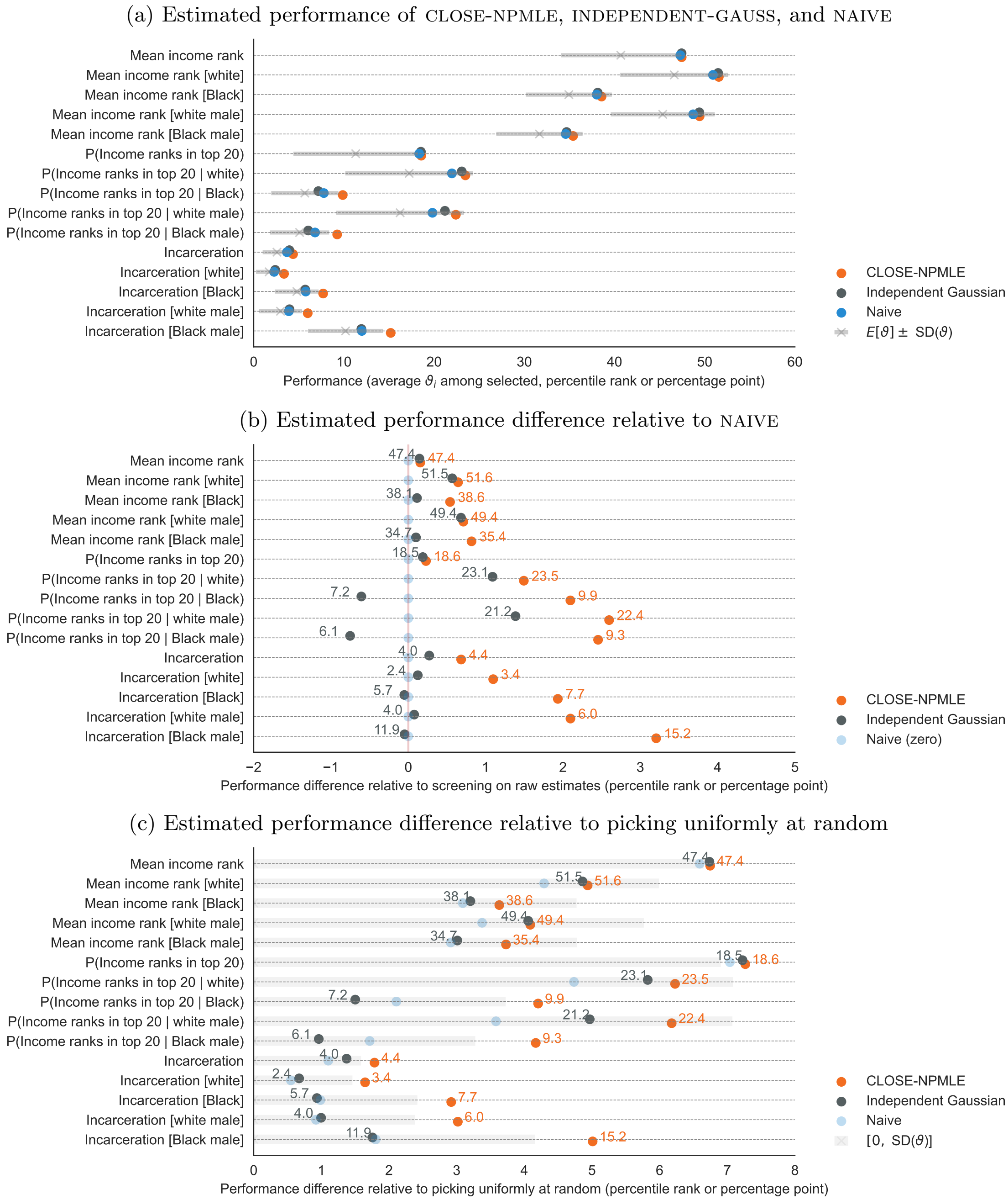
(b) Performance difference against NAIVE



Notes. In panel (a), each column is an empirical Bayes strategy that we consider, and each row is a different definition of θ_i. The table shows relative performance, defined as the squared error improvement over NAIVE, normalized as a multiple of the improvement of INDEPENDENT-GAUSS over NAIVE. By definition, such a measure is zero for NAIVE and one for INDEPENDENT-GAUSS. The last row shows the column median. The mean-squared error estimates average over 100 coupled bootstrap draws. For the variable INCARCERATION for white individuals, the strategy INDEPENDENT-GAUSS underperform NAIVE, and the resulting ratio is thus undefined.
Panel (b) shows the difference in MSE against NAIVE. □

FIGURE B.4. Estimated MSE Bayes risk for various empirical Bayes strategies in the validation exercise.

Commuting Zones. We still pick the top third of every Commuting Zone. Our first exercise repeats Figure 5 in this setting, shown in Figure B.5. The results are extremely similar.



Notes. These figures show the estimated performance of various decision rules over 100 coupled bootstrap draws. Performance is measured as the mean ϑ_i among selected Census tracts. All decision rules select the top third of Census tracts within each Commuting Zone. Figure (a) plots the estimated performance, averaged over 100 coupled bootstrap draws, with the estimated unconditional mean and standard deviation shown as the grey interval. Figure (b) plots the estimated performance *gap* relative to NAIVE, where we annotate with the estimated performance for CLOSE-NPMLE and INDEPENDENT-GAUSS. Figure (c) plots the estimated performance gap relative to picking uniformly at random; we continue to annotate with the estimated performance. The shaded regions in Figure (c) have lengths equal to the unconditional standard deviation of the underlying parameter ϑ. □

FIGURE B.5. Performance of decision rules in top-m selection exercise

Separately, we consider the version of this exercise without covariates in Figure B.6. We see that covariates are extremely important for the performance of INDEPENDENT-GAUSS, as it frequently underperforms NAIVE without covariates.[69] By comparison, they are less important for the performance of CLOSE-NPMLE, as σ_i contains a lot of the signal in the tract-level covariates.

B.7. **The tradeoff between accurate targeting and estimation precision.** In this section, we investigate the tradeoff between accurate targeting and estimation precision. That is, suppose θ_i, Y_i, σ_i and $\vartheta_i, \Upsilon_i, \varsigma_i$ are two sets variables corresponding to two measures of economic mobility. For instance, perhaps θ_i is MEAN RANK for Black individuals and ϑ_i is MEAN RANK pooling over all individuals. Suppose the decision maker would like to select populations with high θ_i, but the estimates Y_i are noisier than the estimates Υ_i. It is plausible that screening on posterior means for ϑ_i might outperform screening on posterior means for θ_i.

We investigate this question via coupled bootstrap in the Bergman et al. (2023) exercise. In particular, we let the subscript b (resp. w) denote quantities for Black (resp. white) individuals. We assume that $Y_{ib} \perp\!\!\!\perp Y_{iw} \mid \theta_{ib}, \theta_{iw}$. For each tract, we construct $\pi_i = n_{ib}/n_i$, where n_i (resp. n_i) is the number of (resp. Black) children under 18 living in the given tract with parents whose household income was below the national median.[70] Let $\theta_i = \pi_i\theta_{ib} + (1-\pi_i)\theta_{iw}$ be a pooled measure, where

$$Y_i = \pi_i Y_{ib} + (1-\pi_i) Y_{iw} \mid \theta_i \sim \mathcal{N}(0, \pi_i^2\sigma_{ib}^2 + (1-\pi_i)^2\sigma_{iw}^2).$$

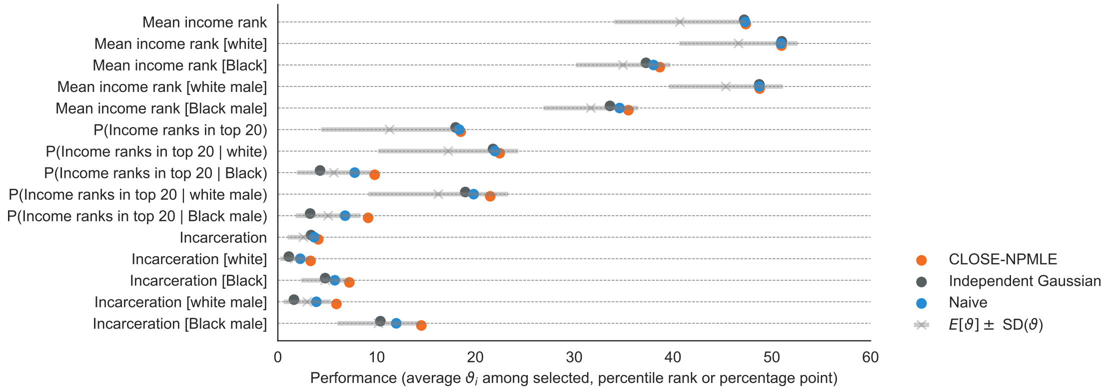
Each coupled bootstrap draw adds and subtracts noise Z_{ib}, Z_{iw} to Y_{ib} and Y_{iw}, where $Z_{ib} \perp\!\!\!\perp Z_{iw}$. Bootstrap draws for Y_i are constructed by taking the π_i-combination of bootstrap draws for Y_{ib}, Y_{iw}.

Here, we investigate whether screening tracts based on posterior mean estimates for θ_{iw} or θ_i generates better decisions in terms of θ_{ib}, owing to the precision in Y_{iw} and Y_i. Figure B.7 shows estimated performances of different empirical Bayes methods by different proxy variables that the screening targets. For each measure of economic mobility for Black individuals, dots on the thick black dashed line correspond to screening on the corresponding θ_{ib}. Dots on the red (resp. blue) dashed line correspond to screening on θ_{iw} (resp. θ_i). We see that for all three measures of economic mobility, using CLOSE-NPMLE to screen on the original parameter θ_{ib} performs best. In other words, the benefits of higher precision are insufficient to offset inaccurate targeting.

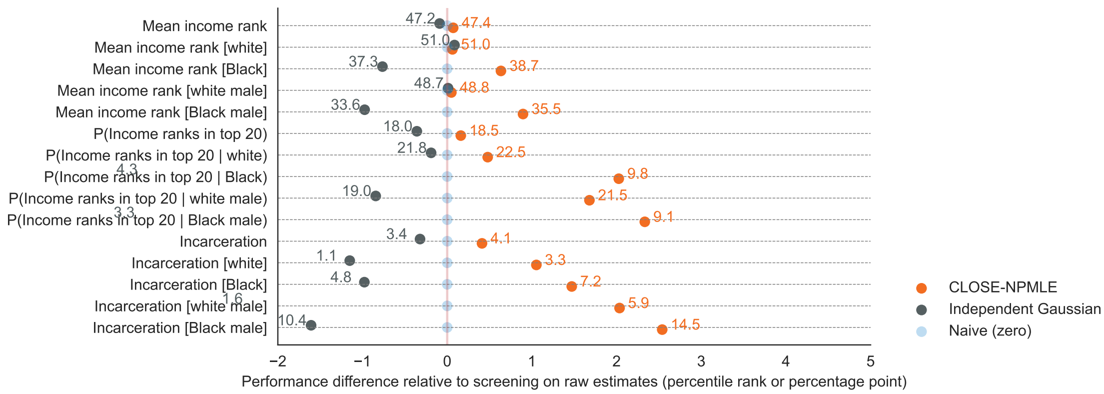
[69]This is in part since our implementation of INDEPENDENT-GAUSS uses weighted means for estimating the prior parameters, worsening the misspecification. See Footnote 54.

[70]This is the demographic weighting variable used in Chetty et al. (2020). We use this weighting to construct a pooled variable, rather than use the pooled variable in the Opportunity Atlas directly for the following reasons. The pooled estimates of Chetty et al. (2020) unfortunately frequently lies outside the convex hull of the white and Black estimates, making it difficult to infer the relative weights for Black individuals in a tract.

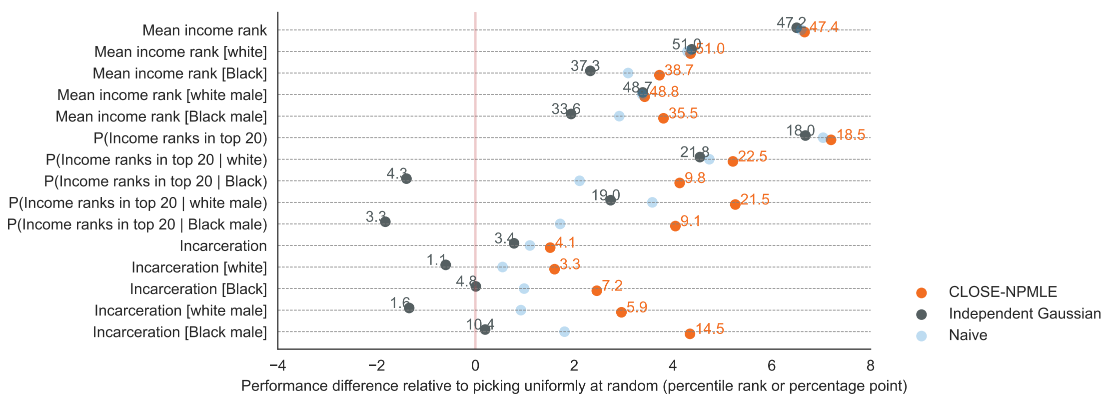
(a) Estimated performance of CLOSE-NPMLE, INDEPENDENT-GAUSS, and NAIVE



(b) Estimated performance difference relative to NAIVE

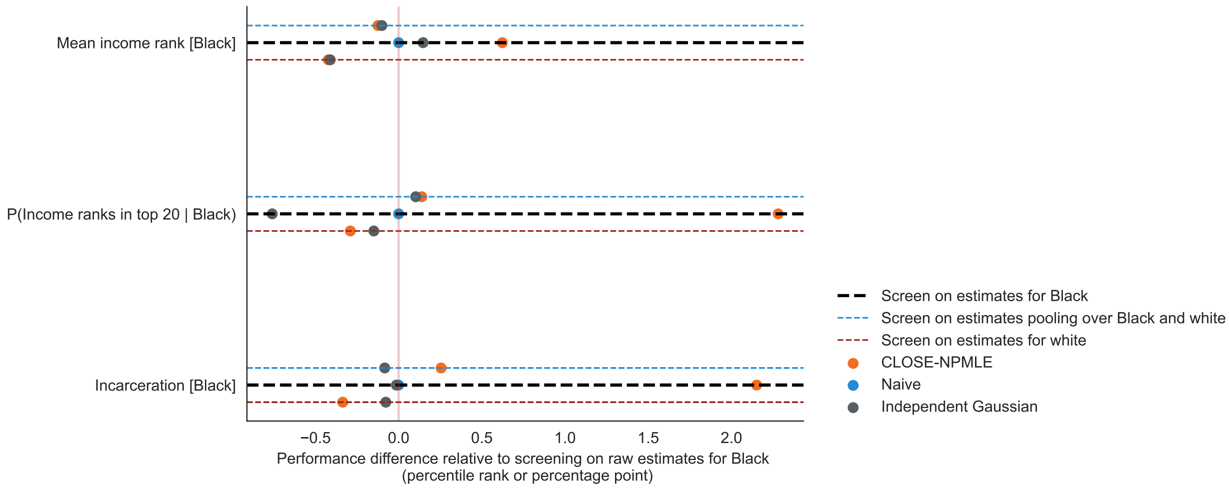


(c) Estimated performance difference relative to picking uniformly at random



Notes. These figures show the estimated performance of various decision rules over 100 coupled bootstrap draws. There are no covariates to residualize against. Performance is measured as the mean ϑ_i among selected Census tracts. All decision rules select the top third of Census tracts within each Commuting Zone. Figure (a) plots the estimated performance, averaged over 100 coupled bootstrap draws, with the estimated unconditional mean and standard deviation shown as the grey interval. Figure (b) plots the estimated performance *gap* relative to NAIVE, where we annotate with the estimated performance for CLOSE-NPMLE and INDEPENDENT-GAUSS. Figure (c) plots the estimated performance gap relative to picking uniformly at random; we continue to annotate with the estimated performance. The shaded regions in Figure (c) have lengths equal to the unconditional standard deviation of the underlying parameter ϑ. □

FIGURE B.6. Performance of decision rules in top-m selection exercise (No covariates)



Notes. Estimated performance for different empirical Bayes methods by different proxy parameters. The performance of screening based on the raw Y_{ib} is normalized to zero. All results are over 100 coupled bootstrap draws. □

FIGURE B.7. Performances of strategies that screen on posterior means for more precisely estimated parameters

Part 3. Regret control proofs

Appendix C. Setup, assumptions, and notation

We recall some notation in the main text, and introduce additional notation. Recall that we assume $n \geq 7$. We observe $(Y_i, \sigma_i)_{i=1}^n, (Y_i, \sigma_i) \in \mathbb{R} \times \mathbb{R}_{>0}$ such that

$$Y_i \mid (\theta_i, \sigma_i) \sim \mathcal{N}(\theta_i, \sigma_i^2)$$

and $(Y_i, \theta_i, \sigma_i)$ are mutually independent. Assume that the joint distribution for (θ_i, σ_i) takes the location-scale form (2.6)

$$\theta_i \mid (\sigma_1, \ldots, \sigma_n) \sim G_0 \left(\frac{\theta_i - m_0(\sigma_i)}{s_0(\sigma_i)} \right)$$

Define shorthands $m_{0i} = m_0(\sigma_i)$ and $s_{0i} = s_0(\sigma_i)$. Define the transformed parameter $\tau_i = \frac{\theta_i - m_{0i}}{s_{0i}}$, the transformed data $Z_i = \frac{Y_i - m_{0i}}{s_{0i}}$, and the transformed variance $\nu_i^2 = \frac{\sigma_i^2}{s_{0i}^2}$. By assumption,

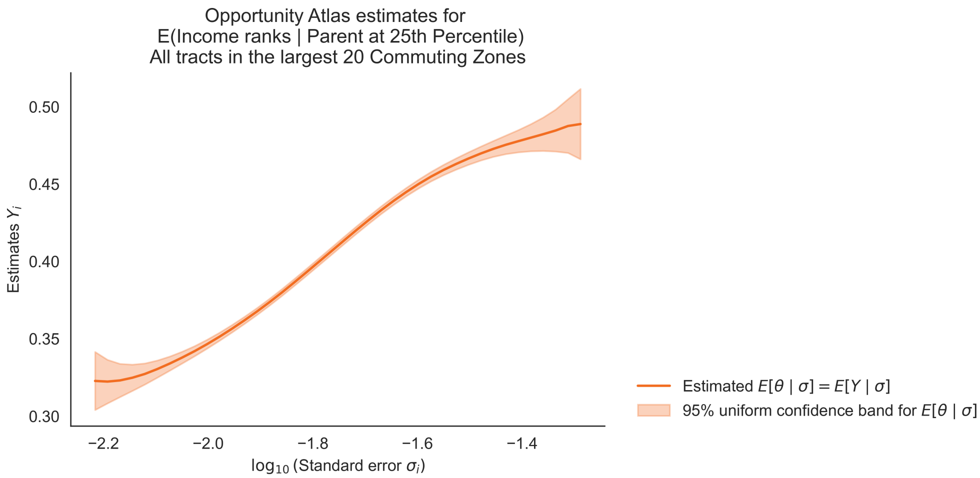
$$Z_i \mid (\tau_i, \nu_i) \sim \mathcal{N}(\tau_i, \nu_i^2) \quad \tau_i \mid \nu_1, \ldots, \nu_n \overset{\text{i.i.d.}}{\sim} G_0.$$

Let $\hat{\eta} = (\hat{m}, \hat{s})$ denote estimates of m_0 and s_0. Likewise, let $\hat{\eta}_i = (\hat{m}_i, \hat{s}_i) = (\hat{m}(\sigma_i), \hat{s}(\sigma_i))$. For a given $\hat{\eta}$, define

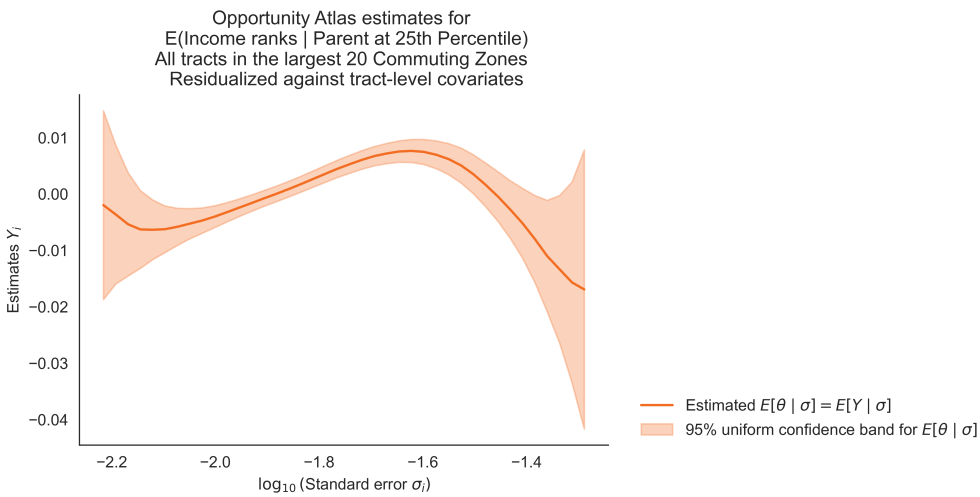
$$\hat{Z}_i = \hat{Z}_i(\hat{\eta}) = \hat{Z}_i(Z_i, \hat{\eta}) = \frac{Y_i - \hat{m}_i}{\hat{s}_i} = \frac{s_{0i} Z_i + m_{0i} - \hat{m}_i}{\hat{s}_i} \quad \hat{\nu}_i^2 = \hat{\nu}_i^2(\hat{\eta}) = \frac{\sigma_i^2}{\hat{s}_i^2}.$$

We will condition on $\sigma_{1:n}$ throughout, and hence we treat them as fixed.

(a) Not residualized by covariates



(b) Residualized by covariates



Notes. This figure shows the estimated $\mathbb{E}[\theta \mid \sigma]$ for mean income rank, pooling over all demographic groups. This is the measure of economic mobility used by Bergman et al. (2023). The estimation and the confidence band procedures are the same as those in Figure 1. In panel (a), θ_i, Y_i are defined as unresidualized measures of mean income rank. In panel (b), we treat θ_i, Y_i as residualized against a vector of tract-level covariates as specified in Appendix B.3. □

FIGURE B.8. Estimated $\mathbb{E}[\theta \mid \sigma]$ for mean income rank among those with parents at the 25$^{\text{th}}$ percentile

For generic G and $\nu > 0$, define

$$f_{G,\nu}(z) = \int_{-\infty}^{\infty} \varphi\left(\frac{z-\tau}{\nu}\right) \frac{1}{\nu}\, G(d\tau).$$

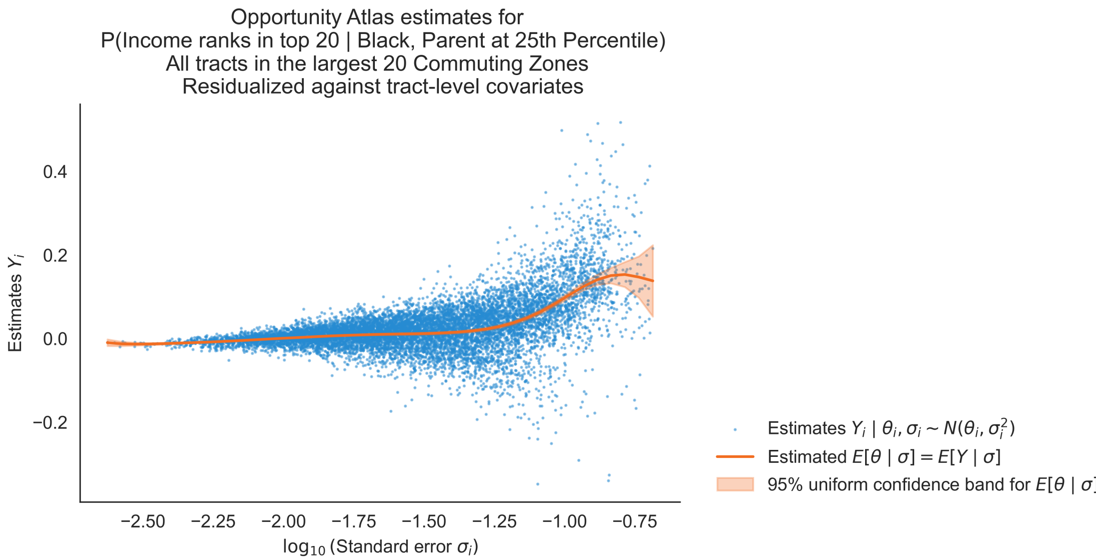


FIGURE B.9. The analogue of Figure 1 where Y_i, θ_i are treated as residualized against a vector of covariates as specified in Appendix B.3.

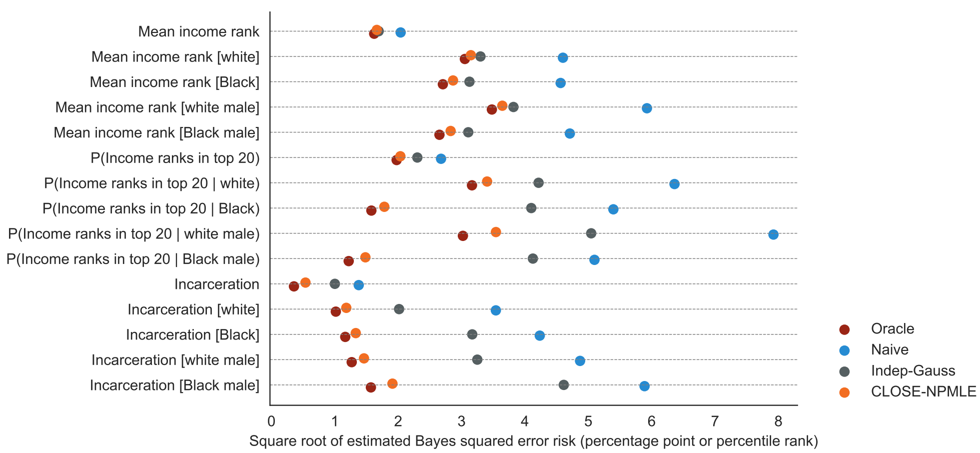


FIGURE B.10. Absolute mean-squared error risk of key methods for the calibrated simulation in Figure 4.

to be the marginal density of some mixed normal deviate $Z \mid \tau \sim \mathcal{N}(\tau, \nu^2)$ with mixing distribution $\tau \sim G$. As a shorthand, we write

$$f_{i,G} = f_{G,\nu_i}(Z_i) \quad f'_{i,G} = f'_{G,\nu_i}(Z_i)$$

Let the average squared Hellinger distance be

$$\bar{h}^2(f_{G_1,\cdot}, f_{G_2,\cdot}) = \frac{1}{n}\sum_{i=1}^n h^2\left(f_{G_1,\nu_i}, f_{G_2,\nu_i}\right).$$

For generic values $\eta = (m, s)$ and distribution G, define the log-likelihood function

$$\psi_i(z, \eta, G) = \psi_i(z, (m, s), G) = \log \int_{-\infty}^{\infty} \varphi\left(\frac{\hat{Z}_i(\eta) - \tau}{\hat{\nu}_i(\eta)}\right) G(d\tau) = \log\left(\hat{\nu}_i(\eta) \cdot f_{G, \hat{\nu}_i(\eta)}(\hat{Z}_i(\eta))\right)$$

Define

$$\mathrm{Sub}_n(G) = \left(\frac{1}{n}\sum_{i=1}^n \psi_i(Z_i, \eta_0, G) - \frac{1}{n}\sum_{i=1}^n \psi_i(Z_i, \eta_0, G_0)\right)_+ \tag{C.1}$$

as the log-likelihood suboptimality of G against the true distribution G_0, evaluated on the true, but unobserved, transformed data Z_i, ν_i.

Fix some generic G and $\eta = (m, s)$. The empirical Bayes posterior mean ignores the fact that G, η are potentially estimated. The posterior mean for $\theta_i = s_i\tau + m_i$ is

$$\hat{\theta}_{i,G,\eta} = m_i + s_i \mathbf{E}_{G, \hat{\nu}_i(\eta)}[\tau \mid \hat{Z}_i(\eta)].$$

Here, we define $\mathbf{E}_{G,\nu}\left[h(\tau, Z) \mid z\right]$ as the function of z that equals the posterior mean for $h(\tau, Z)$ under the data-generating model $\tau \sim G$ and $Z \mid \tau \sim \mathcal{N}(\tau, \nu)$. Explicitly,

$$\mathbf{E}_{G,\nu}\left[h(\tau, Z) \mid z\right] = \frac{1}{f_{G,\nu}(z)} \int h(\tau, z)\varphi\left(\frac{z - \tau}{\nu}\right)\frac{1}{\nu}\, G(d\tau).$$

Explicitly, by Tweedie's formula,

$$\mathbf{E}_{G,\hat{\nu}_i(\eta)}[\tau_i \mid \hat{Z}_i(\eta)] = \hat{Z}_i(\eta) + \hat{\nu}_i^2(\eta)\frac{f'_{G,\hat{\nu}_i(\eta)}(\hat{Z}_i(\eta))}{f_{G,\hat{\nu}_i(\eta)}(\hat{Z}_i(\eta))}.$$

Hence, since $\hat{Z}_i(\eta) = \frac{Y_i - m_i}{s_i}$,

$$\hat{\theta}_{i,G,\eta} = Y_i + s_i\hat{\nu}_i^2(\eta)\frac{f'_{G,\hat{\nu}_i(\eta)}(\hat{Z}_i(\eta))}{f_{G,\hat{\nu}_i(\eta)}(\hat{Z}_i(\eta))}.$$

Define $\theta_i^* = \hat{\theta}_{i,G_0,\eta_0}$ to be the oracle Bayesian's posterior mean. Fix some positive number $\rho > 0$, define a regularized posterior mean as

$$\hat{\theta}_{i,G,\eta,\rho} = Y_i + s_i\hat{\nu}_i^2(\eta)\frac{f'_{G,\hat{\nu}_i(\eta)}(\hat{Z}_i(\eta))}{f_{G,\hat{\nu}_i(\eta)}(\hat{Z}_i(\eta)) \vee \frac{\rho}{\hat{\nu}_i(\eta)}} \tag{C.2}$$

and define $\theta_{i,\rho}^* = \hat{\theta}_{i,G_0,\eta_0,\rho}$ correspondingly.

Lastly, we will also define

$$\varphi_+(\rho) = \varphi^{-1}(\rho) = \sqrt{\log\frac{1}{2\pi\rho^2}} \quad \rho \in (0, (2\pi)^{-1/2}) \tag{C.3}$$

so that $\varphi(\varphi_+(\rho)) = \rho$. Observe that $\varphi_+(\rho) \lesssim \sqrt{\log(1/\rho)}$.

C.1. **Assumptions.** Recall the assumptions we stated in the main text.

Assumption 1. *Let* $\psi_i(Z_i, \hat{\eta}, G) \equiv \log\left(\int_{-\infty}^{\infty} \varphi\left(\frac{\hat{Z}_i - \tau}{\hat{\nu}_i}\right) G(d\tau)\right)$ *be the objective function in* (2.11), *ignoring a constant factor* $1/\hat{\nu}_i$. *We assume that* $\hat{G}_n$ *satisfies*

$$\frac{1}{n}\sum_{i=1}^{n} \psi_i(Z_i, \hat{\eta}, \hat{G}_n) \geq \sup_{H \in \mathcal{P}(\mathbb{R})} \frac{1}{n}\sum_{i=1}^{n} \psi_i(Z_i, \hat{\eta}, H) - \kappa_n \tag{3.2}$$

for tolerance κ_n

$$\kappa_n = \frac{2}{n} \log\left(\frac{n}{\sqrt{2\pi e}}\right). \tag{3.3}$$

Moreover, we require that $\hat{G}_n$ *has support points within* $[\min_i \hat{Z}_i, \max_i \hat{Z}_i]$. *To ensure that* κ_n *is positive, we assume that* $n \geq 7 = \lceil \sqrt{2\pi e} \rceil$.[71]

Assumption 2. *The distribution* G_0 *is has zero mean, unit variance, and admits simultaneous moment control with parameter* $\alpha \in (0, 2]$: *There exists a constant* $A_0 > 0$ *such that for all* $p > 0$,

$$(\mathbb{E}_{\tau \sim G_0}[|\tau|^p])^{1/p} \leq A_0 p^{1/\alpha}. \tag{3.4}$$

Assumption 3. *The variances* $(\sigma_{1:n}, s_0)$ *admit lower and upper bounds:*

$$\sigma_\ell < \sigma_i < \sigma_u \text{ and } s_\ell < s_0(\cdot) < s_u,$$

where $0 < \sigma_\ell, \sigma_u, s_{0\ell}, s_{0u} < \infty$. *This implies that* $0 < \nu_\ell \leq \nu_i = \frac{\sigma_i}{s_0(\sigma_i)} \leq \nu_u < \infty$ *for some* ν_ℓ, ν_u.

Assumption 4. *Let* $C^p_{A_1}([\sigma_\ell, \sigma_u])$ *be the Hölder class of order* $p \geq 1$ *with maximal Hölder norm* $A_1 > 0$ *supported on* $[\sigma_\ell, \sigma_u]$.[72] *We assume that*

(1) *The true conditional moments are Hölder-smooth:* $m_0, s_0 \in C^p_{A_1}([\sigma_\ell, \sigma_u])$.

