# Empirical Bayes When Estimation Precision Predicts Parameters


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



ABSTRACT. Empirical Bayes 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. 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. Census tracts selected by CLOSE are substantially more mobile on average than those selected by the standard shrinkage method.



JEL CODES. C10, C11, C44

KEYWORDS. Empirical Bayes, $g$-modeling, regret, heteroskedasticity, nonparametric maximum likelihood, Opportunity Atlas, Creating Moves to Opportunity


---


*Date*: April 8, 2024. This paper is based on the second chapter of my Ph.D. thesis. It is also 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, Bryan Graham, Jiaying Gu, Aditya Guntuboyina, 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, Azeem Shaikh, Suproteem Sarkar, Ashesh Rambachan, David Ritzwoller, Brad Ross, Jonathan Roth, Neil Shephard, Rahul Singh, Asher Spector, Harald Uhlig, Winnie van Dijk, Davide Viviano, Christopher Walker, Chris Walters, and workshop and seminar participants at Brown, Harvard, Penn State, Philadelphia Fed, Rutgers, Princeton, Stanford, the University of Chicago, Berkeley, UCLA, and Yale. 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 decisions.[1] Commonly used empirical Bayes methods assume *prior independence*—that the known precisions of the noisy estimates do not predict the underlying unknown parameters. However, prior independence is economically questionable and empirically rejected in many contexts; inappropriately imposing it can harm empirical Bayes decisions, possibly even making them underperform decisions 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., 2024) performs shrinkage on estimates of economic mobility of low-income children[2] published in the Opportunity Atlas (Chetty et al., 2020). Here, prior independence assumes that the 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 mobility. Consequently, the standard errors of mobility estimates and true mobility are positively correlated, violating prior independence.

Bergman et al. (2024) select high-mobility Census tracts by choosing those with high shrinkage estimates. Empirical Bayes methods under prior independence shrink all estimates to their *unconditional* mean, and they shrink noisier estimates more severely. However, since Census tracts with high standard errors also tend to be more mobile, their *conditional mean* given high standard errors is greater than the unconditional mean. As a result, conventional methods erroneously shrink these noisier estimates to a target that is too low. This can harm subsequent decisions: For a few measures of economic mobility where prior

[1]Empirical Bayes methods are applicable 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., 2024) 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; Aloni and Avivi, 2023). Empirical Bayes methods are also applicable in studies of discrimination (Kline et al., 2022; Kline et al., 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).

[2]Throughout this paper, measures of economic mobility for a place are defined as certain average adult outcomes of children from low-income households growing up in the place (Chetty et al., 2020). One example is the probability that a Black person have incomes in the top 20 percentiles, whose parents have household incomes at the $25^{\text{th}}$ percentile. See Section 4 for details.

independence is severely violated, we find that screening on conventional shrinkage estimates selects *less* economically mobile tracts, on average, than screening on the unshrunk estimates.[3] In contrast, screening on shrinkage estimates computed by our method selects substantially more mobile tracts.

To describe our method, let us formalize the setup and introduce empirical Bayes methods in more detail. For $i = 1, \ldots, n$, suppose we observe $(Y_i, \sigma_i)$. $Y_i$ are noisy estimates for parameters $\theta_i$, and $\sigma_i$ are the corresponding standard errors for the point estimates $Y_i$. In our empirical application, $(Y_i, \sigma_i)$ are published in the Opportunity Atlas for each Census tract $i$, and are designed to measure true economic mobility $\theta_i$. Motivated by the central limit theorem applied to the underlying micro-data—which is unavailable to the public in the case of the Opportunity Atlas—we assume that $Y_i$ is approximately Gaussian:

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

Under this setup, empirical Bayes methods are rationalized as approximations of unknown optimal decisions. If we knew the distribution of $(\theta_i, \sigma_i)$, then we can do no better than *oracle Bayes* decisions, which use the distribution of $(\theta_i, \sigma_i)$ as a prior and optimize actions with respect to the corresponding posterior distribution $\theta_i \mid \sigma_i, Y_i$. Empirical Bayes methods emulate such optimal decisions by estimating the oracle's prior—the distribution of $(\theta_i, \sigma_i)$. As an example, *shrinkage estimation*, discussed so far, corresponds to using the (estimated) posterior means of $\theta_i$ as a decision rule.

Prior independence—the assumption that $\theta_i \perp\!\!\!\perp \sigma_i$—simplifies the problem of estimating the oracle's prior. However, empirical Bayes methods based on this assumption can have poor performance when it fails to hold. In Section 2, we relax prior independence by modeling the prior distribution $\theta_i \mid \sigma_i$ flexibly. We model $\theta_i \mid \sigma_i$ as a conditional location-scale family, controlled by $\sigma_i$-dependent location and scale hyperparameters and a $\sigma_i$-independent shape hyperparameter. Under this assumption, different values of the standard errors $\sigma_i$ 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 location and scale parameters are constant functions of $\sigma_i$.

This <u>c</u>onditional <u>lo</u>cation-<u>s</u>cale assumption leads naturally to a family of <u>e</u>mpirical 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 simply estimates these parameters

[3] Fortunately, for the measure of economic mobility (mean income rank pooling over all demographic groups whose parents are at the $25^{\text{th}}$ percentile of household income) used in Bergman et al. (2024), the violation of prior independence is sufficiently mild, so that screening on these empirical Bayes shrinkage estimates still outperforms screening on the raw estimates.

flexibly 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 $\theta_i \mid \sigma_i$.

Section 3 provides three statistical guarantees for our preferred method, CLOSE-NPMLE. First and foremost, CLOSE-NPMLE emulates the oracle as well as possible, in terms of squared error loss. Specifically, Theorems 1 and 2 establish that CLOSE-NPMLE is minimax rate-optimal, up to logarithmic factors, 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 $(\theta_i, \sigma_i)$. From a technical perspective, these results extend existing regret guarantees for NPMLE-based empirical Bayes to accommodate estimated nuisance parameters.

Second, our guarantee for squared error regret also bounds the Bayes regret for two ranking-related decision problems, including the problem of selecting high-mobility tracts in Bergman et al. (2024). Theorem 3 shows that the Bayes regret in squared error dominates the Bayes regret for these decision problems. Coupled with Theorem 1, this implies that CLOSE-NPMLE has good performance for these ranking-related problems as well.

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

Lastly, since practitioners may want to assess how and whether CLOSE-NPMLE provides improvements in specific applications, Section 3.4 proposes an out-of-sample validation procedure by extending the *coupled bootstrap* in Oliveira et al. (2021). This procedure provides unbiased loss estimates for arbitrary decision rules. In particular, this procedure allows practitioners to evaluate whether CLOSE provides improvements by comparing loss estimates for CLOSE and those for other procedures.

To illustrate our method, Section 4 applies CLOSE to two empirical exercises, building on Chetty et al. (2020) and Bergman et al. (2024). The first exercise is a Monte Carlo simulation calibrated to the Opportunity Atlas. For all 15 measures of economic mobility that we consider, CLOSE-NPMLE improves over all alternatives and 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.

The second exercise evaluates the out-of-sample performance of various procedures for the selection policy problem in Bergman et al. (2024). Bergman et al. (2024) use empirical Bayes procedures to select high-mobility Census tracts in Seattle. In an exercise that mimics theirs, we find that CLOSE-NPMLE selects more economically mobile tracts than conventional methods. 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. Because the value of data is likely economically significant, the additional improvements are substantial as well.

## 2. Model and proposed method

**2.1. Empirical Bayes assumptions.** We observe estimates $Y_i$ and their (estimated) standard errors $\sigma_i$ for parameters $\theta_i$, over populations $i \in \{1, \ldots, n\}$. We maintain two assumptions that are standard in the empirical Bayes literature (Gilraine et al., 2020; Jiang, 2020; Soloff et al., 2021; Gu and Koenker, 2023; Gu and Walters, 2022).

First, we assume throughout that the estimates are conditionally Gaussian (1.1) and independent across $i$. The Gaussian assumption (1.1) is motivated by a central limit theorem applied to the underlying micro-data. Despite being standard in the empirical Bayes literature, this assumption is not without loss, as we ignore the fact that the central limit theorem only provides asymptotic approximations and instead treat the Normality as exact.[4]

Second, empirical Bayes methods estimate the distribution of $(\theta_i, \sigma_i)$. For that to be well-defined, naturally, we assume that $(\theta_i, \sigma_i)$ are sampled i.i.d. from some distribution.[5] As a minor technical perspective, throughout, we condition on $\sigma_{1:n} = (\sigma_1, \ldots, \sigma_n)$ and treat them as fixed. Thus, we think of $\theta_i$ as drawn independently but not necessarily identically:

$$\theta_i \mid \sigma_i \overset{\text{i.n.i.d.}}{\sim} G_{(i)}. \tag{2.1}$$

Let $P_0 \equiv (G_{(1)}, \ldots, G_{(n)})$ denote the conditional distribution $\theta_{1:n} \mid \sigma_{1:n}$.

[4]Following the empirical Bayes literature, this paper abstracts away from the micro-data and treats the estimates $(Y_i, \sigma_i)$ as primitives. See Remark 2 for a concrete example on how $Y_i, \sigma_i$ are related to the micro-data and whether (1.1) is appropriate.

[5]This is analogous to (correlated) random effects in panel data.

Some further discussions of these assumptions—e.g., on the Gaussian setup (Remark 2), the independence of $(Y_i, \theta_i, \sigma_i)$ across $i$ (Remark 3), and the role of additional covariates $X_i$ (Remark 4)—are deferred to Section 2.5.

Under these assumptions, empirical Bayes methods are desirable for decision making: They approximate optimal but infeasible decision rules. To see this, consider a decision problem with loss function $L(\boldsymbol{\delta}, \theta_{1:n})$ that evaluates an action $\boldsymbol{\delta}$ at a vector of parameters $\theta_{1:n}$. The optimal decision—in terms of expected loss $\mathbb{E}_{P_0}[L(\cdot, \theta_{1:n}) \mid \sigma_{1:n}]$ over $(Y_i, \theta_i) \mid \sigma_i$—is the *oracle Bayes decision rule* $\boldsymbol{\delta}^\star$. At a realization $Y_{1:n}$, $\boldsymbol{\delta}^\star$ chooses actions that minimize the posterior expected loss under the *oracle prior* $P_0$:

$$\boldsymbol{\delta}^\star(Y_{1:n}, \sigma_{1:n}; P_0) \in \arg\min_{\boldsymbol{\delta}} \mathbb{E}_{P_0}[L(\boldsymbol{\delta}, \theta_{1:n}) \mid Y_{1:n}, \sigma_{1:n}]. \tag{2.2}$$

Unfortunately, $\boldsymbol{\delta}^\star$ is infeasible since we do not know $P_0$. To remedy, empirical Bayes methods seek to approximate the oracle Bayes rule $\boldsymbol{\delta}^\star$. Naturally, one recipe for generating empirical Bayes decision rules is to plug an estimate $\hat{P}$ for $P_0$ into (2.2):[6]

$$\boldsymbol{\delta}_{\mathrm{EB}}(Y_{1:n}, \sigma_{1:n}; \hat{P}) \in \arg\min_{\boldsymbol{\delta}} \mathbf{E}_{\hat{P}}[L(\boldsymbol{\delta}, \theta_{1:n}) \mid Y_{1:n}, \sigma_{1:n}]. \tag{2.3}$$

For the leading decision problem where $L(\cdot, \cdot)$ is mean-squared error, the recipe (2.3) generates empirical Bayes posterior means—posterior means under $\hat{P}$. They are often referred to as *shrinkage estimates*.[7]

2.2. **Prior independence and its violation.** To simplify $P_0$, popular empirical Bayes methods often assume *prior independence*: $\theta_i \perp\!\!\!\perp \sigma_i$, or, equivalently, $G_{(1)} = \cdots = G_{(n)} \equiv G_{(0)}$.

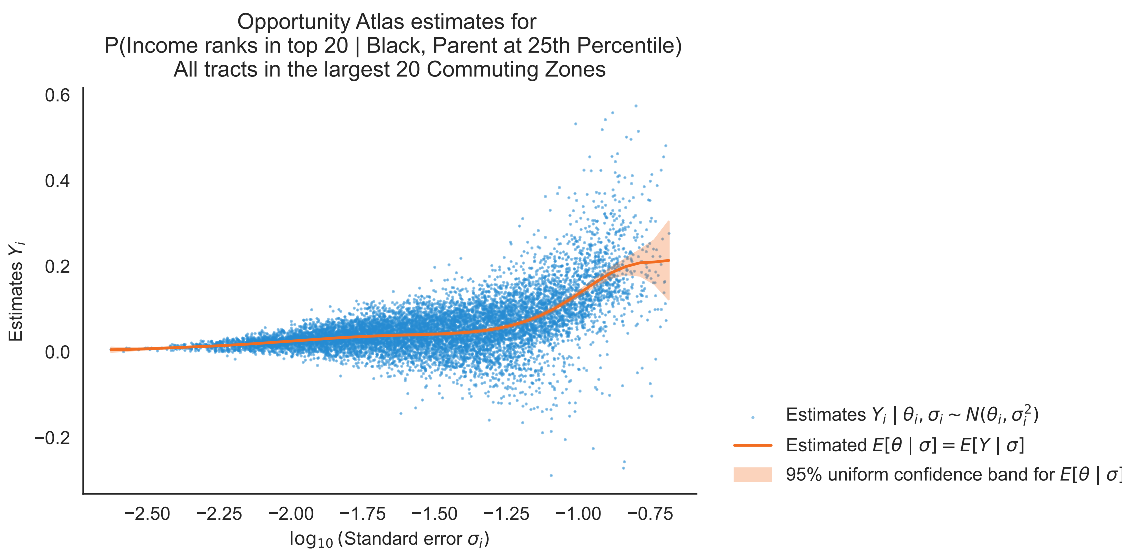
For instance, the standard parametric empirical Bayes method models $G_{(i)}$ as i.i.d. Gaussian, $G_{(0)} \sim \mathcal{N}(m_0, s_0^2)$ (Morris, 1983). Following the recipe (2.3), this approach estimates $P_0$ by estimating its mean and variance $(m_0, s_0^2)$. Henceforth, we shall refer to this method as INDEPENDENT-GAUSS. State-of-the-art empirical Bayes methods (Jiang, 2020; Gilraine et al., 2020; Soloff et al., 2021) relax the parametric assumptions on $G_{(0)}$ and estimate $G_{(0)}$ with *nonparametric maximum likelihood* (NPMLE). We refer to this method as INDEPENDENT-NPMLE. The "INDEPENDENT" here emphasizes that they assume prior independence.

[6]To emphasize the distinction between the true expectation with respect to the data-generating process (2.1) 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.

[7]A complementary view of shrinkage estimation and compound decisions, dating to James and Stein (1961) and Robbins (1956), does not impose (2.1) and instead views $\theta_{1:n}$ as fixed or conditioned upon. In this paper, however, we do impose (2.1) and refer to empirical Bayes posterior means and shrinkage estimates interchangeably.

Despite being convenient for estimating $P_0$, prior independence may be economically implausible, statistically rejected, and even decision-theoretically harmful. We illustrate this with an empirical application to the Opportunity Atlas. There, one published measure of economic mobility defines the economic mobility of a Census tract ($\theta_i$) as the probability that a relatively poor Black child from tract $i$ grows up to be relatively rich. More precisely, $\theta_i$ is 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.

Intuitively, Census tracts with more low-income Black households should have *more precise* estimates of $\theta_i$, simply because there is a larger sample size to estimate $\theta_i$. However, we might also expect that these Census tracts are on average poorer and are less likely to generate favorable outcomes for low-income Black children than wealthier tracts. Thus, these Census tracts should have smaller $\sigma_i$ but also lower $\theta_i$, meaning that $(\sigma_i, \theta_i)$ are positively correlated.




*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 implemented by 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. See Appendix SM8 for details on estimating conditional moments of $\theta_i$ given $\sigma_i$. □

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

As this economic intuition predicts, prior independence is readily rejected for this measure of economic mobility. Figure 1 plots $Y_i$—the Opportunity Atlas estimates of $\theta_i$—against their standard errors ($\log_{10}(\sigma_i)$). Figure 1 also displays an 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 $\theta_i$ were independent of $\sigma_i$, then the true 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 mobility.[8]

What happens if we apply empirical Bayes methods that assume prior independence here? Figure 2 overlays empirical Bayes posterior means on the scatterplot. In the top panel, INDEPENDENT-GAUSS shrinks estimates $Y_i$ towards a common estimated mean $\hat{m}_0$, depicted as the black line. When $\sigma_i$ and $\theta_i$ are positively correlated—as is the case here—estimated posterior means under INDEPENDENT-GAUSS systematically undershoot $\theta_i$ for populations with imprecise estimates. Similarly, the middle panel of Figure 2 shows that INDEPENDENT-NPMLE suffers from the same undershooting. 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. On average, these high-mobility tracts are exactly those with high $\sigma_i$. Shrinking these tracts severely towards the estimated common mean leads to suboptimal selections that may even underperform screening directly based on $Y_i$.[9]

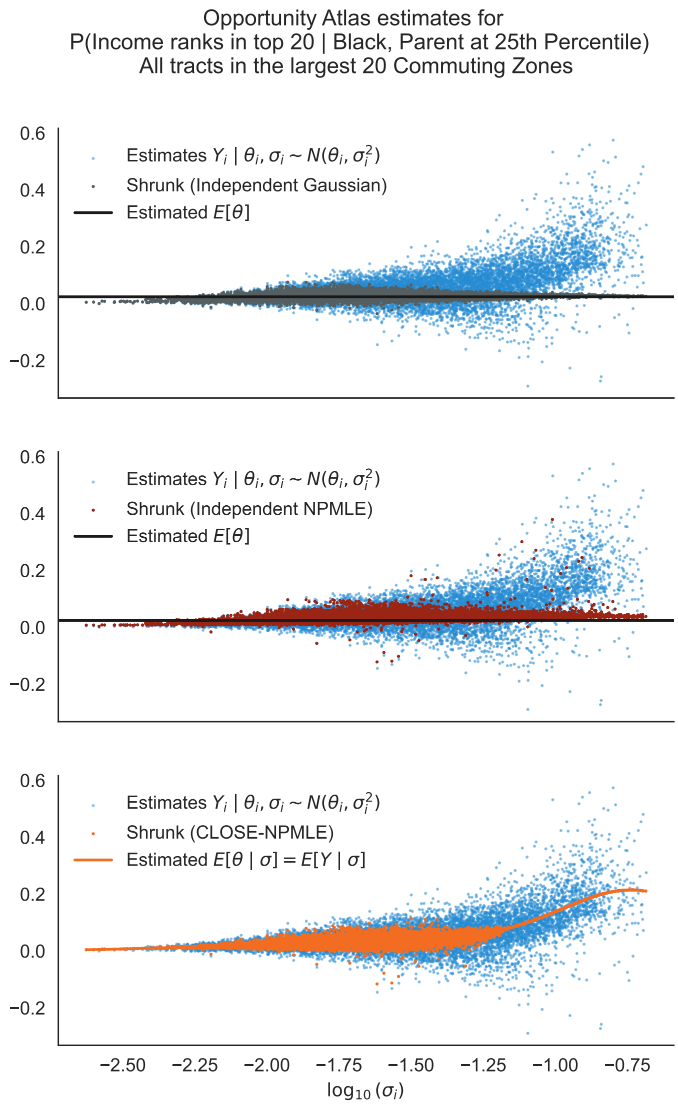
2.3. **Conditional location-scale relaxation of prior independence.** To remedy these issues, we propose the following *conditional location-scale model* as a relaxation:

$$\mathrm{P}\left(\theta_i \le t \mid \sigma_i\right) = G_0\left(\frac{t - m_0(\sigma_i)}{s_0(\sigma_i)}\right), \quad \eta_0(\cdot) \equiv (m_0(\cdot), s_0(\cdot)) \tag{2.4}$$

where $G_0$ is normalized to have mean zero and variance one. Equation (2.4) states that the conditional distribution of $\theta_i \mid \sigma_i$ follows a location-scale family, controlled by $\sigma_i$-dependent location parameter $m_0(\cdot)$ and scale parameter $s_0(\cdot)$. By the normalization of $G_0$, we can think of $m_0(\cdot)$ as the conditional mean of $\theta_i \mid \sigma_i$ and $s_0^2(\cdot)$ as the conditional variance. To be clear, despite generalizing prior independence, (2.4) is still restrictive: We

[8]Moreover, $\log \sigma_i$ remains predictive of $Y_i$ even if we residualize $Y_i$ against a vector of tract-level covariates (Figure OA5.8). We note that this definition of $\theta_i$ is not the measure used in Bergman et al. (2024). Prior independence is also readily rejected for the mobility measure used in Bergman et al. (2024), but its violation is not as severe once adjusted for tract-level covariates (see Section 4 and Figure OA5.7).

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




*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 $1/\sigma_i^2$ following Bergman et al. (2024). □

FIGURE 2. Posterior mean estimates under prior independence

discuss a rationale of this assumption—as well as comparing it to some alternatives—in Remarks 6 and 7; Theorem 4 provides additional robustness assurances.[10]

[10]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

Applying the empirical Bayes recipe (2.3) amounts to estimating $P_0$. Under (2.4), it suffices to estimate the unknown hyperparameters $(\eta_0, G_0)$. Estimating $\eta_0 = (m_0, s_0)$ is relatively straightforward, as $\eta_0$ can be written as conditional moments of $Y_i$ as a function of $\sigma_i$:

$$m_0(\sigma) = \mathbb{E}[\theta \mid \sigma] = \mathbb{E}[Y \mid \sigma] \text{ and } s_0^2(\sigma) = \mathrm{Var}(\theta \mid \sigma) = \mathbb{E}[(Y - m_0(\sigma))^2 \mid \sigma] - \sigma^2. \quad (2.5)$$

Estimating these conditional moments reduces to nonparametric regression, for which convenient off-the-shelf methods exist (e.g., Calonico et al. (2019) and Appendix SM8).

Estimating $G_0$ is more complicated. To introduce our procedure, we can write (2.4) equivalently as the following representation with transformed parameters $\tau_i \equiv \frac{\theta_i - m_0(\sigma_i)}{s_0(\sigma_i)}$ and similarly transformed estimates $Z_i \equiv \frac{Y_i - m_0(\sigma_i)}{s_0(\sigma_i)}$ and $\nu_i \equiv \sigma_i / s_0(\sigma_i)$:

$$\begin{aligned} \theta_i &= m_0(\sigma_i) + s_0(\sigma_i)\tau_i & \tau_i \mid \sigma_i &\overset{\text{i.i.d.}}{\sim} G_0 \\ Y_i &= m_0(\sigma_i) + s_0(\sigma_i) Z_i & Z_i \mid \tau_i, \nu_i &\sim \mathcal{N}(\tau_i, \nu_i^2). \end{aligned} \quad (2.6)$$

The representation (2.6) makes clear that, first, the transformed triplet $(Z_i, \tau_i, \nu_i)$ obeys an analogue of the Gaussian model (1.1), where $Z_i$ is a noisy Gaussian signal on $\tau_i$ with variance $\nu_i^2$. Second, prior independence holds in this model, in terms of the transformed parameters, as $\tau_i \mid \nu_i \overset{\text{i.i.d.}}{\sim} G_0$.

This observation motivates the following strategy: We first transform $(Y_i, \sigma_i)$ into $(Z_i, \nu_i)$; we then apply empirical Bayes methods that assume prior independence on these transformed quantities to estimate $G_0$. Indeed, given estimates $\hat{m}$ and $\hat{s}$ of $m_0(\cdot)$ and $s_0(\cdot)$, we can then form the estimated transformed data $\hat{Z}_i, \hat{\nu}_i$ as

$$\hat{Z}_i = \frac{Y_i - \hat{m}(\sigma_i)}{\hat{s}(\sigma_i)} \quad \text{and} \quad \hat{\nu}_i = \frac{\sigma_i}{\hat{s}(\sigma_i)}. \quad (2.7)$$

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:

**CLOSE–STEP 1** Nonparametrically estimate $m_0(\sigma), s_0^2(\sigma)$ according to (2.5).

**CLOSE–STEP 2** With the estimates $\hat{\eta} = (\hat{m}, \hat{s})$, transform the data according to (2.7). 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 $(\hat{\eta}, \hat{G}_n)$ and hence having obtained $\hat{P}$, we then form empirical Bayes decision rules following (2.3).

additional parametric restrictions on $s(\cdot)$) and George et al. (2017) impose the location scale model (2.4) with $G_0 \sim \mathcal{N}(0, 1)$ (as well as additional parametric restrictions on $s_0(\cdot), m_0(\cdot)$).

This framework produces a family of empirical Bayes strategies, since $\boxed{\text{CLOSE–STEP 2}}$ 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 $\hat{Z}_i$, whose marginal distribution is the convolution $G_0 \star \mathcal{N}(0, \hat{\nu}_i^2)$, treating $\hat{m}, \hat{s}$ as fixed: For $\varphi(\cdot)$ the Gaussian probability density function and $\mathcal{P}(\mathbb{R})$ the set of all distributions supported on $\mathbb{R}$, we maximize

$$\hat{G}_n \in \operatorname*{arg\,max}_{G \in \mathcal{P}(\mathbb{R})} \frac{1}{n} \sum_{i=1}^{n} \log \int_{-\infty}^{\infty} \varphi\left(\frac{\hat{Z}_i - \tau}{\hat{\nu}_i}\right) \frac{1}{\hat{\nu}_i} G(d\tau). \tag{2.8}$$

When the estimated moments $\hat{m}, \hat{s}$ are constant functions of $\sigma$, 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 properties.[11] Our subsequent results in Section 3 extend some of these favorable properties to CLOSE-NPMLE under (2.4).

The next subsection introduces three decision problems that our theory shall focus on. We defer to Section 2.5 some discussions on the implementation of CLOSE-NPMLE (Remark 5), the rationale of (2.4) (Remark 6), alternatives to CLOSE (Remark 7), and another method in the CLOSE-family (Remark 8).

2.4. **Decision problems.** To prepare for our theoretical results in Section 3, we review 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*. Recall that $L(\boldsymbol{\delta}, \theta_{1:n})$ denotes a *loss function* mapping actions and parameters to a scalar. Let $R_{\mathrm{B}}(\boldsymbol{\delta}; P_0) = \mathbb{E}_{P_0}[L(\boldsymbol{\delta}, \theta_{1:n}) \mid \sigma_{1:n}]$ be the *Bayes risk* of $\boldsymbol{\delta}$ under $P_0$.

The oracle Bayes decision rule $\boldsymbol{\delta}^\star$ (2.2) is optimal in the sense that it minimizes $R_{\mathrm{B}}$. Thus, a natural performance measure for the empirical Bayesian (2.3)—who tries to mimic the oracle Bayesian by estimating $P_0$—is the gap between the Bayes risks of $\boldsymbol{\delta}_{\mathrm{EB}}$ and $\boldsymbol{\delta}^\star$.

[11] 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), and 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 and lack of tuning parameters are advocated in Koenker and Gu (2019), Koenker and Mizera (2014), and Koenker and Gu (2017).
One might also consider a sieve likelihood estimator for $G_0$ à la Efron (2016). We conjecture that by an argument analogous to ours, one could likewise obtain regret guarantees for this sieve likelihood approach as we do for NPMLE, though we are unfamiliar with such results even in the homoskedastic setting.

We refer to this quantity as *Bayes regret*:

$$\begin{aligned}\mathrm{BayesRegret}_n(\boldsymbol{\delta}_{\mathrm{EB}}) &= R_{\mathrm{B}}(\boldsymbol{\delta}_{\mathrm{EB}}; P_0) - R_{\mathrm{B}}(\boldsymbol{\delta}^\star; P_0) \\ &= \mathbb{E}_{P_0}[L(\boldsymbol{\delta}_{\mathrm{EB}}, \theta_{1:n}) - L(\boldsymbol{\delta}^\star, \theta_{1:n}) \mid \sigma_{1:n}] \end{aligned} \tag{2.9}$$

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.[12]

We introduce a few concrete decision problems by specifying the actions $\boldsymbol{\delta}$ and loss functions $L$ and state the corresponding oracle 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 $\delta_i$ for parameters $\theta_i$, evaluated with MSE: $L(\boldsymbol{\delta}, \theta_{1:n}) = \frac{1}{n}\sum_{i=1}^n (\delta_i - \theta_i)^2$. 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 $\theta_i^* = \theta_{i,P_0}^* \equiv \mathbb{E}_{P_0}[\theta_i \mid Y_i, \sigma_i]$ 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 economically relevant, such as replacing low value-added teachers and recommending high-mobility tracts (Gilraine et al., 2020; Bergman et al., 2024).[13]

**Decision Problem 2** (UTILITY MAXIMIZATION BY SELECTION)**.** Suppose $\boldsymbol{\delta}$ consists of binary selection decisions $\delta_i \in \{0, 1\}$. For each population, selecting that population has net benefit $\theta_i$. The decision maker wishes to maximize utility (i.e., negative loss): $-L(\boldsymbol{\delta}, \theta_{1:n}) = \frac{1}{n}\sum_{i=1}^n \delta_i \theta_i$. The oracle Bayes rule selects all whose posterior mean net benefit $\theta_i$ is nonnegative: $\delta_i^\star = \mathbb{1}\left(\theta_{i,P_0}^* \geq 0\right)$. One natural empirical Bayes decision rule replaces $\theta_{i,P_0}^*$ with $\theta_{i,\hat{P}}^*$, following (2.3).[14] ■

---

[12]Bayes regret is likewise the focus of the empirical Bayes literature 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 $R_{\mathrm{F}} = \mathbb{E}[L(\boldsymbol{\delta}, \theta_{1:n}) \mid \sigma_{1:n}, \theta_{1:n}]$, but limit attention to squared error and to a restricted class of methods.

[13]We analyze these problems from a decision-theoretic perspective imposing the sampling assumption (2.1). For a different and complementary perspective in terms of frequentist hypothesis testing without imposing (2.1), see Mogstad et al. (2020) and Mogstad et al. (2023). For additional ranking-related decision problems, see Gu and Koenker (2023).

[14]In a context where the parameters are conditional average treatment effects for a particular covariate cell, $\theta_i = \mathrm{CATE}(i) \equiv \mathbb{E}[Y(1) - Y(0) \mid X = i]$, and $\delta_i$ are treatment decisions, this problem is an instance of welfare maximization by treatment choice (Manski, 2004; Stoye, 2009; Kitagawa and Tetenov, 2018; Athey

**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 $\theta_i$ of the selected set:

$$-L(\boldsymbol{\delta}, \theta_{1:n}) = \frac{1}{m}\sum_{i=1}^{n} \delta_i \theta_i. \tag{2.10}$$

Oracle Bayes selects the populations corresponding to the $m$ largest posterior means $\theta^*_{i,P_0}$ (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 (2.3) replaces $P_0$ with the estimate $\hat{P}$. ■

**Remark 1.** The utility function (2.10) rationalizes the widespread practice of screening based on empirical Bayes posterior means (Gilraine et al., 2020; Chetty et al., 2014; Kane and Staiger, 2008; Hanushek, 2011; Bergman et al., 2024). In Bergman et al. (2024), for instance, where housing voucher holders are incentivized to move to Census tracts selected according to economic mobility, (2.10) represents the expected economic mobility of a mover were they to move randomly to one of the selected tracts. Our theoretical results can accommodate slightly less restrictive mover behavior (Remark B.1). ■

2.5. **Discussions.** Before presenting our theoretical results in Section 3, we end this section with several self-contained remarks.

**Remark 2** (Gaussian approximation and estimated standard errors)**.** In our empirical application, the $\sigma_i$'s are the estimated standard errors that Chetty et al. (2020) publish. We might object to (1.1) treating $\sigma_i$ as the true standard deviation of $Y_i$. This should be viewed as an objection to the quality of the Gaussian approximation in (1.1).

To be more concrete, consider the following example that specifies how the $Y_i$'s are generated. Suppose $\theta_i = \mathbb{E}_{Q_i}[Y_{ij}]$ is the population mean of some variable $Y_{ij} \sim Q_i$ drawn from population $Q_i$. Suppose $Y_i$ is the sample mean of $Y_{i1}, \ldots, Y_{in_i} \overset{\text{i.i.d.}}{\sim} Q_i$ and suppose $\sigma_i$ is the estimated standard error for $Y_i$. Then, standard large-sample approximations yield that for large $n_i$, approximately, $\sigma_i^{-1}(Y_i - \theta_i) \sim \mathcal{N}(0, 1)$. Taking this approximation as exact yields $Y_i \mid \theta_i, \sigma_i \sim \mathcal{N}(\theta_i, \sigma_i^2)$. Note that this approximation is valid even though $\sigma_i$ is the estimated standard error.

and Wager, 2021). In this setting, $\delta_i$ is a decision to treat individuals with covariate values in the $i^{\text{th}}$ cell. The average benefit of treating these individuals is their conditional average treatment effect $\theta_i$. However, we note that the literature on treatment choice uses a different notion of regret compared to this paper (based on $R_{\mathrm{F}} = \mathbb{E}[L(\boldsymbol{\delta}, \theta_{1:n}) \mid \sigma_{1:n}, \theta_{1:n}]$ rather than $R_{\mathrm{B}}$).

Thus, if we were comfortable with treating the Gaussian approximation as exact, we should then also be comfortable with treating the estimated standard error as the true standard deviation of $Y_i$. Assessing the robustness of empirical Bayes methods to the failure of the Gaussian approximation is an important avenue for future work, though beyond the scope of this paper. ■

**Remark 3** (Interpretation when $(Y_i, \theta_i, \sigma_i)$ are dependent across $i$)**.** The assumptions (1.1) and (2.1) imply that $(Y_i, \theta_i, \sigma_i)$ are i.i.d. across $i$. This may be violated when the measurements $(Y_i, Y_j)$ are correlated conditional on $\theta_{1:n}, \sigma_{1:n}$, or when $((\theta_i, \sigma_i), (\theta_j, \sigma_j))$ are posited to be correlated in the underlying sampling process for the heterogeneous populations.

When $(Y_i, \theta_i, \sigma_i)$ are not i.i.d., consider a *separable* decision problem—in the sense that $L(\boldsymbol{\delta}, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \ell(\delta_i, \theta_i)$. Here, the oracle Bayes decision rule (2.2) that erroneously assumes independence across $i$—and optimizes decision with respect to the distribution $\theta_i \mid \sigma_i, Y_i$ instead of $\theta_i \mid \sigma_{1:n}, Y_{1:n}$—is nonetheless the best *separable* decision. That is, this erroneous oracle rule minimizes expected loss among all decision rules that make decisions solely using information on the $i^{\text{th}}$ population: $\delta_i(Y_{1:n}, \sigma_{1:n}) = \delta_i(Y_i, \sigma_i)$. Thus, when $(Y_i, \theta_i, \sigma_i)$ are correlated, we may view empirical Bayes under (1.1) and (2.1) as approximating the best separable decision rule, though our statistical guarantees characterizing the quality of approximation are no longer applicable. ■

**Remark 4** (Role of covariates $X_i$)**.** In many settings, we additionally have access to covariates $X_i$ that do not predict the noise in $Y_i$ (i.e., $Y_i \mid X_i, \theta_i, \sigma_i \sim \mathcal{N}(\theta_i, \sigma_i^2)$). When $\theta_i \mid X_i, \sigma_i$ depends substantially on $X_i$, decisions that ignore this dependence on $X_i$ may be similarly suboptimal. Our model (2.4) can easily adapt to accommodate covariates by making $m_0(\cdot)$ and $s_0(\cdot)$ depend on $X_i$. Our subsequent theoretical results also extend naturally.

Nevertheless, this paper primarily focuses on the dependence with respect to $\sigma_i$ because $\sigma_i$ is in a sense a special covariate. First, $\sigma_i$ is always present in any Gaussian empirical Bayes setting. Second, crucially, the likelihood $Y_i \mid X_i, \theta_i, \sigma_i$ depends on $\sigma_i$ but not on $X_i$. As a result, marginalizing out $X_i$ does not change the statistical structure of $Y_i$, but marginalizing out $\sigma_i$ does.

This observation means that we can often afford to be more cavalier with respect to $X_i$—as long as we are satisfied with a procedure that approximates some *constrained* oracle Bayesian who is oblivious about features of $X_i$.[15] To be clear, these constrained oracle Bayesians make worse decisions than the oracle Bayesian who fully knows $\theta_i \mid X_i, \sigma_i$.

[15]For instance, empirical Bayes procedures that ignore $X_i$ entirely can be rationalized as approximating the constrained oracle who does not have access to $X_i$. Similarly, procedures—which we employ in Section 4—that linearly residualize $Y_i$ against $X_i$ and perform empirical Bayes on the projection residuals are approximations of constrained oracles who only have access to the projection residuals.

Still, their decisions outperform naive decisions using $Y_i$. In contrast, because $\sigma_i$ enters the likelihood, all oracle Bayesians have access to $\sigma_i$, and therefore empirical Bayesians must model $\theta_i \mid \sigma_i$ accurately. ■

**Remark 5** (Practical issues when implementing CLOSE-NPMLE)**.** We highlight two issues when implementing CLOSE-NPMLE, both of which have default solutions in our accompanying software. First, analogue estimators for $s_0^2(\sigma_i) = \mathrm{Var}(Y_i \mid \sigma_i) - \sigma_i^2$ may take negative values.[16] In our experience, truncating $\hat{s}$ at zero does not seem to cause bad performance when computing posterior means. Nevertheless, in Appendix SM8 and the software implementation, we propose a heuristic but data-driven truncation rule that produces strictly positive $\hat{s}$, borrowing from a statistics literature on estimating non-centrality parameters for non-central $\chi^2$ distributions (Kubokawa et al., 1993).Second, in optimizing the NPMLE objective (2.8), following Koenker and Mizera (2014) and Koenker and Gu (2017), we approximate the set of all distributions $\mathcal{P}(\mathbb{R})$ with the set of all probability mass functions over a finite set of grid points. These grid points need to be chosen, though, theoretically, the only downside of a finer grid is computational burden. Ideally, adjacent grid points should have a sufficiently small and economically insignificant gap between them.[17] ■

**Remark 6** (Rationale for the location-scale assumption)**.** Our strategy CLOSE is certainly not the only method to relax prior independence. Our assumption (2.4) is motivated in part by a desire to take advantage of the appealing theoretical and computational properties of INDEPENDENT-NPMLE, while mitigating the bad consequences of imposing prior independence. Subjected to this goal, there is a sense in which (2.4) is maximally flexible (see Appendix OA4.1.2 for details). Suppose we consider the class of procedures that transform $(Y_i, \sigma_i)$ in some way, and then apply INDEPENDENT-NPMLE to the transformed estimates. In order to preserve conditional Normality of the transformed estimates, we are limited to considering affine transformations of $Y_i$. Thus, if the post-transformation estimates were to satisfy prior independence, then (2.4) is the most flexible specification we can allow. ■

**Remark 7** (Alternatives to CLOSE)**.** Among alternative relaxations of prior independence, which we discuss at length in Appendix OA4.1, one may seem particularly natural (Gu and Koenker, 2017; Fu et al., 2020). In many cases, the sample-size-scaled estimated standard error $S_i^2 = n_i \sigma_i^2$ is an estimator for some population variance $\sigma_{i0}^2$. Note that, under a

[16]The negative estimated variance phenomenon is in part caused by estimation noise in $\mathrm{Var}(Y_i \mid \sigma_i)$. However, in our empirical application, there is some evidence that observations with large estimated $\sigma_i$'s are underdispersed for the measures of economic mobility in the Opportunity Atlas (see Appendix OA5.1.)

[17]Since the true prior $G_0$ for $\tau_i$ 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 $[\min_i \hat{Z}_i, \max_i \hat{Z}_i] \setminus [-6, 6]$ (e.g., 100 equally spaced grid points), performs well.

