HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Economics

have examined a dissertation entitled

"Essays in Econometrics"

presented by **Jiafeng Chen**

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature ______________________

Typed name: Prof. Isaiah Andrews

Signature ______________________

Typed name: Prof. Edward Glaeser

Signature ______________________

Typed name: Prof. Elie Tamer

Signature Jesse Shapiro

Typed name: Prof. Jesse Shapiro

Date: March 4, 2024

Essays in Econometrics

A dissertation presented

by

Jiafeng Chen

to

The Department of Economics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Business Economics

Harvard University

Cambridge, Massachusetts

March 2024

Dissertation Advisor:
Professor Isaiah Andrews

Author:
Jiafeng Chen

Essays in Econometrics

Abstract

This dissertation consists of three essays in econometrics. A common throughline is decision theory, defined here broadly as the formal considerations justifying, informing, or undermining choices of statistical procedures.

The first chapter proposes a new interpretation of synthetic control methods as instances of online convex optimization algorithms. Viewed in a certain way, synthetic control methods implement an algorithm called *Follow-The-Leader*. Mathematical guarantees of Follow-The-Leader then translate into new guarantees for synthetic control. Specifically, over long time horizons, synthetic control methods predict almost as well as the best weighted average of untreated units chosen ex post.

The second chapter proposes new empirical Bayes methods that improve statistical decision-making. It shows that conventional empirical Bayes methods embed an assumption called *prior independence*; this assumption frequently fails to hold; and imposing this assumption incorrectly can harm the performance of standard empirical Bayes methods. Motivated by these observations, the chapter proposes new empirical Bayes methods and proves some new decision-theoretic guarantees.

The third chapter is a paper co-authored with Jonathan Roth. Empirical researchers frequently want to estimate some causal effect in terms of the log transformation of their outcome variables. However, when the outcome variable can take the value zero, its log is not well-defined. In such situations, empirical researchers often resort to certain "logarithm-like" transformations that are defined at zero and continue to interpret results as approximate log or percentage effects. We show that such interpretations are inappropriate. We also show that there exists no estimand satisfying certain desirable properties simultaneously, and one has to forgo at least one of these properties.

Contents

List of Tables

List of Figures

Acknowledgments

I am deeply indebted to my advisor, Isaiah Andrews, for his continued support and guidance that always go above and beyond. I am also profoundly grateful to my committee members—Elie Tamer, Jesse Shapiro, and Edward Glaeser—for their generosity and kindness. In some form or other, I have also learned greatly and benefited enormously from the following wonderful individuals: Susan Athey, the late Gary Chamberlain, Xiaohong Chen, Raj Chetty, Bryan Graham, Guido Imbens, Pat Kline, Scott Duke Kominers, Greg Lewis, Neil Shephard, Jann Spiess, and Jonathan Roth.

Graduate school would be extremely lonely if not for fellow graduate students and friends—among them, Ross Mattheis, Suproteem Sarkar, David Ritzwoller, Brad Ross, Chris Walker, Jennifer Walsh, and Lingxuan Wu. Talking to Suproteem, David, Brad, Chris, and Lingxuan kept me sane during the worst throes of the pandemic in late 2020 and early 2021. I have also learned an absurd amount of econometrics and statistics just from hashing things out with Brad and David and from organizing reading groups[1] with Chris.

Words fail to convey my immense gratitude towards my parents. They supported and loved me so generously, so unconditionally, and so firmly to the point of relentlessness. I love them with all my heart. It breaks my heart that I was an ocean and a continent away from them during graduate school, and was only able to visit home once after pandemic restrictions loosened in China.

Lastly, to my wife Lotus, the year that we were separated by the pandemic was probably the worst year of my life. That year made me realize two things. First, our cat heidi—whom you insisted on adopting—probably kept me from going insane. Second, and more importantly, you're the love of my life, the apple of my eye, and the one person I cannot live without. Thank you—none of this is possible without your support and sacrifice, and none of this would have meant anything to me without having you in my life.

[1] Let's face it, it's more like reading *pairs* than reading groups.

To my parents

and

To the memory of my grandfather, who passed away in Shanghai in 2023 due to COVID-19

When I got started in comedy in 1985, someone asked me what I dreamed about, and I said I dreamed about having a body of work. People can like what I do or not like what I do, but Jesus, there's a body of work there, and I'm really proud of a lot of it.

Conan O'Brien, *Vulture* interview, June 23, 2021

Chapter 1

Synthetic control as online linear regression[1]

[1] Jiafeng Chen, "Synthetic control as online linear regression", *Econometrica*, 2023, Volume 91, Issue 2, pages 465–491. The copyright to this article is held by the Econometric Society, http://www.econometricsociety.org/. It may be downloaded, printed and reproduced only for personal or classroom use. Absolutely no downloading or copying may be done for, or on behalf of, any for-profit commercial firm or for other commercial purpose without the explicit permission of the Econometric Society.
This paper would not have been possible without the generous encouragement and guidance of Isaiah Andrews. I also thank two anonymous referees, Alberto Abadie, Susan Athey, Xiaohong Chen, Avi Feller, Bruno Ferman, Wayne Gao, Edward Glaeser, Sukjin Han, Patrick Kline, Scott Kominers, Apoorva Lal, Namrata Narain, Ross Mattheis, David Ritzwoller, Brad Ross, Jonathan Roth, Bas Sanders, Suproteem Sarkar, Jesse Shapiro, Neil Shephard, Jann Spiess, Elie Tamer, Jaume Vives-i-Bastida, Davide Viviano, Chris Walker, and Jennifer Walsh for helpful discussions. I am also grateful for comments from participants of the Harvard Graduate Student Workshop in Econometrics, the 2022 Synthetic Control Workshop at the Princeton University Center for Statistics and Machine Learning, and the 2022 NBER Labor Studies Summer Institute Meeting.

Dissertation Advisor:
Professor Isaiah Andrews

Author:
Jiafeng Chen

Essays in Econometrics

Abstract

This paper notes a simple connection between synthetic control and online learning. Specifically, we recognize synthetic control as an instance of *Follow-The-Leader* (FTL). Standard results in online convex optimization then imply that, even when outcomes are chosen by an adversary, synthetic control predictions of counterfactual outcomes for the treated unit perform almost as well as an oracle weighted average of control units' outcomes. Synthetic control on differenced data performs almost as well as oracle weighted difference-in-differences, potentially making it an attractive choice in practice. We argue that this observation further supports the use of synthetic control estimators in comparative case studies.

1.1 Introduction

Synthetic control (Abadie and Gardeazabal, 2003; Abadie *et al.*, 2015) is an increasingly popular method for causal inference among policymakers, private institutions, and social scientists alike. In parallel, there is a rapidly growing methodological literature providing statistical guarantees for synthetic control methods.[2] Existing results for synthetic control—and for modifications thereof—are typically derived under a low-rank linear factor model or a vector autoregressive model of the outcomes (see, among others, Abadie *et al.*, 2010; Ben-Michael *et al.*, 2019, 2021; Ferman and Pinto, 2021; Viviano and Bradic, 2019).[3] While these statistical guarantees formally hold under these outcome models, a number of authors have expressed optimism that the synthetic control method is robust to these modeling assumptions.[4]

On the other hand, in empirical settings where synthetic control is commonly applied—where the treated unit is an aggregate entity like a country or a U.S. state—plausible outcome modeling may be challenging. Manski and Pepper (2018), in studying the effect of gun laws in the United States using state-level crime rates, provocatively ask, "what random process should be assumed to have generated the existing United States, with its realized state-year crime rates?" Granted, the low-rank linear factor model is a general class of data-generating processes and may even arise under finer-grained models on the individual outcomes contained in the aggregate data (Shi *et al.*, 2022). But to pessimists and skeptics, perhaps even such a model is implausible for the settings considered by many synthetic control studies. Indeed, if practitioners were willing to fully commit to an outcome model, perhaps they should estimate the outcome model directly—e.g., use factor model-based methods (Bai and Ng, 2002; Bai, 2003; Xu, 2017; Athey *et al.*, 2021a)—instead of using synthetic control?

As a result, existing methodological results seem to leave practitioners in a somewhat awkward

[2] See the review by Abadie (2021) as well as the special section on synthetic control methods in the *Journal of the American Statistical Association* (Abadie and Cattaneo, 2021).

[3] Notably, like this paper, Bottmer *et al.* (2021) consider a design-based framework which conditions on the outcomes and considers randomness arising solely from assignment of the treated unit or the treatment time period.

[4] For instance, Ben-Michael *et al.* (2019) write, "Outcome modeling can also be sensitive to model mis-specification, such as selecting an incorrect number of factors in a factor model. Finally, [... synthetic control] can be appropriate under multiple data generating processes (e.g., both the autoregressive model and the linear factor model) so that it is not necessary for the applied researcher to take a strong stand on which is correct." Abadie and Vives-i-Bastida (2021) write, "Synthetic controls are intuitive, transparent, and produce reliable estimates for a variety of data generating processes."

position. On the one hand, synthetic control is intuitively appealing, and it is conjectured to have good properties under a variety of outcome models. On the other hand, perhaps existing outcome models that have so far proved sufficiently analytically tractable are not always compelling in common empirical settings. To address this tension, this paper provides a few theoretical results and offers a novel interpretation of synthetic control methods. In particular, we seek guarantees for synthetic control that do not rely on any outcome model. Consequently, our results complement existing, model-based ones.

It is unlikely that nontrivial guarantees on the performance of synthetic control exist without any structure on the outcomes. However, we can derive guarantees of synthetic control's performance *relative* to a class of alternatives, such as weighted matching or weighted difference-in-differences (DID) estimators, which practitioners may otherwise choose. Our first main result shows that, on average over *hypothetical treatment timings*, synthetic control predictions are never much worse than the predictions made by any weighted matching estimator. Our second main result shows that the same is true for synthetic control on differenced data versus any weighted DID estimator. These results imply that if there is a weighted matching or DID estimator that performs well, synthetic control likewise performs well. To be clear, these *regret guarantees* average over hypothetical treatment timings, which can be interpreted as expected loss under random treatment timing, a design-based assumption.

Taken together, our results provide reassurances for practitioners, as they offer justifications for synthetic control that do not rely on particular statistical models of the outcomes. At least on average over hypothetical treatment timings, regardless of outcomes, variations of synthetic control are competitive against common estimators, such as weighted matching and weighted DID estimators. Additionally, our second result introduces a novel version of synthetic control that is competitive against DID. Since DID is extremely popular in practice (Currie *et al.*, 2020) and is thus a natural benchmark, this version of synthetic control may be particularly attractive.

We derive our results by casting prediction with panel data as an instance of *online convex optimization*, and by recognizing synthetic control as an online regression algorithm known as *Follow-The-Leader* (FTL, a name coined by Kalai and Vempala, 2005).[5] Regret guarantees on FTL in the

[5] For an introduction to online convex optimization, see Hazan (2019), Orabona (2019), Cesa-Bianchi and Lugosi (2006), and Shalev-Shwartz (2011).

online convex optimization literature translate directly to guarantees for synthetic control against a class of alternative estimators. Since most results in online convex optimization have been derived under an adversarial model—where an imagined adversary generates the data—these results translate to guarantees on synthetic control without any structure on the outcome process.

This paper is perhaps closest to Viviano and Bradic (2019). They propose an ensemble scheme to aggregate predictions from multiple predictive models, which can include synthetic control, interactive fixed effects models, and random forests. Using results from the online learning literature, Viviano and Bradic (2019)'s ensemble scheme has the no-regret property, making the ensemble predictions competitive against the predictions of any fixed predictive model in the ensemble. Under sampling processes that yield good performance for some predictive model in the ensemble, Viviano and Bradic (2019) then derive performance guarantees for the ensemble learner. In contrast, we study synthetic control directly in the worst-case setting, and connect corresponding worst-case results to guarantees on statistical risk in a design-based framework. We show that synthetic control algorithms *themselves* are no-regret online algorithms and are in fact competitive against a wide class of matching or DID estimators.

Section 1.2 sets up the notation and the decision protocol and presents our main results for synthetic control. Section 1.3 presents several extensions that show alternative guarantees on modifications of synthetic control; in particular, we show that synthetic control on differenced data is competitive against a class of difference-in-differences estimators. Section 1.4 concludes the paper.

1.2 Setup and main results

Consider a simple setup for synthetic control, following Doudchenko and Imbens (2016). There are T time periods and $N+1$ units. To simplify convergence rate expressions, we assume $T > N$ unless noted otherwise, but this assumption is not strictly necessary for our results. Let unit 0 be the only treated unit, first treated at some time $S \in \{1, \ldots, T\} \equiv [T]$. The other N units are referred to as control units. Since we observe the treated potential outcomes for the treated unit after S, estimating causal effects for unit 0 amounts to predicting the unobserved, post-S untreated potential outcomes of this unit. Thus, we focus on untreated potential outcomes.

Let the full panel of untreated potential outcomes be $\mathbf{Y} = \mathbf{Y}(0)$ with representative entry $y_{it} = y_{it}(0)$. Since we focus on untreated potential outcomes, it is convenient to omit the "(0)" symbol going forward. We let (i) $\mathbf{Y}_{1:s} = (y_{0t}, \ldots, y_{Nt})_{t=1}^{s}$ collect all untreated potential outcomes until and including time s, and we let (ii) $\mathbf{y}_t = (y_{1t}, \ldots, y_{Nt})'$ be the vector of control unit outcomes at time t. Additionally, we let $\mathbf{y}(1) = (y_1(1), \ldots, y_T(1))'$ denote the treated potential outcomes of unit 0, which are only observable for times $t \geq S$. Similarly, we let $\mathbf{y}(0) = (y_{01}(0), \ldots, y_{0T}(0))' = (y_{01}, \ldots, y_{0T})'$ denote the untreated potential outcomes of unit 0, which are observable for $t < S$. The analyst is tasked with predicting y_{0S} from observed data, which typically consist of pre-treatment outcomes of unit 0 and outcomes of untreated units. Like the main analysis in Doudchenko and Imbens (2016), we do not consider covariates extensively, though Section 1.3.3 considers matching on covariates as a form of regularization.[6]

Synthetic control (Abadie and Gardeazabal, 2003; Abadie *et al.*, 2010), in its basic form, chooses some convex weights $\hat{\theta}_S$ that minimize past prediction errors

$$\hat{\theta}_S \in \arg\min_{\theta \in \Theta} \sum_{t=1}^{S-1} (y_{0t} - \theta' \mathbf{y}_t)^2, \tag{1.1}$$

where $\Theta \equiv \{(\theta_1, \ldots, \theta_N) \in \mathbb{R}^N : \theta_i \geq 0, 1'\theta = 1\}$ is the simplex. For a one-step-ahead forecast for y_{0S}, synthetic control outputs the weighted average $\hat{y}_S \equiv \hat{\theta}_S' \mathbf{y}_S$, and forms the treatment effect estimate $\hat{\tau}_S \equiv y_S(1) - \hat{y}_S$.

Theoretical guarantees for treatment effect estimates $\hat{\tau}_S$ often rely on statistical models of the outcomes $\mathbf{Y}$. While synthetic control has good performance under a range of outcome models, one may still doubt whether these models are plausible—and whether the underlying repeated sampling thought experiments are appropriate—in the spirit of comments by Manski and Pepper (2018). In contrast to the usual outcome modeling approach, we instead consider a worst-case setting where

[6] To extend our analysis to cases with covariates, at a minimum, we can interpret $\mathbf{Y}$ as the residuals of the untreated potential outcomes against some fixed regression function of the covariates, i.e. $y_{it} = y_{it}^* - h_t(x_i)$, for fixed h_t (perhaps estimated from auxiliary data), outcomes y_{it}^*, and covariate vectors x_i. The residualization is similar to Section 5.5 in Doudchenko and Imbens (2016) and expression (16) in Abadie (2021), but is stronger due to h_t being fixed for different adversarial choices of $\mathbf{Y}$. Our results apply so long as these residuals obey the boundedness assumption $\|\mathbf{Y}\|_\infty \leq 1$ that we impose later.

the outcomes are generated by an adversary.[7] Doing so has the appeal of giving decision-theoretic justification for methods while being entirely agnostic towards the data-generating process. Since a dizzying range of reasonable data-generating models and identifying assumptions are possible in panel data settings—yet perhaps none are unquestionably realistic—this worst-case view is valuable, and worst-case guarantees can be comforting.

In particular, we assume an adversary picks the outcomes $\mathbf{Y}$—or, equivalently, we derive results that hold uniformly over $\{\mathbf{Y} : \|\mathbf{Y}\|_\infty \le 1\}$. Specifically, we consider the following protocol between an analyst and an adversary:

(P1) The analyst commits to a class of linear prediction rules $\hat{y}_t \equiv f(\mathbf{y}_t; \theta_t(\mathbf{Y}_{1:t-1})) = \theta_t' \mathbf{y}_t$, parametrized by some $\theta_t \in \Theta$ that may be chosen as a function of the past data $\mathbf{Y}_{1:t-1}$. We refer to the maps $\sigma \equiv \{\theta_t(\cdot) \colon t \in [T]\}$ as the analyst's *strategy*. This means that if the treatment time S is equal to t, then the analyst reports $\hat{y}_t$ as their prediction for the untreated potential outcome at the first period after treatment.

(P2) The adversary chooses the matrix of outcomes $\mathbf{Y}$. In order to obtain nontrivial bounds, we assume that the adversary cannot choose arbitrarily large outcomes, and without further loss of generality, we assume $\|\mathbf{Y}\|_\infty \le 1$. Since we are interested in the worst case, the adversary may choose $\mathbf{Y}$ with knowledge of σ.

(P3) The analyst suffers loss equal to squared prediction error at time S: i.e., $\ell(\hat{y}_S, y_{0S}) \equiv (\hat{y}_S - y_{0S})^2$.[8]

If we just consider the prediction error at a specific value of $S = s_0$ known to the adversary, then there is little hope of obtaining nontrivial guarantees. In this case, the adversary is simply too powerful: They can choose outcomes at s_0 such that any method performs badly on any metric. Motivated by this difficulty, we consider a different performance criterion: Under (P1) to (P3), the analyst's average

[7] The adversarial framework, popular in online learning, dates to the works of Hannan (1958) and Blackwell (1956).

[8] The protocol (P1) to (P3) easily generalizes when we replace $f(\mathbf{y}_t, \theta_t)$ with any known scalar function and $\ell(\cdot, \cdot)$ with any loss function, so long as $\theta \mapsto \ell(f(\mathbf{y}_t, \theta), y_{0t})$ is convex and bounded. Our results in Section 1.3.3 allow for general loss functions.

squared loss, averaging over *hypothetical values of S*, is

$$\frac{1}{T}\sum_{S=1}^{T}(y_{0S}-\hat{y}_S)^2=\frac{1}{T}\sum_{S=1}^{T}(y_{0S}-\theta_S'\mathbf{y}_S)^2=E_{S\sim\mathrm{Unif}[T]}\left[(y_{0S}-\hat{y}_S)^2\right]. \tag{1.2}$$

Most results in this paper are guarantees in terms of the decision criterion (1.2) for synthetic control, where synthetic control (1.1) is viewed as a particular strategy σ under (P1) to (P3).

As the second equality in (1.2) indicates, under an additional assumption that treatment timing is *uniformly random*, $S \sim \mathrm{Unif}[T]$, the average loss over hypothetical treatment timings is equal to the expected squared loss over S. This additional assumption is a design-based perspective (Doudchenko and Imbens, 2016; Bottmer *et al.*, 2021) on the panel causal inference problem. This perspective enables us to interpret average prediction loss over hypothetical treatment timings as expected prediction loss under the random treatment time S. The latter can in turn be thought of as design-based risk. Uniformly random assignment of S is restrictive, but we shall relax this requirement in Sections A.2.1 and 1.3.1.

We now make clear the connection with online convex optimization (see Section 1.1 in Hazan, 2019). Online convex optimization works with the following general protocol. Time t increments sequentially for T periods, and at time t:

(O1) An online player chooses some $\theta_t \in \Theta$, where $\Theta \subset \mathbb{R}^d$ is a bounded convex set. The choice θ_t may depend on the loss functions $\{\ell_s : s < t\}$ chosen by the adversary in the past.

(O2) After θ_t is chosen, an adversary chooses a loss function $\ell_t : \Theta \to \mathbb{R}$ from some given set of loss functions, which may be further parametrized. These loss functions are constrained to be convex and bounded but can otherwise be quite general. They are often further constrained in order to obtain specific regret results.

(O3) The player suffers loss $\ell_t(\theta_t)$ and observes the entire loss function $\ell_t(\cdot)$.[9] The player may update their decision θ_{t+1} based on $\ell_1(\cdot), \ldots, \ell_t(\cdot)$.

At the end of the game, the online player suffers total loss $\sum_{t=1}^{T}\ell_t(\theta_t)$.

[9] A closely related setting, where the player only observes $\ell_t(\theta_t)$ instead of the entire loss function $\ell_t(\cdot)$, is known as *bandit convex optimization* (see Chapter 6 in Hazan, 2019), of which the adversarial multi-armed bandit problem (Robbins, 1952; Bubeck and Cesa-Bianchi, 2012) is a special case.

Our setup of the panel prediction protocol, (P1) to (P3), is then an instance of online convex optimization, (O1) to (O3). To see this, the most important step is to recognize that the analyst's loss (1.2) is analogous to the online player's loss, and therefore to think of the analyst as making sequential decisions where $\mathbf{Y}$ is sequentially revealed to them. This change in perspective relies on (i) our choice of decision criterion (1.2) and (ii) the fact that the analyst's decisions $\theta_t(\cdot)$ only require outcomes prior to t. Indeed, by fixing $\mathbf{Y}$ and considering the hypothetical values of $S = 1, \ldots, T$ sequentially, we can treat the analyst as if they were solving an online problem and learning from data in the past—even though, for any particular value of S, they are only confronted with a static, offline problem. To be clear, we are not considering some online version of synthetic control; the connection to online convex optimization comes from considering hypothetical, unrealized values of S.

After viewing the analyst's problem as an online problem, we may straightforwardly establish the remaining correspondences. First, note that the simplex Θ is convex and bounded. Second, note that we may imagine the adversary in the panel prediction game as picking loss functions $\ell_t(\cdot)$ of the form $\theta \mapsto (y_{0t} - \theta' \mathbf{y}_t)^2$, parametrized by the potential outcomes $(y_{0t}, \mathbf{y}_t)$. These loss functions are indeed convex in θ and bounded, since both θ and $\mathbf{Y}$ are bounded. Finally, note that the average loss (1.2) is equal to $\frac{1}{T} \sum_{t=1}^{T} \ell_t(\theta_t)$, which is simply the total loss in the online protocol scaled by $\frac{1}{T}$.[10]

Having recognized our setup as an instance of online convex optimization, the main observation of this paper recognizes that synthetic control is an online learning algorithm known as *Follow-the-Leader* (FTL). FTL, under (O1) to (O3), is the algorithm that, when prompted for a decision in (O1), simply

[10] It may be tempting to ask whether the same argument applies to "horizontal regression" (Athey *et al.*, 2021a), where one regresses y_{iS} on $y_{i1}, \ldots, y_{iS-1}$, perhaps constraining the coefficients to some bounded, convex set. Since synthetic control can be viewed as a "vertical regression," where one regresses y_{0t} on $y_{1t}, \ldots, y_{Nt}$, it seems we may apply our argument to the transposed $\mathbf{Y}$ matrix. Indeed, we may formulate analogous claims by replacing t with i, s with j, S with some randomly chosen unit $M \in [N]$, and T with N. However, a difficulty with this interpretation is that synthetic control (1.1) naturally only uses information in the past ($t < S$), but the analogous restriction in horizontal regression, $i < M$, for a randomly chosen treated unit $M \in [N]$, is much less natural.

chooses θ_t to minimize past losses:[11,12]

$$\theta_t \in \arg\min_{\theta \in \Theta} \sum_{s<t} \ell_s(\theta).$$

Observation 1.2.1. *Synthetic control* (1.1) *is an instance of FTL applied to the panel prediction protocol (P1) to (P3).*

Standard online convex optimization results on *regret* then apply to synthetic control as well. Before introducing these results, let us define regret as the gap between the total loss of a strategy σ and the best fixed weights θ in hindsight:

$$\begin{aligned}
\text{Regret}_T(\sigma; \mathbf{Y}) &\equiv \sum_{t=1}^{T} \ell_t(\theta_t) - \min_{\theta \in \Theta} \sum_{t=1}^{T} \ell_t(\theta) && (1.3) \\
&= \sum_{S=1}^{T} (y_{0S} - \theta_S' \mathbf{y}_S)^2 - \min_{\theta \in \Theta} \sum_{S=1}^{T} (y_{0S} - \theta' \mathbf{y}_S)^2 && (1.4) \\
&= T \left(E_S[(y_{0S} - \theta_S' \mathbf{y}_S)^2] - \min_{\theta \in \Theta} E_S[(y_{0S} - \theta' \mathbf{y}_S)^2] \right) && (1.5) \\
&\geq T \left(E_S[(y_{0S} - \theta_S' \mathbf{y}_S)^2] - E_S[(y_{0S} - \theta' \mathbf{y}_S)^2] \right) \text{ for any } \theta \in \Theta. && (1.6)
\end{aligned}$$

(1.4) observes that, in our setting, regret is the difference between total squared prediction error of a strategy σ and that of the best fixed weights θ chosen in hindsight, summing over hypothetical treatment times S. (1.5) interprets the sum of losses as T times the expected loss under random treatment timing. Finally, (1.6) observes that regret is an upper bound of the expected error gap between the strategy σ and any fixed weights θ. We refer to $\arg\min_{\theta \in \Theta} \sum_{S=1}^{T} (y_{0S} - \theta' \mathbf{y}_S)^2$ as the *oracle weighted match*—the best set of weights for a given realization of the data $\mathbf{Y}$.

Focusing on regret rather than loss shifts the goalposts from performance to *competition*, which is a more fruitful perspective in our adversarial setting. After all, we cannot hope to obtain meaningful loss control as the all-powerful adversary can make the analyst miserable. However, the crucial

[11] FTL is also known as fictitious play in game theory (Brown, 1951). The name "follow-the-leader," coined by Kalai and Vempala (2005), is popular in the recent computer science literature. For an introduction to FTL and similar algorithms, see Chapter 5 in Hazan (2019) and Chapters 1 and 7 in Orabona (2019).