Additionally, let $\beta_0 > 0$ *be a constant. Let* $\mathcal{V}$ *be a set of bounded functions supported on* $[\sigma_\ell, \sigma_u]$ *that (i) admits the uniform bound* $\sup_{f \in \mathcal{V}} \|f\|_\infty \leq C_{A_1}$ *and (ii) admits the metric entropy bound*

$$\log N(\epsilon, \mathcal{V}, \|\cdot\|_\infty) \leq C_{A_1, p, \sigma_\ell, \sigma_u} (1/\epsilon)^{1/p}.$$

We assume that the estimators for m_0 *and* s_0, $\hat{\eta} = (\hat{m}, \hat{s})$, *satisfy the following assumptions.*

(2) *For any* $\epsilon > 0$, *there exists a sufficiently large* $C = C(\epsilon)$, *independently of* n, *such that for all* n,

$$\mathrm{P}\left(\max\left(\|\hat{m} - m_0\|_\infty, \|\hat{s} - s_0\|_\infty\right) > C(\epsilon) n^{-\frac{p}{2p+1}} (\log n)^{\beta_0}\right) < \epsilon.$$

[71]The constants κ_n also feature in Jiang (2020) to ensure that the fitted likelihood is bounded away from zero. The particular constants in κ_n are chosen to simplify expressions and are not material to the result.

[72]We recall the definition of a Hölder class from van der Vaart and Wellner (1996), Section 2.7.1. We specialize its definition to functions of one real variable. For an integer p, Hölder-p functions are $(p-1)$-times differentiable, with a Lipschitz continuous $(p-1)^{\text{st}}$ derivative.

Definition 2. For some set $\mathcal{X} \subset \mathbb{R}$ and constant $A > 0$, $p > 0$, let $C^p_A(\mathcal{X})$ be the set of continuous functions $f : \mathcal{X} \to \mathbb{R}$ with $\|f\|_{(p)} \leq A$. The norm $\|\cdot\|_{(p)}$ is defined as follows. Let $\underline{p}$ be the greatest integer strictly smaller than p. Define

$$\|f\|_{(p)} = \max_{k \leq \underline{p}} \sup_{x \in \mathcal{X}} \left|f^{(k)}(x)\right| + \sup_{x, y \in \mathcal{X}} \frac{\left|f^{(\underline{p})}(x) - f^{(\underline{p})}(y)\right|}{|x - y|^{p - \underline{p}}}.$$

We refer to $C^p_A(\mathcal{X})$ as a Hölder class of order p and $\|f\|_{(p)}$ as the Hölder norm.

(3) The nuisance estimators take values in $\mathcal{V}$ almost surely: $\mathrm{P}(\hat{m} \in \mathcal{V}, \hat{s} \in \mathcal{V}) = 1$.

(4) The conditional variance estimator respects the conditional variance bounds in Assumption 3: $\mathrm{P}\left(\frac{s_{0\ell}}{2} < \hat{s} < 2s_{0u}\right) = 1$.

C.2. **Regret control: result statement.** Define the regret as the difference between the mean-squared error of some feasible posterior means $\hat{\theta}_{i,G,\eta}$ against the mean-squared error of the oracle posterior means

$$\begin{aligned}\mathrm{MSERegret}_n(G,\eta) &= \frac{1}{n}\sum_{i=1}^n(\hat{\theta}_{i,G,\eta}-\theta_i)^2 - \frac{1}{n}\sum_{i=1}^n(\theta_i^*-\theta_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^n(\hat{\theta}_{i,G,\eta}-\theta_i^*)^2 + \frac{2}{n}\sum_{i=1}^n(\theta_i^*-\theta_i)(\hat{\theta}_{i,G,\eta}-\theta_i^*) \end{aligned} \tag{C.4}$$

(C.4) decomposes the MSE regret into a mean term that equals the mean-squared distance between the feasible posterior means and the oracle posterior means, as well as a term that is mean zero conditional on the data $Y_1, \ldots, Y_n$, since $\theta_i^* - \theta_i$ represents irreducible noise.

Fix sequences $\Delta_n > 0$ and $M_n > 0$. Define the following "good" event which we use in Theorem F.1:

$$A_n = \left\{ \|\hat{\eta}-\eta\|_\infty \equiv \max(\|\hat{m}-m_0\|_\infty, \|\hat{s}-s_0\|_\infty) \le \Delta_n,\, \overline{Z}_n \equiv \max_{i\in[n]}(|Z_i| \vee 1) \le M_n \right\}. \tag{C.5}$$

On the event A_n, the nuisance estimates $\hat{\eta}$ are good, and the data Z_i are not too large. Note that, with $\Delta_n = C_1 n^{-\frac{p}{2p+1}}(\log n)^{\beta_0}$,

$$A_n = \mathbf{A}_n(C_1) \cap \left\{\overline{Z}_n \le M_n\right\},$$

where $\mathbf{A}_n$ is the event in (3.5).

Here, we prove the version of our result stated in the main text.

Theorem 1. *Assume Assumptions 1 to 4 hold. Then, for any $\delta \in (0, \frac{1}{2})$, there exists universal constants $C_{1,\mathcal{H},\delta} > 0$ and $C_{0,\mathcal{H},\delta} > 0$ such that (i) $\mathrm{P}(\mathbf{A}_n(C_{1,\mathcal{H},\delta})) \ge 1-\delta$ and that (ii) the expected regret conditional on $\mathbf{A}_n(C_{1,\mathcal{H},\delta})$ is dominated by the rate function*

$$\mathbb{E}\left[\mathrm{MSERegret}_n(\hat{G}_n, \hat{\eta}) \mid \mathbf{A}_n(C_{1,\mathcal{H},\delta})\right] \le C_{0,\mathcal{H},\delta} n^{-\frac{2p}{2p+1}}(\log n)^{\frac{2+\alpha}{\alpha}+3+2\beta_0}. \tag{3.6}$$

Proof. Immediately by Assumption 4(2–3), we can choose $C_{1,\mathcal{H}}$ so that $\mathrm{P}(\mathbf{A}_n(C_{1,\mathcal{H}})) \ge 1-\delta$. Let $\Delta_n = C_{1,\mathcal{H}} n^{-\frac{p}{2p+1}}(\log n)^{\beta_0}$ and $M_n = C(\log n)^{1/\alpha}$ for some C to be chosen. Both $C_{1,\mathcal{H}}$ and C may depend on δ. Moreover, we can decompose

$$\begin{aligned}&\mathbb{E}\left[\mathrm{MSERegret}_n(\hat{G}_n, \hat{\eta}) \mid \mathbf{A}_n(C_{1,\mathcal{H}})\right] \\ &\le \frac{1}{1-\delta}\left\{\mathbb{E}\left[\mathrm{MSERegret}_n(\hat{G}_n,\hat{\eta})\mathbb{1}(A_n)\right] + \mathbb{E}\left[\mathrm{MSERegret}_n(\hat{G}_n,\hat{\eta})\underbrace{\mathbb{1}(\mathbf{A}_n(C_{1,\mathcal{H}}), \overline{Z}_n > M_n)}_{\mathbf{A}_n \setminus A_n}\right]\right\} \\ &\lesssim_{\mathcal{H}} n^{-\frac{2p}{2p+1}}(\log n)^{\frac{2+\alpha}{\alpha}+3+2\beta_0} + \frac{1}{n}(\log n)^{2/\alpha} \qquad \text{(Theorem F.1 and Lemma F.1)} \\ &\lesssim_{\mathcal{H}} n^{-\frac{2p}{2p+1}}(\log n)^{\frac{2+\alpha}{\alpha}+3+2\beta_0}\end{aligned}$$

Note that the application of Theorem F.1 and Lemma F.1 implicitly picks some constant for $M_n = C(\log n)^{1/\alpha}$. This concludes the proof. $\square$

Corollary 1. *Assume the same setting as Theorem 1. Suppose, additionally, for all sufficiently large $C_{1,\mathcal{H}} > 0$, $\mathrm{P}(\mathbf{A}_n(C_{1,\mathcal{H}})) \geq 1 - n^{-2}$. Then, there exists a constant $C_{0,\mathcal{H}} > 0$ such that the expected regret is dominated by the rate function*

$$\text{BayesRegret}_n = \mathbb{E}\left[\text{MSERegret}_n(\hat{G}_n, \hat{\eta})\right] \leq C_{0,\mathcal{H}} n^{-\frac{2p}{2p+1}} (\log n)^{\frac{2+\alpha}{\alpha}+3+2\beta_0}.$$

Proof. Let Δ_n, M_n as in the proof of Theorem 1. Decompose

$$\begin{aligned}
\mathbb{E}[\text{MSERegret}_n(\hat{G}_n, \hat{\eta})] &= \mathbb{E}[\text{MSERegret}_n(\hat{G}_n, \hat{\eta})\mathbb{1}(A_n)] + \mathbb{E}[\text{MSERegret}_n(\hat{G}_n, \hat{\eta})\mathbb{1}(A_n^{\mathrm{C}})] \\
&= \mathbb{E}[\text{MSERegret}_n(\hat{G}_n, \hat{\eta})\mathbb{1}(A_n)] + \mathbb{E}[\text{MSERegret}_n(\hat{G}_n, \hat{\eta})\mathbb{1}(\mathbf{A}_n^{\mathrm{C}} \cup \{Z_n > M_n\})] \\
&\leq \mathbb{E}[\text{MSERegret}_n(\hat{G}_n, \hat{\eta})\mathbb{1}(A_n)] + \mathbb{E}[\text{MSERegret}_n(\hat{G}_n, \hat{\eta})\mathbb{1}(\mathbf{A}_n^{\mathrm{C}})] \\
&\quad + \mathbb{E}[\text{MSERegret}_n(\hat{G}_n, \hat{\eta})\mathbb{1}(\overline{Z}_n > M_n)] \\
&\lesssim_{\mathcal{H}} n^{-\frac{2p}{2p+1}} (\log n)^{\frac{2+\alpha}{\alpha}+3+2\beta_0} + \frac{2}{n}(\log n)^{2/\alpha} \\
&\qquad \text{(Theorem F.1 and Lemma F.1)} \\
&\lesssim_{\mathcal{H}} n^{-\frac{2p}{2p+1}} (\log n)^{\frac{2+\alpha}{\alpha}+3+2\beta_0},
\end{aligned}$$

where our application of Lemma F.1 uses the assumption that $\mathrm{P}(\mathbf{A}_n(C_{1,\mathcal{H}})^{\mathrm{C}}) = \mathbb{1}(\|\hat{\eta} - \eta\|_\infty > \Delta_n) \leq \frac{1}{n^2}$. $\square$

Remark C.1 (Relaxing Assumption 4(4))**.** Note that the event $\mathbf{A}_n(C)$ implies $s_{0\ell}/2 \leq \hat{s} \leq 2s_{0u}$ for all sufficiently large $n > N_{C,s_{0\ell},s_{0u},p,\beta_0}$. Since we condition on $\mathbf{A}_n(C)$ in Theorem 1, we can drop Assumption 4(3) by only requiring (3.6) to hold for all sufficiently large n. This is a minor modification since Theorem 1 is an upper bound on the convergence rate. On the other hand, dropping Assumption 4(4) does affect regret control on the event $\mathbf{A}_n^{\mathrm{C}}(C_1)$ below. Our truncation rule for $\hat{s}(\cdot)$ in Appendix G ensures that $\hat{s}(\cdot) \geq \frac{c}{n}$. We show in Appendix G that this is sufficient for the conclusion of Corollary 1.[73] ■

C.3. **Regret control: proof ideas.** We now discuss the main ideas and the structure of our argument. Existing work (Soloff et al., 2021) controls the following quantity, in our notation,

$$\mathbb{E}\left[\text{MSERegret}_n^{\tau}(\hat{G}_n, \eta_0)\right] \equiv \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{\tau}_{i,\hat{G}_n^*,\eta_0} - \tau_i^*)^2\right] \tag{C.6}$$

where $\hat{\tau}_{i,\hat{G}_n,\eta_0} = \mathbf{E}_{\hat{G}_n,\nu_i}[\tau \mid Z_i]$ and $\hat{G}_n^*$ is an approximate NPMLE on the data $(Z_i, \nu_i)_{i=1}^n$ (Theorem 8 in Soloff et al. (2021)).

They do so by showing that, loosely speaking,

[73] This lower bound on $\hat{s}$ also adds enough regularity to avoid writing "sufficiently large n" for the statement analogous to Theorem 1 as well. See Appendix G for details.

(i) For some constant C and rate function δ_n, with high probability, the NPMLE achieves low average squared Hellinger distance:

$$\mathrm{P}\left(\overline{h}^2(f_{\hat{G}_n^*,\cdot}, f_{G_0,\cdot}) > C\delta_n^2\right) < \frac{1}{n}.$$

This is because distributions G that achieve high likelihood—which G_n^* does by construction—tend to have low average squared Hellinger distance with respect to G_0 (Theorem 6 in Soloff et al. (2021)). Roughly speaking, the rate function is linked to likelihood suboptimality (C.1):

$$\delta_n^2 \asymp \max\left(\mathrm{Sub}_n(\hat{G}_n^*), \frac{1}{n}(\log n)^{\frac{2\alpha}{2+\alpha}+1}\right). \tag{C.7}$$

(ii) For a *fixed* distribution G, the deviation from oracle between the *regularized* posterior means (C.2) is bounded by the average squared Hellinger distance:

$$\mathbb{E}[(\hat{\tau}_{i,G,\eta_0,\rho_n} - \tau_{i,\rho_n}^*)^2] \lesssim (\log(1/\rho_n))^3\overline{h}^2(f_{G,\cdot}, f_{G_0,\cdot}). \tag{C.8}$$

Therefore, we should expect that the rate attained is $\log(1/\rho_n)^3\delta_n^2$, subjected to resolving the following two issues.

(iii) Additional arguments can handle the difference between (C.8) and (C.6).

(iv) Additional empirical process arguments can handle the fact that $\hat{G}_n^*$ is estimated.

Our proof adapts this argument, where the key challenge is that we only observe $(\hat{Z}_i, \hat{\nu}_i)$ instead of (Z_i, ν_i). As an outline,

- Appendix D (Theorem D.1 and Corollary D.1) establishes that $\hat{G}_n$, estimated off $(\hat{Z}_i, \hat{\nu}_i)$, achieves high likelihood (i.e., low $\mathrm{Sub}_n(\hat{G}_n)$) on the data (Z_i, ν_i), with high probability. This is an oracle inequality in the sense that it bounds the performance degredation of $\hat{G}_n$ relative to a setting where η_0 is known.
- Appendix E (Theorem E.1 and Corollary E.2) establishes that $\hat{G}_n$, with high probability, achieves low Hellinger distance. This is a result of independent interest, as it characterizes the quality of $f_{\hat{G}_n,\nu_i}$ as an estimate of the true density f_{G_0,ν_i}.
- Appendix F (Theorem F.1) establishes that the regret of $\hat{\theta}_{i,\hat{G}_n,\hat{\eta}}$ is low, using the argument controlling (C.6).

C.3.1. *Intuition for Appendix D.* The argument in Appendix D is our most novel theoretical contribution. Note that, by (C.7), to obtain a rate of the form $\delta_n^2 = n^{-\frac{2p}{2p+1}}(\log n)^\gamma$,[74] we would require that $\mathrm{Sub}_n(\hat{G}_n) \lesssim n^{-\frac{2p}{2p+1}}(\log n)^\gamma$. However, such a rate is not immediately attainable. To see this, note that a direct Taylor expansion in η of the log-likelihood yields

$$\begin{aligned}
&\frac{1}{n}\sum_i \psi_i(Z_i, \hat{\eta}, \hat{G}_n) - \frac{1}{n}\sum_i \psi_i(Z_i, \eta_0, \hat{G}_n) \\
&\approx \frac{1}{n}\sum_i \left(\frac{\partial \psi_i}{\partial \eta_i}\right)'(\eta_i - \eta_{0i}) + \frac{1}{2n}\sum_i (\eta_i - \eta_{0i})'\frac{\partial^2\psi_i}{\partial \eta_i^2}(\eta_i - \eta_{0i}).
\end{aligned} \tag{C.9}$$

[74]We let $(\log n)^\gamma$ denote a generic logarithmic factor, and we will not keep track of γ throughout this heuristic discussion.

$$\lessapprox (\log n)^{\gamma}\left\{\frac{1}{n}\sum_i \frac{\partial \psi_i}{\partial \eta_i} O\left(n^{-\frac{p}{2p+1}}\right) + n^{-\frac{2p}{2p+1}} \sum_i \left\| \frac{\partial^2 \psi_i}{\partial \eta_i^2} \right\| \right\}$$

Thus, without somehow showing that the first-order term $\frac{\partial \psi_i}{\partial \eta_i}$ converges to zero, we would only be able to obtain $\mathrm{Sub}_n(\hat{G}_n) \lesssim n^{-\frac{p}{2p+1}}(\log n)^{\gamma}$, which is insufficient.

Fortunately, it is easy to compute that the expected first derivative, *evaluated at* G_0, is zero:

$$\mathbb{E}\left[\frac{\partial \psi_i(Z, G_0, \eta_0)}{\partial \eta}\right] = 0.$$

As a result, we expect that if $\hat{G}_n$ is close to G_0, then the corresponding first-order terms for $\hat{G}_n$ will also be small. More precisely, it is possible to bound the first-order term in terms of the average squared Hellinger distance, yielding

$$\left| \frac{1}{n}\sum_i \left(\frac{\partial \psi_i}{\partial \eta_i}\right)' (\eta_i - \eta_{0i}) \right| \lesssim n^{-\frac{p}{p+1}} (\log n)^{\gamma} \overline{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}).$$

To summarize, through our calculation, the rate we obtain (Corollary D.1, (D.3)) for $\mathrm{Sub}_n(\hat{G}_n)$ is

$$\varepsilon_n = (\log n)^{\gamma}\left\{ n^{-\frac{p}{2p+1}} \overline{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}) + n^{-\frac{2p}{2p+1}} \right\}.$$

A more detailed breakdown is presented in Appendix D.2.4.

C.3.2. *Intuition for Appendix E.* Since the rate for $\mathrm{Sub}_n(\hat{G}_n)$ from Appendix D itself includes $\overline{h}$, it is necessary to adapt the argument in the literature on Hellinger rate control (See, e.g., Theorem 4 in Jiang, 2020).

Our argument proceeds by observing that, with high probability,

$$\mathrm{Sub}_n(\hat{G}_n) \lesssim \gamma_n^2 + \overline{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot})\lambda_n.$$

for some rates γ_n, λ_n. Then, we separately bound, for $k = 1, \ldots, K$,

$$\begin{aligned} &\mathrm{P}\left[C\lambda_n^{1-2^{-k}} \le \overline{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}) \le C\lambda_n^{1-2^{-k+1}},\ \mathrm{Sub}_n(\hat{G}_n) \lesssim \gamma_n^2 + \overline{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot})\lambda_n\right] \\ &\le \mathrm{P}\left[C\lambda_n^{1-2^{-k}} \le \overline{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}),\ \mathrm{Sub}_n(\hat{G}_n) \lesssim \gamma_n^2 + \lambda_n^{1-2^{-k+1}}\lambda_n\right] \end{aligned} \tag{C.10}$$

using standard arguments in the literature. This is now feasible since the event (C.10) comes with an upper bound for $\overline{h}$. Thus, by a union bound,

$$\mathrm{P}\left(\overline{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}) > C\lambda_n \cdot \lambda_n^{-2^{-K}}\right) \lessapprox \frac{K}{n}.$$

We can choose $K \to \infty$ appropriately slowly so as to obtain $\overline{h}^2 \lesssim \delta_n^2$ with high probability.

C.3.3. *Intuition for Appendix F.* All that is remaining before we can use the bound (C.6) directly is dealing with the difference between $\hat{\theta}_{i,\hat{G}_n,\hat{\eta}}$ and $\tau_{i,\hat{G}_n,\eta_0}$. In Appendix F.3, we can use a Taylor expansion to control the distance

$$\left|\hat{\theta}_{i,\hat{G}_n,\hat{\eta}} - \hat{\theta}_{i,\hat{G}_n,\eta_0}\right| = \sigma_i^2 \left| \frac{f'_{\hat{G}_n,\hat{\nu}_i}(\hat{Z}_i)}{\hat{s}_i f_{\hat{G}_n,\hat{\nu}_i}(\hat{Z}_i)} - \frac{f'_{\hat{G}_n,\nu_i}(Z_i)}{s_{0i} f_{\hat{G}_n,\nu_i}(Z_i)} \right| = \sigma_i \left| \frac{\partial \psi_i}{\partial m}\bigg|_{\hat{G}_n,\hat{\eta}} - \frac{\partial \psi_i}{\partial m}\bigg|_{\hat{G}_n,\eta_0} \right|.$$

Doing so requires bounding the second derivatives of ψ_i, which are posterior moments under $\hat{G}_n$ (Appendix D.10), and hence bounded due to assuming that $\hat{G}_n$ has supported bounded within the range of the data $\hat{Z}_i$ (Lemma D.14). We then immediately find that

$$\left|\hat{\theta}_{i,\hat{G}_n,\eta_0} - \theta^*_{i,G_0,\eta_0}\right|$$

is proportional to the difference in τ-space. Therefore, the existing argument for (C.6) controls the regret.

Appendix D. An oracle inequality for the likelihood

Recall that for some fixed Δ_n, M_n, we define $A_n = \left\{\|\hat{\eta} - \eta\|_\infty \le \Delta_n, \overline{Z}_n \le M_n\right\}$. In this section, we bound

$$\mathrm{P}\left[A_n, \mathrm{Sub}_n(\hat{G}_n) \gtrsim_{\mathcal{H}} \epsilon_n\right]$$

for some rate function ϵ_n. It is convenient to state a set of high-level assumptions on the rates Δ_n, M_n. These are satisfied for $\Delta_n \asymp n^{-p/(2p+1)}(\log n)^\beta, M_n \asymp (\log n)^{1/\alpha}$.

Assumption D.1. *Assume that*

(1) $\frac{1}{\sqrt{n}} \lesssim_{\mathcal{H}} \Delta_n \lesssim_{\mathcal{H}} \frac{1}{M_n^3} \lesssim_{\mathcal{H}} 1$
(2) $\sqrt{\log n} \lesssim_{\mathcal{H}} M_n$

Note that there exists ρ_n by Lemma D.9 that lower bounds the density $f_{\hat{G}_n,\nu_i}(z)$ for all Z_i. Then our main result is an oracle inequality.

Theorem D.1. *Let* $\|\hat{\eta} - \eta\|_\infty = \max(\|\hat{m} - m_0\|_\infty, \|\hat{s} - s_0\|_\infty)$ *and* $\overline{Z}_n = \max_{i\in[n]} |Z_i| \vee 1$. *Suppose* $\hat{G}_n$ *satisfies Assumption 1. Under Assumptions 2 to 4 and D.1, there exists constants* $C_{1,\mathcal{H}}, C_{2,\mathcal{H}} > 0$ *such that the following tail bound holds: Let*

$$\epsilon_n = M_n\sqrt{\log n}\Delta_n \frac{1}{n}\sum_{i=1}^n h\left(f_{\hat{G}_n,\nu_i}, f_{G_0,\nu_i}\right) + \Delta_n M_n \sqrt{\log n} e^{-C_{2,\mathcal{H}}M_n^\alpha} + \Delta_n^2 M_n^2 \log n + M_n^2 \frac{\Delta_n^{1-\frac{1}{2p}}}{\sqrt{n}}. \tag{D.1}$$

Then,

$$\mathrm{P}\left[\overline{Z}_n \le M_n, \|\hat{\eta} - \eta\|_\infty \le \Delta_n, \mathrm{Sub}_n(\hat{G}_n) > C_{1,\mathcal{H}}\epsilon_n\right] \le \frac{9}{n}.$$

The following corollary plugs in some concrete rates for Δ_n, M_n and verifies that they satisfy Assumption D.1.

Corollary D.1. *For* $\beta \ge 0$, *suppose*

$$\Delta_n = C_{\mathcal{H}} n^{-\frac{p}{2p+1}}(\log n)^\beta \text{ and } M_n = (C_{\mathcal{H}} + 1)(C_{2,\mathcal{H}}^{-1} \log n)^{1/\alpha}. \tag{D.2}$$

Then there exists a $C^*_{\mathcal{H}}$ *such that the following tail bound holds. Suppose* $\hat{G}_n$ *satisfies Assumption 1. Under Assumptions 2 to 4, define* ε_n *as:*

$$\varepsilon_n = n^{-\frac{p}{2p+1}}(\log n)^{\frac{2+\alpha}{2\alpha}+\beta}\overline{h}\left(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}\right) + n^{-\frac{2p}{2p+1}}(\log n)^{\frac{2+\alpha}{\alpha}+2\beta}, \tag{D.3}$$

we have that,

$$\mathrm{P}\left[\overline{Z}_n \le M_n, \|\hat{\eta} - \eta\|_\infty \le \Delta_n, \mathrm{Sub}_n(\hat{G}_n) > C^*_{\mathcal{H}}\varepsilon_n\right] \le \frac{9}{n}.$$

The constant $C_{\mathcal{H}}$ in Δ_n, M_n affects the conclusion of the statement only through affecting the constant $C^*_{\mathcal{H}}$.

D.1. **Proof of Corollary D.1.** We first show that the specification of Δ_n and M_n means that the requirements of Assumption D.1 are satisfied. Among the requirements of Assumption D.1:

(1) is satisfied since the polynomial part of Δ_n converges to zero slower than $n^{-1/2}$, but converges to zero faster than any logarithmic rate. M_n is a logarithmic rate.

(2) is satisfied since $\alpha \le 2$.

We also observe that by Jensen's inequality,

$$\frac{1}{n}\sum_i h(f_{\hat{G}_n,\nu_i}, f_{G_0,\nu_i}) \le \overline{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}),$$

and so we can replace the corresponding factor in ϵ_n by $\overline{h}$. Now, we plug the rates Δ_n, M_n into ϵ_n. We find that the term

$$\Delta_n M_n^2 e^{-C_{2,\mathcal{H}} M_n^\alpha} = \Delta_n M_n^2 e^{-(C_{\mathcal{H}}+1)^\alpha (\log n)} \le \Delta_n M_n^2 n^{-1} \le \frac{1}{n}\Delta_n M_n^2 \lesssim_{\mathcal{H}} \Delta_n^2 M_n^2 \log n$$

since $\log n > 1$ as $n > \sqrt{2\pi} e$ by Assumption 1. Plugging in the rates for the other terms, we find that

$$\epsilon_n \lesssim_{\mathcal{H}} \varepsilon_n.$$

Therefore, Corollary D.1 follows from Theorem D.1.

D.2. **Proof of Theorem D.1.**

D.2.1. *Decomposition of* $\mathrm{Sub}_n(\hat{G}_n)$. Observe that, by definition of $\hat{G}_n$ in (3.2),

$$\frac{1}{n}\sum_{i=1}^n \psi_i(Z_i, \hat{\eta}, \hat{G}_n) - \frac{1}{n}\sum_{i=1}^n \psi_i(Z_i, \hat{\eta}, G_0) \ge \kappa_n$$

For random variables a_n, b_n such that almost surely

$$\left|\frac{1}{n}\sum_{i=1}^n \psi_i(Z_i, \hat{\eta}, \hat{G}_n) - \psi_i(Z_i, \eta_0, \hat{G}_n)\right| \le a_n$$

$$\left|\frac{1}{n}\sum_{i=1}^n \psi_i(Z_i, \hat{\eta}, G_0) - \psi_i(Z_i, \eta_0, G_0)\right| \le b_n$$

we have

$$\frac{1}{n}\sum_{i=1}^n \psi_i(Z_i, \eta_0, \hat{G}_n) - \frac{1}{n}\sum_{i=1}^n \psi_i(Z_i, \eta_0, G_0) \ge -a_n - b_n - \kappa_n$$

and

$$\mathrm{Sub}_n(\hat{G}_n) \le a_n + b_n + \kappa_n.$$

Therefore, it suffices to show large deviation results for a_n and b_n.

D.2.2. *Taylor expansion of* $\psi_i(Z_i, \hat{\eta}, \hat{G}_n) - \psi_i(Z_i, \eta_0, \hat{G}_n)$. Define $\Delta_{mi} = \hat{m}_i - m_{0i}$, $\Delta_{si} = \hat{s}_i - s_{0i}$, and $\Delta_i = [\Delta_{mi}, \Delta_{si}]'$. Recall $\|\hat{\eta} - \eta\|_\infty = \max(\|s - s_0\|_\infty, \|m - m_0\|_\infty)$ as in (C.5). Since $\psi_i(Z_i, \eta, G)$

is smooth in $(m_i, s_i) \in \mathbb{R} \times \mathbb{R}_{>0}$, we can take a second-order Taylor expansion:

$$\psi_i\left(Z_i, \hat{\eta}, \hat{G}_n\right) - \psi_i\left(Z_i, \eta_0, \hat{G}_n\right) = \frac{\partial \psi_i}{\partial m_i}\bigg|_{\eta_0, \hat{G}_n} \Delta_{mi} + \frac{\partial \psi_i}{\partial s_i}\bigg|_{\eta_0, \hat{G}_n} \Delta_{si} + \underbrace{\frac{1}{2}\Delta_i' H_i(\tilde{\eta}_i, \hat{G}_n)\Delta_i}_{R_{1i}} \tag{D.4}$$

where $H_i(\tilde{\eta}_i, \hat{G}_n)$ is the Hessian matrix $\frac{\partial^2 \psi_i}{\partial \eta_i \partial \eta_i'}$ evaluated at some intermediate value $\tilde{\eta}_i$ lying on the line segment between $\hat{\eta}_i$ and η_{0i}.

We further decompose the first-order terms into an empirical process term and a mean-component term. By Lemma D.9, (D.26), and (D.28), for

$$\rho_n = \frac{1}{n^3} e^{-C_{\mathcal{H}} M_n^2 \Delta_n} \wedge \frac{1}{e\sqrt{2\pi}}, \tag{D.5}$$

we have that the numerators to the first derivatives can be truncated at ρ_n, as the truncation does not bind:

$$\begin{aligned}
\frac{\partial \psi_i}{\partial m_i}\bigg|_{\eta_0, \hat{G}_n} &= -\frac{1}{s_i} \frac{f'_{i,\hat{G}_n}}{f_{i,\hat{G}_n} \vee \frac{\rho_n}{\nu_i}} \equiv D_{m,i}(Z_i, \hat{G}_n, \eta_0, \rho_n) \\
\frac{\partial \psi_i}{\partial s_i}\bigg|_{\eta_0, \hat{G}_n} &= \frac{s_i}{\sigma_i^2} \frac{Q_i(Z_i, \eta_0, \hat{G}_n)}{f_{i,\hat{G}_n} \vee \frac{\rho_n}{\nu_i}} \equiv D_{s,i}(Z_i, \hat{G}_n, \eta_0, \rho_n).
\end{aligned}$$

Let

$$\overline{D}_{k,i}(\hat{G}_n, \eta_0, \rho_n) = \int D_{k,i}(z, \hat{G}_n, \eta_0, \rho_n)\, f_{G_0,\nu_i}(z) dz \quad \text{for } k \in \{m, s\}$$

be the mean of $D_{k,i}$. Then, for $k \in \{m, s\}$,

$$\frac{\partial \psi_i}{\partial k_i}\bigg|_{\eta_0, \hat{G}_n} \Delta_{ki} = \left[D_{k,i}(Z_i, \hat{G}_n, \eta_0, \rho_n) - \overline{D}_{k,i}(\hat{G}_n, \eta_0, \rho_n)\right] \Delta_{ki} + \overline{D}_{k,i}(\hat{G}_n, \eta_0, \rho_n)\Delta_{ki}$$

Hence, we can decompose the first-order terms in a_n as

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \frac{\partial \psi_i}{\partial k_i}\bigg|_{\eta_0, \hat{G}_n} \Delta_{ki} &= \frac{1}{n}\sum_{i=1}^n \overline{D}_{k,i}(\hat{G}_n, \eta_0, \rho_n)\Delta_{ki} + \frac{1}{n}\sum_{i=1}^n \left[D_{k,i}(Z_i, \hat{G}_n, \eta_0, \rho_n) - \overline{D}_{k,i}(\hat{G}_n, \eta_0, \rho_n)\right]\Delta_{ki} \\
&\equiv U_{1k} + U_{2k}
\end{aligned}$$

Let the second order term be $R_1 = \frac{1}{n}\sum_i R_{1i}$. We let $a_n = |R_1| + \sum_{k\in\{m,s\}} |U_{1k}| + |U_{2k}|$

D.2.3. *Taylor expansion of $\psi_i(Z_i, \hat{\eta}, G_0) - \psi_i(Z_i, \eta_0, G_0)$.* Like (D.4), we similarly decompose

$$\begin{aligned}
\psi_i(Z_i, \hat{\eta}, G_0) - \psi_i(Z_i, \eta_0, G_0) &= \frac{\partial \psi_i}{\partial m_i}\bigg|_{\eta_0, G_0} \Delta_{mi} + \frac{\partial \psi_i}{\partial s_i}\bigg|_{\eta_0, G_0} \Delta_{si} + \underbrace{\frac{1}{2}\Delta_i' H_i(\tilde{\eta}_i, G_0)\Delta_i}_{R_{2i}} \qquad \text{(D.6)} \\
&= \sum_{k\in\{m,s\}} D_{k,i}(Z_i, G_0, \eta_0, 0)\Delta_{ki} + R_{2i} \\
&\equiv U_{3mi} + U_{3si} + R_{2i}. \qquad \text{(D.7)}
\end{aligned}$$

Let $U_{3k} = \frac{1}{n}\sum_i U_{3ki}$ for $k \in \{m, s\}$ and let $R_2 = \frac{1}{n}\sum_i R_{2i}$. We let $b_n = |R_2| + \sum_{k\in\{m,s\}} |U_{3k}| + |U_{3k}|$

D.2.4. *Bounding each term individually.* By our decomposition, we can write

$$a_n + b_n + \kappa_n \le \kappa_n + |R_1| + |R_2| + \sum_{k\in\{m,s\}} |U_{1k}| + |U_{2k}| + |U_{3k}|$$

The ensuing subsections bound each term individually. Here we give an overview of the main ideas:

(1) We bound $\mathbb{1}(A_n)|U_{1m}|$ in almost sure terms in Lemma D.1 by observing that $|\overline{D}_{mi}|$ is small when $\hat{G}_n$ is close to G_0, since $\overline{D}_{mi}(G_0, \eta_0, 0) = 0$. To do so, we need to control the differences

$$\overline{D}_{mi}(\hat{G}_n, \eta_0, \rho_n) - \overline{D}_{mi}(G_0, \eta_0, \rho_n)$$

and

$$\overline{D}_{mi}(G_0, \eta_0, \rho_n) - \underbrace{\overline{D}_{mi}(G_0, \eta_0, 0)}_{=0} = \overline{D}_{mi}(G_0, \eta_0, \rho_n).$$

Controlling the first difference features the Hellinger distance, while controlling the second relies on the fact that $\mathrm{P}_{X\sim f(X)}(f(X) \le \rho)$ cannot be too large, by a Chebyshev's inequality argument in Lemma D.12. Similarly, we bound $\mathbb{1}(A_n)|U_{1s}|$ in Lemma D.2.