Gaussian approximation similar to (1.1),

$$\begin{bmatrix} Y_i \\ S_i^2 \end{bmatrix} \mid \Sigma_i \sim \mathcal{N}\left(\begin{bmatrix} \theta_i \\ \sigma_{i0}^2 \end{bmatrix}, \Sigma_i\right)$$

for some known or estimable $\Sigma_i$. One could perform bivariate empirical Bayes by estimating the bivariate distribution of $(\theta_i, \sigma_{i0}^2)$ flexibly. This accounts for the dependence between $\theta_i$ and the population variance $\sigma_{i0}^2$. However, it remains possible that $(\theta_i, \sigma_{i0}^2) \mid \Sigma_i$ depends on $\Sigma_i$, which is often a function of the sample sizes $n_i$ underlying $(Y_i, S_i^2)$. Thus, though this bivariate formulation controls some channels of how $\sigma_i$ predicts $\theta_i$, prior independence issues still apply. ■

**Remark 8** (CLOSE-GAUSS, a "lite" version of CLOSE-NPMLE)**.** If we assume that the shape parameter $G_0$ is Gaussian, the oracle Bayes posterior means are simply

$$\theta^*_{i,\mathcal{N}(0,1),\eta_0} = \frac{\sigma_i^2}{s_0^2(\sigma_i) + \sigma_i^2} m_0(\sigma_i) + \frac{s_0^2(\sigma_i)}{s_0^2(\sigma_i) + \sigma_i^2} Y_i. \tag{2.11}$$

Equation (2.11) is the conditional-on-$\sigma_i$ analogue of posterior means under INDEPENDENT-GAUSS. Despite being rationalized under the assumption $\theta_i \mid \sigma_i \sim \mathcal{N}(m_0(\sigma_i), s_0^2(\sigma_i))$, the oracle (2.11) enjoys strong robustness properties even without (2.4). First, (2.11) is the optimal decision rule for estimating $\theta_i$ when we restrict to the class of decision rules that are linear in $Y_i$ (Weinstein et al., 2018). Second, (2.11) is minimax in the sense that it minimizes the worst-case mean squared error over choices of $G_{(1)}, \ldots, G_{(n)}$.[18]

However, the Normality assumption does imply that (2.11), unlike CLOSE-NPMLE, fails to approximate the optimal decision (2.2) when the location-scale assumption (2.4) holds. Since we also show a sense in which CLOSE-NPMLE is worst-case robust (Theorem 4), we recommend CLOSE-NPMLE over CLOSE-GAUSS, unless the researcher is extremely concerned about the misspecification of the location-scale model. ■

## 3. Theoretical results

We observe $(Y_i, \sigma_i)_{i=1}^n$, where $(\theta_i, \sigma_i)$ satisfies the location-scale assumption (2.4) and $(Y_i, \theta_i, \sigma_i)$ obeys the Gaussian location model (1.1). Our recommended procedure, CLOSE-NPMLE, transforms the data $(Y_i, \sigma_i)$ 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 (2.8) on $(\hat{Z}_i, \hat{\nu}_i)_{i=1}^n$.

[18]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 $\eta_0(\sigma_i)$.

Our leading result shows that CLOSE-NPMLE mimics the oracle Bayesian as well as possible, for the problem of estimation under squared error loss (Decision Problem 1), in the sense that its Bayes regret (2.9) 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 for Decision Problem 1, then they likewise have low regret for Decision Problems 2 and 3.

Since our main results assume the location-scale model, one may be concerned about its potential misspecification. Theorem 4 bounds the worst-case Bayes risk of an idealized version of CLOSE-NPMLE (i.e. with known $\eta_0$ and fixed but misspecified $\hat{G}_n$) as a multiple of a notion of minimax risk, without assuming (2.4). Thus, even under misspecification, CLOSE-NPMLE does not perform arbitrarily badly relative to the minimax procedure. Lastly, Section 3.4 introduces a validation procedure that produces out-of-sample estimates of the performance of arbitrary decision procedures, without assuming any additional structure beyond Normality (1.1), which is helpful for evaluating various empirical Bayes methods in practice.

**Remark 9** (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 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 $n$.[19] 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 $\theta_{1:n} \mid \sigma_{1:n} \sim P_0$, and omitting the subscript $P_0$ or the conditioning on $\sigma_{1:n}$. ■

3.1. **Regret rate in squared error.** Define $\mathrm{MSERegret}_n$ as the excess loss of the empirical Bayes posterior means relative to that of the oracle Bayes posterior means:

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

where $\theta_i^*$ are the oracle posterior means and $\hat{\theta}_{i,G,\eta}$ are the posterior means under a prior parametrized by $(G, \eta)$. The corresponding Bayes regret (2.9) for CLOSE-NPMLE is then

[19] In logical statements, appearances of $\lesssim$ implicitly prepend "there exists a universal constant" to the statement. 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 $n$, $\mathrm{P}(A_n \leq CB_n) \geq c_0$."

the $P_0$-expectation of $\mathrm{MSERegret}_n$:

$$\mathrm{BayesRegret}_n = \mathbb{E}\left[\mathrm{MSERegret}_n(\hat{G}_n, \hat{\eta})\right] = \mathbb{E}_{P_0}\left[\frac{1}{n}\sum_{i=1}^n (\theta_i^* - \hat{\theta}_{i,\hat{G}_n,\hat{\eta}})^2\right]. \tag{3.1}$$

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 $\theta_i^*$.

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 $(C, \beta)$ that depend solely on $\mathcal{P}_0$, the Bayes regret in squared error decays at the same rate as $\|\hat{\eta} - \eta_0\|_\infty^2 \equiv \max\left(\|\hat{m} - m_0\|_\infty, \|\hat{s} - s_0\|_\infty\right)^2$:

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

This result continues a recent statistics literature on empirical Bayes methods via NPMLE by characterizing the effect of an estimated nuisance parameter $\hat{\eta}$.[20]

Moreover, we show that controlling the Bayes regret is no easier than estimating $m$ in $\|\cdot\|_2^2$, which is a corresponding lower bound on regret. There exists $c$ such that for any estimator of $\theta_i$, its worst-case regret is bounded below[21]

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

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.[22]

---

[20]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 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.

[21]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.

[22]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).

**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* (2.8), *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* $\kappa_n$

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

*Moreover, we require that* $\hat{G}_n$ *has support points within* $[\min_i \hat{Z}_i, \max_i \hat{Z}_i]$. *To ensure that* $\kappa_n$ *is positive, we assume that* $n \geq 7 = \lceil \sqrt{2\pi e} \rceil$.[23]

We now state further assumptions on $\mathcal{P}_0$ beyond (2.4). First, we assume that $G_0$ is sufficiently thin-tailed for its moments to grow slowly.[24]

**Assumption 2.** *The distribution* $G_0$ *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* $p > 0$,

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

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:* $\sigma_i \in (\sigma_\ell, \sigma_u)$ *and* $s_0(\cdot) \in (s_\ell, s_u)$, *where* $\sigma_\ell, \sigma_u, s_{0\ell}, s_{0u} > 0$.

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

**Assumption 4.** *Let* $C_{A_1}^p([\sigma_\ell, \sigma_u])$ *be the Hölder class of order* $p \geq 1$ *with maximal Hölder norm* $A_1 > 0$ *supported on* $[\sigma_\ell, \sigma_u]$.[25] *We assume that*

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

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

[24] 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 $G_0$ 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).

[25] We follow the definition of Hölder classes in van der Vaart and Wellner (1996), Section 2.7.1.

*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) is uniformly bounded $\sup_{f \in \mathcal{V}} \|f\|_\infty \le C_{A_1}$ and (ii) admits the entropy bound $\log N(\epsilon, \mathcal{V}, \|\cdot\|_\infty) \le C_{A_1, p, \sigma_\ell, \sigma_u} (1/\epsilon)^{1/p}$.*

*We assume that the estimators for $m_0$ and $s_0$, $\hat\eta = (\hat m, \hat s)$, satisfy:*

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

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

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

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

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 uniform entropy behavior as the Hölder class $C^p_{A_1}([\sigma_\ell, \sigma_u])$.[26]

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 uniform for a fixed $\mathcal{H}$.

3.1.2. *MSE 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 are accurate within some radius in $\|\cdot\|_\infty$. Our main result bounds the Bayes regret in MSE, assuming that some $C_{1,\mathcal{H}}$ can be chosen so that $\mathbf{A}_n(C_{1,\mathcal{H}})$ has sufficiently high probability.[27]

[26]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 SM8). 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 show in Appendix SM8 that our truncation rule satisfies weaker conditions that are nonetheless sufficient for the conclusion of Theorem 1.

[27]Note that by Assumption 4(2), $C$ can be chosen such that $\mathbf{A}_n(C)$ occurs with arbitrarily high probability. The additional assumption $\mathrm{P}(\mathbf{A}_n) \ge 1 - n^{-2}$ imposes that $\|\hat\eta - \eta_0\|_\infty$ concentrates around $\mathbb{E}[\|\hat\eta - \eta_0\|_\infty]$ so that $\mathrm{P}[\|\hat\eta - \eta_0\|_\infty > C\mathbb{E}[\|\hat\eta - \eta_0\|_\infty]]$ decays sufficiently fast with $n$ for some $C > 1$. Appendix SM8

**Theorem 1.** *Assume Assumptions 1 to 4 hold. 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 following upper bound holds:*

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

Second, we show a corresponding lower bound on the Bayes regret—i.e., a lower bound on the worst-case regret when an adversary picks $G_0, \eta_0$—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 $\theta_i^*$.[28]

**Theorem 2.** *Fix a set of valid hyperparameters $\mathcal{H}$. Let $\mathcal{P}(\mathcal{H}, \sigma_{1:n})$ be the set of distributions $P_0$ on support points $\sigma_{1:n}$ which satisfy (2.4) and Assumptions 2 to 4 corresponding to $\mathcal{H}$.[29] 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*

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

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

Therefore, 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 upper bound in Theorem 1 is a finite-sample statement, holding uniformly over all distributions $P_0$ delineated by the problem parameters $\mathcal{H}$. However, the usefulness of Theorem 1 still lies in the convergence rate, as the constants implied by the proofs are likely large. The proof of Theorem 1 is deferred to the Online Appendix, but its main ideas are outlined in Appendix A. The key theoretical novelty is characterizing the behavior of nuisance parameter estimation on the NPMLE objective function, as well as its subsequent impact on Bayes regret.

We mention two generalizations here. First, these regret upper bounds readily extend to the case where covariates are present and the location-scale assumption is with respect to

verifies that local linear regression satisfies a weakening of these assumptions that are also sufficient for the conclusion of Theorem 1.

[28] A similar argument is considered in Ignatiadis and Wager (2019) for a related but distinct setting.

[29] This result additionally imagines the adversary picking the support points $\sigma_{1:n}$. This is because the nonparametric regression problem would be "too easy" for certain configurations of $\sigma_{1:n}$. For instance, when $\sigma_{1:n}$ only takes $m \ll n$ unique values, nonparametric regression is possible at rate $\sqrt{m/n}$. For the proof, it suffices to consider $\sigma_{1:n}$ being equally spaced in $[\sigma_\ell, \sigma_u]$.

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 smoothness assumptions 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 $X$. Second, Appendix SM10 generalizes the regret rate to the problem of estimating higher moments $(\theta_i^v, v \in \mathbb{N})$ in squared error loss. Regret rates for these objects are novel in the empirical Bayes literature and are useful for decision problems where we wish to model risk aversion. For instance, the oracle's posterior variance is of the form $\mathbb{E}[\theta^2 \mid Y_i, \sigma_i] - \mathbb{E}[\theta \mid Y_i, \sigma_i]^2$.

Taken together, Theorems 1 and 2 are strong statistical optimality guarantees for CLOSE-NPMLE for MSE. That is, the worst-case MSE 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.

So far, our regret guarantees are only about estimation in MSE (Decision Problem 1). The next subsection analyzes regret for empirical Bayes decision rules targeted to the ranking-related problems (Decision Problems 2 and 3), and relates their performances to those for Decision Problem 1.

3.2. **Other decision objectives and relation to squared-error loss.** The oracle Bayes decision rules $\boldsymbol{\delta}^\star$ in Decision Problems 2 and 3 depend solely on the vector of oracle Bayes posterior means $\theta^*_{1:n}$. Therefore, for these problems, the natural empirical Bayes decision rules simply replace oracle Bayes posterior means $(\theta_i^*)$ with empirical Bayes ones $(\hat{\theta}_i)$ in the oracle decision rules. 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, $\mathrm{BayesRegret}_n$ (2.9) 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 $\mathrm{UMRegret}_n$ denote $\mathrm{BayesRegret}_n$ for Decision Problem 2 and we let $\mathrm{TopRegret}_n^{(m)}$ denote $\mathrm{BayesRegret}_n$ for Decision Problem 3. The following result relates $\mathrm{UMRegret}_n$ and $\mathrm{TopRegret}_n^{(m)}$ to $\mathrm{MSERegret}_n$ by showing that if $\hat{\theta}_i$ are close to $\theta_i^*$ in MSE, then decisions plugging in $\hat{\theta}_i$ are also close to their oracle counterparts in terms of Bayes risk.

**Theorem 3.** *Suppose* (2.1) *holds. Let $\hat{\delta}_i$ be the plug-in decisions with any vector of estimates $\hat{\theta}_i$. Then,* *(1) For* UTILITY MAXIMIZATION BY SELECTION,

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

*(2) For* TOP-$m$ SELECTION,

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

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, in the sense that the regret of the latter dominates those of the former. As a result, our convergence rates from Theorem 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 Theorem 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 10** (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. The proof of Theorem 3 exploits the fact that when the empirical Bayesian makes a selection mistake, the size of the mistake is not large if the square-error regret is low. It does not exploit the fact that if squared error regret is low, then the empirical Bayesian may be unlikely to make mistakes in the first place.[30] Nevertheless, Theorem 3 is competitive with recent results. For instance, in nonparametric settings, the rate in Theorem 3 is more favorable than the upper bound derived in Coey and Hung (2022), who also study Decision Problem 3. ■

3.3. **Robustness to the location-scale assumption** (2.4)**.** We prove Theorems 1 and 2 imposing the location-scale model (2.4). This is an optimistic assessment of the performance of CLOSE-NPMLE. While (2.4) nests prior independence, it may still be misspecified. We now explore the worst-case behavior of CLOSE-NPMLE without (2.4).

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. We will assume that $m_0, s_0$ are known. Without the location-scale model, $G_0$ is ill-defined, but we

[30] Upper and lower bounds are derived in related but distinct settings by Audibert and Tsybakov (2007), Bonvini et al. (2023), and Liang (2000); some upper bounds, under possibly stronger assumptions, appear better than implied by Theorem 3. We speculate that the bound for UTILITY MAXIMIZATION BY SELECTION can be tightened by verifying a margin condition, using Proposition 2 in Bonvini et al. (2023).

assume that we obtain some pseudo-true value $G_0^*$ that has zero mean and unit variance.[31] 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 $G_0^* \neq G_i$, where $G_0^*$ agrees with $G_i$ in the 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 $\eta_0$. 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 $\tau_i \mid \sigma_i$.

**Theorem 4.** *Under (2.1) but not (2.4), 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 means under a prior satisfying (2.4) with parameters $G_0^*, \eta_0$, for some fixed $G_0^*$ with mean zero and variance one. 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 $C_{\overline{\rho}} < \infty$,*

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

*where the infimum on the right-hand side is over all (possibly randomized) estimators of $\theta_i$ 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 comes 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 ($\delta_i = m_0(\sigma_i)$) 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 $s_0^2 = \frac{1}{n}\sum_{i=1}^n (m_0(\sigma_i) - m_0)^2 + s_0^2(\sigma_i)$.[32]

[31] This is a reasonable condition to impose, since every conditional prior distribution $\tau_i \mid \sigma_i$ obeys this moment constraint. We do not know if the maximizer $G$ of the population analogue to (2.8) 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 OA5.3).

[32] 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.

For additional reassurance under misspecification, Appendix OA4.2 discusses an interpretation of CLOSE-NPMLE under misspecification of (2.4). We end this section with a validation procedure that provides unbiased evaluation without relying on the location-scale model.

3.4. **Validating performance by coupled bootstrap.** Here, we describe a procedure that provides unbiased estimates of the loss of *arbitrary* decision procedures for Decision Problems 1 to 3. Practitioners can use this procedure to evaluate the gain of CLOSE-NPMLE relative to other alternatives—we do so extensively in Section 4.

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 under (1.1):

$$\begin{bmatrix} Y_i^{(1)} \\ Y_i^{(2)} \end{bmatrix} \mid \theta_i, \sigma_i^2 \sim \mathcal{N}\left( \begin{bmatrix} \theta_i \\ \theta_i \end{bmatrix}, \begin{bmatrix} (1+\omega)\sigma_i^2 & 0 \\ 0 & (1+\omega^{-1})\sigma_i^2 \end{bmatrix} \right). \tag{3.10}$$

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 $\sigma^2_{i,(1)}$ and $\sigma^2_{i,(2)}$.

The coupled bootstrap can be thought of as approximating sample-splitting the micro-data without needing access. We could imagine splitting the micro-data into training and testing sets, and think of $Y_i^{(1)}$ as estimates computed on the training data and $Y_i^{(2)}$ as estimates computed on the testing data. We might compute decisions based on $Y_i^{(1)}$ and evaluate them honestly with fresh data $Y_i^{(2)}$. The coupled bootstrap precisely emulates this sample-splitting procedure.

To see this, suppose $Y_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$ is a sample mean of i.i.d. micro-data $Y_{ij} : j = 1, \ldots, n_i$, as in Remark 2. Suppose we split the micro-data 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, (3.10) holds for $Y_i^{(1)}$ and $Y_i^{(2)}$. 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.

The following proposition formalizes this idea and states unbiased estimators for the loss of these decision rules, as well as their accompanying standard errors.[33]

[33]Oliveira et al. (2021) state the unbiased estimation result for the mean-squared error estimation problem. They connect the coupled bootstrap estimator to Stein's unbiased risk estimate. Our calculation for other

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 | $-\frac{1}{n}\sum_{i=1}^n \delta_i(Y_{1:n}^{(1)})Y_i^{(2)}$ | $\frac{1}{n^2}\sum_{i=1}^n \delta_i(Y_{1:n}^{(1)})\sigma_{i,(2)}^2$ |
| 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 $(Y_i, \sigma_i)$ obey (1.1). 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 displayed in 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 evaluate the performance of competing methods relative to some estimated prior—say the one learned by CLOSE-NPMLE. Doing so likely tips the scale in favor of a particular method, and we advocate for the coupled bootstrap instead.

## 4. 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., 2024).

Using longitudinal Census micro-data, Chetty et al. (2020) estimate a suite of tract-level children's outcomes in adulthood and publish the estimates, along with corresponding standard errors, in a collection of datasets called the Opportunity Atlas. Taking these estimates from the Opportunity Atlas, Bergman et al. (2024) conducted a program in Seattle called

loss functions extends their unbiased estimation result. 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 $\theta_i$. For unbiased estimation of functions of Gaussian parameters, see Table A1 in Voinov and Nikulin, 2012).

Creating Moves to Opportunity. They provided assistance to treated low-income individualsto move to Census tracts with empirical Bayes posterior means in the top third.[34] We view Bergman et al. (2024)'s 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, following Bergman et al. (2024), 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 $\vartheta_i$ and let $(Y_i, \theta_i)$ be their residualized counterparts against a vector of tract-level covariates $X_i$, with regression coefficient $\beta$.[35] We can apply the empirical Bayes procedures in this paper to $(Y_i, \sigma_i^2)$ and obtain an estimated posterior for $\theta_i$. This estimated posterior for the residualized parameter $\theta_i$ then implies an estimated posterior for 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 15 measures of economic mobility $\vartheta_i$. Each $\vartheta_i$ 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 (MEAN RANK) (2) an indicator for whether the individual has incomes in the top 20 percentiles (TOP-20 PROBABILITY), and (3) an indicator for whether the individual is incarcerated (INCARCERATION) for the following five demographic subgroups: (a) all individuals (POOLED), (b) white individuals, (c) white men, (d) Black individuals, and (e) Black men. Under these shorthands, the outcome we use in Section 2 is TOP-20 PROBABILITY for Black individuals, while Bergman et al. (2024) consider MEAN RANK POOLED.

The remainder of this section compares several empirical Bayes approaches on two exercises. In the first exercise, a calibrated simulation, we compare MSE performance of various methods to the that of the oracle posterior. The second exercise is an empirical

[34] There are also adjustments to make the selected tracts geographically contiguous. See Bergman et al. (2024) for details.

[35] 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 $\vartheta_i$, with accompanying standard errors $\sigma_i$. Let $\beta$ 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. Since $\beta$ is precisely estimated, we ignore its estimation noise. Then, the residualized objects $(Y_i, \theta_i)$ obey the Gaussian location model $Y_i \mid \theta_i, \sigma_i \sim \mathcal{N}(\theta_i, \sigma_i^2)$. Figure OA5.10 contains empirical results without residualizing against covariates. See Appendix OA5.2 for details on the covariates included.

application to a scale-up of the exercise in Bergman et al. (2024). It uses the coupled bootstrap (Section 3.4) to evaluate whether CLOSE-NPMLE selects more economically mobile tracts than INDEPENDENT-GAUSS.

4.1. **Calibrated simulation.** To devise a data-generating process that does not impose the location-scale assumption, we partition $\sigma$ into vingtiles, fit a location-scale model within each vingtile, and draw from the estimated model. The sampling process is detailed in Appendix OA5.2. Since the location-scale model is only imposed within each vingtile, this data-generating process does not impose (2.4).

On the simulated data, we then implement various empirical Bayes strategies. We consider the feasible procedures NAIVE, INDEPENDENT-GAUSS, INDEPENDENT-NPMLE, CLOSE-GAUSS, and CLOSE-NPMLE.[36] Here, NAIVE sets $\hat{\theta}_i = Y_i$. Since we know the ground truth data-generating process, we can also compute the ORACLE procedure, as well as an ORACLE-GAUSS procedure that computes (2.11) with the true $m_0, s_0$.

Figure 3 plots the main results from this calibrated simulation, focusing on MSE performance. 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 normalized by 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 3 shows how much of this pie each method captures.

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 perform worse than methods based on CLOSE.[37] Across the 15 variables, the median 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,

[36]We note that none of the feasible procedures have access to the true projection coefficient $\beta$ of $\tilde{Y}_i$ onto $X_i$, which they must estimate by residualizing against covariates on the data. Additionally, we weigh the estimation of $m_0$ and $s_0$ in INDEPENDENT-GAUSS by the precision $1/\sigma_i^2$, following Bergman et al. (2024).

[37]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 NAIVE. We note that, as in Bergman et al. (2024), 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 $\theta$ and $\sigma$ holds, these 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 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.

**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 $\vartheta_i$. The table shows relative performance, defined as the squared error improvement over NAIVE, normalized as a percentage of the improvement of ORACLE over NAIVE. 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 OA5.9. □

FIGURE 3. Table of relative squared error Bayes risk for various empirical Bayes approaches

CLOSE-NPMLE uniformly dominates all three other methods. This indicates that the standard error $\sigma_i$ 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. For MEAN RANK, after covariate residualization, there appears to be little dependence between $\theta_i$ and $\sigma_i$. INDEPENDENT-NPMLE and CLOSE-NPMLE perform similarly, capturing almost all of the available gains. Both slightly outperform the Gaussian methods for MEAN RANK.[38] For the other two outcome variables, TOP-20 PROBABILITY and INCARCERATION, the dependence between $\theta_i$ and $\sigma_i$ is stronger, and CLOSE-based methods display substantial improvements over methods that assume prior independence. CLOSE-NPMLE achieves near-oracle performance across the different definitions of $\theta_i$ 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 3.4) estimator of performance.

4.2. **Validation exercise via coupled bootstrap.** Our second empirical exercise uses the coupled bootstrap described in Section 3.4 for the policy problem in Bergman et al. (2024). Viewing the policy problem in Bergman et al. (2024) as TOP-$m$ SELECTION, can CLOSE-NPMLE make better selections?

Specifically, we imagine scaling up Bergman et al. (2024)'s exercise and perform empirical Bayes procedures for all Census tracts in the largest 20 Commuting Zones. We then select the top third of tracts *within* each Commuting Zone, according to empirical Bayesian posterior means for $\vartheta_i$. Additionally, to faithfully mimic Bergman et al. (2024), here we perform all empirical Bayes procedures *within Commuting Zone*. That is, for each of the 20 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.[39] Throughout, we choose $\omega$ to emulate a 90-10 train-test split on the micro-data.

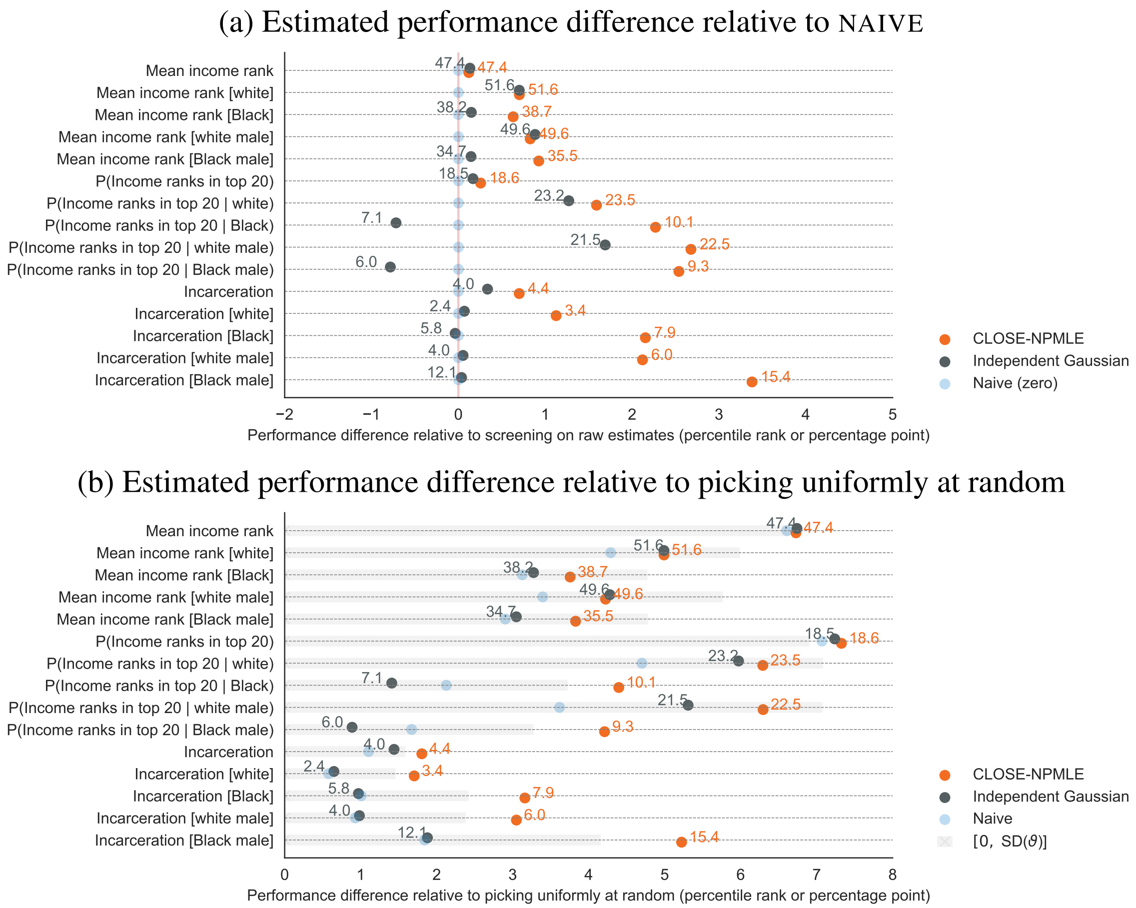
Figure 4(a) shows the estimated performance gap between a given empirical Bayes method and NAIVE as the $x$-position of the dots. According to these estimates, CLOSE-NPMLE generally improves over INDEPENDENT-GAUSS.[40]

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, at the income level for experiment participants in Bergman et al. (2024), an incremental percentile rank amounts to

[38] Appendix OA5.4 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.

[39] Appendix OA5.6 contains results where we perform empirical Bayes pooling over all Commuting Zones and select the top third within each Commuting Zone. We obtain very similar results. Appendix OA5.6 also contains results without residualizing against covariates, and INDEPENDENT-GAUSS performs very poorly in that setting. Appendix OA5.5 contains results on estimating $\vartheta_i$ in MSE (Decision Problem 1) in this context.

[40] 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 men. In either case, the estimated disimprovement is small.




*Notes.* These figures show the estimated performance of various decision rules over 1,000 draws of coupled bootstrap. Performance is measured as the mean $\vartheta_i$ 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 $\vartheta$. □

FIGURE 4. Performance of decision rules in top-$m$ selection exercise

about \$1,000 per annum. Thus, the estimated gain in terms of mean income rank is roughly \$500–800. For the other two outcomes, TOP-20 PROBABILITY and INCARCERATION,[41] the gains are even more sizable, especially for Black individuals.These gains are as high as 2–3 percentage points on average.

[41]We consider a policy objective of encouraging people to move *out* of high-incarceration areas.

Bergman et al. (2024) 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. (2024)'s setting, one might consider providing more personalized recommendations by targeting measures of economic mobility for finer demographic subgroups.[42] If we select tracts based on these demographic-specific measures, CLOSE-NPMLE then provides economically significant improvements: Appendix OA5.7 shows that screening with mobility measures for Black individuals outperforms screening mobility for Black individuals with the POOLED estimate.

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 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 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 4(b). Figure 4(b) plots the difference between a given method's performance and the estimated mean $\vartheta_i$ 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 $\vartheta_i$ as the *value of data*, since choosing the tracts randomly in the absence of data has expected performance equal to mean $\vartheta_i$. 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 sizable. These relative gains are more substantial for the binarized outcome variables TOP-20 PROBABILITY and

[42]When such personalized policies face legal and ethical barriers, Aloni and Avivi (2023) propose a minimax regret-type objective that takes the worst case over a customer's demographic identity.

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.

## 5. **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 $\theta_i$ as independent from $\sigma_i$, 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. (2024). 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. (2024), 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.

## **Appendix** A. **Proof outline for Theorem 1**

The proof of Theorem 1 depends on numerous results deferred to the Online Appendix. An outline is stated here.

To prove Theorem 1, we consider the events $\mathbf{A}_n, \mathbf{A}_n^{\mathrm{C}}$ separately. On $\mathbf{A}_n^{\mathrm{C}}$, we use the fact that the empirical Bayes posterior means $\hat{\theta}_i$ and the oracle posterior means $\theta_i^*$ are no farther than the range of the data $\max Y_i - \min Y_i$, which is logarithmic in $n$ under Assumption 2 (Lemma OA3.2). Since $\mathbf{A}_n^{\mathrm{C}}$ is assumed to be unlikely, regret on $\mathbf{A}_n^{\mathrm{C}}$ is sufficiently small.

The bulk of the argument controls regret on $\mathbf{A}_n$, stated separately in the following theorem (Theorem A.1), whose proof is deferred to Appendix OA3. Fix sequences $\Delta_n > 0$ and $M_n > 0$. Define the following "good" event which we use in Theorem A.1:

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

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

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

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

**Theorem A.1.** *Suppose Assumptions 1 to 4 hold. Fix some $\beta > 0, C_1 > 0$, there exists choices of constants $C_{\mathcal{H},2}$ such that, for $\Delta_n = C_1 n^{-p/(2p+1)} (\log n)^\beta$, $M_n = C_{\mathcal{H},2} (\log n)^{1/\alpha}$, and corresponding $A_n$,*

$$\mathbb{E}\left[ \mathrm{MSERegret}_n(\hat{G}_n, \hat{\eta}) \mathbb{1}(A_n) \right] \lesssim_{\mathcal{H}} n^{-\frac{2p}{2p+1}} (\log n)^{\frac{2+\alpha}{\alpha} + 3 + 2\beta}.$$

A.1. **Step 1: convert regret on $\theta_i$ to regret on $\tau_i$.** To prove Theorem A.1, note that the empirical Bayes posterior mean is of the form

$$\hat{\theta}_{i,\hat{G}_n,\hat{\eta}} = \hat{m}(\sigma_i) + \hat{s}(\sigma_i) \cdot \hat{\tau}_{i,\hat{G}_n,\hat{\eta}}$$

where $\hat{\tau}_{i,\hat{G}_n,\hat{\eta}}$ denotes the posterior mean of $\tau_i \mid \hat{Z}_i, \hat{\nu}_i$, where $\tau_i \sim \hat{G}_n$ and $\hat{Z}_i \mid \tau_i, \hat{\nu}_i \sim \mathcal{N}(\tau_i, \hat{\nu}_i^2)$. On the event $\mathbf{A}_n$, $\hat{m}, \hat{s}$ are close to $m_0, s_0$, and thus controlling $\mathrm{MSERegret}_n$ amounts to controlling MSE on $\tau$'s: $\mathbb{E}\left[ (\tau_i^* - \hat{\tau}_{i,\hat{G}_n,\hat{\eta}})^2 \right]$, where $\tau_i^* = \hat{\tau}_{i,G_0,\eta_0}$ is the oracle posterior mean for $\tau_i$.

To do so, we adapt the argument in Soloff et al. (2021) and Jiang (2020). To introduce this argument, recall that $\psi_i$ denotes the log-likelihood in Assumption 1 and define

$$\mathrm{Sub}_n(G) = \left( \frac{1}{n} \sum_{i=1}^n \psi_i(Z_i, \eta_0, G) - \frac{1}{n} \sum_{i=1}^n \psi_i(Z_i, \eta_0, G_0) \right)_+ \tag{A.2}$$

as the log-likelihood suboptimality of $G$ against the true distribution $G_0$, evaluated on the unobserved transformed data $Z_i, \nu_i$. For generic $G$ and $\nu > 0$, define

$$f_{G,\nu}(z) = \int_{-\infty}^{\infty} \varphi\left(\frac{z-\tau}{\nu}\right)\frac{1}{\nu}\, G(d\tau). \tag{A.3}$$

to be the marginal density of some mixed Gaussian deviate $Z \sim \mathcal{N}(0, \nu^2) \star G$. As a shorthand, we write $f_{i,G} = f_{G,\nu_i}(Z_i)$ and $f'_{i,G} = f'_{G,\nu_i}(Z_i)$. Let the average squared Hellinger distance be

$$\overline{h}^2(f_{G_1,\cdot}, f_{G_2,\cdot}) = \frac{1}{n}\sum_{i=1}^n h^2\left(f_{G_1,\nu_i}, f_{G_2,\nu_i}\right). \tag{A.4}$$

Loosely speaking, Soloff et al. (2021), following Jiang and Zhang (2009), show that

(1) With high probability, all approximate maximizers of the likelihood have low average Hellinger distance:

$$\mathrm{P}\left[\text{There exists } G \text{ where } \mathrm{Sub}_n(G) < C_1\delta_n^2 \text{ but } \overline{h}^2(f_{G,\cdot}, f_{G_0,\cdot}) > C_2\delta_n^2\right] < \frac{1}{n} \tag{A.5}$$

for some rate function $\delta_n^2 = \tilde{O}(1/n)$.

(2) For a given $G$, $\mathbb{E}[(\tau_i^* - \hat{\tau}_{i,G,\eta_0})^2] = \tilde{O}\left(\overline{h}^2(f_{G,\cdot}, f_{G_0,\cdot})\right)$ where $\hat{\tau}_{i,G,\eta_0} = \mathbf{E}_{G,\nu_i}[\tau \mid Z_i]$ are posterior means for $\tau_i$ under $\tau_i \sim G$ and $Z_i \mid \tau_i, \nu_i \sim \mathcal{N}(\tau_i, \nu_i^2)$.

Therefore, an approximate maximizer $\hat{G}_n^*$ of the likelihood $\mathrm{Sub}_n(G)$ should have low average Hellinger distance to $G_0$ and thus should output similar posterior means.[43]

A.2. **Step 2: show $\hat{G}_n$ is an approximate maximizer of true likelihood.** To use this argument for Theorem A.1, a key challenge is that $\hat{G}_n$ only maximizes the *approximate* likelihood $\frac{1}{n}\sum_i \psi_i(Z_i, \hat{\eta}, G)$, which only has $\hat{\eta} \approx \eta_0$ on $\mathbf{A}_n$, but $\hat{\eta} \neq \eta_0$. A key result is an oracle inequality for the likelihood (Corollary SM6.1), where, loosely speaking,

$$\mathrm{P}\left[\mathbf{A}_n, \mathrm{Sub}_n(\hat{G}_n) \gtrsim_{\mathcal{H}} \varepsilon_n\right] = O(1/n) \tag{A.6}$$

for some $\varepsilon_n = \tilde{O}\left(n^{-2p/(2p+1)} + n^{-p/(2p+1)}\overline{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot})\right)$. This result states that the likelihood suboptimality of the feasible NPMLE $\hat{G}_n$ cannot be much higher than its average Hellinger distance to $G_0$.

The bound (A.6) is a refinement of a simple linearization argument applied to $\eta \mapsto \frac{1}{n}\sum_{i=1}^n \psi_i(Z_i, \eta, \hat{G}_n)$. Heuristically speaking, a first-order Taylor expansion yields

$$\frac{1}{n}\sum_{i=1}^n \psi_i(Z_i, \hat{\eta}, \hat{G}_n) - \frac{1}{n}\sum_{i=1}^n \frac{\partial \psi_i}{\partial \eta}\bigg|_{\eta=\eta_0}(\hat{\eta}_i - \eta_{0i}) \approx \frac{1}{n}\sum_{i=1}^n \psi_i(Z_i, \eta_0, \hat{G}_n)$$

[43] Subjected to additional empirical process arguments that accommodate the fact that $\hat{G}_n^*$ is estimated.

where $\frac{1}{n}\sum_{i=1}^{n}\psi_i(Z_i,\hat{\eta},\hat{G}_n)$ is large by definition of $\hat{G}_n$. Thus, the right-hand side would be large following a bound on

$$\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\psi_i}{\partial\eta}\bigg|_{\eta=\eta_0}(\hat{\eta}_i-\eta_{0i})\right|.$$

A naive bound on this term, using only the fact that $|\hat{\eta}_i-\eta_{0i}|\le\|\hat{\eta}-\eta_0\|_\infty$, would lead to a suboptimal regret rate of $\tilde{O}(n^{-p/(2p+1)})$. Our more refined analysis additionally leverages the fact that

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

and thus the derivative $\frac{\partial\psi_i}{\partial\eta}$ is sufficiently small if $\hat{G}_n\approx G_0$.

A.3. **Step 3: adapt Hellinger distance bound.** Corollary SM6.1 makes sure that $\hat{G}_n$ probably achieves high likelihood, but the bound depends on $\overline{h}^2$. Since (A.5) uses a likelihood bound for $G$ to control $\overline{h}^2$, we need to additionally finesse (A.5) to accommodate the fact that the likelihood bound depends on $\overline{h}^2$.

Second, we adapt (A.5) to show that, loosely speaking, with high probability $\hat{G}_n$ has low average Hellinger distance to $G_0$ (Corollary OA3.1):

$$\mathrm{P}\left[\mathbf{A}_n,\overline{h}^2(f_{\hat{G}_n,\cdot},f_{G_0,\cdot})\gtrsim_{\mathcal{H}} n^{-p/(2p+1)}(\log n)^C\right]=\tilde{O}(1/n).$$

Thus, this allows us to show that $\mathbb{E}[(\tau_i^*-\hat{\tau}_{i,G,\eta_0})^2\mathbb{1}(\mathbf{A}_n)]$ is small, after additional empirical process arguments in Appendix OA3.

This section concludes with a proof for Theorem 1 given these results.

*Proof of Theorem 1.* 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 $C$ chosen by our application of Theorem A.1. Decompose

$$\begin{aligned}
&\mathbb{E}[\mathrm{MSERegret}_n(\hat{G}_n,\hat{\eta})]\\
&=\mathbb{E}[\mathrm{MSERegret}_n(\hat{G}_n,\hat{\eta})\mathbb{1}(A_n)]+\mathbb{E}[\mathrm{MSERegret}_n(\hat{G}_n,\hat{\eta})\mathbb{1}(\mathbf{A}_n^{\mathrm{C}}\cup\{\overline{Z}_n>M_n\})]\\
&\le\mathbb{E}[\mathrm{MSERegret}_n(\hat{G}_n,\hat{\eta})\mathbb{1}(A_n)]+\mathbb{E}[\mathrm{MSERegret}_n(\hat{G}_n,\hat{\eta})\mathbb{1}(\mathbf{A}_n^{\mathrm{C}})]\\
&\quad+\mathbb{E}[\mathrm{MSERegret}_n(\hat{G}_n,\hat{\eta})\mathbb{1}(\overline{Z}_n>M_n)]\\
&\lesssim_{\mathcal{H}} n^{-\frac{2p}{2p+1}}(\log n)^{\frac{2+\alpha}{\alpha}+3+2\beta_0}+\frac{2}{n}(\log n)^{2/\alpha} \qquad \text{(Theorem A.1 and Lemma OA3.2)}\\
&\lesssim_{\mathcal{H}} n^{-\frac{2p}{2p+1}}(\log n)^{\frac{2+\alpha}{\alpha}+3+2\beta_0},
\end{aligned}$$

where our application of Lemma OA3.2 uses the assumption that $\mathrm{P}(\mathbf{A}_n(C_{1,\mathcal{H}})^{\mathrm{C}})=\mathrm{P}(\|\hat{\eta}-\eta\|_\infty>\Delta_n)\le\frac{1}{n^2}$. ☐

## Appendix B. Proofs of other results stated in the main text

*Proof of Theorem 2.* 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. Under our assumptions, the oracle posterior means $\theta_i^*$ 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 $\hat{m}(\sigma_i) = \frac{s_0^2+\sigma_i^2}{\sigma_i^2}\left(\tilde{\theta}_i - \frac{s_0^2}{s_0^2+\sigma_i^2} Y_i\right)$. Note that, for this choice,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\tilde{\theta}_i - \theta_i^*)^2\right] \gtrsim_{\sigma_\ell, s_u} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\hat{m}(\sigma_i) - m_0(\sigma_i))^2\right].$$

Therefore, the minimax rate must be lower bounded by the minimax rate of estimating $m_0$ at $\sigma_{1:n}$, where the right-hand side takes the infimum over all estimators of $m_0$ with data $(Y_i, \sigma_i)$:

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

Using classical minimax results, Lemma SM9.1 shows that the right-hand side is lower bounded by $n^{-2p/(2p+1)}$, which completes the proof. □

*Proof of Theorem 3.* (1) By the law of iterated expectations, since $\hat{\theta}_i, \theta_i^*$ are both measurable with respect to the data,[44]

$$\mathbb{E}[\text{UMRegret}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left\{\mathbb{1}(\theta_i^* \ge 0) - \mathbb{1}(\hat{\theta}_i \ge 0)\right\} \theta_i^*\right]$$

Note that, for $\mathbb{1}(\theta_i^* \ge 0) - \mathbb{1}(\hat{\theta}_i \ge 0)$ to be nonzero, $0$ is between $\hat{\theta}_i$ and $\theta_i^*$. Hence, $|\theta_i^*| \le |\theta_i^* - \theta_i|$ and thus by Jensen's inequality

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

[44] 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 $\mathbb{E}[\theta_i \mid U, Y_i, \sigma_i] = \theta_i^*$.