[12] When there are multiple minima, the choice of θ_t does not affect our theoretical guarantees. Nevertheless, it seems sensible in practice to take the minimum that is smallest in some norm, e.g. $\|\cdot\|_2$.

insight of regret analysis is that, for certain strategies σ, the adversary cannot simultaneously make the analyst suffer high loss while letting some fixed strategy θ perform well—in other words, if any fixed θ performs well, then σ performs almost as well over time. Indeed, if regret is sublinear, i.e., $\text{Regret}_T \leq o(T)$,[13] then the strategy σ never performs much worse than any fixed weights θ, on average over hypothetical treatment timing S. In this case, we can interpret σ as a strategy that is *competitive* against the class of weighted matching estimators.

It may appear surprising that these no-regret strategies σ exist in the first place. We emphasize that σ can output different weights θ_t, chosen adaptively over time, while σ is compared to an oracle that uses the best fixed weights. As a result, σ can compensate for its lack of oracle access by changing its choices judiciously over time.

The main result of this paper shows that the regret of synthetic control under quadratic loss is logarithmic in T. The result follows from a direct application of Hazan *et al.* (2007)'s regret bound for FTL (Theorem 5 in their paper, reproduced as Theorem A.1.1 in the appendix).

Theorem 1.2.2. *With bounded outcomes $\|\mathbf{Y}\|_\infty \leq 1$, synthetic control (1.1), denoted σ, satisfies the regret bound*[14]

$$\text{Regret}_T(\sigma, \mathbf{Y}) \leq 16N(\log(\sqrt{N}T) + 1) = O(N \log T).$$

Theorem 1.2.2 shows that the synthetic control strategy (1.1) achieves logarithmic regret—and as a result, the average difference between the losses of synthetic control and losses of the oracle weighted match vanishes quickly as a function of T.[15] In particular, if there exists a weighted average of the

[13] We mean $\text{Regret}_T \leq o(T)$ in the sense that $\limsup_{T\to\infty} \frac{1}{T}\text{Regret}_T \leq 0$, since it is possible for Regret_T to be negative. Following the online convex optimization literature, we sometimes refer to σ as no-regret if it has sublinear regret.

[14] We say $f(N,T) = O(g(N,T))$ for $g(N,T) > 0$ if, for any sequence $N_T < T$ and $T \to \infty$,

$$\limsup_{T\to\infty} \frac{f(N_T,T)}{g(N_T,T)} < \infty.$$

In the conclusion of Theorem 1.2.2, the inequality does not require $T > N$. The assumption $T > N$ is only used for the simplification $16N(\log(\sqrt{N}T) + 1) = O(N\log N + N \log T) = O(N \log T)$. Of course, the regret bound is less interesting if $\limsup N \log T/T > 0$.

[15] Restricting θ to the simplex Θ—a debated choice in the synthetic control literature—is somewhat important for the dependence on N, in so far as the simplex is bounded in $\|\cdot\|_1$. This is a consequence of the assumption that the outcomes $\mathbf{Y}$ are bounded in the dual norm $\|\cdot\|_\infty$, which implies a bound on $\theta'\mathbf{y}_t$ that is free of N, T. In contrast, if we let $\Theta = \{\theta : \|\theta\|_2 \leq D/2\}$ be an ℓ_2-ball, then the regret bound worsens to $O(D^2N^2\log(T))$.

untreated units' outcomes that tracks $\mathbf{y}(0)$ well, then the average one-step-ahead loss of synthetic control estimates is only worse by $O\left(\frac{N\log T}{T}\right)$.

On its own, Theorem 1.2.2 is purely an optimization result; we now offer a few comments on its statistical implications. As a preview, under random treatment timing, Theorem 1.2.2 implies that the *risk* of estimating the causal effect at time S for synthetic control is not too much higher than that for any weighted matching estimator. Indeed, if any weighted matching estimator performs well, then synthetic control achieves low risk as well. Our discussion below translates Theorem 1.2.2 into guarantees on the expected loss at treatment time—expressing regret as (1.5)—which relies on the design assumption that S is randomly assigned. Nevertheless, we stress that we could view Theorem 1.2.2 purely as guarantees of average loss over hypothetical timings S—expressing regret only as (1.4)—which does not require a treatment timing assumption.

We can interpret regret as a gap in the design-based risk of estimating treatment effects. Specifically, we can interpret the expected loss of predicting the untreated outcome as the risk of estimating the treatment effect:

$$\begin{aligned}\mathrm{Risk}(\sigma, \mathbf{Y}, \mathbf{y}(1)) &\equiv E_S\left[(\tau_S - \hat{\tau}_S(\sigma))^2\right] \\ &\equiv E_S\left[((y_S(1) - y_{0S}) - (y_S(1) - \hat{y}_S))^2\right] \\ &= E_S[(y_{0S} - \hat{y}_S)^2]. \end{aligned} \tag{1.7}$$

Hence, (1.5) and (1.7), combined with Theorem 1.2.2, imply that the risk of using synthetic control is no more than $N\log T/T$ worse than the risk of the oracle weighted match,[16] regardless of the potential outcomes $\mathbf{Y}, \mathbf{y}(1)$:

$$\mathrm{Risk}(\sigma, \mathbf{Y}, \mathbf{y}(1)) - \min_{\theta\in\Theta} \mathrm{Risk}(\theta, \mathbf{Y}, \mathbf{y}(1)) = \frac{1}{T}\mathrm{Regret}_T(\sigma, \mathbf{Y}) = O\left(\frac{N\log T}{T}\right). \tag{1.8}$$

This observation connects regret on prediction of the untreated potential outcome with differences in the risk of estimating treatment effects. Roughly speaking, (1.8) shows that synthetic control estimates of one-step-ahead causal effects are competitive against that of any fixed weighted match, for any realization of $\mathbf{Y}, \mathbf{y}(1)$, on average over S.

[16] We slightly abuse notation and use θ to denote the strategy that outputs θ every period.

Of course, since the guarantee (1.8) holds for every $\mathbf{Y}$, it continues to hold when we average over $\mathbf{Y}$ and $\mathbf{y}(1)$, over a joint distribution P that respects the boundedness condition $\|\mathbf{Y}\|_\infty \le 1$. In this sense, analyzing regret in the adversarial framework not only does not preclude statistical interpretations, but rather facilitates analysis in a wide range of outcome models.[17] Formally, let $\mathcal{P}$ be a family of distributions for $\mathbf{Y}, \mathbf{y}(1)$ such that $P(\|\mathbf{Y}\|_\infty \le 1) = 1$ for all $P \in \mathcal{P}$. Under an outcome model P, we may understand $\operatorname{Risk}(\sigma, \mathbf{Y}, \mathbf{y}(1))$ as *conditional risk* and $E_P \operatorname{Risk}(\sigma, \mathbf{Y}, \mathbf{y}(1))$ as *unconditional risk*. Then, (1.8) implies that[18]

$$\sup_{P \in \mathcal{P}} E_P \left[\operatorname{Risk}(\sigma, \mathbf{Y}, \mathbf{y}(1)) - \min_{\theta \in \Theta} \operatorname{Risk}(\theta, \mathbf{Y}, \mathbf{y}(1)) \right] = O\left(\frac{N \log T}{T} \right). \tag{1.9}$$

Therefore, the unconditional risk of synthetic control is never much worse than the risk of the oracle weighted match

$$R^*_\Theta \equiv E_P \left[\min_{\theta \in \Theta} \operatorname{Risk}(\theta, \mathbf{Y}, \mathbf{y}(1)) \right].$$

Hence, if the data-generating process P guarantees that R^*_Θ is small, then synthetic control achieves low expected risk as well. Concretely speaking, this latter requirement is that, for most realizations of the data, had we observed all the potential outcomes, we could find a weighted match that tracks the potential outcomes $y_{01}, \ldots, y_{0T}$ well, so that[19]

$$E_P \left[\min_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^T (y_{0t} - \theta' \mathbf{y}_t)^2 \right] \approx 0.$$

In many empirical settings, it seems plausible that the oracle weighted match performs well.[20]

[17] The technique of "online-to-batch conversion" in the online learning literature exploits this intuition to prove results in batch (i.i.d.) settings via results in online adversarial settings.

[18] Abernethy *et al.* (2009) show that a minimax theorem applies, and

$$\sup_P \inf_\sigma E_P \left[\operatorname{Risk}(\sigma, \mathbf{Y}, \mathbf{y}(1))] - \min_{\theta \in \Theta} \operatorname{Risk}(\theta, \mathbf{Y}, \mathbf{y}(1)) \right] = \frac{1}{T} \inf_\sigma \sup_{\mathbf{Y}} \operatorname{Regret}_T(\sigma, \mathbf{Y}).$$

Note that the $\le$ direction is immediate via the min-max inequality. This result shows that the worst-case optimal risk differences in a stochastic setting (i.e. the analyst knows P and responds to it optimally) is equal to minimax regret. In this sense, worst-case regret analysis is not by itself conservative for a stochastic setting—minimax regret is a tight upper bound for performance in stochastic settings.

[19] Also, observe that $E_P[\min_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^T (y_{0t} - \theta' \mathbf{y}_t)^2] \le \min_{\theta \in \Theta} E_P[\frac{1}{T} \sum_{t=1}^T (y_{0t} - \theta' \mathbf{y}_t)^2]$, and thus the guarantee (1.9) is stronger in the sense that it allows the oracle θ to depend on the realization of the data.

[20] We recognize that under many data-generating models, there is unforecastable, idiosyncratic randomness in y_{0t}. As a

Abadie (2021) states the following intuition in many comparative case studies: "[T]he effect of an intervention can be inferred by comparing the evolution of the outcome variables of interest between the unit exposed to treatment and a group of units that are similar to the exposed unit but were not affected by the treatment." More formally speaking, a well-fitting oracle weighted match also resembles—and implies—Abadie *et al.* (2010)'s assumption that there exists a perfect pre-treatment fit of the outcomes. When the oracle weighted match performs well, our regret guarantees imply a guarantee on the loss of the feasible synthetic control estimator, making it an attractive option for causal inference in comparative case studies.

Even if no weighted average of the untreated units tracks y_{0t} closely, synthetic control continues to enjoy the assurance that it performs almost as well as the best weighted match. Moreover, in the general online learning setup (O1) to (O3), this no-regret property cannot be attained without choosing θ_t in some data-dependent manner.[21] This observation rules out alternatives such as simple difference-in-differences, which does not aggregate the control units in a data-dependent manner. In contrast, in Section 1.3, we additionally show that synthetic control on differenced data performs almost as well as the best *weighted* difference-in-differences estimator, a popular class of estimators in practice.

1.3 Extensions

1.3.1 Non-uniform treatment timing

The previous interpretations—in (1.5) and (1.7)—rely on interpreting average loss over hypothetical values of S as expected loss over S, which requires uniform treatment timing $S \sim \text{Unif}[T]$. Despite being plausible in certain settings and appearing elsewhere in the literature (Doudchenko and Imbens,

result, there may not exist a synthetic match that perfectly tracks the *realized* series y_{0t} (even though such a match may exist that tracks various conditional expectations of y_{0t} quite well). In many such cases, since squared error can be orthogonally decomposed, risk differences for estimating y_{0t} are also risk differences for estimating conditional means μ_t of y_{0t}. We discuss these results in Section A.2.3.

[21] See Section A.1.2 for a simple argument in a general setup with unspecified $\ell(\cdot)$. Since simple DID does not choose weights adaptively, it fails to control regret against the class of weighted DID estimators that we discuss in Section 1.3.

2016; Bottmer *et al.*, 2021), this assumption is perhaps crude.[22] To some extent this is inevitable: Since we are agnostic on the outcome generation process, it is unavoidable to make treatment timing assumptions in order to obtain nontrivial statistical results on estimation of causal quantities. Nevertheless, note that such an assumption is only necessary for interpreting average losses as expected losses. The *a priori* proposition that *it is reasonable to expect a causal estimator to predict well relative to some oracle, at least on average over hypothetical treatment timings,* strikes us as defensible. Accepting this dictum relieves us of any need to model treatment timing.

Even if we wish to maintain the interpretation of average loss as expected loss, we can relax the uniform treatment timing assumption. In this subsection, we show that if the treatment timing distribution is known, then a weighted version of synthetic control achieves logarithmic weighted regret. Moreover, even if the treatment timing distribution is non-uniform, unknown, and possibly chosen by the adversary, we continue to show that synthetic control performs well if some weighted average of untreated units predicts y_{0S} accurately. Both results have constants that worsen if the treatment timing distribution deviates far from $\mathrm{Unif}[T]$.

Suppose the conditional distribution $(S \mid \mathbf{Y})$ is denoted by $\pi = (\pi_1, \dots, \pi_T)'$, where $\sum_{t=1}^T \pi_t = 1$, which may depend on $\mathbf{Y}$. Note that, for a known π, we may apply the same argument in Theorem 1.2.2 to the following weighted synthetic control estimator:

$$\hat{\theta}_S^{\pi} \in \arg\min_{\theta \in \Theta} \sum_{t<S} \pi_t (y_{0t} - \theta' \mathbf{y}_t)^2, \tag{1.10}$$

by redefining the loss functions $\ell_t(\cdot)$. This argument shows that (1.10) achieves $\log T$ weighted regret, stated in the following corollary. Note that (1.10) implements FTL with loss functions $\ell_t(\theta) \equiv \pi_t (y_{0t} - \theta' \mathbf{y}_t)^2$, and hence the argument of Hazan *et al.* (2007) applies.

Corollary 1.3.1. *Suppose $S \sim \pi$, $\frac{1}{CT} \le \pi_t \le \frac{C}{T}$ for some C, and $\|\mathbf{Y}\|_\infty \le 1$. Then weighted*

[22] Doudchenko and Imbens (2016) discuss inference in synthetic control via randomization of the treatment timing in their Section 6.2. Bottmer *et al.* (2021) consider randomization of the treated period in their Assumption 2, though, in their setting, the treatment lasts only one period. We also note that the randomness *per se* of S conditional on $\mathbf{Y}$ can be realistic, but that its distribution is uniform and known is restrictive.

synthetic control (1.10)*, denoted* σ_π*, achieves weighted regret bound*

$$\begin{aligned}\operatorname{Regret}_T(\sigma_\pi; \pi, \mathbf{Y}) &\equiv T \cdot \left(E_{S\sim\pi}[(y_{0S} - \hat\theta_S' \mathbf{y}_S)^2] - \min_{\theta\in\Theta} E_{S\sim\pi}[(y_{0S} - \theta'\mathbf{y}_S)^2] \right) \\ &\le 16C^3 N \left[\log \frac{\sqrt{N}T}{C^2} + 1\right] = O(C^3 N \log T).\end{aligned} \tag{1.11}$$

Theorem 1.3.1 shows that the weighted regret—a difference in π-expected loss—is logarithmic in T, thereby controlling the worst-case gap between weighted synthetic control and the oracle weighted match for the expected loss. Assuming a known π could be reasonable. With a known dynamic treatment regime, π can depend on $\mathbf{Y}_{1:S-1}$, but is known whenever the analyst is prompted for a prediction at time S.[23] We can also interpret Theorem 1.3.1 as providing guarantees on differences in Bayes risk under the analyst's prior $S \sim \pi$, independent of $\mathbf{Y}$.

Even when π is *unknown* and chosen by the adversary, we can bound the loss of unweighted synthetic control, so long as π is not too far from uniform.

Corollary 1.3.2. *Suppose* $S \sim \pi$, $\pi_t \le C/T$ *for some* C*, and* $\|\mathbf{Y}\|_\infty \le 1$*. Then synthetic control* (1.1)*, denoted* σ*, achieves the following bound on* the expected loss

$$E_{S\sim\pi}\left[(y_{0S} - \hat\theta_S'\mathbf{y}_S)^2\right] \le C\left(\min_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^T (y_{0t} - \theta'\mathbf{y}_t)^2 + \frac{1}{T}\operatorname{Regret}_T(\sigma; \mathbf{Y})\right), \tag{1.12}$$

where $\operatorname{Regret}_T(\sigma; \mathbf{Y})$ *is defined by* (1.4)*. Hence, for any joint distribution* Q *of* $(\mathbf{Y}, S)$ *where* $Q(S = t \mid \mathbf{Y}) \le C/T$ *for all* t*, and* $Q(\|\mathbf{Y}\|_\infty \le 1) = 1$*, we have the average loss bound*

$$E_Q[(y_{0S} - \hat\theta_S'\mathbf{y}_S)^2] \le C\left(E_Q\left[\min_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^T (y_{0t} - \theta'\mathbf{y}_t)^2\right] + O\left(\frac{N\log T}{T}\right)\right). \tag{1.13}$$

The result (1.12) shows that, uniformly over all bounded $\mathbf{Y}$ and bounded treatment distributions π, the expected squared error is bounded by the average loss of the oracle weighted match plus the regret, all scaled with a constant C that indexes how far π deviates from the uniform distribution. Under the

[23] Since the bound is for a fixed $\mathbf{Y}$, we can allow π to depend on $\mathbf{Y}$, so long as $\pi_t(\mathbf{Y})$ is known at time $t+1$ so that the analyst can compute (1.10). This allows for Theorem 1.3.1 to be applied in the following example, which is a more realistic design-based setting. There is a known *dynamic treatment regime* (Chakraborty and Murphy, 2014) parametrizing the treatment hazard: That is,

$$P(S = t \mid S \ge t, \mathbf{Y}) = r_t(\mathbf{Y}_{1:t-1})$$

for some known $r_t(\cdot)$. Then $\pi_t(\mathbf{Y}) = P(S = t \mid \mathbf{Y}) = (1 - r_1)\cdots(1 - r_{t-1})r_t$ is a function of $\mathbf{Y}_{1:t-1}$. We thank Davide Viviano for suggesting this extension.

same assumption that the oracle weighted match performs well on average, (1.12) continues to show that the treatment estimation risk of synthetic control is small. Since such a result is valid for all $\mathbf{Y}$ and π, we may understand (1.12) as a bound that holds even in a setting where the adversary picks both $\mathbf{Y}$ and π, with the restriction that $\pi_t \le C/T$, but otherwise unrestricted in the dependence between $\mathbf{Y}$ and π.

As before, since (1.12) is a guarantee uniformly over $\mathbf{Y}$, it is also a guarantee when we average over $\mathbf{Y}$ under an outcome model, yielding (1.13). Again, (1.13) shows that *for any* joint distribution of the bounded outcomes and the treatment timing, the unconditional risk of synthetic control is small when the expected oracle conditional risk, $E_Q[\min_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^T (y_{0t} - \theta' \mathbf{y}_t)^2]$, is small—so long as S has sufficient randomness conditional on $\mathbf{Y}$ so that C is not too large.

So far, we have considered weighted averages of untreated units as the class of competing estimators. These competing estimators are matching estimators. However, a more common class of competing estimators in applications are difference-in-differences (DID) estimators. It turns out that synthetic control on preprocessed data has regret guarantees against a class of DID estimators, which we turn to in the next subsection.

1.3.2 Competing against DID

Section 1.2 shows that the original synthetic control estimator is competitive against a class of matching estimators that use weighted averages of untreated units as matches for the treated unit. However, in many applications in economics, matching estimators are much less popular than DID estimators, since the latter accounts for unobserved confounders that are additive and constant over time. In this subsection, we show that synthetic control on differenced data is competitive against a large class of DID estimators. Additionally, Section A.1.3 offers regret guarantees against other flavors of DID estimators.

In practice, a common DID specification is the following two-way fixed effects regression:

$$\min_{\mu_i, \alpha_t, \lambda} \sum_{i=0}^{N} \sum_{t=1}^{S} \left(y_{it}^{\text{obs}} - \mu_i - \alpha_t - \lambda \mathbb{1}\left[(i,t) = (0,S)\right] \right)^2,$$

where the observed outcome $y_{it}^{\text{obs}} = y_{it}$ for all $(i,t) \neq (0,S)$, and $y_{0S}^{\text{obs}} = y_S(1)$. This specification regresses the observed outcomes on unit and time fixed effects, and uses the estimated coefficient λ as an estimate of the treatment effect $y_S(1) - y_{0S}$. Implicitly, this regression uses the estimated fixed effects $\mu_0 + \alpha_S$ as a forecast for the unobserved y_{0S}. We consider a weighted generalization of this regression, a special case of the synthetic DID estimators in Arkhangelsky *et al.* (2021):[24],[25]

$$\min_{\mu_i,\alpha_t,\lambda} \sum_{i=0}^{N}\sum_{t=1}^{S} w_i(y_{it}^{\text{obs}} - \mu_i - \alpha_t - \lambda \mathbb{1}\left[(i,t) = (0,S)\right])^2 \quad w_0 = 1, \sum_{i=1}^{N} w_i = 1, w_i \geq 0. \tag{1.14}$$

For convex weights $w = (w_1, \ldots, w_N)'$, denote by $\sigma_{\text{TWFE}}(w)$ the strategy that estimates (1.14) on the data $(\mathbf{Y}_{1:t-1}, \mathbf{y}_t)$ at time t,[26] and outputs the estimated coefficients $\mu_0 + \alpha_t$ as a prediction for y_{0t}. By varying over $w \in \Theta$, we obtain a class of competing DID strategies, where conventional DID corresponds to picking uniform weights $w = (1/N, \ldots, 1/N)'$. We calculate in Section A.1.6 that the prediction that $\sigma_{\text{TWFE}}(w)$ makes is

$$\hat{y}_t(\sigma_{\text{TWFE}}(w)) = \frac{1}{t-1}\sum_{s=1}^{t-1} y_{0s} + w'\left(\mathbf{y}_t - \frac{1}{t-1}\sum_{s=1}^{t-1}\mathbf{y}_s\right) \qquad t \geq 2,$$

which simply uses the outcome difference against historical averages of untreated units to forecast that of unit 0. Note that this strategy amounts to using a weighted match with weight w on the *differenced data*

$$\tilde{y}_{i1} = y_{i1} \qquad \tilde{y}_{it} \equiv y_{it} - \frac{1}{t-1}\sum_{s=1}^{t-1} y_{is} \qquad |\tilde{y}_{it}| \leq 2$$

to forecast the same differences of unit 0, $\tilde{y}_{0t}$. Therefore, we may apply Theorem 1.2.2 and show the following regret bound.

Theorem 1.3.3. *Consider synthetic control on the differenced data, where the analyst computes*

$$\hat{\theta}_t \in \arg\min_{\theta\in\Theta} \sum_{s<t} \left(\tilde{y}_{0s} - \theta'\tilde{\mathbf{y}}_s\right)^2$$

and predicts $\hat{y}_t = \frac{1}{t-1}\sum_{s<t} y_{0s} + \hat{\theta}_t'\tilde{\mathbf{y}}_t$. *Here,* $\tilde{y}_{it} = y_{it} - \frac{1}{t-1}\sum_{s<t} y_{is}$ *is the difference against*

[24] The weight w_0 does not affect $\mu_0 + \alpha_S$ achieving the optimum in the least-squares problem, per the calculation in Section A.1.6. As a result, we normalize $w_0 = 1$. Moreover, specifically, (1.14) is a special case of synthetic DID, (1) in Arkhangelsky *et al.* (2021), with only unit-level weights and no time-level weights.

[25] (1.14) is underdetermined if $S = 1$. The ensuing discussion assumes $\sum_{i=1}^N w_i y_{i1}$ is the weighted two-way fixed effects prediction for y_{01}.

[26] The value of y_{0t} does not enter $\alpha_S + \mu_0$ since it is absorbed by the coefficient λ.

historical means, and $\tilde{\mathbf{y}}_t = (\tilde{y}_{1,t}, \dots, \tilde{y}_{N,t})'$*. Then we have the following regret guarantee against the oracle* σ_{TWFE}*, whose weights are chosen ex post:*

$$\sum_{t=1}^{T} (y_{0t} - \hat{y}_t)^2 - \min_{\theta \in \Theta} \sum_{t=1}^{T} (y_{0t} - \hat{y}_t(\sigma_{\text{TWFE}}(\theta)))^2 \le CN \log T$$

for some constant C*.*

Theorem 1.3.3 shows that synthetic control on differenced data controls regret against the class of DID estimators (1.14).[27] In particular, the class of DID benchmarks corresponds to weighted two-way fixed effects regressions, and synthetic control is competitive against any fixed weighting. In this sense, Theorem 1.3.3 builds on the intuition that synthetic control is a generalization of DID (Doudchenko and Imbens, 2016) to show that a version of synthetic control performs as well as any weighted DID estimator. Again, if any weighted DID estimator performs well, then Theorem 1.3.3 becomes a performance guarantee on synthetic control. Moreover, since (1.14) is a popular alternative for many practitioners—setting aside whether there is a weighted DID that performs well—Theorem 1.3.3 shows that it is without much loss to use synthetic control in such settings instead. Since DID is more popular in practice than weighted matching, competitive performance against DID is a more relevant consideration, which suggests prioritizing synthetic control on differenced data $\tilde{y}_{it}$ over classic synthetic control (1.1).[28]

To the best of our knowledge, the difference scheme $\tilde{y}_{it}$ has yet to be considered in the literature. We do note that since the resulting predictions are equivalent to a weighted two-way fixed effects regression, this proposed synthetic control scheme can be thought of as synthetic DID (Arkhangelsky *et al.*, 2021) with weights chosen by constrained least-squares on $\tilde{y}_{it}$. We also note that $\tilde{y}_{it}$ is slightly different from Ferman and Pinto (2021)'s demeaned synthetic control, which takes the difference $\dot{y}_{it} \equiv y_{it} - \frac{1}{t}\sum_{s=1}^{t} y_{is}$. In Section A.1.3, we show that Ferman and Pinto (2021)'s demeaned synthetic

[27] The benchmark class of DID estimators in Theorem 1.3.3 output predictions in a sequential manner, in so far as the coefficients in the regression (1.14) depend on S. In contrast, Theorem A.1.3 compares synthetic control against a class of static DID estimators that do not exhibit this feature.