(2) The empirical process terms U_{2m}, U_{2s} are bounded probabilistically in Lemmas D.3 and D.4 with statements of the form

$$\mathrm{P}(A_n, |U_{2k}| > c_1) \le c_2.$$

To do so, we upper bound $\mathbb{1}(A_n)U_{2k} \le \overline{U}_{2k}$ in almost sure terms. The upper bound is obtained by projecting $\hat{G}_n$ onto a ω-net of $\mathcal{P}(\mathbb{R})$ in terms of some pseudo-metric d_{k,∞,M_n} induced by $\overline{D}_{k,i}$. The upper bound $\overline{U}_{2k}$ then takes the form

$$\omega\Delta_n + \max_{j\in[N]} \sup_{\eta\in S} \left| \frac{1}{n}\sum_i (D_{ki} - \overline{D}_{ki})(\eta_i - \eta_{0i}) \right| \quad N \le N(\omega, \mathcal{P}(\mathbb{R}), d_{k,\infty,M_n}).$$

Large deviation of $\overline{U}_{2k}$ is further controlled by applying Dudley's chaining argument (Vershynin, 2018), since the entropy integral over Hölder spaces is well-behaved. The covering number N is controlled via Proposition D.1 and Proposition D.2, which are minor extensions to Lemma 4 and Theorem 7 in Jiang (2020). The covering number is of a manageable size since the induced distributions f_{G,ν_i} are very smooth.

(3) Since $\overline{D}_{k,i}(G_0, \eta_0, 0) = 0$. U_{3m}, U_{3s} are effectively also empirical process terms, without the additional randomness in $\hat{G}_n$. Thus the ω-net argument above is unnecessary for U_{3m}, U_{3s}, whereas the bounding follows from the same Dudley's chaining argument. Lemma D.6 bounds U_{3k}.

(4) For the second derivative terms R_1, R_2, we observe that the second derivatives take the form of functions of posterior moments. The posterior moments under prior $\hat{G}_n$ is bounded within constant factors of M_n^q since the support of G_n is restricted. The posterior moments under prior G_0 is bounded by $|Z_i|^q \lesssim_{\mathcal{H}} M_n^q$ as we show in Lemma D.18, thanks to the simultaneous moment control for G_0. Hence $\mathbb{1}(A_n)R_1$ can be bounded in almost sure terms. We bound $\mathbb{1}(A_n)R_2$ probabilistically. The second derivatives are bounded in Lemmas D.5 and D.7.

(1) and (4) above bounds U_{1k}, R_1, R_2 almost surely under A_n. (2) and (3) bounds U_{2k}, U_{3k} probabilistically. By a union bound in Lemma D.17, we can simply add the rates. Doing so, we find that the first term in ϵ_n (D.1) comes from U_{1s}, which dominates U_{1m}. The second term comes

from U_{2s}, which dominates U_{2m}. The third term comes from R_1, which dominates R_2. The fourth term comes from U_{3s}. The leading terms in ϵ_n dominate κ_n, recalling (3.3). This completes the proof.

Before we proceed to the individual lemmas, we highlight a few convenient facts:

- The support of $\hat{G}_n$ is within $[-\overline{M}_n, \overline{M}_n]$, where $\overline{M}_n = \max_i |\hat{Z}_i(\hat{\eta})| \vee 1$. Under Assumption D.1, $\mathbb{1}(A_n)\overline{M}_n \lesssim_{\mathcal{H}} M_n$ by Lemma D.11(3).
- As a result, moments of $\hat{G}_n$ and $f_{\hat{G}_n,\nu_i}$ is bounded by appropriate moments of M_n, up to constants, under A_n.

D.3. **Bounding U_{1m}.**

Lemma D.1. *Under Assumptions 1 to 4, assume additionally that $\|\hat{\eta} - \eta\|_\infty \le \Delta_n, \overline{Z}_n \le M_n$. Assume that the rates satisfy Assumption D.1. Then*

$$|U_{1m}| \equiv \left| \frac{1}{n} \sum_{i=1}^n \overline{D}_{mi}(\hat{G}_n, \eta_0, \rho_n)\Delta_{mi} \right| \lesssim_{\mathcal{H}} \Delta_n \left[\frac{\log n}{n} \sum_{i=1}^n h(f_{G_0,\nu_i}, f_{\hat{G}_n,\nu_I}) + \frac{M_n^{1/3}}{n} \right]. \tag{D.8}$$

Proof. Note that

$$\begin{aligned} |\overline{D}_{m,i}(\hat{G}_n, \eta_0, \rho_n)| &\lesssim_{s_{0\ell}} \left| \int \frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}} f_{G_0,\nu_i}(z) dz \right| \\ &= \left| \int \frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}} [f_{G_0,\nu_i}(z) - f_{\hat{G}_n,\nu_i}(z) + f_{\hat{G}_n,\nu_i}(z)] dz \right| \\ &\le \left| \int \frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}} [f_{G_0,\nu_i}(z) - f_{\hat{G}_n,\nu_i}(z)]\, dz \right| \qquad \text{(D.9)} \\ &\quad + \left| \int \frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}} f_{\hat{G}_n,\nu_i}(z) dz \right| \qquad \text{(D.10)} \end{aligned}$$

By the bounds for (D.9) and (D.10) below, we have that

$$|U_{1m}| \lesssim_{\mathcal{H}} \Delta_n \left\{ \frac{\sqrt{\log n}}{n} \sum_{i=1}^n h(f_{G_0,\nu_i}, f_{\hat{G}_n,\hat{\nu}_i}) + \frac{M_n^{1/3}}{n} \right\}$$

by Assumption D.1. □

D.3.1. *Bounding* (D.9). Consider the first term (D.9):

$$\begin{aligned} &\left| \int \frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}} \left(f_{G_0,\nu_i}(z) - f_{\hat{G}_n,\nu_i}(z) \right) dz \right| \\ &= \left| \int \frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}} \left(\sqrt{f_{G_0,\nu_i}(z)} - \sqrt{f_{\hat{G}_n,\nu_i}(z)} \right) \left(\sqrt{f_{G_0,\nu_i}(z)} + \sqrt{f_{\hat{G}_n,\nu_i}(z)} \right) dz \right| \end{aligned}$$

$$\le \left\{ \underbrace{\int \left(\sqrt{f_{G_0,\nu_i}(z)} - \sqrt{f_{\hat{G}_n,\nu_i}(z)} \right)^2 dz}_{2h^2} \cdot \int \left(\frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}} \right)^2 \left(\sqrt{f_{G_0,\nu_i}(z)} + \sqrt{f_{\hat{G}_n,\nu_i}(z)} \right)^2 dz \right\}^{1/2}$$

(Cauchy–Schwarz)

$$\lesssim h(f_{G_0,\nu_i}, f_{\hat{G}_n,\nu_i}) \left\{ \int \left(\frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}} \right)^2 (f_{G_0,\nu_i}(z) + f_{\hat{G}_n,\nu_i}(z))\, dz \right\}^{1/2} \tag{D.11}$$

By Lemmas D.9 and D.10,

$$\left(\frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}} \right)^2 \lesssim \frac{1}{\nu_i} \log(1/\rho_n) \lesssim_{\mathcal{H}} \log n.$$

Hence,

$$\text{(D.9)} \lesssim_{\mathcal{H}} h(f_{G_0,\nu_i}, f_{\hat{G}_n,\nu_i}) \sqrt{\log n}$$

D.3.2. *Bounding* (D.10). The second term (D.10) is

$$\begin{aligned}
&\left| \int \frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}} f_{\hat{G}_n,\nu_i}(z)\, dz \right| \\
&= \left| \int \frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z)} \left(\frac{f_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}} - 1 \right) f_{\hat{G}_n,\nu_i}(z)\, dz \right| \\
&\le \int \left| \frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z)} \right| \mathbb{1}\left(f_{\hat{G}_n,\nu_i}(z) \le \rho_n/\nu_i \right) f_{\hat{G}_n,\nu_i}(z)\, dz \\
&\le \underbrace{\left(\mathbb{E}_{Z\sim f_{\hat{G}_n,\nu_i}} \left[\left(\mathbf{E}_{\hat{G}_n,\nu_i} \left[\frac{(\tau - Z)}{\nu_i^2} \mid Z \right] \right)^2 \right] \right)^{1/2}}_{\le \mathbb{E}_{\tau\sim\hat{G}_n, Z\sim\mathcal{N}(\tau,\nu_i)}[(\tau-Z)^2/\nu_i^4]^{1/2} = \nu_i^{-1}} \cdot \sqrt{\mathrm{P}_{f_{\hat{G}_n,\nu_i}}[f_{\hat{G}_n,\nu_i}(Z) \le \rho_n/\nu_i]}.
\end{aligned}$$

(Cauchy–Schwarz and (D.33))

By Jensen's inequality and law of iterated expectations, the first term is bounded by $\frac{1}{\nu_i}$. By Lemma D.12, the second term is bounded by $\rho_n^{1/3} \operatorname{Var}_{Z\sim f_{\hat{G}_n,\nu_i}}(Z)^{1/6}$. Now,

$$\operatorname{Var}_{Z\sim f_{\hat{G}_n,\nu_i}}(Z) \le \nu_i^2 + \mu_2^2(\hat{G}_n) \lesssim_{\mathcal{H}} M_n^2.$$

Hence,

$$\left| \int \frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}} f_{\hat{G}_n,\nu_i}(z)\, dz \right| \lesssim_{\mathcal{H}} M_n^{1/3} \rho_n^{1/3} \lesssim_{\mathcal{H}} M_n^{1/3} n^{-1}. \qquad \text{(Lemma D.9)}$$

D.4. **Bounding U_{1s}.**

Lemma D.2. *Under Assumptions 1 to 4 and D.1, if $\|\hat{\eta} - \eta\|_\infty \le \Delta_n$, $\overline{Z}_n \le M_n$, then*

$$|U_{1s}| \lesssim_{\mathcal{H}} \Delta_n \left[\frac{M_n \sqrt{\log n}}{n} \sum_{i=1}^n h(f_{\hat{G}_n,\nu_i}, f_{G_0,\nu_i}) + \frac{M^{4/3}}{n} \right]. \tag{D.12}$$

Proof. Similar to our computation with $\overline{D}_{m,i}$, we decompose

$$|\overline{D}_{s,i}(\hat{G}_n, \eta_0, \rho_n)| \lesssim_{\sigma_\ell,\sigma_u,s_{0\ell},s_{0u}} \left| \int \frac{Q_i(z, \eta_0, \hat{G}_n)}{f_{\hat{G}_n,\nu_i}(z) \vee (\rho_n/\nu_i)} (f_{G_0,\nu_i}(z) - f_{\hat{G}_n,\nu_i}(z))\, dz \right| \tag{D.13}$$

$$+ \left| \int \frac{Q_i(z, \eta_0, \hat{G}_n)}{f_{\hat{G}_n,\nu_i}(z) \vee (\rho_n/\nu_i)} f_{\hat{G}_n,\nu_i}(z)\, dz \right|. \tag{D.14}$$

We conclude the proof by plugging in our subsequent calculations. □

D.4.1. *Bounding* (D.13). The first term (D.13) is bounded by

$$\left(\int \frac{Q_i(z, \eta_0, \hat{G}_n)}{f_{\hat{G}_n,\nu_i}(z) \vee (\rho_n/\nu_i)} \left[f_{G_0,\nu_i}(z) - f_{\hat{G}_n,\nu_i}(z) \right] dz \right)^2$$

$$\lesssim h^2(f_{G_0,\nu_i}, f_{\hat{G}_n,\nu_i}) \int \left(\frac{Q_i(z, \eta_0, \hat{G}_n)}{f_{\hat{G}_n,\nu_i}(z) \vee (\rho_n/\nu_i)} \right)^2 \left[f_{G_0,\nu_i}(z) + f_{\hat{G}_n,\nu_i}(z) \right] dz,$$

similar to the computation in (D.11).

By Lemmas D.9 and D.13,

$$\left(\frac{Q_i(z, \eta_0, \hat{G}_n)}{f_{\hat{G}_n,\nu_i}(z) \vee (\rho_n/\nu_i)} \right)^2 \lesssim_{\sigma_\ell,\sigma_u,s_{0\ell},s_{0u}} (\sqrt{\log n} M_n + \log n)^2 \lesssim_{\mathcal{H}} M_n^2 \log n$$

Hence

$$\int \left(\frac{Q(z, \nu_i)}{f_{\hat{G}_n,\hat{\nu}_i}(z) \vee (\rho_n/\nu_i)} \right)^2 \left[f_{G_0,\nu_i}(z) + f_{\hat{G}_n,\nu_i}(z) \right] dz \lesssim_{\mathcal{H}} M_n^2 \log n.$$

Hence

$$\text{(D.13)} \lesssim_{\sigma_\ell,\sigma_u,s_{0\ell},s_{0u}} M_n \sqrt{\log n} h(f_{G_0,\nu_i}, f_{\hat{G}_n,\nu_i}). \tag{D.15}$$

D.4.2. *Bounding* (D.14). Observe that

$$\text{(D.14)} = \left| \int \frac{Q_i(z, \eta_0, \hat{G}_n)}{f_{\hat{G}_n,\nu_i}(z)} \left(\frac{f_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee (\rho_n/\nu_i)} - 1 \right) f_{\hat{G}_n,\nu_i}(z)\, dz \right|$$

Similar to our argument for (D.10), by Cauchy–Schwarz,

$$\text{(D.14)} \le \left(\mathbb{E}_{f_{\hat{G}_n,\nu_i}(z)} \left[(\mathbf{E}_{\hat{G}_n,\nu_i} [(Z-\tau)\tau \mid Z])^2 \right] \right)^{1/2} \sqrt{\mathrm{P}_{f_{\hat{G}_n,\nu_i}(z)}(f_{\hat{G}_n,\nu_i}(z) \le \rho_n/\nu_i)}$$

$$\lesssim_{\mathcal{H}} M_n \cdot \rho_n^{1/3} M_n^{1/3} \lesssim_{\mathcal{H}} \frac{M_n^{4/3}}{n}.$$

D.5. **Bounding U_{2m}.**

Lemma D.3. *Under Assumptions 1 to 4 and D.1,*

$$\mathrm{P}\left[\|\hat{\eta} - \eta\|_\infty \le \Delta_n, \overline{Z}_n \le M_n, |U_{2m}| \gtrsim_{\mathcal{H}} \sqrt{\log n} \Delta_n \left\{ e^{-C_{\mathcal{H}} M_n^\alpha} + \frac{\log n}{\sqrt{n}} + \frac{1}{(n\Delta_n^{1/p})^{1/2}} \right\} \right] \le \frac{2}{n}$$

Proof. We prove this claim by first showing that if $\|\hat{\eta} - \eta\|_\infty \le \Delta_n$ and $\overline{Z}_n \le M_n$, we can upper bound $|U_{2m}|$ by some stochastic quantity $\overline{U}_{2m}$. Now, observe that

$$\mathrm{P}\left[\|\hat{\eta} - \eta\|_\infty \le \Delta_n, \overline{Z}_n \le M_n, |U_{2m}| > t\right] \le \mathrm{P}\left[\|\hat{\eta} - \eta\|_\infty \le \Delta_n, \overline{Z}_n \le M_n, \overline{U}_{2m} > t\right] \le \mathrm{P}[\overline{U}_{2m} > t].$$

Hence, a stochastic upper bound on $\overline{U}_{2m}$ would verify the claim.

We now construct $\overline{U}_{2m}$ assuming $\|\hat{\eta} - \eta\|_\infty \le \Delta_n$ and $\overline{Z}_n \le M_n$. Let

$$D_{m,i,M_n}(Z_i, \hat{G}_n, \hat{\eta}, \rho_n) = D_{m,i}(Z_i, \hat{G}_n, \hat{\eta}, \rho_n)\mathbb{1}(|Z_i| \le M_n)$$

and let

$$\overline{D}_{m,i,M_n}(\hat{G}_n, \hat{\eta}, \rho_n) = \int D_{m,i}(z, \hat{G}_n, \hat{\eta}, \rho_n)\mathbb{1}(|z| \le M_n) f_{G_0,\nu_i}(z)\, dz.$$

On the event $\overline{Z}_n \le M_n$, $D_{m,i,M_n} = D_{m,i}$. We recall that

$$\begin{aligned} |U_{2m}| &= \left|\frac{1}{n}\sum_{i=1}^n (D_{m,i} - \overline{D}_{m,i})\Delta_{mi}\right| \\ &\le \left|\frac{1}{n}\sum_{i=1}^n (D_{m,i,M_n} - \overline{D}_{m,i,M_n})\Delta_{mi}\right| + \left|\frac{1}{n}\sum_{i=1}^n (\overline{D}_{m,i} - \overline{D}_{m,i,M_n})\Delta_{mi}\right|. \end{aligned}$$

Note that

$$\begin{aligned} |\overline{D}_{m,i} - \overline{D}_{m,i,M_n}| &\lesssim_{\sigma_\ell,\sigma_u,s_{0\ell},s_{0u}} \left|\int_{|z|>M_n} \underbrace{\frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee (\rho_n/\nu_i)}}_{\lesssim_{\mathcal{H}}\sqrt{\log n}, \quad \text{Lemmas D.9 and D.10}} f_{G_0,\nu_i}(z)\, dz\right| \\ &\lesssim_{\mathcal{H}} \sqrt{\log n}\, \mathrm{P}_{G_0,\nu_i}(|Z_i| > M_n) \end{aligned}$$

By Lemma D.16, $\mathrm{P}_{G_0,\nu_i}(|Z_i| > M_n) \le \exp\left(-C_{\alpha,A_0,\nu_u} M_n^\alpha\right)$. Hence, the second term $|\frac{1}{n}\sum_{i=1}^n (\overline{D}_{m,i} - \overline{D}_{m,i,M_n})\Delta_{mi}|$ is bounded above by $e^{-C_{\mathcal{H}} M_n^\alpha}\sqrt{\log n}\Delta_n$, up to constants.

Note that under our assumptions, $\max_i |\hat{Z}_i| \vee 1 \le C_{\mathcal{H}} M_n$. Let $\mathcal{L} = [-C_{\mathcal{H}} M_n, C_{\mathcal{H}} M_n] \equiv [-\overline{M}, \overline{M}]$. Define

$$S = \left\{(m, s) : \|m - m_0\| \le \Delta_n, \|s - s_0\| \le \Delta_n, (m, s) \in C^p_{A_1}([\sigma_\ell, \sigma_u])\right\}. \tag{D.16}$$

For two distributions G_1, G_2, define the following pseudo-metric

$$d_{m,\infty,M_n}(G_1, G_2) = \max_{i \in [n]} \sup_{|z| \le M_n} |D_{m,i}(z, G_1, \eta_0, \rho_n) - D_{m,i}(z, G_2, \eta_0, \rho_n)| \tag{D.17}$$

Let $G_1, \ldots, G_N$ be an ω-net of $\mathcal{P}(\mathcal{L})$ in terms of $d_{m,\infty,M_n}(G_1, G_2)$, where N is taken to be the covering number

$$N = N\left(\omega, \mathcal{P}(\mathcal{L}), d_{m,\infty,M_n}(\cdot, \cdot)\right).$$

Let G_{j^*} be a G_j where $d_{m,\infty,M_n}(\hat{G}_n, G_{j^*}) \le \omega$.

By construction, $|\overline{D}_{m,i,M_n}(\hat{G}_n, \hat{\eta}, \rho_n) - \overline{D}_{m,i,M_n}(G_{j^*}, \hat{\eta}, \rho_n)| \le \omega$ as well, since the integrand is bounded uniformly. Hence, by projecting $\hat{G}_n$ to G_{j^*}, we obtain

$$\left|\frac{1}{n}\sum_{i=1}^n (D_{m,i,M_n}(Z_i, \hat{G}_n, \eta_0, \rho_n) - \overline{D}_{m,i,M_n}(\hat{G}_n, \eta_0, \rho_n))(\hat{m}(\sigma_i) - m_0(\sigma_i))\right|$$

$$\leq 2\omega\Delta_n + \max_{j\in[N]}\left|\frac{1}{n}\sum_{i=1}^n (D_{m,i,M_n}(Z_i, G_j, \eta_0, \rho_n) - \overline{D}_{m,i,M_n}(G_j, \eta_0, \rho_n))(\hat{m}(\sigma_i) - m_0(\sigma_i))\right| \quad \text{(D.18)}$$

Next, consider the process

$$\eta \mapsto \frac{1}{n}\sum_{i=1}^n (D_{m,i,M_n}(Z_i, G_j, \eta_0, \rho_n) - \overline{D}_{m,i,M_n}(G_j, \eta_0, \rho_n))(m(\sigma_i) - m_0(\sigma_i))$$
$$\equiv \frac{1}{n}\sum_{i=1}^n v_{i,j}(\eta) \equiv V_{n,j}(\eta)$$

so that, when $\|\hat{\eta} - \eta\|_\infty \leq \Delta_n, \overline{Z}_n \leq M_n$,

$$\text{(D.18)} \lesssim \omega\Delta_n + \max_{j\in[N]}\sup_{\eta\in S}|V_{n,j}(\eta)|.$$

Thus, we can take

$$\overline{U}_{2m} = C_{\mathcal{H}}\left\{e^{-C_{\mathcal{H}}M_n^\alpha}\sqrt{\log n}\Delta_n + \omega\Delta_n + \max_{j\in[N]}\sup_{\eta\in S}|V_{n,j}(\eta)|\right\}$$

where we shall prove a stochastic upper bound and optimize ω shortly.

By the results in Appendix D.5.1 via Dudley's chaining argument, with probability at least $1 - 2/n$,

$$\max_{j\in[N]}\sup_{\eta\in S}|V_{n,j}(\eta)| \lesssim_{\mathcal{H}} \frac{\Delta_n\sqrt{\log n}}{\sqrt{n}}\left[\Delta_n^{-1/(2p)} + \sqrt{\log N} + \sqrt{\log n}\right]$$

By Appendix D.5.2, we can pick ω such that

$$\omega\Delta_n + \max_{j\in[N]}\sup_{\eta\in S}V_{nj}(\eta) \lesssim_{\mathcal{H}} \Delta_n\sqrt{\log n}\left(\frac{\log n}{\sqrt{n}} + \frac{1}{\sqrt{n\Delta_n^{1/p}}}\right) \quad \text{(D.19)}$$

with probability at least $1 - 2/n$. Putting these observations together, we have that

$$\mathrm{P}\left[\overline{U}_{2m} \gtrsim_{\mathcal{H}} \sqrt{\log n}\Delta_n\left\{e^{-C_{\mathcal{H}}M_n^\alpha} + \frac{\log n}{\sqrt{n}} + \frac{1}{(n\Delta_n^{1/p})^{1/2}}\right\}\right] \leq \frac{2}{n}.$$

This concludes the proof. $\square$

D.5.1. *Bounding* $\max_{j\in[N]}\sup_{\eta\in S}|V_{n,j}(\eta)|$. Note that $\mathbb{E}v_{ij}(\eta) = 0$. Moreover, by Lemmas D.9 and D.10,

$$\max\left(D_{m,i,M_n}(Z_i, G_j, \eta_0, \rho_n), \overline{D}_{m,i,M_n}(G_j, \eta_0, \rho_n)\right) \lesssim_{\mathcal{H}} \sqrt{\log(1/\rho_n)} \lesssim_{\mathcal{H}} \sqrt{\log n}$$

Recall that $\|\eta_1 - \eta_2\|_\infty = \max\left(\|m_1 - m_2\|_\infty, \|s_1 - s_2\|_\infty\right)$. Then,

$$|v_{ij}(\eta_1) - v_{ij}(\eta_2)| \lesssim_{\mathcal{H}} \sqrt{\log n}\|\eta_1 - \eta_2\|_\infty$$

As a result,[75]

$$\|V_{n,j}(\eta_1) - V_{n,j}(\eta_2)\|_{\psi_2} \lesssim_{\mathcal{H}} \frac{\sqrt{\log n}}{\sqrt{n}}\|\eta_1 - \eta_2\|_\infty.$$

[75]See Definition 2.5.6 in Vershynin (2018) for a definition of the ψ_2-norm (subgaussian norm).

Hence $V_{n,j}(\eta)$ is a mean-zero process with subgaussian increments[76] with respect to $\|\eta_1 - \eta_2\|_\infty$. Note that the diameter of S under $\|\eta_1 - \eta_2\|_\infty$ is at most $2\Delta_n$. Hence, by an application of Dudley's tail bound (Theorem 8.1.6 in Vershynin (2018)), for all $u > 0$,

$$\mathrm{P}\left[\sup_{\eta \in S} |V_{n,j}(\eta)| \gtrsim_{\mathcal{H}} \frac{\sqrt{\log n}}{\sqrt{n}} \left\{ \int_0^{2\Delta_n} \sqrt{\log N(\epsilon, S, \|\cdot\|_\infty)}\, d\epsilon + u\Delta_n \right\}\right] \leq 2e^{-u^2}.$$

Note that

$$\sqrt{\log N(\epsilon, S, \|\cdot\|_\infty)} \leq \sqrt{2 \log N(\epsilon, C^p_{A_1}([-\sigma_\ell, \sigma_u]), \|\cdot\|_\infty)} \leq \sqrt{2 \log N(\epsilon/A_1, C^p_1([-\sigma_\ell, \sigma_u]), \|\cdot\|_\infty)}$$

By Theorem 2.7.1 in van der Vaart and Wellner (1996),

$$\log N(\epsilon/A_1, C^p_1([-\sigma_\ell, \sigma_u]), \|\cdot\|_\infty) \lesssim_{p,\sigma_\ell,\sigma_u} \left(\frac{A_1}{\epsilon}\right)^{1/p} \lesssim_{\mathcal{H}} \left(\frac{1}{\epsilon}\right)^{1/p}.$$

Hence, plugging in these calculations, we obtain

$$\mathrm{P}\left[\sup_{\eta \in S} |V_{n,j}(\eta)| \gtrsim_{\mathcal{H}} \frac{\sqrt{\log n}}{\sqrt{n}} \left\{ \Delta_n^{1-\frac{1}{2p}} + u\Delta_n \right\}\right] \leq 2e^{-u^2}.$$

This implies that

$$\sup_{\eta \in S} |V_{n,j}(\eta)| \lesssim_{\mathcal{H}} \frac{\sqrt{\log n}}{\sqrt{n}} \Delta_n^{1-\frac{1}{2p}} + \tilde{V}_{n,j},$$

for some random variable $\tilde{V}_{n,j} \geq 0$ and $\|\tilde{V}_{n,j}\|_{\psi_2} \lesssim_{\mathcal{H}} \frac{\Delta_n}{\sqrt{n}} \sqrt{\log n}$.[77] Thus,

$$\text{(D.18)} \lesssim_{\mathcal{H}} \Delta_n \left[\omega + \frac{\sqrt{\log n}}{\sqrt{n\Delta_n^{1/p}}}\right] + \max_{j \in [N]} \tilde{V}_{n,j}.$$

Finally, note that by Lemma D.15 with the choice $t = \sqrt{\log n}$,

$$\mathrm{P}\left[\max_{j \in [N]} \tilde{V}_{n,j} \gtrsim_{\mathcal{H}} \frac{\Delta_n}{\sqrt{n}} \sqrt{\log n} \left[\sqrt{\log N} + \sqrt{\log n}\right]\right] \leq \frac{2}{n}.$$

D.5.2. *Selecting ω.* The rate function that involves ω and $\log N$ is of the form

$$\omega + \sqrt{\frac{\log N}{n}} \sqrt{\log n}$$

Reparametrizing $\omega = \delta \log(1/\delta) \frac{\sqrt{\log(1/\rho_n)}}{\rho_n}$, by Proposition D.2, shows that

$$\log N \leq \log N\left(\delta \log(1/\delta) \frac{\sqrt{\log(1/\rho_n)}}{\rho_n}, \mathcal{P}(\mathbb{R}), d_{m,\infty,M}\right) \lesssim_{\mathcal{H}} \log(1/\delta)^2 \max\left(1, \frac{M_n}{\sqrt{\log(1/\delta)}}\right)$$

[76] See Definition 8.1.1 in Vershynin (2018).

[77] We can take

$$\tilde{V}_{n,j} = \left\{ \sup_{\eta \in S} |V_{n,j}(\eta)| - C_{\mathcal{H}} \frac{M_n}{\sqrt{n}} \Delta_n^{1-\frac{1}{2p}} \right\}_+.$$

The tail bound $\mathrm{P}(\tilde{V}_{n,j} \gtrsim_{\mathcal{H}} u \frac{\Delta_n}{\sqrt{n}} M_n) \leq 2e^{-u^2}$ implies the ψ_2-norm bound by expression (2.14) in Vershynin (2018).

Consider picking $\delta = \rho_n \frac{1}{\sqrt{n}} \le 1/e$ so that $\log(1/\delta) = \log(1/\rho_n) + \frac{1}{2}\log n \lesssim_{\mathcal{H}} \log n$. Since $\log(1/\rho_n) \gtrsim M_n^2$, we conclude that $\max\left(1, \frac{M_n}{\sqrt{\log(1/\delta)}}\right) \lesssim_{\mathcal{H}} 1$. Hence,

$$\log N \lesssim_{\mathcal{H}} \log^2 n.$$

Note too that $\omega \lesssim_{\mathcal{H}} \frac{(\log n)^{3/2}}{\sqrt{n}}$. Thus, under Assumption D.1,

$$\omega + \sqrt{\log N}\frac{1}{\sqrt{n}}\sqrt{\log n} \lesssim_{\mathcal{H}} \frac{(\log n)^{3/2}}{\sqrt{n}}.$$

D.6. **Bounding U_{2s}.**

Lemma D.4. *Under Assumptions 1 to 4 and D.1,*

$$P\left[\|\hat{\eta} - \eta\|_\infty \le \Delta_n, \overline{Z}_n \le M_n, |U_{2s}| \gtrsim_{\mathcal{H}} \Delta_n M_n \sqrt{\log n}\left\{e^{-C_{\mathcal{H}} M_n^\alpha} + \frac{\log n}{\sqrt{n}} + \frac{1}{\sqrt{n\Delta_n^{1/p}}}\right\}\right] \le \frac{2}{n}$$

Proof. This proof operates much like the proof of Lemma D.3. We observe that we can come up with an upper bound $\overline{U}_{2s}$ of U_{2s} under the event $\|\hat{\eta} - \eta\|_\infty \le \Delta_n$ and $\overline{Z}_n \le M_n$. A stochastic upper bound on $\overline{U}_{2s}$ then implies the lemma.