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

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

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

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

Divide through by $m/n$ to obtain the result. □

**Proposition B.1.** *Suppose $\sigma(\cdot)$ is a permutation such that $\hat{\theta}_{\sigma(n)} \geq \cdots \geq \hat{\theta}_{\sigma(1)}$. Then*

$$\frac{1}{n}\sum_{i=1}^n w_i\theta_{(i)}^* - \frac{1}{n}\sum_{i=1}^n w_i\theta_{\sigma(i)}^* \leq \frac{2\|w\|_2}{\sqrt{n}}\sqrt{\frac{1}{n}\sum_{i=1}^n(\hat{\theta}_i - \theta_i^*)^2}.$$

*Proof.* We compute

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

The last step follows from the observation that $\sum_{i=1}^n(\theta_{(i)}^* - \hat{\theta}_{\sigma(i)})^2 \leq \sum_{i=1}^n(\hat{\theta}_i - \theta_i^*)^2$, which is true by the rearrangement inequality.[45] □

**Remark B.1** (Mover interpretation of Theorem 3)**.** Recall that we can think of TOP-$m$ SELECTION as the decision problem in Bergman et al. (2024). 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 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

[45]That is, for all real numbers $x_1 \leq \cdots \leq x_n, y_1 \leq \cdots \leq y_n$, we have that $\sum_i x_i y_{\pi(i)} \leq \sum_i x_i y_i$ for any permutation $\pi(\cdot)$.

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 $\pi_1$ of moving to the highest-ranked tract, $\pi_2$ 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). ■

*Proof of Theorem 4.* Note that $\hat{\theta}_{i,G_0^*,\eta_0} = s_0(\sigma_i)\hat{\tau}_{i,G_0^*,\eta_0} + m_0(\sigma_i)$, where $\tau_{i,G,\eta}^*$ is the posterior mean for $\tau_i$ under $(G,\eta)$, 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

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

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 (2.11)). 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

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

□

*Proof of Proposition 1.* 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

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

The conditional variance statement follows from

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

(3) (Decision Problem 3) The loss function for Decision Problem 3 is the same as that for Decision Problem 2 up to a factor of $n/m$. Since we condition on $Y_{1:n}^{(1)}$, the argument is thus analogous. □

## References


Abadie, A., A. Agarwal, G. Imbens, S. Jia, J. McQueen, and S. Stepaniants (2023). "Estimating the Value of Evidence-Based Decision Making". In: *arXiv preprint arXiv:2306.13681* (pp. 2).

Abadie, A. and M. Kasy (2019). "Choosing among regularized estimators in empirical economics: The risk of machine learning". In: *Review of Economics and Statistics* 101.5, pp. 743–762 (pp. 12).

Abaluck, J., M. Caceres Bravo, P. Hull, and A. Starc (2021). "Mortality effects and choice across private health insurance plans". In: *The quarterly journal of economics* 136.3, pp. 1557–1610 (pp. 2).

Aloni, T. and H. Avivi (2023). "One Land, Many Promises: The Unequal Consequences of Childhood Location for Natives and Immigrants in Israel". In: (pp. 2, 32).

Andrews, I. and M. Kasy (2019). "Identification of and correction for publication bias". In: *American Economic Review* 109.8, pp. 2766–94 (pp. 2).

Angrist, J. D., P. D. Hull, P. A. Pathak, and C. R. Walters (2017). "Leveraging lotteries for school value-added: Testing and estimation". In: *The Quarterly Journal of Economics* 132.2, pp. 871–919 (pp. 2).

Arellano, M. and S. Bonhomme (2009). "Robust priors in nonlinear panel data models". In: *Econometrica* 77.2, pp. 489–536 (pp. 2).

Arnold, D., W. Dobbie, and P. Hull (2022). "Measuring racial discrimination in bail decisions". In: *American Economic Review* 112.9, pp. 2992–3038 (pp. 2).

Athey, S. and S. Wager (2021). “Policy learning with observational data”. In: *Econometrica* 89.1, pp. 133–161 (pp. 12).

Audibert, J.-Y. and A. Tsybakov (2007). “Fast learning rates for plug-in classifiers”. In: *Annals of Statistics* (pp. 23).

Azevedo, E. M., A. Deng, J. L. Montiel Olea, J. Rao, and E. G. Weyl (2020). “A/b testing with fat tails”. In: *Journal of Political Economy* 128.12, pp. 4614–000 (pp. 2).

Baum-Snow, N. and L. Han (2019). “The microgeography of housing supply”. In: *Work in progress, University of Toronto* (pp. 2).

Bergman, P., R. Chetty, S. DeLuca, N. Hendren, L. F. Katz, and C. Palmer (2024). “Creating Moves to Opportunity: Experimental Evidence on Barriers to Neighborhood Choice”. In: *American Economic Review* (pp. 2–5, 8, 9, 12, 13, 26–28, 30, 32, 33, 38).

Bonhomme, S., K. Holzheu, T. Lamadon, E. Manresa, M. Mogstad, and B. Setzler (2020). “How much should we trust estimates of firm effects and worker sorting?” In: (pp. 2).

Bonhomme, S. and E. Manresa (2015). “Grouped patterns of heterogeneity in panel data”. In: *Econometrica* 83.3, pp. 1147–1184 (pp. 2).

Bonvini, M., E. H. Kennedy, and L. J. Keele (2023). “Minimax optimal subgroup identification”. In: *arXiv preprint arXiv:2306.17464* (pp. 23).

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2019). “nprobust: Nonparametric kernel-based estimation and robust bias-corrected inference”. In: *arXiv preprint arXiv:1906.00198* (pp. 7, 10).

Chamberlain, G. (1984). “Panel data”. In: *Handbook of econometrics* 2, pp. 1247–1318 (pp. 2).

Chandra, A., A. Finkelstein, A. Sacarny, and C. Syverson (2016). “Health care exceptionalism? Performance and allocation in the US health care sector”. In: *American Economic Review* 106.8, pp. 2110–44 (pp. 2).

Chen, J. (2023). “Mean-variance constrained priors have finite maximum Bayes risk in the normal location model”. In: *arXiv preprint arXiv:2303.08653* (pp. 24, 39).

Chetty, R., J. N. Friedman, N. Hendren, M. R. Jones, and S. R. Porter (2020). *The opportunity atlas: Mapping the childhood roots of social mobility*. Tech. rep. (pp. 2, 4, 7, 13, 26).

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014). “Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates”. In: *American economic review* 104.9, pp. 2593–2632 (pp. 13).

Chetty, R. and N. Hendren (2018). “The impacts of neighborhoods on intergenerational mobility I: Childhood exposure effects”. In: *The Quarterly Journal of Economics* 133.3, pp. 1107–1162 (pp. 2).

Chyn, E. and L. F. Katz (2021). “Neighborhoods matter: Assessing the evidence for place effects”. In: *Journal of Economic Perspectives* 35.4, pp. 197–222 (pp. 2).

Coey, D. and K. Hung (2022). “Empirical Bayesian Selection for Value Maximization”. In: *arXiv preprint arXiv:2210.03905* (pp. 23).

DellaVigna, S. and E. Linos (2022). “RCTs to scale: Comprehensive evidence from two nudge units”. In: *Econometrica* 90.1, pp. 81–116 (pp. 2).

Diamond, R. and E. Moretti (2021). *Where is standard of living the highest? Local prices and the geography of consumption*. Tech. rep. National Bureau of Economic Research (pp. 2).

Dimick, J. B., D. O. Staiger, and J. D. Birkmeyer (2010). “Ranking hospitals on surgical mortality: the importance of reliability adjustment”. In: *Health services research* 45.6p1, pp. 1614–1629 (pp. 2).

Doyle, J. J., J. A. Graves, J. Gruber, et al. (2017). *Evaluating measures of hospital quality*. Tech. rep. National Bureau of Economic Research (pp. 2).

Efron, B. (2016). “Empirical Bayes deconvolution estimates”. In: *Biometrika* 103.1, pp. 1–20 (pp. 11).

Egan, M., G. Matvos, and A. Seru (2022). “When Harry fired Sally: The double standard in punishing misconduct”. In: *Journal of Political Economy* 130.5, pp. 1184–1248 (pp. 2).

Einav, L., A. Finkelstein, and N. Mahoney (2022). *Producing Health: Measuring Value Added of Nursing Homes*. Tech. rep. National Bureau of Economic Research (pp. 2).

Elliott, G., N. Kudrin, and K. Wüthrich (2022). “Detecting p-hacking”. In: *Econometrica* 90.2, pp. 887–906 (pp. 2).

Fay, R. E. and R. A. Herriot (1979). “Estimates of income for small places: an application of James-Stein procedures to census data”. In: *Journal of the American Statistical Association* 74.366a, pp. 269–277 (pp. 27).

Finkelstein, A., M. Gentzkow, and H. Williams (2021). “Place-based drivers of mortality: Evidence from migration”. In: *American Economic Review* 111.8, pp. 2697–2735 (pp. 2).

Fu, L. J., G. M. James, and W. Sun (2020). “Nonparametric empirical bayes estimation on heterogeneous data”. In: *arXiv preprint arXiv:2002.12586* (pp. 15).

George, E. I., V. Ročková, P. R. Rosenbaum, V. A. Satopää, and J. H. Silber (2017). “Mortality rate estimation and standardization for public reporting: Medicare’s hospital compare”. In: *Journal of the American Statistical Association* 112.519, pp. 933–947 (pp. 10).

Giacomini, R., S. Lee, and S. Sarpietro (2023). *A Robust Method for Microforecasting and Estimation of Random Effects*. arXiv: 2308.01596 [econ.EM] (pp. 2).

Gilraine, M., J. Gu, and R. McMillan (2020). *A new method for estimating teacher value-added*. Tech. rep. National Bureau of Economic Research (pp. 5, 6, 12, 13).

Gu, J. and R. Koenker (2017). "Empirical Bayesball remixed: Empirical Bayes methods for longitudinal data". In: *Journal of Applied Econometrics* 32.3, pp. 575–599 (pp. 15).

– (2023). "Invidious comparisons: Ranking and selection as compound decisions". In: *Econometrica* 91.1, pp. 1–41 (pp. 5, 12).

Gu, J. and C. Walters (July 2022). *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 (pp. 5).

Hanushek, E. A. (2011). "The economic value of higher teacher quality". In: *Economics of Education review* 30.3, pp. 466–479 (pp. 13).

Heckman, J. and B. Singer (1984). "A method for minimizing the impact of distributional assumptions in econometric models for duration data". In: *Econometrica: Journal of the Econometric Society*, pp. 271–320 (pp. 11).

Hull, P. (2018). "Estimating hospital quality with quasi-experimental data". In: *Available at SSRN 3118358* (pp. 2).

Ignatiadis, N. and S. Wager (2019). "Covariate-powered empirical Bayes estimation". In: *Advances in Neural Information Processing Systems* 32 (pp. 21).

James, W. and C. Stein (1961). "Estimation with quadratic loss". In: *Breakthroughs in statistics: Foundations and basic theory*, pp. 443–460 (pp. 6).

Jiang, W. (2020). "On general maximum likelihood empirical Bayes estimation of heteroscedastic IID normal means". In: *Electronic Journal of Statistics* 14.1, pp. 2272–2297 (pp. 5, 6, 11, 12, 18, 19, 34).

Jiang, W. and C.-H. Zhang (2009). "General maximum likelihood empirical Bayes estimation of normal means". In: *The Annals of Statistics* 37.4, pp. 1647–1684 (pp. 4, 11, 18, 19, 35).

Jing, B.-Y., Z. Li, G. Pan, and W. Zhou (2016). "On sure-type double shrinkage estimation". In: *Journal of the American Statistical Association* 111.516, pp. 1696–1704 (pp. 12).

Kane, T. J. and D. O. Staiger (2008). *Estimating teacher impacts on student achievement: An experimental evaluation*. Tech. rep. National Bureau of Economic Research (pp. 13).

Kiefer, J. and J. Wolfowitz (1956). "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters". In: *The Annals of Mathematical Statistics*, pp. 887–906 (pp. 4, 11).

Kitagawa, T. and A. Tetenov (2018). “Who should be treated? empirical welfare maximization methods for treatment choice”. In: *Econometrica* 86.2, pp. 591–616 (pp. 12).

Kline, P., E. K. Rose, and C. Walters (2023). “A Discrimination Report Card”. In: *arXiv preprint arXiv:2306.13005* (pp. 2, 9).

Kline, P., E. K. Rose, and C. R. Walters (2022). “Systemic discrimination among large US employers”. In: *The Quarterly Journal of Economics* 137.4, pp. 1963–2036 (pp. 2).

Koenker, R. and J. Gu (2017). “REBayes: an R package for empirical Bayes mixture methods”. In: *Journal of Statistical Software* 82, pp. 1–26 (pp. 11, 15).

– (2019). “Comment: Minimalist g-modeling”. In: (pp. 11).

Koenker, R. and I. Mizera (2014). “Convex optimization, shape constraints, compound decisions, and empirical Bayes rules”. In: *Journal of the American Statistical Association* 109.506, pp. 674–685 (pp. 4, 11, 15).

Kubokawa, T., C. Robert, and A. M. E. Saleh (1993). “Estimation of noncentrality parameters”. In: *Canadian Journal of Statistics* 21.1, pp. 45–57 (pp. 15).

Kwon, S. (2021). “Optimal Shrinkage Estimation of Fixed Effects in Linear Panel Data Models”. In: *EliScholar–A Digital Platform for Scholarly Publishing at Yal e*, p. 1 (pp. 12).

Liang, T. (2000). “On an empirical Bayes test for a normal mean”. In: *Annals of statistics*, pp. 648–655 (pp. 23).

Lindsay, B. G. (1995). “Mixture models: theory, geometry, and applications”. In: Ims (pp. 11).

Liu, L., H. R. Moon, and F. Schorfheide (2020). “Forecasting with dynamic panel data models”. In: *Econometrica* 88.1, pp. 171–201 (pp. 2).

Manski, C. F. (2004). “Statistical treatment rules for heterogeneous populations”. In: *Econometrica* 72.4, pp. 1221–1246 (pp. 12).

Meager, R. (2022). “Aggregating distributional treatment effects: A Bayesian hierarchical analysis of the microcredit literature”. In: *American Economic Review* 112.6, pp. 1818–47 (pp. 2).

Mehta, N. (2019). “Measuring quality for use in incentive schemes: The case of “shrinkage” estimators”. In: *Quantitative Economics* 10.4, pp. 1537–1577 (pp. 8).

Mogstad, M., J. P. Romano, A. Shaikh, and D. Wilhelm (2020). *Inference for ranks with applications to mobility across neighborhoods and academic achievement across countries*. Tech. rep. National Bureau of Economic Research (pp. 12).

– (2023). “A Comment on:“Invidious Comparisons: Ranking and Selection as Compound Decisions” by Jiaying Gu and Roger Koenker”. In: *Econometrica* 91.1, pp. 53–60 (pp. 12).

Montiel Olea, J. L., B. O’Flaherty, and R. Sethi (2021). “Empirical Bayes Counterfactuals in Poisson Regression with an Application to Police Use of Deadly Force". In: *Available at SSRN 3857213* (pp. 2).

Morris, C. N. (1983). "Parametric empirical Bayes inference: theory and applications". In: *Journal of the American statistical Association* 78.381, pp. 47–55 (pp. 6).

Mountjoy, J. and B. R. Hickman (2021). *The returns to college (s): Relative value-added and match effects in higher education*. Tech. rep. National Bureau of Economic Research (pp. 2).

Oliveira, N. L., J. Lei, and R. J. Tibshirani (2021). "Unbiased risk estimation in the normal means problem via coupled bootstrap techniques". In: *arXiv preprint arXiv:2111.09447* (pp. 4, 25, 33).

Polyanskiy, Y. and Y. Wu (2020). "Self-regularizing property of nonparametric maximum likelihood estimator in mixture models". In: *arXiv preprint arXiv:2008.08244* (pp. 11).

– (2021). "Sharp regret bounds for empirical Bayes and compound decision problems". In: *arXiv preprint arXiv:2109.03943* (pp. 11).

Rambachan, A. (2021). "Identifying prediction mistakes in observational data". In: (pp. 2).

Robbins, H. (1956). *An empirical Bayes approach to statistics* (pp. 6, 12).

Saha, S. and A. Guntuboyina (2020). "On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising". In: *The Annals of Statistics* 48.2, pp. 738–762 (pp. 11, 18).

Shen, Y. and Y. Wu (2022). "Empirical Bayes estimation: When does $g$-modeling beat $f$-modeling in theory (and in practice)?" In: *arXiv preprint arXiv:2211.12692* (pp. 11).

Soloff, J. A., A. Guntuboyina, and B. Sen (2021). "Multivariate, heteroscedastic empirical bayes via nonparametric maximum likelihood". In: *arXiv preprint arXiv:2109.03466* (pp. 5, 6, 11, 12, 18, 19, 34, 35).

Stone, C. J. (1980). "Optimal rates of convergence for nonparametric estimators". In: *The annals of Statistics*, pp. 1348–1360 (pp. 20).

Stoye, J. (2009). "Minimax regret treatment choice with finite samples". In: *Journal of Econometrics* 151.1, pp. 70–81 (pp. 12).

Tsybakov, A. (2008). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York. ISBN: 9780387790527. URL: https://books.google.com/books?id=mwB8rUBsbqoC (pp. 20).

van der Vaart, A. W. and J. A. Wellner (1996). *Weak convergence*. Springer, pp. 16–28 (pp. 19).

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press (pp. 19).

Voinov, V. G. and M. S. Nikulin (2012). *Unbiased estimators and their applications: volume 1: univariate case*. Vol. 263. Springer Science & Business Media (pp. 26).

Weinstein, A., Z. Ma, L. D. Brown, and C.-H. Zhang (2018). "Group-linear empirical Bayes estimates for a heteroscedastic normal mean". In: *Journal of the American Statistical Association* 113.522, pp. 698–710 (pp. 16).

Wernerfelt, N., A. Tuchman, B. Shapiro, and R. Moakler (2022). "Estimating the Value of Offsite Data to Advertisers on Meta". In: *University of Chicago, Becker Friedman Institute for Economics Working Paper* 114 (pp. 2).

Xie, X., S. Kou, and L. D. Brown (2012). "SURE estimates for a heteroscedastic hierarchical model". In: *Journal of the American Statistical Association* 107.500, pp. 1465–1479 (pp. 12).

# Online Appendix to
# "Empirical Bayes When Estimation Precision Predicts Parameters"

Jiafeng Chen

April 8, 2024

**Contents**

## Part 1 Proof of Theorem 1

### Appendix OA3. Review of notation and proof of Lemma OA3.2 and Theorem A.1

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 $(\theta_i, \sigma_i)$ takes the location-scale form (2.4)

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

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. Let $\nu_\ell, \nu_u$ be the corresponding bounds on $\nu_i = \frac{\sigma_i}{s_0(\sigma_i)}$, implied by Assumption 3.

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

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

where we recall $f_{G,\nu}$ from (A.3).

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

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

Here, we define $\mathbf{E}_{G,\nu}[h(\tau, Z) \mid z]$ 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,

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

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

$$\hat{\theta}_{i,G,\eta,\rho} = Y_i + s_i\hat{\nu}_i^2(\eta)\frac{f'_{G,\hat{\nu}_i(\eta)}(\hat{Z}_i(\eta))}{f_{G,\hat{\nu}_i(\eta)}(\hat{Z}_i(\eta)) \vee \frac{\rho}{\hat{\nu}_i(\eta)}} \tag{OA3.1}$$

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

$$\hat{\tau}_{i,G,\eta,\rho} = \hat{Z}_i(\eta) + \hat{\nu}_i^2(\eta)\frac{f'_{G,\hat{\nu}_i(\eta)}(\hat{Z}_i(\eta))}{f_{G,\hat{\nu}_i(\eta)}(\hat{Z}_i(\eta)) \vee \frac{\rho}{\hat{\nu}_i(\eta)}} \quad \tau_{i,\rho}^* = \hat{\tau}_{i,G_0,\eta_0,\rho} \tag{OA3.2}$$

Lastly, we will also define

$$\varphi_+(\rho) = \sqrt{\log\frac{1}{2\pi\rho^2}} \quad \rho \in (0, (2\pi)^{-1/2}) \tag{OA3.3}$$

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

Recall the event $A_n$ in (A.1) as well as $\overline{Z}_n$. Many of the following statements are true for $A_n$ defined with generic $\Delta_n, M_n$. However, to obtain our rate expression in the end, we shall set $\Delta_n, M_n$ to be of the following form:

$$\Delta_n = C_{\mathcal{H}} n^{-\frac{p}{2p+1}}(\log n)^\beta \text{ and } M_n = (C_{\mathcal{H}} + 1)(C_{2,\mathcal{H}}^{-1}\log n)^{1/\alpha}. \tag{OA3.4}$$

Here, $C_{\mathcal{H}}$ is to be chosen, and $C_{2,\mathcal{H}}$ is some constant determined by Theorem SM6.1. Correspondingly, we also have a choice

$$\rho_n = \frac{1}{n^3}e^{-C_{\mathcal{H},\rho}M_n^2\Delta_n} \wedge \frac{1}{e\sqrt{2\pi}}, \tag{OA3.5}$$

where the constant $C_{\mathcal{H},\rho}$ is chosen to satisfy the following result, proved in Appendix SM6.

**Lemma OA3.1.** *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 SM6.1*,[46]

*(1)* $|\hat{Z}_i \vee 1| \lesssim_{\mathcal{H}} M_n$

[46]This assumption is satisfied with our choices in (OA3.4).

*(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 $\rho_n$ satisfies $\log(1/\rho_n) \asymp_{\mathcal{H}} \log n$, $\varphi_+(\rho_n) \asymp_{\mathcal{H}} \sqrt{\log n}$, and $\rho_n \lesssim_{\mathcal{H}} n^{-3}$.*

We now state and prove Lemma OA3.2 and Theorem A.1, which are crucial claims in the proof of Theorem 1. The first claim, Lemma OA3.2, controls regret on the event $A_n^{\mathrm{C}}$.

**Lemma OA3.2.** *Under Assumptions 1 to 4, for $\beta \ge 0$, suppose $\Delta_n, M_n$ are of the form (OA3.4) such that $\mathrm{P}(\overline{Z}_n > M_n) \le n^{-2}$, we can decompose*

$$\mathbb{E}[\mathrm{MSERegret}_n(\hat{G}_n, \hat{\eta})\mathbb{1}(\|\hat{\eta} - \eta\|_\infty > \Delta_n)] \lesssim_{\mathcal{H}} \mathrm{P}(\|\hat{\eta} - \eta\|_\infty > \Delta_n)^{1/2}(\log n)^{2/\alpha}$$

$$\mathbb{E}[\mathrm{MSERegret}_n(\hat{G}_n, \hat{\eta})\mathbb{1}(\overline{Z}_n > M_n)] \lesssim_{\mathcal{H}} \frac{1}{n}(\log n)^{2/\alpha}.$$

*Proof.* Observe that, for an event $A$ on the data $Z_{1:n}$,

$$\begin{aligned}\mathbb{E}\left[\mathrm{MSERegret}_n(\hat{G}_n, \hat{\eta})\mathbb{1}(A)\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_{i,\hat{G},\hat{\eta}} - \theta_i^*)^2 \mathbb{1}(A)\right] \\ &\le \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_{i,\hat{G},\hat{\eta}} - \theta_i^*)^2\right)^2\right]^{1/2} \mathrm{P}(A)^{1/2}\end{aligned}$$

by Cauchy–Schwarz. A crude bound (Lemma OA3.6) 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 OA3.7 to find that $\mathbb{E}[\overline{Z}_n^4] \lesssim_{\mathcal{H}} (\log n)^{4/\alpha}$. This proves both claims. ☐

The main theorem of this part in the Online Appendix is stated and proved in the following section. It characterizes regret behavior on the event $A_n$, for $\Delta_n, M_n$ chosen as in (OA3.4).

**OA3.1 Proof of Theorem A.1.** We first state a result that is key to our remaining arguments, which we verify in the Supplementary Material (Appendix SM7).

**Corollary OA3.1.** *Assume Assumptions 1 to 4 hold and suppose $\Delta_n, M_n$ take the form (OA3.4). Define the rate function*

$$\delta_n = n^{-p/(2p+1)}(\log n)^{\frac{2+\alpha}{2\alpha}+\beta}. \tag{OA3.6}$$

*Then, there exists some constant $B_{\mathcal{H}}$, depending solely on $C^*_{\mathcal{H}}$ in Corollary SM6.1, $\beta$, and $p, \nu_\ell, \nu_u$ such that*

$$\mathrm{P}\left[A_n, \overline{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}) > B_{\mathcal{H}}\delta_n\right] \le \left(\frac{\log\log n}{\log 2} + 10\right)\frac{1}{n}.$$

**Theorem A.1.** *Suppose Assumptions 1 to 4 hold. Fix some $\beta > 0, C_1 > 0$, there exists choices of constants $C_{\mathcal{H},2}$ such that, for $\Delta_n = C_1 n^{-p/(2p+1)}(\log n)^\beta$, $M_n = C_{\mathcal{H},2}(\log n)^{1/\alpha}$, and corresponding $A_n$,*

$$\mathbb{E}\left[\mathrm{MSERegret}_n(\hat{G}_n, \hat{\eta})\mathbb{1}(A_n)\right] \lesssim_{\mathcal{H}} n^{-\frac{2p}{2p+1}}(\log n)^{\frac{2+\alpha}{\alpha}+3+2\beta}.$$

*Proof.* We choose $M_n$ to be of the form (OA3.4). Note that we can decompose

$$\begin{aligned}\mathrm{MSERegret}_n(G,\eta) &= \frac{1}{n}\sum_{i=1}^n(\hat{\theta}_{i,G,\eta} - \theta_i)^2 - \frac{1}{n}\sum_{i=1}^n(\theta_i^* - \theta_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^n(\hat{\theta}_{i,G,\eta} - \theta_i^*)^2 + \frac{2}{n}\sum_{i=1}^n(\theta_i^* - \theta_i)(\hat{\theta}_{i,G,\eta} - \theta_i^*) \qquad \text{(OA3.7)}\end{aligned}$$

Note that the second term in the decomposition (OA3.7), truncated to $A_n$, is mean zero:

$$\mathbb{E}\left[\mathbb{1}(A_n)\frac{2}{n}\sum_{i=1}^n(\theta_i^* - \theta_i)(\hat{\theta}_{i,\hat{G}_n,\hat{\eta}} - \theta_i^*)\right] = 0,$$

since $\mathbb{E}[(\theta_i^* - \theta_i) \mid Y_1, \ldots, Y_n] = 0$. Thus, we can focus on

$$\mathbb{E}[\mathrm{MSERegret}_n(\hat{G}_n,\hat{\eta})\mathbb{1}(A_n)] = \mathbb{E}\left[\frac{\mathbb{1}(A_n)}{n}\sum_{i=1}^n(\hat{\theta}_{i,\hat{G}_n,\hat{\eta}} - \theta_i^*)^2\right] \equiv \frac{1}{n}\mathbb{E}[\mathbb{1}(A_n)\|\hat{\theta}_{\hat{G}_n,\hat{\eta}} - \theta^*\|^2], \qquad \text{(OA3.8)}$$

where we let $\hat{\theta}_{\hat{G}_n,\hat{\eta}}$ denote the vector of estimated posterior means and let $\theta^*$ denote the corresponding vector of oracle posterior means. Let the subscript $\rho_n$ denote a vector of regularized posterior means as in (OA3.1). Here, we set $\rho_n$ as in (OA3.5). Thus, we may further decompose by triangle inequality:

$$\|\hat{\theta}_{\hat{G}_n,\hat{\eta}} - \theta^*\| \le \|\hat{\theta}_{\hat{G}_n,\hat{\eta}} - \hat{\theta}_{\hat{G}_n,\eta_0}\| + \|\hat{\theta}_{\hat{G}_n,\eta_0} - \hat{\theta}_{\hat{G}_n,\eta_0,\rho_n}\| + \|\hat{\theta}_{\hat{G}_n,\eta_0,\rho_n} - \theta^*_{\rho_n}\| + \|\theta^*_{\rho_n} - \theta^*\|.$$

We denote each term in the decomposition of (OA3.8) by $\xi_1, \ldots, \xi_4$:

$$\xi_1 = \frac{\mathbb{1}(A_n)}{n}\|\hat{\theta}_{\hat{G}_n,\hat{\eta}} - \hat{\theta}_{\hat{G}_n,\eta_0}\|^2 \qquad \text{(OA3.9)}$$

$$\xi_2 = \frac{\mathbb{1}(A_n)}{n}\|\hat{\theta}_{\hat{G}_n,\eta_0} - \hat{\theta}_{\hat{G}_n,\eta_0,\rho_n}\|^2 \qquad \text{(OA3.10)}$$

$$\xi_3 = \frac{\mathbb{1}(A_n)}{n}\|\hat{\theta}_{\hat{G}_n,\eta_0,\rho_n} - \theta^*_{\rho_n}\|^2 \qquad \text{(OA3.11)}$$

$$\xi_4 = \frac{\mathbb{1}(A_n)}{n}\|\theta^*_{\rho_n} - \theta^*\|^2. \tag{OA3.12}$$

We have that

$$\text{(OA3.8)} \le 4(\mathbb{E}\xi_1 + \mathbb{E}\xi_2 + \mathbb{E}\xi_3 + \mathbb{E}\xi_4) = 4(\mathbb{E}\xi_1 + \mathbb{E}\xi_3 + \mathbb{E}\xi_4).$$

The individual $\xi_j$'s are bounded by the arguments in the remainder of this section. The key term leading to the final rate is $\mathbb{E}[\xi_3]$:

- We show in Lemma OA3.3 that $\xi_1 \lesssim_{\mathcal{H}} M_n^2(\log n)^2\Delta_n^2$, and thus $\mathbb{E}\xi_1 \lesssim_{\mathcal{H}} M_n^2(\log n)^2\Delta_n^2$.
- Lemma OA3.1 implies that, given the choice $\rho_n$ in (OA3.5), the regularized posterior means and the unregularized posterior means are equal $\hat{\theta}_{\hat{G}_n,\eta_0,\rho_n} = \hat{\theta}_{\hat{G}_n,\eta_0}$, since the truncation does not bind. Therefore, $\xi_2 = 0$.
- We show in Appendix OA3.2 that $\mathbb{E}\xi_3 \lesssim_{\mathcal{H}} (\log n)^3\delta_n^2$. Here, $\delta_n$ is the rate in (OA3.6).
- Finally, we show in Lemma OA3.4 that $\mathbb{E}\xi_4 \lesssim_{\mathcal{H}} \frac{1}{n}$.

Lastly, we observe that by the definition of $\delta_n$ in (OA3.6), the upper bound for $\mathbb{E}[\xi_3]$ is the dominating rate. Plugging the definition of $\delta_n^2$ yields that

$$\text{(OA3.8)} = \mathbb{E}[\mathrm{MSERegret}_n(\hat{G}_n, \hat{\eta})\mathbb{1}(A_n)] \lesssim_{\mathcal{H}} n^{-\frac{2p}{2p+1}}(\log n)^{\frac{2+\alpha}{\alpha}+3+2\beta_1}. \qquad \square$$

**Remark OA3.1** (Remainder of proof)**.** The proof for Theorem A.1 hinges on the key result in Appendix OA3.2 for bounding $\xi_3$. Effectively, the argument first relates $\xi_3$ to the corresponding regret for the transformed parameters $\tau_i$ (OA3.2):

$$\|\tau_{\hat{G}_n,\eta_0,\rho_n} - \tau^*_{\rho_n}\|^2.$$

To prove a bound for this object, we truncate to the event where $\bar{h}^2(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot})$ is small and use the fact that—loosely speaking—the $\|\tau_{\hat{G}_n,\eta_0,\rho_n} - \tau^*_{\rho_n}\|^2$ can be bounded by $\bar{h}^2(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot})$. For this argument to work, the key is that the event where $\bar{h}^2(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot})$ is small has high probability, which is shown in Corollary OA3.1. Lastly, to prove Corollary OA3.1, we need to first establish that $\hat{G}_n$—estimated off $(\hat{Z}_i, \hat{\nu}_i)$—does not have high likelihood suboptimality $\mathrm{Sub}_n(\hat{G}_n)$. This is the most laborious part of the proof and shown in Corollary SM6.1. ■

**Lemma OA3.3.** *Under the assumptions of Theorem A.1, in the proof of Theorem A.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,

$$\left|\hat{\theta}_{i,\hat{G}_n,\hat{\eta}} - \hat{\theta}_{i,\hat{G}_n,\eta_0}\right| = \sigma_i^2\left|\frac{f'_{\hat{G}_n,\hat{\nu}_i}(\hat{Z}_i)}{\hat{s}_i f_{\hat{G}_n,\hat{\nu}_i}(\hat{Z}_i)} - \frac{f'_{\hat{G}_n,\nu_i}(Z_i)}{s_{0i} f_{\hat{G}_n,\nu_i}(Z_i)}\right| = \sigma_i^2\left|\left(\frac{\partial\psi_i}{\partial m_i}\bigg|_{\hat{G}_n,\hat{\eta}} - \frac{\partial\psi_i}{\partial m_i}\bigg|_{\hat{G}_n,\eta_0}\right)\right|$$

$$= \sigma_i^2 \left| \frac{\partial^2 \psi_i}{\partial m_i \partial s_i}\bigg|_{\hat{G}_n, \tilde{\eta}_i} (\hat{s}_i - s_{0i}) + \frac{\partial^2 \psi_i}{\partial m_i^2}\bigg|_{\hat{G}_n, \tilde{\eta}_i} (\hat{m}_i - m_{0i}) \right|,$$

where we use $\tilde{\eta}_i$ to denote some intermediate value lying on the line segment between $\hat{\eta}_i$ and $\eta_{0i}$. By Lemma SM6.13, we can bound the two derivative terms,

$$\mathbb{1}(A_n) \left| \hat{\theta}_{i,\hat{G}_n,\hat{\eta}} - \hat{\theta}_{i,\hat{G}_n,\eta_0} \right| \lesssim_{\mathcal{H}} M_n (\log n) \Delta_n.$$

Hence, squaring both sides, we obtain $\xi_1 \lesssim_{\mathcal{H}} M_n^2 (\log n)^2 \Delta_n^2$. □

**Lemma OA3.4.** *Under the assumptions of Theorem A.1, in the proof of Theorem A.1,* $\mathbb{E}\xi_4 \lesssim_{\mathcal{H}} \frac{1}{n}$.

*Proof.* Note that

$$\begin{aligned}
\mathbb{E}[(\theta_{i,\rho_n}^* - \theta_i^*)^2] &= \int \left( \nu_i^2 \frac{f'_{G_0,\nu_i}(z)}{f_{G_0,\nu_i}(z)} \right)^2 \left( 1 - \frac{f_{G_0,\nu_i}}{f_{G_0,\nu_i} \vee \frac{\rho_n}{\nu_i}} \right)^2 f_{G_0,\nu_i}(z)\, dz \\
&\le \mathbb{E}\left[ \left( \nu_i^2 \frac{f'_{G_0,\nu_i}(z)}{f_{G_0,\nu_i}(z)} \right)^4 \right]^{1/2} \mathrm{P}\left[ f_{G_0,\nu_i}(Z) < \rho_n/\nu_i \right]^{1/2} \quad \text{(Cauchy–Schwarz)} \\
&\lesssim_{\mathcal{H}} 1 \cdot \rho_n^{1/3}\, \mathrm{Var}(Z)^{1/6} \\
&\qquad\qquad \text{(Tweedie's formula, Jensen's inequality, and Lemma SM6.11)} \\
&\lesssim_{\mathcal{H}} \frac{1}{n}.
\end{aligned}$$

Therefore, $\mathbb{E}[\xi_4] \lesssim_{\mathcal{H}} \frac{1}{n}$. □

**OA3.2 Controlling $\xi_3$.**

**Lemma OA3.5.** *Under the assumptions of Theorem A.1, in the proof of Theorem A.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 $\eta_0$ and $\tau_{i,\rho_n}^* = \hat{\tau}_{i,G_0,\eta_0,\rho_n}$ (where we recall (OA3.2)).

Thus, 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 $\delta_n$ in (OA3.6) and the constant $B_{\mathcal{H}}$ in Corollary OA3.1 (which in turn depends on $C_{\mathcal{H}}^*$ in Corollary SM6.1). 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 $\tau_{\rho_n}^{(j)}$ be the posterior mean vector corresponding to prior $G_j$ with nuisance parameter $\eta_0$ and regularization $\rho_n$.

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

$$\zeta_1^2 = \|\hat{\tau}_{\hat{G}_n,\eta_0,\rho_n} - \tau^*_{\rho_n}\|^2 \mathbb{1}\left(A_n \cap B_n^{\mathrm{C}}\right) \tag{OA3.13}$$

$$\zeta_2^2 = \left(\|\hat{\tau}_{\hat{G}_n,\eta_0,\rho_n} - \tau^*_{\rho_n}\| - \max_{j\in[N]}\|\tau^{(j)}_{\rho_n} - \tau^*_{\rho_n}\|\right)_+^2 \mathbb{1}(A_n \cap B_n) \tag{OA3.14}$$

$$\zeta_3^2 = \max_{j\in[N]}\left(\|\tau^{(j)}_{\rho_n} - \tau^*_{\rho_n}\| - \mathbb{E}\left[\|\tau^{(j)}_{\rho_n} - \tau^*_{\rho_n}\|\right]\right)_+^2 \tag{OA3.15}$$

$$\zeta_4^2 = \max_{j\in[N]}\left(\mathbb{E}\left[\|\tau^{(j)}_{\rho_n} - \tau^*_{\rho_n}\|\right]\right)^2 \tag{OA3.16}$$

The decomposition $\zeta_1$ through $\zeta_4$ is exactly analogous to Section C.3 in Soloff et al. (2021) and to the proof of Theorem 1 in Jiang (2020). In particular, $\zeta_1$ 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 OA3.1. $\zeta_2$ is the distance from the posterior means at $\hat{G}_n$ to the closest posterior mean generated from the net $G_1, \ldots, G_N$; $\zeta_2$ is small if we make the net very fine. $\zeta_3$ measures the distance between $\|\tau^{(j)}_{\rho_n} - \tau^*_{\rho_n}\|$ and its expectation; $\zeta_3$ can be controlled by (i) a large-deviation inequality and (ii) controlling the metric entropy of the net (Proposition SM6.2). Lastly, $\zeta_4$ 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 $\zeta_3$ is slightly different and avoids an argument in Jiang and Zhang (2009) which appears to not apply in the heteroskedastic setting. See Remark OA3.2.

The subsequent subsections control $\zeta_1$ through $\zeta_4$, and find that $\zeta_4 \lesssim_{\mathcal{H}} (\log n)^3\delta_n^2$ is the dominating term. □

*OA3.2.1 Controlling* $\zeta_1$. First, we note that

$$\left(\hat{\tau}_{i,\hat{G}_n,\eta_0,\rho_n} - \tau^*_{i,\rho_n}\right)^2 \mathbb{1}(A_n \cap B_n^{\mathrm{C}}) \lesssim_{\mathcal{H}} \log(1/\rho_n)\mathbb{1}(A_n \cap B_n^{\mathrm{C}}) = \log n \mathbb{1}(A_n \cap B_n^{\mathrm{C}})$$

By Corollary OA3.1, $\mathrm{P}(A_n \cap B_n^{\mathrm{C}}) \le \left(\frac{\log\log n}{\log 2} + 9\right)\frac{1}{n}$, and hence

$$\frac{1}{n}\mathbb{E}\zeta_1^2 \lesssim_{\mathcal{H}} \frac{\log n \log\log n}{n}.$$

*OA3.2.2 Controlling $\zeta_2$.* Choose $G_1, \ldots, G_N$ to be a minimal $\omega$-covering of $\left\{G : \overline{h}(f_{G,\cdot}, f_{G_0,\cdot}) \leq \delta_n\right\}$ under the pseudometric

$$d_{M_n,\rho_n}(H_1, H_2) = \max_{i\in[n]} \sup_{z:|z|\leq M_n} \left| \frac{\nu_i^2 f'_{H_1,\nu_i}(z)}{f_{H_1,\nu_i}(z) \vee \left(\frac{\rho_n}{\nu_i}\right)} - \frac{\nu_i^2 f'_{H_2,\nu_i}(z)}{f_{H_2,\nu_i}(z) \vee \left(\frac{\rho_n}{\nu_i}\right)} \right| \tag{OA3.17}$$

where $N \leq N\left(\omega/2, \mathcal{P}(\mathbb{R}), d_{M_n,\rho_n}\right)$.[47] We note that (OA3.17) and (SM6.25) are different only by constant factors. Therefore, Proposition SM6.2 implies that

$$\log N\left(\frac{\delta \log(1/\delta)}{\rho_n}\sqrt{\log(1/\rho_n)}, \mathcal{P}(\mathbb{R}), d_{M_n,\rho_n}\right) \lesssim_{\mathcal{H}} \log(1/\delta)^2 \max\left(1, \frac{M_n}{\sqrt{\log(1/\delta)}}\right) \tag{OA3.18}$$

for all sufficiently small $\delta > 0$.