[28] This comment is with the caveat that the constant in Theorem 1.3.3 is worse than that in Theorem 1.2.2. It seems possible to further improve the guarantee in Theorem 1.3.3, since in our proof, we solely use the implication $|\tilde{y}_{it}| \le 2$ and do not restrict the adversary from choosing $\tilde{y}_{it}$ where the implied $|y_{it}| > 1$. We leave such a refinement to future work.

Of course, this observation also implies that Theorem 1.3.3 holds without bounded outcomes $\|\mathbf{Y}\|_\infty \le 1$ and solely with bounded differences $\max_{i,t} |\tilde{y}_{it}| \le 2$.

control achieves logarithmic regret against a different class of DID estimators that we call static DID estimators.[29] Another popular alternative is first-differencing (Abadie, 2021), which by similar arguments may be shown to control regret against a class of *two-period* weighted DID strategies that output $\hat{y}_t(\sigma_{\text{2P-DID}}(\theta)) \equiv y_{0t-1} + \theta' (\mathbf{y}_t - \mathbf{y}_{t-1})$ as successive predictions.

1.3.3 Regularization, covariates, and other extensions

Theorem 1.2.2 shows that synthetic control, as FTL, gives logarithmic regret when we consider quadratic loss. However, to some extent this bound is an artifact of using squared losses, whose curvature ensures that the FTL predictions do not move around excessively over time. If we replace the loss function with the absolute loss $|\hat{y} - y|$, then the regret may be linear in T—no better than that of the trivial prediction $\hat{y}_t \equiv 0$ (see Example 2.10 in Orabona, 2019).

Motivated by the lack of general sublinear regret guarantees in FTL, the online learning literature proposes a large class of algorithms called *Follow-The-Regularized-Leader* (FTRL), where regularization helps stabilize the FTL predictions. With linear prediction functions $f(\mathbf{y};\theta) = \theta'\mathbf{y}$, such strategies take the form

$$\theta_t \in \arg\min_{\theta\in\Theta} \sum_{s<t} \ell(\theta'\mathbf{y}_s, y_{0s}) + \frac{1}{\eta}\Phi(\theta) \tag{1.15}$$

for some convex penalty $\Phi(\cdot)$ and regularization strength $\frac{1}{\eta} > 0$. Here, we let $\ell(\cdot,\cdot)$ denote a generic convex and bounded loss function, generalizing our previous framework. Many regularized variants of synthetic control have been proposed (among others, Chernozhukov *et al.*, 2021; Doudchenko and Imbens, 2016; Hirshberg, 2021). These regularized estimators have the form (1.15), though most such estimators are based on quadratic loss.

Observation 1.3.4. *Regularized synthetic control with penalty $\Phi(\cdot)$ is an instance of FTRL, where $\ell(\cdot,\cdot)$ is typically quadratic loss.*

Moreover, we can think of synthetic control with covariates as regularized synthetic control as

[29] Under certain conditions, Ferman and Pinto (2021) (Proposition 3) show that the demeaned synthetic control in Theorem A.1.3 dominates DID with uniform weighting $\theta_i = 1/N$. The results Theorems A.1.3 and 1.3.3 are in a similar flavor, and show that synthetic control is competitive against DID with any fixed weighting, on average over random assignment of treatment time. Of course, Theorems A.1.3 and 1.3.3 are not generalizations of Ferman and Pinto (2021)'s result—for one, we consider average loss under random treatment timing, and Ferman and Pinto (2021) consider a fixed treatment time under an outcome model, with the number of pre-treatment periods tending to infinity.

well. With time-invariant covariates $\mathbf{x}_j = (x_{1j}, \ldots, x_{Nj})'$ for $j = 1, \ldots, J$, synthetic control may choose weights θ to additionally match the covariates (see, e.g., (7) in Abadie, 2021):

$$\hat{\theta}_{S,x} \in \underset{\theta \in \Theta}{\arg\min} \sum_{t<S} (y_{0t} - \theta' \mathbf{y}_t)^2 + \frac{1}{2\eta} \sum_{j=1}^{J} \eta_j (x_{0j} - \theta' \mathbf{x}_j)^2, \tag{1.16}$$

for some given η_j that indexes the importance of matching covariate j. Observe that, for fixed $x_{0j}, \mathbf{x}_j$, (1.16) is a special case of (1.15); specifically, (1.16) uses a quadratic penalty of the form

$$\Phi(\theta) = \frac{1}{2}(\mathbf{x} - \mathbf{X}\theta)' H (\mathbf{x} - \mathbf{X}\theta)$$

for some positive definite H, vector $\mathbf{x}$, and conformable matrix $\mathbf{X}$. Thus, under the assumption that the covariates $x_{0j}, \mathbf{x}_j$ are fixed and not chosen by the adversary, we may analyze synthetic control with time-invariant covariates as a special case of FTRL.

Motivated by the importance of loss function curvature, we slightly generalize and consider regularized synthetic control estimators using generic loss functions. A standard result in online convex optimization (e.g. Corollary 7.9 in Orabona (2019), Theorem 5.2 in Hazan (2019)) shows that choices of η exist to obtain $\sqrt{T}$ regret.[30] The conditions for this result are highly general, explaining the popularity of FTRL in online convex optimization. We specialize to a few choices of the penalty function Φ in the synthetic control setting; see Theorem A.1.4 for a general statement.

Theorem 1.3.5. *Consider regularized synthetic control (1.15), equivalently FTRL, with penalty function $\Phi(\theta)$ and θ restricted to the simplex Θ. Let $\ell(\theta'\mathbf{y}_t, y_{0t})$ be a convex loss function in θ, not necessarily quadratic, to be specified.*

1. *Consider the quadratic penalty $\Phi(\theta) = \frac{1}{2}(\mathbf{x} - \mathbf{X}\theta)' H (\mathbf{x} - \mathbf{X}\theta)$. Assume the Hessian $\nabla_{\theta\theta'}\Phi(\cdot) = \mathbf{X}' H \mathbf{X}$ is positive definite with minimum eigenvalue normalized to 1. Let $K = \sup_{\theta\in\Theta} \Phi(\theta) - \inf_{\theta\in\Theta} \Phi(\theta)$ be the range of $\Phi(\cdot)$. Then, for both squared loss $\ell(\hat{y}, y) = \frac{1}{2}(y - \hat{y})^2$ and linear loss $\ell(\hat{y}, y) = |y - \hat{y}|$, we have*

$$\text{Regret}_T \leq 2\sqrt{2KNT} \quad \textit{with the choice } \eta = \sqrt{K(2NT)^{-1}}.$$

[30] This rate matches the lower bound for linear losses. See Chapter 5 of Orabona (2019).

Moreover, if $\mathbf{x} = 0$ *and* $\mathbf{X} = H = I$*, then* $\Phi(\theta) = \frac{1}{2}\|\theta\|^2$ *is the ridge penalty, for which we obtain*

$$\text{Regret}_T \leq 2\sqrt{NT} \text{ with the choice } \eta = 1/\sqrt{4NT}.$$

2. *For the entropy penalty* $\Phi(\theta) = \sum_i \theta_i \log \theta_i + \log(N)$*, for both squared and linear losses, we have*

$$\text{Regret}_T \leq 3\sqrt{T \log N} \text{ with the choice } \eta = \sqrt{(\log N)/T}.$$

These results hold for any $N, T > 0$ *and allow for* $T \leq N$*.*

Naturally, these choices of $\Phi(\cdot)$ correspond to regularized variants of synthetic control. As we discuss above, quadratic penalties generalize ridge penalization (Hirshberg, 2021) and matching on covariates.[31] The entropy penalty, which is very natural when the parameters lie on the simplex, is a special case of the proposal in Robbins *et al.* (2017); the resulting regret bound has better dependence on N and obtains the no-regret property as long as $\frac{\log N}{T} \to 0$.[32] For these guarantees, the choice of η does require knowledge on the total number of periods T. This may be relaxed via the "doubling trick" (see Shalev-Shwartz (2011), Section 2.3.1), if we allow for different regularization strengths η_S for different realizations of S.

We conclude this section by pointing out a few other extensions. First, another weakening of the uniform treatment timing requirement can be achieved by considering the maximal regret over subperiods of $[T]$, also known as *adaptive regret*. We show in Section A.2.1 that a modification to the synthetic control algorithm—which still outputs a weighted average of untreated units—achieves worst subperiod regret of order $\log T$. Such a result implies that if *we additionally let the adversary pick a subperiod* of length T', and treatment is uniformly randomly assigned on this subperiod, then modified synthetic control is at most $\frac{\log T}{T'}$-worse on expected loss than the oracle weighted match. Of course,

[31] Ridge penalties are a special case of elastic net penalties proposed by (Doudchenko and Imbens, 2016). Theorem A.1.4 applies to elastic net penalties with nonzero ℓ_2 component as well.

Note that when $\mathbf{X} \in \mathbb{R}^{J \times N}$ represents pre-treatment covariates of the control units, $\mathbf{X}'H\mathbf{X}$ being positive definite requires that the dimension of the covariates is at least the number of control units.

[32] Interestingly, ℓ_1-penalty (proposed by,e.g., Chernozhukov *et al.*, 2021) alone is not strongly convex (See Section 9.1.2 of Boyd and Vandenberghe, 2004), and Theorem A.1.4 does not apply. However, Theorem A.1.4 only contains sufficient conditions, and so this alone is not a criticism of ℓ_1-penalty.

this regret guarantee is meaningful only when the subperiod is sufficiently long, i.e., $T' \gg \log T$. Second, under a design-based framework on treatment timing, we can test sharp hypotheses of the form $H_0 : \mathbf{y}(1) - \mathbf{y}(0) = \mathbf{z}$ by leveraging symmetries induced by random treatment timing. We briefly discuss inference in Section A.2.2.

1.4 Conclusion

This paper notes a simple connection between synthetic control methods and online convex optimization. Synthetic control is an instance of Follow-The-Leader, which are well-studied strategies in the online learning literature. We present standard regret bounds for FTL that apply to synthetic control, which have interpretations as bounds for expected regret under random treatment timing. These regret bounds translate to bounds on expected risk gap under outcome models and imply that synthetic control is competitive against a wide class of matching estimators. In cases where some weighted match of untreated units predict the unobserved potential outcomes, these results show that synthetic control achieves low expected loss. Moreover, the regret bounds can be adapted to be regret bounds against difference-in-differences strategies. Lastly, we draw an analogous connection between regularized synthetic control and Follow-the-Regularized-Leader, a popular class of strategies in online learning.

We now point out a few limitations of this paper and directions for future work. First and foremost, the approach we have taken in this paper is deliberately pessimistic. Living in fear of an adversary constrained solely by bounded outcomes is perhaps too paranoid for sound decision-making. For instance, this worst-case perspective is not particularly amenable to incorporating covariates, since matching on covariates is inherently based on the hope that the covariates are predictive of potential outcomes. Further constraining the adversary (Rakhlin *et al.*, 2011) may be an interesting direction for future research. For instance, it may be fruitful to consider an adversary with a fixed budget for how much $y_{0t}, \mathbf{y}_t$ deviate from $y_{0,t-1}, \mathbf{y}_{t-1}$. Constraining the adversary may also render covariates useful, even in a worst-case framework.

It may also be interesting to consider alternative online protocols. So far, we have considered a thought experiment where, before each step t, the analyst only has access to data $\mathbf{Y}_{1:t-1}$ to output a prediction function. In practice, the analyst typically does have access to $\mathbf{y}_1, \ldots, \mathbf{y}_T$. Alternative protocols have been considered in the online learning literature. One example is the Vovk–Azoury–Warmuth forecaster (See Section 7.10 in Orabona, 2019), where we assume the analyst additionally has access to $\mathbf{y}_t$ before they are prompted for a prediction at time t. In this case, regularized strategies can also achieve $\log T$ regret. Additionally, Bartlett *et al.* (2015) consider the fixed design setting in which $\mathbf{y}_1, \dots, \mathbf{y}_T$ is fully accessible to the analyst before they are prompted for a prediction. Bartlett *et al.* (2015) give a simple and explicit minimax regret strategy for online linear regression, which we may adapt into a synthetic control estimator.

We have only considered regret on one-step-ahead prediction for y_{0S}, but synthetic control estimates are often extrapolated multiple time periods ahead in practice. In attempting to extend our results to k-step-ahead prediction, it is natural to consider $\check{y}_{it} = (y_{it}, \dots, y_{i,t+k})$, and to attempt a similar argument on $\check{\mathbf{Y}}$. The chief difficulty in doing so is one of delayed feedback, where the analyst cannot update their time-S decision based on loss from times $1, \dots, S-1$. That is, for k-step-ahead prediction, the analyst, viewed as an online player who is prompted for a forecast of $\check{y}_{0,S} = (y_{0S}, y_{0,S+1}, \dots, y_{0,S+k-1})$, does not have access to their prediction loss for $\check{y}_{0,S-1} = (y_{0,S-1}, y_{0S}, \dots, y_{0,S+k-2})$, since $y_{0,S+k-2}$ is not yet observed. As a result, unlike (O3) in the standard online convex optimization protocol, the analyst does not have access to $\ell_1(\cdot), \dots, \ell_{S-1}(\cdot)$ when making decisions θ_S—rendering our results here insufficient. That said, delayed feedback—where the online player only has knowledge of the loss function after k periods—is studied in online learning (Weinberger and Ordentlich, 2002; Korotin *et al.*, 2018; Flaspohler *et al.*, 2021), and we leave an exploration to future work.

Chapter 2

Empirical Bayes when estimation precision predicts parameters[1]

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

Dissertation Advisor:
Professor Isaiah Andrews

Author:
Jiafeng Chen

Essays in Econometrics

Abstract

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

2.1 Introduction

Applied economists often use empirical Bayes methods to shrink noisy parameter estimates, in hopes of accounting for the imprecision in the estimates and improving subsequent policy decisions.[2] The textbook empirical Bayes method assumes *prior independence*—that the precisions of the noisy estimates do not predict the underlying unknown parameters. However, prior independence is economically questionable and empirically rejected in many contexts. This is frequently because sample sizes associated with the estimates either *select on* or *affect* the underlying parameters, rendering the resulting standard errors highly predictive of the parameters.[3] Inappropriately imposing prior independence can harm empirical Bayes decisions, possibly even making them underperform decisions without using shrinkage. Motivated by these concerns, this paper introduces empirical Bayes methods that relax prior independence.

To be concrete, our primary empirical example (Bergman *et al.*, 2023) computes empirical Bayes posterior means for economic mobility estimates of low-income children[4] published in the Opportunity

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

In terms of policy decisions driven by empirical Bayes posterior means, Gilraine *et al.* (2020) report that by the end of 2017, 39 states require that teacher value-added measures—typically, empirical Bayes posterior means of teacher performance—be incorporated into the teacher evaluation process.

[3] To see this, take value-added modeling as an example. The precision of value-added estimates is usually a function of the number of customers associated with a service provider (e.g. number of students for a teacher). It is possible that customers select into higher quality providers. It is also possible that congestion effects render more popular service providers worse. These channels predict that the sample sizes for a provider are associated with latent value-added, and the direction of association depends on the interplay of the selection and congestion effects. Section B.1.5 outlines a formal discrete choice model to illustrate these effects. Potential failure of prior independence is noted by, among others, Bruhn *et al.* (2022), Kline *et al.* (2023), George *et al.* (2017), and Mehta (2019).

[4] Throughout this paper, measures of economic mobility are defined as certain average outcomes of children from low-income households. There are various definitions of economic mobility provided by Chetty *et al.* (2020), discussed later in the paper. They are all measures of economic outcomes for children from low-income households (households at the 25th percentile of the national income distribution). One example is the probability that a Black person have incomes in the top

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 economic mobility. Consequently, the standard errors of the estimates and true economic mobility are positively correlated, violating prior independence.

Bergman *et al.* (2023) use empirical Bayes posterior means to select high-mobility Census tracts, choosing those with high estimated posterior means. Using a validation procedure that we develop, for a few measures of economic mobility where prior independence is severely violated, we find that screening on conventional empirical Bayes posterior means selects *less* economically mobile tracts, on average, than screening on the unshrunk estimates.[5] In contrast, screening on empirical Bayes posterior means computed by our method selects substantially more mobile tracts.

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

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

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

We relax prior independence by modeling the prior distribution $\theta_i \mid \sigma_i$ flexibly, detailed in Section 2.2. We model $\theta_i \mid \sigma_i$ as a conditional location-scale family, controlled by σ_i-dependent

20 percentiles, whose parents have household incomes at the 25th percentile. As another example, Bergman *et al.* (2023) measure economic mobility as the mean income rank of children growing up in households at the 25th income percentile.

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

location and scale hyperparameters and a σ_i-independent shape hyperparameter. Under this assumption, different values of the standard errors σ_i translate, compress, or dilate the distribution of the parameters $\theta_i \mid \sigma_i$, but the underlying shape of $\theta_i \mid \sigma_i$ does not vary. This model subsumes prior independence as the special case where the unknown location and scale parameters are constant functions of σ_i.

This conditional location-scale assumption leads naturally to a family of empirical Bayes methods that we call CLOSE. Since the unknown prior distribution $\theta_i \mid \sigma_i$ is fully described by its location, scale, and shape hyperparameters, CLOSE estimates these parameters flexibly and plugs the estimated parameters into downstream decision rules. Among different estimation strategies for the hyperparameters, our preferred specification of CLOSE uses nonparametric maximum likelihood (NPMLE, Kiefer and Wolfowitz, 1956; Koenker and Mizera, 2014) to estimate the unknown shape of the prior distribution $\theta_i \mid \sigma_i$. We find that CLOSE-NPMLE inherits the favorable computational and theoretical properties of NPMLE documented in the literature (Soloff *et al.*, 2021; Jiang, 2020; Polyanskiy and Wu, 2020).

Section 2.3 provides three statistical guarantees for CLOSE-NPMLE. First and foremost, CLOSE-NPMLE emulates the oracle as well as possible, at least in terms of squared error loss. Specifically, Theorems 2.3.4 and 2.3.5 establish that CLOSE-NPMLE is minimax rate-optimal—up to logarithmic factors and under the conditional location-scale assumptions—for *Bayes regret in squared error*, a standard performance metric (Jiang and Zhang, 2009). Bayes regret is the performance gap between CLOSE-NPMLE and oracle Bayes decisions made with knowledge of the distribution of (θ_i, σ_i).

Second, our guarantee for squared error regret also controls the Bayes regret for two ranking-related decision problems, including the problem of selecting high-mobility tracts encountered by Bergman *et al.* (2023). Theorem 2.3.7 shows that the Bayes regret in squared error dominates the Bayes regret for these decision problems. Thus, these ranking-related problems are easier than squared error estimation, and our squared error regret result implies upper bounds for the regrets of these problems.

Third, to assess robustness of CLOSE to the location-scale modeling assumption, Theorem 2.3.10 establishes that CLOSE-NPMLE is worst-case robust. Without imposing the location-scale assumptions, for a population version of CLOSE-NPMLE, we show that its worst-case mean-squared error is a bounded multiple of that of the minimax procedure. Since the minimax procedure optimizes its worst-case risk, this result shows that CLOSE-NPMLE does not perform exceedingly poorly even when the location-scale model is misspecified.

Since practitioners may want to assess how and whether CLOSE-NPMLE provides improvements in specific applications, Section 2.4.3 produces an out-of-sample validation procedure by extending the *coupled bootstrap* in Oliveira *et al.* (2021). If one had access to the micro-data, one could split the data into training and testing samples, use one to compute decisions, and use the other to evaluate them. Our validation procedure emulates this sample-splitting without needing access to the underlying micro-data. It provides unbiased loss estimates for any decision rules. In particular, this procedure allows practitioners to evaluate whether CLOSE provides improvements for their setting by comparing loss estimates for CLOSE and those for the standard shrinkage procedure.

To illustrate our method, Section 2.5 applies CLOSE to two empirical exercises, building on Chetty *et al.* (2020) and Bergman *et al.* (2023). The first exercise is a calibrated Monte Carlo simulation, in which we have access to the true distribution of (θ_i, σ_i). We find that CLOSE-NPMLE has mean-squared error (MSE) performance close to that of the oracle posterior, uniformly across the 15 measures of economic mobility that we include. For all 15 measures, CLOSE-NPMLE captures over 90% of possible MSE gains relative to no shrinkage, whereas conventional shrinkage captures only 70% on average and as little as 40% for some measures.

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

2.2 Model and proposed method

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

$$Y_i \mid \theta_i, \sigma_i^2 \sim \mathcal{N}(\theta_i, \sigma_i^2) \qquad i = 1, \ldots, n. \tag{2.2}$$

The Normality in (2.2) is motivated by the central limit theorem applied to the underlying micro-data that generate the estimates Y_i. That is, let n_i denote the underlying sample size in the micro-data which generate (Y_i, σ_i). Standard large-sample approximation implies

$$\frac{Y_i - \theta_i}{\sigma_i} \xrightarrow{d} \mathcal{N}(0, 1) \tag{2.3}$$

as $n_i \to \infty$.[6]

We also assume that the population parameters (θ_i, σ_i) are sampled from some joint distribution. Throughout this paper, we condition on $\sigma_{1:n} = (\sigma_1, \ldots, \sigma_n)$ and treat them as fixed. We assume that (θ_i, σ_i) are independently and identically drawn,[7] but the conditional distribution $\theta_i \mid \sigma_i$ may be different across σ_i:

$$\theta_i \mid \sigma_i \overset{\text{i.n.i.d.}}{\sim} G_{(i)}. \tag{2.4}$$

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

These assumptions imply that the Bayes decision rule with respect to the oracle Bayes prior P_0 is optimal (Lehmann and Casella, 2006). Consider a loss function $L(\boldsymbol{\delta}, \theta_{1:n})$, which evaluates an

[6] Note that, under standard assumptions, the approximation (2.3) holds regardless of whether σ_i is an estimated standard error or its unknown population counterpart. This is because the estimation error in σ_i is typically of order $1/n_i$, which is smaller than that in Y_i, which is of order $1/\sqrt{n_i}$.

[7] Combined with the independence assumption of Y_i across i, we assume that $(\theta_i, \sigma_i, Y_i)$ are independently drawn unconditionally. The independence assumption for the estimates Y_i conditional on (θ_i, σ_i) holds when the underlying micro-data for different estimates Y_i are sampled independently. This assumption does not precisely hold for the Opportunity Atlas, but the correlation between Y_i and Y_j, which arises from individuals who move between tracts, is likely small. Papers imposing this assumption include Mogstad *et al.* (2020) and Andrews *et al.* (2023). Moreover, we discuss an interpretation of the procedure when we erroneously assume that Y_i and/or θ_i are independent across i in Section B.1.6.

action $\boldsymbol{\delta}$ at a vector of parameters $\theta_{1:n}$. For instance, in our empirical application, the loss function may measure how well we estimate true mobility $\theta_{1:n}$ or how well we select high mobility Census tracts.[8] At any realization of the data $(Y_{1:n}, \sigma_{1:n})$, the *oracle Bayes decision rule* $\boldsymbol{\delta}^\star$ picks an action that minimizes the posterior expected loss:

$$\boldsymbol{\delta}^\star(Y_{1:n}, \sigma_{1:n}; P_0) \in \arg\min_{\boldsymbol{\delta}} E_{P_0}[L(\boldsymbol{\delta}, \theta_{1:n}) \mid Y_{1:n}, \sigma_{1:n}]. \tag{2.5}$$

Empirical Bayesians seek to approximate the oracle Bayes rule $\boldsymbol{\delta}^\star$ (Efron, 2014). With an estimate $\hat{P}$ for P_0, it is natural to plug $\hat{P}$ into (2.5):[9]

$$\boldsymbol{\delta}_{\mathrm{EB}}(Y_{1:n}, \sigma_{1:n}; \hat{P}) \in \arg\min_{\boldsymbol{\delta}} \mathbf{E}_{\hat{P}}[L(\boldsymbol{\delta}, \theta_{1:n}) \mid Y_{1:n}, \sigma_{1:n}]. \tag{2.6}$$

Popular empirical Bayes methods impose more structure than (2.4) in order to simplify estimating P_0.[10] The standard parametric empirical Bayes method additionally models $G_{(i)}$ as identical across i and Gaussian: i.e., for all i, $G_{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(m_0, s_0^2)$ (Morris, 1983). Following the recipe (2.6), this approach estimates the prior parameters (m_0, s_0^2). Henceforth, we shall refer to this method as INDEPENDENT-GAUSS. On the other hand, state-of-the-art empirical Bayes methods (Jiang, 2020; Soloff *et al.*, 2021; Jiang and Zhang, 2009; Koenker and Gu, 2019; Gilraine *et al.*, 2020) assume that the marginal distributions are equal to some common, unknown distribution $G_{(0)}$, not necessarily Gaussian: i.e., for all i, $G_{(i)} \overset{\text{i.i.d.}}{\sim} G_{(0)}$. They estimate $G_{(0)}$ with nonparametric maximum likelihood and form decision rules according to (2.6). We refer to this method as INDEPENDENT-NPMLE. The "INDEPENDENT" here emphasizes that these methods assume *prior independence*: $\theta_i \perp\!\!\!\perp \sigma_i$ under the

[8] We formalize the sense of optimality and formalize three decision problems in Section 2.2.3.

[9] To emphasize the distinction between the true expectation with respect to the data-generating process (2.4) and a posterior mean taken with respect to some possibly estimated measure $\hat{P}$, we shall use E to refer to the former and $\mathbf{E}$ to refer to the latter. Subscripts typically make the distinction clear as well. Specifically,

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

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

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

prior P_0.

We relax prior independence by instead modeling $\theta_i \mid \sigma_i$ as a location-scale family,[11] indexed by unknown hyperparameters $(m_0(\cdot), s_0(\cdot), G_0(\cdot))$: Specifically, we assume

$$P\left(\theta_i \leq t \mid \sigma_i\right) = G_0\left(\frac{t - m_0(\sigma_i)}{s_0(\sigma_i)}\right), \tag{2.7}$$

where the distribution G_0 is normalized to have zero mean and unit variance. Under (2.7), different values of σ may translate, compress, or dilate the conditional distribution of $\theta \mid \sigma$ via the location parameter $m_0(\cdot)$ and the scale parameter $s_0(\cdot)$, but the conditional distributions can be normalized to take the same shape $G_0(\cdot)$. Under this model, the oracle prior distribution P_0 is fully described by the hyperparameters $(m_0(\cdot), s_0(\cdot), G_0(\cdot))$. Our method, CLOSE, proposes to estimate P_0 with an estimate $\hat{P}$ derived from estimated hyperparameters $(\hat{m}(\cdot), \hat{s}(\cdot), \hat{G}_n)$. CLOSE then produces empirical Bayes decision rules with respect to the estimated prior $\hat{P}$, following the recipe (2.6).