Let us first assume $\|\hat{\eta} - \eta\|_\infty \le \Delta_n$ and $\overline{Z}_n \le M_n$. Define D_{s,i,M_n} and $\overline{D}_{s,i,M_n}$ analogously to D_{m,i,M_n} and $\overline{D}_{m,i,M_n}$. A similar decomposition shows

$$|U_{2s}| \le \left|\frac{1}{n}\sum_{i=1}^n (D_{s,i,M_n} - \overline{D}_{s,i,M_n})\Delta_{si}\right| + \left|\frac{1}{n}\sum_{i=1}^n (\overline{D}_{s,i} - \overline{D}_{s,i,M_n})\Delta_{si}\right|$$

Lemma D.13 is a uniform bound on the integrand in the second term. Hence, the second term is bounded by

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^n (\overline{D}_{s,i} - \overline{D}_{s,i,M_n})\Delta_{si}\right| \\
&\lesssim_{\mathcal{H}} \Delta_n \sqrt{\log(1/\rho_n)}\frac{1}{n}\sum_{i=1}^n \left(\int_{|Z_i|>M_n} |z| f_{G_0,\nu_i}(z)\,dz + \sqrt{\log(1/\rho_n)}\int_{|Z_i|>M_n} f_{G_0,\nu_i}(z)\,dz\right) \\
&\lesssim_{\mathcal{H}} \Delta_n \sqrt{\log n}\left\{e^{-\frac{C_{\mathcal{H}}}{2}M_n^\alpha}\max_{i\in[n]}\mu_2(f_{G_0,\nu_i}) + \sqrt{\log n}e^{-C_{\mathcal{H}}M_n^\alpha}\right\} \\
&\qquad\qquad\qquad\qquad \text{(Cauchy–Schwarz for the first term and apply Lemmas D.9 and D.16)} \\
&\lesssim_{\mathcal{H}} \Delta_n (\log n) e^{-C_{\mathcal{H}}M_n^\alpha}.
\end{aligned}$$

Note that under our assumptions, $\max_i |\hat{Z}_i| \vee 1 \le C_{\mathcal{H}} M_n$. Let $\mathcal{L} = [-C_{\mathcal{H}}M_n, C_{\mathcal{H}}M_n] \equiv [-\overline{M}, \overline{M}]$. Define $S = \left\{(m,s) : \|m - m_0\| \le \Delta_n, \|s - s_0\| \le \Delta_n, (m,s) \in C_{A_1}^p([\sigma_\ell, \sigma_u])\right\}$. For two distributions G_1, G_2, define the following pseudo-metric

$$d_{s,\infty,M_n}(G_1, G_2) = \max_{i\in[n]}\sup_{|z|\le M_n} |D_{s,i}(z, G_1, \eta_0, \rho_n) - D_{s,i}(z, G_2, \eta_0, \rho_n)| \tag{D.20}$$

Let $G_1, \ldots, G_N$ be an ω-net of $\mathcal{P}(\mathcal{L})$ in terms of $d_{s,\infty,M_n}(G_1, G_2)$, where

$$N = N\left(\omega, \mathcal{P}(\mathcal{L}), d_{s,\infty,M_n}(\cdot,\cdot)\right).$$

Let G_{j^*} be a G_j where $d_{s,\infty,M_n}(\hat{G}_n, G_{j^*}) \le \omega$. By construction, $|\overline{D}_{s,i,M_n}(\hat{G}_n, \eta_0, \rho_n) - \overline{D}_{s,i,M_n}(G_{j^*}, \eta_0, \rho_n)| \le \omega$ as well, since the integrand is bounded uniformly.

Hence

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^n (D_{s,i,M_n}(Z_i, \hat{G}_n, \eta_0, \rho_n) - \overline{D}_{s,i,M_n}(\hat{G}_n, \eta_0, \rho_n))(\hat{s}(\sigma_i) - s_0(\sigma_i))\right| \\ &\le 2\omega\Delta_n + \max_{j\in[N]}\left|\frac{1}{n}\sum_{i=1}^n (D_{s,i,M_n}(Z_i, G_j, \eta_0, \rho_n) - \overline{D}_{s,i,M_n}(G_j, \eta_0, \rho_n))(\hat{s}(\sigma_i) - s_0(\sigma_i))\right| \end{aligned} \tag{D.21}$$

Next, consider the process

$$\eta \mapsto \frac{1}{n}\sum_{i=1}^n (D_{s,i,M_n}(Z_i, G_j, \eta_0, 0) - \overline{D}_{s,i,M_n}(G_j, \eta_0, 0))(s(\sigma_i) - s_0(\sigma_i)) \equiv \frac{1}{n}\sum_{i=1}^n v_{i,j}(\eta) \equiv V_{n,j}(\eta)$$

so that (D.21) $\lesssim \omega\Delta_n + \max_{j\in[N]}\sup_{\eta\in S}|V_{n,j}(\eta)|$. This again upper bounds $|U_{is}|$ with some $\overline{U}_{is}$ that does not depend on the event $\|\hat{\eta} - \eta\|_\infty \le \Delta_n, \overline{Z}_n \le M_n$, on the event $\|\hat{\eta} - \eta\|_\infty \le \Delta_n, \overline{Z}_n \le M_n$. Hence, we can choose

$$\overline{U}_{2s} = C_{\mathcal{H}}\left\{\omega\Delta_n + \max_{j\in[N]}\sup_{\eta\in S}|V_{n,j}(\eta)| + \Delta_n(\log n)e^{-C_{\mathcal{H}}M_n^\alpha}\right\}.$$

It remains to show a tail bound with an appropriate choice of ω for $\overline{U}_{2s}$.

By Lemma D.13, the process $V_{n,j}$ has the subgaussian increment property

$$|V_{n,j}(\eta_1) - V_{n,j}(\eta_2)| \lesssim_{\mathcal{H}} \frac{M_n\sqrt{\log n}}{\sqrt{n}}\|\eta_1 - \eta_2\|_\infty$$

as in Appendix D.5.1, with a different constant for the subgaussianity. Hence, by the same argument as in Appendix D.5.1, with probability at least $1 - 2/n$,

$$\max_{j\in[N]}\sup_{\eta\in S}|V_{n,j}(\eta)| \lesssim_{\mathcal{H}} \frac{\Delta_n M_n\sqrt{\log n}}{\sqrt{n}}\left[\Delta_n^{-1/(2p)} + \sqrt{\log N} + \sqrt{\log n}\right]$$

We turn to selecting ω. The relevant term for selecting ω is $\omega + \frac{M_n\sqrt{\log n}}{\sqrt{n}}\sqrt{\log N}$. Reparametrize $\omega = M_n\sqrt{\log(1/\rho_n)}\delta\log(1/\delta)/\rho_n$. Pick $\delta = \rho_n/\sqrt{n} < 1/e$. The same argument as in Appendix D.5.2 with Proposition D.2 shows that

$$\omega + \frac{M_n\sqrt{\log n}}{\sqrt{n}}\sqrt{\log N} \lesssim_{\mathcal{H}} \frac{M_n(\log n)^{3/2}}{\sqrt{n}}.$$

Therefore, we can select ω such that, overall, with probability at least $1 - 2/n$, under Assumption D.1,

$$\begin{aligned}\overline{U}_{2s} &\lesssim_{\mathcal{H}} \Delta_n\left\{M_n\sqrt{\log n}\exp\left(-C_{\alpha,A_0,\nu_u}M_n^\alpha\right) + \frac{M_n(\log n)^{3/2}}{\sqrt{n}} + M_n\sqrt{\log n}\frac{1}{\sqrt{n\Delta_n^{1/p}}} + \frac{\sqrt{\log n}}{\sqrt{n}}M_n\sqrt{\log n}\right\} \\ &\lesssim_{\mathcal{H}} \Delta_n M_n\sqrt{\log n}\left\{e^{-C_{\mathcal{H}}M_n^\alpha} + \frac{\log n}{\sqrt{n}} + \frac{1}{\sqrt{n\Delta_n^{1/p}}}\right\}.\end{aligned}$$

This concludes the proof. ∎

D.7. **Bounding R_1.**

Lemma D.5. *Recall R_{1i} from (D.4). Then, under Assumptions 1 to 4 and D.1, if $\|\hat{\eta} - \eta\|_\infty \le \Delta_n$ and $\overline{Z}_n \le M_n$, then $R_{1i} \lesssim_{\mathcal{H}} \Delta_n^2 M_n^2 \log n$.*

Proof. Observe that $R_{1i} \lesssim_{\sigma_\ell, \sigma_u, s_{0\ell}, s_{0u}} \max\left(\Delta_{mi}^2, \Delta_{si}^2\right) \cdot \|H_i(\tilde{\eta}_i, \hat{G}_n)\|_\infty$, where $\|\cdot\|_\infty$ takes the largest element from a matrix by magnitude. By assumption, the first term is bounded by Δ_n^2. By Lemma D.14, the second derivatives are bounded by $M_n^2 \log n$. Hence $\|H_i(\tilde{\eta}_i, \hat{G}_n)\|_\infty \lesssim_{\mathcal{H}} M_n^2 \log n$. This concludes the proof. ∎

D.8. **Bounding U_{3m}, U_{3s}.**

Lemma D.6. *Under Assumptions 2 to 4 and D.1,*

$$\mathrm{P}\left[\|\hat{\eta} - \eta\|_\infty \le \Delta_n, \overline{Z}_n \le M_n, |U_{3m}| \gtrsim_{\mathcal{H}} \Delta_n \left\{e^{-C_{\mathcal{H}} M_n^\alpha} + \frac{M_n}{\sqrt{n}}\left(\Delta_n^{-1/(2p)} + \log n\right)\right\}\right] \le \frac{2}{n}$$

$$\mathrm{P}\left[\|\hat{\eta} - \eta\|_\infty \le \Delta_n, \overline{Z}_n \le M_n, |U_{3s}| \gtrsim_{\mathcal{H}} \Delta_n \left\{e^{-C_{\mathcal{H}} M_n^\alpha} + \frac{M_n^2}{\sqrt{n}}\left(\Delta_n^{-1/(2p)} + \log n\right)\right\}\right] \le \frac{2}{n}.$$

Proof. The proof structure follows that of Lemmas D.3 and D.4.

Recall that

$$\begin{aligned} U_{3m} &= \frac{1}{n}\sum_{i=1}^n D_{m,i}(Z_i, G_0, \eta_0, 0)(\hat{m}_i - m_0). \\ &= \frac{1}{n}\sum_{i=1}^n (D_{m,i,M_n} - \overline{D}_{m,i,M_n})(\hat{m}_i - m_0) + \overline{D}_{m,i,M_n}(\hat{m}_i - m_0) \end{aligned}$$

Note that

$$\begin{aligned} |\overline{D}_{m,i,M_n}| &= \left|\int_{|z| \le M_n} \frac{f'_{G_0,\nu_i}(z)}{f_{G_0,\nu_i}(z)} f_{G_0,\nu_i}(z)\, dz\right| \\ &= \left|\int \mathbb{1}\left(|z| > M_n\right) \cdot \frac{f'_{G_0,\nu_i}(z)}{f_{G_0,\nu_i}(z)} f_{G_0,\nu_i}(z)\, dz\right| \\ &\lesssim_{\sigma_\ell, \sigma_u, s_{0\ell}, s_{0u}} \mathrm{P}(|z| > M_n)^{1/2} \\ &\qquad \text{(Cauchy–Schwarz, Jensen, and law of iterated expectations via (D.33))} \\ &\lesssim_{\mathcal{H}} e^{-C_{\mathcal{H}} M_n^\alpha}. \end{aligned} \tag{D.22}$$

Recall S in (D.16). Define the process $V_n(\eta) = \frac{1}{n}\sum_i v_{n,i}(\eta) \equiv \frac{1}{n}\sum_{i=1}^n (D_{m,i,M_n} - \overline{D}_{m,i,M_n})(\hat{m}_i - m_0)$. Therefore, if $\|\hat{\eta} - \eta\|_\infty \le \Delta_n, \overline{Z}_n \le M_n$,

$$|U_{3m}| \lesssim_{\mathcal{H}} \Delta_n e^{-C_{\mathcal{H}} M_n^\alpha} + \sup_{\eta \in S} |V_n(\eta)| \equiv \overline{U}_{3m}.$$

Therefore, to bound U_{3m} it suffices to show a tail bound for $\sup_{\eta \in S} |V_n(\eta)|$. Observe that

$$V_n(\eta_1) - V_n(\eta_2) = \frac{1}{n}\sum_i (D_{m,i,M_n} - \overline{D}_{m,i,M_n})(\eta_{1i} - \eta_{2i})$$

Now, by Lemma 2.6.8 in Vershynin (2018), since $|D_{m,i,M_n}| \lesssim_{\mathcal{H}} M_n$ by Lemma D.18,

$$\|v_{ni}(\eta_1) - v_{ni}(\eta_2)\|_{\psi_2} \lesssim \|D_{m,i,M_n}(\eta_{1i} - \eta_{2i})\|_{\psi_2} \lesssim_{\mathcal{H}} M_n \|\eta_1 - \eta_2\|_\infty.$$

Since $v_{ni}(\eta_1) - v_{ni}(\eta_2)$ is mean zero, we have that

$$\|V_n(\eta_1) - V_n(\eta_2)\|_{\psi_2} \lesssim_{\mathcal{H}} \frac{M_n}{\sqrt{n}} \|\eta_1 - \eta_2\|_\infty \tag{D.23}$$

Hence, by the same Dudley's chaining calculation in Appendix D.5.1, with probability at least $1 - 2/n$,

$$\overline{U}_{3m} \lesssim_{\mathcal{H}} \Delta_n \left\{ e^{-C_{\mathcal{H}} M_n^\alpha} + \frac{M_n}{\sqrt{n}} \left(\Delta_n^{-1/(2p)} + \log n \right) \right\}.$$

This concludes the proof for U_{3m}.

The proof for U_{3s} is similar. We need to establish the analogue of (D.22) and (D.23). For the tail bound (analogue of (D.22)), we have the same bound

$$|\overline{D}_{s,i,M_n}| \lesssim \mathrm{P}\left(|z| > M_n\right)^{1/2} \left(\mathbb{E}_{f_{G_0,\nu_i}(z)} \left[(\mathbf{E}_{G_0,\nu_i}\left[(Z-\tau)\tau \mid Z\right])^2 \right] \right)^{1/2} \lesssim_{\mathcal{H}} e^{-C_{\mathcal{H}} M_n^\alpha}.$$

For the analogue of (D.23), since Lemma D.18 implies that $|D_{s,i,M_n}| \lesssim_{\mathcal{H}} Z_i^2 \mathbb{1}(Z_i \le M_n) \le M_n^2$,

$$\|V_n(\eta_1) - V_n(\eta_2)\|_{\psi_2} \lesssim_{\mathcal{H}} \frac{M_n^2}{\sqrt{n}} \|\eta_1 - \eta_2\|_\infty.$$

Hence, with probability at most $2/n$

$$\overline{U}_{3s} \gtrsim_{\mathcal{H}} \Delta_n \left\{ e^{-C_{\mathcal{H}} M_n^\alpha} + \frac{M_n^2}{\sqrt{n}} (\Delta_n^{-1/(2p)} + \log n) \right\}.$$

$\square$

D.9. **Bounding R_2.**

Lemma D.7. *Under Assumptions 2 to 4 and D.1, then* $\mathrm{P}\left(\|\hat{\eta} - \eta\|_\infty \le \Delta_n, \overline{Z}_n \le M_n, |R_2| \gtrsim_{\mathcal{H}} \Delta_n^2\right) \le \frac{1}{n}$.

Proof. Recall that $\mathbb{1}(A_n) = \mathbb{1}(\|\hat{\eta} - \eta\|_\infty \le \Delta_n, \overline{Z}_n \le M_n)$. Note that

$$\mathbb{1}(A_n)|R_2| \lesssim_{\mathcal{H}} \Delta_n^2 \frac{1}{n} \sum_{i=1}^n \mathbb{1}(A_n) \|H_i\|_\infty.$$

by $(1, \infty)$-Hölder inequality. Moreover, note that the second derivatives that occur in entries of H_i are functions of posterior moments. By Lemma D.18, under G_0, these posterior moments are bounded by above by corresponding moments of $\hat{Z}_i(\tilde{\eta}_i)$. By Lemma D.18, under G_0, these posterior moments are bounded by above by corresponding moments of $\hat{Z}_i(\tilde{\eta}_i)$. Hence,

$$\mathbb{1}(A_n)\|H_i\|_\infty \lesssim_{\mathcal{H}} \mathbb{1}(A_n) \left(\hat{Z}_i(\tilde{\eta}_i) \vee 1 \right)^4 \lesssim_{\mathcal{H}} (Z_i \vee 1)^4. \tag{D.24}$$

Hence,

$$\mathbb{1}(A_n)|R_2| \lesssim_{\mathcal{H}} \Delta_n^2 \frac{1}{n} \sum_{i=1}^n (Z_i \vee 1)^4.$$

By Chebyshev's inequality,

$$P\left(\frac{1}{n}\sum_{i=1}^n (Z_i \vee 1)^4 > \mathbb{E}[(Z_i \vee 1)^4] + t\right) \le \frac{1}{t^2}\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^n (Z_i \vee 1)^4\right) = \frac{\operatorname{Var}(Z_i^4 \vee 1)}{nt^2}.$$

Picking $t^2 = \operatorname{Var}(Z_i^4 \vee 1)$ yields that

$$P\left(\frac{1}{n}\sum_{i=1}^n (Z_i \vee 1)^4 \gtrsim_{\mathcal{H}} 1\right) \le \frac{1}{n}.$$

Hence,

$$P\left(\|\hat{\eta} - \eta\|_\infty \le \Delta_n, \overline{Z}_n \le M_n, |R_2| \gtrsim_{\mathcal{H}} \Delta_n^2\right) \le \frac{1}{n}.$$

□

D.10. **Derivative computations.** It is sometimes useful to relate the derivatives of ψ_i to $\mathbf{E}_{G,\eta}$.

We compute the following derivatives. Since they are all evaluated at G, η, we let $\hat{\nu} = \hat{\nu}_i(\eta)$ and $\hat{z} = \hat{Z}_i(\eta)$ as a shorthand.

$$\left.\frac{\partial \psi_i}{\partial m_i}\right|_{\eta,G} = -\frac{1}{s_i}\frac{f'_{G,\hat{\nu}}(\hat{z})}{f_{G,\hat{\nu}}(\hat{z})} \tag{D.25}$$

$$= \frac{s_i}{\sigma_i^2}\mathbf{E}_{G,\hat{\nu}}[Z - \tau \mid \hat{z}] \tag{D.26}$$

$$\left.\frac{\partial \psi_i}{\partial s_i}\right|_{\eta,G} = \frac{1}{\sigma_i \hat{\nu}_i(\eta) f_{G,\hat{\nu}(\eta)}(\hat{Z}_i(\eta))} \underbrace{\int (\hat{Z}_i(\eta) - \tau)\tau\varphi\left(\frac{\hat{Z}_i(\eta) - \tau}{\hat{\nu}_i(\eta)}\right)\frac{1}{\hat{\nu}_i(\eta)}\, G(d\tau)}_{Q_i(Z_i,\eta,G)} \tag{D.27}$$

$$= \frac{1}{\sigma_i \hat{\nu}}\mathbf{E}_{G,\hat{\nu}}[(Z - \tau)\tau \mid \hat{z}] \tag{D.28}$$

$$\left.\frac{\partial^2 \psi_i}{\partial m_i^2}\right|_{\eta,G} = \frac{1}{s_i^2}\left[\frac{f''_{G,\hat{\nu}}(\hat{z})}{f_{G,\hat{\nu}}(\hat{z})} - \left(\frac{f'_{G,\hat{\nu}}(\hat{z})}{f_{G,\hat{\nu}}(\hat{z})}\right)^2\right] \tag{D.29}$$

$$= \frac{1}{s_i^2}\left[\frac{1}{\hat{\nu}^4}\mathbf{E}_{G,\hat{\nu}}[(\tau - Z)^2 \mid \hat{z}] - \frac{1}{\hat{\nu}^2} - \frac{1}{\hat{\nu}^4}\left(\mathbf{E}_{G,\hat{\nu}}[(\tau - Z) \mid \hat{z}]\right)^2\right] \tag{D.30}$$

$$\left.\frac{\partial^2 \psi_i}{\partial m_i \partial s_i}\right|_{\eta,G} = \left(\frac{1}{\sigma_i^2}\mathbf{E}_{G,\hat{\nu}}[(Z - \tau)\tau \mid \hat{z}] - \frac{1}{s_i^2}\right)\frac{1}{\hat{\nu}^2}\mathbf{E}_{G,\hat{\nu}}[(\tau - Z) \mid \hat{z}] + \frac{\mathbf{E}_{G,\hat{\nu}}[(\tau - Z)^2\tau \mid \hat{z}]}{\hat{\nu}\sigma_i s_i} \tag{D.31}$$

$$\left.\frac{\partial^2 \psi}{\partial s^2}\right|_{\eta,G} = \frac{1}{\sigma_i^2}\left\{\mathbf{E}_{G,\hat{\nu}}\left[\left(\frac{s_i^2}{\sigma_i}(Z - \tau)^2 - 1\right)\tau^2 \mid \hat{z}\right] - \frac{1}{\hat{\nu}^2}\left(\mathbf{E}_{G,\hat{\nu}}[(Z - \tau)\tau \mid \hat{z}]\right)^2\right\} \tag{D.32}$$

It is useful to note that

$$\frac{f'_{G,\nu}(z)}{f_{G,\nu}(z)} = \frac{1}{\nu^2}\mathbf{E}_{G,\nu}[(\tau - Z) \mid z] \tag{D.33}$$

$$\frac{f''_{G,\nu}(z)}{f_{G,\nu}(z)} = \frac{1}{\nu^4}\mathbf{E}_{G,\nu}[(\tau - Z)^2 \mid z] - \frac{1}{\nu^2} \tag{D.34}$$

D.11. **Metric entropy of $\mathcal{P}(\mathbb{R})$ under moment-based distance.** The following is a minor generalization of Lemma 4 and Theorem 7 in Jiang (2020). In particular, Jiang (2020)'s Lemma

4 reduces to the case $q = 0$ below, and Jiang (2020)'s Theorem 7 relies on the results below for $q = 0, 1$. The proof largely follows the proofs of these two results of Jiang (2020).

We first state the following fact readily verified by differentiation.

Lemma D.8. *For all integer $m \geq 0$:*

$$\sup_{t \in \mathbb{R}} |t^m \varphi(t)| = m^{m/2} \varphi(\sqrt{m}).$$

As a corollary, there exists absolute $C_m > 0$ such that $t \mapsto t^m \varphi(t)$ is C_m-Lipschitz.

Proposition D.1. *Fix some $q \in \mathbb{N} \cup \{0\}$ and $M > 1$. Consider the pseudometric*

$$d_{\infty,M}^{(q)}(G_1, G_2) = \max_{i \in [n]} \underbrace{\max_{0 \leq v \leq q} \sup_{|x| \leq M} \left| \int \frac{(u-x)^v}{\nu_i^v} \varphi\left(\frac{x-u}{\nu_i}\right) (G_1 - G_2)(du) \right|}_{d_{q,i,m}(G_1, G_2)}.$$

Let ν_ℓ, ν_u be the lower and upper bounds of ν_i. Then, for all $0 < \delta < \exp(-q/2) \wedge e^{-1}$,

$$\log N(\delta \log^{q/2}(1/\delta), \mathcal{P}(\mathbb{R}), d_{\infty,M}^{(q)}) \lesssim_{q,\nu_u,\nu_\ell} \log^2(1/\delta) \max\left(\frac{M}{\sqrt{\log(1/\delta)}}, 1 \right).$$

Proof. The proof strategy is as follows. First, we discretize $[-M, M]$ into a union of small intervals I_j. Fix G. There exists a finitely supported distribution G_m that matches moments of G on every I_j. It turns out that such a G_m is close to G in terms of $\|\cdot\|_{q,\infty,M}$. Next, we discipline G_m by approximating G_m with $G_{m,\omega}$, a finitely supported distribution supported on the fixed grid $\{k\omega : k \in \mathbb{Z}\} \cap [-M, M]$. Finally, the set of all $G_{m,\omega}$'s may be approximated by a finite set of distributions, and we count the size of this finite set.

D.11.1. *Approximating G with G_m.* First, let us fix some $\omega < \varphi(\sqrt{q}) \wedge \varphi(1)$.

Let $a = \frac{\nu_u}{\nu_\ell} \varphi_+(\omega) \geq 1$. Let $I_j = [-M + (j-2)a\nu_\ell, -M + (j-1)a\nu_\ell]$ be such that

$$I = [-M - a\nu_\ell, +M + a\nu_\ell] \subset \bigcup_j I_j$$

where I_j is a width $a\nu_\ell$ interval. Let $j^* = \lceil \frac{2M}{a\nu_\ell} + 2 \rceil$ be the number of such intervals.

There exists by Carathéodory's theorem a distribution G_m with support on I and no more than

$$m = (2k^* + q + 1)j^* + 1$$

support points s.t. the moments match

$$\int_{I_j} u^k dG(u) = \int_{I_j} u^k dG_m(u) \text{ for all } k = 0, \ldots, 2k^* + q \text{ and } j = 1, \ldots, j^*.$$

for some k^* to be chosen later.

Then, for some $x \in I_j \cap [-M, M]$, we have

$$d_{q,i,M}(G, G_m) \leq \max_{0 \leq v \leq q} \left[\left| \int_{(I_{j-1} \cup I_j \cup I_{j+1})^{\mathrm{C}}} \left(\frac{u-x}{\nu_i}\right)^v \varphi\left(\frac{x-u}{\nu_i}\right) (G(du) - G_m(du)) \right| \right. \tag{D.35}$$

$$+\left|\int_{I_{j-1}\cup I_j\cup I_{j+1}}\left(\frac{u-x}{\nu_i}\right)^v\varphi\left(\frac{x-u}{\nu_i}\right)(G(du)-G_m(du))\right|\Bigg] \tag{D.36}$$

Note that $t^v\varphi(t)$ is a decreasing function for all $t>\sqrt{v}$. Note that $\omega<\varphi(\sqrt{q})$ implies that $a\nu_u/\nu_\ell=\varphi_+(\omega)>\sqrt{q}$. Hence, the integrand in (D.35) is bounded by $\varphi_+(\omega)^v\omega$, as $\frac{|u-x|}{\nu_i}\geq a\nu_\ell/\nu_u=\varphi_+(\omega)$:

$$(\text{D.35})\leq 2\max_{0\leq v\leq q}\varphi_+(\omega)^v\omega=2\varphi_+(\omega)^q\omega.$$

Note that

$$\varphi(t)=\sum_{k=0}^{\infty}\frac{(-t^2/2)^k}{\sqrt{2\pi}k!}=\sum_{k=0}^{k^*}\frac{(-t^2/2)^k}{\sqrt{2\pi}k!}+R(t)$$

Thus the second term (D.36) can be written as the maximum-over-v of the absolute value of

$$\sum_{k=0}^{k^*}\int\frac{\left(\frac{x-u}{\nu_i}\right)^{v+2k}(-1/2)^k}{\sqrt{2\pi}k!}[G(du)-G_m(du)]+\int R\left(\frac{x-u}{\nu_i}\right)\left(\frac{x-u}{\nu_i}\right)^v[G(du)-G_m(du)]$$

The first term in the line above is zero since the moments match up to $2k^*+q$. Therefore (D.36) is equal to

$$(\text{D.36})=\max_{0\leq v\leq q}\left|\int_{(I_{j-1}\cup I_j\cup I_{j+1}^{\mathrm{C}})}\left(\frac{u-x}{\nu_i}\right)^v R\left(\frac{x-u}{\nu_i}\right)(G(du)-G_m(du))\right|.$$

We know that since $\varphi(t)$ has alternating-signed Taylor expansion,

$$|R(t)|\leq\frac{(t^2/2)^{k^*+1}}{\sqrt{2\pi}(k^*+1)!}$$

We can bound $|\frac{u-x}{\nu_i}|\leq 2a\nu_\ell/\nu_i\leq 2a$. Hence the integral is upper bounded by

$$\begin{aligned}
(\text{D.36}) &\leq 2\cdot(2a)^q\cdot\frac{\left((2a)^2/2\right)^{k^*+1}}{\sqrt{2\pi}(k^*+1)!} && ((2a)^v\leq(2a)^q)\\
&\leq\frac{2(2a)^q}{(2\pi)\sqrt{k^*+1}}\left(\frac{2a^2}{k^*+1}e\right)^{k^*+1}\\
& && \text{(Recall Stirling's formula } (k^*+1)!\geq\sqrt{2\pi(k^*+1)}\left(\tfrac{k^*+1}{e}\right)^{k^*+1}.)\\
&\leq\frac{(2a)^q}{\pi\sqrt{k^*+1}}\left(\frac{e}{3}\right)^{k^*+1} && (\text{Choosing } k^*+1\geq 6a^2\geq 6)\\
&\leq\frac{(2a)^q}{\pi\sqrt{k^*+1}}\exp\left(-\frac{1}{2}\frac{k^*+1}{6}\right) && ((e/3)^6\leq e^{-1/2})\\
&\leq\frac{(2a)^q}{\sqrt{k^*+1}\sqrt{\pi/2}}\underbrace{\varphi(a\nu_\ell/\nu_u)}_{\varphi(\varphi_+(\omega))} && (k^*+1\geq 6a^2\geq 6(a\nu_\ell/\nu_u)^2)\\
&\leq\frac{(2a)^q}{\sqrt{k^*+1}\sqrt{\pi/2}}\omega\\
&\leq\frac{2^q}{\sqrt{3\pi}}\left(\frac{\nu_u}{\nu_\ell}\right)^{q-1}\varphi_+^{q-1}(\omega)\omega && (k^*+1\geq 6a^2)
\end{aligned}$$

This bounds (D.35) + (D.36) uniformly over $|x| \le M$. Therefore,

$$d_{q,i,M}(G, G_m) \le \left(2 + \frac{2^q}{\sqrt{3\pi}}(\nu_u/\nu_\ell)^{q-1}\right) \cdot \varphi_+^q(\omega)\omega \lesssim_{q,\nu_u,\nu_\ell} \log^{q/2}(1/\omega)\omega.$$

D.11.2. *Disciplining G_m onto a fixed grid.* Now, consider a gridding of G_m via $G_{m,\omega}$. We construct $G_{m,\omega}$ to be the following distribution. For a draw $\xi \sim G_m$, let $\tilde{\xi} = \omega\,\mathrm{sgn}(\xi)\lfloor |\xi|/\omega \rfloor$. We let $G_{m,\omega}$ be the distribution of $\tilde{\xi}$. $G_{m,\omega}$ has at most $m = (2k^* + q + 1)j^* + 1$ support points by construction, and all its support points are multiples of ω.

Since

$$\int g(x, u) G_{m,\omega}(du) = \int g(x, \omega\,\mathrm{sgn}(u)\lfloor |u|/\omega \rfloor)\, G_m(du)$$

we have that

$$\left| \int g(x, u) G_{m,\omega}(du) - \int g(x, u) G_m(du) \right| \le \int |g(x, \omega\,\mathrm{sgn}(u)\lfloor |u|/\omega \rfloor) - g(x, u)|\, G_m(du)$$

In the case of $g(x, u) = ((x - u)/\nu_i)^v\, \varphi((x - u)/\nu_i)$, this function is Lipschitz by Lemma D.8, we thus have that,

$$d_{q,i,M}(G_m, G_{m,\omega}) \le \int C_q \frac{\omega}{\nu_i} G_m(du) \lesssim_{\nu_\ell,q} \omega.$$

So far, we have shown that there exists a distribution with at most m support points, supported on the lattice points $\{j\omega : j \in \mathbb{Z}, |j\omega| \in I\}$, that approximates G up to

$$C_{q,\nu_u,\nu_\ell}\omega \log^{q/2}(1/\omega)$$

in $d_{\infty,M}^{(q)}(\cdot, \cdot)$.

D.11.3. *Covering the set of $G_{m,\omega}$.* Let Δ^{m-1} be the $(m-1)$-simplex of probability vectors in m dimensions. Consider discrete distributions supported on the support points of $G_{m,\omega}$, which can be identified with a subset of Δ^{m-1}. Thus, there are at most $N(\omega, \Delta^{m-1}, \|\cdot\|_1)$ such distributions that form an ω-net in $\|\cdot\|_1$. Now, consider a distribution $G'_{m,\omega}$ where

$$\|G'_{m,\omega} - G_{m,\omega}\|_1 \le \omega.$$

Since $t^q\varphi(t)$ is bounded, we have that

$$\|G'_{m,\omega} - G_{m,\omega}\|_{q,i,M} \le \omega \max_{0 \le v \le q} v^{v/2}\varphi(\sqrt{v}) \lesssim_q \omega$$

by Lemma D.8.

There are at most

$$\binom{1 + 2\lfloor (M + a\nu_\ell)/\omega \rfloor}{m}$$

configurations of m support points. Hence there are a collection of at most

$$\binom{1 + 2\lfloor (M + a\nu_\ell)/\omega \rfloor}{m} N(\omega, \Delta^{m-1}, \|\cdot\|_1)$$

distributions $\mathcal{G}$ where

$$\min_{H \in \mathcal{G}} \|G - H\|_{q,\infty,M} \le \underbrace{C_{q,\nu_u,\nu_\ell} \log(1/\omega)^{q/2}\omega}_{\omega^*}.$$

D.11.4. *Putting things together.* In other words,

$$N(\omega^*, \mathcal{P}(\mathbb{R}), \|\cdot\|_{q,\infty,M}) \le \binom{1 + 2\lfloor (M + a\nu_\ell)/\omega \rfloor}{m} N(\omega, \Delta^{m-1}, \|\cdot\|_1)$$
$$\le \left(\frac{(\omega+2)(\omega+2(M+a\nu_\ell))e}{m}\right)^m \omega^{-2m}(2\pi m)^{-1/2}$$

((6.24) in Jiang (2020))

Since $\omega < 1$ and $m \ge 2\frac{12a^2+3+q}{a\nu_\ell}(M + a\nu_\ell)$, the first term is of the form C^m:

$$\frac{(\omega+2)(\omega+2(M+a\nu_\ell))e}{m} \le \frac{3e}{m}(1 + 2(M + a\nu_\ell)) \lesssim \frac{a\nu_\ell}{12a^2+3+q} \lesssim \nu_\ell.$$

Therefore

$$\log N(\omega^*, \mathcal{P}(\mathbb{R}), \|\cdot\|_{q,\infty,M}) \lesssim m \cdot |\log(1/\omega)| + m|\log \nu_\ell| \lesssim_{\nu_\ell,\nu_u,q} m \log(1/\omega).$$

Finally, since $m = (2k^* + q + 1)j^* + 1$. Recall that we have required $k^* + 1 \ge 6a^2$, and it suffices to pick $k^* = \lceil 6a^2 \rceil$. Then

$$m \lesssim_{q,\nu_u,\nu_\ell} \log(1/\omega) \max\left(\frac{M}{\sqrt{\log(1/\omega)}}, 1\right).$$

Hence

$$\log N(\omega^*, \mathcal{P}(\mathbb{R}), \|\cdot\|_{q,\infty,M}) \lesssim_{q,\nu_u,\nu_\ell} \log(1/\omega)^2 \max\left(\frac{M}{\sqrt{\log(1/\omega)}}, 1\right).$$

Lastly, let K equal the constant in $\omega^* = K \log(1/\omega)^{q/2}\omega$. Note that we can take $K \ge 1$. For some $c > 1$ such that $\log(cK)^{q/2} < c$, we plug in $\omega = \frac{\delta}{cK}$ such that whenever $\delta < cK(\varphi(1) \wedge \varphi(\sqrt{q})) \wedge e^{-q/2}$, the covering number bound holds for

$$\omega^* = \frac{\delta}{c}\log(cK/\delta)^{q/2} \le \delta \log(1/\delta)^{q/2}.$$

In this case,

$$N\left(\delta \log(1/\delta)^{q/2}, \mathcal{P}(\mathbb{R}), \|\cdot\|_{q,\infty,M}\right) \le N\left(\omega^* \log(1/\delta)^{q/2}, \mathcal{P}(\mathbb{R}), \|\cdot\|_{q,\infty,M}\right)$$
$$\lesssim_{q,\nu_u,\nu_\ell} \log(1/\omega)^2 \max\left(\frac{M}{\sqrt{\log(1/\omega)}}, 1\right)$$
$$\lesssim_{q,\nu_u,\nu_\ell} \log(1/\delta)^2 \max\left(\frac{M}{\sqrt{\log(1/\delta)}}, 1\right)$$

This bound holds for all sufficiently small δ. Since $\delta \log(1/\delta)^{q/2}$ is increasing over $(0, e^{-q/2} \wedge e^{-1})$ and the right-hand side does not vanish over the interval, we can absorb larger δ's into the constant.