Then

$$\frac{1}{n}\zeta_2^2 \leq \mathbb{1}(A_n \cap B_n)\frac{1}{n}\max_{j\in[N]}\|\hat{\tau}_{\hat{G}_n,\eta_0,\rho_n} - \tau_{\rho_n}^{(j)}\|^2$$

(Triangle inequality : $\|a-b\| - \|b-c\| \leq \|a-c\|$)

$$= \mathbb{1}(A_n \cap B_n)\max_{j\in[N]}\frac{1}{n}\sum_{i=1}^n \mathbb{1}\left(|Z_i| \leq M_n\right)\left(\frac{\nu_i^2 f'_{\hat{G}_n,\nu_i}(Z_i)}{f_{\hat{G}_n,\nu_i}(Z_i) \vee \left(\frac{\rho_n}{\nu_i}\right)} - \frac{\nu_i^2 f'_{G_j,\nu_i}(Z_i)}{f_{G_j,\nu_i}(Z_i) \vee \left(\frac{\rho_n}{\nu_i}\right)}\right)^2$$

$$\leq \omega^2$$

$$\leq \frac{\delta^2\log(1/\delta)^2}{\rho_n^2}\log(1/\rho_n). \qquad \text{(Reparametrize } \omega = 2\delta\log(1/\delta)\rho_n^{-1}\sqrt{\log(1/\rho_n)}\text{)}$$

*OA3.2.3 Controlling $\zeta_3$.* We first observe that $V_{ij} \equiv |\tau_{i,\rho_n}^{(j)} - \tau_{i,\rho_n}^*| \lesssim_{\mathcal{H}} \sqrt{\log n}$, by Lemma SM6.9. Let $V_j = (V_{1j}, \ldots, V_{nj})'$, we have that

$$\zeta_3 = \max_j(\|V_j\| - \mathbb{E}\|V_j\|)_+$$

Let $K_n = C_{\mathcal{H}}\log n \geq \max_{ij}|V_{ij}|$. Since $G_j, G_0$ are both fixed, $V_{1j}, \ldots, V_{nj}$ are mutually independent.

Observe that

$$\mathrm{P}\left(\|V_j\| > \mathbb{E}[\|V_j\|] + u\right) = \mathrm{P}\left(\left\|\frac{V_j}{K_n}\right\| \geq \mathbb{E}\left\|\frac{V_j}{K_n}\right\| + \frac{u}{K_n}\right) \leq \exp\left(-\frac{u^2}{2K_n^2}\right).$$

[47] Note that $N$ is the $\omega$ covering number for $\left\{G : \overline{h}(f_{G,\cdot}, f_{G_0,\cdot}) \leq \delta_n\right\}$, which is bounded above by the $\omega$-packing number of $\left\{G : \overline{h}(f_{G,\cdot}, f_{G_0,\cdot}) \leq \delta_n\right\}$. The packing number is further bounded above by the $\omega$-packing number of $\mathcal{P}(\mathbb{R})$, since packing numbers respect subset ordering. This is in turn bounded above by the $\omega/2$-covering number of $\mathcal{P}(\mathbb{R})$.

by Lemma OA3.8. By a union bound,

$$\mathrm{P}\left(\zeta_3^2 > x\right) \le N \exp\left(-\frac{x}{2K_n^2}\right).$$

Therefore,

$$\begin{aligned}
\mathbb{E}[\zeta_3^2] &= \int_0^\infty \mathrm{P}(\zeta_3^2 > x)\, dx \\
&= \int_0^\infty \min\left(1, N\exp\left(-\frac{x}{2K_n^2}\right)\right) dx \\
&= 2K_n^2 \log N + \int_{2K_n^2 \log N}^\infty N \exp\left(-\frac{x}{2K_n^2}\right) dx \\
&\lesssim_{\mathcal{H}} \log n \log N.
\end{aligned}$$

Now, if we take $\delta = \rho_n / n$, then

$$\frac{1}{n}\mathbb{E}[\zeta_2^2 + \zeta_3^2] \lesssim_{\mathcal{H}} \frac{(\log n)^{2.5} M_n}{n}.$$

**Remark OA3.2.** For the analogous term in the homoskedastic setting, Jiang and Zhang (2009) 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

$$\mathrm{P}\left(\|\tau_{\rho_n}^{(j)} - \tau_{\rho_n}^*\| \ge \mathbb{E}\left[\|\tau_{\rho_n}^{(j)} - \tau_{\rho_n}^*\| \mid \tau_1, \ldots, \tau_n\right] + x\right)$$

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 $\nu_i = \nu$,

$$\mathbb{E}\left[\|\tau_{\rho_n}^{(j)} - \tau_{\rho_n}^*\| \mid \tau_1, \ldots, \tau_n\right] = \mathbb{E}_{G_{0,n}}\left[\|\tau_{\rho_n}^{(j)} - \tau_{\rho_n}^*\|\right] \tag{OA3.19}$$

where $G_{0,n} = \frac{1}{n}\sum_i \delta_{\tau_i}$ is the empirical distribution of the $\tau$'s. However, (OA3.19) 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 argument in Jiang (2020) (p.2289) appears to use the Gaussian concentration of Lipschitz functions argument without the additional step.

Instead, we establish control of $\zeta_3$ 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. ■

*OA3.2.4 Controlling* $\zeta_4$. 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 $\lambda_i$ 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,

$$\begin{aligned}
\mathbb{E}(\tau_{\rho_n}^{(j)} - \tau^*_{\rho_n})^2 &= \nu_i^2 \mathbb{E}\left(\frac{f'_{G_{ji},1}(w_i)}{f_{G_{ji},1}(w_i)\vee \rho_n} - \frac{f'_{G_{0i},1}(w_i)}{f_{G_{0i},1}(w_i)\vee\rho_n}\right)^2 \\
&\lesssim_{\mathcal{H}} \max\left((\log 1/\rho_n)^3, |\log h(f_{G_{ji},1}, f_{G_{0i},1})|\right) h^2(f_{G_{ji},1}, f_{G_{0i},1}) \\
&\qquad\qquad \text{(Lemmas OA3.1 and OA3.9)} \\
&= \max\left((\log 1/\rho_n)^3, |\log h(f_{G_j,\nu_i}, f_{G_0,\nu_i})|\right) h^2(f_{G_j,\nu_i}, f_{G_0,\nu_i}) \\
&\qquad\qquad \text{(Hellinger distance is invariant to change-of-variables)}
\end{aligned}$$

Let $h_i = h(f_{G_j,\nu_i}, f_{G_0,\nu_i})$. Hence,

$$\begin{aligned}
\frac{1}{n}\mathbb{E}[\zeta_4^2] &\lesssim_{\mathcal{H}} \frac{(\log n)^3}{n}\sum_{i:|\log h_i|<(\log 1/\rho_n)^3} h_i^2 + \frac{1}{n}\sum_{i:|\log h_i|>(\log 1/\rho_n)^3} |\log h_i| h_i^2 \\
&\le (\log n)^3 \overline{h}^2(f_{G_j,\cdot}, f_{G_0,\cdot}) + \frac{1}{n}\sum_{i:|\log h_i|>(\log 1/\rho_n)^3} \frac{1}{e} h_i \qquad (x|\log x|\le e^{-1})
\end{aligned}$$

Note that

$$|\log h_i| > (\log 1/\rho_n)^3 \implies h_i < \exp\left(-\log(1/\rho_n)^3\right) < \rho_n^{(\log 1/\rho_n)^2} \lesssim_{\mathcal{H}} \rho_n^3 \lesssim_{\mathcal{H}} n^{-1}.$$

(Assumption SM6.1)

Therefore the first term dominates, and thus $\frac{1}{n}\mathbb{E}[\zeta_4^2] \lesssim_{\mathcal{H}} (\log n)^3\delta_n^2$.

**OA3.3 Auxiliary lemmas.**

**Lemma OA3.6.** *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 SM6.17. 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,

$$\left|\hat{\theta}_{i,\hat{G}_n,\hat{\eta}} - \hat{\theta}_{i,G_0,\eta_0}\right| \lesssim_{\mathcal{H}} s_{\ell n}^{-2}\overline{Z}_n. \qquad \square$$

**Lemma OA3.7.** *Let $\overline{Z}_n = \max_i |Z_i| \vee 1$. Under Assumption 2, for $t > 1$*

$$\mathrm{P}(\overline{Z}_n > t) \le n \exp\left(-C_{A_0,\alpha,\nu_u} t^\alpha\right) \quad \text{and} \quad \mathbb{E}[\overline{Z}_n^p] \lesssim_{p,\mathcal{H}} (\log n)^{p/\alpha}.$$

*Moreover, if $M_n = (C_{\mathcal{H}}+1)(C_{2,\mathcal{H}}^{-1}\log n)^{1/\alpha}$ as in (OA3.4), 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 SM6.15 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 $\mathrm{P}(Z_n > M_n) \le e^{-2\log n} = n^{-2}$. $\square$

**Lemma OA3.8.** *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

$$\|w+a\| \leq \|w\|+\|a\| \quad \|w\| = \|w+a-a\| \leq \|w+a\|+\|a\| \implies |\|w+a\|-\|w\|| \leq \|a\|.$$

Moreover, trivially $\|\lambda w + (1-\lambda)v\| \leq \lambda\|w\| + (1-\lambda)\|v\|$ for $\lambda \in [0,1]$, and hence $w \mapsto \|w\|$ is convex. Convexity implies the 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). ☐

**Lemma OA3.9.** *Let $f_H = f_{H,1}$. Then, for $0 < \rho_n \leq \frac{1}{\sqrt{2\pi e^2}}$,*

$$\int \left[\frac{f'_{H_1}(x)}{f_{H_1}(x) \vee \rho_n} - \frac{f'_{H_0}(x)}{f_{H_0}(x) \vee \rho_n}\right]^2 f_{H_0}(x)\, dx \lesssim \left((\log 1/\rho_n)^3 \vee |\log h\left(f_{H_1}, f_{H_0}\right)|\right) h^2\left(f_{H_1}, f_{H_0}\right)$$

*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 $r^2(f_{H_1}, \rho_n)$. Their subsequent calculation, which involves Lemma 1 of Jiang and Zhang (2009), proceeds to bound

$$r^2(f_{H_1}, \rho_n) \leq 2\sqrt{2}eh(f_{H_1}, f_{H_0}) \max\left(\varphi_+^3(\rho_n), \sqrt{2}a\right) + 2\varphi_+(\rho_n)\sqrt{2}h(f_{H_1}, f_{H_0}),$$

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 $h, \log h$, squaring, and using $\varphi_+(\rho_n) \lesssim \sqrt{\log(1/\rho_n)}$ proves the claim. ☐

## Part 2 Additional discussions and empirical results

### Appendix OA4. Additional discussions

#### OA4.1 Alternatives to CLOSE.

*OA4.1.1 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 OA4.1.

TABLE OA4.1. 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 $\theta_i$ is independent from some other known nuisance parameter, e.g. $n_i$ | | ✗ | ✗ | |
| Parametric restrictions on the micro-data | | ✗ | ✗ | |

**Alternative 1** (Working with $t$-ratios)**.** We may consider normalizing $\sigma_i$ 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 $\theta_i$ (Jiang and Zhang, 2010). However, such an approach approximates the optimal decision rule within a restricted class on a different objective.

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 $\theta_i/\sigma_i$ 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 $\theta_i$ 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 $\omega_i$ 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 $\theta_i$. We do not know of any guarantees for the loss function on $\theta_i$, $\frac{1}{n}\sum_{i=1}^n(\delta_i - \sin^2(\omega_i/2))^2$, when we perform empirical Bayes analysis on $\omega_i$. ■

**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 $\sigma_i$ as noisy as well, in which we know the likelihood of $(Y_i, \sigma_i)$. This approach allows for dependence between $\theta_i$ and $\sigma_i$ but assumes independence between $(\theta_i, \sigma_i)$ 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 $\sigma_{i0}^2$ denote the true variance of observations from population $i$.

Both papers work under Gaussian assumptions on the micro-data. This parametric assumption[48] 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 $\theta_i$ and $\sigma_{i0}^2$. 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 $\theta_i$ and $\sigma_i$ cannot be easily resolved. As summarized in Table OA4.1, existing alternatives compromise on optimality, use a different loss function, or implicitly assume $\theta_i$ is independent from components of $\sigma_i^2$ (e.g., $n_i$). Of course, depending on the empirical context, these may well be reasonable features.

[48] 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, $\Sigma_i$ also depends on $n_i$ and higher moments of $Y_{ij}$, which again may not be independent of $\boldsymbol{\theta}_i$.

In contrast, our approach models $\theta_i \mid \sigma_i$ directly via the location-scale assumption (2.4). 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.

*OA4.1.2 Alternative models for $\theta_i \mid \sigma_i$.* One alternative is simply treating the joint distribution of $(\theta_i, \sigma_i)$ fully nonparametrically. For instance, an $f$-modeling approach with Tweedie's formula[49] 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 $g$-modeling of $\theta_i \mid \sigma_i$ in the vein of the log-spline sieve of Efron (2016).[50] 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 $(\theta_i, \sigma_i)$, it is fair to ask why the conditional location-scale restriction (2.4) 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 OA4.2 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., $Z = a(\sigma) + b(\sigma)Y$). If we further wish that $Z$ obeys a

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

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

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 $\theta_i$.

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 $\sigma_i$ 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.

[50]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 $\alpha_j$. The parameters $\alpha_1, \ldots, \alpha_J$ can be estimated by maximizing the penalized likelihood of $Y_{1:n}$.

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 (2.4). 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.

**OA4.2 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 $(Z_i, \tau_i, \nu_i)$ 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), 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:

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

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-$\sigma$ regime (Chen and Niles-Weed, 2022).

## Appendix OA5. Additional empirical exercises

**OA5.1 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 $Y_i, \sigma_i$ by $\sigma_i$, 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

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

Hence $S_i$ is an overestimate of the successive averages of $s_0(\sigma)$. Figure OA5.1 plot the estimated conditional expectation of $S_i$ given $\sigma_i$, 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.

**OA5.2 Simulation exercise setup.** This section describes the details of the simulation exercise in Section 4. 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 $\beta$ and residuals $Y_i$. Estimate the conditional moments $m_0, s_0$ on $(Y_i, \sigma_i)$ via local linear regression, described in Appendix SM8.

(Sim-2) Partition $\sigma$ 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 $\beta, m_0, s_0$ 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. Figure OA5.2 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.

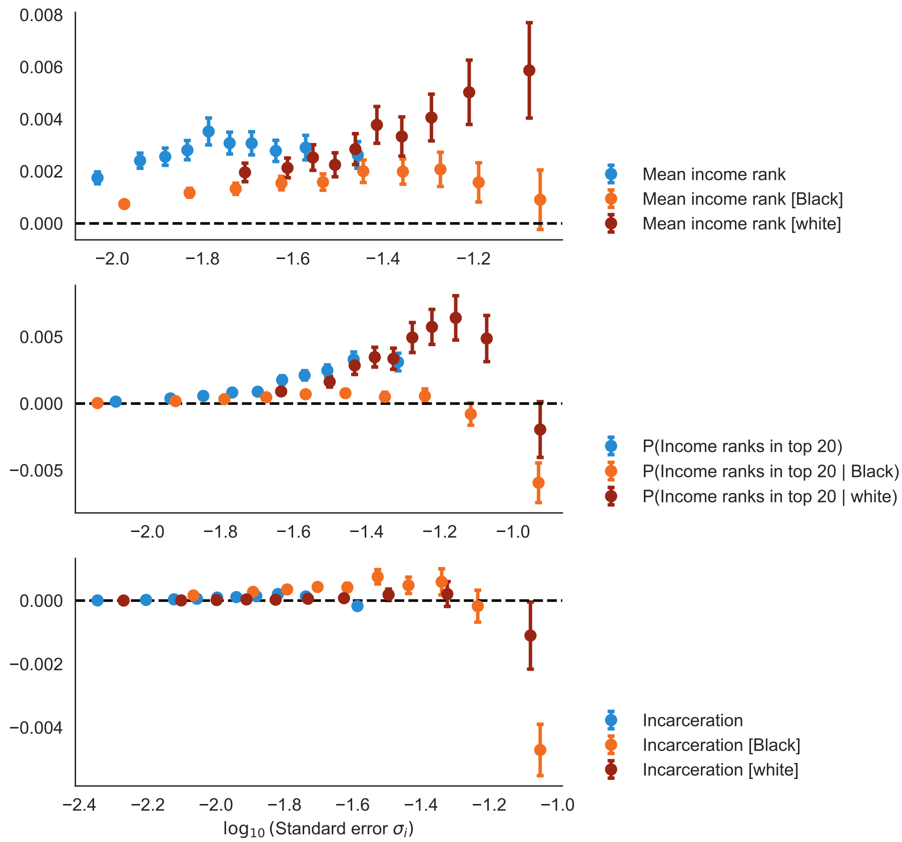



FIGURE OA5.1. Estimated conditional variance $s_0^2(\sigma)$, binned into deciles, with 95% uniform confidence intervals shown.

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. (2020)'s 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. (2024). Bergman et al. (2024) 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 $n_i$.

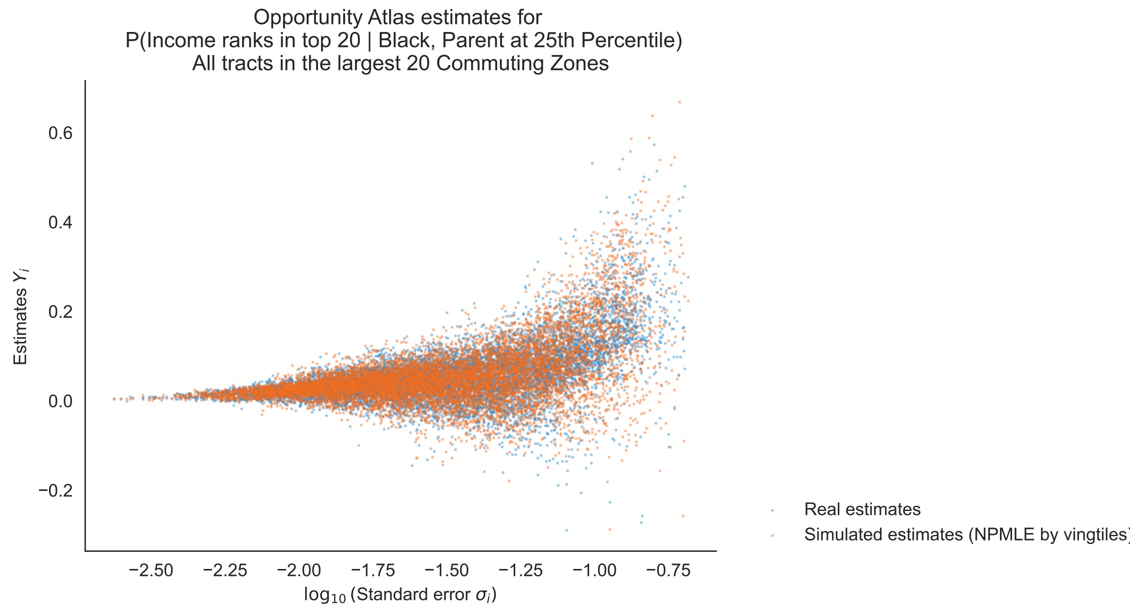



FIGURE OA5.2. A draw of real vs. simulated data for estimates of TOP-20 PROBABILITY for Black individuals

**OA5.3 Robustness checks for the calibration exercise in Section 4.** In Figure OA5.3, 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 $\sigma^2$. 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.

**OA5.4 Different simulation setup.** We have also conducted a Monte Carlo exercise where we replace (Sim-2) with the following step:

- For each $\sigma_i$, 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 $\alpha_i$. 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 $\alpha_i$ corresponds exactly to $\alpha$ in Assumption 2.

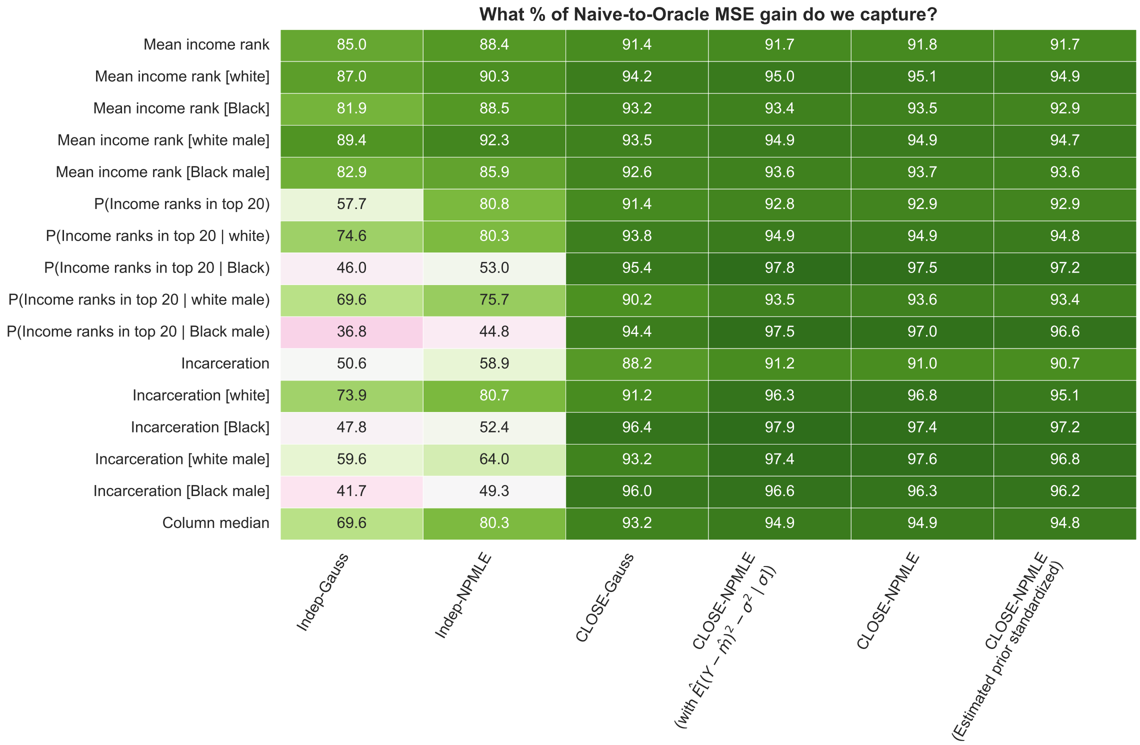



FIGURE OA5.3. Additional CLOSE-NPMLE variants for the calibrated simulation in Section 4. Here the results average over 100 replications.

Our choices of $\alpha_i$ 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 OA5.4 shows the analogue of Figure 3 for this data-generating process. Indeed, we see that INDEPENDENT-NPMLE improves over INDEPENDENT-GAUSS considerably, and similarly for CLOSE-NPMLE and ORACLE-GAUSS.

**OA5.5 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 4. 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

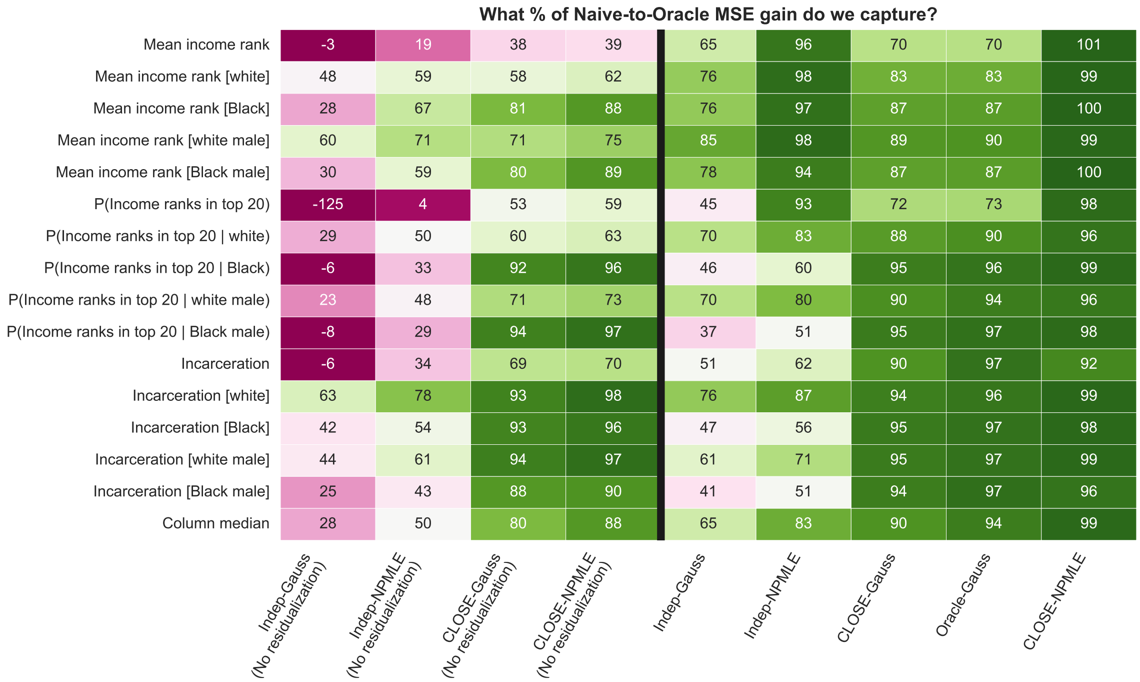



FIGURE OA5.4. Analogue of Figure 3 for the data-generating process in Appendix OA5.4. Here the results average over 100 replications.

against NAIVE as a multiple of this "value of basic empirical Bayes." Figure OA5.5(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 OA5.5(b).

We find that our empirical patterns from the calibrated simulation Figure 3 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 $\vartheta_i$, 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.

**OA5.6 Empirical Bayes pooling over all Commuting Zones in validation exercise.** Here, we repeat the exercise in Figure 4, but we now estimate empirical Bayes methods pooling over all Commuting Zones. We still pick the top third of every Commuting Zone. Our first exercise repeats Figure 4 in this setting, shown in Figure OA5.6. The results are extremely similar.

Separately, we consider the version of this exercise without covariates in Figure OA5.10. We see that covariates are extremely important for the performance of INDEPENDENT-GAUSS, as it frequently underperforms NAIVE without covariates.[51] By comparison, they are less important for the performance of CLOSE-NPMLE, as $\sigma_i$ contains a lot of the signal in the tract-level covariates.

**OA5.7 The tradeoff between accurate targeting and estimation precision.** In this section, we investigate the tradeoff between accurate targeting and estimation precision. That is, suppose $\theta_i, Y_i, \sigma_i$ and $\vartheta_i, \Upsilon_i, \varsigma_i$ are two sets variables corresponding to two measures of economic mobility. For instance, perhaps $\theta_i$ is MEAN RANK for Black individuals and $\vartheta_i$ is MEAN RANK pooling over all individuals. Suppose the decision maker would like to select populations with high $\theta_i$, but the estimates $Y_i$ are noisier than the estimates $\Upsilon_i$. It is plausible that screening on posterior means for $\vartheta_i$ might outperform screening on posterior means for $\theta_i$.

We investigate this question via coupled bootstrap in the Bergman et al. (2024) 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.[52] Let $\theta_i = \pi_i\theta_{ib} + (1-\pi_i)\theta_{iw}$ be a pooled measure, where

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

[51] This is in part since our implementation of INDEPENDENT-GAUSS uses weighted means for estimating the prior parameters, worsening the misspecification. See Footnote 37.

[52] 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.

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 $\pi_i$-combination of bootstrap draws for $Y_{ib}, Y_{iw}$.

Here, we investigate whether screening tracts based on posterior mean estimates for $\theta_{iw}$ or $\theta_i$ generates better decisions in terms of $\theta_{ib}$, owing to the precision in $Y_{iw}$ and $Y_i$. Figure OA5.11 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 $\theta_{ib}$. Dots on the red (resp. blue) dashed line correspond to screening on $\theta_{iw}$ (resp. $\theta_i$). We see that for all three measures of economic mobility, using CLOSE-NPMLE to screen on the original parameter $\theta_{ib}$ performs best. In other words, the benefits of higher precision are insufficient to offset inaccurate targeting.

## References


Bergman, P., R. Chetty, S. DeLuca, N. Hendren, L. F. Katz, and C. Palmer (2024). "Creating Moves to Opportunity: Experimental Evidence on Barriers to Neighborhood Choice". In: *American Economic Review* (pp. 65, 69, 74).

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press (pp. 56, 58, 59).

Brown, L. D. (2008). "In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies". In: (pp. 60).

Brown, L. D. and E. Greenshtein (2009). "Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means". In: *The Annals of Statistics*, pp. 1685–1704 (pp. 62).

Chen, H.-B. and J. Niles-Weed (2022). "Asymptotics of smoothed Wasserstein distances". In: *Potential Analysis* 56.4, pp. 571–595 (pp. 63).

Chetty, R., J. N. Friedman, N. Hendren, M. R. Jones, and S. R. Porter (2020). *The opportunity atlas: Mapping the childhood roots of social mobility*. Tech. rep. (pp. 65, 69).

Efron, B. (2014). "Two modeling strategies for empirical Bayes estimation". In: *Statistical science: a review journal of the Institute of Mathematical Statistics* 29.2, p. 285 (pp. 62).

– (2016). "Empirical Bayes deconvolution estimates". In: *Biometrika* 103.1, pp. 1–20 (pp. 62).

Efron, B. and C. Morris (1975). "Data analysis using Stein's estimator and its generalizations". In: *Journal of the American Statistical Association* 70.350, pp. 311–319 (pp. 60).

Fu, L. J., G. M. James, and W. Sun (2020). "Nonparametric empirical bayes estimation on heterogeneous data". In: *arXiv preprint arXiv:2002.12586* (pp. 61).

Gu, J. and R. Koenker (2017). “Empirical Bayesball remixed: Empirical Bayes methods for longitudinal data”. In: *Journal of Applied Econometrics* 32.3, pp. 575–599 (pp. 61).

Jiang, W. (2020). “On general maximum likelihood empirical Bayes estimation of heteroscedastic IID normal means”. In: *Electronic Journal of Statistics* 14.1, pp. 2272–2297 (pp. 54, 56, 57, 63).

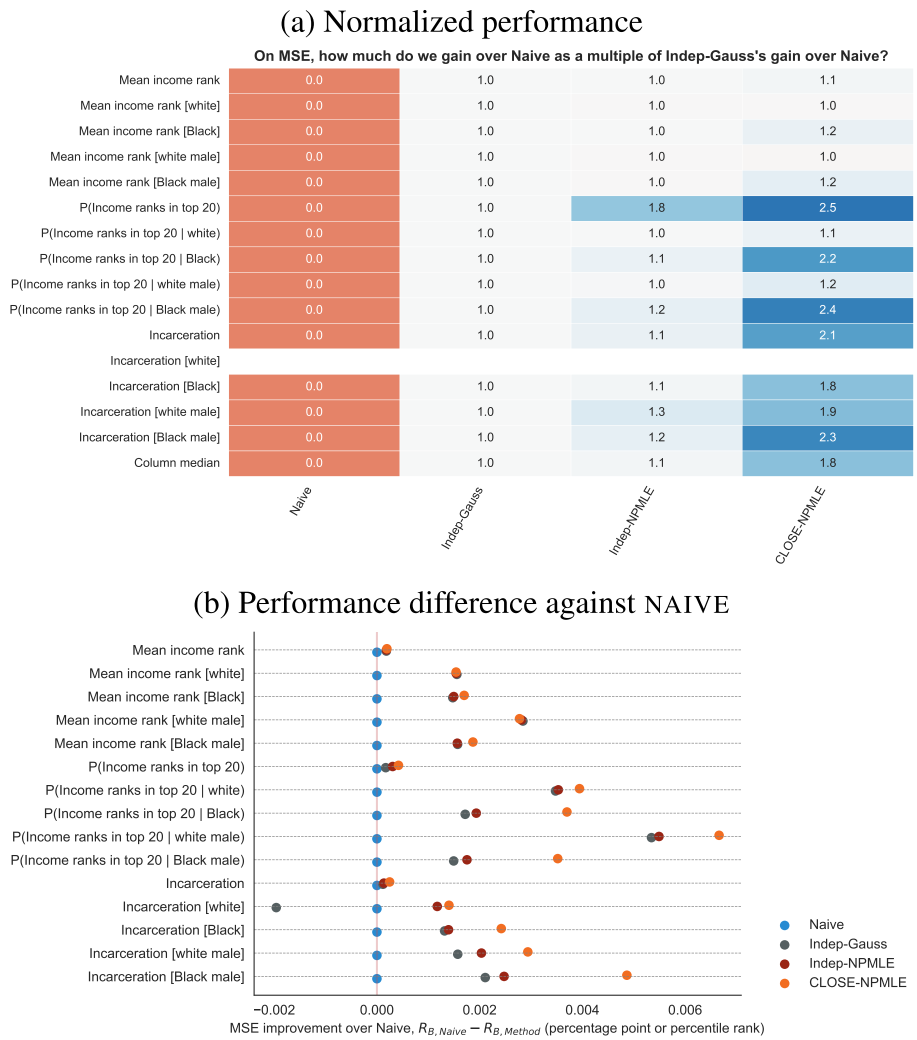
Jiang, W. and C.-H. Zhang (2009). “General maximum likelihood empirical Bayes estimation of normal means”. In: *The Annals of Statistics* 37.4, pp. 1647–1684 (pp. 54, 56, 59, 62, 63).

– (2010). “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, pp. 263–274 (pp. 60).

Liu, L., H. R. Moon, and F. Schorfheide (2020). “Forecasting with dynamic panel data models”. In: *Econometrica* 88.1, pp. 171–201 (pp. 62).

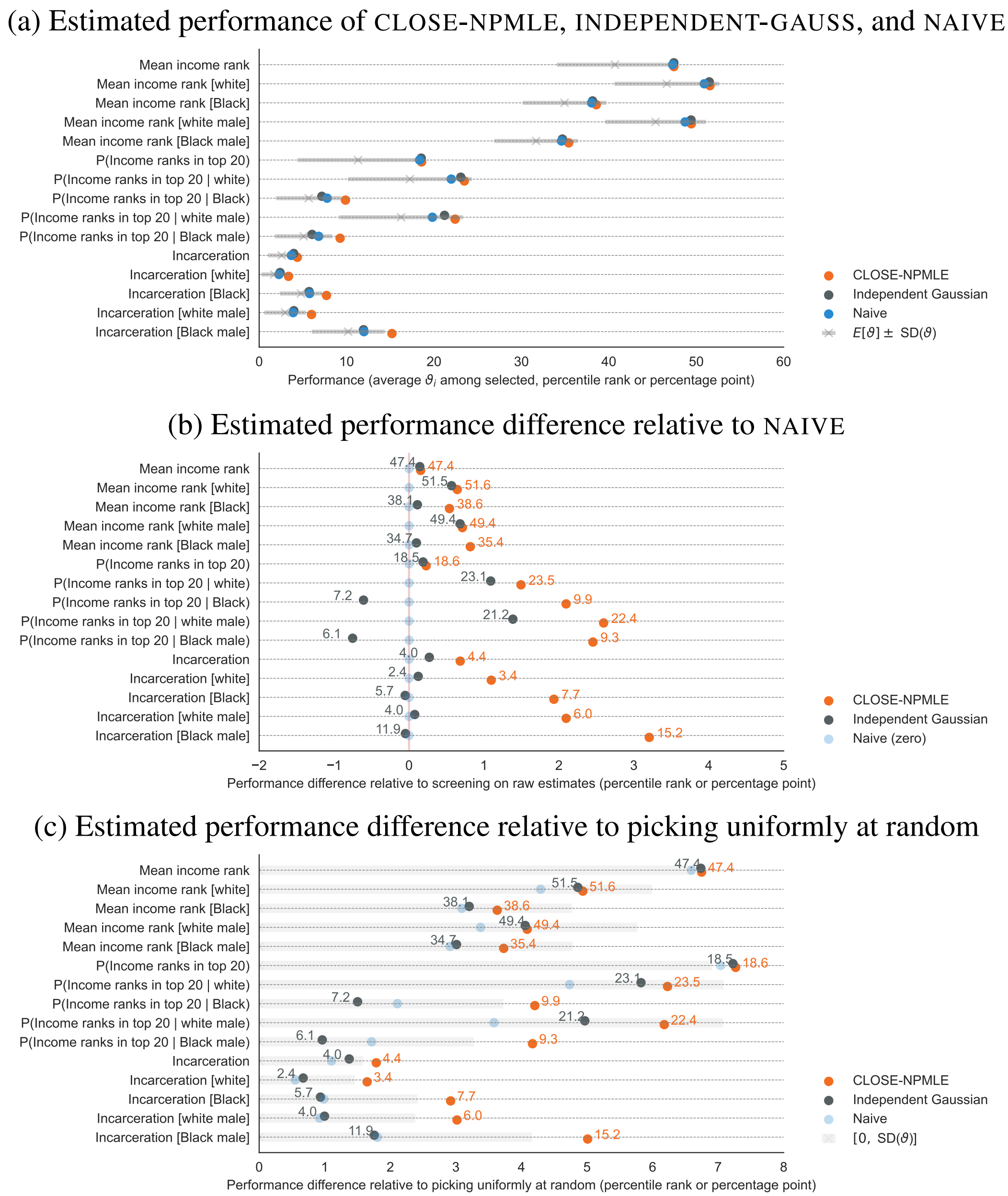
Luo, J., T. Banerjee, G. Mukherjee, and W. Sun (2023). *Empirical Bayes Estimation with Side Information: A Nonparametric Integrative Tweedie Approach*. arXiv: 2308.05883 [stat.ME] (pp. 62).

Soloff, J. A., A. Guntuboyina, and B. Sen (2021). “Multivariate, heteroscedastic empirical bayes via nonparametric maximum likelihood”. In: *arXiv preprint arXiv:2109.03466* (pp. 54, 56).




*Notes.* In panel (a), each column is an empirical Bayes strategy that we consider, and each row is a different definition of $\theta_i$. 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 OA5.5. Estimated MSE Bayes risk for various empirical Bayes strategies in the validation exercise.