Before specifying our procedure in detail in Section 2.2.2, we illustrate with an example where prior independence fails and show what happens to empirical Bayes decision rules that inappropriately impose prior independence.

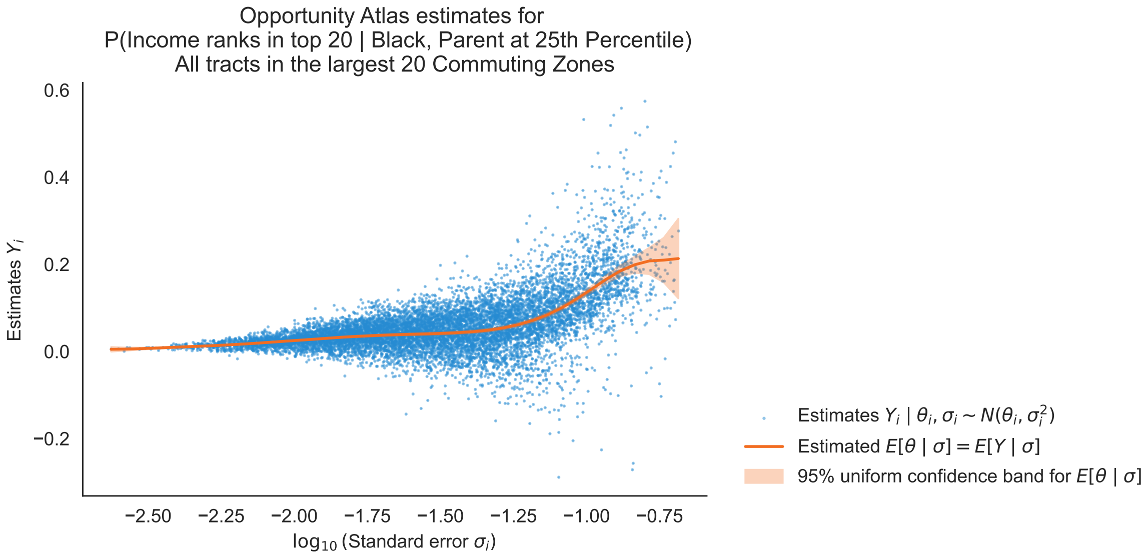
2.2.1 Plausibility of prior independence

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

Prior independence is readily rejected for this measure of economic mobility. Figure 2.1 plots Y_i against $\log_{10}(\sigma_i)$ and imposes a nonparametric regression estimate of the conditional mean function

[11] We explore alternatives to the location-scale model in Section B.1.7. We find that no alternative provides a free-lunch improvement over our assumptions.

More restrictive forms of this assumption also appear in the past and concurrent literature. For instance, Kline *et al.* (2023) model the dependence as a pure scale model $\theta \mid \sigma \sim s(\sigma) \cdot \tau$ for some $\tau \mid \sigma \overset{\text{i.i.d.}}{\sim} G$ (with additional parametric restrictions on $s(\cdot)$) and George *et al.* (2017) impose the location scale model (2.7) with $G_0 \sim \mathcal{N}(0, 1)$ (as well as additional parametric restrictions on $s_0(\cdot), m_0(\cdot)$).



Notes. All tracts within the largest 20 Commuting Zones (CZs) are shown. Due to the regression specification in Chetty *et al.* (2020), point estimates of $\theta_i \in [0, 1]$ do not always lie within $[0, 1]$. The orange line plots nonparametric regression estimates of the conditional mean $E[Y \mid \sigma] = E[\theta \mid \sigma] \equiv m_0(\sigma)$, estimated via local linear regression with automatic bandwidth selection implemented in Calonico *et al.* (2019). The orange shading shows a 95% uniform confidence band, constructed by the max-t confidence set over 50 equally spaced evaluation points. The confidence band excludes any constant function. See Section B.7 for details on estimating conditional moments of θ_i given σ_i. □

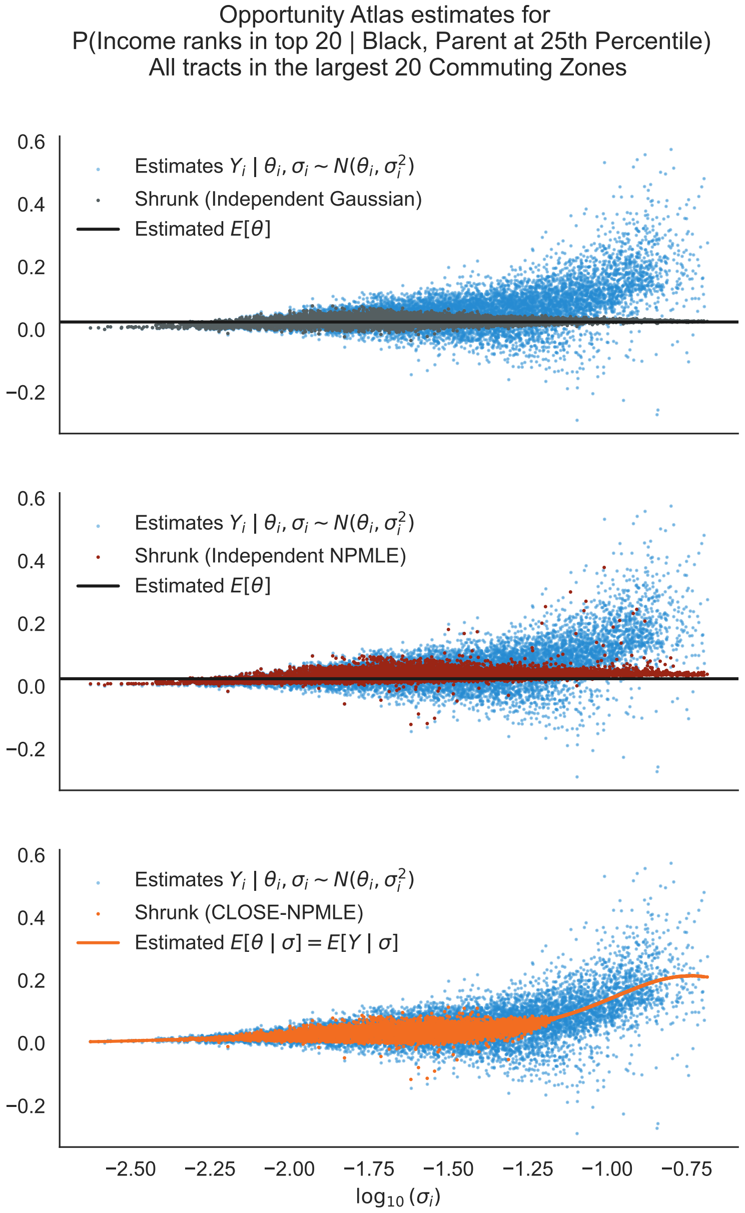
Figure 2.1: Scatter plot of Y_i against $\log_{10}(\sigma_i)$ in the Opportunity Atlas

$m_0(\sigma_i) \equiv E[\theta_i \mid \sigma_i] = E[Y_i \mid \sigma_i]$. If θ_i were independent of σ_i, then the true conditional mean function $m_0(\sigma_i)$ should be constant. Figure 2.1 shows the contrary—tracts with more imprecisely estimated Y_i tend to have higher economic mobility.[12]

This correlation is in part through the following channel. Since θ_i is an average outcome for children from poor Black families, tracts with more poor Black families tend to have more precise estimates of θ_i.[13] However, these tracts also tend to have lower economic mobility θ_i due to the pernicious effects of residential segregation.

[12] Moreover, $\log \sigma_i$ remains predictive of Y_i even if we residualize Y_i against a vector of tract-level covariates (Figure B.9). Prior independence is also readily rejected for the mobility measure used in Bergman *et al.* (2023), but its violation is not as severe once adjusted for tract-level covariates (see Section 2.5 and Figure B.8).

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



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

Figure 2.2: Posterior mean estimates under prior independence

What happens if we apply empirical Bayes methods that assume prior independence here? Figure 2.2 overlays empirical Bayes posterior means on the Y_i-against-$\log \sigma_i$ scatterplot. In the top panel, INDEPENDENT-GAUSS shrinks estimates Y_i towards a common estimated mean $\hat{m}_0$, depicted as the black line. INDEPENDENT-GAUSS shrinks noisier estimates more aggressively. When σ_i and θ_i are positively correlated—as is the case here—estimated posterior means under INDEPENDENT-GAUSS systematically undershoot θ_i for populations with imprecise estimates. Similarly, the middle panel of Figure 2.2 shows that INDEPENDENT-NPMLE suffers from the same undershooting, though less so. In contrast, the bottom panel of Figure 2.2 previews our preferred procedure, CLOSE-NPMLE, which shrinks towards the conditional mean $E[\theta_i \mid \sigma_i]$, thus avoiding the undershooting.

This undershooting is particularly problematic if one would like to select high-mobility Census tracts. These high-mobility tracts are exactly those with high imprecision σ_i, owing to the positive correlation between θ_i and σ_i. By shrinking these tracts severely towards the estimated common mean, empirical Bayes under prior independence makes suboptimal selections that may even underperform screening directly based on Y_i.[14]

For a given empirical context, prior independence can always be checked empirically by plotting à la Figure 2.1. Nevertheless, we discuss the general plausibility of prior independence in the following remark.

Remark 2.2.1 (Plausibility of prior independence)**.** To describe the general channels underlying the potential failure of prior independence, let us write (2.3) in a different form

$$\sqrt{n_i}(Y_i - \theta_i) \xrightarrow{d} \mathcal{N}(0, \sigma_{0i}^2) \quad \text{where } \sigma_i \approx \frac{\sigma_{0i}}{\sqrt{n_i}}. \tag{2.8}$$

Expression (2.8) decomposes the (estimated) standard error into the underlying sample size n_i in the micro-data and the asymptotic variance σ_{0i}^2 of the (properly scaled) estimator. Both n_i and σ_{0i} may predict θ_i in a variety of empirical contexts.

Let us start with the implicit sample sizes n_i. It is possible that n_i is in part determined by θ_i, which we loosely term *selection*. In value-added modeling, n_i is the number of observations associated with a provider. It is possible that n_i selects on the latent quality θ_i of that provider. For instance,

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

Chandra *et al.* (2016) find "higher quality hospitals have higher market shares and grow more over time." If market share and hospital size relate to the underlying sample size n_i (e.g. number of patient observations) for estimating hospital value-added, then this suggests non-independence between θ_i and σ_i (see George *et al.* (2017) for some empirical evidence). As another example, in meta-analysis, suppose θ_i represents the treatment effect of some intervention i. If researchers power studies based on informative priors for θ_i, then we should observe that interventions with larger conjectured effect sizes have smaller sample sizes n_i.

Another channel driving the correlation between n_i and θ_i can be loosely termed *congestion*, where n_i affects the latent feature θ_i. For our primary application, n_i represents the number of poor and minority households in a Census tract, and θ_i represents underlying economic or social mobility. Places with more poor and minority households experience white flight and residential segregation (Cutler *et al.*, 1999; Agan and Starr, 2020; Kain, 1968), develop oppressive institutions (Derenoncourt, 2022; Alesina *et al.*, 2001), and provide worse public goods (Laliberté, 2021; Jackson and Mackevicius, 2021; Colmer *et al.*, 2020). These factors contribute to lower economic mobility θ_i. Section B.1.5 contains more examples of violation of prior independence and outlines a model in which selection and congestion effects drive correlation between n_i and θ_i.

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

2.2.2 Conditional location-scale relaxation of prior independence

Having argued that (i) prior independence is theoretically suspect and empirically rejected and that (ii) inappropriately imposing it can harm empirical Bayes decision rules, we propose the conditional location-scale model (2.7) as a relaxation.[15] Here, we state the location-scale assumption (2.7) equivalently as the following representation with transformed parameters $\tau_i = \frac{\theta_i - m_0(\sigma_i)}{s_0(\sigma_i)}$:

$$\theta_i = m_0(\sigma_i) + s_0(\sigma_i)\tau_i \qquad \tau_i \mid \sigma_i \overset{\text{i.i.d.}}{\sim} G_0 \qquad E_{G_0}[\tau_i] = 0 \qquad \mathrm{Var}_{G_0}(\tau_i) = 1. \tag{2.9}$$

To estimate P_0 under (2.9), it suffices to estimate the unknown hyperparameters (m_0, s_0, G_0). Expression (2.9) makes clear that, under the location-scale model, the transformed parameter $\tau_i \sim G_0$ is independent from σ_i. Analogously, let $Z_i = \frac{Y_i - m_0(\sigma_i)}{s_0(\sigma_i)}$ be the transformed estimates and $\nu_i = \frac{\sigma_i}{s_0(\sigma_i)}$ be their standard errors.

Crucially, (Z_i, τ_i, ν_i) obey an analogue of the Gaussian location model (2.2) in which prior independence holds:

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

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

This strategy is still infeasible, since the transformation depends on unknown location and scale parameters $\eta_0 \equiv (m_0, s_0)$. Fortunately, $m_0(\cdot)$ and $s_0(\cdot)$ are readily estimable from the data (Y_i, σ_i), as they only require conditional expectations and variances of Y given σ:

$$m_0(\sigma) = E[\theta \mid \sigma] = E[Y \mid \sigma] \quad \text{and} \quad s_0^2(\sigma) = \mathrm{Var}(\theta \mid \sigma) = E[(Y - m_0(\sigma))^2 \mid \sigma] - \sigma^2. \tag{2.10}$$

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

$$\hat{Z}_i = \frac{Y_i - \hat{m}(\sigma_i)}{\hat{s}(\sigma_i)} \quad \text{and} \quad \hat{\nu}_i = \frac{\sigma_i}{\hat{s}(\sigma_i)}. \tag{2.11}$$

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:[16]

[15] In the presence of covariates X_i—which do not predict the noise in Y_i, $Y_i \perp\!\!\!\perp X_i \mid \theta_i, \sigma_i$—the assumption (2.7) can be modified to accommodate additional covariates as well. We provide additional discussion of covariates in Section B.1.6.

[16] We give a more detailed walkthrough of these steps in Section 2.4. We also detail a local linear regression estimator in Section B.7 for $\boxed{\text{CLOSE–STEP } \mathbf{1}}$.

CLOSE–STEP 1 Nonparametrically estimate $m_0(\sigma), s_0^2(\sigma)$ according to (2.10).

CLOSE–STEP 2 With the estimates $\hat{\eta} = (\hat{m}, \hat{s})$, transform the data according to (2.11). Apply empirical Bayes methods with prior independence to estimate G_0 with some $\hat{G}_n$ on the transformed data $(\hat{Z}_i, \hat{\nu}_i)$.

CLOSE–STEP 3 Having estimated $(\hat{\eta}, \hat{G}_n)$, which implies an estimate $\hat{P}$ of P_0, we then form empirical Bayes decision rules following (2.6).

This framework produces a family of empirical Bayes strategies, since **CLOSE–STEP 2** can take different forms. To leverage theoretical and computational advances, we will focus on—and recommend—using nonparametric maximum likelihood (NPMLE) to estimate G_0. That is, we maximize the log-likelihood of (an estimated version of) the transformed data Z_i, whose marginal distribution is the convolution $G_0 \star \mathcal{N}(0, \nu_i^2)$:[17]

$$\hat{G}_n \in \operatorname*{arg\,max}_{G \in \mathcal{P}(\mathbb{R})} \frac{1}{n} \sum_{i=1}^{n} \log \int_{-\infty}^{\infty} \varphi\left(\frac{\hat{Z}_i - \tau}{\hat{\nu}_i}\right) \frac{1}{\hat{\nu}_i} G(d\tau). \tag{2.12}$$

When the estimated moments $\hat{m}, \hat{s}$ are constant functions of σ, CLOSE-NPMLE estimates the same prior as INDEPENDENT-NPMLE. In the theoretical literature, under prior independence, INDEPENDENT-NPMLE is state-of-the-art in terms of computational ease and regret properties.[18] Our subsequent results in Section 2.3 extend some of these favorable properties to CLOSE-NPMLE under the conditional location-scale model.

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

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

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

the following remark.

Remark 2.2.2 (CLOSE-GAUSS)**.** Under $G_0 \sim \mathcal{N}(0,1)$, the oracle Bayes posterior means are simply

$$\theta^*_{i,\mathcal{N}(0,1),\eta_0} = \frac{\sigma_i^2}{s_0^2(\sigma_i) + \sigma_i^2} m_0(\sigma_i) + \frac{s_0^2(\sigma_i)}{s_0^2(\sigma_i) + \sigma_i^2} Y_i. \tag{2.13}$$

Equation (2.13) is the analogue of posterior means estimated by INDEPENDENT-GAUSS, where the unconditional mean m_0 and variance s_0^2 are replaced with their conditional counterparts $(m_0(\cdot), s_0^2(\cdot))$. As an empirical Bayes strategy, CLOSE-GAUSS then replaces the unknown conditional moments with their estimated counterparts.[19] Its properties depend on those of the oracle (2.13) it mimics, which we turn to now.

Despite being rationalized under the assumption $\theta_i \mid \sigma_i \sim \mathcal{N}(m_0(\sigma_i), s_0^2(\sigma_i))$, (2.13) enjoys strong robustness properties: It is optimal over a restricted class of decision rules and minimax over all decision rules—without imposing the location-scale assumption (2.7). First, (2.13) is the optimal decision rule for estimating θ_i when we restrict to the class of decision rules that are linear in Y_i (Weinstein *et al.*, 2018). Second, (2.13) is minimax in the sense that it minimizes the worst-case mean squared error, where an adversary chooses $G_{(1)}, \ldots, G_{(n)}$, subjected to the constraint that $G_{(i)}$'s first two moments are $(m_0(\sigma_i), s_0^2(\sigma_i))$.[20]

However, the Normality assumption does imply that (2.13), unlike CLOSE-NPMLE, fails to approximate the optimal decision (2.5) when the location-scale assumption (2.7) holds but $\theta_i \mid \sigma_i$ may not be Gaussian. Since we also show that CLOSE-NPMLE is worst-case robust—though with higher worst-case risk than CLOSE-GAUSS, we recommend CLOSE-NPMLE over CLOSE-GAUSS, unless the researcher is extremely concerned about the misspecification of the location-scale model. ■

[19] (2.13) is first proposed by Weinstein *et al.* (2018). Weinstein *et al.* (2018) propose estimating $m_0(\cdot), s_0(\cdot)$ in a particular manner to ensure the resulting empirical Bayes posterior means dominate the naive estimates Y_i uniformly over $\theta_{1:n}, \sigma_{1:n}$, which are conditioned upon.

[20] Formally,

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

where the supremum is taken over $G_{(i)}$ having moments $\eta_0(\sigma_i)$. To wit, note that the Bayes risk of (2.13) is the same regardless of choices of $G_{(1)}, \ldots, G_{(n)}$ under the moment constraint, and it is equal to the optimal Bayes risk when $G_{(i)} \sim \mathcal{N}(m_0(\sigma_i), s_0^2(\sigma_i))$. We therefore conclude that (2.13) is minimax by observing that the minimax Bayes risk is at least the risk of (2.13).

2.2.3 Decision problems

To prepare for our theoretical results in the next section, we close this one by introducing decision theory notation and formalizing a few decision problems. Let $\boldsymbol{\delta}(Y_{1:n}, \sigma_{1:n})$ be a *decision rule* mapping the data $(Y_{1:n}, \sigma_{1:n})$ to *actions*. Let $L(\boldsymbol{\delta}, \theta_{1:n})$ denote a *loss function* mapping actions and parameters to a scalar. Let $R_{\mathrm{F}}(\boldsymbol{\delta}, \theta_{1:n}) = E[L(\boldsymbol{\delta}, \theta_{1:n}) \mid \theta_{1:n}, \sigma_{1:n}]$ denote the *frequentist risk* associated with the loss function L, which integrates over the randomness in $Y_{1:n}$, keeping $\theta_{1:n}, \sigma_{1:n}$ fixed. Finally, let $R_{\mathrm{B}}(\boldsymbol{\delta}; P_0) = E_{P_0}[R_{\mathrm{F}}(\boldsymbol{\delta}, \theta_{1:n}) \mid \sigma_{1:n}]$ be the *Bayes risk* of $\boldsymbol{\delta}$ under P_0, which additionally integrates over the conditional distribution $\theta_{1:n} \mid \sigma_{1:n}$.[21]

The oracle Bayes decision rule $\boldsymbol{\delta}^\star$ (2.5) is optimal in the sense that it minizes R_{B}. A natural metric of success for the empirical Bayesian (2.6) is thus the gap between the Bayes risks of $\boldsymbol{\delta}_{\mathrm{EB}}$ and $\boldsymbol{\delta}^\star$. We refer to this quantity as *Bayes regret*:

$$\mathrm{BayesRegret}_n(\boldsymbol{\delta}_{\mathrm{EB}}) = R_{\mathrm{B}}(\boldsymbol{\delta}_{\mathrm{EB}}; P_0) - R_{\mathrm{B}}(\boldsymbol{\delta}^\star; P_0) = E[L(\boldsymbol{\delta}_{\mathrm{EB}}, \theta_{1:n}) - L(\boldsymbol{\delta}^\star, \theta_{1:n}) \mid \sigma_{1:n}] \quad (2.14)$$

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.[22]

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

Decision Problem 1 (Squared-error estimation of $\theta_{1:n}$)**.** *The canonical statistical problem (Robbins, 1956) is estimating the parameters $\theta_{1:n}$ under mean-squared error (MSE). That is, the action $\boldsymbol{\delta}$ =*

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

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

$(\delta_1, \ldots, \delta_n)$ *collects estimates* δ_i *for parameters* θ_i*, evaluated with MSE:*

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

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

$$\delta_i^\star = \theta_{i,P_0}^* \equiv E_{P_0}\left[\theta_i \mid Y_i, \sigma_i\right]$$

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

Next, we describe two problems that are likely more relevant for policy-making, such as replacing low value-added teachers and recommending high economic mobility tracts (Gilraine *et al.*, 2020; Bergman *et al.*, 2023).[23]

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

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

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

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

One natural empirical Bayes decision rule replaces θ_{i,P_0}^* *with* $\theta_{i,\hat{P}}^*$*, following (2.6).*

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

[23] We analyze these problems from a decision-theoretic perspective, under the sampling assumption (2.4). For a different and complementary perspective in terms of conditional-on-θ frequentist inference on ranks, see Mogstad *et al.* (2020, 2023). For additional ranking-related decision problems, see Gu and Koenker (2023).

average treatment effect θ_i, and the cost of treatment is c_i.[24] ■

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

$$-L(\boldsymbol{\delta}, \theta_{1:n}) = \frac{1}{m} \sum_{i=1}^{n} \delta_i \theta_i. \tag{2.15}$$

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

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

Again, the empirical Bayes recipe (2.6) suggests replacing P_0 with the estimate $\hat{P}$.

The utility function (2.15) rationalizes the widespread practice of screening based on empirical Bayes posterior means. For instance, this objective may be reasonable for rewarding the top 5% of teachers or replacing the bottom 5%, according to value-added (Gilraine et al.*, 2020; Chetty* et al.*, 2014a; Kane and Staiger, 2008; Hanushek, 2011). In Bergman* et al. *(2023), where housing voucher holders are incentivized to move to Census tracts selected according to economic mobility, (2.15) represents the expected economic mobility of a mover if they move randomly to one of the selected tracts.*[25] ■

2.3 Regret results for CLOSE-NPMLE

We observe $(Y_i, \sigma_i)_{i=1}^n$, where (θ_i, σ_i) satisfies the location-scale assumption (2.7) and $(Y_i, \theta_i, \sigma_i)$ obeys the Gaussian location model (2.2). Our recommended procedure, CLOSE-NPMLE, transforms the data (Y_i, σ_i) into $(\hat{Z}_i, \hat{\nu}_i)$, with estimated nuisance parameters $\hat{\eta} = (\hat{m}, \hat{s})$ for $\eta_0 = (m_0, s_0)$ in **CLOSE–STEP 1**. It then estimates the unknown shape parameter G_0 via NPMLE (2.12) on $(\hat{Z}_i, \hat{\nu}_i)_{i=1}^n$.

Our leading result shows that CLOSE-NPMLE mimics the oracle Bayesian as well as possible, for

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

[25] Our theoretical results in Section 2.3.2 can accommodate a slightly more general decision problem, which allows for an expected mobility interpretation for movers who do not move uniformly randomly. See Theorem 2.3.8.

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

Since our main result assumes the location-scale model, one may be concerned about its potential misspecification. The last result in this section, Theorem 2.3.10, bounds the worst-case Bayes risk of an idealized version of CLOSE-NPMLE (i.e. with known η_0 and fixed but misspecified $\hat{G}_n$) as a multiple of a notion of minimax risk, without assuming (2.7). Thus, even under misspecification, CLOSE-NPMLE does not perform arbitrarily badly relative to the minimax procedure.

The rest of this section states and discusses these results formally. Practitioners who are less interested in the theoretical details are free to skip to Section 2.4, where we discuss a number of practical considerations.

Remark 2.3.1 (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. In logical statements, appearances of $\lesssim$ implicitly prepend "there exists a universal constant" to the statement.[26] 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 $E[\cdot], P(\cdot)$ to denote the expectation and probability over $\theta_{1:n} \mid \sigma_{1:n} \sim P_0$. We omit the P_0 subscript and the conditioning on $\sigma_{1:n}$. ■

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

2.3.1 Regret rate in squared error

Since we consider CLOSE-NPMLE in mean-squared error, we define

$$\operatorname{Regret}(G, \eta) \equiv \frac{1}{n}\sum_{i=1}^n (\hat\theta_{i,G,\eta} - \theta_i)^2 - \frac{1}{n}\sum_{i=1}^n (\theta_i^* - \theta_i)^2$$

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

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

$$\operatorname{BayesRegret}_n = E\left[\operatorname{Regret}(\hat G_n, \hat\eta)\right] = E\left[\frac{1}{n}\sum_{i=1}^n (\theta_i^* - \hat\theta_{i,\hat G_n, \hat\eta})^2\right]. \tag{2.16}$$

Equation (2.16) additionally notes that expected Regret is equal to the expected mean-squared difference between the empirical Bayesian posterior means $\hat\theta_{i,\hat G_n,\hat\eta}$ and the oracle Bayes posterior means.