□

As a consequence, we can control the covering number in terms of $d_{k,\infty,M}$ for $k \in \{m, s\}$

Proposition D.2. *Consider $d^{(q)}_{\infty,M}$ in Proposition D.1, $d_{s,\infty,M}$ in (D.20), and $d_{m,\infty,M}$ in (D.17) for some $M > 1$. Then*

$$d^{(2)}_{\infty,M}(H_1, H_2) \le \delta \implies d_{s,\infty,M}(H_1, H_2) \lesssim_{\mathcal{H}} \frac{M\sqrt{\log(1/\rho_n)} + \log(1/\rho_n)}{\rho_n}\delta.$$

and

$$d^{(2)}_{\infty,M}(H_1, H_2) \le \delta \implies d_{m,\infty,M}(H_1, H_2) \lesssim_{\mathcal{H}} \frac{\sqrt{\log(1/\rho_n)}}{\rho_n}\delta.$$

As a corollary, for all $\delta \in (0, 1/e)$,

$$\log N\left(\frac{\delta\log(1/\delta)}{\rho_n}\sqrt{\log(1/\rho_n)}, \mathcal{P}(\mathbb{R}), d_{m,\infty,M}\right) \lesssim_{\mathcal{H}} \log(1/\delta)^2 \max\left(1, \frac{M}{\sqrt{\log(1/\delta)}}\right)$$

$$\log N\left(\frac{\delta\log(1/\delta)}{\rho_n}\left(M\sqrt{\log(1/\rho_n)} + \log(1/\rho_n)\right), \mathcal{P}(\mathbb{R}), d_{s,\infty,M}\right) \lesssim_{\mathcal{H}} \log(1/\delta)^2 \max\left(1, \frac{M}{\sqrt{\log(1/\delta)}}\right).$$

Proof. Recall that

$$D_{s,i}(z_i, G, \eta_0, \rho_n) = \frac{s_i}{\sigma_i^2}\frac{Q_i(Z_i, \eta_0, G)}{f_{i,G} \vee \frac{\rho_n}{\nu_i}}.$$

Hence

$$\begin{aligned}
&|D_{s,i}(z, G_1, \eta_0, \rho_n) - D_{s,i}(z, G_2, \eta_0, \rho_n)| \\
&\lesssim_{\mathcal{H}} \frac{1}{f_{i,G_1} \vee \frac{\rho_n}{\nu_i}}|Q_i(Z_i, \eta_0, G_1) - Q_i(Z_i, \eta_0, G_2)| + |Q_i(Z_i, \eta_0, G_2)|\left|\frac{1}{f_{i,G_1} \vee \frac{\rho_n}{\nu_i}} - \frac{1}{f_{i,G_1} \vee \frac{\rho_n}{\nu_i}}\right| \\
&\lesssim_{\mathcal{H}} \frac{1}{\rho_n}|f_{i,G_1}\mathbf{E}_{G_1,\nu_i}[(Z-\tau)\tau \mid z] - f_{i,G_2}\mathbf{E}_{G_2,\nu_i}[(Z-\tau)\tau \mid z]| \\
&\quad + \frac{M\sqrt{\log(1/\rho_n)} + \log(1/\rho_n)}{\rho_n}|f_{i,G_1} - f_{i,G_2}|
\end{aligned}$$

where the last inequality follows from the definition of Q_i and Lemma D.13.

Note that

$$f_{i,G_1}\mathbf{E}_{G_1,\nu_i}[(Z-\tau)\tau \mid z] = f_{i,G_1}\mathbf{E}_{G_1,\nu_i}[(Z-\tau)^2 \mid z] - zf_{i,G_1}\mathbf{E}_{G_1,\nu_i}[(Z-\tau) \mid z].$$

Thus we can further upper bound, by the bound on $d^{(2)}_{\infty,M}$,

$$|\mathbf{E}_{G_1,\nu_i}[(Z-\tau)\tau \mid z] - \mathbf{E}_{G_2,\nu_i}[(Z-\tau)\tau \mid z]| \lesssim_{\mathcal{H}} \delta(1+M) \lesssim M\delta.$$

Similarly, $|f_{i,G_1} - f_{i,G_2}| \lesssim_{\mathcal{H}} \delta$. Hence,

$$\begin{aligned}
|D_{s,i}(z, G_1, \eta_0, \rho_n) - D_{s,i}(z, G_2, \eta_0, \rho_n)| &\lesssim_{\mathcal{H}} \left\{\frac{M}{\rho_n} + \rho_n^{-1}\left(M\sqrt{\log(1/\rho_n)} + \log(1/\rho_n)\right)\right\}\delta \\
&\lesssim_{\mathcal{H}} \frac{M\sqrt{\log(1/\rho_n)} + \log(1/\rho_n)}{\rho_n}\delta.
\end{aligned}$$

Similarly,

$$D_{m,i}(z, G, \eta_0) = \frac{s_i}{\sigma_i^2}\frac{f_{i,G}\mathbf{E}_{G,\nu_i}[(Z-\tau) \mid z]}{f_{i,G} \vee \rho_n/\nu_i}.$$

Therefore

$$|D_{m,i}(z, G_1, \eta_0) - D_{m,i}(z, G_2, \eta_0)| \lesssim_{\mathcal{H}} \frac{1}{\rho_n}\delta + \frac{1}{\rho_n}\sqrt{\log(1/\rho_n)}\delta \lesssim \frac{1}{\rho_n}\sqrt{\log(1/\rho_n)}$$

by a similar calculation, involving Lemma D.10.

Thus, for the "corollary" part, note that, letting $C_{\mathcal{H}}$ be the constant in the bound, taken to be at least 1:

$$N\left(\frac{\delta \log(1/\delta)}{\rho_n}\sqrt{\log(1/\rho_n)}, \mathcal{P}(\mathbb{R}), d_{m,\infty,M}\right) \le N\left(\frac{\delta}{C_{\mathcal{H}}}\log(1/(\delta/(C_{\mathcal{H}}))), \mathcal{P}(\mathbb{R}), d^{(2)}_{\infty,M}\right)$$
$$\lesssim_{\mathcal{H}} \log(1/\delta)^2 \max\left(1, \frac{M}{\sqrt{\log(1/\delta)}}\right).$$

for all $0 < \delta < 1/e$. Similarly for the covering number in $d_{s,\infty,M}$. □

D.12. **Auxiliary lemmas.**

Lemma D.9. *Suppose $|\overline{Z}_n| = \max_{i\in[n]} |Z_i| \vee 1 \le M_n$, $\|\hat{s} - s_0\|_\infty \le \Delta_n$, and $\|\hat{m} - m_0\|_\infty \le \Delta_n$. Let $\hat{G}_n$ satisfy Assumption 1. Then, under Assumption D.1,*

(1) $|\hat{Z}_i \vee 1| \lesssim_{\mathcal{H}} M_n$

(2) There exists $C_{\mathcal{H}}$ such that with $\rho_n = \frac{1}{n^3}\exp\left(-C_{\mathcal{H}} M_n^2 \Delta_n\right) \wedge \frac{1}{e\sqrt{2\pi}}$,

$$f_{\hat{G}_n,\nu_i}(Z_i) \ge \frac{\rho_n}{\nu_i}.$$

(3) The choice of ρ_n satisfies $\log(1/\rho_n) \asymp_{\mathcal{H}} \log n$, $\varphi_+(\rho_n) \asymp_{\mathcal{H}} \sqrt{\log n}$, and $\rho_n \lesssim_{\mathcal{H}} n^{-3}$.

Proof. Observe that $|\hat{Z}_i| \vee 1 \lesssim_{\sigma_\ell,\sigma_u,s_{0\ell},s_{0u}} (1 + \Delta_n)M_n + \Delta_n \lesssim (1 + \Delta_n)M_n$ by Lemma D.11(3). Hence by Assumption D.1, $|\hat{Z}_i| \vee 1 \lesssim_{\mathcal{H}} M_n$.

For (2), we note by Theorem 5 in Jiang (2020),

$$f_{\hat{G}_n,\hat{\nu}_i}(\hat{Z}_i) \ge \frac{1}{n^3 \hat{\nu}_i}$$

thanks to κ_n in (3.3). That is,

$$\int \varphi\left(\frac{\hat{Z}_i - \tau}{\hat{\nu}_i}\right) \hat{G}_n(d\tau) \ge \frac{1}{n^3}.$$

Now, note that

$$\frac{\hat{Z}_i - \tau}{\hat{\nu}_i} = \frac{Z_i + \frac{m_{0i} - \hat{m}_i}{s_{0i}} + \left(1 - \frac{\hat{s}_i}{s_{0i}}\right)\tau - \tau}{\nu_i} = \frac{Z_i - \tau}{\nu_i} + \frac{m_{0i} - \hat{m}_i}{\sigma_i} + \frac{1}{\sigma_i}(s_i - s_{0i})\tau = \frac{Z_i - \tau}{\nu_i} + \xi(\tau) \tag{D.37}$$

where $|\xi(\tau)| \lesssim_{\mathcal{H}} \Delta_n M_n$ over the support of τ under $\hat{G}_n$, under our assumptions.

Then, for all Z_i, since $|Z_i| \le M_n$ by assumption,

$$\varphi\left(\frac{\hat{Z}_i - \tau}{\hat{\nu}_i}\right) = \varphi\left(\frac{Z_i - \tau}{\nu_i}\right)\exp\left(-\frac{1}{2}\xi^2(\tau) - \xi(\tau)\frac{Z_i - \tau}{\nu_i}\right)$$
$$\le \varphi\left(\frac{Z_i - \tau}{\nu_i}\right)\exp\left(C_{\mathcal{H}}\Delta_n M_n \left|\frac{Z_i - \tau}{\nu_i}\right|\right)$$

$$\leq \varphi\left(\frac{Z_i-\tau}{\nu_i}\right)\exp\left(C_{\mathcal{H}}\Delta_n M_n^2\right). \qquad \left(\left|\tfrac{Z_i-\tau}{\nu_i}\right| \lesssim_{\mathcal{H}} M_n\right)$$

Therefore,

$$\int \varphi\left(\frac{Z_i-\tau}{\nu_i}\right)\hat{G}_n(d\tau) \geq \frac{1}{n^3}e^{-C_{\mathcal{H}}\Delta_n M_n^2}.$$

Dividing by ν_i on both sides finishes the proof of (2). Claim (3) is immediate by calculating $\log(1/\rho_n) = \left(3\log n - C_{\mathcal{H}}M_n^2\Delta_n^2\right) \vee \log(e\sqrt{2\pi}) \lesssim_{\mathcal{H}} \log n$ and apply Assumption D.1(1) to obtain that $\Delta_n M_n^2 \lesssim_{\mathcal{H}} 1$. □

Lemma D.10 (Lemma 2 Jiang (2020))**.** *For all $x \in \mathbb{R}$ and all $\rho \in (0, 1/\sqrt{2\pi e})$,*

$$\left|\frac{\nu^2 f'_{H,\nu}(x)}{(\rho/\nu) \vee f_{H,\nu}(x)}\right| \leq \nu\varphi_+(\rho).$$

Moreover, for all $x \in \mathbb{R}$ and all $\rho \in (0, e^{-1}/\sqrt{2\pi})$,

$$\left|\left(\frac{\nu^2 f''_{H,\nu}(x)}{f_{H,\nu}(z)}+1\right)\left(\frac{\nu f_{H,\nu}(x)}{(\nu f_{G,\nu}(x))\vee\rho}\right)\right| \leq \varphi_+^2(\rho),$$

where we recall φ_+ from (C.3).

Proof. The first claim is immediate from Lemma 2 in Jiang (2020). The second claim follows from parts of the proof. Lemma 1 in Jiang (2020) shows that

$$0 \leq \frac{\nu^2 f''_{H,\nu}(x)}{f_{H,\nu}(z)} + 1 \leq \underbrace{\log\frac{1}{2\pi\nu^2 f_{H,\nu}(z)^2}}_{\varphi_+^2(\nu f_{H,\nu}(z))}.$$

Case 1 ($\nu f_{H,\nu}(x) \leq \rho < e^{-1}/\sqrt{2\pi}$): Observe that $t\log\frac{1}{2\pi t^2}$ is increasing over $t \in (0, e^{-1}(2\pi)^{-1/2})$. Hence,

$$\left(\frac{\nu^2 f''_{H,\nu}(x)}{f_{H,\nu}(z)}+1\right)\nu f_{H,\nu}(x) \leq \nu f_{H,\nu}\log\frac{1}{2\pi\nu^2 f_{H,\nu}(z)^2} \leq \rho\log\frac{1}{2\pi\rho^2}.$$

Dividing by $(\nu f)\vee\rho = \rho$ confirms the bound for $\nu f < \rho$.

Case 2 ($\nu f > \rho$): Since $\log\frac{1}{2\pi t^2}$ is decreasing in t, we have that

$$\left|\left(\frac{\nu^2 f''_{H,\nu}(x)}{f_{H,\nu}(z)}+1\right)\left(\frac{\nu f_{H,\nu}(x)}{(\nu f_{G,\nu}(x))\vee\rho}\right)\right| = \frac{\nu^2 f''_{H,\nu}(x)}{f_{H,\nu}(z)}+1 \leq \varphi_+^2(\nu f_{H,\nu}) \leq \log\frac{1}{2\pi\rho^2}.$$

□

Lemma D.11. *The following statements are true:*

1. *Under Assumption 4, $1/\hat{\nu}_i \lesssim_{s_{0u},\sigma_\ell} 1$ and $\hat{\nu}_i \lesssim_{s_{0\ell},\sigma_u} 1$*
2. *Under Assumption 4, $|1-\frac{s_{0i}}{\hat{s}_i}| \lesssim_{s_{0\ell}} \|\hat{s}-s_0\|_\infty$*
3. *Under Assumption 4,*

$$\max_i |\hat{Z}_i| \lesssim_{\sigma_\ell,\sigma_u,s_{0\ell},s_{0u}} (1+\|\hat{s}-s_0\|_\infty)\overline{Z}_n + \|\hat{m}-m_0\|_\infty$$

where $\overline{Z}_n$ is defined in (C.5).

Proof. (1) Immediate by $1/\hat{\nu}_i = \hat{s}_i/\sigma_i$ and $\mathrm{P}[s_{0\ell} < \hat{s}_i < s_{0u}] = 1$.

(2) Immediate by observing that $|1 - \frac{s_{0i}}{\hat{s}_i}| = |\frac{\hat{s}_i - s_{0i}}{\hat{s}_i}|$ and $\mathrm{P}[s_{0\ell} < \hat{s}_i < s_{0u}] = 1$.

(3) Immediate by $\hat{Z}_i = \frac{s_{0i}}{\hat{s}_i} Z_i + [m_{0i} - \hat{m}_i]$

□

Lemma D.12 (Zhang (1997), p.186)**.** *Let f be a density and let $\sigma(f)$ be its standard deviation. Then, for any $M, t > 0$,*

$$\int_{-\infty}^{\infty} \mathbb{1}(f(z) \le t) f(z)\, dz \le \frac{\sigma(f)^2}{M^2} + 2Mt.$$

In particular, choosing $M = t^{-1/3}\sigma(f)^{2/3}$ gives

$$\int_{-\infty}^{\infty} \mathbb{1}(f(z) \le t) f(z)\, dz \le 3t^{2/3}\sigma^{2/3}.$$

Proof. Since the value of the integral does not change if we shift $f(z)$ to $f(z-c)$, it is without loss of generality to assume that $\mathbb{E}_f[Z] = 0$.

$$\begin{aligned}
\int_{-\infty}^{\infty} \mathbb{1}(f(z) \le t) f(z)\, dz &\le \int_{-\infty}^{\infty} \mathbb{1}(f(z) \le t, |z| < M) f(z)\, dz + \int_{-\infty}^{\infty} \mathbb{1}(f(z) \le t, |z| > M) f(z)\, dz \\
&\le \int_{-M}^{M} t\, dz + \mathrm{P}(|Z| > M) \\
&\le 2Mt + \frac{\sigma^2(f)}{M^2}. && \text{(Chebyshev's inequality)}
\end{aligned}$$

□

Lemma D.13. *Recall that $Q_i(z, \eta, G) = \int (z-\tau)\tau\varphi\left(\frac{z-\tau}{\nu_i(\eta)}\right)\frac{1}{\nu_i(\eta)}\, G(d\tau)$. Then, for any G, z and $\rho_n \in (0, e^{-1}/\sqrt{2\pi})$,*

$$\left| \frac{Q_i(z, \eta_0, G)}{f_{G,\nu_i}(z) \vee (\rho_n/\nu_i)} \right| \le \varphi_+(\rho_n) \left(\nu_i |z| + \nu_i \varphi_+(\rho_n)\right). \tag{D.38}$$

Proof. We can write

$$Q_i(z, \eta_0, G) = f_{G,\nu_i}(z) \left\{ z\mathbf{E}_{G,\nu_i}[(z-\tau) \mid z] - \mathbf{E}_{\hat{G}_n,\nu_i}[(z-\tau)^2 \mid z] \right\}.$$

From Lemma D.10,

$$\frac{f_{G,\nu_i}(z)}{f_{G,\nu_i}(z) \vee (\rho_n/\nu_i)} \mathbf{E}_{G,\nu_i}[(z-\tau) \mid z] \le \nu_i \varphi_+(\rho_n)$$

and

$$\frac{f_{G,\nu_i}(z)}{f_{G,\nu_i}(z) \vee (\rho_n/\nu_i)} \mathbf{E}_{G,\nu_i}[(z-\tau)^2 \mid z] = \nu_i^2 \left(\frac{\nu_i^2 f''_{i,G}}{f_{i,G}} + 1 \right) \frac{f_{G,\nu_i}(z)}{f_{G,\nu_i}(z) \vee (\rho_n/\nu_i)} \le \nu_i^2 \varphi_+^2(\rho_n).$$

Therefore,

$$\left| \frac{Q_i(z, \eta_0, G)}{f_{G,\nu_i}(z) \vee (\rho_n/\nu_i)} \right| \le \varphi_+(\rho_n)\nu_i \left(|z| + \varphi_+(\rho_n)\right).$$

□

Lemma D.14. *Under the assumptions in Lemma D.9, suppose $\tilde{\eta}_i$ lies on the line segment between η_0 and $\hat{\eta}_i$ and define $\tilde{\nu}_i, \tilde{m}_i, \tilde{s}_i, \tilde{Z}_i$ accordingly. Then, the second derivatives (D.29), (D.31), (D.32),* *evaluated at $\tilde{\eta}_i, \hat{G}_n, \tilde{Z}_i$, satisfy*

$$
\begin{aligned}
|(\text{D.29})| &\lesssim_{\mathcal{H}} \log n \\
|(\text{D.31})| &\lesssim_{\mathcal{H}} M_n \log n \\
|(\text{D.32})| &\lesssim_{\mathcal{H}} M_n^2 \log n.
\end{aligned}
$$

Proof. First, we show that

$$
|\log(f_{\hat{G}_n,\tilde{\nu}_i}(\tilde{Z}_i)\tilde{\nu}_i)| \lesssim_{\mathcal{H}} \log n. \tag{D.39}
$$

Observe that we can write

$$
\hat{Z}_i = \frac{\tilde{s}_i\tilde{Z}_i + \tilde{m}_i - \hat{m}_i}{\hat{s}_i}.
$$

where $\|\tilde{s} - \hat{s}\|_\infty \le \Delta_n$ and $\|\tilde{m} - \hat{m}\|_\infty \le \Delta_n$. This also shows that $|\tilde{Z}_i| \lesssim_{\mathcal{H}} M_n$ under the assumptions.

Note that by the same argument in (D.37) in Lemma D.9, we have that

$$
\varphi\left(\frac{\hat{Z}_i - \tau}{\hat{\nu}_i}\right) \le \varphi\left(\frac{\tilde{Z}_i - \tau}{\tilde{\nu}_i}\right) e^{-C_{\mathcal{H}}\Delta_n M_n^2}.
$$

Hence,

$$
\tilde{\nu}_i f_{\hat{G}_{(i)},\tilde{\nu}_i}(\tilde{Z}_i) \ge \frac{1}{n^3} e^{-C_{\mathcal{H}}\Delta_n M_n^2}.
$$

This shows (D.39).

Now, observe that

$$
\mathbf{E}_{\hat{G}_n,\tilde{\nu}}[(\tau - Z)^2 \mid \tilde{Z}_i] \lesssim_{\mathcal{H}} \log\left(\frac{1}{\tilde{\nu}_i f_{\hat{G}_{(i)},\tilde{\nu}_i}(\tilde{Z}_i)}\right) \lesssim_{\mathcal{H}} \log n
$$

and

$$
\mathbf{E}_{\hat{G}_n,\tilde{\nu}}[|\tau - Z| \mid \tilde{Z}_i] \lesssim_{\mathcal{H}} \sqrt{\log\left(\frac{1}{\tilde{\nu}_i f_{\hat{G}_{(i)},\tilde{\nu}_i}(\tilde{Z}_i)}\right)} \lesssim_{\mathcal{H}} \sqrt{\log n}
$$

by Lemma D.10, since we can always choose $\rho = \tilde{\nu}_i f_{\hat{G}_{(i)},\tilde{\nu}_i}(\tilde{Z}_i) \wedge \frac{1}{\sqrt{2\pi e}}$. Similarly, by Lemma D.13, and plugging in $\rho = \tilde{\nu}_i f_{\hat{G}_{(i)},\tilde{\nu}_i}(\tilde{Z}_i) \wedge \frac{1}{\sqrt{2\pi e}}$,

$$
\left|\mathbf{E}_{\hat{G}_n,\tilde{\nu}}[(\tau - Z)Z \mid \tilde{Z}_i]\right| \lesssim_{\mathcal{H}} \sqrt{\log n}|\tilde{Z}_i| + \log n \lesssim_{\mathcal{H}} \sqrt{\log n} M_n.
$$

Observe that

$$
\left|\mathbf{E}_{\hat{G}_n,\tilde{\nu}_i}[(\tau - Z)^2\tau \mid \tilde{Z}_i]\right| \lesssim_{\mathcal{H}} M_n \mathbf{E}_{\hat{G}_n,\tilde{\nu}_i}[(\tau - Z)^2] \lesssim_{\mathcal{H}} M_n \log n.
$$

since $|\tau| \lesssim_{\mathcal{H}} M_n$ under $\hat{G}_n$. Similarly,

$$
\mathbf{E}_{\hat{G}_n,\tilde{\nu}_i}[(Z - \tau)^2\tau^2 \mid \tilde{Z}_i] \lesssim_{\mathcal{H}} M_n^2 \log n \quad \mathbf{E}_{\hat{G}_n,\tilde{\nu}_i}[\tau^2 \mid \tilde{Z}_i] \lesssim_{\mathcal{H}} M_n^2.
$$

Plugging these intermediate results into (D.29), (D.31), (D.32) proves the claim. □

Lemma D.15. *Let $X_1, \ldots, X_J$ be subgaussian random variables with $K = \max_i \|X_i\|_{\psi_2}$, not necessarily independent. Then for some universal C, for all $t \geq 0$,*

$$\mathrm{P}\left[\max_i |X_i| \geq CK\sqrt{\log J} + CKt\right] \leq 2e^{-t^2}.$$

Proof. By (2.14) in Vershynin (2018), $\mathrm{P}(|X_i| > t) \leq 2e^{-ct^2/\|X_i\|_{\psi_2}^2} \leq 2e^{-ct^2/K}$ for some universal c. By a union bound,

$$\mathrm{P}\left[\max_i |X_i| \geq Ku\right] \leq 2\exp\left(-cu^2 + \log J\right)$$

Choose $u = \frac{1}{\sqrt{c}}(\sqrt{\log J} + t)$ so that $cu^2 = \log J + t^2 + 2t\sqrt{\log J} \geq \log J + t^2$. Hence

$$2\exp\left(-cu^2 + \log J\right) \leq 2e^{-t^2}.$$

Implicitly, $C = 1/\sqrt{c}$. □

Lemma D.16. *Suppose Z has simultaneous moment control $\mathbb{E}[|Z|^p]^{1/p} \leq Ap^{1/\alpha}$. Then*

$$\mathrm{P}(|Z| > M) \leq \exp\left(-C_{A,\alpha}M^\alpha\right).$$

As a corollary, suppose $Z \sim f_{G_0,\nu_i}(\cdot)$ and G_0 obeys Assumption 2, then

$$\mathrm{P}(|Z| > M) \leq \exp\left(-C_{A_0,\alpha,\nu_u}M^\alpha\right).$$

Proof. Observe that

$$\mathrm{P}(|Z| > M) = \mathrm{P}(|Z|^p > M^p) \leq \left\{\frac{Ap^{1/\alpha}}{M}\right\}^p. \qquad \text{(Markov)}$$

Choose $p = (M/(eA))^\alpha$ such that

$$\left\{\frac{Ap^{1/\alpha}}{M}\right\}^p = \exp(-p) = \exp\left(-\left(\frac{1}{eA}\right)^\alpha M^\alpha\right).$$

□

Lemma D.17. *Let E be some event and assume that*

$$\mathrm{P}(E, A > a) \leq p_1 \quad \mathrm{P}(E, B > b) \leq p_2$$

Then $\mathrm{P}(E, A + B > a + b) \leq p_1 + p_2$

Proof. Note that $A + B > a + b$ implies that one of $A > a$ and $B > b$ occurs. Hence

$$\mathrm{P}(E, A + B > a + b) \leq \mathrm{P}(\{E, A > a\} \cup \{E, B > b\}) \leq p_1 + p_2$$

by union bound. □

Lemma D.18. *Let $\tau \sim G_0$ where G_0 satisfies Assumption 2. Let $Z \mid \tau \sim \mathcal{N}(\tau, \nu^2)$. Then the posterior moment is bounded by a power of $|z|$:*

$$\mathbb{E}[|\tau|^p \mid Z = z] \lesssim_p (|z| \vee 1)^p$$

Proof. Let $M \geq |z| \vee 2$. We write

$$\mathbb{E}[|\tau|^p \mid Z = z] = \frac{1}{f_{G_0,\nu}(z)} \int |\tau|^p \varphi\left(\frac{z-\tau}{\nu}\right) \frac{1}{\nu} G_0(d\tau).$$

Note that

$$\begin{aligned}
\int |\tau|^p \varphi\left(\frac{z-\tau}{\nu}\right) \frac{1}{\nu} G_0(d\tau) &\leq (3M)^p f_{G_0,\nu}(z) + \int \mathbb{1}(|\tau| > 3M)|\tau|^p \varphi\left(\frac{z-\tau}{\nu}\right) \frac{1}{\nu} G_0(d\tau) \\
&\leq (3M)^p f_{G_0,\nu}(z) + \int_{|\tau|>3M} |\tau|^p G_0(d\tau) \cdot \frac{1}{\nu} \varphi\left(|2M|/\nu\right) \\
&\qquad\qquad (|z-\tau| \geq 2M \text{ when } |\tau| > 3M)
\end{aligned}$$

Also note that

$$f_{G_0,\nu}(z) = \int \varphi\left(\frac{z-\tau}{\nu}\right) \frac{1}{\nu} G_0(d\tau) \geq \frac{1}{\nu} \varphi\left(|2M|/\nu\right) G_0([-M, M]) \quad (|z-\tau| \leq 2M \text{ if } \tau \in [-M, M])$$

Hence,

$$\mathbb{E}[|\tau|^p \mid Z = z] \leq (3M)^p + \frac{\int |\tau|^p G_0(d\tau)}{G_0([-M, M])}$$

Since G_0 is mean zero and variance 1, by Chebyshev's inequality, $G_0([-M, M]) \geq G_0([-2, 2]) \geq 3/4$.

Hence

$$\mathbb{E}[|\tau|^p \mid Z = z] \lesssim_p M^p \lesssim_p (|z| \vee 1)^p,$$

since we have simultaneous moment control by Assumption 2. □

Appendix E. A large-deviation inequality for the average Hellinger distance

Theorem E.1. *For some $n > \sqrt{2\pi}e$, let $\tau_1, \ldots, \tau_n \mid (\nu_1^2, \ldots, \nu_n^2) \overset{\text{i.i.d.}}{\sim} G_0$ where G_0 satisfies Assumption 2. Let $\nu_u = \max_i \nu_i$ and $\nu_\ell = \min_i \nu_i$. Assume $Z_i \mid \tau_i, \nu_i^2 \sim \mathcal{N}(\tau_i, \nu_i^2)$. Fix positive sequences $\gamma_n, \lambda_n \to 0$ with $\gamma_n, \lambda_n \leq 1$ and constant $\epsilon > 0$. Fix some positive constant C^*. Consider the set of distributions that approximately maximize the likelihood*

$$A(\gamma_n, \lambda_n) = \left\{H : \mathrm{Sub}_n(H) \leq C^* \left(\gamma_n^2 + \overline{h}(f_{H,\cdot}, f_{G_0,\cdot})\lambda_n\right)\right\}.$$

Also consider the set of distributions that are far from G_0 in $\overline{h}$:

$$B(t, \lambda_n, \epsilon) = \left\{H : \overline{h}(f_{H,\cdot}, f_{G_0,\cdot}) \geq tB\lambda_n^{1-\epsilon}\right\}$$

with some constant B to be chosen. Assume that for some C_λ,

$$\lambda_n^2 \geq \left(\frac{C_\lambda}{n}(\log n)^{1+\frac{\alpha+2}{2\alpha}}\right) \vee \gamma_n^2.$$

Then the probability that $A \cap B$ is nonempty is bounded for $t > 1$: There exists a choice of B that depends only on $\nu_\ell, \nu_u, C^*, C_\lambda$ such that

$$\mathrm{P}\left[A(\gamma_n, \lambda_n) \cap B(t, \lambda_n, \epsilon) \neq \varnothing\right] \leq (\log_2(1/\epsilon) + 1)n^{-t^2}. \tag{E.1}$$

Corollary E.1. *Let* $\lambda_n = n^{-\frac{p}{2p+1}}(\log n)^{\gamma_1} \wedge 1$ *and* $\gamma_n = n^{-\frac{p}{2p+1}}(\log n)^{\gamma_2} \wedge 1$ *where* $\gamma_1 \geq \gamma_2 > 0$*. Fix some* $C^*_{\mathcal{H}}$*. Fix* $\epsilon > 0$*. Then there exists* $B_{\mathcal{H}}$ *that depends solely on* $C^*_{\mathcal{H}}, p, \gamma_1, \gamma_2, \nu_\ell, \nu_u$ *such that*

$$\mathrm{P}\left[\text{There exists } H\text{: } \mathrm{Sub}_n(H) \leq C^*_{\mathcal{H}}(\gamma_n^2 + \overline{h}(f_{H,\cdot}, f_{G_0,\cdot})\lambda_n) \text{ and } \overline{h}(f_{H,\cdot}, f_{G_0,\cdot}) \geq tB_{\mathcal{H}} n^{-\frac{p}{2p+1}}(\log n)^{\gamma_1}\right]$$
$$\leq \left(\frac{\log\log n}{\log 2} + 1\right) n^{-t^2}$$

Proof. First, note that $\lambda_n^2 \geq \gamma_n^2$ and $\lambda_n^2 \gtrsim \frac{(\log n)^{1+\frac{\alpha+2}{2\alpha}}}{n}$.