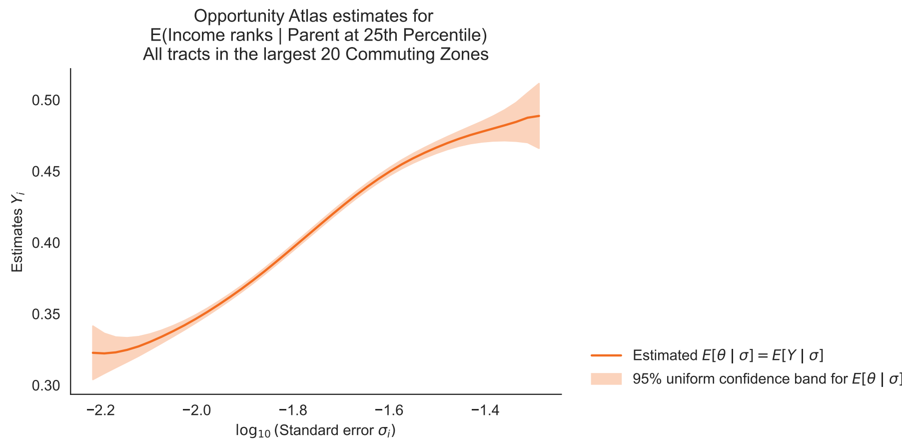



*Notes.* These figures show the estimated performance of various decision rules over 100 coupled bootstrap draws. Performance is measured as the mean $\vartheta_i$ among selected Census tracts. All decision rules select the top third of Census tracts within each Commuting Zone. Figure (a) plots the estimated 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 $\vartheta$. □

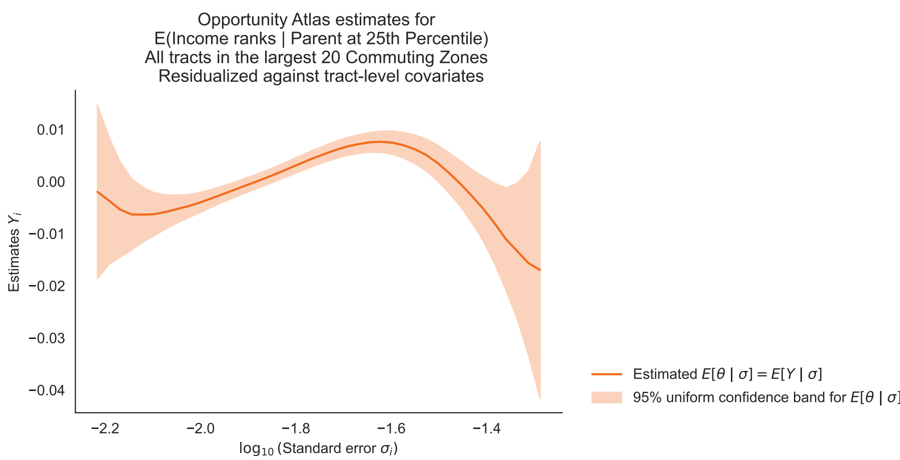


FIGURE OA5.6. Performance of decision rules in top-$m$ selection exercise

(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. (2024). The estimation and the confidence band procedures are the same as those in Figure 1. In panel (a), $\theta_i, Y_i$ are defined as unresidualized measures of mean income rank. In panel (b), we treat $\theta_i, Y_i$ as residualized against a vector of tract-level covariates as specified in Appendix OA5.2. □

FIGURE OA5.7. Estimated $\mathbb{E}[\theta \mid \sigma]$ for mean income rank among those with parents at the 25th percentile

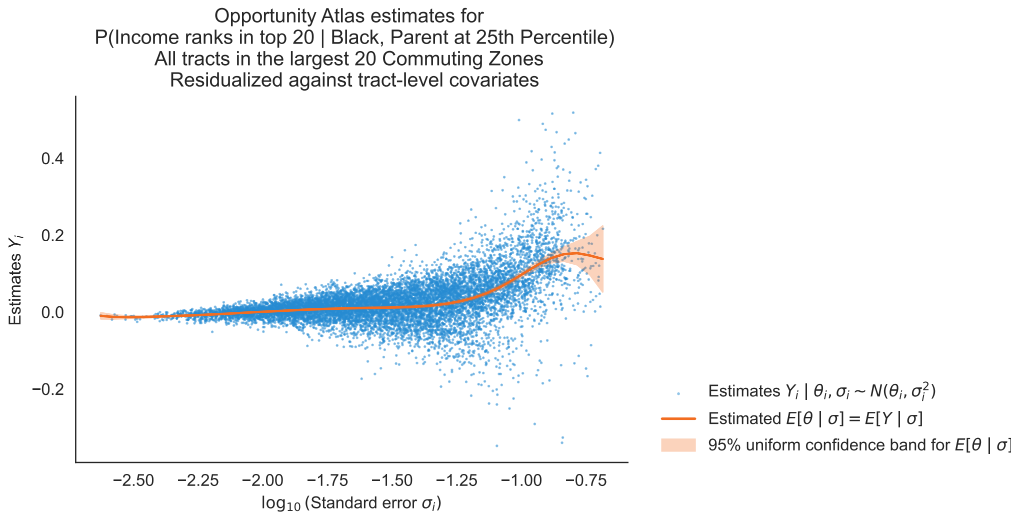



FIGURE OA5.8. The analogue of Figure 1 where $Y_i, \theta_i$ are treated as residualized against a vector of covariates as specified in Appendix OA5.2.

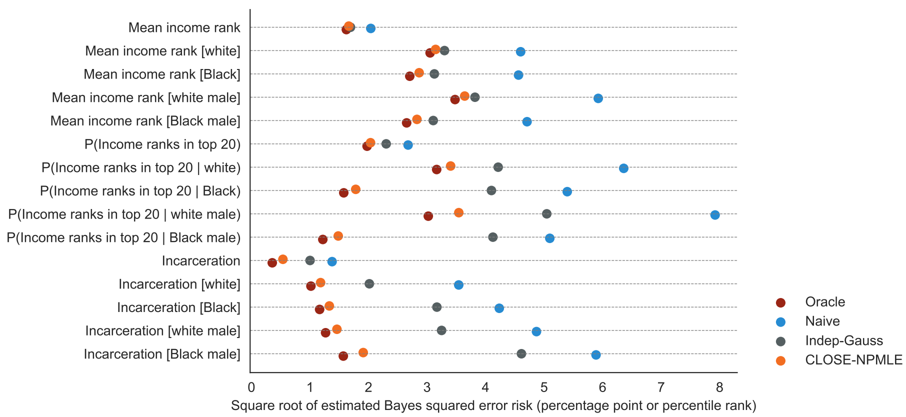
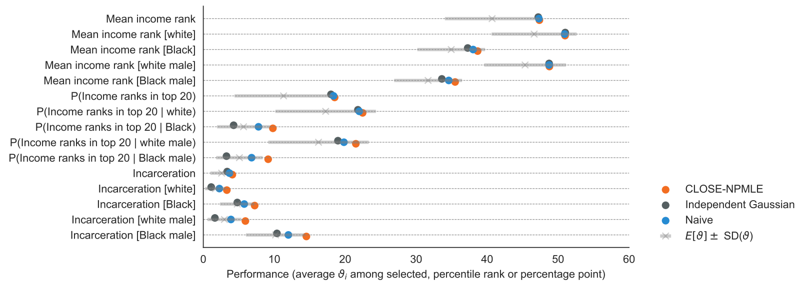
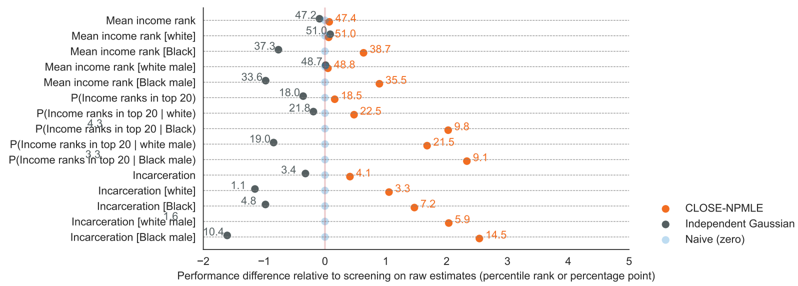



FIGURE OA5.9. Absolute mean-squared error risk of key methods for the calibrated simulation in Figure 3.

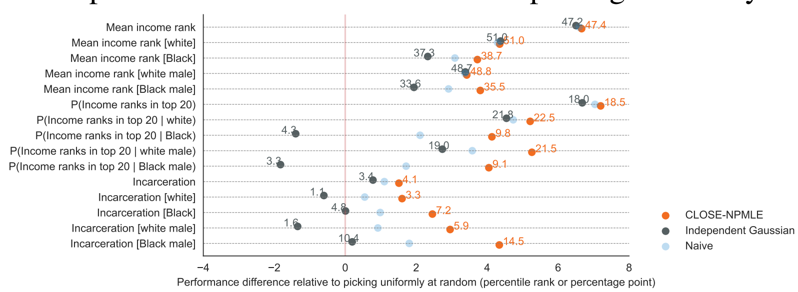
(a) Estimated performance of CLOSE-NPMLE, INDEPENDENT-GAUSS, and NAIVE




(b) Estimated performance difference relative to NAIVE

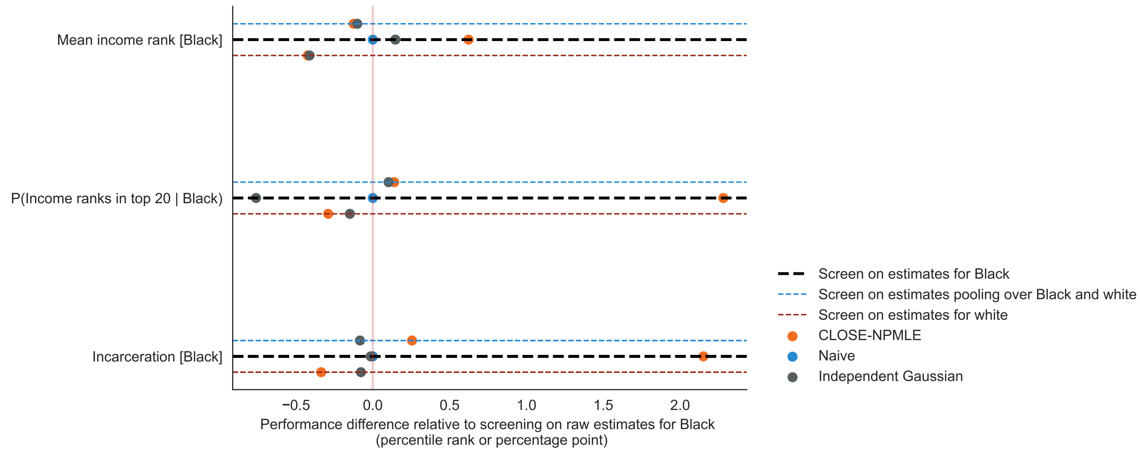



(c) Estimated performance difference relative to picking uniformly at random




*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 $\vartheta_i$ 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 $\vartheta$. □

FIGURE OA5.10. 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 OA5.11. Performances of strategies that screen on posterior means for more precisely estimated parameters

# Supplementary Material to
# "Empirical Bayes When Estimation Precision Predicts Parameters"

Jiafeng Chen

April 8, 2024

**Contents**

## Part 3 Important preliminary results for Theorem 1

### Appendix SM6. An oracle inequality for the likelihood

Recall that for some fixed $\Delta_n, M_n$, we define $A_n = \left\{\|\hat{\eta} - \eta\|_\infty \le \Delta_n, \overline{Z}_n \le M_n\right\}$ in (A.1). In this section, we bound

$$\mathrm{P}\left[A_n, \mathrm{Sub}_n(\hat{G}_n) \gtrsim_{\mathcal{H}} \epsilon_n\right],$$

where we recall $\mathrm{Sub}_n$ from (A.2), for some rate function $\epsilon_n$. It is convenient to state a set of high-level assumptions on the rates $\Delta_n, M_n$. These are satisfied for the choice (OA3.4) for our main results (Theorem A.1).

**Assumption SM6.1.** *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)* $\sqrt{\log n} \lesssim_{\mathcal{H}} M_n$

Our main result in this section is the following oracle inequality.

**Theorem SM6.1.** *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 SM6.1, there exists constants* $C_{1,\mathcal{H}}, C_{2,\mathcal{H}} > 0$ *such that the following tail bound holds: Let*

$$\epsilon_n = M_n \sqrt{\log n}\Delta_n \frac{1}{n}\sum_{i=1}^n h\left(f_{\hat{G}_n,\nu_i}, f_{G_0,\nu_i}\right) + \Delta_n M_n \sqrt{\log n} e^{-C_{2,\mathcal{H}} M_n^\alpha} + \Delta_n^2 M_n^2 \log n + M_n^2 \frac{\Delta_n^{1-\frac{1}{2p}}}{\sqrt{n}}. \tag{SM6.1}$$

*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 concrete rates for $\Delta_n, M_n$ (OA3.4) and verifies that they satisfy Assumption SM6.1.

**Corollary SM6.1.** *For* $\beta \ge 0$, *suppose* $\Delta_n, M_n$ *are of the form (OA3.4). Then there exists a* $C_{\mathcal{H}}^*$ *such that the following tail bound holds. Recall the average Hellinger distance* $\overline{h}$ *from (A.4). Suppose* $\hat{G}_n$ *satisfies Assumption 1. Under Assumptions 2 to 4, define* $\varepsilon_n$ *as:*

$$\varepsilon_n = n^{-\frac{p}{2p+1}} (\log n)^{\frac{2+\alpha}{2\alpha}+\beta} \overline{h}\left(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}\right) + n^{-\frac{2p}{2p+1}} (\log n)^{\frac{2+\alpha}{\alpha}+2\beta}, \tag{SM6.2}$$

*we have that,* $\mathrm{P}\left[A_n, \mathrm{Sub}_n(\hat{G}_n) > C_{\mathcal{H}}^* \varepsilon_n\right] \le \frac{9}{n}$. *The constant* $C_{\mathcal{H}}$ *in* $\Delta_n, M_n$ *affects the conclusion of the statement only through affecting the constant* $C_{\mathcal{H}}^*$.

*Proof.* We first show that the specification of $\Delta_n$ and $M_n$ means that the requirements of Assumption SM6.1 are satisfied. Among the requirements of Assumption SM6.1:

(1) is satisfied since the polynomial part of $\Delta_n$ 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,

$$\frac{1}{n}\sum_i h(f_{\hat{G}_n,\nu_i}, f_{G_0,\nu_i}) \le \overline{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}),$$

where we recall $\overline{h}$ from (A.4), and so we can replace the corresponding factor in $\epsilon_n$ by $\overline{h}$. Now, we plug the rates $\Delta_n, M_n$ into $\epsilon_n$. We find that the term

$$\Delta_n M_n^2 e^{-C_{2,\mathcal{H}} M_n^\alpha} = \Delta_n M_n^2 e^{-(C_{\mathcal{H}}+1)^\alpha (\log n)} \le \Delta_n M_n^2 n^{-1} \lesssim_{\mathcal{H}} \Delta_n^2 M_n^2 \log n$$

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 SM6.1 follows from Theorem SM6.1. □

**SM6.1 Proof of Theorem SM6.1.**

*SM6.1.1 Decomposition of* $\mathrm{Sub}_n(\hat{G}_n)$. Observe that, by (3.2) in Assumption 1,

$$\frac{1}{n}\sum_{i=1}^n \psi_i(Z_i, \hat{\eta}, \hat{G}_n) - \frac{1}{n}\sum_{i=1}^n \psi_i(Z_i, \hat{\eta}, G_0) \ge \kappa_n$$

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 therefore

$$\mathrm{Sub}_n(\hat{G}_n) \le a_n + b_n + \kappa_n.$$

Therefore, it suffices to show large deviation results for $a_n$ and $b_n$, where $a_n$ is chosen to be (SM6.7) and $b_n$ is chosen to be (SM6.10).

*SM6.1.2 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 (A.1). 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:

$$\psi_i\left(Z_i, \hat{\eta}, \hat{G}_n\right) - \psi_i\left(Z_i, \eta_0, \hat{G}_n\right) = \frac{\partial \psi_i}{\partial m_i}\bigg|_{\eta_0, \hat{G}_n} \Delta_{mi} + \frac{\partial \psi_i}{\partial s_i}\bigg|_{\eta_0, \hat{G}_n} \Delta_{si} + \underbrace{\frac{1}{2}\Delta_i' H_i(\tilde{\eta}_i, \hat{G}_n)\Delta_i}_{R_{1i}} \quad \text{(SM6.3)}$$

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 $\eta_{0i}$.

We further decompose the first-order terms into an empirical process term and a mean-component term. By Lemma OA3.1, (SM6.34), and (SM6.36), for $\rho_n$ in (OA3.5) we have that the denominators to the first derivatives can be truncated at $\rho_n$, as $f_{i,\hat{G}_n} \geq \rho_n/\nu_i$ so that the truncation does not bind:

$$\left.\frac{\partial \psi_i}{\partial m_i}\right|_{\eta_0,\hat{G}_n} = -\frac{1}{s_i}\frac{f'_{i,\hat{G}_n}}{f_{i,\hat{G}_n} \vee \frac{\rho_n}{\nu_i}} \equiv D_{m,i}(Z_i, \hat{G}_n, \eta_0, \rho_n) \tag{SM6.4}$$

$$\left.\frac{\partial \psi_i}{\partial s_i}\right|_{\eta_0,\hat{G}_n} = \frac{s_i}{\sigma_i^2}\frac{Q_i(Z_i, \eta_0, \hat{G}_n)}{f_{i,\hat{G}_n} \vee \frac{\rho_n}{\nu_i}} \equiv D_{s,i}(Z_i, \hat{G}_n, \eta_0, \rho_n). \tag{SM6.5}$$

where we recall $Q_i$ from (SM6.40).

Let

$$\overline{D}_{k,i}(\hat{G}_n, \eta_0, \rho_n) = \int D_{k,i}(z, \hat{G}_n, \eta_0, \rho_n)\, f_{G_0,\nu_i}(z)dz \quad \text{for } k \in \{m, s\} \tag{SM6.6}$$

be the population mean of $D_{k,i}$. Then, for $k \in \{m, s\}$, we can decompose

$$\left.\frac{\partial \psi_i}{\partial k_i}\right|_{\eta_0,\hat{G}_n} \Delta_{ki} = \left[D_{k,i}(Z_i, \hat{G}_n, \eta_0, \rho_n) - \overline{D}_{k,i}(\hat{G}_n, \eta_0, \rho_n)\right] \Delta_{ki} + \overline{D}_{k,i}(\hat{G}_n, \eta_0, \rho_n)\Delta_{ki}$$

Hence, we can decompose the first-order terms in (SM6.3):

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \left.\frac{\partial \psi_i}{\partial k_i}\right|_{\eta_0,\hat{G}_n} \Delta_{ki} &= \frac{1}{n}\sum_{i=1}^n \overline{D}_{k,i}(\hat{G}_n, \eta_0, \rho_n)\Delta_{ki} \\ &\quad + \frac{1}{n}\sum_{i=1}^n \left[D_{k,i}(Z_i, \hat{G}_n, \eta_0, \rho_n) - \overline{D}_{k,i}(\hat{G}_n, \eta_0, \rho_n)\right] \Delta_{ki} \\ &\equiv U_{1k} + U_{2k}\end{aligned}$$

Let the second order term in (SM6.3) be $R_1 = \frac{1}{n}\sum_i R_{1i}$. We let

$$a_n = |R_1| + \sum_{k\in\{m,s\}} |U_{1k}| + |U_{2k}| \tag{SM6.7}$$

*SM6.1.3 Taylor expansion of $\psi_i(Z_i, \hat{\eta}, G_0) - \psi_i(Z_i, \eta_0, G_0)$.* Like (SM6.3), we similarly decompose

$$\psi_i(Z_i, \hat{\eta}, G_0) - \psi_i(Z_i, \eta_0, G_0) = \left.\frac{\partial \psi_i}{\partial m_i}\right|_{\eta_0,G_0} \Delta_{mi} + \left.\frac{\partial \psi_i}{\partial s_i}\right|_{\eta_0,G_0} \Delta_{si} + \underbrace{\frac{1}{2}\Delta_i' H_i(\tilde{\eta}_i, G_0)\Delta_i}_{R_{2i}} \tag{SM6.8}$$

$$= \sum_{k\in\{m,s\}} D_{k,i}(Z_i, G_0, \eta_0, 0)\Delta_{ki} + R_{2i}$$

$$\equiv U_{3mi} + U_{3si} + R_{2i}. \tag{SM6.9}$$

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

$$b_n = |R_2| + \sum_{k\in\{m,s\}} |U_{3k}| + |U_{3k}| \tag{SM6.10}$$

*SM6.1.4 Bounding each term individually.* By our decomposition, we can write

$$a_n + b_n + \kappa_n \leq \kappa_n + |R_1| + |R_2| + \sum_{k\in\{m,s\}} |U_{1k}| + |U_{2k}| + |U_{3k}|$$

To summarize, we have that, for $k = m, s$,

$$U_{1k} = \frac{1}{n}\sum_{i=1}^{n} \overline{D}_{k,i}(\hat{G}_n, \eta_0, \rho_n)\Delta_{ki} \tag{SM6.11}$$

$$U_{2k} = \frac{1}{n}\sum_{i=1}^{n} \left[D_{k,i}(Z_i, \hat{G}_n, \eta_0, \rho_n) - \overline{D}_{k,i}(\hat{G}_n, \eta_0, \rho_n)\right]\Delta_{ki} \tag{SM6.12}$$

$$U_{3k} = \frac{1}{n}\sum_{i=1}^{n} D_{k,i}(Z_i, G_0, \eta_0, 0)\Delta_{ki} \tag{SM6.13}$$

$$R_1 = \frac{1}{2n}\sum_{i=1}^{n} \Delta_i' H_i(\tilde{\eta}_i, \hat{G}_n)\Delta_i \tag{SM6.14}$$

$$R_2 = \frac{1}{2n}\sum_{i=1}^{n} \Delta_i' H_i(\tilde{\eta}_i, \hat{G}_0)\Delta_i \tag{SM6.15}$$

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 Lemma SM6.1 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 SM6.11. Similarly, we bound $\mathbb{1}(A_n)|U_{1s}|$ in Lemma SM6.2.

(2) The empirical process terms $U_{2m}, U_{2s}$ are bounded probabilistically in Lemmas SM6.3 and SM6.4 with statements of the form

$$\mathrm{P}(A_n, |U_{2k}| > c_1) \le c_2.$$

To do so, we upper bound $\mathbb{1}(A_n)U_{2k} \le \overline{U}_{2k}$. The upper bound is obtained by projecting $\hat{G}_n$ onto a $\omega$-net of $\mathcal{P}(\mathbb{R})$ in terms of some pseudo-metric $d_{k,\infty,M_n}$ induced by $\overline{D}_{k,i}$. The upper bound $\overline{U}_{2k}$ then takes the form, for $\eta \in S$ over a Hölder space,

$$\omega\Delta_n + \max_{j\in[N]}\sup_{\eta\in S}\left|\frac{1}{n}\sum_i (D_{ki} - \overline{D}_{ki})(\eta_i - \eta_{0i})\right| \quad N \le N(\omega, \mathcal{P}(\mathbb{R}), d_{k,\infty,M_n}).$$

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 SM6.1 and Proposition SM6.2, 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 $f_{G,\nu_i}$ 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 projection-to-$\omega$-net argument above is unnecessary for $U_{3m}, U_{3s}$, whereas the bounding follows from the same Dudley's chaining argument. Lemma SM6.5 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 SM6.17, 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. These second derivatives are bounded in Lemmas SM6.6 and SM6.7.

(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 SM6.16, we can simply add the rates.

Doing so, we find that the first term in $\epsilon_n$ (SM6.1) 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 $\epsilon_n$ dominate $\kappa_n$, recalling (3.3). This completes the proof.

### SM6.2 Bounding $U_{1m}$.

**Lemma SM6.1.** *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 SM6.1. Then*

$$|U_{1m}| \equiv \left|\frac{1}{n}\sum_{i=1}^n \overline{D}_{mi}(\hat{G}_n, \eta_0, \rho_n)\Delta_{mi}\right| \lesssim_{\mathcal{H}} \Delta_n \left[\frac{\log n}{n}\sum_{i=1}^n h(f_{G_0,\nu_i}, f_{\hat{G}_n,\nu_I}) + \frac{M_n^{1/3}}{n}\right]. \tag{SM6.16}$$

*Proof.* Note that

$$\begin{aligned}
|\overline{D}_{m,i}(\hat{G}_n,\eta_0,\rho_n)| = |(\text{SM6.4})| \lesssim_{s_0\ell} & \left|\int \frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z)\vee\frac{\rho_n}{\nu_i}} f_{G_0,\nu_i}(z)dz\right| \\
&= \left|\int \frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z)\vee\frac{\rho_n}{\nu_i}} [f_{G_0,\nu_i}(z) - f_{\hat{G}_n,\nu_i}(z) + f_{\hat{G}_n,\nu_i}(z)]dz\right| \\
&\le \left|\int \frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z)\vee\frac{\rho_n}{\nu_i}} [f_{G_0,\nu_i}(z) - f_{\hat{G}_n,\nu_i}(z)]\,dz\right| \qquad \text{(SM6.17)}\\
&\quad + \left|\int \frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z)\vee\frac{\rho_n}{\nu_i}} f_{\hat{G}_n,\nu_i}(z)dz\right| \qquad \text{(SM6.18)}
\end{aligned}$$

By the bounds for (SM6.17) and (SM6.18) 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 SM6.1. □

*SM6.2.1 Bounding* (SM6.17). Consider the first term (SM6.17):

$$\begin{aligned}
[(\text{SM6.17})]^2 &= \left[\int \frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z)\vee\frac{\rho_n}{\nu_i}}\left(\sqrt{f_{G_0,\nu_i}(z)} - \sqrt{f_{\hat{G}_n,\nu_i}(z)}\right)\left(\sqrt{f_{G_0,\nu_i}(z)} + \sqrt{f_{\hat{G}_n,\nu_i}(z)}\right)\,dz\right]^2 \\
&\le \underbrace{\int \left(\sqrt{f_{G_0,\nu_i}(z)} - \sqrt{f_{\hat{G}_n,\nu_i}(z)}\right)^2 dz}_{2h^2}
\end{aligned}$$

$$\cdot \int \left(\frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}}\right)^2 \left(\sqrt{f_{G_0,\nu_i}(z)} + \sqrt{f_{\hat{G}_n,\nu_i}(z)}\right)^2 dz \qquad \text{(Cauchy–Schwarz)}$$

$$\lesssim h^2(f_{G_0,\nu_i}, f_{\hat{G}_n,\nu_i}) \int \left(\frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}}\right)^2 (f_{G_0,\nu_i}(z) + f_{\hat{G}_n,\nu_i}(z))\, dz \tag{SM6.19}$$

By Lemmas OA3.1 and SM6.9,

$$\left(\frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}}\right)^2 \lesssim \frac{1}{\nu_i} \log(1/\rho_n) \lesssim_{\mathcal{H}} \log n.$$

Hence,

$$\text{(SM6.17)} \lesssim_{\mathcal{H}} h(f_{G_0,\nu_i}, f_{\hat{G}_n,\nu_i})\sqrt{\log n}.$$

*SM6.2.2 Bounding* (SM6.18). The second term (SM6.18) is

$$\begin{aligned}
\text{(SM6.18)} &= \left| \int \frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z)} \left(\frac{f_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}} - 1\right) f_{\hat{G}_n,\nu_i}(z)\, dz \right| \\
&\le \int \left|\frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z)}\right| \mathbb{1}\left(f_{\hat{G}_n,\nu_i}(z) \le \rho_n/\nu_i\right) f_{\hat{G}_n,\nu_i}(z)\, dz \\
&\le \underbrace{\left(\mathbb{E}_{Z\sim f_{\hat{G}_n,\nu_i}}\left[\left(\mathbf{E}_{\hat{G}_n,\nu_i}\left[\frac{(\tau - Z)}{\nu_i^2} \mid Z\right]\right)^2\right]\right)^{1/2}}_{\le \mathbb{E}_{\tau\sim\hat{G}_n, Z\sim\mathcal{N}(\tau,\nu_i)}[(\tau-Z)^2/\nu_i^4]^{1/2} = \nu_i^{-1}} \cdot \sqrt{\mathrm{P}_{f_{\hat{G}_n,\nu_i}}[f_{\hat{G}_n,\nu_i}(Z) \le \rho_n/\nu_i]}.
\end{aligned}$$

(Cauchy–Schwarz and (SM6.41))

By Jensen's inequality and law of iterated expectations, the first term is bounded by $\frac{1}{\nu_i}$. By Lemma SM6.11, 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, by Lemma OA3.1,

$$\text{(SM6.18)} \lesssim_{\mathcal{H}} M_n^{1/3}\rho_n^{1/3} \lesssim_{\mathcal{H}} M_n^{1/3} n^{-1}.$$

**SM6.3 Bounding $U_{1s}$.**

**Lemma SM6.2.** *Under Assumptions 1 to 4 and SM6.1, if $\|\hat{\eta} - \eta\|_\infty \le \Delta_n$, $\overline{Z}_n \le M_n$, then*

$$|U_{1s}| \lesssim_{\mathcal{H}} \Delta_n \left[\frac{M_n\sqrt{\log n}}{n} \sum_{i=1}^n h(f_{\hat{G}_n,\nu_i}, f_{G_0,\nu_i}) + \frac{M^{4/3}}{n}\right]. \tag{SM6.20}$$

*Proof.* Similar to our computation with $\overline{D}_{m,i}$, we decompose

$$|\overline{D}_{s,i}(\hat{G}_n, \eta_0, \rho_n)| \lesssim_{\sigma_\ell,\sigma_u,s_{0\ell},s_{0u}} \left| \int \frac{Q_i(z, \eta_0, \hat{G}_n)}{f_{\hat{G}_n,\nu_i}(z) \vee (\rho_n/\nu_i)} (f_{G_0,\nu_i}(z) - f_{\hat{G}_n,\nu_i}(z))\, dz \right| \tag{SM6.21}$$

$$+ \left| \int \frac{Q_i(z, \eta_0, \hat{G}_n)}{f_{\hat{G}_n,\nu_i}(z) \vee (\rho_n/\nu_i)} f_{\hat{G}_n,\nu_i}(z)\, dz \right|. \tag{SM6.22}$$

We conclude the proof by plugging in our subsequent calculations. □

*SM6.3.1 Bounding* (SM6.21). The first term (SM6.21) is bounded by

$$[(\text{SM6.21})]^2 \lesssim h^2(f_{G_0,\nu_i}, f_{\hat{G}_n,\nu_i}) \int \left( \frac{Q_i(z,\eta_0,\hat{G}_n)}{f_{\hat{G}_n,\nu_i}(z) \vee (\rho_n/\nu_i)} \right)^2 \left[ f_{G_0,\nu_i}(z) + f_{\hat{G}_n,\nu_i}(z) \right] dz,$$

(recall $Q_i$ from (SM6.40)) similar to the computation in (SM6.19).

By Lemmas OA3.1 and SM6.12,

$$\left( \frac{Q_i(z,\eta_0,\hat{G}_n)}{f_{\hat{G}_n,\nu_i}(z) \vee (\rho_n/\nu_i)} \right)^2 \lesssim_{\sigma_\ell,\sigma_u,s_{0\ell},s_{0u}} (\sqrt{\log n} M_n + \log n)^2 \lesssim_{\mathcal{H}} M_n^2 \log n$$

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

$$(\text{SM6.21}) \lesssim_{\sigma_\ell,\sigma_u,s_{0\ell},s_{0u}} M_n(\sqrt{\log n}) h(f_{G_0,\nu_i}, f_{\hat{G}_n,\nu_i}). \tag{SM6.23}$$

*SM6.3.2 Bounding* (SM6.22). Observe that

$$(\text{SM6.22}) = \left| \int \frac{Q_i(z,\eta_0,\hat{G}_n)}{f_{\hat{G}_n,\nu_i}(z)} \left( \frac{f_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee (\rho_n/\nu_i)} - 1 \right) f_{\hat{G}_n,\nu_i}(z)\, dz \right|$$

Similar to our argument for (SM6.18), by Cauchy–Schwarz,

$$\begin{aligned}(\text{SM6.22}) &\le \left( \mathbb{E}_{f_{\hat{G}_n,\nu_i}(z)} \left[ (\mathbf{E}_{\hat{G}_n,\nu_i}[(Z-\tau)\tau \mid Z])^2 \right] \right)^{1/2} \sqrt{\mathrm{P}_{f_{\hat{G}_n,\nu_i}(z)}(f_{\hat{G}_n,\nu_i}(z) \le \rho_n/\nu_i)} \\ &\lesssim_{\mathcal{H}} M_n \cdot \rho_n^{1/3} M_n^{1/3} \lesssim_{\mathcal{H}} \frac{M_n^{4/3}}{n}.\end{aligned}$$

**SM6.4 Bounding $U_{2m}$.**

**Lemma SM6.3.** *Under Assumptions 1 to 4 and SM6.1,*

$$\mathrm{P}\left[ \|\hat{\eta} - \eta\|_\infty \le \Delta_n, \overline{Z}_n \le M_n, |U_{2m}| \gtrsim_{\mathcal{H}} \sqrt{\log n} \Delta_n \left\{ e^{-C_{\mathcal{H}} M_n^\alpha} + \frac{\log n}{\sqrt{n}} + \frac{1}{(n\Delta_n^{1/p})^{1/2}} \right\} \right] \le \frac{2}{n}$$

*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

$$\begin{aligned}\mathrm{P}\left[ \|\hat{\eta} - \eta\|_\infty \le \Delta_n, \overline{Z}_n \le M_n, |U_{2m}| > t \right] &\le \mathrm{P}\left[ \|\hat{\eta} - \eta\|_\infty \le \Delta_n, \overline{Z}_n \le M_n, \overline{U}_{2m} > t \right] \\ &\le \mathrm{P}[\overline{U}_{2m} > t].\end{aligned}$$

Hence, a large-deviation 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

$$\overline{D}_{m,i,M_n}(\hat{G}_n, \hat{\eta}, \rho_n) = \int D_{m,i}(z, \hat{G}_n, \hat{\eta}, \rho_n) \mathbb{1}(|z| \le M_n) f_{G_0,\nu_i}(z)\, dz.$$

$D_{m,i,M_n}, \overline{D}_{m,i,M_n}$ are the analogues of $D_{m,i}$ and $\overline{D}_{m,i}$ that truncates to $\{|z| \le M_n\}$. Note that, on the event $\overline{Z}_n \le M_n$, $D_{m,i,M_n} = D_{m,i}$.

We decompose 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 the second term can be controlled since $|Z_i| \ge M_n$ is unlikely:

$$
\begin{aligned}
|\overline{D}_{m,i} - \overline{D}_{m,i,M_n}| &\lesssim_{\sigma_\ell, \sigma_u, s_{0\ell}, s_{0u}} \left| \int_{|z|>M_n} \underbrace{\frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee (\rho_n/\nu_i)}}_{\lesssim_{\mathcal{H}} \sqrt{\log n}, \quad \text{Lemmas OA3.1 and SM6.9}} f_{G_0,\nu_i}(z)\, dz \right| \\
&\lesssim_{\mathcal{H}} \sqrt{\log n}\, \mathrm{P}_{G_0,\nu_i}(|Z_i| > M_n)
\end{aligned}
$$

By Lemma SM6.15, $\mathrm{P}_{G_0,\nu_i}(|Z_i| > M_n) \le \exp\left(-C_{\alpha,A_0,\nu_u} M_n^\alpha\right)$. Hence, the second term

$$
\left| \frac{1}{n} \sum_{i=1}^n (\overline{D}_{m,i} - \overline{D}_{m,i,M_n}) \Delta_{mi} \right| \lesssim_{\mathcal{H}} e^{-C_{\mathcal{H}} M_n^\alpha} \sqrt{\log n} \Delta_n.
$$

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

$$
S = \left\{ (m, s) : \|m - m_0\| \le \Delta_n, \|s - s_0\| \le \Delta_n, (m, s) \in C^p_{A_1}([\sigma_\ell, \sigma_u]) \right\}. \tag{SM6.24}
$$

For two distributions $G_1, G_2$, define the following pseudo-metric

$$
d_{m,\infty,M_n}(G_1, G_2) = \max_{i \in [n]} \sup_{|z| \le M_n} |D_{m,i}(z, G_1, \eta_0, \rho_n) - D_{m,i}(z, G_2, \eta_0, \rho_n)| \tag{SM6.25}
$$

Let $G_1, \ldots, G_N$ be an $\omega$-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 the projection of $\hat{G}_n$ to the net. Namely, $G_{j^*}$ is a member of the net 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 in $\overline{D}_{m,i,M_n}$ is bounded uniformly. Hence, by projecting $\hat{G}_n$ to $G_{j^*}$, we obtain

$$
\begin{aligned}
&\left| \frac{1}{n} \sum_{i=1}^n (D_{m,i,M_n}(Z_i, \hat{G}_n, \eta_0, \rho_n) - \overline{D}_{m,i,M_n}(\hat{G}_n, \eta_0, \rho_n))(\hat{m}(\sigma_i) - m_0(\sigma_i)) \right| \\
&\le 2\omega\Delta_n + \max_{j \in [N]} \left| \frac{1}{n} \sum_{i=1}^n (D_{m,i,M_n}(Z_i, G_j, \eta_0, \rho_n) - \overline{D}_{m,i,M_n}(G_j, \eta_0, \rho_n))(\hat{m}(\sigma_i) - m_0(\sigma_i)) \right|
\end{aligned} \tag{SM6.26}
$$

Next, consider the process $V_{n,j}(\eta)$ defined by

$$
\eta \mapsto \frac{1}{n} \sum_{i=1}^n (D_{m,i,M_n}(Z_i, G_j, \eta_0, \rho_n) - \overline{D}_{m,i,M_n}(G_j, \eta_0, \rho_n))(m(\sigma_i) - m_0(\sigma_i))
$$

$$\equiv \frac{1}{n}\sum_{i=1}^{n} v_{i,j}(\eta) \equiv V_{n,j}(\eta)$$

so that, when $\|\hat{\eta} - \eta\|_\infty \le \Delta_n, \overline{Z}_n \le M_n$,

$$\text{(SM6.26)} \lesssim \omega\Delta_n + \max_{j\in[N]}\sup_{\eta\in S}|V_{n,j}(\eta)|.$$

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 $\omega$ shortly.

By the results in Appendix SM6.4.1 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 SM6.4.2, we can pick $\omega$ such that

$$\omega\Delta_n + \max_{j\in[N]}\sup_{\eta\in S}V_{nj}(\eta) \lesssim_{\mathcal{H}} \Delta_n\sqrt{\log n}\left(\frac{\log n}{\sqrt{n}} + \frac{1}{\sqrt{n\Delta_n^{1/p}}}\right) \tag{SM6.27}$$

with probability at least $1 - 2/n$. Putting these observations together, we have that

$$\mathrm{P}\left[\overline{U}_{2m} \gtrsim_{\mathcal{H}} \sqrt{\log n}\Delta_n\left\{e^{-C_{\mathcal{H}}M_n^\alpha} + \frac{\log n}{\sqrt{n}} + \frac{1}{(n\Delta_n^{1/p})^{1/2}}\right\}\right] \le \frac{2}{n}.$$

This concludes the proof. $\square$

*SM6.4.1 Bounding* $\max_{j\in[N]}\sup_{\eta\in S}|V_{n,j}(\eta)|$. Note that $\mathbb{E}v_{ij}(\eta) = 0$. Moreover, by Lemmas OA3.1 and SM6.9,

$$\max\left(D_{m,i,M_n}(Z_i, G_j, \eta_0, \rho_n), \overline{D}_{m,i,M_n}(G_j, \eta_0, \rho_n)\right) \lesssim_{\mathcal{H}} \sqrt{\log(1/\rho_n)} \lesssim_{\mathcal{H}} \sqrt{\log n}$$

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,[53]

$$\|V_{n,j}(\eta_1) - V_{n,j}(\eta_2)\|_{\psi_2} \lesssim_{\mathcal{H}} \frac{\sqrt{\log n}}{\sqrt{n}}\|\eta_1 - \eta_2\|_\infty.$$

Hence $V_{n,j}(\eta)$ is a mean-zero process with subgaussian increments[54] 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 $u > 0$,

$$\mathrm{P}\left[\sup_{\eta\in S}|V_{n,j}(\eta)| \gtrsim_{\mathcal{H}} \frac{\sqrt{\log n}}{\sqrt{n}}\left\{\int_0^{2\Delta_n}\sqrt{\log N(\epsilon, S, \|\cdot\|_\infty)}\,d\epsilon + u\Delta_n\right\}\right] \le 2e^{-u^2}.$$

[53]See Definition 2.5.6 in Vershynin (2018) for a definition of the $\psi_2$-norm (subgaussian norm).
[54]See Definition 8.1.1 in Vershynin (2018).

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 $\|\tilde{V}_{n,j}\|_{\psi_2} \lesssim_{\mathcal{H}} \frac{\Delta_n}{\sqrt{n}}\sqrt{\log n}$.[55] Thus,

$$\text{(SM6.26)} \lesssim_{\mathcal{H}} \Delta_n\left[\omega + \frac{\sqrt{\log n}}{\sqrt{n\Delta_n^{1/p}}}\right] + \max_{j\in[N]}\tilde{V}_{n,j}.$$

Finally, note that by Lemma SM6.14 with the choice $t = \sqrt{\log n}$,

$$\mathrm{P}\left[\max_{j\in[N]}\tilde{V}_{n,j} \gtrsim_{\mathcal{H}} \frac{\Delta_n}{\sqrt{n}}\sqrt{\log n}\left[\sqrt{\log N} + \sqrt{\log n}\right]\right] \leq \frac{2}{n}.$$

*SM6.4.2 Selecting $\omega$.* The rate function that involves $\omega$ 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 SM6.2, shows that

$$\log N \leq \log N\left(\delta\log(1/\delta)\frac{\sqrt{\log(1/\rho_n)}}{\rho_n}, \mathcal{P}(\mathbb{R}), d_{m,\infty,M}\right) \lesssim_{\mathcal{H}} \log(1/\delta)^2 \max\left(1, \frac{M_n}{\sqrt{\log(1/\delta)}}\right)$$

Consider picking $\delta = \rho_n\frac{1}{\sqrt{n}} \leq 1/e$ 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 SM6.1,

$$\omega + \sqrt{\log N}\frac{1}{\sqrt{n}}\sqrt{\log n} \lesssim_{\mathcal{H}} \frac{(\log n)^{3/2}}{\sqrt{n}}.$$

[55] 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 $\psi_2$-norm bound by expression (2.14) in Vershynin (2018).