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

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

where we define $\|\eta\|_\infty = \max\left(\|m\|_\infty, \|s\|_\infty\right)$ for $\eta = (m, s)$. This result continues a recent statistics literature on empirical Bayes methods via NPMLE by characterizing the effect of an estimated nuisance parameter $\hat\eta$ in a first step.[27]

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

[27] 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.

regret is bounded below[28]

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

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.

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.[29]

Assumption 2.3.1. *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* (2.12), *ignoring a constant factor* $1/\hat{\nu}_i$. *We assume that* $\hat{G}_n$ *satisfies*

$$\frac{1}{n}\sum_{i=1}^n \psi_i(Z_i, \hat{\eta}, \hat{G}_n) \geq \sup_{H \in \mathcal{P}(\mathbb{R})} \frac{1}{n}\sum_{i=1}^n \psi_i(Z_i, \hat{\eta}, H) - \kappa_n \tag{2.17}$$

for tolerance κ_n

$$\kappa_n = \frac{2}{n}\log\left(\frac{n}{\sqrt{2\pi e}}\right). \tag{2.18}$$

Moreover, we require that $\hat{G}_n$ *has support points within* $[\min_i \hat{Z}_i, \max_i \hat{Z}_i]$. *To ensure that* κ_n *is positive, we assume that* $n \geq 7 = \lceil\sqrt{2\pi e}\rceil$.[30]

We now state further assumptions on the data-generating processes $\mathcal{P}_0$ beyond (2.7). First, we assume that G_0 is exponential-tailed with parameter α that controls the thickness of its tails. We state

[28] 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.

[29] In particular, the support restriction for $\hat{G}_n$ in Assumption 2.3.1 is satisfied by all maximizers of the likelihood function (see Corollary 3 in Soloff *et al.*, 2021).

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

the restriction in an equivalent form of simultaneous moment control.[31]

Assumption 2.3.2. *The distribution G_0 is has zero mean, unit variance, and admits simultaneous moment control with parameter $\alpha \in (0, 2]$: There exists a constant $A_0 > 0$ such that for all $p > 0$,*

$$(E_{\tau \sim G_0}[|\tau|^p])^{1/p} \leq A_0 p^{1/\alpha}. \tag{2.19}$$

Next, Assumption 2.3.3 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 2.3.3. *The variances $(\sigma_{1:n}, s_0)$ admit lower and upper bounds:*

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

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

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

Assumption 2.3.4. *Let $C^p_{A_1}([\sigma_\ell, \sigma_u])$ be the Hölder class of order $p \geq 1$ with maximal Hölder norm $A_1 > 0$ supported on $[\sigma_\ell, \sigma_u]$.[32] We assume that*

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

Additionally, let $\beta_0 > 0$ be a constant. Let $\mathcal{V}$ be a set of bounded functions supported on $[\sigma_\ell, \sigma_u]$

[31] An equivalent statement to Assumption 2.3.2 is that there exists $a_1, a_2 > 0$ such that $P_{G_0}(|\tau| > t) \leq a_1 \exp(-a_2 t^\alpha)$ 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 2.3.2 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).

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

Definition 2.3.2. For some set $\mathcal{X} \subset \mathbb{R}$ and constant $A > 0, p > 0$, let $C^p_A(\mathcal{X})$ be the set of continuous functions $f : \mathcal{X} \to \mathbb{R}$ with $\|f\|_{(p)} \leq A$. The norm $\|\cdot\|_{(p)}$ is defined as follows. Let $\underline{p}$ be the greatest integer strictly smaller than p. Define

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

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

that (i) admits the uniform bound $\sup_{f\in\mathcal{V}}\|f\|_\infty \leq C_{A_1}$ *and (ii) admits the metric entropy bound*

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

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

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

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

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

4. *The conditional variance estimator respects the conditional variance bounds in Assumption 2.3.3:* $P\left(\frac{s_{0\ell}}{2} < \hat{s} < 2s_{0u}\right) = 1$.

Assumption 2.3.4 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 2.3.4 expects that the nuisance parameter estimates $\hat{m}$ and $\hat{s}$ are rate-optimal up to logarithmic factors (Stone, 1980). Assumption 2.3.4 also expects that the nuisance parameter estimates belong to a class $\mathcal{V}$ with the same metric entropy behavior as the Hölder class $C^p_{A_1}([\sigma_\ell, \sigma_u])$.[33]

Assumptions 2.3.2 to 2.3.4 specify a class of distributions $\mathcal{P}_0$ and nuisance estimators $\hat{\eta}$ indexed by a set of hyperparameters $\mathcal{H} = (\sigma_\ell, \sigma_u, s_\ell, s_u, A_0, A_1, \alpha, \beta_0, p)$. Our subsequent theoretical results are finite sample, with implicit constants dependent on these hyperparameters $\mathcal{H}$. To review, $(\sigma_\ell, \sigma_u, s_\ell, s_u)$ are bounds on the variances $(\sigma_i^2, s^2(\sigma_i))$; (A_0, α) control the tails of G_0; and (A_1, p) control the smoothness of η_0; and β_0 is the power of the log factor in the $\|\cdot\|_\infty$ estimation rate for η_0.

[33] Regarding Assumption 2.3.4(2), we note that kernel smoothing estimators attain the rates required for Hölder smooth functions m_0, s_0 (see Tsybakov (2008) and Section B.7). Regarding Assumption 2.3.4(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 2.3.4(4) also expects the nuisance parameter estimates to respect the boundedness constraints for s_0. This is mainly so that our results are easier to state; we discuss this assumption in Theorem B.3.3.

Regret results

Consider the following "good event," indexed by $C > 0$,

$$\mathbf{A}_n(C) \equiv \left\{ \|\hat{\eta} - \eta_0\|_\infty \le C n^{-\frac{p}{2p+1}} (\log n)^{\beta_0} \right\}. \tag{2.20}$$

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

Theorem 2.3.3. *Assume Assumptions 2.3.1 to 2.3.4 hold. Then, for any $\delta \in (0, \frac{1}{2})$, there exists universal constants $C_{1,\mathcal{H},\delta} > 0$ and $C_{0,\mathcal{H},\delta} > 0$ such that (i) $P(\mathbf{A}_n(C_{1,\mathcal{H},\delta})) \ge 1 - \delta$ and that (ii) the expected regret conditional on $\mathbf{A}_n(C_{1,\mathcal{H},\delta})$ is dominated by the rate function*

$$E\left[\mathrm{Regret}(\hat{G}_n, \hat{\eta}) \mid \mathbf{A}_n(C_{1,\mathcal{H},\delta})\right] \le C_{0,\mathcal{H},\delta} n^{-\frac{2p}{2p+1}} (\log n)^{\frac{2+\alpha}{\alpha}+3+2\beta_0}. \tag{2.21}$$

If the event $\mathbf{A}_n(C)$ is sufficiently likely, we can control expected regret on the bad event $\mathbf{A}_n^{\mathrm{C}}$ as well. In Section B.7, we verify that local linear regression satisfies a weakening of these assumptions that are also sufficient for the conclusion of Theorem 2.3.4.

Corollary 2.3.4. *Assume the same setting as Theorem 2.3.3. Suppose, additionally, for all sufficiently large $C_{1,\mathcal{H}} > 0$, $P(\mathbf{A}_n(C_{1,\mathcal{H}})) \ge 1 - n^{-2}$. Then, there exists a constant $C_{0,\mathcal{H}} > 0$ such that the expected regret is dominated by the rate function*

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

We can show a corresponding lower bound on the Bayes regret—i.e., a lower bound on the worst-case Bayes regret when an adversary picks G_0, η_0—by showing that any good posterior mean estimate $\hat{\theta}_i$ implies a good estimate $\hat{m}(\sigma_i)$ for m_0. Minimax lower bounds for estimation of m_0 then imply lower bounds for estimation of the oracle posterior means θ_i^*.[34]

Theorem 2.3.5. *Fix a set of valid hyperparameters $\mathcal{H} = (\sigma_\ell, \sigma_u, s_\ell, s_u, A_0, A_1, \alpha, \beta_0, p)$ for Assumptions 2.3.2 to 2.3.4. Let $\mathcal{P}(\mathcal{H}, \sigma_{1:n})$ be the set of distributions P_0 on support points $\sigma_{1:n}$ which satisfy (2.7) and Assumptions 2.3.2 to 2.3.4 corresponding to $\mathcal{H}$. For a given P_0, let $\theta_i^* = E_{P_0}[\theta_i \mid Y_i, \sigma_i]$*

[34] A similar argument is considered in Ignatiadis and Wager (2019) for a related but distinct setting. See, also, Section B.1.6.

denote the oracle posterior means. Then there exists a constant $c_{\mathcal{H}} > 0$ such that the worst-case Bayes regret of any estimator exceeds $c_{\mathcal{H}} n^{-\frac{2p}{2p+1}}$:

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

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

As a result, the rate (2.21) 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 2.3.3. In any case, this cost is not substantial.

The regret upper bounds Theorems 2.3.3 and 2.3.4 are finite-sample statements. As a result, they hold uniformly over all distributions P_0 delineated by the problem parameters $\mathcal{H}$. However, the usefulness of Theorems 2.3.3 and 2.3.4 still lies in the convergence rate, as the constants implied by the proofs are not sharp.

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

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

under assumption on $m_0, s_0, \hat{m}, \hat{s}$ analogous to Assumption 2.3.4. Of course, the resulting convergence rate would suffer from the curse of dimensionality, and the term $n^{-\frac{2p}{2p+1}}$ would be replaced with $n^{-\frac{2p}{2p+1+d}}$, where d is the dimension of X.

Taken together, Theorems 2.3.4 and 2.3.5 are strong statistical optimality guarantees for CLOSE-NPMLE in the canonical problem of estimation with squared error loss. That is, the worst-case performance gap of CLOSE-NPMLE relative to the oracle contracts at the best possible rate, meaning that CLOSE-NPMLE mimics the oracle as well as possible.

For interested readers, we provide an overview of the proof of our main result Theorem 2.3.3 in the following remark. A more detailed walkthrough is in Section B.3.3.

Remark 2.3.6 (Informal discussion of the proof for Theorem 2.3.3)**.** Regret results assuming prior independence are established by Soloff *et al.* (2021) and Jiang (2020), and we build on these results for Theorem 2.3.3. Applied to (Z_i, ν_i, τ_i), these results state that **(i)** approximate maximizers $\tilde{G}_n$ of the (infeasible) log-likelihood $\Psi_n(\eta_0, G) \equiv \frac{1}{n}\sum_i \psi_i(Z_i, \eta_0, G)$ are close to G_0 in terms of the *average Hellinger distance* of the induced densities of Z_i

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

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

Our results extend this argument by accommodating the fact that η_0 is unknown and must be estimated with $\hat{\eta}$.[35] To apply **(ii)** in the literature, we would like to show that **(i')** $\hat{G}_n$—an approximate maximizer of the feasible log-likelihood $\Psi_n(\hat{\eta}, G) = \frac{1}{n}\sum_i \psi_i(Z_i, \hat{\eta}, G)$—is close to G_0 in terms of $\bar{h}^2(\cdot,\cdot)$. This is not a straightforward task and is the most intricate part of our argument. To show **(i')**, we prove a lower bound for the likelihood $\Psi_n(\eta_0, \hat{G}_n)$ (Theorem B.4.1) and adapt the argument for **(i)** to accommodate our likelihood lower bound (Theorem B.5.1).

To lower bound $\Psi_n(\eta_0, \hat{G}_n)$, we relate the two likelihoods by linearization (formally, see (B.15)):

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

Since $\hat{G}_n$ approximately maximizes the feasible likelihood $\Psi_n(\hat{\eta}, \cdot)$, $\Psi_n(\hat{\eta}, \hat{G}_n)$ is large by construction. Thus, if we can show that the right-hand side is small, then the infeasible likelihood $\Psi_n(\eta_0, \hat{G}_n)$ would be close to $\Psi_n(\hat{\eta}, \hat{G}_n)$ and hence would also be large. To obtain the rate (2.21), it is important to show that the right-hand side vanishes *strictly faster* than $\|\hat{\eta} - \eta_0\|_\infty$. To do so, we additionally need to show that the derivatives $\frac{1}{n}\sum_i \partial \psi_i(Z_i, \eta_0, \hat{G}_n)/\partial \eta$ are small. Without it, we would obtain a worse squared error regret rate of the form $n^{-\frac{p}{2p+1}}(\log n)^\beta$.

In particular, we manage to relate the average derivative to the average Hellinger distance (see Theorems B.4.3 and B.4.4)

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

[35] We also translate the resulting regret guarantee on estimating τ_i to regret guarantees on estimating θ_i. In doing so, we identify an apparent gap in the arguments of Jiang (2020) and Soloff *et al.* (2021). We show a modified argument avoids the gap in our setting, which applies to the setting in Soloff *et al.* (2021) as well. See Theorem B.6.5 for details.

Loosely, this is because the population score in η is mean-zero, $E[\partial\psi_i(Z, \eta_0, G_0)/\partial\eta] = 0$. Thus if $\hat{G}_n$ is close to G_0, then the sample score evaluated at $\hat{G}_n$ should also be approximately zero. This is a key step in Section B.4.

This bound for $\Psi_n(\eta_0, \hat{G}_n)$ creates an additional complication when attempting to apply the claim **(i)**. The claim **(i)** upper bounds the Hellinger distance $\bar{h}(f_{\tilde{G}_n,\cdot}, f_{G_0,\cdot})$ using a lower bound for $\Psi_n(\eta_0, \tilde{G}_n)$. However, now our lower bound for the likelihood $\Psi_n(\eta_0, \hat{G}_n)$ itself depends on $\bar{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot})$, and so we cannot apply **(i)** directly. The proof for **(i')** additionally modifies the argument for **(i)** to accommodate our likelihood bound (Section B.5). ■

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

2.3.2 Other decision objectives and relation to squared-error loss

Notably, the oracle Bayes decision rules $\boldsymbol{\delta}^\star$ in Decision Problems 2 and 3 depend solely on the vector of oracle Bayes posterior means $\theta^*_{1:n}$.[36] Therefore, for these problems, the natural empirical Bayes decision rules simply replace oracle Bayes posterior means (θ_i^*) with empirical Bayes ones ($\hat{\theta}_i$) in the oracle decision rules.[37] For instance, if one is comfortable with the prior estimated by CLOSE-NPMLE, then the corresponding decision rules for Decision Problems 2 and 3 threshold based on estimated posterior means under CLOSE-NPMLE.

In these problems, BayesRegret_n (2.14) is equal to the expected risk gap between using the

[36] In principle, one could consider many other policy problems with a ranking flavor (Koenker and Gu, 2019; Kline *et al.*, 2023). Among these problems, UTILITY MAXIMIZATION BY SELECTION and TOP-m SELECTION are special in that optimal decisions are simple functions of the posterior means. We caution that the worst-case regret rate for ranking-type problems without this property can be unfavorable—as Gu and Koenker (2023) put it, "inherently futile"—since their optimal decisions depend on functionals that are known to be difficult to estimate (i.e., they have logarithmic minimax rates of estimation, Pensky, 2017; Dedecker and Michel, 2013; Cai and Low, 2011), without stronger assumptions on the prior.

In general, the minimax squared error rate of estimating $E[f(\theta)]$ is logarithmic, unless f is an analytic function, by an extension of the argument in Cai and Low (2011). Ranking-type problems often involve f of the form $f(\theta) = \mathbb{1}(\theta > c)$ or $f(\theta) = \max(\theta, c)$, which are not smooth. This observation suggests that these ranking-type problems may also suffer from logarithmic regret rates—though, this observation alone does not rigorously prove this, as difficulties in estimating $Ef(\theta)$ *in squared error* might not preclude a polynomial regret rate for these ranking-type problems.

[37] Theorem 2.3.7 applies to any estimators of the oracle Bayes posterior means—not necessarily derived through an empirical Bayes procedure—and does not impose the location-scale assumption. As a result, it may be of independent interest.

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 Regret.

Theorem 2.3.7. *Suppose* (2.4) *holds, but* (2.7) *may or may not hold. Let $\hat{\delta}_i$ be the plug-in decisions with any vector of estimates $\hat{\theta}_i$, not necessarily from* CLOSE-NPMLE. *We have the following inequalities on the expected regret corresponding to the decision rules $\hat{\delta}_i$:*

1. *For* UTILITY MAXIMIZATION BY SELECTION,

$$E[\mathrm{UMRegret}_n] \le \left(E\left[\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_i - \theta_i^*)^2\right]\right)^{1/2}. \tag{2.22}$$

2. *For* TOP-m SELECTION,

$$E[\mathrm{TopRegret}_n^{(m)}] \le 2\sqrt{\frac{n}{m}}\left(E\left[\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_i - \theta_i^*)^2\right]\right)^{1/2}. \tag{2.23}$$

Theorem 2.3.7 shows that the two decision problems UTILITY MAXIMIZATION BY SELECTION and TOP-m SELECTION are easier than estimating the oracle Bayesian posterior means. As a result, our convergence rates from Theorems 2.3.3 and 2.3.4 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 (2.22) and (2.23) are of the form $n^{-p/(2p+1)}(\log n)^c = o(1)$ under Theorem 2.3.4. 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 2.3.8 (Mover interpretation of Theorem 2.3.7)**.** Recall that we can think of TOP-m SELECTION as the decision problem in Bergman *et al.* (2023). The utility function represents the expected mobility of a mover, assuming that the mover moves randomly into one of the high mobility Census tracts. Our proof of Theorem 2.3.7 in Section B.1.2 allows for a slightly more general decision problem. Suppose the decision now is to provide a full ranking of Census tracts for potential movers and maximize the expected mobility for a mover. Suppose that the probability that a mover moves to a tract depends decreasingly and solely on the tract's rank. To be more concrete, suppose the mover has probability π_1 of moving to the highest-ranked tract, π_2 to the second-highest, and so forth. Then, with the same argument, the corresponding regret is dominated by $2\sqrt{n\sum_{i=1}^n \pi_i^2} \cdot \left(E\left[\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_i - \theta_i^*)^2\right]\right)^{1/2}$, which generalizes (2.23). ■

Remark 2.3.9 (Tightness of Theorem 2.3.7)**.** We suspect that the actual performance of CLOSE-NPMLE for Decision Problems 2 and 3 may be better than predicted by Theorem 2.3.7. Take the bound for $\mathrm{UMRegret}_n$, for instance. As would be clear from the proof, the bound (2.22) holds even when the c_i's are adversarially chosen[38] such that the empirical Bayesian makes every mistake: $\hat{\delta}_i \neq \delta_i^\star$ for every i. However, for a fixed vector c, we expect that $\hat{\delta}_i \neq \delta_i^\star$ only for a vanishing fraction of populations, and thus the actual performance of $\hat{\delta}_i$ may be better than the rate in Section B.1.2 implies.[39]

Though we conjecture that the rate in Theorem 2.3.7 does not match a lower bound, Theorem 2.3.7 is competitive with recent results. TOP-m SELECTION is recently studied by Coey and Hung (2022), who show that under prior independence, if $\hat{\theta}_{1:n}$ are posterior means for some estimate $\hat{G}$ of the prior $G_{(0)}$, then

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

where $W_1(P, Q)$ is the Wasserstein-1 distance between P, Q. Theorem 2.3.7 attains a worse rate in parametric settings, when the prior $G_{(0)}$ can be estimated at fast rates. However, in nonparametric settings, $G_{(0)}$ is often only estimable at logarithmic rates (Dedecker and Michel, 2013), and thus the rate in Theorem 2.3.7 is much more favorable in those settings. ■

[38] That said, if the c_i's are indeed adversarially chosen given knowledge of $(Y_{1:n}, \sigma_{1:n}, P_0)$, then Theorem 2.3.7 does match a corresponding lower bound, derived by choosing $c_i = (\hat{\theta}_i + \theta_i^\star)/2$.

[39] Upper and lower bounds are derived in related but distinct settings by Audibert and Tsybakov (2007); Bonvini *et al.* (2023); Liang (2000); some upper bounds, under possibly stronger assumptions, appear better than implied by Section B.1.2.

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

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

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

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

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

2.3.3 Robustness to the location-scale assumption (2.7)

We prove our regret upper and lower bounds imposing the location-scale model (2.7). This is an optimistic assessment of the performance of CLOSE-NPMLE. While (2.7) nests prior independence, it may still be misspecified. We now consider the worst-case behavior of CLOSE-NPMLE without the location-scale assumption. Since without the location-scale assumption, CLOSE-NPMLE can no longer hope to emulate the oracle Bayes decisions, we focus on worst-case Bayes *risk* here, instead of on regret.

We will do so by considering an idealized version of the procedure. So long as $\theta_i \mid \sigma_i$ has two moments, $\eta_0(\cdot) = (m_0(\cdot), s_0(\cdot))$ are well-defined as conditional moments of $\theta_i \mid \sigma_i$ without imposing the location-scale assumption. We will assume that m_0, s_0 are known. Without the location-scale model, G_0 is ill-defined, but we will assume that we obtain some pseudo-true value G_0^* that has zero mean and unit variance. This is a reasonable condition to impose, since every conditional prior distribution $\tau_i \mid \sigma_i$ obeys this moment constraint.[40] Thus, for estimating $\tau_i = \frac{\theta_i - m_0(\sigma_i)}{s_0(\sigma_i)}$, whose true prior is $\tau_i \mid \sigma_i \sim G_i$, this idealized procedure uses some misspecified prior $G_0^* \neq G_i$, which does have the correct first two moments.

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

Theorem 2.3.10. *Under (2.4) but not (2.7), assume the conditional distribution $\theta_i \mid \sigma_i$ has mean $m_0(\sigma_i)$ and variance $s_0^2(\sigma_i)$. Denote the set of distributions of $\theta_{1:n} \mid \sigma_{1:n}$ which obey these restrictions as $\mathcal{P}(m_0, s_0)$. Let $\hat{\theta}_{i,G_0^*,\eta_0}$ denote the posterior mean estimates with some prior P^* under the location-scale model $P^*(\theta_i \le t \mid \sigma_i) = G_0^*\left(\frac{t - m_0(\sigma_i)}{s_0(\sigma_i)}\right)$, for some fixed G_0^* with zero mean and unit variance. Let $\bar{\rho} = \max_i s_0^2(\sigma_i)/\sigma_i^2 < \infty$ be the maximal conditional signal-to-noise ratio and assume that it is*

[40] We do not know if the maximizer G of the population analogue to (2.12) 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 (Section B.2.2).

bounded. Then, for some $C_{\bar{\rho}} < \infty$ that solely depends on $\bar{\rho}$,

$$\sup_{P_0 \in \mathcal{P}(m_0, s_0)} E_{P_0}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_{i, G_0^*, \eta_0} - \theta_i)^2\right] \le C_{\bar{\rho}} \cdot \inf_{\hat{\theta}_{1:n}} \sup_{P_0 \in \mathcal{P}(m_0, s_0)} E_{P_0}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2\right]. \tag{2.24}$$

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

Theorem 2.3.10 shows that the worst-case behavior of an idealized version of CLOSE-NPMLE must come within a factor of the minimax risk and hence is not arbitrarily unreasonable, even under misspecification. We caution that (2.24) is a fairly weak guarantee, in that the decision rule that simply outputs the prior conditional mean ($\delta_i = \hat{\theta}_{i,\boldsymbol{\delta}_0,\eta_0} = m_0(\sigma_i)$) also satisfies it. Nevertheless, even so, (2.24) *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 $s_0^2 = \frac{1}{n}\sum_{i=1}^n (m_0(\sigma_i) - m_0)^2 + s_0^2(\sigma_i)$.[41] To provide additional reassurance for CLOSE-NPMLE under misspecification, Section B.1.7 discusses an interpretation of CLOSE-NPMLE under misspecification of (2.7), and the validation procedure developed in Section 2.4.3 provides unbiased evaluation without relying on the location-scale model.

2.4 Practical considerations

2.4.1 A detailed recipe

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

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

$$\begin{aligned} m_0(\sigma) &= E[\theta \mid \sigma] = E[Y \mid \sigma] \\ s_0^2(\sigma) &= \mathrm{Var}(\theta \mid \sigma) = E[(Y - m_0(\sigma))^2 \mid \sigma] - \sigma^2, \end{aligned} \tag{2.25}$$

we can estimate them accordingly with off-the-shelf methods (e.g., local polynomial kernel smoothing

[41] 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.

methods implemented by Calonico *et al.*, 2019). Specifically, estimating m_0 with $\hat{m}$ is directly a nonparametric regression of Y_i on σ_i.[42] Estimating $s_0^2(\cdot)$ can be operationalized by first nonparametrically regressing $(Y_i - \hat{m}(\sigma_i))^2$ on σ_i, and then subtracting off σ_i^2. This is a plug-in estimator for s_0^2, as it replaces quantities in (2.25) with their empirical counterparts.[43]

A wrinkle is that the plug-in estimate $\hat{s}$ may be negative.[44] Truncating $\hat{s}$ at zero results in observations whose estimated prior variances $\hat{s}^2(\sigma_i) = 0$. These observations also have implied $\hat{\nu}_i = \infty$. For these observations, an empirical Bayesian taking $\hat{s}^2(\sigma_i) = 0$ at face value has degenerate priors at $\hat{m}(\sigma_i)$. Since observations with $\nu_i = \infty$ do not contribute to the likelihood objective for NPMLE, excluding them from the NPMLE computation does not alter the estimated $\hat{G}_n$. Thus, we can continue to use $(\hat{m}, \hat{s}^2, \hat{G}_n)$ as the estimated posterior—an observation with $\hat{s}^2(\sigma_i) = 0$ would have a point mass at $\hat{m}(\sigma_i)$ as its estimated posterior. In our experience, this simple approach does not appear to affect performance. Nevertheless, in Section B.7, we propose a heuristic but data-driven truncation rule, borrowing from a statistics literature on estimating non-centrality parameters for non-central χ^2 distributions (Kubokawa *et al.*, 1993). Section B.7 also discusses tuning parameter selection for estimating (m_0, s_0) and verifies that our local linear regression estimators satisfy the regularity conditions in Section 2.3.