Note that $tB\lambda_n^{1-\epsilon} \leq tB_{\mathcal{H}} n^{-\frac{p}{2p+1}+\epsilon\frac{p}{2p+1}}(\log n)^{\gamma_1} \leq tB_{\mathcal{H}} n^{-\frac{p}{2p+1}+\epsilon}(\log n)^{\gamma_1}$. Therefore,

$$\left\{H : \overline{h}(f_{H,\cdot}, f_{G_0,\cdot}) \geq tB_{\mathcal{H}} n^{-\frac{p}{2p+1}+\epsilon}(\log n)^{\gamma_1}\right\} \subset \left\{H : \overline{h}(f_{H,\cdot}, f_{G_0,\cdot}) \geq tB\lambda_n^{1-\epsilon}\right\}.$$

As a result, the probability

$$\mathrm{P}\left[\text{There exists } H\text{: } \mathrm{Sub}_n(H) \leq C^*_{\mathcal{H}}(\gamma_n^2 + \overline{h}(f_{H,\cdot}, f_{G_0,\cdot})\lambda_n) \text{ and } \overline{h}(f_{H,\cdot}, f_{G_0,\cdot}) \geq tB_{\mathcal{H}} n^{-\frac{p}{2p+1}+\epsilon}(\log n)^{\gamma_1}\right]$$

is upper bounded by

$$\mathrm{P}\left[A(\gamma_n, \lambda_n) \cap B(t, \lambda_n, \epsilon) \neq \varnothing\right] \leq (\log_2(1/\epsilon) + 1)n^{-t^2}$$

via an application of Theorem E.1.

Finally, set $\epsilon = \frac{1}{\log n}$. Note that $n^\epsilon = n^{\frac{1}{\log n}} = \exp(\log n / \log n) = e$. Hence

$$tB_{\mathcal{H}} n^{-\frac{p}{2p+1}+\epsilon}(\log n)^{\gamma_1} = tB_{\mathcal{H}} e n^{-\frac{p}{2p+1}}(\log n)^{\gamma_1}. \qquad \square$$

Corollary E.2. *Assume the conditions in Corollary D.1. That is,*

(1) The estimate $\hat{G}_n$ *satisfies Assumption 1.*

(2) For $\beta \geq 0$*, and suppose that* Δ_n, M_n *take the form* (D.2)*.*

(3) Suppose Assumptions 2 to 4 hold.

Define the rate function

$$\delta_n = n^{-p/(2p+1)}(\log n)^{\frac{2+\alpha}{2\alpha}+\beta}. \tag{E.2}$$

Then, there exists some constant $B_{\mathcal{H}}$*, depending solely on* $C^*_{\mathcal{H}}$ *in Corollary D.1,* β*, and* p, ν_ℓ, ν_u *such that*

$$\mathrm{P}\left[\overline{Z}_n \leq M_n, \|\hat{\eta} - \eta\|_\infty \leq \Delta_n, h(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}) > B_{\mathcal{H}}\delta_n\right] \leq \left(\frac{\log\log n}{\log 2} + 10\right)\frac{1}{n}.$$

Proof. Let $\gamma = \frac{2+\alpha}{2\alpha} + \beta$. We first verify that, for ε_n in (D.3), we make the choices

$$\lambda_n = n^{-p/(2p+1)}(\log n)^{\frac{2+\alpha}{2\alpha}+\beta} \wedge 1 \quad \gamma_n = n^{-p/(2p+1)}(\log n)^{\frac{2+\alpha}{2\alpha}+\beta} \wedge 1$$

does satisfy $\lambda_n^2 \geq \gamma_n^2$, as required by Corollary D.1. Since $\varepsilon_n \lesssim \lambda_n \overline{h} + \gamma_n^2$, the truncation by 1 only affects our subsequent results by constant factors.

The event in question is a subset of the union of

$$\left\{\overline{Z}_n \leq M_n, \|\hat{\eta} - \eta\|_\infty \leq \Delta_n, \mathrm{Sub}_n(\hat{G}_n) > C^*_{\mathcal{H}}\varepsilon_n\right\}$$

and

$$\left\{\overline{Z}_n \le M_n, \|\hat{\eta}-\eta\|_\infty \le \Delta_n, \mathrm{Sub}_n(\hat{G}_n) \le C^*_{\mathcal{H}}\varepsilon_n, \overline{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}) > B_{\mathcal{H}} n^{-p/(2p+1)}(\log n)^\gamma\right\}.$$

The first event has measure at most $9/n$ by Corollary D.1, and there exists a choice of $B_{\mathcal{H}}$ such that the second has measure at most $n^{-1}\left(\frac{\log\log n}{\log 2}+1\right)$ by Corollary E.1. We conclude the proof by applying a union bound. $\square$

E.1. **Proof of Theorem E.1.**

E.1.1. *Decompose $B(t, \lambda_n, \epsilon)$.* We decompose $B(t,\lambda_n,\epsilon) \subset \bigcup_{k=1}^K B_k(t,\lambda_n)$ where, for some constant B to be chosen,

$$B_k = \left\{H : \overline{h}\left(f_{H,\cdot}, f_{G_0,\cdot}\right) \in \left(tB\lambda_n^{1-2^{-k}}, tB\lambda_n^{1-2^{-k+1}}\right]\right\}.$$

The relation $B(t,\lambda_n,\epsilon) \subset \bigcup_k B_k$ holds if we take $K = \lceil |\log_2(1/\epsilon)|\rceil$, since, in that case, $K \ge \log_2(1/\epsilon) \implies 2^{-K} \le \epsilon \implies \lambda_n^{1-2^{-K}} \le \lambda_n^{1-\epsilon}$.

We will bound

$$\mathrm{P}(A(\gamma_n,\lambda_n) \cap B_k(t,\lambda_n) \neq \varnothing) \le n^{-t^2}$$

which becomes the bound (E.1) by a union bound. For $k \in [K]$, define $\mu_{n,k} = B\lambda_n^{1-2^{-k+1}}$ such that $B_k = \left\{H : \overline{h}\left(f_{H,\cdot}, f_{G_0,\cdot}\right) \in (t\mu_{n,k+1}, t\mu_{n,k}]\right\}$. To that end, fix some k.

E.1.2. *Construct a net for the set of densities f_G.* Fix a positive constant M and define the semi-norm

$$\|G\|_{\infty,M} = \max_{i\in[n]} \sup_{y\in[-M,M]} f_{G,\nu_i}(y).$$

Note that $\|G\|_{\infty,M}$ is proportional to $\|G\|_{0,\infty,M}$ defined in Proposition D.1. Fix $\omega = \frac{1}{n^2} > 0$ and consider an ω-net for the distribution $\mathcal{P}(\mathbb{R})$ under $\|\cdot\|_{\infty,M}$. Let $N = N(\omega, \mathcal{P}(\mathbb{R}), \|\cdot\|_{\infty,M})$ and the ω-net is the distributions $H_1,\dots,H_N$. For each j, let $H_{k,j}$ be the distribution with

$$\overline{h}(f_{H_{k,j},\cdot}, f_{G_0,\cdot}) \ge \mu_{n,k+1}$$

if it exists, and let J_k collect the indices for which $H_{j,k}$ exists.

E.1.3. *Project to the net and upper bound the likelihood.* Fix a distribution $H \in B_k(t,\lambda_n)$. There exists some H_j where $\|H-H_j\|_{\infty,M} \le \omega$. Moreover, H serves as a witness that $H_{k,j}$ exists, with $\|H-H_{k,j}\|_{\infty,M} \le 2\omega$.

We can construct an upper bound for $f_{H,\nu_i}(z)$ via

$$f_{H,\nu_i}(z) \le \begin{cases} f_{H_{k,j},\nu_i}(z) + 2\omega & |z| < M \\ \frac{1}{\sqrt{2\pi}\nu_i} & |z| \ge M \end{cases}.$$

Define

$$v(z) = \omega\mathbb{1}(|z|<M) + \frac{\omega M^2}{z^2}\mathbb{1}(|z|\ge M).$$

Observe that

$$f_{H,\nu_i}(z) \le \frac{f_{H_{k,j},\nu_i}(z) + 2v(z)}{\sqrt{2\pi}\nu_i v(z)} \text{ if } |z| > M$$

$$f_{H,\nu_i}(z) \le f_{H_{k,j},\nu_i}(z) + 2v(z) \text{ if } |z| \le M.$$

Hence, the likelihood ratio between H and G_0 is upper bounded:

$$\begin{aligned}
\prod_{i=1}^n \frac{f_{H,\nu_i}(Z_i)}{f_{G_0,\nu_i}(Z_i)} &\le \prod_{i=1}^n \frac{f_{H_{k,j},\nu_i}(Z_i) + 2v(Z_i)}{f_{G_0,\nu_i}(Z_i)} \prod_{i:|Z_i|>M} \frac{1}{\sqrt{2\pi}\nu_i v(Z_i)} \\
&\le \left(\max_{j \in J_k} \prod_{i=1}^n \frac{f_{H_{k,j},\nu_i}(Z_i) + 2v(Z_i)}{f_{G_0,\nu_i}(Z_i)} \right) \prod_{i:|Z_i|>M} \frac{1}{\sqrt{2\pi}\nu_i v(Z_i)}
\end{aligned}$$

If $H \in A(t, \gamma_n, \lambda_n)$, then the likelihood ratio is lower bounded:

$$\prod_{i=1}^n \frac{f_{H,\nu_i}(Z_i)}{f_{G_0,\nu_i}(Z_i)} \ge \exp\left(-nC^*(\gamma_n^2 + \overline{h}\left(f_{H,\cdot}, f_{G_0,\cdot}\right)\lambda_n)\right) \ge \exp\left(-ntC^*(t\gamma_n^2 + \overline{h}\left(f_{H,\cdot}, f_{G_0,\cdot}\right)\lambda_n)\right). \tag{$t > 1$}$$

Hence,

$$\begin{aligned}
&\mathrm{P}\left[A(t, \gamma_n, \lambda_n) \cap B_k(t, \lambda_n) \neq \varnothing\right] \\
&\le \mathrm{P}\left\{ \left(\max_{j \in J_k} \prod_{i=1}^n \frac{f_{H_{k,j},\nu_i}(Z_i) + 2v(Z_i)}{f_{G_0,\nu_i}(Z_i)} \right) \prod_{i:|Z_i|>M} \frac{1}{\sqrt{2\pi}\nu_i v(Z_i)} \ge \exp\left(-nt^2C^*(\gamma_n^2 + \mu_{n,k}\lambda_n)\right) \right\} \\
&\le \mathrm{P}\left[\max_{j \in J_k} \prod_{i=1}^n \frac{f_{H_{k,j},\nu_i} + 2v(Z_i)}{f_{G_0,\nu_i}(Z_i)} \ge e^{-nt^2aC^*(\gamma_n^2 + \mu_{n,k}\lambda_n)} \right] \qquad \text{(E.3)} \\
&\qquad + \mathrm{P}\left[\prod_{i:|Z_i|>M} \frac{1}{\sqrt{2\pi}\nu_i v(Y_i)} \ge e^{nt^2(a-1)C^*(\gamma_n^2 + \mu_{n,k}\lambda_n)} \right] \qquad \text{(E.4)}
\end{aligned}$$

The first inequality follows from plugging in $\overline{h} \le t\mu_{n,k}$. The second inequality follows from choosing some $a > 1$ and applying union bound.

E.1.4. *Bounding* (E.3). We consider bounding the first term (E.3) now:

$$\begin{aligned}
\text{(E.3)} &\le \sum_{j \in J_k} \mathrm{P}\left[\prod_{i=1}^n \frac{f_{H_{k,j},\nu_i} + 2v(Z_i)}{f_{G_0,\nu_i}(Z_i)} \ge e^{-nat^2C^*(\gamma_n^2 + \mu_{n,k}\lambda_n)} \right] && \text{(Union bound)} \\
&\le \sum_{j \in J_k} \mathbb{E}\left[\prod_{i=1}^n \sqrt{\frac{f_{H_{k,j},\nu_i}(Z_i) + 2v(Z_i)}{f_{G_0,\nu_i}(Z_i)}} \right] e^{nat^2C^*(\gamma_n^2 + \mu_{n,k}\lambda_n)/2} \\
& && \text{(Take square root of both sides, then apply Markov's inequality)} \\
&= \sum_{j \in J_k} e^{nat^2C^*(\gamma_n^2 + \mu_{n,k}\lambda_n)/2} \prod_{i=1}^n \mathbb{E}\left[\sqrt{\frac{f_{H_{k,j},\nu_i}(Z_i) + 2v(Z_i)}{f_{G_0,\nu_i}(Z_i)}} \right] && \text{(E.5)}
\end{aligned}$$

where the last step (E.5) is by independence over i. Note that

$$\mathbb{E}\left[\sqrt{\frac{f_{H_{k,j},\nu_i}(Z_i) + 2v(Z_i)}{f_{G_0,\nu_i}(Y_i)}} \right] = \int_{-\infty}^{\infty} \sqrt{f_{H_{k,j},\nu_i}(x) + 2v(x)} \sqrt{f_{G_0,\nu_i}(x)}\, dx$$

$$\le 1 - h^2(f_{H_{k,j},\nu_i}, f_{G_0,\nu_i}) + \int_{-\infty}^{\infty} \sqrt{2v(x) f_{G_0,\nu_i}(x)}\, dx \qquad (\sqrt{a+b} \le \sqrt{a} + \sqrt{b})$$

$$\le 1 - h^2(f_{H_{k,j},\nu_i}, f_{G_0,\nu_i}) + \left(2\int_{-\infty}^{\infty} v(x)\, dx\right)^{1/2} \qquad \text{(Jensen's inequality)}$$

$$= 1 - h^2(f_{H_{k,j},\nu_i}, f_{G_0,\nu_i}) + \sqrt{8M\eta} \qquad \text{(Direct integration)}$$

Also note that, for $t_i > 0$, we have

$$\prod_i t_i = \exp \sum_i \log t_i \le \exp\left(\sum_i (t_i - 1)\right).$$

and thus

$$\prod_{i=1}^n \mathbb{E}\left[\sqrt{\frac{f_{H_{k,j},\nu_i} + 2v(Z_i)}{f_{G_0,\nu_i}(Z_i)}}\right] \le \exp\left[-n\bar{h}^2(f_{H_{k,j},\cdot}, f_{G_0,\cdot}) + n\sqrt{8M\omega}\right].$$

Thus, we can further bound (E.5):

$$\begin{aligned}
\text{(E.3)} \le \text{(E.5)} &= \sum_{j \in J_k} e^{n\alpha t^2(\gamma_n^2 + \mu_{n,k}\lambda_n)/2} \prod_{i=1}^n \mathbb{E}\left[\sqrt{\frac{f_{H_{k,j},\nu_i} + 2v(Z_i)}{f_{G_0,\nu_i}(Z_i)}}\right] \\
&\le \sum_{j \in J_k} \exp\left\{\frac{nat^2C^*}{2}(\gamma_n^2 + \mu_{n,k}\lambda_n) - n\bar{h}^2(f_{H_{k,j},\cdot}, f_{G,\cdot}) + n\sqrt{8M\omega}\right\} \\
&\le \sum_{j \in J_k} \exp\left\{\frac{nat^2C^*}{2}(\gamma_n^2 + \mu_{n,k}\lambda_n) - nt^2\mu_{n,k+1}^2 + n\sqrt{8M\omega}\right\} \\
&\le \exp\left\{\frac{nat^2C^*}{2}(\gamma_n^2 + \mu_{n,k}\lambda_n) - nt^2\mu_{n,k+1}^2 + n\sqrt{8M\omega} + \log N\right\} && (|J_k| \le N) \\
&\le \exp\left\{\frac{nat^2C^*}{2}(\gamma_n^2 + \mu_{n,k}\lambda_n) - nt^2\mu_{n,k+1}^2 + n\sqrt{8M\omega} + C|\log\omega|^2 \max\left(\frac{M}{\sqrt{|\log\omega|}}, 1\right)\right\} \\
& && (\text{Proposition D.1}, q = 0) \\
&= \exp\left\{\frac{nat^2C^*}{2}(\gamma_n^2 + \mu_{n,k}\lambda_n) - nt^2\mu_{n,k+1}^2 + \sqrt{8M} + C(\log n)^2 \max\left(\frac{M}{\sqrt{\log n}}, 1\right)\right\}. \\
& && (\text{Recall that } \omega = \tfrac{1}{n^2})
\end{aligned}$$

E.1.5. *Bounding* (E.4). We now consider bounding the second term (E.4). By Markov's inequality again (taking $x \mapsto x^{1/(2\log n)}$ on both sides, we can choose to bound

$$\text{(E.4)} \le \mathbb{E}\left[\prod_{i=1}^n \left(\frac{1}{(2\pi\nu_i^2)^{1/4}} \frac{Z_i}{M\sqrt{\omega}}\right)^{\frac{1}{\log n}\mathbb{1}(|Z_i| > M)}\right] \exp\left(-\frac{n(a-1)t^2C^*(\gamma_n^2 + \mu_{n,k}\lambda_n)}{2\log n}\right)$$

instead. Define

$$a_i = \frac{1}{(2\pi\nu_i^2)^{1/4} M\sqrt{\omega}} \le \frac{C_{\nu_\ell} n}{M} \quad \lambda = \frac{1}{\log n}$$

Apply Lemma E.1 to obtain the following. Note that to do so, we require

$$M \ge \nu_u\sqrt{8\log n} \quad p \ge \frac{1}{\log n}$$

Then,

$$\log \mathbb{E}\left[\prod_{i=1}^{n}\left(\frac{1}{(2\pi\nu_i^2)^{1/4}}\frac{Z_i}{M\sqrt{\omega}}\right)^{\frac{1}{\log n}\mathbb{1}(|Z_i|>M)}\right] = \log \mathbb{E}\left[\prod_i (a_i Z_i)^{\lambda \mathbb{1}(|Z_i|\geq M)}\right]$$
$$\lesssim_{\nu_u} \sum_{i=1}^{n}(a_i M)^{\lambda}\left(\frac{1}{Mn}+\frac{2^p\mu_p^p(G_0)}{M^p}\right)$$
$$\leq \sum_{i=1}^{n}(C_{\nu_\ell} n)^{\frac{1}{\log n}}\left(\frac{1}{Mn}+\frac{2^p\mu_p^p(G_0)}{M^p}\right)$$
$$\lesssim_{\nu_u,\nu_\ell} \frac{1}{M}+\frac{2^p n\mu_p^p(G_0)}{M^p}$$

As a result,

$$\log[(\text{E.4})] \leq C_{\nu_u,\nu_\ell}\left(\frac{1}{M}+\frac{2^p n\mu_p^p(G_0)}{M^p}\right) - \frac{n(a-1)}{2\log n}t^2 C^*\left(\gamma_n^2 + B\lambda_n^{2(1-2^{-k})}\right). \tag{E.6}$$

E.1.6. *Choosing p, M, a and verifying conditions.* By Assumption 2, $\mu_p^p(G_0) \leq A_0^p p^{p/\alpha}$. Let $M = 2eA_0(c_m \log n)^{1/\alpha}$ and $p = (M/(2eA_0))^{1/\alpha}$ so that

$$2^p\mu_p^p(G_0)/M^p \leq \exp\left(-c_m \log n\right)$$

We choose $c_m \geq 2$ sufficiently large such that $M = 2eA_0(c_m\log n)^{1/\alpha} > \nu_u\sqrt{8\log n} \vee 1$ and $p \geq 1$ for all $n > 2$ to ensure that our application of Lemma E.1 is correct. Since $\alpha \leq 2$, such a choice is available. Hence,

$$\frac{2^p n\mu_p^p(G_0)}{M^p} \leq \frac{1}{n}.$$

Hence the first term in (E.6) is less than $2C_{\nu_u,\nu_\ell}$.

Choose $a = 1.5$ to obtain that

$$\log[(\text{E.4})] \leq 2C_{\nu_u,\nu_\ell} - \frac{n}{4\log n}t^2 C^*\left(\gamma_n^2 + B\lambda_n^{2(1-2^{-k})}\right)$$
$$\leq t^2\left[2C_{\nu_u,\nu_\ell} - \frac{n}{4\log n}C^* B\lambda_n^2\right] \qquad (t \geq 1, \gamma_n > 0, \lambda_n < 1)$$
$$\leq t^2\left[2C_{\nu_u,\nu_\ell} - \frac{C^* B C_\lambda}{4}(\log n)\right] \qquad (\lambda_n^2 \geq C_\lambda(\log n)^{1+\frac{\alpha+2}{2\alpha}}/n \geq C_\lambda(\log n)^2/n)$$

There exists a sufficiently large B dependent only on $C^*, C_\lambda, C_{\nu_u,\nu_\ell}$ where $2C_{\nu_u,\nu_\ell} - \frac{C^*BC_\lambda}{4}(\log n) \leq -\log n$ for all $n \geq 2$. Hence, for all sufficiently large B,

$$\log[(\text{E.4})] \leq -t^2 \log n.$$

Similarly, under these choices,

$$\log[(\text{E.3})] \leq -nt^2\left[-\frac{3}{4}C^*(\gamma_n^2 + B\lambda_n^{2(1-2^{-k})}) + B^2\lambda_n^{2(1-2^{-k+1})}\right] + C(\log n)^{1+\frac{2+\alpha}{2\alpha}}$$
$$\leq -nt^2\left[-\frac{3}{4}C^*(\lambda_n^2 + B\lambda_n^{2(1-2^{-k})}) + B^2\lambda_n^{2(1-2^{-k+1})}\right] + C(\log n)^{1+\frac{2+\alpha}{2\alpha}}t^2 \quad (\gamma_n \leq \lambda_n, t \geq 1)$$

$$
\le -t^2 \left[n\lambda_n^2 \left(-\frac{3}{4}C^* - \frac{3}{4}C^* B \left(\frac{1}{\lambda_n}\right)^{2^{-k+1}} + B^2 \left(\frac{1}{\lambda_n}\right)^{2^{-k+2}} \right) - C(\log n)^{1+\frac{2+\alpha}{2\alpha}} \right]
$$

$$
\le -t^2 \left[n\lambda_n^2 \left(\frac{1}{\lambda_n}\right)^{2^{-k+2}} \left(-\frac{3}{4}C^* - \frac{3}{4}C^* B + B^2 \right) - C(\log n)^{1+\frac{2+\alpha}{2\alpha}} \right]
$$

$$
(\lambda_n \le 1. \text{ Pick } B \text{ such that } -\tfrac{3}{4}C^* - \tfrac{3}{4}C^* B + B^2 > 0)
$$

$$
\le -t^2 \left[n\lambda_n^2 \left(-\frac{3}{4}C^* - \frac{3}{4}C^* B + B^2 \right) - C(\log n)^{1+\frac{2+\alpha}{2\alpha}} \right]
$$

$$
\le -t^2 (\log n)^{1+\frac{2+\alpha}{2\alpha}} \left[C_\lambda \left(-\frac{3}{4}C^* - \frac{3}{4}C^* B + B^2 \right) - C \right]
$$

There exists choices of B, depending solely on $C^*, C, C_\lambda, C_{\nu_u,\nu_\ell}$ where $\left[C_\lambda\left(-\frac{3}{4}C^* - \frac{3}{4}C^*B + B^2\right) - C\right] > 1$ so that the above is at most $-t^2 \log n - \log 2$.

Putting the union bound together, we obtain that

$$
(\text{E.3}) + (\text{E.4}) \le n^{-t^2}.
$$

This concludes the proof.

E.2. **Auxiliary lemmas.**

Lemma E.1 (Lemma 5 in Jiang (2020))**.** *Suppose $Z_i \mid \tau_i \sim \mathcal{N}(\tau_i, \nu_i^2)$ where $\tau_i \mid \nu_i^2 \sim G_0$ independently across i. Let $0 < \nu_u, \nu_\ell < \infty$ be the upper and lower bounds for ν_i. Then, for all constants $M > 0, \lambda > 0, a_i > 0, p \in \mathbb{N}$ such that $M \ge \nu_u\sqrt{8\log n}$, $\lambda \in (0, p \wedge 1)$, and $a_1, \ldots, a_n > 0$:*

$$
\mathbb{E}\left\{ \prod_i |a_i Z_i|^{\lambda \mathbb{1}(|Z_i| \ge M)} \right\} \le \exp\left\{ \sum_{i=1}^n (a_i M)^\lambda \left(\frac{4\nu_u}{Mn\sqrt{2\pi}} + \left(\frac{2\mu_p(G_0)}{M}\right)^p \right) \right\}.
$$

Appendix F. An oracle inequality for the Bayes squared-error risk

Recall the definition of MSERegret_n in (C.4) and the event A_n in (C.5).

F.1. **Controlling** MSERegret_n **on** A_n^{C}**.** The first term is the regret when a bad event occurs, on which either the nuisance estimates are bad or the data has large values. The probability of this bad event is

$$
\mathrm{P}(A_n^{\mathrm{C}}) \le \mathrm{P}(\|\hat\eta - \eta\|_\infty > \Delta_n) + \mathrm{P}(\overline{Z}_n > M_n) \le \mathrm{P}(\|\hat\eta - \eta\|_\infty > \Delta_n) + n^{-2}.
$$

There exist choices of the constant in (D.2) for M_n such that $\mathrm{P}(\overline{Z}_n > M_n) \le n^{-2}$, by Lemma F.6. Thus, at a minimum, the first term is $o(1)$ for appropriate choices of Δ_n, M_n such that $\mathrm{P}(A_n^{\mathrm{C}}) \to 0$. We can also control the expected value of MSERegret_n on the bad event A_n^{C}.

Lemma F.1. *Under Assumptions 1 to 4. For $\beta \ge 0$, suppose $n > 3$ and suppose Δ_n, M_n satisfies (D.2) such that $\mathrm{P}(\overline{Z}_n > M_n) \le n^{-2}$, we can decompose*

$$
\mathbb{E}[\text{MSERegret}_n(\hat G_n, \hat\eta)\mathbb{1}(\|\hat\eta - \eta\|_\infty > \Delta_n)] \lesssim_{\mathcal{H}} \mathrm{P}(\|\hat\eta - \eta\|_\infty > \Delta_n)^{1/2}(\log n)^{2/\alpha}
$$

$$
\mathbb{E}[\text{MSERegret}_n(\hat G_n, \hat\eta)\mathbb{1}(\overline{Z}_n > M_n)] \lesssim_{\mathcal{H}} \frac{1}{n}(\log n)^{2/\alpha}
$$

Proof. Observe that, for an event A on the data $Z_{1:n}$,

$$\mathbb{E}\left[\text{MSERegret}_n(\hat{G}_n, \hat{\eta})\mathbb{1}(A)\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_{i,\hat{G},\hat{\eta}} - \theta_i^*)^2 \mathbb{1}(A)\right]$$

$$\leq \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_{i,\hat{G},\hat{\eta}} - \theta_i^*)^2\right)^2\right]^{1/2} \text{P}(A)^{1/2}$$

by Cauchy–Schwarz.

A crude bound (Lemma F.5) shows that, almost surely,

$$\left\{\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_{i,\hat{G},\hat{\eta}} - \theta_i^*)^2\right\}^2 \lesssim_{\mathcal{H}} \overline{Z}_n^4.$$

Apply Lemma F.6 to find that $\mathbb{E}[\overline{Z}_n^4] \lesssim_{\mathcal{H}} (\log n)^{4/\alpha}$. This proves both claims. $\square$

F.2. **Controlling** MSERegret_n **on** A_n**.**

Theorem F.1. *Assume the conditions in Corollary E.2. That is,*

(1) Suppose $\hat{G}_n$ satisfies Assumption 1.

(2) For $\beta \geq 0$, suppose Δ_n, M_n satisfies (D.2).

(3) Suppose Assumptions 2 to 4 hold.

Then,

$$\mathbb{E}\left[\text{MSERegret}_n(\hat{G}_n, \hat{\eta})\mathbb{1}(A_n)\right] \lesssim_{\mathcal{H}} n^{-\frac{2p}{2p+1}} (\log n)^{\frac{2+\alpha}{\alpha}+3+2\beta}$$

Proof. Let $C^*_{\mathcal{H}}$ be the constant in Corollary D.1 and $B_{\mathcal{H}}$ be the constant in Corollary E.2. Recall the Hellinger rate δ_n in (E.2).

Recall the decomposition (C.4) for MSERegret_n. Note that the term corresponding to the second term in the decomposition (C.4),

$$\mathbb{E}\left[\mathbb{1}(A_n)\frac{2}{n}\sum_{i=1}^n (\theta_i^* - \theta_i)(\hat{\theta}_{i,\hat{G}_n,\hat{\eta}} - \theta_i^*)\right] = 0,$$

is mean zero, since $\mathbb{E}[(\theta_i^* - \theta_i) \mid Y_1, \ldots, Y_n] = 0$. Thus, we can focus on

$$\mathbb{E}\left[\frac{\mathbb{1}(A_n)}{n}\sum_{i=1}^n (\hat{\theta}_{i,\hat{G}_n,\hat{\eta}} - \theta_i^*)^2\right] \equiv \frac{1}{n}\mathbb{E}[\mathbb{1}(A_n)\|\hat{\theta}_{\hat{G}_n,\hat{\eta}} - \theta^*\|^2], \tag{F.1}$$

where we let $\hat{\theta}_{\hat{G}_n,\hat{\eta}}$ denote the vector of estimated posterior means and let θ^* denote the corresponding vector of oracle posterior means. Let the subscript ρ_n denote a vector of regularized posterior means as in (C.2). Thus, we may further decompose,

$$\|\hat{\theta}_{\hat{G}_n,\hat{\eta}} - \theta^*\| \leq \|\hat{\theta}_{\hat{G}_n,\hat{\eta}} - \hat{\theta}_{\hat{G}_n,\eta_0}\| + \|\hat{\theta}_{\hat{G}_n,\eta_0} - \hat{\theta}_{\hat{G}_n,\eta_0,\rho_n}\| + \|\hat{\theta}_{\hat{G}_n,\eta_0,\rho_n} - \theta^*_{\rho_n}\| + \|\theta^*_{\rho_n} - \theta^*\|.$$

Let

$$\xi_1 = \frac{\mathbb{1}(A_n)}{n}\|\hat{\theta}_{\hat{G}_n,\hat{\eta}} - \hat{\theta}_{\hat{G}_n,\eta_0}\|^2 \tag{F.2}$$

$$\xi_2 = \frac{\mathbb{1}(A_n)}{n}\|\hat{\theta}_{\hat{G}_n,\eta_0} - \hat{\theta}_{\hat{G}_n,\eta_0,\rho_n}\|^2 \tag{F.3}$$

$$\xi_3 = \frac{\mathbb{1}(A_n)}{n}\|\hat{\theta}_{\hat{G}_n,\eta_0,\rho_n} - \theta^*_{\rho_n}\|^2 \tag{F.4}$$

$$\xi_4 = \frac{\mathbb{1}(A_n)}{n}\|\theta^*_{\rho_n} - \theta^*\|^2 \tag{F.5}$$

corresponding to the square of each of the terms, such that

$$(\text{F.1}) \le 4(\mathbb{E}\xi_1 + \mathbb{E}\xi_2 + \mathbb{E}\xi_3 + \mathbb{E}\xi_4) = 4(\mathbb{E}\xi_1 + \mathbb{E}\xi_3 + \mathbb{E}\xi_4).$$

Observe that $\xi_2 = 0$ by Lemma D.9, since the truncation by ρ_n does not bind when A_n occurs.

The ensuing subsections control $\mathbb{E}\xi_1, \mathbb{E}\xi_3, \mathbb{E}\xi_4$ individually. Putting together the rates we obtain, we find that

$$\begin{aligned} \xi_1 \lesssim_{\mathcal{H}} M_n^6\Delta_n^2 \implies \mathbb{E}\xi_1 &\lesssim_{\mathcal{H}} M_n^2(\log n)^2\Delta_n^2 \\ \mathbb{E}\xi_3 &\lesssim_{\mathcal{H}} (\log n)^3\delta_n^2 \\ \mathbb{E}\xi_4 &\lesssim_{\mathcal{H}} \frac{1}{n} \end{aligned}$$

Now, observe that $\delta_n \asymp_{\mathcal{H}} \Delta_n M_n^2 \gtrsim_{\mathcal{H}} \Delta_n M_n \log n$ and $\frac{1}{n} \lesssim_{\mathcal{H}} (\log n)^3\delta_n^2$. Hence, the dominating rate is $(\log n)^3\delta_n^2$. Plugging in δ_n^2 in (E.2) to obtain the rate

$$(\text{F.1}) \lesssim_{\mathcal{H}} n^{-\frac{2p}{2p+1}}(\log n)^{\frac{2+\alpha}{\alpha}+3+2\beta_1}. \qquad \square$$

F.3. **Controlling ξ_1.**

Lemma F.2. *Under the assumptions of Theorem F.1, in the proof of Theorem F.1,* $\xi_1 \lesssim_{\mathcal{H}} M_n^2(\log n)^2\Delta_n^2$.