**SM6.5 Bounding $U_{2s}$.**

**Lemma SM6.4.** *Under Assumptions 1 to 4 and SM6.1,*

$$\mathrm{P}\left[\|\hat{\eta}-\eta\|_\infty \le \Delta_n, \overline{Z}_n \le M_n, |U_{2s}| \gtrsim_{\mathcal{H}} \Delta_n M_n \sqrt{\log n}\left\{e^{-C_{\mathcal{H}} M_n^\alpha} + \frac{\log n}{\sqrt{n}} + \frac{1}{\sqrt{n\Delta_n^{1/p}}}\right\}\right] \le \frac{2}{n}$$

*Proof.* This proof operates much like the proof of Lemma SM6.3. 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 SM6.12 is a uniform bound on the integrand in the second term. Hence, the second term is bounded by

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^n (\overline{D}_{s,i} - \overline{D}_{s,i,M_n})\Delta_{si}\right| \\
&\lesssim_{\mathcal{H}} \Delta_n \sqrt{\log(1/\rho_n)} \frac{1}{n}\sum_{i=1}^n \left(\int_{|Z_i|>M_n} |z| f_{G_0,\nu_i}(z)\,dz + \sqrt{\log(1/\rho_n)}\int_{|Z_i|>M_n} f_{G_0,\nu_i}(z)\,dz\right) \\
&\lesssim_{\mathcal{H}} \Delta_n \sqrt{\log n}\left\{e^{-\frac{C_{\mathcal{H}}}{2}M_n^\alpha} \max_{i\in[n]} \mu_2(f_{G_0,\nu_i}) + \sqrt{\log n}\, e^{-C_{\mathcal{H}} M_n^\alpha}\right\} \\
&\qquad\qquad\qquad \text{(Cauchy–Schwarz for the first term and apply Lemmas OA3.1 and SM6.15)} \\
&\lesssim_{\mathcal{H}} \Delta_n (\log n) e^{-C_{\mathcal{H}} M_n^\alpha}.
\end{aligned}$$

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 $S = \left\{(m,s) : \|m-m_0\| \le \Delta_n, \|s-s_0\| \le \Delta_n, (m,s) \in C^p_{A_1}([\sigma_\ell, \sigma_u])\right\}$. For two distributions $G_1, G_2$, define the following pseudo-metric

$$d_{s,\infty,M_n}(G_1,G_2) = \max_{i\in[n]} \sup_{|z|\le M_n} |D_{s,i}(z,G_1,\eta_0,\rho_n) - D_{s,i}(z,G_2,\eta_0,\rho_n)| \tag{SM6.28}$$

Let $G_1,\ldots,G_N$ be an $\omega$-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

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^n (D_{s,i,M_n}(Z_i,\hat{G}_n,\eta_0,\rho_n) - \overline{D}_{s,i,M_n}(\hat{G}_n,\eta_0,\rho_n))(\hat{s}(\sigma_i)-s_0(\sigma_i))\right| \\
&\le 2\omega\Delta_n + \max_{j\in[N]}\left|\frac{1}{n}\sum_{i=1}^n (D_{s,i,M_n}(Z_i,G_j,\eta_0,\rho_n) - \overline{D}_{s,i,M_n}(G_j,\eta_0,\rho_n))(\hat{s}(\sigma_i)-s_0(\sigma_i))\right| \qquad \text{(SM6.29)}
\end{aligned}$$

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 (SM6.29) $\lesssim \omega\Delta_n + \max_{j\in[N]}\sup_{\eta\in S}|V_{n,j}(\eta)|$. 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 for $V_{n,j}$, as well as an appropriate choice of $\omega$, for $\overline{U}_{2s}$.

By Lemma SM6.12, 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 SM6.4.1, with a different constant for the subgaussianity. Hence, by the same argument as in Appendix SM6.4.1, 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 $\omega$. The relevant term for selecting $\omega$ 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 SM6.4.2 with Proposition SM6.2 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, plugging in these choices, we can compute that with probability at least $1-2/n$, under Assumption SM6.1,

$$\overline{U}_{2s} \lesssim_{\mathcal{H}} \Delta_n M_n\sqrt{\log n}\left\{e^{-C_{\mathcal{H}}M_n^\alpha} + \frac{\log n}{\sqrt{n}} + \frac{1}{\sqrt{n\Delta_n^{1/p}}}\right\}.$$

This concludes the proof. □

### SM6.6 Bounding $U_{3m}, U_{3s}$.

**Lemma SM6.5.** *Under Assumptions 2 to 4 and SM6.1,*

$$\mathrm{P}\left[\|\hat{\eta}-\eta\|_\infty \le \Delta_n, \overline{Z}_n \le M_n, |U_{3m}| \gtrsim_{\mathcal{H}} \Delta_n\left\{e^{-C_{\mathcal{H}}M_n^\alpha} + \frac{M_n}{\sqrt{n}}\left(\Delta_n^{-1/(2p)} + \log n\right)\right\}\right] \le \frac{2}{n}$$

$$\mathrm{P}\left[\|\hat{\eta}-\eta\|_\infty \le \Delta_n, \overline{Z}_n \le M_n, |U_{3s}| \gtrsim_{\mathcal{H}} \Delta_n\left\{e^{-C_{\mathcal{H}}M_n^\alpha} + \frac{M_n^2}{\sqrt{n}}\left(\Delta_n^{-1/(2p)} + \log n\right)\right\}\right] \le \frac{2}{n}.$$

*Proof.* The proof structure follows that of Lemmas SM6.3 and SM6.4.

Recall that

$$U_{3m} = \frac{1}{n}\sum_{i=1}^{n} D_{m,i}(Z_i, G_0, \eta_0, 0)(\hat{m}_i - m_0).$$

$$= \frac{1}{n}\sum_{i=1}^{n}(D_{m,i,M_n} - \overline{D}_{m,i,M_n})(\hat{m}_i - m_0) + \overline{D}_{m,i,M_n}(\hat{m}_i - m_0)$$

Note that

$$\begin{aligned}
|\overline{D}_{m,i,M_n}| &= \left|\int_{|z|\le M_n} \frac{f'_{G_0,\nu_i}(z)}{f_{G_0,\nu_i}(z)} f_{G_0,\nu_i}(z)\,dz\right| \\
&= \left|\int \mathbb{1}(|z| > M_n)\cdot \frac{f'_{G_0,\nu_i}(z)}{f_{G_0,\nu_i}(z)} f_{G_0,\nu_i}(z)\,dz\right| \\
&\lesssim_{\sigma_\ell,\sigma_u,s_{0\ell},s_{0u}} \mathrm{P}(|z| > M_n)^{1/2} \\
&\qquad \text{(Cauchy–Schwarz, Jensen, and law of iterated expectations via (SM6.41))} \\
&\lesssim_{\mathcal{H}} e^{-C_{\mathcal{H}} M_n^\alpha}. \qquad \text{(SM6.30)}
\end{aligned}$$

Recall $S$ in (SM6.24). 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 SM6.17,

$$\|v_{ni}(\eta_1) - v_{ni}(\eta_2)\|_{\psi_2} \lesssim \|D_{m,i,M_n}(\eta_{1i} - \eta_{2i})\|_{\psi_2} \lesssim_{\mathcal{H}} M_n\|\eta_1 - \eta_2\|_\infty.$$

Since $v_{ni}(\eta_1) - v_{ni}(\eta_2)$ is mean zero, we have that

$$\|V_n(\eta_1) - V_n(\eta_2)\|_{\psi_2} \lesssim_{\mathcal{H}} \frac{M_n}{\sqrt{n}}\|\eta_1 - \eta_2\|_\infty \qquad \text{(SM6.31)}$$

Hence, by the same Dudley's chaining calculation in Appendix SM6.4.1, 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 (SM6.30) and (SM6.31). For the tail bound (analogue of (SM6.30)), we have the same bound

$$|\overline{D}_{s,i,M_n}| \lesssim \mathrm{P}(|z| > M_n)^{1/2}\left(\mathbb{E}_{f_{G_0,\nu_i}(z)}\left[(\mathbf{E}_{G_0,\nu_i}[(Z-\tau)\tau \mid Z])^2\right]\right)^{1/2} \lesssim_{\mathcal{H}} e^{-C_{\mathcal{H}} M_n^\alpha}.$$

For the analogue of (SM6.31), since Lemma SM6.17 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 least $1 - 2/n$

$$\overline{U}_{3s} \lesssim_{\mathcal{H}} \Delta_n \left\{ e^{-C_{\mathcal{H}} M_n^\alpha} + \frac{M_n^2}{\sqrt{n}}(\Delta_n^{-1/(2p)} + \log n)\right\}. \qquad \square$$

**SM6.7 Bounding $R_1, R_2$.**

**Lemma SM6.6.** *Recall $R_{1i}$ from (SM6.3). Then, under Assumptions 1 to 4 and SM6.1, 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 $\Delta_n^2$. By Lemma SM6.13, 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. □

**Lemma SM6.7.** *Under Assumptions 2 to 4 and SM6.1, 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 SM6.17, under $G_0$, these posterior moments are bounded by above by corresponding moments of $\hat{Z}_i(\tilde{\eta}_i)$. Hence,

$$\mathbb{1}(A_n)\|H_i\|_\infty \lesssim_{\mathcal{H}} \mathbb{1}(A_n)\left(\hat{Z}_i(\tilde{\eta}_i) \vee 1\right)^4 \lesssim_{\mathcal{H}} (Z_i \vee 1)^4. \tag{SM6.32}$$

Hence, $\mathbb{1}(A_n)|R_2| \lesssim_{\mathcal{H}} \Delta_n^2 \frac{1}{n}\sum_{i=1}^n (Z_i \vee 1)^4$. Note that Chebyshev's inequality implies that there exists some choice $C_{\mathcal{H}}$ such that

$$\mathrm{P}\left[\frac{1}{n}\sum_{i=1}^n (Z_i \vee 1)^4 \ge C_{\mathcal{H}}\right] \le \frac{1}{n},$$

since $\mathrm{Var}(\frac{1}{n}\sum_{i=1}^n (Z_i \vee 1)^4) \lesssim_{\mathcal{H}} \frac{1}{n}$. Hence,

$$\mathrm{P}\left(\|\hat{\eta}-\eta\|_\infty \le \Delta_n, \overline{Z}_n \le M_n, |R_2| \gtrsim_{\mathcal{H}} \Delta_n^2\right) \le \frac{1}{n}.$$

□

**SM6.8 Derivative computations.** In the remainder of the proof, it is sometimes useful to relate the derivatives of $\psi_i$ to $\mathbf{E}_{G,\eta}$. We compute the following derivatives. Since they are all evaluated at $G, \eta$, we let $\hat{\nu} = \hat{\nu}_i(\eta)$ and $\hat{z} = \hat{Z}_i(\eta)$ as a shorthand.

$$\left.\frac{\partial \psi_i}{\partial m_i}\right|_{\eta,G} = -\frac{1}{s_i}\frac{f'_{G,\hat{\nu}}(\hat{z})}{f_{G,\hat{\nu}}(\hat{z})} \tag{SM6.33}$$

$$= \frac{s_i}{\sigma_i^2}\mathbf{E}_{G,\hat{\nu}}[Z-\tau \mid \hat{z}] \tag{SM6.34}$$

$$\left.\frac{\partial \psi_i}{\partial s_i}\right|_{\eta,G} = \frac{1}{\sigma_i \hat{\nu}_i(\eta) f_{G,\hat{\nu}(\eta)}(\hat{Z}_i(\eta))} \underbrace{\int (\hat{Z}_i(\eta)-\tau)\tau\varphi\left(\frac{\hat{Z}_i(\eta)-\tau}{\hat{\nu}_i(\eta)}\right)\frac{1}{\hat{\nu}_i(\eta)}\,G(d\tau)}_{Q_i(Z_i,\eta,G)} \tag{SM6.35}$$

$$= \frac{1}{\sigma_i \hat{\nu}}\mathbf{E}_{G,\hat{\nu}}[(Z-\tau)\tau \mid \hat{z}] \tag{SM6.36}$$

$$\left.\frac{\partial^2\psi_i}{\partial m_i^2}\right|_{\eta,G} = \frac{1}{s_i^2}\left[\frac{f''_{G,\hat{\nu}}(\hat{z})}{f_{G,\hat{\nu}}(\hat{z})} - \left(\frac{f'_{G,\hat{\nu}}(\hat{z})}{f_{G,\hat{\nu}}(\hat{z})}\right)^2\right] \tag{SM6.37}$$

$$= \frac{1}{s_i^2}\left[\frac{1}{\hat{\nu}^4}\mathbf{E}_{G,\hat{\nu}}[(\tau-Z)^2 \mid \hat{z}] - \frac{1}{\hat{\nu}^2} - \frac{1}{\hat{\nu}^4}\left(\mathbf{E}_{G,\hat{\nu}}[(\tau-Z)\mid\hat{z}]\right)^2\right] \tag{SM6.38}$$

$$\left.\frac{\partial^2\psi_i}{\partial m_i \partial s_i}\right|_{\eta,G} = \left(\frac{1}{\sigma_i^2}\mathbf{E}_{G,\hat{\nu}}[(Z-\tau)\tau \mid \hat{z}] - \frac{1}{s_i^2}\right)\frac{1}{\hat{\nu}^2}\mathbf{E}_{G,\hat{\nu}}[(\tau-Z)\mid\hat{z}] + \frac{\mathbf{E}_{G,\hat{\nu}}[(\tau-Z)^2\tau \mid \hat{z}]}{\hat{\nu}\sigma_i s_i} \tag{SM6.39}$$

$$\left.\frac{\partial^2\psi_i}{\partial s_i^2}\right|_{\eta,G} = \frac{1}{\sigma_i^2}\left\{\mathbf{E}_{G,\hat{\nu}}\left[\left(\frac{s_i^2}{\sigma_i}(Z-\tau)^2 - 1\right)\tau^2 \mid \hat{z}\right] - \frac{1}{\hat{\nu}^2}\left(\mathbf{E}_{G,\hat{\nu}}[(Z-\tau)\tau\mid\hat{z}]\right)^2\right\} \tag{SM6.40}$$

It is also useful to note that

$$\frac{f'_{G,\nu}(z)}{f_{G,\nu}(z)} = \frac{1}{\nu^2}\mathbf{E}_{G,\nu}[(\tau-Z)\mid z] \tag{SM6.41}$$

$$\frac{f''_{G,\nu}(z)}{f_{G,\nu}(z)} = \frac{1}{\nu^4}\mathbf{E}_{G,\nu}[(\tau-Z)^2\mid z] - \frac{1}{\nu^2} \tag{SM6.42}$$

**SM6.9 Complexity 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 SM6.8.** *For all integer $m \ge 0$:*

$$\sup_{t\in\mathbb{R}} |t^m \varphi(t)| = m^{m/2}\varphi(\sqrt{m}).$$

*As a corollary, there exists absolute $C_m > 0$ such that $t \mapsto t^m\varphi(t)$ is $C_m$-Lipschitz.*

**Proposition SM6.1.** *Fix some $q \in \mathbb{N}\cup\{0\}$ and $M > 1$. Consider the pseudometric*

$$d^{(q)}_{\infty,M}(G_1,G_2) = \max_{i\in[n]} \underbrace{\max_{0\le v\le q}\sup_{|x|\le M}\left|\int \frac{(u-x)^v}{\nu_i^v}\varphi\left(\frac{x-u}{\nu_i}\right)(G_1-G_2)(du)\right|}_{d_{q,i,m}(G_1,G_2)}.$$

*Let $\nu_\ell, \nu_u$ be the lower and upper bounds of $\nu_i$. Then, for all $0 < \delta < \exp(-q/2)\wedge e^{-1}$,*

$$\log N(\delta \log^{q/2}(1/\delta), \mathcal{P}(\mathbb{R}), d^{(q)}_{\infty,M}) \lesssim_{q,\nu_u,\nu_\ell} \log^2(1/\delta)\max\left(\frac{M}{\sqrt{\log(1/\delta)}}, 1\right).$$

*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.

*SM6.9.1 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 such that the moments match

$$\int_{I_j} u^k dG(u) = \int_{I_j} u^k dG_m(u) \text{ for all } k = 0, \dots, 2k^* + q \text{ and } j = 1, \dots, j^*.$$

for some $k^*$ to be chosen later.

Then, for some $x \in I_j \cap [-M, M]$, we have

$$d_{q,i,M}(G, G_m) \leq \max_{0 \leq v \leq q} \left[ \left| \int_{(I_{j-1} \cup I_j \cup I_{j+1})^{\mathrm{C}}} \left( \frac{u - x}{\nu_i} \right)^v \varphi\left( \frac{x - u}{\nu_i} \right) (G(du) - G_m(du)) \right| \right. \tag{SM6.43}$$

$$\left. + \left| \int_{I_{j-1} \cup I_j \cup I_{j+1}} \left( \frac{u - x}{\nu_i} \right)^v \varphi\left( \frac{x - u}{\nu_i} \right) (G(du) - G_m(du)) \right| \right] \tag{SM6.44}$$

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 (SM6.43) is bounded by $\varphi_+(\omega)^v \omega$, as $\frac{|u-x|}{\nu_i} \geq a\nu_\ell/\nu_u = \varphi_+(\omega)$:

$$\text{(SM6.43)} \leq 2 \max_{0 \leq v \leq q} \varphi_+(\omega)^v \omega = 2\varphi_+(\omega)^q \omega.$$

For (SM6.44), 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 (SM6.44) 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 (SM6.44) is equal to

$$\text{(SM6.44)} = \max_{0 \leq v \leq q} \left| \int_{(I_{j-1} \cup I_j \cup I_{j+1}^{\mathrm{C}})} \left( \frac{u - x}{\nu_i} \right)^v R\left( \frac{x - u}{\nu_i} \right) (G(du) - G_m(du)) \right|. \tag{SM6.45}$$

We know that, since $\varphi(t)$ has alternating-signed Taylor expansion, its remainder is bounded by the first term of truncation

$$|R(t)| \leq \frac{(t^2/2)^{k^*+1}}{\sqrt{2\pi}(k^* + 1)!}.$$

We can bound $|\frac{u-x}{\nu_i}| \leq 2a\nu_\ell/\nu_i \leq 2a$. Hence the integral (SM6.45) is upper bounded by

$$\text{(SM6.44)} \leq 2 \cdot (2a)^q \cdot \frac{\left((2a)^2/2\right)^{k^*+1}}{\sqrt{2\pi}(k^* + 1)!} \qquad ((2a)^v \leq (2a)^q)$$

$$
\begin{aligned}
&\leq \frac{2(2a)^q}{(2\pi)\sqrt{k^*+1}}\left(\frac{2a^2}{k^*+1}e\right)^{k^*+1} && \text{(Stirling's formula } (k^*+1)! \geq \sqrt{2\pi(k^*+1)}\left(\tfrac{k^*+1}{e}\right)^{k^*+1}\text{)} \\
&\leq \frac{(2a)^q}{\pi\sqrt{k^*+1}}\left(\frac{e}{3}\right)^{k^*+1} && \text{(Choosing } k^*+1 \geq 6a^2 \geq 6\text{)} \\
&\leq \frac{(2a)^q}{\pi\sqrt{k^*+1}}\exp\left(-\frac{1}{2}\frac{k^*+1}{6}\right) && ((e/3)^6 \leq e^{-1/2}) \\
&\leq \frac{(2a)^q}{\sqrt{k^*+1}\sqrt{\pi/2}}\underbrace{\varphi(a\nu_\ell/\nu_u)}_{\varphi(\varphi_+(\omega))} && (k^*+1 \geq 6a^2 \geq 6(a\nu_\ell/\nu_u)^2) \\
&\leq \frac{(2a)^q}{\sqrt{k^*+1}\sqrt{\pi/2}}\omega \\
&\leq \frac{2^q}{\sqrt{3\pi}}\left(\frac{\nu_u}{\nu_\ell}\right)^{q-1}\varphi_+^{q-1}(\omega)\omega && (k^*+1 \geq 6a^2)
\end{aligned}
$$

This bounds (SM6.43) + (SM6.44) uniformly over $|x| \leq M$. Therefore,

$$
d_{q,i,M}(G, G_m) \leq \left(2 + \frac{2^q}{\sqrt{3\pi}}(\nu_u/\nu_\ell)^{q-1}\right) \cdot \varphi_+^q(\omega)\omega \lesssim_{q,\nu_u,\nu_\ell} \log^{q/2}(1/\omega)\omega.
$$

*SM6.9.2 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 $\tilde{\xi} = \omega \operatorname{sgn}(\xi)\lfloor|\xi|/\omega\rfloor$. 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 $\omega$.

Since

$$
\int g(x,u)G_{m,\omega}(du) = \int g(x, \omega \operatorname{sgn}(u)\lfloor|u|/\omega\rfloor)\, G_m(du)
$$

we have that

$$
\left|\int g(x,u)G_{m,\omega}(du) - \int g(x,u)G_m(du)\right| \leq \int |g(x, \omega \operatorname{sgn}(u)\lfloor|u|/\omega\rfloor) - g(x,u)|\, G_m(du)
$$

In the case of $g(x,u) = ((x-u)/\nu_i)^v\, \varphi((x-u)/\nu_i)$, this function is Lipschitz by Lemma SM6.8, we thus have that,

$$
d_{q,i,M}(G_m, G_{m,\omega}) \leq \int C_q \frac{\omega}{\nu_i} G_m(du) \lesssim_{\nu_\ell,q} \omega.
$$

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 $d^{(q)}_{\infty,M}(\cdot,\cdot)$.

*SM6.9.3 Covering the set of $G_{m,\omega}$.* Let $\Delta^{m-1}$ 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 $\Delta^{m-1}$. Thus, there are at most $N(\omega, \Delta^{m-1}, \|\cdot\|_1)$ such distributions that form an $\omega$-net in $\|\cdot\|_1$. Now, consider a distribution $G'_{m,\omega}$ where

$$
\|G'_{m,\omega} - G_{m,\omega}\|_1 \leq \omega.
$$

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 SM6.8.

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

$$\min_{H\in\mathcal{G}} \|G-H\|_{q,\infty,M} \le \underbrace{C_{q,\nu_u,\nu_\ell}\log(1/\omega)^{q/2}\omega}_{\omega^*}.$$

*SM6.9.4 Putting things together.* In other words,

$$\begin{aligned} N(\omega^*, \mathcal{P}(\mathbb{R}), \|\cdot\|_{q,\infty,M}) &\le \binom{1+2\lfloor (M+a\nu_\ell)/\omega\rfloor}{m} N(\omega, \Delta^{m-1}, \|\cdot\|_1) \\ &\le \left(\frac{(\omega+2)(\omega+2(M+a\nu_\ell))e}{m}\right)^m \omega^{-2m}(2\pi m)^{-1/2} \quad \text{((6.24) in Jiang (2020))} \end{aligned}$$

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 $\delta$. 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 $\delta$'s into the constant. $\square$

As a consequence, we can control the covering number in terms of $d_{k,\infty,M}$ for $k \in \{m, s\}$

**Proposition SM6.2.** *Consider $d^{(q)}_{\infty,M}$ in Proposition SM6.1, $d_{s,\infty,M}$ in (SM6.28), and $d_{m,\infty,M}$ in (SM6.25) 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 $D_{s,i}(z_i, G, \eta_0, \rho_n) = \frac{s_i}{\sigma_i^2}\frac{Q_i(Z_i,\eta_0,G)}{f_{i,G} \vee \frac{\rho_n}{\nu_i}}$ from (SM6.5). 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 SM6.12.

Note that

$$f_{i,G_1}\mathbf{E}_{G_1,\nu_i}[(Z-\tau)\tau \mid z] = f_{i,G_1}\mathbf{E}_{G_1,\nu_i}[(Z-\tau)^2 \mid z] - z f_{i,G_1}\mathbf{E}_{G_1,\nu_i}[(Z-\tau) \mid z].$$

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,

$$|D_{s,i}(z, G_1, \eta_0, \rho_n) - D_{s,i}(z, G_2, \eta_0, \rho_n)| \lesssim_{\mathcal{H}} \left\{\frac{M}{\rho_n} + \rho_n^{-1}\left(M\sqrt{\log(1/\rho_n)} + \log(1/\rho_n)\right)\right\}\delta$$

$$\lesssim_{\mathcal{H}} \frac{M\sqrt{\log(1/\rho_n)}+\log(1/\rho_n)}{\rho_n}\delta.$$

Similarly, recall (SM6.4), and we have

$$|D_{m,i}(z,G_1,\eta_0)-D_{m,i}(z,G_2,\eta_0)| \lesssim_{\mathcal{H}} \frac{1}{\rho_n}\delta + \frac{1}{\rho_n}\sqrt{\log(1/\rho_n)}\delta \lesssim \frac{1}{\rho_n}\sqrt{\log(1/\rho_n)}$$

by a similar calculation, involving Lemma SM6.9.

Thus, for the "corollary" part, note that, letting $C_{\mathcal{H}}$ be the constant in the bound, taken to be at least 1:

$$\begin{aligned} N\left(\frac{\delta\log(1/\delta)}{\rho_n}\sqrt{\log(1/\rho_n)}, \mathcal{P}(\mathbb{R}), d_{m,\infty,M}\right) &\le N\left(\frac{\delta}{C_{\mathcal{H}}}\log(1/(\delta/(C_{\mathcal{H}}))), \mathcal{P}(\mathbb{R}), d^{(2)}_{\infty,M}\right) \\ &\lesssim_{\mathcal{H}} \log(1/\delta)^2 \max\left(1, \frac{M}{\sqrt{\log(1/\delta)}}\right). \end{aligned}$$

for all $0<\delta<1/e$. Similarly for the covering number in $d_{s,\infty,M}$. □

### SM6.10 Auxiliary lemmas.

**Lemma OA3.1.** *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 SM6.1,*[56]

*(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 $\rho_n$ 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 SM6.10(3). Hence by Assumption SM6.1, $|\hat{Z}_i|\vee 1 \lesssim_{\mathcal{H}} M_n$.

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 $\kappa_n$ 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

$$\frac{\hat{Z}_i-\tau}{\hat{\nu}_i} = \frac{Z_i-\tau}{\nu_i} + \frac{m_{0i}-\hat{m}_i}{\sigma_i} + \frac{1}{\sigma_i}(s_i-s_{0i})\tau \equiv \frac{Z_i-\tau}{\nu_i} + \xi(\tau) \tag{SM6.46}$$

where $|\xi(\tau)| \lesssim_{\mathcal{H}} \Delta_n M_n$ over the support of $\tau$ under $\hat{G}_n$, under our assumptions.

Then, for all $Z_i$, since $|Z_i| \le M_n$ by assumption,

$$\begin{aligned} \varphi\left(\frac{\hat{Z}_i-\tau}{\hat{\nu}_i}\right) &= \varphi\left(\frac{Z_i-\tau}{\nu_i}\right)\exp\left(-\frac{1}{2}\xi^2(\tau) - \xi(\tau)\frac{Z_i-\tau}{\nu_i}\right) \\ &\le \varphi\left(\frac{Z_i-\tau}{\nu_i}\right)\exp\left(C_{\mathcal{H}}\Delta_n M_n\left|\frac{Z_i-\tau}{\nu_i}\right|\right) \end{aligned}$$

($C_{\mathcal{H}}$ is defined by optimizing over $|\xi(\tau)| \lesssim_{\mathcal{H}} \Delta_n M_n$)

[56]This assumption is satisfied with our choices in (OA3.4).

$$\leq \varphi\left(\frac{Z_i-\tau}{\nu_i}\right)\exp\left(C_{\mathcal{H}}\Delta_n M_n^2\right). \qquad \left(\left|\frac{Z_i-\tau}{\nu_i}\right| \lesssim_{\mathcal{H}} M_n\right)$$

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 $\nu_i$ 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 SM6.1(1) to obtain that $\Delta_n M_n^2 \lesssim_{\mathcal{H}} 1$. $\square$

**Lemma SM6.9** (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 $\varphi_+$ from (OA3.3).*

*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

$$\left|\left(\frac{\nu^2 f''_{H,\nu}(x)}{f_{H,\nu}(z)}+1\right)\left(\frac{\nu f_{H,\nu}(x)}{(\nu f_{G,\nu}(x)) \vee \rho}\right)\right| = \frac{\nu^2 f''_{H,\nu}(x)}{f_{H,\nu}(z)}+1 \leq \varphi_+^2(\nu f_{H,\nu}) \leq \log\frac{1}{2\pi\rho^2}. \quad \square$$

**Lemma SM6.10.** *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 (A.1).*

*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 $\hat{Z}_i = \frac{s_{0i}}{\hat{s}_i} Z_i + [m_{0i} - \hat{m}_i]$

$\square$

**Lemma SM6.11** (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 $\mathbb{E}_f[Z] = 0$.

$$\begin{aligned}
\int_{-\infty}^{\infty} \mathbb{1}(f(z) \le t) f(z)\, dz &\le \int_{-\infty}^{\infty} \mathbb{1}(f(z) \le t, |z| < M) f(z)\, dz + \int_{-\infty}^{\infty} \mathbb{1}(f(z) \le t, |z| > M) f(z)\, dz \\
&\le \int_{-M}^{M} t\, dz + \mathrm{P}(|Z| > M) \\
&\le 2Mt + \frac{\sigma^2(f)}{M^2}. && \text{(Chebyshev's inequality)}
\end{aligned}$$

$\square$

**Lemma SM6.12.** *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 $\rho_n \in (0, e^{-1}/\sqrt{2\pi})$,*

$$\left|\frac{Q_i(z, \eta_0, G)}{f_{G,\nu_i}(z) \vee (\rho_n/\nu_i)}\right| \le \varphi_+(\rho_n)\left(\nu_i|z| + \nu_i\varphi_+(\rho_n)\right). \tag{SM6.47}$$

*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 SM6.9,

$$\frac{f_{G,\nu_i}(z)}{f_{G,\nu_i}(z) \vee (\rho_n/\nu_i)}\mathbf{E}_{G,\nu_i}[(z-\tau) \mid z] \le \nu_i\varphi_+(\rho_n)$$

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,

$$\left|\frac{Q_i(z, \eta_0, G)}{f_{G,\nu_i}(z) \vee (\rho_n/\nu_i)}\right| \le \varphi_+(\rho_n)\nu_i\left(|z| + \varphi_+(\rho_n)\right).$$

$\square$

**Lemma SM6.13.** *Under the assumptions in Lemma OA3.1 and Assumption 4, suppose $\tilde{\eta}_i$ lies on the line segment between $\eta_0$ and $\hat{\eta}_i$ and define $\tilde{\nu}_i, \tilde{m}_i, \tilde{s}_i, \tilde{Z}_i$ accordingly. Then, the second derivatives (SM6.37), (SM6.39), (SM6.40), evaluated at $\tilde{\eta}_i, \hat{G}_n, \tilde{Z}_i$, satisfy*

$$|\text{(SM6.37)}| \lesssim_{\mathcal{H}} \log n$$

$$|(\text{SM6.39})| \lesssim_{\mathcal{H}} M_n \log n$$
$$|(\text{SM6.40})| \lesssim_{\mathcal{H}} M_n^2 \log n.$$

*Proof.* First, we show that

$$|\log(f_{\hat{G}_n,\tilde{\nu}_i}(\tilde{Z}_i)\tilde{\nu}_i)| \lesssim_{\mathcal{H}} \log n. \tag{SM6.48}$$

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 shows that $|\tilde{Z}_i| \lesssim_{\mathcal{H}} M_n$ under the assumptions since $\hat{s} > s_\ell$. Having verified that $|\tilde{Z}_i| \lesssim_{\mathcal{H}} M_n$, note that by the same argument in (SM6.46) in Lemma OA3.1, we have that

$$\varphi\left(\frac{\hat{Z}_i-\tau}{\hat{\nu}_i}\right) \le \varphi\left(\frac{\tilde{Z}_i-\tau}{\tilde{\nu}_i}\right)e^{C_{\mathcal{H}}\Delta_n M_n^2}. \implies \tilde{\nu}_i f_{\hat{G}_{(i)},\tilde{\nu}_i}(\tilde{Z}_i) \ge \frac{1}{n^3}e^{-C_{\mathcal{H}}\Delta_n M_n^2}.$$

This shows (SM6.48).

Now, observe that

$$\mathbf{E}_{\hat{G}_n,\tilde{\nu}}[(\tau-Z)^2 \mid \tilde{Z}_i] \lesssim_{\mathcal{H}} \log\left(\frac{1}{\tilde{\nu}_i f_{\hat{G}_{(i)},\tilde{\nu}_i}(\tilde{Z}_i)}\right) \lesssim_{\mathcal{H}} \log n$$
$$\mathbf{E}_{\hat{G}_n,\tilde{\nu}}[|\tau-Z| \mid \tilde{Z}_i] \lesssim_{\mathcal{H}} \sqrt{\log\left(\frac{1}{\tilde{\nu}_i f_{\hat{G}_{(i)},\tilde{\nu}_i}(\tilde{Z}_i)}\right)} \lesssim_{\mathcal{H}} \sqrt{\log n}$$

by Lemma SM6.9, 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 SM6.12, and plugging in $\rho = \tilde{\nu}_i f_{\hat{G}_{(i)},\tilde{\nu}_i}(\tilde{Z}_i) \wedge \frac{1}{\sqrt{2\pi e}}$,

$$\left|\mathbf{E}_{\hat{G}_n,\tilde{\nu}}[(\tau-Z)Z \mid \tilde{Z}_i]\right| \lesssim_{\mathcal{H}} \sqrt{\log n}|\tilde{Z}_i| + \log n \lesssim_{\mathcal{H}} M_n\sqrt{\log n}.$$

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 (SM6.37), (SM6.39), (SM6.40) proves the claim. ☐

**Lemma SM6.14.** *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 $t \ge 0$,*

$$\mathrm{P}\left[\max_i |X_i| \ge CK\sqrt{\log J} + CKt\right] \le 2e^{-t^2}.$$

*Proof.* By (2.14) in Vershynin (2018), $\mathrm{P}(|X_i| > t) \le 2e^{-ct^2/\|X_i\|_{\psi_2}^2} \le 2e^{-ct^2/K}$ for some universal $c$. By a union bound,

$$\mathrm{P}\left[\max_i |X_i| \ge Ku\right] \le 2\exp\left(-cu^2 + \log J\right)$$

Choose $u = \frac{1}{\sqrt{c}}(\sqrt{\log J} + t)$ so that $cu^2 = \log J + t^2 + 2t\sqrt{\log J} \ge \log J + t^2$. Hence

$$2\exp\left(-cu^2 + \log J\right) \le 2e^{-t^2}.$$

Implicitly, $C = 1/\sqrt{c}$. $\square$

**Lemma SM6.15.** *Suppose $Z$ has simultaneous moment control $\mathbb{E}[|Z|^p]^{1/p} \le Ap^{1/\alpha}$. Then*

$$\mathrm{P}(|Z| > M) \le \exp\left(-C_{A,\alpha} M^\alpha\right).$$

*As a corollary, suppose $Z \sim f_{G_0,\nu_i}(\cdot)$ and $G_0$ obeys Assumption 2, then*

$$\mathrm{P}(|Z| > M) \le \exp\left(-C_{A_0,\alpha,\nu_u} M^\alpha\right).$$

*Proof.* Observe that

$$\mathrm{P}(|Z| > M) = \mathrm{P}(|Z|^p > M^p) \le \left\{\frac{Ap^{1/\alpha}}{M}\right\}^p. \qquad \text{(Markov's inequality)}$$

Choose $p = (M/(eA))^\alpha$ such that

$$\left\{\frac{Ap^{1/\alpha}}{M}\right\}^p = \exp\left(-p\right) = \exp\left(-\left(\frac{1}{eA}\right)^\alpha M^\alpha\right). \qquad \square$$

**Lemma SM6.16.** *Let $E$ be some event and assume that*

$$\mathrm{P}(E, A > a) \le p_1 \quad \mathrm{P}(E, B > b) \le p_2$$

*Then $\mathrm{P}(E, A + B > a + b) \le p_1 + p_2$*

*Proof.* Note that $A + B > a + b$ implies that one of $A > a$ and $B > b$ occurs. Hence

$$\mathrm{P}(E, A + B > a + b) \le \mathrm{P}(\{E, A > a\} \cup \{E, B > b\}) \le p_1 + p_2$$

by union bound. $\square$

**Lemma SM6.17.** *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 $|z|$:*

$$\mathbb{E}[|\tau|^p \mid Z = z] \lesssim_{p,\alpha,A_0} (|z| \vee 1)^p.$$

*Proof.* Let $M = |z| \vee 2$. 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 we can decompose based on $|\tau| > 3M$:

$$\begin{aligned}\int |\tau|^p \varphi\left(\frac{z-\tau}{\nu}\right) \frac{1}{\nu} G_0(d\tau) &\le (3M)^p f_{G_0,\nu}(z) + \int \mathbb{1}(|\tau| > 3M)|\tau|^p \varphi\left(\frac{z-\tau}{\nu}\right) \frac{1}{\nu} G_0(d\tau) \\ &\le (3M)^p f_{G_0,\nu}(z) + \int_{|\tau|>3M} |\tau|^p G_0(d\tau) \cdot \frac{1}{\nu}\varphi\left(|2M|/\nu\right) \\ &\qquad (|z-\tau| \ge 2M \text{ when } |\tau| > 3M)\end{aligned}$$

Also note that

$$f_{G_0,\nu}(z) = \int \varphi\left(\frac{z-\tau}{\nu}\right) \frac{1}{\nu} G_0(d\tau) \ge \frac{1}{\nu}\varphi\left(|2M|/\nu\right) G_0([-M, M]) \quad (|z-\tau| \le 2M \text{ if } \tau \in [-M, M])$$

Hence,

$$\mathbb{E}[|\tau|^p \mid Z = z] \le (3M)^p + \frac{\int |\tau|^p G_0(d\tau)}{G_0([-M, M])}$$

Since $G_0$ is mean zero and variance 1, by Chebyshev's inequality, $G_0([-M, M]) \ge G_0([-2, 2]) \ge 3/4$. Hence

$$\mathbb{E}[|\tau|^p \mid Z = z] \lesssim_{p,\alpha,A_0} M^p \lesssim_{p,\alpha,A_0} (|z| \vee 1)^p,$$

since we have simultaneous moment control by Assumption 2. □

## Appendix SM7. A large-deviation inequality for the average Hellinger distance

**Theorem SM7.1.** *For some $n \ge 7$, let $\tau_1, \dots, \tau_n \mid (\nu_1^2, \dots, \nu_n^2) \overset{\text{i.i.d.}}{\sim} G_0$ 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 $\gamma_n, \lambda_n \le 1$ and constant $\epsilon > 0$. Fix some positive constant $C^*$. Consider the set of distributions that approximately maximize the likelihood*

$$A(\gamma_n, \lambda_n) = \left\{H : \mathrm{Sub}_n(H) \le C^* \left(\gamma_n^2 + \overline{h}(f_{H,\cdot}, f_{G_0,\cdot})\lambda_n\right)\right\}.$$

*Also consider the set of distributions that are far from $G_0$ in $\overline{h}$:*

$$B(t, \lambda_n, \epsilon) = \left\{H : \overline{h}(f_{H,\cdot}, f_{G_0,\cdot}) \ge tB\lambda_n^{1-\epsilon}\right\}$$

*with some constant $B$ to be chosen. Assume that for some $C_\lambda$,*

$$\lambda_n^2 \ge \left(\frac{C_\lambda}{n}(\log n)^{1+\frac{\alpha+2}{2\alpha}}\right) \vee \gamma_n^2. \tag{SM7.1}$$

*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*

$$\mathrm{P}\left[A(\gamma_n, \lambda_n) \cap B(t, \lambda_n, \epsilon) \neq \varnothing\right] \le (\log_2(1/\epsilon) + 1)n^{-t^2}. \tag{SM7.2}$$

**Corollary OA3.1.** *Assume Assumptions 1 to 4 hold and suppose $\Delta_n, M_n$ take the form (OA3.4). Define the rate function*

$$\delta_n = n^{-p/(2p+1)}(\log n)^{\frac{2+\alpha}{2\alpha}+\beta}. \tag{OA3.6}$$

*Then, there exists some constant $B_{\mathcal{H}}$, depending solely on $C_{\mathcal{H}}^*$ in Corollary SM6.1, $\beta$, and $p, \nu_\ell, \nu_u$ such that*

$$\mathrm{P}\left[A_n, \overline{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}) > B_{\mathcal{H}}\delta_n\right] \le \left(\frac{\log\log n}{\log 2} + 10\right)\frac{1}{n}.$$

*Proof.* Let $\gamma = \frac{2+\alpha}{2\alpha} + \beta$. We first note that, for $\varepsilon_n$ in (SM6.2), the choices

$$\lambda_n = n^{-p/(2p+1)}(\log n)^{\frac{2+\alpha}{2\alpha}+\beta} \wedge 1 = \gamma_n$$

do satisfy (SM7.1). Note that the choices of $\lambda_n, \gamma_n$ are such that $\varepsilon_n \le C_{\mathcal{H}}(\lambda_n\overline{h} + \gamma_n^2)$

The event $\left\{A_n, \overline{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}) > t\delta_n\right\}$ is a subset of the union of

$$E_1 = \left\{A_n, \mathrm{Sub}_n(\hat{G}_n) > C_{\mathcal{H}}^*\varepsilon_n\right\}$$

and

$$E_2 = \left\{A_n, \mathrm{Sub}_n(\hat{G}_n) \le C_{\mathcal{H}}^*\varepsilon_n, \overline{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}) > tn^{-p/(2p+1)}(\log n)^{\gamma}\right\}.$$

Thus $\mathrm{P}\left[A_n, \overline{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}) > t\delta_n\right] \leq \mathrm{P}(E_1) + \mathrm{P}(E_2)$. Corollary SM6.1 implies that $\mathrm{P}(E_1) \leq 9/n$.