Next, in the second step **CLOSE–STEP 2**, we form the transformed estimates $\hat{Z}_i = \frac{Y_i - \hat{m}(\sigma_i)}{\hat{s}(\sigma_i)}$ and the transformed standard errors $\hat{\nu}_i = \sigma_i / \hat{s}(\sigma_i)$. We then estimate the NPMLE on the data $(\hat{Z}_i, \hat{\nu}_i)$ by maximizing (2.12). In practice, the infinite-dimensional optimization problem (2.12) is approximated with a finite-dimensional one by discretizing distributions on a grid. To be precise, let $\min_i \hat{Z}_i = \tau_{(1)} \le \cdots \le \tau_{(J)} = \max_i \hat{Z}_i$ be a pre-specified grid of points, not necessarily equally spaced, and denote it by $\boldsymbol{\tau}$.[45] The feasible version of (2.12) maximizes the concave program

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

[43] Since (2.25) can be written in different forms, there are other reasonable plug-in estimators for s_0. We investigate one such alternative estimator in Section B.2.2 and find very similar performance in our empirical exercise.

[44] The negative estimated variance phenomenon similarly may occur with estimating the prior variance with INDEPENDENT-GAUSS and with conditional variance estimation in Armstrong *et al.* (2022). This is in part caused by estimation noise in $\text{Var}(Y_i \mid \sigma_i)$. However, there is some evidence that observations with large estimated σ_i's are underdispersed for the measures of economic mobility in the Opportunity Atlas (see Section B.2.1.)

[45] Since the gridding is a computational approximation to the infinite dimensional optimization problem, the sole downside

$\pi^\star \equiv \max_{\pi \in \mathbb{R}^J_{\geq 0}, \pi' 1 = 1} \sum_{i=1}^n \log \left(\sum_{j=1}^J \pi_j \varphi \left(\frac{\hat{Z}_i - \tau_{(j)}}{\hat{\nu}_i} \right) \right)$. The estimated NPMLE $\hat{G}_n$ is a discrete distribution with support points $\tau_{(j)}$ and corresponding masses $\pi^\star_j$.

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

$$P_{\hat{G}_n}(\tau_i = \tau_{(j)} \mid \hat{Z}_i = z, \hat{\nu}_i = \nu) \propto \pi^\star_j \exp\left(-\frac{1}{2\nu^2}(z - \tau_{(j)})^2 \right),$$

normalized so that the probabilities sum to 1. This probability mass function can be plugged into (2.6) to compute the empirical Bayes decision rule for any loss function L.[46]

2.4.2 When does relaxing prior independence matter?

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

When prior independence does not hold, but when the conditional location-scale model (2.7) approximately holds, we expect CLOSE to outperform methods that assume prior independence. Qualitatively speaking, we expect the improvement of CLOSE-methods to be large when the conditional expectation accounts for large portions of the unconditional signal variance $\mathrm{Var}(\theta_i)$. Since we can decompose $\mathrm{Var}(\theta_i) = E[s_0^2(\sigma_i)] + \mathrm{Var}(m_0(\sigma_i))$, we expect the improvement of CLOSE-methods to be large when the variance of the conditional expectation $\mathrm{Var}(m_0(\sigma_i))$ is large compared to $E[s_0^2(\sigma_i)]$. Intuitively, this is the case when σ_i is highly predictive of θ_i. Whether this is the case can be easily checked by plotting Y_i against σ_i, as in Figure 2.1, and inspecting the estimated conditional moments.

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

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

Finally, when the conditional distributions $\theta_i \mid \sigma_i$ are non-Gaussian, and in particular when they are discrete, skewed, or thick-tailed, we expect CLOSE-NPMLE to additionally outperform INDEPENDENT-GAUSS due to not assuming Normality of θ_i. When the conditional priors are Gaussian, estimating it via the NPMLE pays a modest statistical price. Admittedly, it is often difficult to diagnose whether the underlying conditional distributions $\theta_i \mid \sigma_i$ have these properties, since we only observe (Y_i, σ_i). Likewise, so far the discussion in this subsection is heuristic. To be more certain of the extent of improvement of CLOSE-NPMLE over other methods, it is helpful to have out-of-sample validation. The next subsection proposes a minor extension of Oliveira *et al.* (2021), which allows for an unbiased estimate of loss and serves as a validation procedure.

2.4.3 A formal validation procedure via coupled bootstrap

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

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

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

$$\begin{bmatrix} Y_i^{(1)} \\ Y_i^{(2)} \end{bmatrix} \mid \theta_i, \sigma_i^2 \sim \mathcal{N}\left(\begin{bmatrix} \theta_i \\ \theta_i \end{bmatrix}, \begin{bmatrix} (1+\omega)\sigma_i^2 & 0 \\ 0 & (1+\omega^{-1})\sigma_i^2 \end{bmatrix} \right). \tag{2.26}$$

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

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

$$Y_i^{(1)} \mid \theta_i, \sigma_i^2 \sim \mathcal{N}\left(\theta_i, (1+\omega)\sigma_i^2\right) \quad Y_i^{(2)} \mid \theta_i, \sigma_i^2 \sim \mathcal{N}\left(\theta_i, (1+\omega^{-1})\sigma_i^2\right) \tag{2.27}$$

independently. Note that the two representations (2.26) and (2.27) are equivalent, and hence coupled bootstrap emulates sample-splitting. For instance, coupled bootstrap with a value of $\omega = 1/9$ is statistically equivalent to splitting the micro-data with a 90-10 train-test split.

Just as we can perform out-of-sample validation with sample-splitting on the micro-data, we can also do so with the coupled bootstrap emulation of sample-splitting. The following proposition formalizes this and states unbiased estimators for the loss of these decision rules, as well as their accompanying standard errors.[47]

Table 2.1: Unbiased estimators for loss of decision rules and associated conditional variance expressions (Theorem 2.4.1)

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

Proposition 2.4.1. *Suppose (Y_i, σ_i) obey the Gaussian heteroskedastic location model, assumed to be independent across i (2.4). 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 2.1 are unbiased for the corresponding loss:*

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

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

[47] Oliveira *et al.* (2021) state the unbiased estimation result for the mean-squared error estimation problem. They develop the result further by connecting the coupled bootstrap estimator to Stein's unbiased risk estimate. Our analogous calculation for other loss functions and for the standard errors is a minor extension of their results. Theorem 2.4.1 can also be easily generalized to other loss functions that admit unbiased estimators (Effectively, the loss is a function of a Gaussian location θ_i. For unbiased estimation of functions of Gaussian parameters, see Table A1 in Voinov and Nikulin, 2012).

Table 2.1.

Theorem 2.4.1 allows for an out-of-sample evaluation of decision rules, as well as uncertainty quantification around the estimate of loss, solely imposing the heteroskedastic Gaussian model. This is a useful property in practice for comparing different empirical Bayes methods. The alternative is to take some estimated prior—say the one learned by CLOSE-NPMLE—as the true prior, and evaluate performance of competing methods. Doing so likely tips the scale in favor of a particular method, and we advocate for the coupled bootstrap instead.

2.5 Empirical illustration

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

2.5.1 The Opportunity Atlas and Creating Moves to Opportunity

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

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

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

Specifically, using longitudinal Census micro-data, Chetty *et al.* (2020) estimate tract-level children's outcomes in adulthood and publish the estimates in a collection of datasets called the Opportunity Atlas. Each dataset contains estimates and standard errors for some particular definition of the economic parameter of interest, at the Census tract level. Taking these estimates from the Opportunity Atlas, Bergman *et al.* (2023) conducted a program in Seattle called Creating Moves to Opportunity. They provided assistance to treated low-income individuals[48] to move to "Opportunity Areas"—Census tracts with empirical Bayes posterior means in the top third.[49] We view Bergman *et al.*'s (2023) objectives as TOP-m SELECTION (Decision Problem 3), for m equal to one third of the number of tracts in King County, Washington (Seattle).

The Opportunity Atlas also includes tract-level covariates, a complication that we have so far abstracted away from. In the ensuing empirical exercises—as well as in Bergman *et al.* (2023)—the estimates and parameters are residualized against the covariates as a preprocessing step. We now let $\tilde{Y}_i$ denote the raw Opportunity Atlas estimates for a pre-residualized parameter ϑ_i and let (Y_i, θ_i) be their residualized counterparts against a vector of tract-level covariates X_i, with regression coefficient β.[50] We can apply the empirical Bayes procedures in this paper to (Y_i, σ_i^2) and obtain an estimated posterior for θ_i. This estimated posterior for the residualized parameter θ_i then implies an estimated posterior for the original parameter $\vartheta_i = \theta_i + X_i'\beta$, by adding back the fitted values $X_i'\beta$ (Fay and Herriot, 1979). When there are no covariates, $\vartheta_i = \theta_i$ and $Y_i = \tilde{Y}_i$.

We consider several measures of economic mobility ϑ_i. For our purposes, these definitions of

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

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

[50] Precisely speaking, let X_i be a vector of tract-level covariates. Let $\tilde{Y}_i$ be the raw Opportunity Atlas estimates of a parameter ϑ_i, with accompanying standard errors σ_i. Let β be some vector of coefficients, typically estimated by weighted least-squares of Y_i on X_i. Let $Y_i = \tilde{Y}_i - X_i'\beta$ and $\theta_i = \vartheta_i - X_i'\beta$ be the residuals. We assume that the tract-level covariates do not predict the estimation noise in $\tilde{Y}_i$: i.e., $X_i \perp\!\!\!\perp \tilde{Y}_i \mid \theta_i, \sigma_i^2$. Since β is precisely estimated, we ignore its estimation noise. Then, the residualized objects (Y_i, θ_i) obey the Gaussian location model $Y_i \mid \theta_i, \sigma_i \sim \mathcal{N}(\theta_i, \sigma_i^2)$. See additional discussion on covariates in Section B.1.6. Figure B.6 contains empirical results without residualizing against covariates.

ϑ_i take the following form: ϑ_i is the population mean of *some* outcome for individuals of *some* demographic subgroup growing up in tract i, whose parents are at the 25th income percentile. We will consider three types of outcomes:

1. Percentile rank of adult income

2. An indicator for whether the individual has incomes in the top 20 percentiles

3. An indicator for whether the individual is incarcerated

for the following demographic subgroups:[51] 1. all individuals (POOLED), 2. white individuals, 3. white men, 4. Black individuals, and 5. Black men. As shorthands, we refer to the three types of outcomes as MEAN RANK, TOP-20 PROBABILITY, and INCARCERATION, respectively. The outcome we use in Section 2.2 corresponds to TOP-20 PROBABILITY for Black individuals, while Bergman *et al.* (2023) consider MEAN RANK POOLED.[52]

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

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

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

2.5.2 Calibrated simulation

Our first empirical exercise is a calibrated simulation. To devise a data-generating process that does not impose the location-scale assumption, we partition σ into vingtiles, fit a location-scale model within each vingtile, and draw from the estimated model (see Section B.2.3 for details). Since the location-scale model is only imposed within each vingtile, this data-generating process does not impose (2.7) on the entire dataset. Figure 2.3 shows an overlay of real and simulated data for one of the variables we consider. Visually, at least, the simulated data resemble the real estimates.

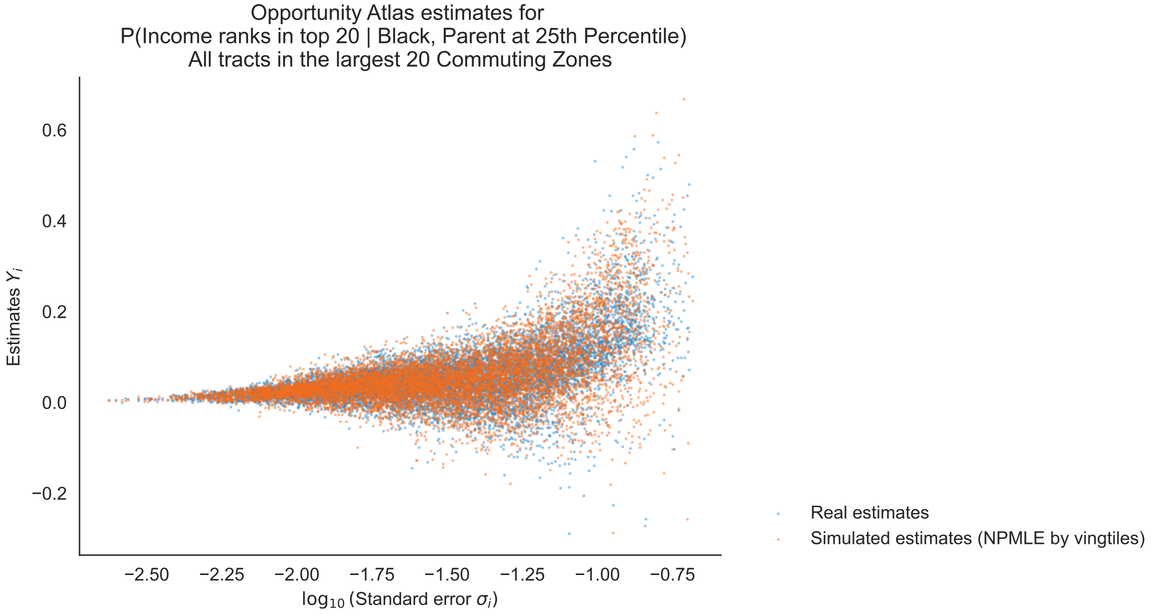


Figure 2.3: A draw of real vs. simulated data for estimates of TOP-20 PROBABILITY for Black individuals

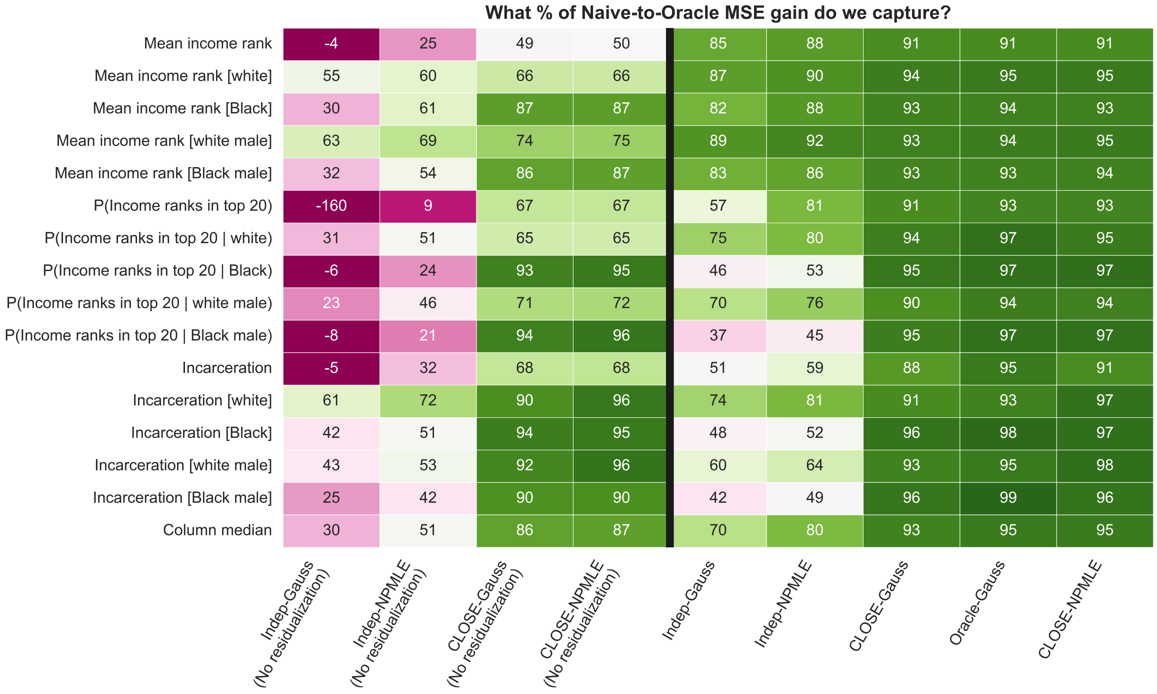
On the simulated data, we then put various empirical Bayes strategies to test. We consider the feasible procedures NAIVE, INDEPENDENT-GAUSS, INDEPENDENT-NPMLE, CLOSE-GAUSS, and CLOSE-NPMLE, where NAIVE sets $\hat{\theta}_i = Y_i$.[53] Because we have the ground truth data-generating process, we additionally have two infeasible benchmarks:

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

- ORACLE: A Bayesian who has access to the distribution of (θ_i, σ_i) and uses the true posterior means for θ_i.[54]

- ORACLE-GAUSS: A Bayesian who knows (m_0, s_0) and uses (2.13).

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



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

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

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

For absolute, un-normalized performance of INDEPENDENT-GAUSS, INDEPENDENT-NPMLE, CLOSE-NPMLE, and ORACLE, see Figure B.10. □

Figure 2.4: Table of relative squared error Bayes risk for various empirical Bayes approaches

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

Figure 2.4 plots the main results from this calibrated simulation. For each method and each target variable, we display a relative measure of gain in terms of mean-squared error. For each method, we calculate its squared error gain over NAIVE, as a percentage of the squared error gain of ORACLE over NAIVE. If we think of the ORACLE–NAIVE difference as the total size of the "statistical pie," then Figure 2.4 shows how much of this pie each method captures. A value of 70 in Figure 2.4, for instance, indicates that a particular method captures 70% of the possible extent of risk gains for a particular parameter definition.

The first four columns show the relative mean-squared error performance *without* residualizing against covariates, applying empirical Bayes methods directly on $(\tilde{Y}_i, \sigma_i)$. We see that methods which assume prior independence—INDEPENDENT-GAUSS and INDEPENDENT-NPMLE—perform worse than methods based on CLOSE.[55] 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, CLOSE-NPMLE uniformly dominates all three other methods. This is because the standard error σ_i contains much of the predictive power of the covariates, and using that information can be very helpful when the researcher does not have rich covariate information.

The next five columns show performance when the methods do have access to covariate information. Compared to their no-covariates counterparts, the methods that assume prior independence do substantially better, since the covariates absorb some dependence between ϑ_i and σ_i. For MEAN RANK, after covariate residualization, there appears to be little dependence between θ_i and σ_i. INDEPENDENT-NPMLE and CLOSE-NPMLE perform similarly, capturing almost all of the available gains. Both methods slightly outperform the Gaussian methods for MEAN RANK.[56]

[55] It may be surprising that INDEPENDENT-GAUSS can perform worse than NAIVE on MSE, since Gaussian empirical Bayes typically has a connection to the James–Stein estimator, which dominates the MLE. We note that, as in Bergman *et al.* (2023), when we estimate the prior mean and prior variance, we *weight* the data with precision weights proportional to $1/\sigma_i^2$. When the independence between θ and σ holds, these precision weights typically improve efficiency. However, the weighting does break the connection between Gaussian empirical Bayes and James–Stein, and the resulting posterior mean does not always dominate the MLE (i.e., NAIVE). To take an extreme example, if a particular observation has $\sigma_i \approx 0$, then that observation is highly influential for the prior mean estimate. If $E[\theta_i \mid \sigma_i]$ is very different for that observation than the other observations, then the estimated prior mean is a bad target to shrink towards.

[56] Section B.2.4 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

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

So far, we have tested the methods in a synthetic environment set up to imitate the real data. Next, we turn to an empirical application that uses the coupled bootstrap (Section 2.4.3) estimator of performance.

2.5.3 Validation exercise via coupled bootstrap

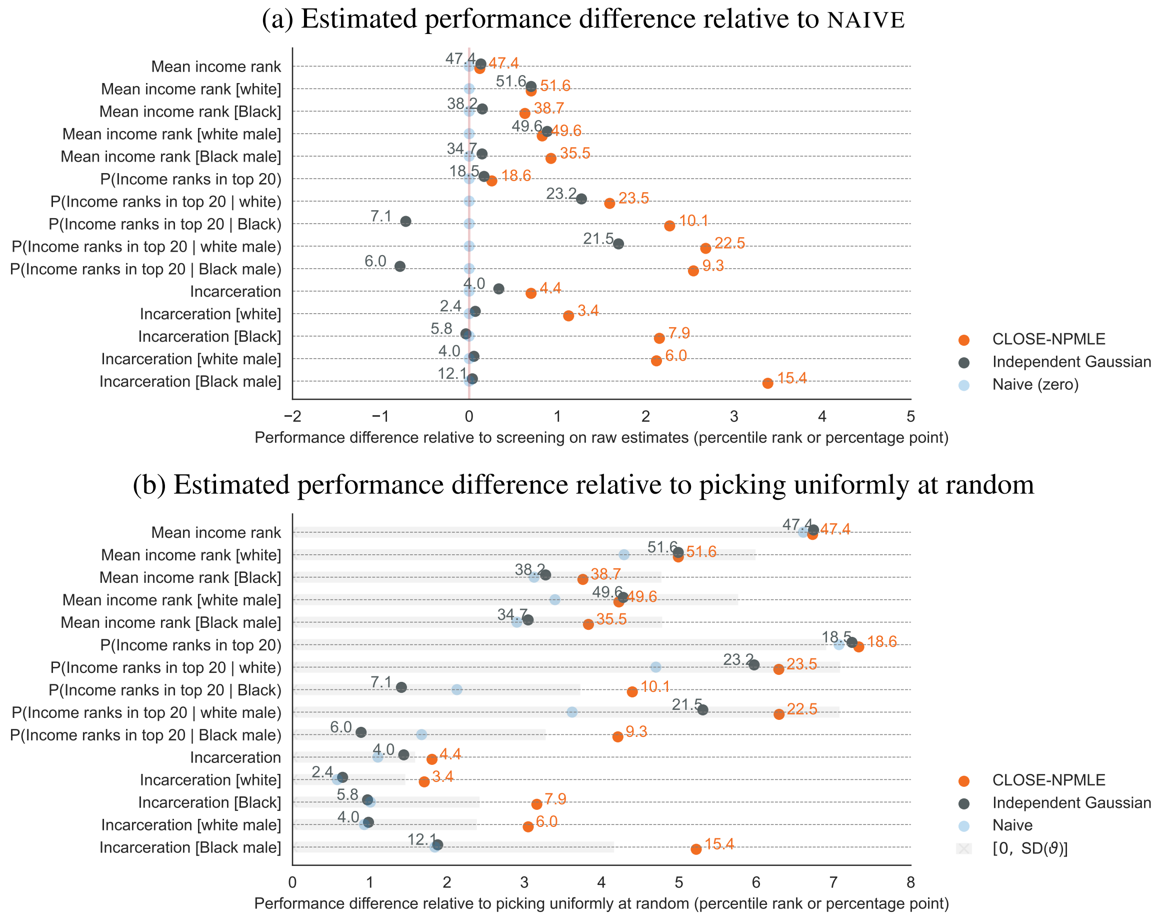
Our second empirical exercise uses the coupled bootstrap described in Section 2.4.3 for the ranking policy problem in Bergman *et al.* (2023). Throughout, we choose ω to emulate a 90-10 train-test split on the micro-data.

Bergman *et al.* (2023) use empirical Bayes methods to select the top third Census tracts in Seattle, based on economic mobility—which we view as a TOP-m SELECTION problem (Decision Problem 3). Can CLOSE-NPMLE make better selections—can it select tracts with higher ϑ_i on average? Specifically, we imagine scaling up Bergman *et al.* (2023)'s exercise and perform INDEPENDENT-GAUSS and CLOSE-NPMLE for all Census tracts in the largest twenty Commuting Zones. We then select the top third of tracts *within* each Commuting Zone, according to empirical Bayesian posterior means for ϑ_i. Additionally, to faithfully mimic Bergman *et al.* (2023), here we perform all empirical Bayes procedures *within Commuting Zone*. That is, for each of the twenty Commuting Zones that we consider, we execute all empirical Bayes methods—including the residualization by covariates—with only $\tilde{Y}_i, \sigma_i$ of tracts within the Commuting Zone.[57]

Figure 2.5(a) shows the estimated performance gap between a given empirical Bayes method

priors.

[57] Section B.2.6 contains results where we perform empirical Bayes pooling over all Commuting Zones and select the top third within each Communting Zone. We obtain very similar results. Section B.2.6 also contains results without residualizing against covariates, and INDEPENDENT-GAUSS performs very poorly in that setting. Section B.2.5 contains results on estimating ϑ_i in MSE (Decision Problem 1) in this context.


(a) Estimated performance difference relative to NAIVE

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

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

Figure 2.5: Performance of decision rules in top-m selection exercise

and NAIVE as the x-position of the dots. The estimated performance of each method,[58] defined as

[58] By virtue of Theorem 2.4.1, these estimated performances are unbiased for the true (negative) Bayes risk. Despite being averaged over 1,000 coupled bootstrap draws, these estimates are not free of sampling error, since, for one, the stochastic components in Y_i are not redrawn.

the average ϑ_i among those selected (2.15), is shown in the annotated figures. According to these estimates, CLOSE-NPMLE generally improves over INDEPENDENT-GAUSS.[59]

For the MEAN RANK variables, using CLOSE-NPMLE generates substantial gains for mobility measures for Black individuals (0.8 percentile ranks for Black men and 0.5 percentile ranks for Black individuals). To put these gains in dollar terms, the Housing Choice Voucher holders in Bergman *et al.* (2023) have incomes around $19,000, and for these individuals, an incremental percentile rank amounts to about $1,000. Thus, the estimated gain in terms of mean income rank is roughly $500–800. For the other two outcomes, TOP-20 PROBABILITY and INCARCERATION,[60] the gains are even more sizable, especially for Black individuals. These gains are as high as 2–3 percentage points on average in terms of these two variables.

Bergman *et al.* (2023) select tracts based on MEAN RANK POOLED. For this measure, there is little additional gain from using CLOSE-NPMLE, at least when residualized against sufficiently rich covariates. Nevertheless, since about half of the trial participants are Black in Bergman *et al.*'s (2023) setting, one might consider providing more personalized recommendations by targeting measures of economic mobility for finer demographic subgroups. If we select tracts based on these demographic-specific measures of economic mobility, CLOSE-NPMLE then provides economically significant improvements.[61]

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

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

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

[61] Section B.2.7 shows that screening with mobility measures for Black individuals outperforms screening mobility for Black individuals with the POOLED estimate.