Proof. Note that, by an application of Taylor's theorem,

$$\begin{aligned} \left|\hat{\theta}_{i,\hat{G}_n,\hat{\eta}} - \hat{\theta}_{i,\hat{G}_n,\eta_0}\right| &= \sigma_i^2\left|\frac{f'_{\hat{G}_n,\hat{\nu}_i}(\hat{Z}_i)}{\hat{s}_i f_{\hat{G}_n,\hat{\nu}_i}(\hat{Z}_i)} - \frac{f'_{\hat{G}_n,\nu_i}(Z_i)}{s_{0i}f_{\hat{G}_n,\nu_i}(Z_i)}\right| \\ &= \sigma_i^2\left|\left(\left.\frac{\partial\psi_i}{\partial m_i}\right|_{\hat{G}_n,\hat{\eta}} - \left.\frac{\partial\psi_i}{\partial m_i}\right|_{\hat{G}_n,\eta_0}\right)\right| \\ &= \sigma_i^2\left|\left.\frac{\partial^2\psi_i}{\partial m_i\partial s_i}\right|_{\hat{G}_n,\tilde{\eta}_i}(\hat{s}_i - s_{0i}) + \left.\frac{\partial^2\psi_i}{\partial m_i^2}\right|_{\hat{G}_n,\tilde{\eta}_i}(\hat{m}_i - m_{0i})\right|, \end{aligned}$$

where we use $\tilde{\eta}_i$ to denote some intermediate value lying on the line segment between $\hat{\eta}_i$ and η_{0i}. By Lemma D.14,

$$\mathbb{1}(A_n)\left|\hat{\theta}_{i,\hat{G}_n,\hat{\eta}} - \hat{\theta}_{i,\hat{G}_n,\eta_0}\right| \lesssim_{\mathcal{H}} M_n \log n\Delta_n.$$

Hence, squaring both sides, we obtain $\xi_1 \lesssim_{\mathcal{H}} M_n^2(\log n)^2\Delta_n^2$. $\square$

F.4. **Controlling ξ_3.**

Lemma F.3. *Under the assumptions of Theorem F.1, in the proof of Theorem F.1,* $\mathbb{E}\xi_3 \lesssim_{\mathcal{H}} (\log n)^3\delta_n^2$.

Proof. Observe that

$$\left|\hat{\theta}_{i,\hat{G}_n,\eta_0,\rho_n} - \theta^*_{i,\rho_n}\right| = s_{0i}\left|\hat{\tau}_{i,\hat{G}_n,\eta_0,\rho_n} - \tau^*_{i,\rho_n}\right|$$

where $\hat{\tau}_{i,\hat{G}_n,\eta_0,\rho_n}$ is the regularized posterior with prior $\hat{G}_n$ at nuisance parameter η_0 and $\tau^*_{i,\rho_n} = \hat{\tau}_{i,G_0,\eta_0,\rho_n}$.

We shall focus on controlling

$$\mathbb{1}(A_n)\|\hat{\tau}_{\hat{G}_n,\eta_0,\rho_n} - \tau^*_{\rho_n}\|^2$$

Fix the rate function δ_n in (E.2) and the constant $B_{\mathcal{H}}$ in Corollary E.2 (which in turn depends on $C^*_{\mathcal{H}}$ in Corollary D.1). Let $B_n = \{\bar{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}) < B_{\mathcal{H}}\delta_n\}$ be the event of a small average squared Hellinger distance. Let $G_1, \ldots, G_N$ be a finite set of prior distributions (chosen to be a net of $\mathcal{P}(\mathbb{R})$ in some distance), and let $\tau^{(j)}_{\rho_n}$ be the posterior mean vector corresponding to prior G_j with nuisance parameter η_0 and regularization ρ_n.

Then

$$\frac{\mathbb{1}(A_n)}{n}\|\hat{\tau}_{\hat{G}_n,\eta_0,\rho_n} - \tau^*_{\rho_n}\|^2 \le \frac{4}{n}\left(\zeta_1^2 + \zeta_2^2 + \zeta_3^2 + \zeta_4^2\right)$$

where

$$\zeta_1^2 = \|\hat{\tau}_{\hat{G}_n,\eta_0,\rho_n} - \tau^*_{\rho_n}\|^2 \mathbb{1}\left(A_n \cap B_n^{\mathrm{C}}\right) \tag{F.6}$$

$$\zeta_2^2 = \left(\|\hat{\tau}_{\hat{G}_n,\eta_0,\rho_n} - \tau^*_{\rho_n}\| - \max_{j\in[N]}\|\tau^{(j)}_{\rho_n} - \tau^*_{\rho_n}\|\right)_+^2 \mathbb{1}(A_n \cap B_n) \tag{F.7}$$

$$\zeta_3^2 = \max_{j\in[N]}\left(\|\tau^{(j)}_{\rho_n} - \tau^*_{\rho_n}\| - \mathbb{E}\left[\|\tau^{(j)}_{\rho_n} - \tau^*_{\rho_n}\|\right]\right)_+^2 \tag{F.8}$$

$$\zeta_4^2 = \max_{j\in[N]}\left(\mathbb{E}\left[\|\tau^{(j)}_{\rho_n} - \tau^*_{\rho_n}\|\right]\right)^2 \tag{F.9}$$

The decomposition ζ_1 through ζ_4 is exactly analogous to Section C.3 in Soloff et al. (2021) and to the proof of Theorem 1 in Jiang (2020). In particular, ζ_1 is the gap on the "bad event" where the average squared Hellinger distance is large, which is manageable since $\mathbb{1}(A_n \cap B_n^{\mathrm{C}})$ has small probability by Corollary E.2. ζ_2 is the distance from the posterior means at $\hat{G}_n$ to the closest posterior mean generated from the net $G_1, \ldots, G_N$; ζ_2 is small if we make the net very fine. ζ_3 measures the distance between $\|\tau^{(j)}_{\rho_n} - \tau^*_{\rho_n}\|$ and its expectation; ζ_3 can be controlled by (i) a large-deviation inequality and (ii) controlling the metric entropy of the net (Proposition D.2). Lastly, ζ_4 measures the expected distance between $\tau^{(j)}_{\rho_n}$ and $\tau^*_{\rho_n}$; it is small since G_j are fixed priors with small average squared Hellinger distance.

However, our argument for ζ_3 is slightly different and avoids an argument in Jiang and Zhang (2009) which appears to not apply in the heteroskedastic setting. See Remark F.1.

The subsequent subsections control ζ_1 through ζ_4, and find that $\zeta_4 \lesssim_{\mathcal{H}} (\log n)^3\delta_n^2$ is the dominating term. □

F.4.1. *Controlling* ζ_1. First, we note that

$$\left(\hat{\tau}_{i,\hat{G}_n,\eta_0,\rho_n} - \tau^*_{\rho_n}\right)^2 \mathbb{1}(A_n \cap B_n^{\mathrm{C}}) \lesssim_{\mathcal{H}} \log(1/\rho_n)\mathbb{1}(A_n \cap B_n^{\mathrm{C}}) = \log n \mathbb{1}(A_n \cap B_n^{\mathrm{C}})$$

By Corollary E.2, $\mathrm{P}(A_n \cap B_n^{\mathrm{C}}) \le \left(\frac{\log\log n}{\log 2} + 9\right)\frac{1}{n}$, and hence

$$\frac{1}{n}\mathbb{E}\zeta_1^2 \lesssim_{\mathcal{H}} \frac{\log n \log\log n}{n}.$$

F.4.2. *Controlling* ζ_2. Choose $G_1, \ldots, G_N$ to be a minimal ω-covering of $\left\{G : \overline{h}(f_{G,\cdot}, f_{G_0,\cdot}) \le \delta_n\right\}$ under the pseudometric

$$d_{M_n,\rho_n}(H_1, H_2) = \max_{i\in[n]} \sup_{z:|z|\le M_n} \left| \frac{\nu_i^2 f'_{H_1,\nu_i}(z)}{f_{H_1,\nu_i}(z) \vee \left(\frac{\rho_n}{\nu_i}\right)} - \frac{\nu_i^2 f'_{H_2,\nu_i}(z)}{f_{H_2,\nu_i}(z) \vee \left(\frac{\rho_n}{\nu_i}\right)} \right| \tag{F.10}$$

where $N \le N\left(\omega, \mathcal{P}(\mathbb{R}), d_{M_n,\rho_n}\right)$. We note that (F.10) and (D.17) are different only by constant factors. Therefore, Proposition D.2 implies that

$$\log N\left(\frac{\delta\log(1/\delta)}{\rho_n}\sqrt{\log(1/\rho_n)}, \mathcal{P}(\mathbb{R}), d_{M_n,\rho_n}\right) \lesssim_{\mathcal{H}} \log(1/\delta)^2 \max\left(1, \frac{M_n}{\sqrt{\log(1/\delta)}}\right) \tag{F.11}$$

for all sufficiently small $\delta > 0$.

Then

$$\begin{aligned}
\frac{1}{n}\zeta_2^2 &\le \mathbb{1}(A_n \cap B_n) \max_{j\in[N]} \|\hat{\tau}_{\hat{G}_n,\eta_0,\rho_n} - \tau_{\rho_n}^{(j)}\|^2 \qquad (\text{Triangle inequality}: \|a-b\| - \|b-c\| \le \|a-c\|) \\
&= \mathbb{1}(A_n \cap B_n) \max_{j\in[N]} \sum_{i=1}^n \mathbb{1}\left(|Z_i| \le M_n\right) \left(\frac{\nu_i^2 f'_{\hat{G}_n,\nu_i}(Z_i)}{f_{\hat{G}_n,\nu_i}(Z_i) \vee \left(\frac{\rho_n}{\nu_i}\right)} - \frac{\nu_i^2 f'_{G_j,\nu_i}(Z_i)}{f_{G_j,\nu_i}(Z_i) \vee \left(\frac{\rho_n}{\nu_i}\right)} \right)^2 \\
&\le \omega^2 \\
&\le \frac{\delta^2 \log(1/\delta)^2}{\rho_n^2} \log(1/\rho_n). \qquad (\text{Reparametrize } \omega = \delta\log(1/\delta)\rho_n^{-1}\sqrt{\log(1/\rho_n)})
\end{aligned}$$

F.4.3. *Controlling* ζ_3. We first observe that $V_{ij} \equiv |\tau_{i,\rho_n}^{(j)} - \tau_{i,\rho_n}^*| \lesssim_{\mathcal{H}} \sqrt{\log n}$, by Lemma D.10. Let $V_j = (V_{1j}, \ldots, V_{nj})'$, we have that

$$\zeta_3 = \max_j (\|V_j\| - \mathbb{E}\|V_j\|)_+$$

Let $K_n = C_{\mathcal{H}} \log n \ge \max_{ij} |V_{ij}|$. Since G_j, G_0 are both fixed, $V_{1j}, \ldots, V_{nj}$ are mutually independent.

Observe that

$$\mathrm{P}\left(\|V_j\| > \mathbb{E}[\|V_j\|] + u\right) = \mathrm{P}\left(\left\|\frac{V_j}{K_n}\right\| \ge \mathbb{E}\left\|\frac{V_j}{K_n}\right\| + \frac{u}{K_n}\right) \le \exp\left(-\frac{u^2}{2K_n^2}\right).$$

by Lemma F.7. By a union bound,

$$\mathrm{P}\left(\zeta_3^2 > x\right) \le N \exp\left(-\frac{x}{2K_n^2}\right).$$

Therefore

$$\mathbb{E}[\zeta_3^2] = \int_0^\infty \mathrm{P}(\zeta_3^2 > x)\, dx$$

$$
\begin{aligned}
&= \int_0^\infty \min\left(1, N\exp\left(-\frac{x}{2K_n^2}\right)\right) dx \\
&= 2K_n^2 \log N + \int_{2K_n^2\log N}^\infty N\exp\left(-\frac{x}{2K_n^2}\right) dx \\
&\lesssim_{\mathcal{H}} \log n \log N.
\end{aligned}
$$

Now, if we take $\delta = \rho_n/n$, then

$$
\frac{1}{n}\mathbb{E}[\zeta_2^2 + \zeta_3^2] \lesssim_{\mathcal{H}} \frac{(\log n)^3}{n}.
$$

Remark F.1. For the analogous term in the homoskedastic setting, Jiang and Zhang (2009) (and, later on, Saha and Guntuboyina (2020)) observe that $\|\tau_{\rho_n}^{(j)} - \tau_{\rho_n}^*\|$ is a Lipschitz function of the noise component $Z_i - \tau_i$. As a result, a Gaussian isoperimetric inequality (Theorem 5.6 in Boucheron et al. (2013)) establishes that

$$
\mathrm{P}\left(\|\tau_{\rho_n}^{(j)} - \tau_{\rho_n}^*\| \geq \mathbb{E}\left[\|\tau_{\rho_n}^{(j)} - \tau_{\rho_n}^*\| \mid \tau_1, \ldots, \tau_n\right] + x\right)
$$

is small, independently of n—a fact used in Proposition 4 of Jiang and Zhang (2009). Note that the concentration of $\|\tau_{\rho_n}^{(j)} - \tau_{\rho_n}^*\|$ is towards its conditional mean $\mathbb{E}\left[\|\tau_{\rho_n}^{(j)} - \tau_{\rho_n}^*\| \mid \tau_1, \ldots, \tau_n\right]$. In the homoskedastic setting where $\nu_i = \nu$,

$$
\mathbb{E}\left[\|\tau_{\rho_n}^{(j)} - \tau_{\rho_n}^*\| \mid \tau_1, \ldots, \tau_n\right] = \mathbb{E}_{G_{0,n}}\left[\|\tau_{\rho_n}^{(j)} - \tau_{\rho_n}^*\|\right] \tag{F.12}
$$

where $G_{0,n} = \frac{1}{n}\sum_i \delta_{\tau_i}$ is the empirical distribution of the τ's. However, (F.12) no longer holds in the heteroskedastic setting, and to adapt this argument, we need to additionally control the difference between $\mathbb{E}\left[\|\tau_{\rho_n}^{(j)} - \tau_{\rho_n}^*\| \mid \tau_1, \ldots, \tau_n\right]$ and $\mathbb{E}\left[\|\tau_{\rho_n}^{(j)} - \tau_{\rho_n}^*\|\right]$. The arguments of Jiang (2020) (p.2289) and Soloff et al. (2021) (Section C.3.3, arXiv:2109.03466v1) appear to use the Gaussian concentration of Lipschitz functions argument without the additional step.

Instead, we establish control of ζ_3 by observing that entries of $\tau_{\rho_n}^{(j)} - \tau_{\rho_n}^*$ are bounded and applying the convex Lipschitz concentration inequality. Since, like Soloff et al. (2021), we seek regret control in terms of mean-squared error, this argument applies to their setting as well. Jiang (2020), on the other hand, seeks regret control in terms of root-mean-squared error, and it is unclear if similar fixes apply.

■

F.4.4. *Controlling* ζ_4. Consider a change of variables where we let $w_i = z/\nu_i$ and $\lambda_i = \tau/\nu_i$. Let $G_{(i)}$ be the distribution of λ_i under G, where

$$
G_{(i)}(d\lambda) = G(d\tau)
$$

Then

$$
f_{G,\nu_i}(z) = \int \frac{1}{\nu_i}\varphi\left(w_i - \lambda_i\right) G(d\tau) = \frac{1}{\nu_i}\int \varphi\left(w_i - \lambda_i\right) G_{(i)}(d\lambda_i) = \frac{1}{\nu_i} f_{G_{(i)},1}(w_i)
$$

and

$$
f'_{G,\nu_i}(z) = \frac{1}{\nu_i^2} f'_{G_{(i)},1}(w_i).
$$

Hence,

$$\mathbb{E}(\tau_{\rho_n}^{(j)} - \tau_{\rho_n}^*)^2 = \nu_i^2 \mathbb{E}\left(\frac{f'_{G_{ji},1}(w_i)}{f_{G_{ji},1}(w_i) \vee \rho_n} - \frac{f'_{G_{0i},1}(w_i)}{f_{G_{0i},1}(w_i) \vee \rho_n}\right)^2$$
$$\lesssim_{\mathcal{H}} \max\left((\log 1/\rho_n)^3, |\log h(f_{G_{ji},1}, f_{G_{0i},1})|\right) h^2(f_{G_{ji},1}, f_{G_{0i},1}) \qquad \text{(Lemmas D.9 and F.8)}$$
$$= \max\left((\log 1/\rho_n)^3, |\log h(f_{G_j,\nu_i}, f_{G_0,\nu_i})|\right) h^2(f_{G_j,\nu_i}, f_{G_0,\nu_i}) \qquad \text{(Hellinger distance is invariant to change-of-variables)}$$

Let $h_i = h(f_{G_j,\nu_i}, f_{G_0,\nu_i})$.

Hence,

$$\frac{1}{n}\mathbb{E}[\zeta_4^2] \lesssim_{\mathcal{H}} \frac{(\log n)^3}{n} \sum_{i:|\log h_i| < (\log 1/\rho_n)^3} h_i^2 + \frac{1}{n} \sum_{i:|\log h_i| > (\log 1/\rho_n)^3} |\log h_i| h_i^2$$
$$\le (\log n)^3 \overline{h}^2(f_{G_j,\cdot}, f_{G_0,\cdot}) + \frac{1}{n} \sum_{i:|\log h_i| > (\log 1/\rho_n)^3} \frac{1}{e} h_i \qquad (x|\log x| \le e^{-1})$$

Note that

$$|\log h_i| > (\log 1/\rho_n)^3 \implies h_i < \exp\left(-\log(1/\rho_n)^3\right) < \rho_n^{(\log 1/\rho_n)^2} \lesssim_{\mathcal{H}} \rho_n^3 \lesssim_{\mathcal{H}} n^{-1}. \qquad \text{(Assumption D.1)}$$

Therefore the first term dominates, and

$$\frac{1}{n}\mathbb{E}[\zeta_4^2] \lesssim_{\mathcal{H}} (\log n)^3 \delta_n^2.$$

F.5. **Controlling ξ_4.**

Lemma F.4. *Under the assumptions of Theorem F.1, in the proof of Theorem F.1,* $\mathbb{E}\xi_4 \lesssim_{\mathcal{H}} \frac{1}{n}$.

Proof. Note that

$$\mathbb{E}[(\theta_{i,\rho_n}^* - \theta_i^*)^2] = \int \left(\nu_i^2 \frac{f'_{G_0,\nu_i}(z)}{f_{G_0,\nu_i}(z)}\right)^2 \left(1 - \frac{f_{G_0,\nu_i}}{f_{G_0,\nu_i} \vee \frac{\rho_n}{\nu_i}}\right)^2 f_{G_0,\nu_i}(z)\, dz$$
$$\le \mathbb{E}\left[\left(\nu_i^2 \frac{f'_{G_0,\nu_i}(z)}{f_{G_0,\nu_i}(z)}\right)^4\right]^{1/2} \mathrm{P}\left[f_{G_0,\nu_i}(Z) < \rho_n/\nu_i\right]^{1/2} \qquad \text{(Cauchy–Schwarz)}$$
$$\lesssim_{\mathcal{H}} \rho_n^{1/3} \operatorname{Var}(Z)^{1/6} \qquad \text{(Lemma D.12)}$$
$$\lesssim_{\mathcal{H}} \frac{1}{n}.$$

Therefore, $\mathbb{E}[\xi_4] \lesssim_{\mathcal{H}} \frac{1}{n}$. ∎

F.6. **Auxiliary lemmas.**

Lemma F.5. *Let $\hat{\theta}_{i,\hat{G},\hat{\eta}}$ be the posterior mean at prior $\hat{G}$ and nuisance parameter estimate at $\hat{\eta}$. Let $\theta_i^* = \hat{\theta}_{i,G_0,\eta_0}$ be the true posterior mean. Assume that $\hat{G}$ is supported within $[-\overline{M}_n, \overline{M}_n]$ where $\overline{M}_n = \max_i |\hat{Z}_i(\hat{\eta}) \vee 1|$. Let $\|\hat{\eta} - \eta\|_\infty = \max(\|\hat{m} - m_0\|_\infty, \|\hat{s} - s_0\|_\infty)$.*

Then, suppose

(1) $\|\hat{\eta} - \eta\|_\infty \lesssim_{\mathcal{H}} 1$.
(2) *Assumptions 2 and 3 holds.*
(3) $\hat{s} \gtrsim_{\mathcal{H}} s_{\ell n}$ *for some fixed sequence* $s_{\ell n} > 0$.

Then

$$\left|\hat{\theta}_{i,\hat{G},\hat{\eta}} - \theta_i^*\right| \lesssim_{\mathcal{H}} \overline{s}_{\ell n}^{-2} \overline{Z}_n.$$

Moreover, the assumptions are satisfied by Assumptions 1 to 4 with $s_{\ell n} = s_{0\ell} \asymp 1$.

Proof. Observe that

$$\begin{aligned}\left|\hat{\theta}_{i,\hat{G}_n,\hat{\eta}} - \hat{\theta}_{i,G_0,\eta_0}\right| &= \left| \frac{1}{\hat{s}_i} \frac{\hat{\nu}_i^2 f'_{\hat{G}_n,\hat{\nu}_i}(\hat{Z}_i)}{f_{\hat{G}_n,\hat{\nu}_i}(\hat{Z}_i)} - \frac{1}{s_{0i}} \frac{v_i^2 f'_{G_0\nu_i}(Z_i)}{f_{G_0,\nu_i}(Z_i)} \right| \\ &\lesssim_{\mathcal{H}} s_{\ell n}^{-1} \overline{M}_n + \overline{Z}_n.\end{aligned}$$

by the boundedness of $\hat{G}_n$ and Lemma D.18. Note that

$$|\hat{Z}_i(\hat{\eta})| = \left| \frac{s_{0i}}{\hat{s}_i} Z_i + \frac{m_{0i} - \hat{m}_i}{\hat{s}_i} \right| \lesssim_{\mathcal{H}} s_{\ell n}^{-1} |Z_i|.$$

Therefore,

$$\left|\hat{\theta}_{i,\hat{G}_n,\hat{\eta}} - \hat{\theta}_{i,G_0,\eta_0}\right| \lesssim_{\mathcal{H}} s_{\ell n}^{-2} \overline{Z}_n.$$

□

Lemma F.6. *Let* $\overline{Z}_n = \max_i |Z_i| \vee 1$. *Under Assumption 2, for* $t > 1$

$$\mathrm{P}(\overline{Z}_n > t) \le n \exp\left(-C_{A_0,\alpha,\nu_u} t^\alpha\right).$$

and

$$\mathbb{E}[\overline{Z}_n^p] \lesssim_{p,\mathcal{H}} (\log n)^{p/\alpha}.$$

Moreover, if $M_n = (C_{\mathcal{H}} + 1)(C_{2,\mathcal{H}}^{-1} \log n)^{1/\alpha}$ *as in (D.2), then for all sufficiently large choices of* $C_{\mathcal{H}}$, $\mathrm{P}(\overline{Z}_n > M_n) \le n^{-2}$.

Proof. The first claim is immediate under Lemma D.16 and a union bound.

The second claim follows from the observation that

$$\mathbb{E}[\max_i (|Z_i| \vee 1)^p] \le \left(\sum_i \mathbb{E}[(|Z_i| \vee 1)^{pc}] \right)^{1/c} \le n^{1/c} C_{\mathcal{H}}^p (pc)^{p/\alpha}.$$

where the last inequality follows from simultaneous moment control. Choose $c = \log n$ with $n^{1/\log n} = e$ to finish the proof.

For the "moreover" part, we have that

$$\mathrm{P}(Z_n > M_n) \le \exp\left(\log n - C_{A_0,\alpha,\nu_u} (C_{\mathcal{H}} + 1)^\alpha C_{2,\mathcal{H}}^{-1} \log n \right)$$

and it suffices to choose $C_{\mathcal{H}}$ such that $(C_{\mathcal{H}} + 1)^\alpha > \frac{3C_{2,\mathcal{H}}}{C_{A_0,\alpha,\nu_u}}$ so that $\mathrm{P}(Z_n > M_n) \le e^{-2\log n} = n^{-2}$. □

Lemma F.7. *Let $W = (W_1, \ldots, W_n)$ be a vector containing independent entries, where $W_i \in [0,1]$. Let $\|\cdot\|$ be the Euclidean norm. Then, for all $t > 0$*

$$\mathrm{P}\left[\|W\| > \mathbb{E}\|W\| + t\right] \le e^{-t^2/2}.$$

Proof. We wish to use Theorem 6.10 of Boucheron et al. (2013), which is a dimension-free concentration inequality for convex Lipschitz functions of bounded random variables. To do so, we observe that $w \mapsto \|w\|$ is Lipschitz with respect to $\|\cdot\|$, since

$$\|w+a\| \le \|w\| + \|a\| \quad \|w\| = \|w + a - a\| \le \|w+a\| + \|a\| \implies |\|w+a\| - \|w\|| \le \|a\|.$$

Moreover, trivially $\|\lambda w + (1-\lambda)v\| \le \lambda\|w\| + (1-\lambda)\|v\|$ for $\lambda \in [0,1]$, and hence $w \mapsto \|w\|$ is convex. Convexity implies separate convexity required in Theorem 6.10 of Boucheron et al. (2013). This checks all conditions and the claim follows by applying Theorem 6.10 of Boucheron et al. (2013). $\square$

Lemma F.8. *Let $f_H = f_{H,1}$. Then, for $0 < \rho_n \le \frac{1}{\sqrt{2\pi e^2}}$,*

$$\int \left(\frac{f'_{H_1}(x)}{f_{H_1}(x) \vee \rho_n} - \frac{f'_{H_0}(x)}{f_{H_0}(x) \vee \rho_n} \right)^2 f_{H_0}(x)\, dx$$
$$\lesssim \max\left((\log 1/\rho_n)^3, |\log h\left(f_{H_1}, f_{H_0}\right)| \right) h^2\left(f_{H_1}, f_{H_0}\right)$$

where we define the right-hand side to be zero if $H_1 = H_0$.

Proof. This claim is an intermediate step of Theorem 3 of Jiang and Zhang (2009). In (3.10) in Jiang and Zhang (2009), the left-hand side of this claim is defined as $r(f_{H_1}, \rho_n)$. Their subsequent calculation, which involves Lemma 1 of Jiang and Zhang (2009), proceeds to bound

$$r(f_{H_1}, \rho_n) \le 4e^2 h^2(f_{H_1}, f_{H_0}) \max\left(\varphi_+^6(\rho_n), 2a^2\right) + 2\varphi_+(\rho_n)\sqrt{2} h(f_{H_1}, f_{H_0}),$$

for $a^2 = \max\left(\varphi_+^2(\rho_n) + 1, |\log h^2\left(f_{H_1}, f_{H_0}\right)|\right)$. Collecting the powers on $h, \log h$ and using $\varphi_+(\rho_n) \lesssim \sqrt{\log(1/\rho_n)}$ proves the claim. $\square$

Appendix G. Estimating η_0 by local linear regression

In this section, we verify that estimating η_0 by local linear regression satisfies the conditions we require for the nuisance estimators, when the true nuisance parameters belong to a Hölder class of order $p = 2$: $m_0(\sigma), s_0(\sigma) \in C^2_{A_1}([\sigma_\ell, \sigma_u])$.

In our empirical application, we estimate m_0, s_0 by nonparametrically regressing Y_i on $x_i \equiv \log_{10}(\sigma_i)$.[78] Since $\log(\cdot)$ is a smooth transformation on strictly positive compact sets, Hölder smoothness conditions for (m_0, s_0) translate to the same conditions on $(\mathbb{E}[Y \mid x], \mathrm{Var}(Y \mid x) - \sigma^2(x))$, with potentially different constants. Moreover, scaling and translating x_i linearly do not affect our technical results. As a result, we assume, without essential loss of generality, $x_i \in [0,1]$. We abuse and recycle notation to write $m_0(x) = \mathbb{E}[Y_i \mid x_i = x], s_0(x) = \mathrm{Var}(\theta_i \mid x_i = x)$. We also note that $m_0(x), s_0(x) \in C^2_{A_3}([0,1])$ for some $A_3 \lesssim_{\mathcal{H}} A_1$.

[78]Correspondingly, let $\sigma(x) = 10^x$.

We will consider the following local linear regression of Y_i on x_i. There are many steps imposed for ease of theoretical analysis, but we conjecture are unnecessary in practice. In our empirical exercises, omitting these steps do not affect performance.

(LLR-1) Fix some kernel $K(\cdot)$. Use the direct plug-in procedure of Calonico et al. (2019) to estimate a bandwidth $\hat{h}_{n,m}$.

(LLR-2) For some $C_h > 1$, project $\hat{h}_{n,m}$ to some interval $[C_h^{-1} n^{-1/5}, C_h n^{-1/5}]$ so as to enforce that it converges at the optimal rate:[79]

$$\hat{h}_{n,m} \leftarrow (\hat{h}_{n,m} \vee C_h^{-1} n^{-1/5}) \wedge C_h n^{-1/5}.$$

(LLR-3) Using $\hat{h}_{n,m}$, estimate m_0 with the local linear regression estimator $\hat{m}_{\mathrm{raw}}$ under kernel $K(\cdot)$ and bandwidth $\hat{h}_{n,m}$.

(LLR-4) Project the resulting estimator $\hat{m}$ to the Hölder class $C^2_{A_3}([0,1])$:

$$\hat{m} \in \operatorname*{arg\,min}_{m \in C^2_{A_3}([0,1])} \|m - \hat{m}_{\mathrm{raw}}\|_\infty.$$

We obtain $\hat{m}$ through this procedure.

(LLR-5) Form estimated squared residuals $\hat{R}_i^2 = (Y_i - \hat{m}(x_i))^2$.

(LLR-6) Repeat (LLR-1) on data $(\hat{R}_i^2, x_i)$ to obtain a bandwidth $\hat{h}_{n,s}$.

(LLR-7) Repeat (LLR-2) to project $\hat{h}_{n,s}$.

(LLR-8) Using $\hat{h}_{n,s}$, estimate $v(x) = \mathbb{E}[R_i^2 \mid X = x]$ with the local linear regression estimator $\hat{v}$ under kernel $K(\cdot)$.

(LLR-9) Since $\hat{v}$ is a local linear regression estimator, it can be written as a linear smoother $\hat{v}(x) = \sum_{i=1}^n \ell_i(x; \hat{h}_{n,s}) \hat{R}_i^2$. Let an estimate of the effective sample size be

$$p_n = \frac{1}{n} \sum_{i=1}^n \frac{1}{\sum_{j=1}^n \ell_i^2(x_j, \hat{h}_{n,s})}. \tag{G.1}$$

(LLR-10) Truncate the estimated conditional standard deviation:

$$\hat{s}_{\mathrm{raw}}(x) = \sqrt{\hat{v}(x) - \sigma^2(x)} \vee \sqrt{\frac{2}{p_n + 2} \hat{v}(x)}. \tag{G.2}$$

(LLR-11) Finally, project the resulting estimate to the Hölder class as in (LLR-4):

$$\hat{s}(x) \in \operatorname*{arg\,min}_{\substack{s \in C^2_{A_3}([0,1]) \\ s^2(\cdot) \ge \frac{2}{p_n+2} \min_i \sigma_i^2}} \|s - \hat{s}_{\mathrm{raw}}\|_\infty.$$

In practice, we expect the projection steps (LLR-3), (LLR-4), (LLR-7), and (LLR-11) to be unnecessary, at least with exceedingly high probability, since (i) Calonico et al. (2019)'s procedure is consistent for the optimal bandwidth, which contracts at $n^{-1/5}$, and (ii) local linear regression estimated functions are likely sufficiently smooth to obey Assumption 4(3). Hence, in our empirical implementation, we do not enforce these steps and simply set $\hat{m} = \hat{m}_{\mathrm{raw}}, \hat{s} = \hat{s}_{\mathrm{raw}}$. Omitting the projection steps does not appear to affect performance.

[79] We use the $\leftarrow$ notation to reassign a variable so that we can reduce notation clutter.

To ensure we always have a positive estimate of s_0, we truncate at a particular point (G.2). This truncation rule is a heuristic (and improper) application of results from the literature on estimating non-centrality parameters. We digress and discuss the truncation rule in the next remark.

Remark G.1 (The truncation rule in (G.2))**.** The truncation rule in (G.2) is an ad hoc adjustment without affecting asymptotic performance.[80] It is based on a literature on the estimation of non-central χ^2 parameters (Kubokawa et al., 1993). Specifically, let $U_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\lambda_i, 1)$ and let $V = \sum_{i=1}^p U_i^2$ be a noncentral χ^2 random variable with p degrees of freedom and noncentrality parameter $\lambda = \sum_{i=1}^p \lambda_i^2$. The UMVUE for λ is $V - p$, which is dominated by its positive part $(V-p)_+$. Kubokawa et al. (1993) derive a class of estimators of the form $V - \phi(V;p)$ that dominate $(V-p)_+$ in squared error risk. An estimator in this class is $(V-p) \vee \frac{2}{p+2}V$.[81]

This setting is loosely connected to ours. Suppose m_0 is known, and we were using a Nadaraya–Watson estimator with uniform kernel. Then, for a given evaluation point x_0, we would be averaging nearby R_i^2's. Each R_i is conditionally Gaussian, $R_i \mid (\theta_i, \sigma_i) \sim \mathcal{N}(\theta_i - m_0(\sigma_i), \sigma_i^2)$ with approximately equal variance $\sigma_i^2 \approx \sigma(x_0)^2$. If there happens to be p_0 R_i^2's that we are averaging, the Nadaraya–Watson estimator is of the form

$$\hat{v}(x_0) = \frac{\sigma(x_0)^2}{p_0} \sum_{i=1}^{p} \left(\frac{R_i}{\sigma(x_0)} \right)^2$$

Conditional on σ_i^2, θ_i, the quantity $\sum_{i=1}^p \left(\frac{R_i}{\sigma(x_0)}\right)^2$ is (approximately) noncentral χ^2 with p degrees of freedom and noncentrality parameter

$$\lambda = \sum_{i=1}^{p_0} \left(\frac{\theta_i - m_0(x_i)}{\sigma(x_0)} \right)^2$$

Therefore, correspondingly, applying the truncation rule from Kubokawa et al. (1993), an estimator for the sample variance of θ_i, $\frac{1}{p_0}\sum_{i=1}^{p_0}(\theta_i - m_0(x_i))^2$, is

$$\big(\hat{v}(x_0) - \sigma^2(x_0)\big) \vee \frac{2}{p_0 + 2}\hat{v}(x_0).$$

Here, we apply this truncation rule (improperly) to the case where $\hat{v}(x_0)$ is a weighted average of the squared residuals, with potentially negative weights due to higher-order polynomials (equiv. higher-order kernels). To do so, we would need to plug in an analogue of p_0. We note that when independent random variables V_i have unit variance, the weighted average has variance equal to the squared length of the weights

$$\mathrm{Var}\left(\sum_i \ell_i(x) V_i \right) = \sum_{i=1}^{n} \ell_i^2(x).$$

[80] Indeed, since we already assumed that the true conditional variance $s_0(x) > s_\ell$, we can truncate by any vanishing sequence. Given any vanishing sequence, eventually it is lower than s_ℓ, and eventually $|\hat{s} - s_0|$ is small enough for the truncation to not bind. This is, in some sense, silly, since finite sample performance is likely affected if we truncate by, say, $\frac{1}{\log \log n}$, reflected in a large constant in the corresponding rate expression. Our following argument assumes that the truncation of order $O(n^{-4/5})$. Doing so is likely to achieve a smaller constant in the rate expression, despite not mattering asymptotically.