Now, note that

$$\mathrm{P}(E_2) \leq \mathrm{P}\left[A_n, \mathrm{Sub}_n(\hat{G}_n) \leq C^*_{\mathcal{H}} C_{\mathcal{H}}(\lambda_n \overline{h} + \gamma_n^2), \overline{h} \geq t\lambda_n\right].$$

Observe that, for $\epsilon = 1/\log(n)$

$$\begin{aligned} t\lambda_n^{1-\epsilon} &= t\left[n^{-\frac{p}{2p+1}(1-\epsilon)}(\log n)^{\gamma(1-\epsilon)} \wedge 1\right] \\ &= t\left(n^{-\frac{p}{2p+1}}(\log n)^{\gamma}\left[n^{\frac{\epsilon p}{2p+1}}(\log n)^{-\gamma\epsilon}\right] \wedge 1\right) \\ &= t\left(n^{-\frac{p}{2p+1}}(\log n)^{\gamma}\left[e^{\frac{p}{2p+1}}(\log n)^{-\gamma\epsilon}\right] \wedge 1\right) \\ &\leq C_{p,\gamma} t\lambda_n \qquad\qquad (e^{\frac{p}{2p+1}}(\log n)^{-\gamma\epsilon} \text{ is bounded by a constant}) \end{aligned}$$

Thus, by Theorem SM7.1, for all sufficiently large $t$,

$$\begin{aligned} \mathrm{P}(E_2) &\leq \mathrm{P}\left[\mathrm{Sub}_n(\hat{G}_n) \leq C^*_{\mathcal{H}} C_{\mathcal{H}}(\lambda_n \overline{h} + \gamma_n^2), \overline{h} \geq \frac{t}{C_{p,\gamma}} \lambda_n^{1-\epsilon}\right] \\ &\leq \mathrm{P}\left[A(\gamma_n, \lambda_n) \cap B\left(\frac{t}{BC_{p,\lambda}}, \lambda_n, \epsilon\right) \neq \varnothing\right] \leq (\log_2(\log n) + 1)\, n^{-t^2/C_{\mathcal{H}}} \end{aligned}$$

We can pick $t = B_{\mathcal{H}}$ sufficiently large such that $n^{-t^2/C_{\mathcal{H}}} \leq 1/n$ and

$$\mathrm{P}\left[A_n, \overline{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}) > t\delta_n\right] \leq \mathrm{P}(E_1) + \mathrm{P}(E_2) \leq \left(\frac{\log\log n}{\log 2} + 10\right)\frac{1}{n}. \qquad \square$$

**SM7.1 Proof of Theorem SM7.1.**

*SM7.1.1 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 $K = \lceil|\log_2(1/\epsilon)|\rceil$.

In the remainder, we will bound

$$\mathrm{P}(A(\gamma_n, \lambda_n) \cap B_k(t, \lambda_n) \neq \varnothing) \leq n^{-t^2}$$

which becomes the bound (SM7.2) by a union bound. This argument follows the argument for Theorem 6 in Soloff et al. (2021) (arXiv: 2109.03466v1) and Theorem 4 in Jiang (2020). 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 $k$.

*SM7.1.2 Construct a net for the set of densities* $f_G$. Fix a positive constant $M$ and define the pseudonorm

$$\|G\|_{\infty,M} = \max_{i \in [n]} \sup_{y \in [-M,M]} f_{G,\nu_i}(y).$$

Note that $\|G\|_{\infty,M}$ is proportional to $\|G\|_{0,\infty,M}$ defined in Proposition SM6.1. Fix $\omega = \frac{1}{n^2} > 0$ and consider an $\omega$-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 $\omega$-net is the distributions $H_1, \ldots, H_N$. For each $j$, let $H_{k,j}$ be the distribution with

$$\|H_{k,j} - H_j\|_{\infty,M} \leq \omega \quad \overline{h}(f_{H_{k,j},\cdot}, f_{G_0,\cdot}) \geq t\mu_{n,k+1} \tag{SM7.3}$$

if it exists, and let $J_k$ collect the indices $j$ for which $H_{j,k}$ exists.

*SM7.1.3 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 also lower bounded:

$$\begin{aligned}\prod_{i=1}^n \frac{f_{H,\nu_i}(Z_i)}{f_{G_0,\nu_i}(Z_i)} &\ge \exp\left(-nC^*(\gamma_n^2 + \overline{h}\left(f_{H,\cdot}, f_{G_0,\cdot}\right)\lambda_n)\right) \\ &\ge \exp\left(-ntC^*(t\gamma_n^2 + \overline{h}\left(f_{H,\cdot}, f_{G_0,\cdot}\right)\lambda_n)\right) && (t > 1) \\ &\ge \exp\left(-nC^*(t^2\gamma_n^2 + t\overline{h}\lambda_n)\right) \\ &\ge \exp\left(-nC^*(t^2\gamma_n^2 + t^2\mu_{n,k}\lambda_n)\right)\end{aligned}$$

Hence,

$$\begin{aligned}&\mathrm{P}\left[A(t,\gamma_n,\lambda_n) \cap B_k(t,\lambda_n) \ne \varnothing\right] \\ &\le \mathrm{P}\left\{\left(\max_{j\in J_k} \prod_{i=1}^n \frac{f_{H_{k,j},\nu_i}(Z_i) + 2v(Z_i)}{f_{G_0,\nu_i}(Z_i)}\right) \prod_{i:|Z_i|>M} \frac{1}{\sqrt{2\pi}\nu_i v(Z_i)} \ge \exp\left(-nt^2C^*(\gamma_n^2 + \mu_{n,k}\lambda_n)\right)\right\} \\ &\le \mathrm{P}\left[\max_{j\in J_k} \prod_{i=1}^n \frac{f_{H_{k,j},\nu_i} + 2v(Z_i)}{f_{G_0,\nu_i}(Z_i)} \ge e^{-nt^2aC^*(\gamma_n^2+\mu_{n,k}\lambda_n)}\right] && \text{(SM7.4)} \\ &\quad + \mathrm{P}\left[\prod_{i:|Z_i|>M} \frac{1}{\sqrt{2\pi}\nu_i v(Y_i)} \ge e^{nt^2(a-1)C^*(\gamma_n^2+\mu_{n,k}\lambda_n)}\right] && \text{(SM7.5)}\end{aligned}$$

The second inequality follows from choosing some $a > 1$ and applying union bound.

*SM7.1.4 Bounding* (SM7.4). We consider bounding the first term (SM7.4) now:

$$\text{(SM7.4)} \le \sum_{j\in J_k} \mathrm{P}\left[\prod_{i=1}^{n} \frac{f_{H_{k,j},\nu_i} + 2v(Z_i)}{f_{G_0,\nu_i}(Z_i)} \ge e^{-nat^2C^*(\gamma_n^2+\mu_{n,k}\lambda_n)}\right] \qquad \text{(Union bound)}$$

$$\le \sum_{j\in J_k} \mathbb{E}\left[\prod_{i=1}^{n} \sqrt{\frac{f_{H_{k,j},\nu_i}(Z_i) + 2v(Z_i)}{f_{G_0,\nu_i}(Z_i)}}\right] e^{nat^2C^*(\gamma_n^2+\mu_{n,k}\lambda_n)/2}$$

(Take square root of both sides, then apply Markov's inequality)

$$= \sum_{j\in J_k} e^{nat^2C^*(\gamma_n^2+\mu_{n,k}\lambda_n)/2} \prod_{i=1}^{n} \mathbb{E}\left[\sqrt{\frac{f_{H_{k,j},\nu_i}(Z_i) + 2v(Z_i)}{f_{G_0,\nu_i}(Z_i)}}\right] \qquad \text{(SM7.6)}$$

where the last step (SM7.6) is by independence over $i$. Note that

$$\mathbb{E}\left[\sqrt{\frac{f_{H_{k,j},\nu_i}(Z_i) + 2v(Z_i)}{f_{G_0,\nu_i}(Y_i)}}\right] = \int_{-\infty}^{\infty} \sqrt{f_{H_{k,j},\nu_i}(x) + 2v(x)}\sqrt{f_{G_0,\nu_i}(x)}\, dx$$

$$\le 1 - h^2(f_{H_{k,j},\nu_i}, f_{G_0,\nu_i}) + \int_{-\infty}^{\infty} \sqrt{2v(x) f_{G_0,\nu_i}(x)}\, dx$$

$$(\sqrt{a+b} \le \sqrt{a} + \sqrt{b})$$

$$\le 1 - h^2(f_{H_{k,j},\nu_i}, f_{G_0,\nu_i}) + \left(2\int_{-\infty}^{\infty} v(x)\, dx\right)^{1/2} \qquad \text{(Jensen's inequality)}$$

$$= 1 - h^2(f_{H_{k,j},\nu_i}, f_{G_0,\nu_i}) + \sqrt{8M\eta} \qquad \text{(Direct integration)}$$

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 (SM7.6):

$$\text{(SM7.4)} \le \text{(SM7.6)} = \sum_{j\in J_k} e^{n\alpha t^2(\gamma_n^2+\mu_{n,k}\lambda_n)/2} \prod_{i=1}^{n} \mathbb{E}\left[\sqrt{\frac{f_{H_{k,j},\nu_i} + 2v(Z_i)}{f_{G_0,\nu_i}(Z_i)}}\right]$$

$$\le \sum_{j\in J_k} \exp\left\{\frac{nat^2C^*}{2}(\gamma_n^2 + \mu_{n,k}\lambda_n) - n\bar{h}^2(f_{H_{k,j},\cdot}, f_{G_0,\cdot}) + n\sqrt{8M\omega}\right\}$$

$$\le \sum_{j\in J_k} \exp\left\{\frac{nat^2C^*}{2}(\gamma_n^2 + \mu_{n,k}\lambda_n) - nt^2\mu_{n,k+1}^2 + n\sqrt{8M\omega}\right\}$$

$$(\bar{h}^2(f_{H_{k,j},\cdot}, f_{G_0,\cdot}) \ge t\mu_{n,k+1} \text{ by (SM7.3)})$$

$$\le \exp\left\{\frac{nat^2C^*}{2}(\gamma_n^2 + \mu_{n,k}\lambda_n) - nt^2\mu_{n,k+1}^2 + n\sqrt{8M\omega} + \log N\right\} \qquad (|J_k| \le N)$$

$$\leq \exp\left\{\frac{nat^2C^*}{2}(\gamma_n^2+\mu_{n,k}\lambda_n) - nt^2\mu_{n,k+1}^2 + n\sqrt{8M\omega} + C|\log\omega|^2 \max\left(\frac{M}{\sqrt{|\log\omega|}}, 1\right)\right\}$$

(Proposition SM6.1, $q = 0$)

$$= \exp\left\{\frac{nat^2C^*}{2}(\gamma_n^2+\mu_{n,k}\lambda_n) - nt^2\mu_{n,k+1}^2 + \sqrt{8M} + C(\log n)^2 \max\left(\frac{M}{\sqrt{\log n}}, 1\right)\right\}.$$

(Recall that $\omega = \frac{1}{n^2}$)

*SM7.1.5 Bounding* (SM7.5). We now consider bounding the second term (SM7.5). By Markov's inequality again (taking $x \mapsto x^{1/(2\log n)}$ on both sides, we can choose to bound

$$\text{(SM7.5)} \leq \mathbb{E}\left[\prod_{i=1}^n \left(\frac{1}{(2\pi\nu_i^2)^{1/4}}\frac{Z_i}{M\sqrt{\omega}}\right)^{\frac{1}{\log n}\mathbb{1}(|Z_i|>M)}\right] \exp\left(-\frac{n(a-1)t^2C^*(\gamma_n^2+\mu_{n,k}\lambda_n)}{2\log n}\right)$$

instead. Define

$$a_i = \frac{1}{(2\pi\nu_i^2)^{1/4}M\sqrt{\omega}} \leq \frac{C_{\nu_\ell}n}{M} \qquad \lambda = \frac{1}{\log n}$$

Apply Lemma SM7.1 to obtain the following. Note that to do so, we require $M \geq \nu_u\sqrt{8\log n}$ and $p \geq \frac{1}{\log n}$. Then,

$$\begin{aligned}
\log\mathbb{E}\left[\prod_{i=1}^n \left(\frac{1}{(2\pi\nu_i^2)^{1/4}}\frac{Z_i}{M\sqrt{\omega}}\right)^{\frac{1}{\log n}\mathbb{1}(|Z_i|>M)}\right] &= \log\mathbb{E}\left[\prod_i (a_iZ_i)^{\lambda\mathbb{1}(|Z_i|\geq M)}\right] \\
&\lesssim_{\nu_u} \sum_{i=1}^n (a_iM)^\lambda \left(\frac{1}{Mn} + \frac{2^p\mu_p^p(G_0)}{M^p}\right) \quad \text{(Lemma SM7.1)} \\
&\leq \sum_{i=1}^n (C_{\nu_\ell}n)^{\frac{1}{\log n}} \left(\frac{1}{Mn} + \frac{2^p\mu_p^p(G_0)}{M^p}\right) \\
&\lesssim_{\nu_u,\nu_\ell} \frac{1}{M} + \frac{2^pn\mu_p^p(G_0)}{M^p}
\end{aligned}$$

As a result,

$$\log[\text{(SM7.5)}] \leq C_{\nu_u,\nu_\ell}\left(\frac{1}{M} + \frac{2^pn\mu_p^p(G_0)}{M^p}\right) - \frac{n(a-1)}{2\log n}t^2C^*\left(\gamma_n^2 + B\lambda_n^{2(1-2^{-k})}\right). \tag{SM7.7}$$

To conclude, note that 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

$$2^p\mu_p^p(G_0)/M^p \leq \exp(-c_m\log n)$$

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 SM7.1 is correct. We also choose $a = 1.5$.

Plugging in these choices, we can verify that, via (SM7.1),

$$\log[\text{(SM7.5)}] \leq t^2\left[2C_{\nu_u,\nu_\ell} - \frac{C^*BC_\lambda}{4}(\log n)\right]$$

$$\log[\text{(SM7.4)}] \leq -t^2(\log n)^{1+\frac{2+\alpha}{2\alpha}}\left[C_\lambda\left(-\frac{3}{4}C^* - \frac{3}{4}C^*B + B^2\right) - C\right]$$

There exists a sufficiently large choice of $B$ such that $\log[\text{(SM7.5)}] \leq -t^2\log n$ and $\log[\text{(SM7.4)}] \leq -t^2\log n - \log 2$. Thus, we obtain that (SM7.4) + (SM7.5) $\leq n^{-t^2}$. This concludes the proof.

**SM7.2 Auxiliary lemmas.**

**Lemma SM7.1** (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* $\nu_i$. *Then, for all constants* $M > 0, \lambda > 0, a_i > 0, p \in \mathbb{N}$ *such that* $M \geq \nu_u\sqrt{8\log n}$, $\lambda \in (0, p \wedge 1)$, *and* $a_1, \ldots, a_n > 0$*:*

$$\mathbb{E}\left\{\prod_i |a_i Z_i|^{\lambda \mathbb{1}(|Z_i| \geq M)}\right\} \leq \exp\left\{\sum_{i=1}^{n} (a_i M)^\lambda \left(\frac{4\nu_u}{Mn\sqrt{2\pi}} + \left(\frac{2\mu_p(G_0)}{M}\right)^p\right)\right\}.$$

**Part 4 Additional theoretical results**

## Appendix SM8. Estimating $\eta_0$ by local linear regression

In this section, we verify that estimating $\eta_0$ 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 $x_i \equiv \log_{10}(\sigma_i)$.[57] 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 $A_3 \lesssim_{\mathcal{H}} A_1$.

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:[58]

$$\hat{h}_{n,m} \leftarrow (\hat{h}_{n,m} \vee C_h^{-1} n^{-1/5}) \wedge C_h n^{-1/5}.$$

(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

$$p_n = \frac{1}{n} \sum_{i=1}^n \frac{1}{\sum_{j=1}^n \ell_i^2(x_j, \hat{h}_{n,s})}. \tag{SM8.1}$$

[57] Correspondingly, let $\sigma(x) = 10^x$.

[58] We use the $\leftarrow$ notation to reassign a variable so that we can reduce notation clutter.

(LLR-10) Truncate the estimated conditional standard deviation:

$$\hat{s}_{\text{raw}}(x) = \sqrt{\hat{v}(x) - \sigma^2(x)} \vee \sqrt{\frac{2}{p_n + 2}\hat{v}(x)}. \tag{SM8.2}$$

(LLR-11) Finally, project the resulting estimate to the Hölder class as in (LLR-4):

$$\hat{s}(x) \in \underset{\substack{s \in C^2_{A_3}([0,1]) \\ s^2(\cdot) \geq \frac{2}{p_n+2} \min_i \sigma_i^2}}{\arg\min} \|s - \hat{s}_{\text{raw}}\|_\infty.$$

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 $\hat{m} = \hat{m}_{\text{raw}}, \hat{s} = \hat{s}_{\text{raw}}$. Omitting the projection steps does not appear to affect performance.

To ensure we always have a positive estimate of $s_0$, we truncate at a particular point (SM8.2). 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 SM8.1** (The truncation rule in (SM8.2))**.** The truncation rule in (SM8.2) is an ad hoc adjustment without affecting asymptotic performance.[59] It is based on a literature on the estimation of non-central $\chi^2$ 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 $\chi^2$ random variable with $p$ degrees of freedom and noncentrality parameter $\lambda = \sum_{i=1}^p \lambda_i^2$. The UMVUE for $\lambda$ 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 $(V-p) \vee \frac{2}{p+2}V$.[60]

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

$$\hat{v}(x_0) = \frac{\sigma(x_0)^2}{p_0} \sum_{i=1}^p \left(\frac{R_i}{\sigma(x_0)}\right)^2$$

[59] 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 $s_\ell$, 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.

[60] Though, since neither $(V-p)_+$ and $(V-p) \vee \frac{2}{p+2}V$ is differentiable in $V$, they are not admissible.

Conditional on $\sigma_i^2, \theta_i$, the quantity $\sum_{i=1}^p \left(\frac{R_i}{\sigma(x_0)}\right)^2$ is (approximately) noncentral $\chi^2$ 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 $\theta_i$, $\frac{1}{p_0}\sum_{i=1}^{p_0}(\theta_i - m_0(x_i))^2$, is

$$\left(\hat{v}(x_0) - \sigma^2(x_0)\right) \vee \frac{2}{p_0+2}\hat{v}(x_0).$$

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

$$\mathrm{Var}\left(\sum_i \ell_i(x) V_i\right) = \sum_{i=1}^n \ell_i^2(x).$$

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 (SM8.1) 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 SM8.1.** *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 SM8.1 is automatically satisfied by the projection steps (LLR-3) and (LLR-7).

**Assumption SM8.2.** *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 (SM8.1) 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 SM8.2(1–3) is nearly the same as Assumption (LP) in Tsybakov (2008). The only difference is that Assumption SM8.2(1) requires the lower bound $\lambda_0$ to hold uniformly over a range of bandwidth choices, relative to LP-1 in Tsybakov (2008), which requires $\lambda_0$ 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 SM8.2(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 SM8.2 is satisfied by regular design points $x_i = i/n$.

**Assumption SM8.3.** *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 $Q$,*

$$N(\epsilon, \mathcal{F}_k, L_2(Q)) \leq C_K (1/\epsilon)^{V_K}$$

*for $C_K, V_K$ that do not depend on $Q$.*

Assumption SM8.3 is satisfied for a wide range of kernels, e.g. the Epanechnikov kernel. By Lemma 7.22 in Sen (2018), reproduced as Lemma SM8.2 below, so long as the function $t \mapsto t^{k-1} K(t)$ is bounded (assumed in Assumption SM8.2(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 SM8.1 on the bias and variance of local linear regression estimators. Proposition SM8.1 is uniform in both the evaluation point $x$ and the bandwidth $h$, as long as the latter converges at the optimal rate.

**Theorem SM8.1.** *Suppose the conditional distribution $\theta_i \mid \sigma_i$ and the design points $\sigma_{1:n}$ satisfy Assumptions 2, 3, and SM8.2. Moreover, suppose $m_0, s_0$ satisfies Assumption 4(1) with $p = 2$. Suppose the kernel $K(\cdot)$ satisfies Assumption SM8.3. 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 $t > 1$,*

$$\mathrm{P}\left(\max\left(\|\hat{m} - m_0\|_\infty, \|\hat{s} - s_0\|\right) \ge Ctn^{-\frac{2}{5}}(\log n)^{1+2/\alpha}\right) \le \frac{1}{n^{10}t^2}. \tag{SM8.3}$$

*(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

$$\mathrm{P}\left(\max\left(\|\hat{m}_{\text{raw}} - m_0\|_\infty, \|\hat{s}_{\text{raw}} - s_0\|_\infty\right) > Ctn^{-2/5}(\log n)^{\beta}\right) \le \frac{1}{n^{10}t^2}$$

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 $\xi_i$:

$$\max_i \mathbb{E}[|\xi_i|^p]^{1/p} \lesssim p^{1/\alpha}$$

where $\alpha$ is the constant in Assumption 2. Therefore, we can apply Proposition SM8.1 to obtain

$$\mathrm{P}\left(\|\hat{m}_{\text{raw}} - m_0\|_\infty > Ctn^{-2/5}(\log n)^{1+1/\alpha}\right) \le \frac{1}{2n^{10}t^2}$$

for the local linear regression estimator $\hat{m}_{\text{raw}}$.

The same argument to control $\|\hat{s}_{\text{raw}} - s_0\|_\infty$ is more involved. First observe that

$$|\hat{s}^2_{\text{raw}} - s_0^2| = |\hat{s}_{\text{raw}} - s_0|(\hat{s}_{\text{raw}} + s_0) \ge s_{0\ell}|\hat{s}_{\text{raw}} - s_0|.$$

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

$$\|\hat{s}_{\mathrm{raw}} - s_0\|_\infty \leq \frac{1}{s_{0\ell}} \left( \|\hat{v} - v_0\|_\infty \vee \left( \frac{2}{2+p_n} \hat{v} \right) \right) \qquad (v_0(x) \equiv \mathrm{Var}(Y_i \mid x_i = x))$$

$$\lesssim \|\hat{v} - v_0\|_\infty \vee \frac{\|\hat{v} - v_0\|_\infty + \|v_0\|_\infty}{2 + n^{4/5}} \qquad \text{(Assumption SM8.2)}$$

$$\lesssim \|\hat{v} - v_0\|_\infty \tag{SM8.4}$$

Now, observe that $\hat{R}_i^2 = R_i^2 + (m_0 - \hat{m})^2 - 2(m_0 - \hat{m})\xi_i$. Hence,

$$|\hat{v}(x) - v_0(x)| \leq \left| \sum_{i=1}^n \ell_i(x, \hat{h}_{n,s}) R_i^2 - v_0(x) \right| + \left\{ \|m_0 - \hat{m}\|_\infty^2 + 2\|m_0 - \hat{m}\|_\infty \left( \max_{i \in [n]} |\xi_i| \right) \right\} \sum_{i=1}^n |\ell_i(x, \hat{h}_{n,s})|$$

$$\leq \left| \sum_{i=1}^n \ell_i(x, \hat{h}_{n,s}) R_i^2 - v_0(x) \right| + C \left\{ \|m_0 - \hat{m}\|_\infty^2 + 2\|m_0 - \hat{m}\|_\infty \left( \max_{i \in [n]} |\xi_i| \right) \right\}. \tag{SM8.5}$$

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 SM8.1 and taking care to plug in $\tilde{\xi}, \tilde{\alpha}$, we can bound the first term in (SM8.5)

$$\mathrm{P}\left( \left\| \sum_{i=1}^n \ell_i(x, \hat{h}_{n,s}) R_i^2 - v_0(x) \right\|_\infty \geq Ctn^{-2/5} (\log n)^{1+2/\alpha} \right) \leq \frac{1}{4n^{10} t^2}.$$

Note that by an application of Lemma OA3.7, 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 (SM8.5) 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 $\hat{s}_{\mathrm{raw}}$ by (SM8.4).

Combining the bounds for $\hat{m}$ and $\hat{s}$, we obtain (SM8.3). This concludes the proof. ☐

**Theorem SM8.2.** *Under the assumptions of Theorem SM8.1, 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 (A.1) 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 $\tilde{A}_n = A_n \cap \{s_{0\ell}/2 \le \hat{s} \le 2s_{0u}\}$. 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 $C_1, \beta, s_\ell, s_u$, the event $A_n$ implies $\{s_{0\ell}/2 \le \hat{s} \le 2s_{0u}\}$ and hence $A_n = \tilde{A}_n$. Thus, by Theorem SM8.1, 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 A.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 OA3.6 and Theorem SM8.1(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 OA3.2,

$$\mathbb{E}\left[\mathrm{MSERegret}_n(\hat{G}_n, \hat{\eta})\mathbb{1}(\tilde{A}_n^{\mathrm{C}})\right] \lesssim \mathrm{P}(\tilde{A}_n^{\mathrm{C}})^{1/2} n^4 (\log n)^{2/\alpha},$$

where we apply Lemma OA3.7 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 SM8.1 to bound $\|\hat{\eta} - \eta_0\|_\infty$. We can apply Lemma OA3.7 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 $n$. □

**SM8.1 Auxiliary lemmas.**

**Proposition SM8.1.** *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 SM8.2 and the (possibly data-driven) bandwidths $h_n$ satisfy Assumption SM8.1. Assume the kernel additionally satisfies Assumption SM8.3.*

*Assume that the residuals $\xi_i$ 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 $p \geq 2$. 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 SM8.1. 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 $t\geq 1$*

$$\mathrm{P}\left(\sup_{x\in[0,1],h\in H_n} |v(x,h)| > C(a,b)\cdot t\cdot (\log n)^{1+1/\alpha} n^{-2/5}\right) \leq an^{-b}\frac{1}{t^2}.$$

*(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 $t\geq 1$,*

$$\mathrm{P}\left(\|\hat{f} - f_0\|_\infty > C(a,b) t (\log n)^{1+1/\alpha} n^{-2/5}\right) \leq an^{-b}\frac{1}{t^2}.$$

*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

$$\sup_{x\in[0,1]} |b(x,h_n)| \leq Ch_n^2,$$

which is uniformly bounded by $Cn^{-2/5}$ by Assumption SM8.1. 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

$$\xi_{i,<M} = \xi_i \mathbb{1}(|\xi_i|\leq M) - \mathbb{E}[\xi_i \mathbb{1}(|\xi_i|\leq M)] \quad \xi_{i,>M} = \xi_i \mathbb{1}(|\xi_i|> M) - \mathbb{E}[\xi_i \mathbb{1}(|\xi_i|> M)]$$

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 $x, h_n$,

$$\begin{aligned} V_{1n}^2 &\leq \sum_{i=1}^n \ell_i(x,h_n)^2 \sum_{i=1}^n \xi_{i,>M}^2 \\ &\lesssim \frac{1}{h_n^2}\frac{1}{n}\sum_{i=1}^n \xi_{i,>M}^2 && \text{(Lemma 1.3(i) in Tsybakov (2008) shows that } |\ell_i(x,h_n)| \leq \tfrac{C}{nh_n}) \\ &\lesssim n^{2/5}\frac{1}{n}\sum_{i=1}^n \xi_{i,>M}^2 \end{aligned}$$

Now, for some $C$ related to the implicit constant in the above display,

$$\mathrm{P}\left(\sup_{x\in[0,1],h_n\in H_n} V_{1n}^2(x,h_n) > Ct^2\right) \le \mathrm{P}\left(\frac{1}{n}\sum_{i=1}^n \xi_{i,>M}^2 > t^2 n^{-2/5}\right) \le \frac{\max_i \mathbb{E}\xi_{i,>M}^2}{t^2} n^{2/5}.$$

(Markov's inequality)

We note that by Cauchy–Schwarz,

$$\mathbb{E}[\xi_{i,>M}^2] \le \sqrt{\mathbb{E}[\xi_i^4]}\sqrt{\mathrm{P}(|\xi_i|>M)} \lesssim \sqrt{\mathrm{P}(|\xi_i|>M)} \le \exp\left(-cM^\alpha\right) \qquad \text{(Lemma SM6.15)}$$

where $c$ depends on $A_\xi$. Hence, for a potentially different constant $C$,

$$\mathrm{P}\left(\sup_{x\in[0,1],h_n\in H_n} |V_{1n}(x,h_n)| > Ct\right) \le \exp\left(-cM^\alpha - 2\log t + \frac{2}{5}\log n\right). \tag{SM8.6}$$

Next, consider the process

$$\begin{aligned}
V_{2n}(x,h_n) &= \sum_{i=1}^n \ell_i(x,h_n)\xi_{i,<M} \\
&= \frac{1}{nh_n}\sum_{i=1}^n \underbrace{u(0)'B_{nx}^{-1}\begin{bmatrix}1\\0\end{bmatrix}}_{A_1(x,h_n)} K\left(\frac{x_i-x}{h_n}\right)\xi_{i,<M} \\
&\quad + \frac{1}{nh_n}\sum_{i=1}^n \underbrace{u(0)'B_{nx}^{-1}\begin{bmatrix}0\\1\end{bmatrix}}_{A_2(x,h_n)} K\left(\frac{x_i-x}{h_n}\right)\left(\frac{x_i-x}{h_n}\right)\xi_{i,<M} \\
&\equiv \frac{A_1(x,h_n)}{h_n}\frac{1}{n}\sum_{i=1}^n K\left(\frac{x_i-x}{h_n}\right)\xi_{i,<M} + \frac{A_2(x,h_n)}{h_n}\frac{1}{n}\sum_{i=1}^n K\left(\frac{x_i-x}{h_n}\right)\left(\frac{x_i-x}{h_n}\right)\xi_{i,<M}.
\end{aligned}$$

Note that, by Assumption SM8.2(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 $\psi_2$-norm of the left-hand side. Note that it suffices to control the $\psi_2$-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 SM8.1. The analogue of $f$ in Lemma SM8.1 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 SM8.2(2). Therefore, by Assumption SM8.3, we apply Lemma SM8.1 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

$$\left\| \sup_{x \in [0,1], h \in H_n} |V_{2n}(x,h)| \right\|_{\psi_2} \lesssim \frac{1}{\sqrt{n} h_n n^{1/10}} M \sqrt{\log n} \lesssim n^{-2/5} M \sqrt{\log n}. \tag{SM8.7}$$

Putting things together, we can choose $M = (c_m \log n)^{1/\alpha}$ for sufficiently large $c_m$ so that by (SM8.6),

$$\mathrm{P}\left( \sup_{x \in [0,1], h \in H_n} |V_{1n}(x,h)| > C t n^{-2/5} \right) \le \frac{a}{2} n^{-b} \frac{1}{t^2},$$

where $c_m$ depends on $a, b$. The bound (SM8.7) 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,

$$\mathrm{P}\left( \sup_{x \in [0,1], h_n \in H_n} |v(x,h_x)| > C(a,b) t (\log n)^{1+1/\alpha} n^{-2/5} \right) \le a n^{-b} \frac{1}{t^2}.$$

□

**Lemma SM8.1.** *Suppose $\xi_i$ are bounded by $M \geq 1$ 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:*

$$N(\epsilon, \mathcal{F}, \|\cdot\|_n) \leq C(1/\epsilon)^V \quad \epsilon \in [0,1]$$

*where $C, V > 0$ depend solely on $\mathcal{F}$. Then*

$$\left\| \sup_{f \in \mathcal{F}} |V_n(f)| \right\|_{\psi_2} \lesssim M \operatorname{diam}(\mathcal{F}) \sqrt{\log(1/\operatorname{diam}(\mathcal{F}))},$$

*where $\operatorname{diam}(\mathcal{F}) = \sup_{f_1, f_2 \in \mathcal{F}} \|f_1 - f_2\|_n$.*

*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 $f_0 \in \mathcal{F}$,

$$\left\| \sup_f V_n(f) \right\|_{\psi_2} \leq \|V_n(f_0)\|_{\psi_2} + CM \int_0^{\operatorname{diam}(\mathcal{F})} \sqrt{\log N(\delta, \mathcal{F}, \|\cdot\|_n)}\, d\delta.$$

Note that (i) the metric entropy integral is bounded by $C\operatorname{diam}(\mathcal{F})\sqrt{\log(1/\operatorname{diam}(\mathcal{F}))}$, and (ii) for a fixed $f_0$, $\|V_n(f_0)\|_{\psi_2} \lesssim \|f_0\|_n M \leq \operatorname{diam}(\mathcal{F}) M$ since $0 \in \mathcal{F}$. Therefore,

$$\left\| \sup_f V_n(f) \right\|_{\psi_2} \lesssim M \operatorname{diam}(\mathcal{F}) \sqrt{\log(1/\operatorname{diam}(\mathcal{F}))}.$$

□

**Lemma SM8.2** (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.*

## Appendix SM9. Minimax lower bound for nonparametric regression

**Lemma SM9.1.** *Under the setup of Theorem 2, suppose $Y_i \sim \mathcal{N}(m_0(\sigma_i), s_0^2 + \sigma_i^2)$, then*

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

*where the infimum is over all estimators of $m_0$ from $(Y_i, \sigma_i)_{i=1}^n$ and the supremum is over all Hölder continuous $m_0(\cdot)$.*

*Proof.* First, note that learning $m_0$ from $(Y_i, \sigma_i)$ is a nonparametric regression problem with heteroskedastic variances. This problem is more difficult than a corresponding problem with homoskedastic variances $\sigma_\ell^2 + s_0^2$. Since we may represent

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

for independent Gaussians $W_i, U_i \sim \mathcal{N}(0,1)$. Let $V_i = \theta_i + \sigma_\ell W_i$. Note that we can do no worse for estimating $m_0$ with $(V_i, \sigma_i)$ than with $(Y_i, \sigma_i)$, and estimating $m_0$ from $(V_i, \sigma_i)$ is a homoskedastic regression problem, where $V_i \sim \mathcal{N}(m_0(\sigma_i), \sigma_\ell^2 + s_0^2)$. It remains to show that the minimax rate for estimating $m_0$ on the grid points $\sigma_{1:n}$ from $(V_i, \sigma_i)$ is $n^{-2p/(2p+1)}$.

Since we simply have a nonparametric regression problem, 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. Corollary 2.3 in Tsybakov (2008) shows a lower bound for integrated MSE:

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

where the infimum is over all (randomized) estimators using $(V_i, \sigma_i)$. It thus suffices to connect the MSE objective over $\sigma_1, \ldots, \sigma_n$ to the integrated MSE.

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 $\ell_2$ 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

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

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 $p \geq 1$ in Assumption 4. Therefore,

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

## Appendix SM10. Regret for $\theta^v$

This section contains novel results on the squared-error regret for higher moments $\theta^v$ for some positive integer $v$. Our main theoretical results correspond to the case $v = 1$. These results are novel for the setting with homoskedasticity or prior independence as well (Jiang and Zhang, 2009; Soloff et al., 2021). The key to this result is a generalization (Theorem SM10.2) of Theorem 3 in Jiang and Zhang (2009), and a generalization (Lemma SM10.4) of Lemma A.1 in Jiang and Zhang (2009). These results then provide a suitable generalization for Lemma OA3.9, which is key in controlling the distance between posterior means of two priors as a function of their (smoothed) Hellinger distance.

We will prove a result similar to Theorem A.1. Recall $A_n$ from (A.1) defined given some $\Delta_n, M_n$. The following theorem controls the regret truncated by the event $A_n$. With $\beta = \beta_0$ in Assumption 4, with appropriate choice of constants in $\Delta_n, M_n$, we can ensure that $\mathrm{P}(A_n) \geq 1 - \delta$ for any $\delta \in (0, 1/2)$.

**Theorem SM10.1.** *Suppose Assumptions 1 to 4 hold. Fix some $\beta > 0, C_1 > 0$, there exists choices of constants $C_{\mathcal{H},m,2}$ such that, for $\Delta_n = C_1 n^{-p/(2p+1)}(\log n)^\beta$, $M_n = C_{\mathcal{H},m,2}(\log n)^{1/\alpha}$, and corresponding $A_n$,*

$$\mathbb{E}\left[\frac{\mathbb{1}(A_n)}{n}\sum_{i=1}^n \left(\mathbf{E}_{\hat{G}_n,\hat{\eta}}[\theta_i^v \mid Y_i, \sigma_i] - \mathbb{E}[\theta_i^v \mid Y_i, \sigma_i]\right)^2\right] \lesssim_{\mathcal{H},v} n^{-\frac{2p}{2p+1}} (\log n)^{\frac{2+\alpha}{\alpha} + 2\beta + \frac{2(v-1)}{\alpha} \vee (3v)}.$$

*Proof.* We let $\theta^*_{v,i} = \mathbb{E}[\theta_i^v \mid Y_i, \sigma_i]$ be the oracle posterior means and let $\hat{\theta}_{v,i} = \mathbf{E}_{\hat{G}_n,\hat{\eta}}[\theta_i^v \mid Y_i, \sigma_i]$ denote the estimated posterior means from our procedure.

Note that, for a given $G, \eta$, the posterior means take the form

$$\begin{aligned}
&|\mathbf{E}_{G,\eta}[\theta^v \mid Y, \sigma] - \mathbb{E}[\theta^v \mid Y, \sigma]| \\
&= |\mathbf{E}_{G,\eta}\left[(s(\sigma)\tau + m(\sigma))^v \mid Y, \sigma\right] - \mathbb{E}[(s_0(\sigma)\tau + m_0(\sigma))^v \mid Y, \sigma]| \\
&\leq \sum_{k=0}^v C_{k,v} \left| s(\sigma)^k m(\sigma)^{v-k} \mathbf{E}_{G,\eta}[\tau^k \mid Y, \sigma] - s_0(\sigma)^k m_0(\sigma)^{v-k} \mathbb{E}[\tau^k \mid Y, \sigma]\right| \qquad \text{(Expand } (a+b)^v\text{)} \\
&\lesssim_v \sum_{k=0}^v \max\left(\left|\mathbf{E}_{G,\eta}[\tau^k \mid Y, \sigma]\right|, \left|\mathbb{E}[\tau^k \mid Y, \sigma]\right|\right) \left| s(\sigma)^k m(\sigma)^{v-k} - s_0(\sigma)^k m_0(\sigma)^{v-k}\right| \\
&\quad + \sum_{k=0}^v \max\left(s(\sigma)^k |m(\sigma)|^{v-k}, s_0(\sigma)^k |m_0(\sigma)|^{v-k}\right) \left|\mathbf{E}_{G,\eta}[\tau^k \mid Y, \sigma] - \mathbb{E}[\tau^k \mid Y, \sigma]\right| \qquad \text{(SM10.1)}
\end{aligned}$$

We first observe that if $\|\eta - \eta_0\|_\infty \leq \Delta_n$ and $\|\eta_0\|_\infty \lesssim_{\mathcal{H}} 1$, then

$$\left| s(\sigma)^k m(\sigma)^{v-k} - s_0(\sigma)^k m_0(\sigma)^{v-k}\right| \lesssim_v \Delta_n$$

and

$$\max\left(s(\sigma)^k|m(\sigma)|^{v-k}, s_0(\sigma)^k|m_0(\sigma)|^{v-k}\right) \lesssim_{\mathcal{H},v} 1$$

Next, note that on $A_n$,

$$\left|\mathbf{E}_{\hat{G}_n,\hat{\eta}}[\tau^k \mid Y,\sigma]\right| \lesssim_{\mathcal{H}} M_n^k$$

since the prior $G_n$ is bounded within $[\min \hat{Z}_i, \max \hat{Z}_i]$. By Lemma SM6.17, we have that, likewise on $A_n$, the true posterior means are bounded $\left|\mathbb{E}[\tau^k \mid Y,\sigma]\right| \lesssim_{\mathcal{H}} M_n^k$. Thus, we conclude that

$$\begin{aligned}
&\mathbb{1}(A_n)\left(\mathbf{E}_{\hat{G}_n,\hat{\eta}}[\theta_i^v \mid Y_i,\sigma_i] - \mathbb{E}[\theta_i^v \mid Y_i,\sigma_i]\right)^2 \\
&\lesssim_{\mathcal{H},v} \Delta_n^2 M_n^{2v} + \mathbb{1}(A_n)\max_{1\le k\le v}\left(\mathbf{E}_{\hat{G}_n,\hat{\eta}}[\tau^k \mid Y,\sigma] - \mathbb{E}[\tau^k \mid Y,\sigma]\right)^2
\end{aligned}$$

Summing over $i$,

$$\mathbb{E}\left[\frac{\mathbb{1}(A_n)}{n}\sum_{i=1}^n\left(\mathbf{E}_{\hat{G}_n,\hat{\eta}}[\theta_i^v \mid Y_i,\sigma_i] - \mathbb{E}[\theta_i^v \mid Y_i,\sigma_i]\right)^2\right] \lesssim_{\mathcal{H},v} \Delta_n^2 M_n^{2v} + \sum_{1\le k\le v}\mathbb{E}\left[\frac{\mathbb{1}(A_n)}{n}\|\tau_v^* - \hat{\tau}_v\|^2\right], \tag{SM10.2}$$

where we define

$$\tau_{v,i}^* = \mathbb{E}[\tau^v \mid Y,\sigma] = \mathbb{E}_{G_0}[\tau^v \mid Z,\nu] \quad \hat{\tau}_{v,i} = \mathbf{E}_{\hat{G}_n,\hat{\eta}}[\tau^v \mid Y,\sigma] = \mathbf{E}_{\hat{G}_n}[\tau^v \mid \hat{Z},\hat{\nu}]$$

and define $\tau_v^*, \hat{\tau}_v$ as $\mathbb{R}^n$ vectors collecting these entries.