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 2.5(b). Figure 2.5(b) plots the difference between a given method's performance and the estimated mean ϑ_i for a given measure. Analogous to the value of basic empirical Bayes, we think of the difference between INDEPENDENT-GAUSS's performance and the estimated mean ϑ_i as the *value of data*, since choosing the tracts randomly in the absence of data has expected performance equal to mean ϑ_i. If the mobility estimates are at all useful for decision-making, the value of data must be economically significant.

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

2.6 Conclusion

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

Instead of treating θ_i as independent from σ_i, we model its conditional distribution as a location-scale family. This assumption leads naturally to a family of empirical Bayes strategies that we call

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

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

We close by highlighting some future directions. In Section 2.5, we use kernel smoothing methods to estimate the unknown conditional moments $\eta_0 = (m_0, s_0)$. These methods presume a given level of smoothness and do not adapt to the true smoothness of η_0. We can imagine replacing the conditional moment estimation with adaptive methods (e.g., van der Vaart and van Zanten, 2009). With cross-fitting, the regret result should similarly adapt to the $\|\cdot\|_\infty$ rate of the estimators. Additionally, for the purpose of frequentist inference, the procedure of Armstrong *et al.* (2022) apply in our setting as well and provide confidence sets for the vector of parameters $\theta_{1:n}$ with average coverage guarantees. For frequentist inference on the oracle posterior mean $E_{P_0}[\theta_i \mid Y_i, \sigma_i]$, we conjecture that a version of Ignatiadis and Wager's (2022) procedure—which so far only applies in the homoskedastic Gaussian case—is valid under the location-scale model (2.7).

Chapter 3

Logs with zeros? Some problems and solutions[1]

[1] Jiafeng Chen and Jonathan Roth, "Logs with zeros? Some problems and solutions," *The Quarterly Journal of Economics*, forthcoming, online advanced article qjad054, by permission of Oxford University Press.
An earlier draft of this paper was titled "Log-like? Identified ATEs defined with zero-valued outcomes are (arbitrarily) scale-dependent." We thank Isaiah Andrews, Kirill Borusyak, Jonathan Cohn, Amy Finkelstein, Edward Glaeser, Nick Hagerty, Peter Hull, Jetson Leder-Luis, Erzo Luttmer, Giovanni Mellace, John Mullahy, Edward Norton, David Ritzwoller, Brad Ross, Pedro Sant'Anna, Jesse Shapiro, Neil Thakral, Casey Wichman, and seminar participants at BU, Georgetown, Harvard/MIT, Southern Denmark University, Vanderbilt, Stanford, UC-Irvine, UCLA, UCSD, and the SEA annual conference for helpful comments and suggestions. Bruno Lagomarsino provided superb research assistance.

Dissertation Advisor:
Professor Isaiah Andrews

Author:
Jiafeng Chen

Essays in Econometrics

Abstract

When studying an outcome Y that is weakly-positive but can equal zero (e.g. earnings), researchers frequently estimate an average treatment effect (ATE) for a "log-like" transformation that behaves like $\log(Y)$ for large Y but is defined at zero (e.g. $\log(1+Y)$, $\operatorname{arcsinh}(Y)$). We argue that ATEs for log-like transformations should not be interpreted as approximating percentage effects, since unlike a percentage, they depend on the units of the outcome. In fact, we show that if the treatment affects the extensive margin, one can obtain a treatment effect of any magnitude simply by re-scaling the units of Y before taking the log-like transformation. This arbitrary unit-dependence arises because an individual-level percentage effect is not well-defined for individuals whose outcome changes from zero to non-zero when receiving treatment, and the units of the outcome implicitly determine how much weight the ATE for a log-like transformation places on the extensive margin. We further establish a trilemma: when the outcome can equal zero, there is no treatment effect parameter that is an average of individual-level treatment effects, unit-invariant, and point-identified. We discuss several alternative approaches that may be sensible in settings with an intensive and extensive margin, including (i) expressing the ATE in levels as a percentage (e.g. using Poisson regression), (ii) explicitly calibrating the value placed on the intensive and extensive margins, and (iii) estimating separate effects for the two margins (e.g. using Lee bounds). We illustrate these approaches in three empirical applications.

3.1 Introduction

When the outcome of interest Y is strictly positive, researchers often estimate an average treatment effect (ATE) in logs of the form $E_P[\log(Y(1)) - \log(Y(0))]$, which has the appealing feature that its units approximate percentage changes in the outcome.[2] A practical challenge in many economic settings, however, is that the outcome may sometimes equal zero, and thus the ATE in logs is not well-defined. When this is the case, it is common for researchers to estimate ATEs for alternative transformations of the outcome such as $\log(1+Y)$ or $\text{arcsinh}(Y) = \log\left(\sqrt{1+Y^2}+Y\right)$, which behave similarly to $\log(Y)$ for large values of Y but are well-defined at zero. The treatment effects for these alternative transformations are typically interpreted like the ATE in logs, i.e. as (approximate) average percentage effects. For example, among the 11 papers published in the *American Economic Review* since 2018 that interpret a treatment effect for $\text{arcsinh}(Y)$, all but one interpret the result as a percentage effect or elasticity.[3]

The main point of this paper is that identified ATEs that are well-defined with zero-valued outcomes should not be interpreted as percentage effects, at least if one imposes the logical requirement that a percentage effect does not depend on the baseline units in which the outcome is measured (e.g. dollars, cents, or yuan).

Our first main result shows that if $m(y)$ is a function that behaves like $\log(y)$ for large values of y but is defined at zero, then the ATE for $m(Y)$ will be *arbitrarily sensitive* to the units of Y. Specifically, we consider continuous, increasing functions $m(\cdot)$ that approximate $\log(y)$ for large values of y in the sense that $m(y)/\log(y) \to 1$ as $y \to \infty$. The common $\log(1+y)$ and $\text{arcsinh}(y)$ transformations satisfy this property. We show that if the treatment affects the extensive margin (i.e. $P(Y(1)=0) \neq P(Y(0)=0)$), then one can obtain any magnitude for the ATE for $m(Y)$ by rescaling the outcome by some positive factor a. It is therefore inappropriate to interpret the ATE for $m(Y)$ as a percentage effect, since a percentage is inherently a unit-invariant quantity, while the ATE for $m(Y)$ depends arbitrarily on the units of Y.

[2] That is, $\log(Y(1)/Y(0)) \approx \frac{Y(1)-Y(0)}{Y(0)}$ when $Y(1)/Y(0) \approx 1$.

[3] We found 17 papers overall using $\text{arcsinh}(Y)$ as an outcome variable, of which 11 interpret the units; see Table C.1.

The intuition for this result is that a "percentage" treatment effect is not well-defined for an individual for whom treatment increases their outcome from zero to a positive value. For example, in our application to Carranza *et al.* (2022) in Section 3.5, the treatment induces more people to have positive hours worked. The percentage change in hours is then not well-defined for individuals who would work positive hours under the treatment condition but zero hours under the control condition. Any average treatment effect that is well-defined with zero-valued outcomes must therefore implicitly assign a value for a change along the extensive margin. For logarithm-like transformations $m(\cdot)$, the importance of the extensive margin is determined implicitly by the units of Y. To see why this is the case, consider an individual who works positive hours only if they are treated, so that $Y(1) > 0$ and $Y(0) = 0$. Their treatment effect for the transformed outcome $m(Y)$ is $m(Y(1)) - m(0)$, which becomes larger if the units of Y are re-scaled by some $a > 1$, e.g. if we convert from weekly hours worked to yearly hours worked. When the treatment has an extensive margin effect, the ATE for $m(Y)$ can thus be made large in magnitude by re-scaling Y by a large factor a. By contrast, if we re-scale Y by a small factor $a \approx 0$, such that the resulting outcomes are close to zero, then $m(Y) \approx m(0)$, and so the ATE for $m(Y)$ will be small. By varying the units of the outcome, we can thus obtain any magnitude for the ATE for $m(Y)$.

Our theoretical results also imply that if we re-scale the units of the outcome by a finite factor $a > 0$, the ATE for a log-like transformation $m(Y)$ will change by approximately $\log(a)$ times the effect of the treatment on the extensive margin. This result implies that sensitivity analyses that explore how the estimated ATE for $m(Y)$ changes with finite changes in the units of Y—or equivalently, how the ATE for $\log(c + Y)$ changes with the constant c—are essentially indirectly measuring the size of the treatment effect on the extensive margin.

We illustrate the practical importance of these results by systematically replicating recent papers published in the *American Economic Review* that estimate treatment effects for $\operatorname{arcsinh}$-transformed outcomes. In line with our theoretical results, we find that treatment effect estimates using $\operatorname{arcsinh}(Y)$ are sensitive to changes in the units of the outcome, particularly when the extensive margin effect is large. In half of the papers that we replicated, multiplying the original outcome by a factor of 100 (e.g. converting from dollars to cents) changes the estimated treatment effect by more than 100% of the original estimate. We obtain similar results using $\log(1+Y)$ instead of $\operatorname{arcsinh}(Y)$.

What, then, are alternative options in settings with zero-valued outcomes? Our second main result delineates the possibilities. We show that when there are zero-valued outcomes, there is no treatment effect parameter that satisfies all three of the following properties:

(a) The parameter is an average of individual-level treatment effects, i.e. takes the form $\theta_g = E_P[g(Y(1), Y(0))]$, where g is increasing in $Y(1)$.

(b) The parameter is invariant to re-scaling of the units of the outcome (i.e. $g(y_1, y_0) = g(ay_1, ay_0)$).

(c) The parameter is point-identified from the marginal distributions of the potential outcomes.

This "trilemma" implies that any target parameter that is well-defined with zero-valued outcomes must necessarily jettison at least one of the three properties above. Of course, the choice of target parameter should depend on the economic question of interest. Which of the three properties the researcher prefers to forgo will thus generally depend on their context-specific motivation for using a log-like transformation in the first place.

To that end, Section 3.4 highlights a menu of parameters that may be attractive depending on the researcher's core motivation. We first consider the case where the researcher is interested in obtaining a causal parameter with an intuitive "percentage" interpretation. In this case, it may be natural to consider a parameter outside of the class of individual-level averages of the form $E_P[g(Y(1), Y(0))]$. One prominent option is $\theta_{\text{ATE}\%} = \frac{E[Y(1)-Y(0)]}{E[Y(0)]}$, the ATE in levels as a percentage of the baseline mean, which in many cases can be estimated via Poisson regression (Santos Silva and Tenreyro, 2006; Wooldridge, 2010). The researcher might also consider alternative normalizations of the outcome that lead to intuitive units, e.g. expressing the outcome in per-capita units or converting it to a rank with respect to some reference distribution. Next, we suppose the researcher would like to capture concave preferences over the outcome; for example, the researcher might consider income gains to be more meaningful for individuals who are initially poor. In this case, it is natural to directly specify how much the researcher values a change along the extensive margin relative to the intensive margin—e.g., that a change from 0 to 1 is worth an x percent change along the intensive margin. Finally, suppose the researcher is interested in separately understanding the effects of the treatment along both the intensive and extensive margins. In this case, the researcher may target separate parameters for the two margins—e.g., $E[\log(Y(1)) - \log(Y(0)) \mid Y(1) > 0, Y(0) > 0]$, the average effect in logs for individuals with positive outcomes under both treatments, captures the intensive margin. Separate effects for the two margins are not generally point-identified, but can be can be bounded using the method in Lee (2009) or point-identified with additional assumptions (Zhang *et al.*, 2008, 2009).

Section 3.5 provides a blueprint for estimating these alternative parameters in practice by applying our recommended approaches to three recent empirical applications, including a randomized controlled trial (RCT) (Carranza *et al.*, 2022), a difference-in-differences (DiD) setting (Sequeira, 2016), and an instrumental variables (IV) setting (Berkouwer and Dean, 2022).

Related work. The use of log-like transformations for dealing with zero-valued outcomes has a long history. The use of the $\log(1 + Y)$ transformation dates to at least Williams (1937), while Bartlett (1947) considers both the $\log(1+Y)$ and inverse hyperbolic sine transformations.[4] More recent papers by Burbidge *et al.* (1988) and Bellemare and Wichman (2020), among others, provide results for $\operatorname{arcsinh}(Y)$ that are frequently cited in economics papers using this transformation.[5]

Previous work has illustrated in simulations or selected empirical applications that results for particular transformations such as $\log(1 + Y)$ or $\operatorname{arcsinh}(Y)$ may be sensitive to the units of the outcome (Aihounton and Henningsen, 2021; de Brauw and Herskowitz, 2021). In concurrent work, Mullahy and Norton (2023) show theoretically that the marginal effects from linear regressions using $\log(1+Y)$ or $\operatorname{arcsinh}(Y)$ are sensitive to the scaling of the outcome, with the the limits of the marginal effects approaching those of either a levels regression or a (normalized) linear probability model, depending on whether the units are made small or large. We complement this work by proving that scale-dependence is a necessary feature of *any* identified ATE that is well-defined with zero-valued outcomes, and that the dependence on units is arbitrarily bad for transformations that approximate $\log(Y)$ for large values of Y. Thus, it is not possible to fix the issues with $\log(1 + Y)$ or $\operatorname{arcsinh}(Y)$ by choosing a "better" transformation or using a different estimator. We also complement previous

[4] Bartlett (1947) proposes using $\operatorname{arcsinh}(\sqrt{Y})$.

[5] MacKinnon and Magee (1990) propose transformations of the form $\operatorname{arcsinh}(y\zeta)/\zeta$, where ζ is estimated by assuming $\operatorname{arcsinh}(y\zeta)/\zeta$ is normally distributed conditional on covariates.

empirical examples by providing a systematic analysis of the sensitivity to scaling for papers in the *American Economic Review* using $\text{arcsinh}(Y)$.

Other work has considered the interpretation of regressions using $\text{arcsinh}(Y)$ or $\log(1 + Y)$ from the perspective of structural equations models, as opposed to the potential outcomes model considered here. This literature has reached diverging conclusions: For example, Bellemare and Wichman (2020) conclude that coefficients from $\text{arcsinh}(Y)$ regressions have an interpretation as a semi-elasticity, while Cohn *et al.* (2022) conclude that these estimators are inconsistent and advocate for Poisson regression instead. Thakral and Tô (2023) show that the semi-elasticities implied by OLS regressions using $\text{arcsinh}(Y)$ or $\log(1 + Y)$ are sensitive to scale; they recommend instead the use of power functions Y^k, which they show are the only transformations (besides $\log$) for which the implied semi-elasticities for OLS regressions are scale-invariant. In Section C.3, we show that these diverging conclusions stem from the fact that the structural equations considered in these papers implicitly impose different restrictions on the potential outcomes—some of which are incompatible with zero-valued outcomes—and consider different target causal parameters. This highlights the value of a potential outcomes framework such as ours, which makes transparent what causal parameters are identifiable and what properties they can have.

Finally, there is a long history in econometrics of explicitly modeling the intensive and extensive margins in settings with zero-valued outcomes, such as Tobin (1958) and Heckman (1979). Broadly speaking, these methods impose parametric structure on the joint distribution of the potential outcomes, which allows one to separate out the intensive and extensive margin effects of a treatment (see Section C.3 for technical details). Of course, the parametric restrictions underlying these approaches may often be difficult to justify in practice, which perhaps has contributed to the growth in the use of log-like transformations in place of approaches that explicitly model the extensive margin. Our paper shows that the presence of an extensive margin should not simply be ignored by taking a log-like transformation. It also clarifies what parameters can be learned in such cases without imposing restrictions on the joint distribution of the potential outcomes.

3.1.1 Setup and notation

Let $D \in \{0, 1\}$ be a binary treatment and let $Y \in [0, \infty)$ be a weakly positively-valued outcome.[6] We assume that $Y = DY(1) + (1-D)Y(0)$, where $Y(1)$ and $Y(0)$ are respectively the potential outcomes under treatment and control. We suppose that in some (sub-)population of interest, $(Y(1), Y(0)) \sim P$ for some (unknown) joint distribution P. We denote the marginal distribution of $Y(d)$ under P by $P_{Y(d)}$ for $d = 0, 1$. We assume that neither $P_{Y(0)}$ nor $P_{Y(1)}$ is a degenerate distribution at zero.

3.2 Sensitivity to scaling for transformations that behave like $\log(Y)$

We first consider average treatment effects of the form $\theta = E_P[m(Y(1)) - m(Y(0))]$ for an increasing function m. We note that θ corresponds to the ATE among the (sub-)population indexed by P; if P refers to the sub-population of compliers for an instrument, for instance, then θ is the local average treatment effect (LATE), rather than the ATE in the full population. We are interested in how θ changes if we change the units of Y by a factor of a. That is, how does

$$\theta(a) = E_P[m(aY(1)) - m(aY(0))]$$

depend on a? Setting $a = 100$, for example, might correspond with a change in units between dollars and cents. Of course, if Y is strictly positive and $m(y) = \log(y)$, then $\theta(a)$ is the ATE in logs and does not depend on the value of a.

We consider "log-like" functions $m(y)$ that are well-defined at zero but behave like $\log(y)$ for large values of y, in the sense that $m(y)/\log(y) \to 1$ as $y \to \infty$. This property is satisfied by $\log(1 + y)$ and $\operatorname{arcsinh}(y)$, for example. Our first main result shows that if the treatment affects the extensive margin, then $|\theta(a)|$ can be made to take any desired value through the appropriate choice of a.

Proposition 3.2.1. *Suppose that:*

1. *(The function m is continuous and increasing) $m : [0, \infty) \to \mathbb{R}$ is a continuous, weakly increasing function.*

[6] The $\operatorname{arcsinh}$ transformation is sometimes used in settings where Y can be negative. We impose that $Y \in [0, \infty)$, and thus do not consider this case. See Section C.2.2 for extensions of our results to settings with continuous treatments.

2. *(The function m behaves like $\log$ for large values) $m(y)/\log(y) \to 1$ as $y \to \infty$.*

3. *(Treatment affects the extensive margin) $P(Y(1) = 0) \neq P(Y(0) = 0)$.*

4. *(Finite expectations) $E_{P_{Y(d)}}[|\log(Y(d))| \mid Y(d) > 0] < \infty$ for $d = 0, 1$.*[7]

Then, for every $\theta^* \in (0, \infty)$, there exists an $a > 0$ such that $|\theta(a)| = \theta^*$. In particular, $\theta(a)$ is continuous with $\theta(a) \to 0$ as $a \to 0$ and $|\theta(a)| \to \infty$ as $a \to \infty$.

Theorem 3.2.1 casts serious doubt on the interpretation of ATEs for functions like $\log(1 + Y)$ or $\text{arcsinh}(Y)$ as (approximate) average percentage effects. While a percent (or log point) is entirely invariant to the units of the outcome, Theorem 3.2.1 shows that, in sharp contrast, the ATEs for these transformations are *arbitrarily* dependent on units.

3.2.1 Intuition for Theorem 3.2.1

Loosely speaking, the result in Theorem 3.2.1 follows from the fact that a "percentage" treatment effect is not well-defined for individuals who have $Y(0) = 0$ but $Y(1) > 0$.[8] Any ATE that is well-defined with zero-valued outcomes must implicitly determine how much weight to place on changes along the extensive margin relative to proportional changes along the intensive margin.

When $m(Y)$ behaves like $\log(Y)$ for large values of Y, the importance of the extensive margin is implicitly determined by the units of Y. For intuition, suppose that we re-scale the outcomes so that the non-zero values of Y are very large. Then for an individual for whom treatment changes the outcome from zero to non-zero, the treatment effect will be very large, since $m(Y(1)) \gg m(Y(0)) = m(0)$. Extensive margin treatment effects thus have a large impact on the ATE when the values of Y are made large. By contrast, changing the units of Y does not change the importance of treatment effects along the intensive margin by much, since for $Y(1) > 0$ and $Y(0) > 0$, we have that $m(Y(1)) - m(Y(0)) \approx \log(Y(1)/Y(0))$, which does not depend on the units of the outcome.

To see the roles of the extensive and intensive margins more formally, for simplicity consider the

[7] This assumption simply ensures that $E_{P_{Y(d)}}[|m(aY(d))| \mid Y > 0]$ exists for all values of $a > 0$.

[8] See Delius and Sterck (2020) for an intuitive discussion of this difficulty in the context of the $\text{arcsinh}(\cdot)$ transformation. They write, "the concept of elasticity itself does not make sense with zeros" (p. 21).

case where $P(Y(1)=0, Y(0)>0)=0$, so that, for example, everyone who has positive income without receiving a training also has positive income when receiving the training.[9] Then, by the law of iterated expectations, we can write

$$\begin{aligned}
&E[m(aY(1)) - m(aY(0))] \\
&\quad = P(Y(1)>0, Y(0)>0)\, \underbrace{E_P[m(aY(1)) - m(aY(0)) \mid Y(1)>0, Y(0)>0]}_{\text{Intensive margin}} \\
&\quad + P(Y(1)>0, Y(0)=0)\, \underbrace{E_P[m(aY(1)) - m(0) \mid Y(1)>0, Y(0)=0]}_{\text{Extensive margin}}.
\end{aligned}$$

When a is large, $m(ay) \approx \log(ay)$ for non-zero values of y, and thus the intensive margin effect in the previous display is approximately equal to $E_P[\log(Y(1)) - \log(Y(0)) \mid Y(1)>0, Y(0)>0]$, the treatment effect in logs for individuals with positive outcomes under both treatment and control. This, of course, does not depend on the scaling of the outcome. However, the extensive margin effect grows with a, since $m(aY(1)) \approx \log(a) + \log(Y(1))$ is increasing in a while $m(0)$ does not depend on a. Thus, as a grows large, the ATE for $m(aY)$ places more and more weight on the extensive margin effect of the treatment relative to the intensive margin. We can therefore make $|\theta(a)|$ arbitrarily large by sending $a \to \infty$. By contrast, if $a \approx 0$, then $m(aY(d)) \approx 0$ with very high probability, and thus the ATE for $m(aY)$ is approximately equal to 0.

It is worth emphasizing that the arbitrary scale-dependence described in Theorem 3.2.1 exists whenever the treatment affects the probability that the outcome is zero, regardless of whether the extensive margin is of direct economic interest or not.[10] In some settings, the presence of zeros may correspond to a discrete economic choice (e.g. not participating in the labor market), and thus may be of direct interest. In other settings—for example, if the outcome is a yearly count of publications which is sometimes zero for idiosyncratic reasons—the extensive margin may be a "nuisance" rather than a direct economic object of interest.[11] The result in Theorem 3.2.1 highlights that regardless

[9] A similar argument goes through without this restriction, but then there are two extensive margins, one for individuals with $Y(1)>0=Y(0)$, and the other for those with $Y(0)>Y(1)=0$.

[10] Without an extensive margin, ATEs for transformations $m(\cdot)$ defined at zero still exhibit scale-dependence, though perhaps not arbitrarily so. See Section 3.3.1 below for further discussion.

[11] One setting where nuisance zeros may arise is when the observed outcome Y is actually a mis-measured version of

of the source of the zeros, an ATE for a log-like transformation is not interpretable as a percentage, since the presence of the extensive margin effect makes it arbitrarily dependent on the units. Indeed, a percentage effect is not a well-defined for individuals moving from zero to non-zero outcomes. Whether the zeros correspond to a discrete economic choice or not will be relevant, however, when considering the choice of alternative target parameter, a topic we return to in Section 3.4 below.

Intuition for the special case of $\log(1+Y)$

We can also develop some intuition for Theorem 3.2.1 by considering the special case where $m(y) = \log(1+y)$. In that case, we have that

$$\theta(a) = E[\log(1+aY(1)) - \log(1+aY(0))] = E\left[\log\left(\frac{1+aY(1)}{1+aY(0)}\right)\right]. \tag{3.1}$$

Note that

$$\lim_{a\to\infty} \log\left(\frac{1+aY(1)}{1+aY(0)}\right) = \begin{cases} \log\left(\frac{Y(1)}{Y(0)}\right) & \text{if } Y(1) > 0, Y(0) > 0 \\ 0 & \text{if } Y(1) = 0, Y(0) = 0 \\ \infty & \text{if } Y(1) > 0, Y(0) = 0 \\ -\infty & \text{if } Y(1) = 0, Y(0) > 0. \end{cases}$$

We thus see that the term inside the expectation in (3.1) diverges to ∞ for individuals with $Y(1) > 0, Y(0) = 0$, and likewise diverges to $-\infty$ when $Y(1) = 0, Y(0) > 0$. If on average the extensive margin effect is positive, then there are more individuals for whom the limit is $+\infty$ rather than $-\infty$, and thus (under appropriate regularity conditions) the ATE diverges to ∞. Analogously, if the extensive margin effect is negative, then the ATE diverges to $-\infty$. Hence, we see that the magnitude of the ATE for $\log(1+aY)$ diverges as $a \to \infty$ when the average effect on the extensive margin is non-zero. By contrast, as $a \to 0$, $\log(1+aY(d)) \to \log(1) = 0$ for both $d = 0$ and $d = 1$, and thus the treatment effect converges to 0. Theorem 3.2.1 shows that this dependence on units occurs for *any* log-like transformation, not just $\log(1+Y)$, and thus this issue cannot be fixed by choosing a different log-like

the true economic object of interest. For example, publications Y may be a noisy measure of true researcher productivity $Y^* > 0$. One possible remedy in this setting is to model the measurement error to recover the treatment effect on Y^* rather than on Y. In a similar vein, Gandhi *et al.* (2023) models the measurement error in product shares in demand estimation, which are sometimes zero in finite samples.

transformation ($\log(c+Y)$, $\text{arcsinh}(Y)$, $\text{arcsinh}(\sqrt{Y})$, etc.)