[81] Though, since neither $(V-p)_+$ and $(V-p) \vee \frac{2}{p+2}V$ is differentiable in V, they are not admissible.

Since a simple average has variance equal to $1/n$, we can take $\left(\sum_{i=1}^n \ell_i^2(x)\right)^{-1}$ to be an effective sample size. Our rule simply takes the average effective sample size over evaluation points in (G.1) and use it as a candidate for p. ■

The goal in this section is to control the following probability as a function of $t > 0$

$$\mathrm{P}\left(\|\hat{\eta} - \eta_0\|_\infty > C_{\mathcal{H}} t n^{-2/5} (\log n)^\beta\right)$$

for some constants $\beta, C_{\mathcal{H}}$ to be chosen. Since we treat $x_1, \ldots, x_n$ as fixed (fixed design), we shall do so placing some assumptions on sequences of the design points $x_{1:n}$ as a function of n. These assumptions are mild and satisfied when the design points are equally spaced. They are also satisfied with high probability when the design points are drawn from a well-behaved density $f(\cdot)$.

Before doing so, we introduce some notation on the local linear regression estimator. Note that, by translating and scaling if necessary, it is without essential loss of generality to assume x_i take values in $[0, 1]$. Let h_n denote some (possibly data-driven) choice of bandwidth. Let $u(x) = [1, x]'$ and let $B_{nx} = B_{nx}(h_n) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{x_i - x}{h_n}\right) u\left(\frac{x_i - x}{h_n}\right) u\left(\frac{x_i - x}{h_n}\right)'$. Then, it is easy to see that the local linear regression weights can be written in terms of B_{nx} and $u(\cdot)$:

$$s_n \equiv nh_n \quad \ell_i(x) = \ell_i(x, h_n) \equiv \frac{1}{s_n} u(0)' B_{nx}^{-1} u\left(\frac{x_i - x}{h_n}\right) K\left(\frac{x_i - x}{h_n}\right).$$

We shall maintain the following assumptions on the design points. The following assumptions introduce constants $(C_h, n_0, \lambda_0, a_0, K_0, K(\cdot), c, C, C_K, V_K)$ which we shall take as primitives like those in $\mathcal{H}$. The symbols $\lesssim, \gtrsim, \asymp$ are relative to these constants, and we will not keep track of exact dependencies through subscripts.

Assumption G.1. *For some constant $C_h > 1$, the data-driven bandwidth h_n is almost surely contained in the set $H_n \equiv [C_h^{-1} n^{-1/5} \vee \frac{1}{2n}, C_h n^{-1/5}]$.*

Assumption G.1 is automatically satisfied by the projection steps (LLR-3) and (LLR-7).

Assumption G.2. *The sequence of design points $(x_i : i = 1, \ldots, n)$ satisfy:*

(1) There exists a real number $\lambda_0 > 0$ and integer $n_0 > 0$ such that, for all $n \geq n_0$, any $x \in [0, 1]$, and any $\tilde{h} \in [C_h^{-1} n^{-1/5} \vee \frac{1}{2n}, C_h n^{-1/5}]$, the smallest eigenvalue $\lambda_{\min}(B_{nx}(\tilde{h})) \geq \lambda_0$.

(2) There exists a real number $a_0 > 0$ such that for any interval $I \subset [0, 1]$ and all $n \geq 1$,

$$\frac{1}{n} \sum_{i=1}^n \mathbb{1}(x_i \in I) \leq a_0 \left(\lambda(I) \vee \frac{1}{n}\right)$$

where $\lambda(I)$ is the Lebesgue measure of I.

(3) The kernel K is supported on $[-1, 1]$ and uniformly bounded by some positive constant K_0.

(4) There exists $c, C > 0$ such that for all $n \geq n_0$, the choice of p_n in (G.1) satisfies $cn^{4/5} \leq p_n(\tilde{h}) \leq Cn^{4/5}$ for all $\tilde{h} \in [C_h^{-1} n^{-1/5} \vee \frac{1}{2n}, C_h n^{-1/5}]$.

Assumption G.2(1–3) is nearly the same as Assumption (LP) in Tsybakov (2008). The only difference is that Assumption G.2(1) requires the lower bound λ_0 to hold uniformly over a range of bandwidth choices, relative to LP-1 in Tsybakov (2008), which requires λ_0 to hold for some deterministic sequence h_n. This is a mild strengthening of LP-1: Note that if x_i are drawn from a

Lipschitz-continuous, everywhere-positive density $f(x)$, then for $h \to 0, nh \to \infty$,

$$B_{nx}(h) \approx \int K(t)u(t)u(t)' f(x)\, dt \succeq \int K(t)u(t)u(t)'\, dt \left(\min_{x\in[0,1]} f(x) \right)$$

where $\succ$ denotes the positive-definite matrix order. Thus the minimum eigenvalue of $B_{nx}(h)$ should be positive irrespective of x and h. See, also, Lemma 1.5 in Tsybakov (2008).

Assumption G.2(2)–(3) are the same as (LP-2)–(LP-3) in Tsybakov (2008). (2) expects that the design points are sufficiently spread out, and (3) is satisfied by, say, the Epanechnikov kernel.

Lastly, (4) expects that the average effective sample size is about $s_n = nh_n \asymp n^{-4/5}$. Again, heuristically, if x_i are drawn from a Lipschitz and everywhere-positive density $f(x)$, then

$$\sum_{i=1}^n \ell_i^2(x_j) \approx n\frac{1}{s_n^2} h_n \cdot \int (u(0)' B_{n,x_j}^{-1} u(t)K(t))^2 f(x_j)\, dt = \frac{1}{s_n} \int (u(0)' B_{n,x_j}^{-1} u(t)K(t))^2 f(x_j)\, dt.$$

Hence the mean reciprocal p_n is of order s_n. We also remark that Assumption G.2 is satisfied by regular design points $x_i = i/n$.

Assumption G.3. *The kernel satisfies the following VC subgraph-type conditions. Let*

$$\mathcal{F}_k = \left\{ y \mapsto \left(\frac{y-x}{h}\right)^{k-1} K\left(\frac{y-x}{h}\right) : x \in [0,1], h \in H_n \right\}$$

for $k = 1, 2$. For any finitely supported measure Q,

$$N(\epsilon, \mathcal{F}_k, L_2(Q)) \le C_K (1/\epsilon)^{V_K}$$

for C_K, V_K that do not depend on Q.

Assumption G.3 is satisfied for a wide range of kernels, e.g. the Epanechnikov kernel. By Lemma 7.22 in Sen (2018), reproduced as Lemma G.2 below, so long as the function $t \mapsto t^{k-1}K(t)$ is bounded (assumed in Assumption G.2(3)) and of bounded variation (satisfied by any absolutely continuous kernel function), the covering number conditions hold by exploiting the finite VC dimension of subgraphs of these functions.

We now state and prove the main results in this section. The key to these arguments is Proposition G.1 on the bias and variance of local linear regression estimators. Proposition G.1 is uniform in both the evaluation point x and the bandwidth h, as long as the latter converges at the optimal rate.

Theorem G.1. *Suppose the conditional distribution $\theta_i \mid \sigma_i$ and the design points $\sigma_{1:n}$ satisfy Assumptions 2, 3, and G.2. Moreover, suppose m_0, s_0 satisfies Assumption 4(1) with $p = 2$. Suppose the kernel $K(\cdot)$ satisfies Assumption G.3. Let $\hat{m}, \hat{s}$ denote the estimators computed by (LLR-1) through (LLR-11). Then:*

(1) $\mathrm{P}\left(\hat{m}, \hat{s} \in C^2_{A_3}([0,1])\right) = 1$

(2) For some C depending only on the parameters in the assumptions, for all $n \ge 7$ and $t > 1$,

$$\mathrm{P}\left(\max\left(\|\hat{m} - m_0\|_\infty, \|\hat{s} - s_0\|\right) \ge Ctn^{-\frac{2}{5}} (\log n)^{1+2/\alpha}\right) \le \frac{1}{n^{10}t^2}. \tag{G.3}$$

(3) *For some c depending only on the parameters in the assumptions, for all $n \ge 7$,*

$$\mathrm{P}\left(\frac{c}{n} \le \hat{s}\right) = 1.$$

Proof. The first claim is true automatically by the projection to the Hölder space. The third claim is true automatically by (LLR-11), since $p_n \asymp n^{4/5}$ and $n^{-4/5} \gtrsim n^{-1}$.

Now, we show the second claim. Since we assume that m_0, s_0 lies in the Hölder space with $s_0 > s_{0\ell}$, then projection to the Hölder space (and truncation by $2/(2+p_n)\min_i \sigma_i^2$) worsens performance by at most a factor of two for all sufficiently large n. The projection to the Hölder space ensures that $\|\hat{\eta} - \eta_0\|_\infty$ is bounded a.s. for all n, so that we can remove "for all sufficiently large n" at the cost of enlarging a constant so as to accommodate the first finitely many values of n. As a result, it suffices to show that

$$\mathrm{P}\left(\max\left(\|\hat{m}_{\mathrm{raw}} - m_0\|_\infty, \|\hat{s}_{\mathrm{raw}} - s_0\|_\infty\right) > Ctn^{-2/5}(\log n)^\beta\right) \le \frac{1}{n^{10}t^2}$$

for some C and $\beta = 1 + 2/\alpha$.

Let $Y_i = m_0(x_i) + \xi_i$ where $\xi_i = \theta_i - m_0(x_i) + (Y_i - \theta_i)$. Note that we have simultaneous moment control for ξ_i:

$$\max_i \mathbb{E}[|\xi_i|^p]^{1/p} \lesssim p^{1/\alpha}$$

where α is the constant in Assumption 2. Therefore, we can apply Proposition G.1 to obtain

$$\mathrm{P}\left(\|\hat{m}_{\mathrm{raw}} - m_0\|_\infty > Ctn^{-2/5}(\log n)^{1+1/\alpha}\right) \le \frac{1}{2n^{10}t^2}$$

for the local linear regression estimator $\hat{m}_{\mathrm{raw}}$.

The same argument to control $\|\hat{s}_{\mathrm{raw}} - s_0\|_\infty$ is more involved. First observe that

$$|\hat{s}_{\mathrm{raw}}^2 - s_0^2| = |\hat{s}_{\mathrm{raw}} - s_0|(\hat{s}_{\mathrm{raw}} + s_0) \ge s_{0\ell}|\hat{s}_{\mathrm{raw}} - s_0|.$$

Also observe that for a positive f_0,

$$|\hat{f} \vee g - f_0| \le |\hat{f} - f_0| \vee |g|.$$

As a result, it suffices to control the upper bound in

$$\begin{aligned}
\|\hat{s}_{\mathrm{raw}} - s_0\|_\infty &\le \frac{1}{s_{0\ell}}\left(\|\hat{v} - v_0\|_\infty \vee \left(\frac{2}{2+p_n}\hat{v}\right)\right) && (v_0(x) \equiv \mathrm{Var}(Y_i \mid x_i = x)) \\
&\lesssim \|\hat{v} - v_0\|_\infty \vee \frac{\|\hat{v} - v_0\|_\infty + \|v_0\|_\infty}{2 + n^{4/5}} && (\text{Assumption G.2}) \\
&\lesssim \|\hat{v} - v_0\|_\infty && \text{(G.4)}
\end{aligned}$$

Now, observe that $\hat{R}_i^2 = R_i^2 + (m_0 - \hat{m})^2 - 2(m_0 - \hat{m})\xi_i$. Hence,

$$\begin{aligned}
|\hat{v}(x) - v_0(x)| &\le \left|\sum_{i=1}^n \ell_i(x, \hat{h}_{n,s})R_i^2 - v_0(x)\right| + \left\{\|m_0 - \hat{m}\|_\infty^2 + 2\|m_0 - \hat{m}\|_\infty\left(\max_{i\in[n]}|\xi_i|\right)\right\}\sum_{i=1}^n |\ell_i(x, \hat{h}_{n,s})| \\
&\le \left|\sum_{i=1}^n \ell_i(x, \hat{h}_{n,s})R_i^2 - v_0(x)\right| + C\left\{\|m_0 - \hat{m}\|_\infty^2 + 2\|m_0 - \hat{m}\|_\infty\left(\max_{i\in[n]}|\xi_i|\right)\right\}.
\end{aligned} \tag{G.5}$$

By Lemma 1.3 in Tsybakov (2008), the term $\sum_{i=1}^{n} |\ell_i(x, \hat{h}_{n,s})|$ is bounded uniformly in h and x by a constant. Note that

$$\tilde{\xi}_i \equiv R_i^2 - v_0(x_i)$$

has simultaneous moment control with a different parameter ($\tilde{\alpha} = \alpha/2$):

$$\max_i (\mathbb{E}|\tilde{\xi}_i|^p)^{1/p} \lesssim p^{2/\alpha}.$$

Thus, applying Proposition G.1 and taking care to plug in $\tilde{\xi}, \tilde{\alpha}$, we can bound the first term in (G.5)

$$\mathrm{P}\left(\left\| \sum_{i=1}^{n} \ell_i(x, \hat{h}_{n,s}) R_i^2 - v_0(x) \right\|_\infty \geq Ctn^{-2/5} (\log n)^{1+2/\alpha} \right) \leq \frac{1}{4n^{10}t^2}.$$

Note that by an application of Lemma F.6, for any $a, b > 0$, we have that

$$\mathrm{P}\left(\max_i |\xi_i| > C(a,b)t(\log n)^{1/\alpha} \right) < an^{-b}e^{-t^2}$$

As a result, the second term in (G.5) admits

$$\mathrm{P}\left(\|m_0 - \hat{m}\|_\infty^2 + 2\|m_0 - \hat{m}\|_\infty \left(\max_{i \in [n]} |\xi_i| \right) > Ctn^{-2/5}(\log n)^{1+2/\alpha} \right) \leq \frac{1}{4n^{10}t^2}$$

Finally, putting these bounds together, we have that

$$\mathrm{P}\left(\|\hat{v} - v_0\|_\infty > Ctn^{-2/5}(\log n)^{1+2/\alpha} \right) \leq \frac{1}{2n^{10}t^2},$$

where the same bound (with a different constant) holds for $\hat{s}_{\text{raw}}$ by (G.4).

Combining the bounds for $\hat{m}$ and $\hat{s}$, we obtain (G.3). This concludes the proof. □

Theorem G.2. *Under the assumptions of Theorem G.1, let $\hat{\eta} = (\hat{m}, \hat{s})$ denote estimators computed by (LLR-1) through (LLR-11). Then,*

$$\mathbb{E}\left[\mathrm{MSERegret}_n(\hat{G}_n, \hat{\eta}) \right] \lesssim n^{-2/5}(\log n)^{1+2/\alpha}.$$

Proof. Recall the event A_n in (C.5) for $\Delta_n = C_1 n^{-2/5}(\log n)^\beta$ and $M_n = C_2(\log n)^{1/\alpha}$, where C_1, C_2 are to be chosen and $\beta = 1 + 2/\alpha$. Define $\tilde{A}_n = A_n \cap \{s_{0\ell}/2 \leq \hat{s} \leq 2s_{0u}\}$. Decompose

$$\mathbb{E}\left[\mathrm{MSERegret}_n(\hat{G}_n, \hat{\eta}) \right] = \mathbb{E}\left[\mathrm{MSERegret}_n(\hat{G}_n, \hat{\eta}) \mathbb{1}(\tilde{A}_n) \right] + \mathbb{E}\left[\mathrm{MSERegret}_n(\hat{G}_n, \hat{\eta}) \mathbb{1}(\tilde{A}_n^{\mathrm{C}}) \right].$$

Note that, for all sufficiently large $n > N$, such that N depends only on C_1, β, s_ℓ, s_u, the event A_n implies $\{s_{0\ell}/2 \leq \hat{s} \leq 2s_{0u}\}$ and hence $A_n = \tilde{A}_n$. Thus, by Theorem G.1, for all sufficiently large n, on the event A_n, statements analogous to Assumption 4(2–4) hold for the estimator $\hat{\eta}$. As a result, we may apply Theorem F.1, *mutatis mutandis*, to obtain that

$$\mathbb{E}\left[\mathrm{MSERegret}_n(\hat{G}_n, \hat{\eta}) \mathbb{1}(\tilde{A}_n) \right] \lesssim n^{-4/5}(\log n)^{\frac{2+\alpha}{\alpha}+3+2\beta}$$

for all sufficiently large choices of C_1, C_2.

To control $\mathbb{E}\left[\mathrm{MSERegret}_n(\hat{G}_n, \hat{\eta}) \mathbb{1}(\tilde{A}_n^{\mathrm{C}}) \right]$, we observe that under Lemma F.5 and Theorem G.1(1 and 3), we have that almost surely,

$$\mathrm{MSERegret}_n(\hat{G}_n, \hat{\eta}) \lesssim n^4 \overline{Z}_n^2.$$

Hence, by Cauchy–Schwarz as in Lemma F.1,

$$\mathbb{E}\left[\mathrm{MSERegret}_n(\hat{G}_n, \hat{\eta})\mathbb{1}(\tilde{A}_n^{\mathrm{C}})\right] \lesssim \mathrm{P}(\tilde{A}_n^{\mathrm{C}})^{1/2} n^4 (\log n)^{2/\alpha},$$

where we apply Lemma F.6 to bound $\mathbb{E}[\overline{Z}_n^4]$.

For all sufficiently large $n > N$,

$$\mathrm{P}(A_n^{\mathrm{C}}) = \mathrm{P}(\tilde{A}_n^{\mathrm{C}}) \le \mathrm{P}(\overline{Z}_n > M_n) + \mathrm{P}(\|\hat{\eta} - \eta_0\|_\infty > \Delta_n).$$

Sufficiently large C_1, C_2 can be chosen such that the right-hand side is bounded by n^{-10}. To wit, we can apply Theorem G.1 to bound $\|\hat{\eta} - \eta_0\|_\infty$. We can apply Lemma F.6 to bound $\mathrm{P}(\overline{Z}_n > M_n)$.

As a result, we would obtain

$$\mathbb{E}\left[\mathrm{MSERegret}_n(\hat{G}_n, \hat{\eta})\mathbb{1}(\tilde{A}_n^{\mathrm{C}})\right] \lesssim \frac{1}{n}(\log n)^{2/\alpha}$$

for all sufficiently large n.

Since $\mathbb{E}[\mathrm{MSERegret}_n(\hat{G}_n, \hat{\eta})] \lesssim n^4 (\log n)^{2/\alpha}$ is finite for all n, at the cost of enlarging the implicit constant, we have the result of the theorem holding for all n. □

G.1. **Auxiliary lemmas.**

Proposition G.1. *Consider the local linear regression of data $Y_i = f_0(x_i) + \xi_i$ on the design points x_i, for $i = 1, \ldots, n$. Suppose f_0 belongs to a Hölder class of order two: $f_0 \in C_L^2([0,1])$ for some $L > 0$. Suppose that the design points satisfy Assumption G.2 and the (possibly data-driven) bandwidths h_n satisfy Assumption G.1. Assume the kernel additionally satisfies Assumption G.3.*

Assume that the residuals ξ_i are mean zero, and there exists a constant $A_\xi > 0, \alpha > 0$ such that

$$\max_{i=1,\ldots,n} (\mathbb{E}[|\xi_i|^p])^{1/p} \le A_\xi p^{1/\alpha}$$

for all $p \ge 2$. Let $\ell_i(x,h)$ be the weights corresponding to local linear regression, and define the bias part $b(x, h_n) = (\sum_{i=1}^n \ell_i(x, h_n) f_0(x_i)) - f_0(x_i)$ and the stochastic part $v(x,h) = \sum_{i=1}^n \ell_i(x,h)\xi_i$. Recall that H_n is the interval for h_n in Assumption G.1. Then:

(1) The bias term is of order $n^{-2/5}$:

$$\sup_{x \in [0,1], h \in H_n} |b(x,h)| \lesssim n^{-2/5}.$$

(2) The variance term admits the following large-deviation inequality: For any $a, b > 0$, there exists a constant $C(a,b)$, which may additionally depend on the constants in the assumptions, such that for all $n > 1$ and $t \ge 1$

$$\mathrm{P}\left(\sup_{x \in [0,1], h \in H_n} |v(x,h)| > C(a,b) \cdot t \cdot (\log n)^{1+1/\alpha} n^{-2/5}\right) \le a n^{-b} \frac{1}{t^2}.$$

(3) As a result, let $\hat{f}(\cdot) = b(\cdot, h_n) + v(\cdot, h_n) + f_0(\cdot)$, we have that for any $a, b > 0$, there exists a constant $C(a,b)$ such that for all $n > 1$ and $t \ge 1$,

$$\mathrm{P}\left(\|\hat{f} - f_0\|_\infty > C(a,b) t (\log n)^{1+1/\alpha} n^{-2/5}\right) \le a n^{-b} \frac{1}{t^2}.$$

Proof. Note that (3) follows immediately from (1) and (2) since the bounds in (1) and (2) are uniform over all $h \in H_n$. We now verify (1) and (2).

(1) This claim follows immediately from the bound for $b(x_0)$ in Proposition 1.13 in Tsybakov (2008). The argument in Tsybakov (2008) shows that

$$\sup_{x\in[0,1]} |b(x,h_n)| \le Ch_n^2,$$

which is uniformly bounded by $Cn^{-2/5}$ by Assumption G.1. Hence

$$\sup_{x\in[0,1],h\in H_n} |b(x,h)| \lesssim n^{-2/5}.$$

(2) Let M be a truncation point to be defined. Let

$$\xi_{i,<M} = \xi_i \mathbb{1}(|\xi_i| \le M) - \mathbb{E}[\xi_i \mathbb{1}(|\xi_i| \le M)] \quad \xi_{i,>M} = \xi_i \mathbb{1}(|\xi_i| > M) - \mathbb{E}[\xi_i \mathbb{1}(|\xi_i| > M)]$$

be truncated and demeaned variables. Note that

$$\xi_i = \xi_{i,<M} + \xi_{i,>M}.$$

First, let $V_{1n}(x,h_n) = \sum_{i=1}^n \ell_i(x,h_n)\xi_{i,>M}$. Note that by Cauchy–Schwarz, uniformly over x, h_n,

$$\begin{aligned} V_{1n}^2 &\le \sum_{i=1}^n \ell_i(x,h_n)^2 \sum_{i=1}^n \xi_{i,>M}^2 \\ &\lesssim \frac{1}{h_n^2}\frac{1}{n}\sum_{i=1}^n \xi_{i,>M}^2 && \text{(Lemma 1.3(i) in Tsybakov (2008) shows that } |\ell_i(x,h_n)| \le \tfrac{C}{nh_n}) \\ &\lesssim n^{2/5}\frac{1}{n}\sum_{i=1}^n \xi_{i,>M}^2 \end{aligned}$$

Now, for some C related to the implicit constant in the above display,

$$P\left(\sup_{x\in[0,1],h_n\in H_n} V_{1n}^2(x,h_n) > Ct^2\right) \le P\left(\frac{1}{n}\sum_{i=1}^n \xi_{i,>M}^2 > t^2 n^{-2/5}\right) \le \frac{\max_i \mathbb{E}\xi_{i,>M}^2}{t^2} n^{2/5}.$$

(Markov's inequality)

We note that by Cauchy–Schwarz,

$$\mathbb{E}[\xi_{i,>M}^2] \le \sqrt{\mathbb{E}[\xi_i^4]}\sqrt{P(|\xi_i| > M)} \lesssim \sqrt{P(|\xi_i| > M)} \le \exp\left(-cM^\alpha\right) \qquad \text{(Lemma D.16)}$$

where c depends on A_ξ. Hence, for a potentially different constant C,

$$P\left(\sup_{x\in[0,1],h_n\in H_n} |V_{1n}(x,h_n)| > Ct\right) \le \exp\left(-cM^\alpha - 2\log t + \frac{2}{5}\log n\right). \tag{G.6}$$

Next, consider the process

$$V_{2n}(x,h_n) = \sum_{i=1}^n \ell_i(x,h_n)\xi_{i,<M}$$

$$
\begin{aligned}
&= \frac{1}{nh_n}\sum_{i=1}^n \underbrace{u(0)'B_{nx}^{-1}\begin{bmatrix}1\\0\end{bmatrix}}_{A_1(x,h_n)} K\left(\frac{x_i-x}{h_n}\right)\xi_{i,<M} \\
&\quad + \frac{1}{nh_n}\sum_{i=1}^n \underbrace{u(0)'B_{nx}^{-1}\begin{bmatrix}0\\1\end{bmatrix}}_{A_2(x,h_n)} K\left(\frac{x_i-x}{h_n}\right)\left(\frac{x_i-x}{h_n}\right)\xi_{i,<M} \\
&\equiv \frac{A_1(x,h_n)}{h_n}\frac{1}{n}\sum_{i=1}^n K\left(\frac{x_i-x}{h_n}\right)\xi_{i,<M} + \frac{A_2(x,h_n)}{h_n}\frac{1}{n}\sum_{i=1}^n K\left(\frac{x_i-x}{h_n}\right)\left(\frac{x_i-x}{h_n}\right)\xi_{i,<M}.
\end{aligned}
$$

Note that, by Assumption G.2(1), uniformly over $x \in [0,1]$ and $h_n \in H_n$,

$$
|A_k(x,h_n)| \le \|u(0)'B_{nx}^{-1}\| \le \frac{1}{\lambda_0}.
$$

By triangle inequality,

$$
\begin{aligned}
V_{2n}(x,h_n) &\lesssim \frac{1}{h_n}\left|\frac{1}{n}\sum_{i=1}^n K\left(\frac{x_i-x}{h_n}\right)\xi_{i,<M}\right| + \frac{1}{h_n}\left|\frac{1}{n}\sum_{i=1}^n K\left(\frac{x_i-x}{h_n}\right)\left(\frac{x_i-x}{h_n}\right)\xi_{i,<M}\right| \\
&\equiv \frac{1}{\sqrt{n}h_n}V_{2n,1}(x,h_n) + \frac{1}{\sqrt{n}h_n}V_{2n,2}(x,h_n).
\end{aligned}
$$

We will aim to control the ψ_2-norm of the left-hand side. Note that it suffices to control the ψ_2-norm of both terms on the right-hand side:

$$
\left\|\sup_{x\in[0,1],h_n\in H_n}|V_{2n}(x,h_n)|\right\|_{\psi_2} \lesssim \frac{1}{\sqrt{n}h_n}\max_{k=1,2}\left(\left\|\sup_{x\in[0,1],h_n\in H_n}|V_{2n,k}(x,h_n)|\right\|_{\psi_2}\right).
$$

The above display follows from replacing the sum with two times the maximum and Lemma 2.2.2 in van der Vaart and Wellner (1996).

We will do so by applying Lemma G.1. The analogue of f in Lemma G.1 is

$$
t \mapsto f(t;x,h) = \left(\frac{t-x}{h}\right)^{k-1} K\left(\frac{t-x}{h}\right)
$$

for $V_{2n,k}, k=1,2$. Naturally, the analogues of $\mathcal{F}$ is

$$
\mathcal{F}_k = \{t \mapsto f(t;x,h) : x \in [0,1], h \in H_n\} \cup \{t \mapsto 0\}.
$$

Note that

$$
f(t;x,h) \le \mathbb{1}(|t-x| \le h)K_0
$$

and thus the diameter of $\mathcal{F}_k$ is at most

$$
\sup_{A\subset[0,1]:\lambda(A)\le 4C_h n^{-1/5}} K_0\sqrt{\frac{1}{n}\sum_{i=1}^n \mathbb{1}(x_i \in A)} \lesssim n^{-1/10}
$$

by Assumption G.2(2). Therefore, by Assumption G.3, we apply Lemma G.1 and obtain that for $k = 1, 2$

$$\left\| \sup_{x\in[0,1],h\in H_n} |V_{2n,k}(x,h)| \right\|_{\psi_2} \lesssim M n^{-1/10}\sqrt{\log n}.$$

Finally, this argument shows that

$$\left\| \sup_{x\in[0,1],h\in H_n} |V_{2n}(x,h)| \right\|_{\psi_2} \lesssim \frac{1}{\sqrt{n}h_n n^{1/10}} M\sqrt{\log n} \lesssim n^{-2/5}M\sqrt{\log n}. \tag{G.7}$$

Putting things together, we can choose $M = (c_m \log n)^{1/\alpha}$ for sufficiently large c_m so that by (G.6),

$$\mathrm{P}\left(\sup_{x\in[0,1],h\in H_n} |V_{1n}(x,h)| > Ctn^{-2/5} \right) \le \frac{a}{2}n^{-b}\frac{1}{t^2},$$

where c_m depends on a, b. The bound (G.7) in turns shows that

$$\mathrm{P}\left(\sup_{x\in[0,1],h_n\in H_n} |V_{2n}(x,h_n)| > C(a,b)t(\log n)^{\frac{2+\alpha}{2\alpha}} n^{-2/5} \right) \le 2e^{-t^2}$$

Taking $t = \sqrt{b\log n + \log(a/4)}s$ gives

$$\mathrm{P}\left(\sup_{x\in[0,1],h_n\in H_n} |V_{2n}(x,h_n)| > C(a,b)s(\log n)^{1+1/\alpha} n^{-2/5}e^{-s^2} \right) \le \frac{a}{2}n^{-b}e^{-s^2} < \frac{a}{2}n^{-b}\frac{1}{s^2}$$

for all $s > 1$.

Therefore, combining the two bounds,

$$\mathrm{P}\left(\sup_{x\in[0,1],h_n\in H_n} |v(x,h_x)| > C(a,b)t(\log n)^{1+1/\alpha} n^{-2/5} \right) \le an^{-b}\frac{1}{t^2}.$$

□

Lemma G.1. *Suppose ξ_i are bounded by $M \ge 1$ and mean zero. Consider the process*

$$V_n(f) = \frac{1}{\sqrt{n}}\sum_{i=1}^n f(x_i)\xi_i$$

over a class of real-valued functions $f \in \mathcal{F}$ and evaluation points $x_1, \ldots, x_n \in [0,1]$. Define the seminorm $\|\cdot\|_n$ relative to $x_1, \ldots, x_n$ by

$$\|f\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^n f(x_i)^2}.$$

Suppose $0 \in \mathcal{F}$ and $\mathcal{F}$ has polynomial covering numbers:

$$N(\epsilon, \mathcal{F}, \|\cdot\|_n) \le C(1/\epsilon)^V \quad \epsilon \in [0,1]$$

where $C, V > 0$ depend solely on $\mathcal{F}$. Then

$$\left\| \sup_{f \in \mathcal{F}} |V_n(f)| \right\|_{\psi_2} \lesssim M \mathrm{diam}(\mathcal{F}) \sqrt{\log(1/\mathrm{diam}(\mathcal{F}))},$$

where $\mathrm{diam}(\mathcal{F}) = \sup_{f_1, f_2 \in \mathcal{F}} \|f_1 - f_2\|_n$.

Proof. The process $V_n(f)$ has subgaussian increments with respect to $\|\cdot\|_n$:

$$\|V_n(f_1) - V_n(f_2)\|_{\psi_2} \lesssim M \|f_1 - f_2\|_n.$$

Hence, by Dudley's chaining argument (e.g. Corollary 2.2.5 in van der Vaart and Wellner (1996)), for some fixed $f_0 \in \mathcal{F}$,

$$\left\| \sup_{f} V_n(f) \right\|_{\psi_2} \leq \|V_n(f_0)\|_{\psi_2} + CM \int_0^{\mathrm{diam}(\mathcal{F})} \sqrt{\log N(\delta, \mathcal{F}, \|\cdot\|_n)}\, d\delta.$$

Note that (i) the metric entropy integral is bounded by $C\mathrm{diam}(\mathcal{F})\sqrt{\log(1/\mathrm{diam}(\mathcal{F}))}$, and (ii) for a fixed f_0, $\|V_n(f_0)\|_{\psi_2} \lesssim \|f_0\|_n M \leq \mathrm{diam}(\mathcal{F}) M$ since $0 \in \mathcal{F}$. Therefore,

$$\left\| \sup_{f} V_n(f) \right\|_{\psi_2} \lesssim M \mathrm{diam}(\mathcal{F}) \sqrt{\log(1/\mathrm{diam}(\mathcal{F}))}.$$

□

Lemma G.2 (Lemma 7.22(ii) in Sen (2018))**.** *Let $q(\cdot)$ be a real-valued function of bounded variation on $\mathbb{R}$. The covering number of $\mathcal{F} = \{x \mapsto q(ax + b) : (a, b) \in \mathbb{R}\}$ satisfies*

$$N(\epsilon, \mathcal{F}, L_2(Q)) \leq K_1 \epsilon^{-V_1}$$

for some K_1 and V_1 and for a constant envelope.