The rest of the proof in Appendix SM10.1 focuses on on showing that

$$\mathbb{E}\left[\frac{\mathbb{1}(A_n)}{n}\|\tau_v^* - \hat{\tau}_v\|^2\right] \lesssim_{v,\mathcal{H}} \text{(SM10.3)},$$

which dominates the rate in (SM10.2). Plugging the rates for $\delta_n^2, M_n$ concludes the proof. □

**SM10.1 Bounding $\mathbb{E}\left[\frac{\mathbb{1}(A_n)}{n}\|\tau_v^* - \hat{\tau}_v\|^2\right]$.** We now decompose

$$\mathbb{1}(A_n)\|\tau_v^*-\hat{\tau}_v\|^2 \lesssim \mathbb{1}(A_n)\left[\|\hat{\tau}_{v,\eta_0} - \hat{\tau}_v\|^2 + \|\hat{\tau}_{v,\eta_0,\rho_n} - \hat{\tau}_{v,\eta_0}\|^2 + \|\hat{\tau}_{v,\eta_0,\rho_n} - \tau^*_{v,\eta_0,\rho_n}\|^2 + \|\tau^*_{v,\eta_0} - \tau^*_{v,\eta_0,\rho_n}\|^2\right]$$

where $\hat{\tau}_{v,\eta_0}$ is the posterior mean for $\tau$ under $(\hat{G}_n,\eta_0)$, and $\hat{\tau}_{v,\eta_0,\rho_n}, \tau^*_{v,\rho_n}$ are defined in (SM10.5). This decomposition is analogous to the $\xi_1$ through $\xi_4$ decomposition in the proof for Theorem A.1. The following subsections bound each term individually (here, the second term is zero). The dominant rate is in the third term, where we show that

$$\mathbb{E}\left[\frac{\mathbb{1}(A_n)}{n}\|\tau_v^* - \hat{\tau}_v\|^2\right] \lesssim_{v,\mathcal{H}} \delta_n^2(\log n)^{\frac{2(v-1)}{\alpha}\vee(3v)}. \tag{SM10.3}$$

*SM10.1.1 Bounding $\|\hat{\tau}_{v,\eta_0} - \hat{\tau}_v\|^2$.* For a given $G,\eta$, note that

$$\begin{aligned}
\mathbf{E}_{G,\eta}[\tau^v \mid Z(\eta),\nu(\eta)] &= \mathbf{E}_{G,\eta}[(\tau - Z + Z)^v \mid Z,\nu] \\
&= \sum_{k=0}^v C_{k,v} Z(\eta)^{v-k}\mathbf{E}_{G,\eta}[(\tau-Z)^k \mid Z(\eta),\nu(\eta)]
\end{aligned} \tag{SM10.4}$$

Thus, define

$$U_k(m,s,Z) = \frac{\int \left(\frac{Z-\tau}{\nu(\eta)}\right)^k \varphi(\frac{Z-\tau}{\nu(\eta)})\,\hat{G}_n(d\tau)}{\nu(\eta) f_{\hat{G}_n,\nu(\eta)}(Z(\eta))} = \mathbf{E}_{\hat{G}_n,\eta}[(Z-\tau)^k \mid Z(\eta),\nu(\eta)]\frac{1}{\nu(\eta)^k}.$$

We have that

$$\begin{aligned}
&\mathbb{1}(A_n)\left|\mathbf{E}_{\hat{G}_n,\hat{\eta}}[\tau^v \mid \hat{Z},\hat{\nu}] - \mathbf{E}_{\hat{G}_n,\eta_0}[\tau^v \mid Z,\nu]\right| \\
&\lesssim_{v,\mathcal{H}} \mathbb{1}(A_n)\sum_{k=0}^{v} \max\left(|Z(\hat{\eta})|^{v-k}, |Z(\eta_0)|^{v-k}\right)|U_k(\hat{m},\hat{s},Z) - U_k(m_0,s_0,Z)| \\
&\quad + \mathbb{1}(A_n)\sum_{k=0}^{v-1}|Z(\hat{\eta})^{v-k} - Z(\eta_0)^{v-k}|\max\left(|U_k(\hat{m},\hat{s},Z)|, |U_k(m_0,s_0,Z)|\right)
\end{aligned}$$

Now, observe that

$$\begin{aligned}
\mathbb{1}(A_n)|Z(\hat{\eta})^{v-k} - Z(\eta_0)^{v-k}| &\lesssim_{v,\mathcal{H}} \Delta_n |Z(\eta_0)|^{v-k-1} \\
\mathbb{1}(A_n)\max\left(|Z(\hat{\eta})|^{v-k}, |Z(\eta_0)|^{v-k}\right) &\lesssim_{v,\mathcal{H}} |Z(\eta_0)|^{v-k} \\
\mathbb{1}(A_n)\max\left(|U_k(\hat{m},\hat{s},Z)|, |U_k(m_0,s_0,Z)|\right) &\lesssim_{v,\mathcal{H}} \log^{k/2}(n) \qquad \text{(Lemmas OA3.1 and SM10.3)}
\end{aligned}$$

For $|U_k(\hat{m},\hat{s},Z) - U_k(m_0,s_0,Z)|$, let $U_{km}(m,s,Z), U_{ks}(m,s,Z)$ be its partial derivative with respect to $m$ and $s$, respectively. Take some intermediate $\tilde{m},\tilde{s}$ lying on the line segment between $\hat{\eta},\eta_0$. By Taylor's theorem

$$|U_k(\hat{\eta},Z) - U_k(\eta_0,Z)| \lesssim \|\hat{\eta}-\eta_0\|_\infty \max\left(|U_{km}(\tilde{m},\tilde{s},Z)|, |U_{ks}(\tilde{m},\tilde{s},Z)|\right)$$

Differentiating, we find that $U_{km}$ takes the form of posterior moments of $(Z-\tau)/\nu$ to the at most $k+1^{\text{st}}$ power. We find that $U_{ks}$ takes the form of posterior means for $\tau\,((Z-\tau)/\nu)^{k+1}$ as its leading term. By the argument in the proof to Lemma SM6.13, we find that $f_{\hat{G}_n,\tilde{\nu}_i}(\tilde{Z}_i)$, evaluated at $\tilde{\eta}$, is lower bounded such that $|\log f_{\hat{G}_n,\tilde{\nu}_i}(\tilde{Z}_i)|$ grows like $\log(n)$. As a result, we can apply Lemma SM10.3 to show that posterior means of $((Z-\tau)/\nu)^{k+1}$ is bounded by $\log^{(k+1)/2}(n)$. Thus, bounding $|\tau|$ by $M_n$ since $\hat{G}_n$ has finite support as in Lemma SM6.13, we have that

$$\max\left(|U_{km}(\tilde{m},\tilde{s},Z)|, |U_{ks}(\tilde{m},\tilde{s},Z)|\right) \lesssim_{\mathcal{H},v} M_n \log^{(k+1)/2}(n).$$

Putting these calculations together yields that

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathbb{1}(A_n)}{n}\|\hat{\tau}_{v,\eta_0} - \hat{\tau}_v\|^2\right] &= \frac{\mathbb{1}(A_n)}{n}\sum_{i=1}^n \mathbb{E}\left[\left(\mathbf{E}_{\hat{G}_n,\hat{\eta}}[\tau^v \mid \hat{Z},\hat{\nu}] - \mathbf{E}_{\hat{G}_n,\eta_0}[\tau^v \mid Z,\nu]\right)^2\right] \\
&\lesssim_{\mathcal{H},v} \Delta_n^2 M_n^2 (\log n)^{v+1}.
\end{aligned}$$

*SM10.1.2 Bounding* $\|\hat{\tau}_{v,\eta_0,\rho_n} - \tau^*_{v,\rho_n}\|$. Define

$$\hat{\tau}_{i,v,\eta_0,\rho_n} = \hat{\tau}_{i,v,\eta_0}\frac{f_{\hat{G}_n,\nu_i}(Z_i)}{f_{\hat{G}_n,\nu_i}(Z_i) \vee (\rho_n/\nu_i)}. \qquad \text{(SM10.5)}$$

Note that for the choice of $\rho_n$ in (OA3.5), by Lemma OA3.1, the truncation doesn't bind and $\hat{\tau}_{i,v,\eta_0} = \hat{\tau}_{i,v,\eta_0,\rho_n}$. Analogously, define the truncated oracle posterior means $\tau^*_{v,\rho_n}$. We next bound

$$\xi_{3v} = \frac{\mathbb{1}(A_n)}{n}\|\tau^*_{v,\rho_n} - \hat{\tau}_{v,\rho_n}\|^2.$$

As in Appendix OA3.2, we decompose $\xi_{3v}$ into the following terms

$$\zeta_{1v}^2 = \mathbb{1}(A_n \cap B_n^{\mathrm{C}})\|\tau^*_{v,\rho_n} - \hat{\tau}_{v,\rho_n}\|^2 \tag{SM10.6}$$

$$\zeta_{2v}^2 = \mathbb{1}(A_n \cap B_n)\left(\|\tau^*_{v,\rho_n} - \hat{\tau}_{v,\rho_n}\| - \max_{j\in[N]}\|\tau^*_{v,\rho_n} - \tau^{(j)}_{v,\rho_n}\|\right)_+^2 \tag{SM10.7}$$

$$\zeta_{3v}^2 = \max_{j\in[N]}\left(\|\tau^*_{v,\rho_n} - \tau^{(j)}_{v,\rho_n}\|^2 - \mathbb{E}\left[\|\tau^*_{v,\rho_n} - \tau^{(j)}_{v,\rho_n}\|\right]\right)_+^2 \tag{SM10.8}$$

$$\zeta_{4v}^2 = \max_{j\in[N]}\left(\mathbb{E}[\|\tau^*_{v,\rho_n} - \tau^{(j)}_{v,\rho_n}\|]\right)^2. \tag{SM10.9}$$

where:

- $B_n = \left\{\overline{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}) < B_{\mathcal{H}}\delta_n\right\}$ for the rate function $\delta_n$ in (OA3.6) and the constant $B_{\mathcal{H}}$ is chosen by Corollary OA3.1
- $\tau^{(j)}_{i,v,\rho_n} = \mathbf{E}_{G_j,\nu}[\tau^v \mid Z_i, \nu_i]$ for some finite set of prior distributions $G_1, \ldots, G_N$ to be chosen.

The following calculation shows that

$$\frac{\mathbb{1}(A_n)}{n}\|\tau^*_{v,\rho_n} - \hat{\tau}_{v,\rho_n}\|^2 \lesssim_{v,\mathcal{H}} \delta_n^2(\log n)^{\frac{2(v-1)}{\alpha}\vee(3v)}. \tag{SM10.10}$$

For $\zeta_{1v}$, by Lemma SM10.2 and Corollary OA3.1,

$$\frac{1}{n}\mathbb{E}\zeta_{1v}^2 \lesssim_{\mathcal{H},v} (\log n)^{v/2}\,\mathrm{P}\left(A_n \cap B_n^{\mathrm{C}}\right) \lesssim (\log n)^{v/2}\frac{\log\log n}{n}.$$

For $\zeta_{2v}, \zeta_{3v}$, let $G_1, \ldots, G_N$ be minimal $\omega$-covering of $\left\{G : \overline{h}(f_{G,\cdot}, f_{G_0,\cdot}) \le \delta_n\right\}$ under the pseudometric

$$\begin{aligned} d_{v,M_n,\rho_n}(H_1,H_2) = \max_{i\in[n]}\max_{0\le k\le v}\sup_{|z|\le M_n}\Bigg|&\mathbf{E}_{H_1}\left[\left(\frac{\tau - Z_i}{\nu_i}\right)^k \mid Z_i = z, \nu_i\right]\frac{f_{H_1,\nu_i}(z)}{f_{H_1,\nu_i}(z)\vee(\rho_n/\nu_i)} \\ &- \mathbf{E}_{H_2}\left[\left(\frac{\tau - Z_i}{\nu_i}\right)^k \mid Z_i = z, \nu_i\right]\frac{f_{H_2,\nu_i}(z)}{f_{H_2,\nu_i}(z)\vee(\rho_n/\nu_i)}\Bigg| \end{aligned} \tag{SM10.11}$$

such that

$$N \le N(\omega/2, \mathcal{P}(\mathbb{R}), d_{v,M_n,\rho_n}).$$

Under this choice, note that by (SM10.4),

$$\begin{aligned} |\tau^*_{i,v,\rho_n} - \hat{\tau}_{i,v,\rho_n}| &\le \sum_{k=1}^{v} C_{k,v}|Z_i|^{v-k}\nu_i^k\Bigg|\mathbf{E}_{\hat{G}_n,\eta_0}\left[\left(\frac{\tau - Z_i}{\nu_i}\right)^k \mid Z_i, \nu_i\right]\frac{f_{\hat{G}_n,\nu_i}}{f_{\hat{G}_n,\nu_i}\vee(\rho_n/\nu_i)} \\ &\qquad - \mathbf{E}_{G_0,\eta_0}\left[\left(\frac{\tau - Z_i}{\nu_i}\right)^k \mid Z_i, \nu_i\right]\frac{f_{G_0,\nu_i}}{f_{G_0,\nu_i}\vee(\rho_n/\nu_i)}\Bigg| \\ &\lesssim_{v,\mathcal{H}} |Z_i|^v\omega \end{aligned}$$

Therefore,

$$\frac{1}{n}\mathbb{E}[\zeta_{2v}^2] \lesssim_{v,\mathcal{H}} \mathbb{E}|Z_i|^v\omega^2 \lesssim_{v,\mathcal{H}} \omega^2.$$

Analogous to Appendix OA3.2, under appropriate choices of $\omega$, we find that for some positive constant $C_v$

$$\frac{1}{n}\mathbb{E}\left[\zeta_{2v}^2 + \zeta_{3v}^2\right] \lesssim_{\mathcal{H},v} \frac{M_n(\log n)^{C_v}}{n}.$$

Here, the covering number $N(\omega/2, \mathcal{P}(\mathbb{R}), d_{v,M_n,\rho_n})$ is bounded by an application of Proposition SM6.1 analogous to Proposition SM6.2. The term $\mathbb{E}[\zeta_{3v}^2]$ is controlled analogously to Appendix OA3.2.3, where $K_n$ in Appendix OA3.2.3 is now of the order $(\log(n))^{v/2}$ thanks to Lemma SM10.3.

Finally, for $\zeta_{4v}^2$, by Jensen's inequality

$$\left(\mathbb{E}[\|\tau_{v,\rho_n}^* - \tau_{v,\rho_n}^{(j)}\|]\right)^2 \le \mathbb{E}[\|\tau_{v,\rho_n}^* - \tau_{v,\rho_n}^{(j)}\|^2].$$

Following (SM10.4)

$$\left(\tau_{i,v,\rho_n}^* - \tau_{i,v,\rho_n}^{(j)}\right)^2 \lesssim_{v,\mathcal{H}} \sum_{k=1}^{v}\int_{-\infty}^{\infty} z^{2(v-k)}\Delta_k^2(z;\nu_i,\rho_n,G_j) f_{G_0,\nu_i}(z)\,dz$$

where

$$\Delta_k(z;\nu_i,\rho_n,G_j) = \mathbf{E}_{G_j,\eta_0}\left[\left(\frac{\tau - Z_i}{\nu_i}\right)^k \mid Z_i = z, \nu_i\right]\frac{f_{\hat{G}_n,\nu_i}}{f_{\hat{G}_n,\nu_i} \vee (\rho_n/\nu_i)}$$
$$- \mathbf{E}_{G_0,\eta_0}\left[\left(\frac{\tau - Z_i}{\nu_i}\right)^k \mid Z_i = z, \nu_i\right]\frac{f_{G_0,\nu_i}}{f_{G_0,\nu_i} \vee (\rho_n/\nu_i)}.$$

We decompose

$$\begin{aligned}
&\int_{-\infty}^{\infty} z^{2(v-k)}\Delta_k^2(z;\nu_i,\rho_n,G_j) f_{G_0,\nu_i}(z)\,dz \\
&\le \int_{|z|>M_n} z^{2(v-k)}\Delta_k^2(z;\nu_i,\rho_n,G_j) f_{G_0,\nu_i}(z)\,dz + M_n^{2(v-k)}\int_{-\infty}^{\infty}\Delta_k^2(z;\nu_i,\rho_n,G_j) f_{G_0,\nu_i}(z)\,dz \\
&\lesssim_{v,\mathcal{H}} \log^k(n)\int_{|z|>M_n} z^{2(v-k)} f_{G_0,\nu_i}(z)\,dz + M_n^{2(v-k)}\int_{-\infty}^{\infty}\Delta_k^2(z;\nu_i,\rho_n,G_j) f_{G_0,\nu_i}(z)\,dz \\
&\hfill \text{(Lemma SM10.3)} \\
&\lesssim_{v,\mathcal{H}} \log^k(n)\sqrt{\mathbb{E}[Z_i^{4(v-k)}]\,\mathrm{P}[|Z_i| > M_n]} + M_n^{2(v-k)}\int_{-\infty}^{\infty}\Delta_k^2(z;\nu_i,\rho_n,G_j) f_{G_0,\nu_i}(z)\,dz \\
&\lesssim_{v,\mathcal{H}} \frac{\log^k(n)}{n} + M_n^{2(v-k)}\int_{-\infty}^{\infty}\Delta_k^2(z;\nu_i,\rho_n,G_j) f_{G_0,\nu_i}(z)\,dz \qquad \text{(Lemma OA3.7)}
\end{aligned}$$

Let $G_{j,\nu}, G_{0,\nu}$ be the distribution of $\mu = \tau/\nu$ when $\tau \sim G$, respectively. Let $X = Z/\nu$ be such that $X \sim \mathcal{N}(\tau/\nu, 1)$. Let $f_{G_{0,\nu}}, f_{G_{j,\nu}}$ be the density of $X$ and note that $f_X(x) = \nu f_Z(z)$. Note that Hellinger distance is invariant to this reparametrization

$$h^2(f_{G_{0,\nu}}, f_{G_{j,\nu}}) = h^2(f_{G_j,\nu}, f_{G_0,\nu}).$$

and

$$
\begin{aligned}
&\int_{-\infty}^{\infty} \Delta_k^2(z;\nu_i,\rho_n,G_j) f_{G_0,\nu_i}(z)\,dz \\
&= \int_{-\infty}^{\infty} \left( \mathbb{E}_{G_{j,\nu}}[(X-\mu)^k \mid X_i] \frac{f_{G_{j,\nu}}}{f_{G_{j,\nu}} \vee \rho_n} - \mathbb{E}_{G_{0,\nu}}[(X-\mu)^k \mid X_i] \frac{f_{G_{0,\nu}}}{f_{G_{0,\nu}} \vee \rho_n} \right)^2 f_{G_{0,\nu}}(x)\,dx
\end{aligned} \tag{SM10.12}
$$

Let $f_G(x) = f_{G,1}(x)$. We note that, by repeated differentiation,

$$
\frac{f_G^{(k)}(z)}{f_G(z)} = \sum_{\ell=0}^{k} C_{\ell,k} \mathbf{E}_G \left[ (X-\mu)^k \mid X = x \right] \tag{SM10.13}
$$

where $C_{k,k} > 0$. As a result, we can write, for some different constants $C$,

$$
\mathbf{E}_G[(X-\mu)^k \mid X = x] = \sum_{\ell=0}^{k} C_{\ell,k} \frac{f_G^{(\ell)}(x)}{f_G(x)}.
$$

Therefore, to bound (SM10.12), it suffices to bound, for $k \le v$ a positive integer, $G = G_j$ and $\rho = \rho_n$,

$$
\int \left( \frac{f_G^{(k)}(x)}{f_G(x) \vee \rho} - \frac{f_{G_0}^{(k)}(x)}{f_{G_0}(x) \vee \rho} \right)^2 f_{G_0}(x)\,dx. \tag{SM10.14}
$$

Our key result in Theorem SM10.2 yields that

$$
\left( \tau^*_{i,v,\rho_n} - \tau^{(j)}_{i,v,\rho_n} \right)^2 \lesssim_{v,\mathcal{H}} \sum_{1 \le k \le v} M_n^{2(v-k)} \max\left( \log^{3k}(n), |\log h_i^2|^k \right) h_i^2
$$

where $h_i^2 = h^2(f_{G_j,\nu_i}, f_{G_0,\nu_i})$. Note that if $|\log h_i^2| = \log(1/h_i^2) \gtrsim \log^3(n)$, then $h_i \lesssim e^{-C\log^3(n)/2} \lesssim 1/n$. In this case, $|\log h_i^2|^k h_i^2 \lesssim \frac{1}{n} h_i |2\log(1/h_i)|^k \lesssim_k \frac{1}{n}$. Thus,

$$
\begin{aligned}
\frac{1}{n} \|\tau^*_{v,\rho_n} - \tau^{(j)}_{v,\rho_n}\|^2 &\lesssim_{v,\mathcal{H}} \frac{\log^v(n)}{n} + \sum_{1 \le k \le v} M_n^{2(v-k)} \max\left( \log^{3k}(n) \overline{h}^2, \frac{1}{n} \right) \\
&\lesssim_{v,\mathcal{H}} \delta_n^2 \left[ \max_{1 \le k \le v} M_n^{2(v-k)} \log^{3k}(n) \right] \\
&\lesssim_{v,\mathcal{H}} \delta_n^2 (\log n)^{\frac{2(v-1)}{\alpha} \vee (3v)}.
\end{aligned}
$$

*SM10.1.3 Bounding* $\|\tau^*_{v,\rho_n} - \tau^*_v\|^2$. By an argument analogous to Lemma OA3.4, we have that

$$
\frac{1}{n} \mathbb{E}\left[ \|\tau^*_{v,\rho_n} - \tau^*_v\|^2 \right] \lesssim_{\mathcal{H}} \frac{1}{n}.
$$

**SM10.2 Regret of misspecified Bayes rules for higher moments.** The key ingredient to our argument in this section is a bound for (SM10.14). We rely on the argument in Theorem 3 of Jiang and Zhang (2009) (Lemma OA3.9) and generalize it to higher derivatives.

**Theorem SM10.2.** *For an integer $m \ge 1$ and $\rho < e^{-1}/\sqrt{2\pi}$,*

$$
\int \left( \frac{f_G^{(m)}}{f_G \vee \rho} - \frac{f_{G_0}^{(m)}}{f_{G_0} \vee \rho} \right)^2 f_{G_0}\,dx \lesssim_m \max\left[ \log^{3m}\left( \frac{1}{\sqrt{2\pi}\rho} \right), |\log h^2(f_G, f_{G_0})|^m \right] \cdot h^2\left( f_G, f_{G_0} \right). \tag{SM10.15}
$$

*Proof.* Let $w_* = \frac{1}{f_{G_0}\vee\rho + f_G\vee\rho}$. We can add and subtract $2(f_G^{(m)} - f_{G_0}^{m})w_*$ to the integrand. This means that

$$\begin{aligned}
&\int \left(\frac{f_G^{(m)}}{f_G\vee\rho} - \frac{f_{G_0}^{(m)}}{f_{G_0}\vee\rho}\right)^2 f_{G_0}\,dx \\
&\lesssim \left\|\frac{f_G^{(m)}}{f_G\vee\rho} - 2f_G^{(m)}w_*\right\|_{f_{G_0}}^2 + \left\|\frac{f_{G_0}^{(m)}}{f_{G_0}\vee\rho} - 2f_{G_0}^{(m)}w_*\right\|_{f_{G_0}}^2 + \left\|(f_G^{(m)} - f_{G_0}^{(m)})w_*\right\|_{f_{G_0}}^2, \qquad \text{(SM10.16)}
\end{aligned}$$

where $\|g\|_f^2 = \int g^2 f\,dx$.

Note that

$$\begin{aligned}
\left\|\frac{f_G^{(m)}}{f_G\vee\rho} - 2f_G^{(m)}w_*\right\|_{f_{G_0}}^2 &= \left\|\frac{f_G^{(m)}}{f_G\vee\rho + f_{G_0}\vee\rho}\frac{f_G\vee\rho + f_{G_0}\vee\rho}{f_G\vee\rho} - 2f_G^{(m)}w_*\right\|_{f_{G_0}}^2 \\
&= \left\|\left(\frac{f_G\vee\rho + f_{G_0}\vee\rho}{f_G\vee\rho} - 2\right) f_G^{(m)}w_*\right\|_{f_{G_0}}^2 \\
&= \left\|\frac{f_G^{(m)}}{f_G\vee\rho}(f_G\vee\rho - f_{G_0}\vee\rho)w_*\right\|_{f_{G_0}}^2. \\
&\lesssim_m \log^m\left(\frac{1}{\sqrt{2\pi}\rho}\right)\left\|(f_G - f_{G_0})w_*\right\|_{f_{G_0}}^2 \qquad \text{(Lemma SM10.3)}
\end{aligned}$$

Now,

$$\begin{aligned}
\left\|(f_G - f_{G_0})w_*\right\|_{f_{G_0}}^2 &= \int (\sqrt{f_G} - \sqrt{f_{G_0}})^2 \underbrace{[(\sqrt{f_G} + \sqrt{f_{G_0}})^2 w_*]}_{\le 2}\underbrace{w_* f_{G_0}}_{\le 1}\,dx \\
&\lesssim h^2(f_G, f_{G_0})
\end{aligned}$$

Thus,

$$\begin{aligned}
\text{(SM10.16)} &\lesssim_m h^2(f_G, f_{G_0})\log^{m/2}\left(\frac{1}{\sqrt{2\pi}\rho}\right) + \left\|(f_G^{(m)} - f_{G_0}^{(m)})w_*\right\|_{f_{G_0}}^2 \\
&\lesssim_m h^2(f_G, f_{G_0})\log^{m/2}\left(\frac{1}{\sqrt{2\pi}\rho}\right) + \left\|(f_G^{(m)} - f_{G_0}^{(m)})\right\|_{w^*}^2 \qquad (f_{G_0}w_* \le 1)
\end{aligned}$$

Define

$$\Delta_m^2 = \left\|(f_G^{(m)} - f_{G_0}^{(m)})\right\|_{w^*}^2 = \int \frac{(f_G^{(m)} - f_{G_0}^{(m)})^2}{f_G\vee\rho + f_{G_0}\vee\rho}\,dx. \qquad \text{(SM10.17)}$$

This is bounded by Proposition SM10.1 below. Plugging in the result from Proposition SM10.1, we obtain (SM10.15). ☐

**Proposition SM10.1.** *For all $m \ge 0$ and $\rho \in (0, 1/\sqrt{2\pi})$, in the proof of Theorem SM10.2,*

$$\Delta_m^2 \lesssim_m \max\left[\log^{3m}\left(\frac{1}{\sqrt{2\pi}\rho}\right), |\log h^2(f_G, f_{G_0})|^m\right]\cdot h^2\left(f_G, f_{G_0}\right).$$

*Proof.* We prove this by induction. See the proof of Lemma 1 in Jiang and Zhang for the base case ($m = 0$). The inductive step is immediate with Lemma SM10.1, where we note that

$$\Delta_{m-1}^{c_0/(c_0+1)} \lesssim_m \max\left[\log^{3(m-1)}\left(\frac{1}{\sqrt{2\pi}\rho}\right), |\log h^2(f_G, f_{G_0})|^{m-1}\right](h^2)^{\frac{c_0}{c_0+1}}. \qquad \square$$

**Lemma SM10.1.** *Define* $\sqrt{\log\left(\frac{1}{\sqrt{2\pi}\rho}\right)} = L(\rho)$*. For all positive integers* $m$*, there exists some* $c_0 > 0$*,*

$$\Delta_m^2 \lesssim_m \max\left(\max(L^2(\rho), |\log h^2|)(\Delta_{m-1}^2)^{\frac{c_0}{c_0+1}}(h^2)^{1/(c_0+1)}, L(\rho)^6\Delta_{m-1}^2\right)$$

*Proof.* The proof to Lemma 1 in Jiang and Zhang prove the following for $\rho < 1/\sqrt{2\pi}$:

(1) We have

$$\int (f_G^{(m)} - f_{G_0}^{(m)})^2\,dx \le \frac{4a^{2m}}{\sqrt{2\pi}}h^2(f_G, f_{G_0}) + \frac{4a^{2m-1}}{\pi}e^{-a^2}.$$

for integers $m \ge 0$ and $a \ge \sqrt{2m-1}$.

(2) By integration by parts and Cauchy–Schwarz,

$$\Delta_k^2 \le \Delta_{k-1}\Delta_{k+1} + 2L(\rho)\Delta_{k-1}\Delta_k, \tag{SM10.18}$$

and hence

$$\frac{\Delta_k}{\Delta_{k-1}} \le \left(\frac{\Delta_{k+1}}{\Delta_k} + 2L(\rho)\right).$$

(3) $\Delta_0^2 \le 2h^2(f_G, f_{G_0})$

(4) We have

$$\Delta_k^2 \le \frac{1}{2\rho}\int (f_G^{(m)} - f_{G_0}^{(m)})^2\,dx.$$

First, let us consider the case that, for some $Q$ to be chosen,

$$\Delta_{k+1}/\Delta_k > Q \implies \Delta_{k+1} > Q\Delta_k$$

Then (SM10.18) implies that

$$\Delta_k^2 \le \Delta_{k-1}\Delta_{k+1}(1 + 2L(\rho)/Q) \implies \frac{\Delta_k}{\Delta_{k-1}} \le \frac{\Delta_{k+1}}{\Delta_k}(1 + 2L(\rho)/Q).$$

On the other hand, if

$$\Delta_{k+1}/\Delta_k \le Q \implies \frac{\Delta_k}{\Delta_{k-1}} \le [Q + 2L(\rho)]$$

Fix any integer $m$, let $k(Q) \ge m$ be the first $k$ exceeding $m$ (if it exists) where $\Delta_{k+1}/\Delta_k \le Q$ occurs. On the other hand, fix some $k_0 \ge m$.

**Case 1:** If $k(Q) > k_0$, then for all $m \le k \le k_0 + 1$,

$$\frac{\Delta_m}{\Delta_{m-1}} \le \left(1 + \frac{2L(\rho)}{Q}\right)^{k-m}\frac{\Delta_k}{\Delta_{k-1}}.$$

Thus,

$$\frac{\Delta_m}{\Delta_{m-1}} \le \left(\prod_{k=m}^{k_0+1}\left(1 + \frac{2L(\rho)}{Q}\right)^{k-m}\frac{\Delta_k}{\Delta_{k-1}}\right)^{1/(k_0-m+2)}$$

$$\leq \left(1+\frac{2L(\rho)}{Q}\right)^{(k_0-m+1)/2} (\Delta_{k_0+1}/\Delta_{m-1})^{1/(k_0-m+2)}.$$

Let $c_0 = k_0 - m + 1$. We then have that

$$\Delta_m^2 \leq \left(1+\frac{2L(\rho)}{Q}\right)^{c_0} (\Delta_{m-1}^2)^{\frac{c_0}{c_0+1}} (\Delta_{k_0+1}^2)^{\frac{1}{c_0+1}}.$$

$$\Delta_{k_0+1}^2 \lesssim \frac{a^{2k_0+2}}{\sqrt{2\pi}\rho}\left(h^2(f_G, f_{G_0}) + e^{-a^2}/a\right) \qquad (a \geq \sqrt{2k_0+1})$$

$$\lesssim \frac{a^{2k_0+2}}{\sqrt{2\pi}\rho} h^2(f_G, f_{G_0}) \qquad (a \geq \max\left(\sqrt{2k_0+1}, \sqrt{|\log h^2|}\right))$$

Thus,

$$\Delta_m^2 \leq \left(1+\frac{2L(\rho)}{Q}\right)^{c_0} (\Delta_{m-1}^2)^{\frac{c_0}{c_0+1}} a^{2\left(1+\frac{m}{c_0+1}\right)} \left(\frac{1}{\sqrt{2\pi}\rho}\right)^{\frac{1}{c_0+1}} (h^2)^{1/(c_0+1)}.$$

**Case 2:** On the other hand, if $k(Q) \leq k_0$, then

$$\Delta_m \leq \left(1+\frac{2L(\rho)}{Q}\right)^{k(Q)+1-m} [Q + 2L(\rho)]\Delta_{m-1}.$$

Consider the following choices:

(1) $c_0 = \lceil (m-1) \vee (L^2(\rho)-1) \rceil \geq (m-1) \vee (L^2(\rho)-1)$

(2) $Q = 2L(\rho)c_0$

(3) $a = \max\left(\sqrt{2(m+c_0)+1}, \sqrt{|\log h^2|}\right)$.

Therefore, if $k(Q) > k_0$, then

$$\Delta_m^2 \lesssim_m \max(L^2(\rho), |\log h^2|)(\Delta_{m-1}^2)^{\frac{c_0}{c_0+1}} (h^2)^{1/(c_0+1)}$$

Otherwise,

$$\Delta_m^2 \lesssim_m L(\rho)^6 \Delta_{m-1}^2.$$

Taking the maximum of these two yields the bound in the statement of the lemma. □

### SM10.3 Auxiliary lemmas.

**Lemma SM10.2.** *Under the assumptions in Theorem SM10.1, in the proof of Theorem SM10.1,* $\frac{\mathbb{1}(A_n)}{n}\|\tau_{v,\rho_n}^* - \hat{\tau}_{v,\rho_n}\|^2 \lesssim_{\mathcal{H},m} (\log n)^v$.

*Proof.* By (SM10.4), we have that

$$|\tau_{i,v,\rho_n}^* - \hat{\tau}_{i,v,\rho_n}| \leq \sum_{k=1}^{v} C_{k,v}|Z_i|^{v-k}\nu_i^k \Bigg| \mathbf{E}_{\hat{G}_n,\eta_0}\left[\left(\frac{\tau - Z_i}{\nu_i}\right)^k \mid Z_i, \nu_i\right] \frac{f_{\hat{G}_n,\nu_i}}{f_{\hat{G}_n,\nu_i} \vee (\rho_n/\nu_i)} - \mathbf{E}_{G_0,\eta_0}\left[\left(\frac{\tau - Z_i}{\nu_i}\right)^k \mid Z_i, \nu_i\right] \frac{f_{G_0,\nu_i}}{f_{G_0,\nu_i} \vee (\rho_n/\nu_i)}\Bigg|$$

$$\lesssim_{v,\mathcal{H}} \log^{v/2}(n)|Z_i|^v \qquad \text{(Lemma SM10.3)}$$

Thus,

$$\frac{\mathbb{1}(A_n)}{n}\|\tau_{v,\rho_n}^* - \hat{\tau}_{v,\rho_n}\|^2 \lesssim_{v,\mathcal{H}} \log^v(n)\frac{1}{n}\sum_{i=1}^{n}|Z_i|^{2v} \lesssim_{v,\mathcal{H}} \log^v(n).$$

This completes the proof. $\square$

**Lemma SM10.3.** *Let $Z \mid \tau, \nu \sim \mathcal{N}(\tau, \nu^2)$ and let $k$ be a positive integer. For all $\rho < \frac{1}{e\sqrt{2\pi}}$ Then*

$$\left| \mathbf{E}_G \left[ \left( \frac{Z-\tau}{\nu} \right)^k \mid Z = z, \nu \right] \frac{f_{G,\nu}(z)}{f_{G,\nu}(z) \vee (\rho/\nu)} \right| \lesssim_k \log^{k/2} \left( \frac{1}{\sqrt{2\pi}\rho} \right).$$

*Proof.* Now, by Jensen's inequality and Lemma SM10.4,

$$\left| \mathbf{E}_G \left[ \left( \frac{Z-\tau}{\nu} \right)^k \mid Z = z, \nu \right] \right| \lesssim_k \max \left( \log^{k/2} \left( \frac{1}{\sqrt{2\pi}\nu f_{G,\nu}(z)} \right), 1 \right)$$

Observe that the function

$$t \mapsto \log^{k/2} \left( \frac{1}{\sqrt{2\pi} t} \right) \frac{t}{t \vee \rho} \quad t \in (0, 1/\sqrt{2\pi})$$

is decreasing in $t$ for $t > \rho$ and increasing in $t$ for $t \leq \rho$. Thus, it attains a maximum at $\log^{k/2} \frac{1}{\sqrt{2\pi}\rho} > 1$. This completes the proof. $\square$

**Lemma SM10.4.** *Let $Z \mid \tau, \nu \sim \mathcal{N}(\tau, \nu^2)$ and let $k$ be a positive integer. Then*

$$\mathbf{E}_G \left[ \left( \frac{Z-\tau}{\nu} \right)^{2k} \mid Z, \nu \right] \leq e \left( \frac{2k}{e} \right)^k \max \left( \log^k \left( \frac{1}{\sqrt{2\pi}\nu f_{G,\nu}(z)} \right), 1 \right).$$

*Proof.* Let $W(\tau) = \frac{Z-\tau}{\nu} \sim \mathcal{N}(0,1)$. Note that for $t \in (0, 1]$,

$$\begin{aligned}
\mathbf{E} \left[ e^{tW^2/2} \mid Z, \nu \right] &= \frac{1}{\sqrt{2\pi}\nu f_{G,\nu}(z)} \int \exp \left( -\frac{1}{2}(1-t) W(\tau) \right) G(d\tau) \\
&\leq \frac{1}{\sqrt{2\pi}\nu f_{G,\nu}(z)} \int \left( \exp \left( -\frac{1}{2} W(\tau) \right) \right)^{1-t} G(d\tau) \\
&\leq \frac{1}{\sqrt{2\pi}\nu f_{G,\nu}(z)} (\sqrt{2\pi}\nu)^{1-t} \left\{ \int \frac{1}{\sqrt{2\pi}\nu} \exp \left( -\frac{1}{2} W(\tau) \right) G(d\tau) \right\}^{1-t} \\
&\qquad \text{(Jensen's inequality)} \\
&= \left( \sqrt{2\pi} f_{G,\nu}(z) \right)^{-t}
\end{aligned}$$

The left-hand side is the moment-generating function of $W^2(\tau)/2$ under the law $\tau \mid Z, \nu$ induced by $\tau \sim G$.

Thus, by Lemma SM10.5,

$$\mathbb{E}[W^{2k}] \leq \left( \frac{2k}{te} \right)^k \left( \sqrt{2\pi} f_{G,\nu}(z) \right)^{-t}$$

Optimize this bound with

$$t = \frac{1}{1 \vee \log \left( \frac{1}{\sqrt{2\pi}\nu f_{G,\nu}(z)} \right)} \in (0, 1],$$

then the above yields

$$\mathbf{E}_G \left[ \left( \frac{Z-\tau}{\nu} \right)^{2k} \mid Z, \nu \right] \leq e \left( \frac{2k}{e} \right)^k \max \left( \log^k \left( \frac{1}{\sqrt{2\pi}\nu f_{G,\nu}(z)} \right), 1 \right).$$

This completes the proof. $\square$

**Lemma SM10.5.** *For a nonnegative random variable $X$ whose moment-generating function exists, for any $k$ and $t > 0$ such that $\mathbb{E}[e^{tX}]$ exists,*

$$\mathbb{E}[X^k] \leq \left(\frac{k}{te}\right)^k \mathbb{E}[e^{tX}]$$

*Proof.* This is immediate from the observation that $x^m \leq \left(\frac{m}{te}\right)^m e^{tx}$ for all $x \geq 0, t > 0$. This inequality is included on the Wikipedia page on moment-generating functions (Section "Other properties", accessed 2024-04-15). $\square$

## References


Calonico, S., M. D. Cattaneo, and M. H. Farrell (2019). "nprobust: Nonparametric kernel-based estimation and robust bias-corrected inference". In: *arXiv preprint arXiv:1906.00198* (pp. 109, 110).

Jiang, W. (2020). "On general maximum likelihood empirical Bayes estimation of heteroscedastic IID normal means". In: *Electronic Journal of Statistics* 14.1, pp. 2272–2297 (pp. 82, 93, 96, 98, 99, 104, 108).

Jiang, W. and C.-H. Zhang (2009). "General maximum likelihood empirical Bayes estimation of normal means". In: *The Annals of Statistics* 37.4, pp. 1647–1684 (pp. 121, 126).

Kubokawa, T., C. Robert, and A. M. E. Saleh (1993). "Estimation of noncentrality parameters". In: *Canadian Journal of Statistics* 21.1, pp. 45–57 (pp. 110, 111).

Sen, B. (2018). "A gentle introduction to empirical process theory and applications". In: *Lecture Notes, Columbia University* 11, pp. 28–29 (pp. 112, 119).

Soloff, J. A., A. Guntuboyina, and B. Sen (2021). "Multivariate, heteroscedastic empirical bayes via nonparametric maximum likelihood". In: *arXiv preprint arXiv:2109.03466* (pp. 104, 121).

Tsybakov, A. (2008). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York. ISBN: 9780387790527. URL: https://books.google.com/books?id=mwB8rUBsbqoC (pp. 112, 114, 116, 120).

van der Vaart, A. W. and J. A. Wellner (1996). *Weak convergence*. Springer, pp. 16–28 (pp. 88, 117, 119).

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press (pp. 82, 87, 88, 91, 101).

Zhang, C.-H. (1997). "Empirical Bayes and compound estimation of normal means". In: *Statistica Sinica* 7.1, pp. 181–193 (pp. 100).