3.2.2 Additional remarks and extensions

Remark 3.2.2 (ATEs for $\log(c+Y)$)**.** In some settings, researchers consider the ATE for $\log(c+Y)$ and investigate sensitivity to the parameter c. Observe that $\log(1+aY) = \log(a(1/a+Y)) = \log(a) + \log(1/a+Y)$, and thus the ATE for $\log(1+aY)$ is equal to the ATE for $\log(1/a+Y)$. Hence, varying the constant term for $\log(c+Y)$ is equivalent to varying the scaling of the outcome when using $m(y) = \log(1+y)$. Theorem 3.2.1 thus implies that if treatment affects the extensive margin, one can obtain any desired magnitude for the ATE for $\log(c+Y)$ via the choice of c. In particular, the ATE for $\log(c+Y)$ grows large in magnitude as $c \to 0$, and small as $c \to \infty$.

Remark 3.2.3 (Finite changes in scaling)**.** Theorem 3.2.1 shows that any magnitude of $|\theta(a)|$ can be achieved via the appropriate choice of a. How much does $\theta(a)$ change for finite changes in the scaling a? Theorem C.2.1 in the appendix shows that the change in the ATE from multiplying the outcome by a large factor a is approximately $\log(a)$ times the treatment effect on the extensive margin,[12]

$$E_P[m(aY(1)) - m(aY(0))] = (P(Y(1)>0) - P(Y(0)>0)) \cdot \log(a) + o(\log(a)). \tag{3.2}$$

Thus, the ATE for $m(Y)$ will tend to be more sensitive to finite changes in scale the larger is the extensive margin treatment effect. This implies that sensitivity analyses that assess how treatment effect estimates for $m(Y)$ change under finite changes in the units of Y—or equivalently, under finite changes of c in $\log(c+Y)$—are roughly equivalent to measuring the size of the extensive margin.

Remark 3.2.4 (Extension to continuous treatments)**.** We focus on ATEs for binary treatments for expositional simplicity, although similar results apply with continuous treatments. In Section C.2.2, we show that when d is a continuous treatment, any treatment effect contrast that averages $m(aY(d))$ across possible values of d (i.e. a parameter of the form $\int \omega(d)E[m(aY(d))]$) is arbitrarily sensitive to scaling when there is an extensive margin effect.

Remark 3.2.5 (Extension to OLS estimands)**.** It is worth noting that the results in this section show that

[12] We say $f(a) = o(g(a))$ if $\lim_{a\to\infty} |f(a)/g(a)| = 0$. That is, as $a \to \infty$, $|f(a)|$ grows strictly slower than $|g(a)|$.

population ATEs for $m(Y)$ are sensitive to the units of Y. These results are about *estimands*, and thus any consistent *estimator* of the ATE for $m(Y)$ will be sensitive to scaling (at least asymptotically). Thus, our results apply to ordinary least squares (OLS) estimators when they have a causal interpretation, but also to non-linear estimators such as inverse-probability weighting or doubly-robust methods. Nevertheless, given the prominence of OLS in applied work, and the fact that OLS is sometimes used for non-causal estimands, in Section C.2.3 we provide a result specifically on the scale-sensitivity of the population regression coefficient for a random variable of the form $m(Y)$ on an arbitrary random variable X. Our result shows that the coefficients on X will be arbitrarily sensitive to the scaling of Y when the coefficients of a regression of $\mathbb{1}[Y > 0]$ on X are non-zero. Thus, the OLS estimand using a logarithm-like function on the left-hand side will be sensitive to scaling even when it does not have a causal interpretation.

Remark 3.2.6 (Statistical significance)**.** Equation (3.2) shows that $P(Y(1) > 0) - P(Y(0) > 0)$ is the dominant term in $\theta(a)$ for large a, which suggests that the t-statistic for an estimator of $\theta(a)$ will generally converge to that for the analogous estimator of the extensive margin effect, $P(Y(1) > 0) - P(Y(0) > 0)$. Theorem C.2.4 in the appendix formalizes this intuition when the treatment effects are estimated via OLS: As a is made large, the t-statistic for $\hat{\theta}(a)$ converges to that for the extensive margin estimate. In our empirical analysis of papers in the *American Economic Review* below, we find that indeed the t-statistics for estimates of the ATE using $\text{arcsinh}(Y)$ are typically close to those for the extensive margin effect.

Remark 3.2.7 (When most values are large)**.** Researchers often have the intuition that if most of the values of the outcome are large, then ATEs for transformations like $\log(1 + Y)$ or $\text{arcsinh}(Y)$ will approximate elasticities, since $m(Y) \approx \log(Y)$ for most values of Y. Indeed, in an influential paper, Bellemare and Wichman (2020) recommend that researchers using the $\text{arcsinh}(Y)$ transformation should transform the units of their outcome so that most of the non-zero values of Y are large. The results in this section suggest—perhaps somewhat counterintuitively—that if one rescales the outcome such that the non-zero values are all large, the behavior of the ATE will be driven nearly entirely by the effect of the treatment on the extensive margin and *not* by the distribution of the potential outcomes conditional on being positive. Moreover, the rescaling can be chosen to generate any magnitude for the ATE if the treatment affects the extensive margin.

Remark 3.2.8 (Zero extensive margin)**.** Theorem 3.2.1 applies to settings where treatment has a non-zero effect on average on the extensive margin. This raises the question of whether the use of log-like transformations is justified in the absence of an extensive margin treatment effect. Our Theorem C.1.3 below implies that the ATE for any log-like transformation will be sensitive to the units of the outcome for at least some distribution with *strictly* positive outcomes, but perhaps not arbitrarily so in the sense of Theorem 3.2.1 (see Section 3.3.1 for further discussion). Moreover, if one were confident that the extensive margin effect were exactly zero for all individuals, one could recover the ATE in logs for individuals with positive outcomes by simply dropping individuals with $Y = 0$. The use of log-like transformations is thus somewhat difficult to justify even in settings without an extensive margin.

3.2.3 Empirical illustrations from the *American Economic Review*

We illustrate the results in this section by evaluating the sensitivity to scaling of estimates using the $\operatorname{arcsinh}(Y)$ transformation in recent papers in the *American Economic Review* (*AER*). In November 2022, we used Google Scholar to search for "inverse hyperbolic sine" among papers published in the *AER* since 2018. We searched for papers using $\operatorname{arcsinh}(Y)$ rather than $\log(1 + Y)$ since the former are easier to find with a simple keyword search. Our search returned 17 papers that estimate treatment effects for an $\operatorname{arcsinh}$-transformed outcome.[13] Of these, 10 explicitly interpret the results as percentage changes or elasticities, and 6 of the remaining 7 do not directly interpret the units. See Table C.1 for a list of the papers and relevant quotes. Of the 17 total papers using $\operatorname{arcsinh}(Y)$, 10 had publicly available replication data that allowed us to replicate the original estimates and assess their sensitivity to scaling.[14] For our replications, we focus on the first specification using $\operatorname{arcsinh}(Y)$ presented in a table in the paper, which we view as a reasonable proxy for the paper's main specification using

[13] We consider papers with both binary and non-binary treatments, as our theoretical results extend easily to non-binary treatments; see Theorem 3.2.4. Seven of the 10 papers we replicated used a binary treatment.

[14] We include one paper where there was a slight discrepancy between our replication of the original result and the result reported in the paper that only affected the third decimal place.

$\operatorname{arcsinh}(Y)$.[15]

We assess the sensitivity of these results by re-running exactly the same procedure as in the original paper, except replacing $\operatorname{arcsinh}(Y)$ with $\operatorname{arcsinh}(100 \cdot Y)$. Thus, for example, if the original paper estimated a treatment effect for the arcsinh of an outcome measured in dollars, we use the same procedure to re-estimate the treatment effect for the arcsinh of the outcome measured in cents. Since (3.2) shows that the sensitivity to scaling depends on the size of the extensive margin effect, we also estimate the extensive margin effect by using the same procedure as in the original paper but with the outcome $\mathbb{1}[Y > 0]$.

Table 3.1: Change in estimated treatment effects from re-scaling the outcome by a factor of 100 in papers published in the *AER* using $\operatorname{arcsinh}(Y)$

	Treatment Effect Using:			Change from rescaling units:	
Paper	$\operatorname{arcsinh}(Y)$	$\operatorname{arcsinh}(100 \cdot Y)$	Ext. Margin	Raw	%
Azoulay et al (2019)	0.003	0.017	0.003	0.014	464
Fetzer et al (2021)	-0.177	-0.451	-0.059	-0.273	154
Johnson (2020)	-0.179	-0.448	-0.057	-0.269	150
Carranza et al (2022)	0.201	0.453	0.055	0.252	125
Cao and Chen (2022)	0.038	0.082	0.010	0.044	117
Rogall (2021)	1.248	2.150	0.195	0.902	72
Moretti (2021)	0.054	0.068	0.000	0.013	24
Berkouwer and Dean (2022)	-0.498	-0.478	0.010	0.020	-4
Arora et al (2021)	0.113	0.110	-0.001	-0.003	-3
Hjort and Poulsen (2019)	0.354	0.354	0.000	0.000	0

Note: This table replicates treatment effect estimates using $\operatorname{arcsinh}(Y)$ as the outcome in recent papers published in the *AER*, and explores their sensitivity to the units of Y. The first column shows the author(s) and date of the paper. The second column shows the treatment effect on $\operatorname{arcsinh}(Y)$ using the units originally reported in the paper. The third column shows a treatment effect estimate constructed identically to the estimate in column 2 except using $\operatorname{arcsinh}(100 \cdot Y)$ as the outcome instead of $\operatorname{arcsinh}(Y)$, e.g. converting Y from dollars to cents before taking the arcsinh transformation. The fourth column shows an estimate of the size of the extensive margin, obtained using $\mathbb{1}[Y > 0]$ as the outcome. The final two columns show the raw difference and percentage difference between the second and third columns. The table is sorted on the magnitude of the percentage difference.

The results of this exercise, shown in Table 3.1, illustrate that treatment effect estimates can be

[15] We use the first coefficient presented in a figure for one paper without any tables in the main text using $\operatorname{arcsinh}(Y)$. If the first specification is a validation check (e.g. a pre-trends test), we use the first specification of causal interest.

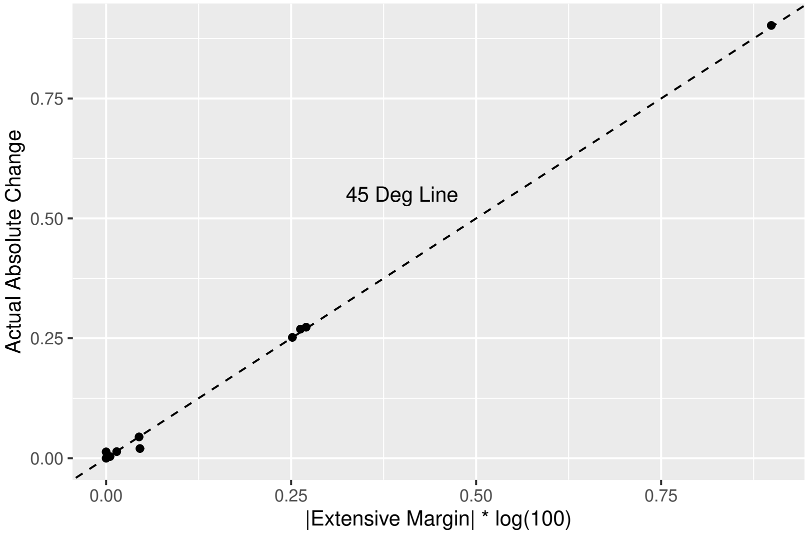


Figure 3.1: Change from multiplying outcome by 100 versus extensive margin effect

Note: This figure shows the relationship between the sensitivity of treatment effects using $\operatorname{arcsinh}(Y)$ to re-scaling the units of Y and and the size of the extensive margin. For each replicated paper, this figure plots the absolute value of the change in the estimated treatment effect from multiplying the outcome by 100 (i.e. the absolute value of the Raw Change column in Table 3.1) on the y-axis against $\log(100)$ times the absolute value of the extensive margin effect on the x-axis. If the approximation in (3.2) were exact, all points would lie on the 45 degree line.

quite sensitive to the scaling of the outcome when the extensive margin is not approximately zero. Indeed, in 5 of the 10 replicable papers, multiplying the outcome by a factor of 100 changes the estimated treatment effect by more than 100% of the original estimate. The change in the estimated treatment effect is less than 10% only in three papers, all of which have either zero or near-zero (<1 p.p.) effects on the extensive margin. Figure 3.1 shows that the (absolute) change in the estimated treatment effect is larger when the extensive margin effect is larger, with the change lining up very closely with the approximation given in (3.2).[16]

Using the same 10 papers, we also estimate treatment effects using $\log(1 + Y)$ as the outcome, and analogously explore how the results change when we multiply the units of Y by 100. (Four of

[16] In Figure C.1, we plot the t-statistics for the treatment effects estimates as well as those for the extensive margin effect. In line with the discussion in Theorem 3.2.6, we find that the t-statistics for the treatment effect using $\operatorname{arcsinh}(Y)$ tend to be similar to those for the extensive margin, except when the extensive margin is very small, and become even closer when multiplying the units by 100.

the 10 papers that we replicate report an alternative specification using $\log(1 + Y)$ in the paper.) The results, shown in Table C.2, are qualitatively quite similar those in Table 3.1, with five of the 10 treatment effect estimates again changing by more than 100%. These results underscore the fact that Theorem 3.2.1 applies to *all* log-like transformations, including both $\text{arcsinh}(Y)$ and $\log(c + Y)$ for any constant c.

3.3 Sensitivity to scaling for other ATEs

Our results so far show that ATEs for transformations that are defined at zero and approximate $\log(y)$ are arbitrarily sensitive to scaling. What other options are available when there are zero-valued outcomes? To help delineate alternative options, in this section we provide a result showing what properties a parameter defined with zero-valued outcomes can have. Specifically, we establish a "trilemma": When there are zero-valued outcomes, there is no parameter that is (a) an average of individual-level treatment effects of the form $\theta_g = E_P[g(Y(1), Y(0))]$, (b) scale-invariant, and (c) point-identified.[17] Any approach for settings with zero-valued outcomes must therefore abandon one of the properties (a)–(c); in Section 3.4 below we discuss several approaches that relax one (or more) of these requirements.

Before stating our formal result, we must make precise what we mean by scale-invariance and point-identification. We say that g is scale-invariant if its value is the same under any re-scaling of the units of y by a positive constant a.

Definition 3.3.1. We say that the function g is *scale-invariant* if it is homogeneous of degree zero, i.e. $g(y_1, y_0) = g(ay_1, ay_0)$ for all $a, y_1, y_0 > 0$.

We next describe point-identification. We consider parameters that are identified without placing restrictions on treatment effect heterogeneity. As in Fan *et al.* (2017), this is formalized by considering parameters that can be learned if we know the marginal distributions of $Y(1)$ and $Y(0)$, but not the

[17] Of course, not all parameters of the form $E_P[g(Y(1), Y(0))]$ can be interpreted as an average of individual treatment effects. For example $E[\mathbb{1}[Y(1) > 0, Y(0) > 0]]$ is the fraction of individuals whose outcomes is positive under both treatments, rather than a treatment effect. Our results apply to all parameters of this form, regardless of whether they are average treatment effects *per se*.

full joint distribution of $(Y(1), Y(0))$.

To connect treatment effect heterogeneity to the joint distribution of potential outcomes, consider the simple case of a randomized experiment. By examining the outcome distribution for the treated group, we can learn the marginal distribution of $Y(1)$. Likewise, by examining the outcome distribution for the control group, we can learn the marginal distribution of $Y(0)$. If treatment effects were assumed to be constant, then for each observed treated unit with outcome $Y(1)$, we could infer their untreated outcome as $Y(0) = Y(1) - \tau$, where τ is the average treatment effect. Hence, the joint distribution of $(Y(1), Y(0))$ would be identified. However, if we allow for treatment effect heterogeneity, then for an observed treated unit with outcome $Y(1)$, we do not know what their value of $Y(0)$ would be, and thus we do not know the joint distribution of $(Y(1), Y(0))$. This winds up being especially important in settings with an extensive margin, since when we observe the distribution of outcomes for treated units, it means that we do not know *which* of the treated units would have had a zero outcome under the control condition, and thus it is difficult to disentangle the intensive and extensive margins.[18]

With that intuition in mind, we now give a formal definition. Recall that P denotes the joint distribution of $(Y(1), Y(0))$, while $P_{Y(d)}$ denotes the marginal distribution of $Y(d)$. We then say θ_g is point-identified if it depends on P only through the marginals $P_{Y(1)}, P_{Y(0)}$.

Definition 3.3.2 (Identification)**.** We say that θ_g is *point-identified from the marginals at* P if for every joint distribution Q with the same marginals as P (i.e. such that $Q_{Y(d)} = P_{Y(d)}$ for $d = 0, 1$), $E_P[g(Y(1), Y(0))] = E_Q[g(Y(1), Y(0))]$. For a class of distributions $\mathcal{P}$, we say that θ_g is *point-identified over* $\mathcal{P}$ if for every $P \in \mathcal{P}$, θ_g is point-identified from the marginals at P.

We will denote by $\mathcal{P}_+$ the set of distributions on $[0, \infty)^2$. Thus, θ_g is point-identified over $\mathcal{P}_+$ if it is always identified when Y takes on zero or weakly positive values. Our next result formalizes that it is not possible to have a parameter of the form $E_P[g(Y(1), Y(0))]$ that is both scale-invariant and point-identified over $\mathcal{P}_+$.

Proposition 3.3.3 (A trilemma)**.** *The following three properties cannot hold simultaneously:*

[18] In Section C.3, we discuss a variety of structural approaches that impose assumptions restricting the joint distribution, thus allowing us to separately point-identify the effects for the two margins.

(a) $\theta_g = E_P[g(Y(1), Y(0))]$ *for a non-constant function* $g : [0, \infty)^2 \to \mathbb{R}$ *that is weakly increasing in its first argument.*

(b) *The function* g *is scale-invariant.*

(c) θ_g *is point-identified over* $\mathcal{P}_+$.[19]

Any parameter defined with zero-valued outcomes must therefore abandon one of properties (a)–(c).

As a special case, Theorem 3.3.3 implies that the ATE for any increasing function $m(Y)$ defined at zero cannot be scale-invariant. This is because the ATE for $m(Y)$ takes the form in (a) with $g(y_1, y_0) = m(y_1) - m(y_0)$, and is also point-identified (part (c)). It follows that property (b) must be violated, i.e. there is some $c, y_0, y_1 > 0$ such that $m(cy_1) - m(cy_0) \neq m(y_1) - m(y_0)$. Theorem 3.3.3 thus formalizes the sense in which it is not possible to "fix" the issues with ATEs for log-like transformations described above by taking alternative transformations of the outcome (e.g. $\sqrt{Y}$).

3.3.1 Implications for settings without an extensive margin

The trilemma in Theorem 3.3.3 applies for transformations of the outcome defined at zero. To prove Theorem 3.3.3, however, we establish an even stronger result: the only parameter satisfying properties (a) and (b) that is point-identified over distributions for which Y is *strictly* positively-valued is the ATE in logs.[20] This result, which is formalized in Theorem C.1.3 in the Appendix, has some useful implications for settings in which the outcome is strictly positive.

First, it implies that the ATE for any transformation of the outcome other than $\log(Y)$ will depend on the units of the outcome for at least some DGP where the outcome is strictly positive. The scale-dependence of log-like transformations such as $\log(1 + Y)$ or $\text{arcsinh}(Y)$ is thus not entirely limited

[19] A minor technical complication arises from the fact that $E_P[g(Y(1), Y(0)]$ could be infinite for some P. For the purposes of our result, it suffices to trivially define θ_g to be identified in this case. Alternatively, the same result holds if part (c) is modified to impose only that θ_g is point-identified over all distributions in $\mathcal{P}_+$ with finite support, thus avoiding issues related to undefined expectations.

[20] More precisely, the only such treatment effect is the ATE in logs or an affine tranformations thereof.

to settings with an extensive margin.[21] We note, however, that while the ATE for such transformations may depend on the units of the outcome even without zero-valued outcomes, the dependence need not be arbitrarily bad in the sense of Theorem 3.2.1. Indeed, (3.2) shows that if there is no extensive margin, the ATE for a log-like transformation will be approximately insensitive to scaling once the values of Y are made large. This is intuitive, since if Y is strictly positively-valued, the ATE for a log-like transformation will be approximately equal to the ATE in logs when the values of Y are made large.

Second, Theorem C.1.3 implies that even when $Y(1)$ and $Y(0)$ are strictly-positively valued, the average proportional effect $\theta_{\text{Avg}\%} = E[(Y(1) - Y(0))/Y(0)]$ is not point-identified. This parameter is empirically relevant: For instance, Andrews and Miller (2013) show that in the Baily (1978)–Chetty (2006) model with heterogeneous consumption responses to unemployment, the optimal level of unemployment insurance depends on a parameter of the form $\theta_{\text{Avg}\%}$, where Y is consumption and D is unemployment. Although the ATE in logs may approximate $\theta_{\text{Avg}\%}$ when the proportional effect of the treatment is approximately constant, our results imply that it is not possible to point-identify $\theta_{\text{Avg}\%}$ when allowing for arbitrarily heterogeneous proportional effects.

3.4 Empirical approaches with zero-valued outcomes

Our theoretical results above imply that when there are zero-valued outcomes, the researcher should not take a log-like transformation of the outcome and interpret the resulting ATE as an average percentage effect: Unlike a percentage, such an ATE depends on the units of the outcome. In this section, we highlight some other parameters that are well-defined and easily interpreted when there are zero-valued outcomes; in Section 3.5 below, we show how these parameters can be estimated in three empirical applications. Of course, any alternative parameter must necessarily drop one of the requirements in the trilemma in Theorem 3.3.3, but the choice of which to drop may depend on the researcher's motivation.

To inform our discussion of alternative parameters, it is therefore useful to first enumerate several reasons why empirical researchers may target treatment effects for a log-transformed outcome rather

[21] There is thus no conflict between our results and those in Thakral and Tô (2023), who note that semi-elasticities for OLS regressions using log-like transformations may depend on the units of the outcome even when Y is strictly positively-valued.

than the ATE in levels:

(i) The researcher is interested in reporting a treatment effect parameter with easily-interpretable units, such as "percentage changes."

(ii) The researcher believes that there are decreasing returns to the outcome, and thus wants to place more weight on treatment effects for individuals with low initial outcomes. For instance, the researcher may perceive it to be more meaningful to raise income from $Y(0) =$ \$10,000 to $Y(1) =$ \$20,000 than from $Y(0) =$ \$100,000 to $Y(1) =$ \$110,000, yet both of these treatment effects contribute equally to the ATE in levels.

(iii) The researcher is interested in both the intensive and extensive margin effects of the treatment, and is using the ATE for a log-like transformation as an approximation to the proportional effect along the intensive margin.

These three motivations suggest different ways of breaking out of the trilemma in Theorem 3.3.3. If the goal is to achieve a percentage interpretation, then one can consider scale-invariant parameters outside of the class $E_P[g(Y(1), Y(0))]$. For instance, researchers can consider the ATE in levels expressed as a percentage of the control mean, or the ATE for a normalized parameter $\tilde{Y}$ that already has a percentage interpretation. Alternatively, if the goal is to capture concave social preferences over the outcome, then it is natural to specify how much we value the intensive margin relative to the extensive margin—thus abandoning scale-invariance. Finally, if the goal is to separately understand the intensive margin effect, the researcher can abandon point-identification (from the marginal distributions) and directly target the partially identified parameter $E\left[\log(Y(1)) - \log(Y(0)) \mid Y(0) > 0, Y(1) > 0\right]$, the effect in logs for individuals with positive outcomes under both treatments. We address each of these cases in turn below, with a summary of some possible parameters in Table 3.2.

Remark 3.4.1 (Statistical reasons for transforming the outcome)**.** We focus on settings where the researcher is interested in a parameter other than the ATE in levels. In some settings, the researcher may be interested in the ATE in levels, but simple regression estimators may be noisy owing to a long right-tail of the outcome (Athey *et al.*, 2021b). The researcher might then try to estimate the ATE in levels by first estimating the ATE for a log-like transformation, and then multiplying by the baseline

Table 3.2: Summary of alternative target parameters

Description	Parameter	Main property sacrificed?	Pros/Cons
Normalized ATE	$E[Y(1) - Y(0)]/E[Y(0)]$	$E[g(Y(1), Y(0))]$	*Pro:* Percent interpretation *Con:* Does not capture decreasing returns
Normalized outcome	$E[Y(1)/X - Y(0)/X]$	$E[g(Y(1), Y(0))]$	*Pro:* Per-unit-X interpretation *Con:* Need to find sensible X
Explicit tradeoff of intensive/extensive margins	ATE for $m(y) = \begin{cases} \log(y) & y > 0 \\ -x & y = 0 \end{cases}$	Scale-invariance	*Pro:* Explicit tradeoff of two margins *Con:* Need to choose x; Monotone only if support excludes $(0, e^{-x})$
Intensive margin effect	$E\left[\log\left(\frac{Y(1)}{Y(0)}\right) \mid Y(1) > 0, Y(0) > 0\right]$	Point-identification	*Pro:* ATE in logs for the intensive margin *Con:* Partial identification

mean. However, since the ATE for a log-like transformation depends on the units of the outcome—and is thus not a true "percentage" effect—the validity of this approach for recovering the ATE in levels will depend on the initial units of Y.[22] We refer the reader to Athey *et al.* (2021b) and Müller (2023) for alternative approaches to estimation and inference targeted to settings where the ATE in levels is of interest but the outcome has heavy tails.

Remark 3.4.2 (Transformation-specific identification)**.** Another reason that researchers may consider taking a transformation of the outcome is that a parametric assumption used for identification may be more plausible for some functional forms than others. For example, when the outcome is strictly positive, parallel trends in logs may be more plausible than parallel trends in levels if time-varying factors are thought to have a multiplicative impact on the outcome. We note that justifying parallel trends for a log-like transformation is especially tricky, however, since if parallel trends holds for the $\operatorname{arcsinh}$ of an outcome measured in dollars, say, it will not generally hold for the $\operatorname{arcsinh}$ of the outcome measured in cents (Roth and Sant'Anna, 2023). Thus, the parallel trends assumption is specific to both the transformation $m(\cdot)$ *and* the units of the outcome. Moreover, even if the researcher is confident in parallel trends for a particular log-like transformation and unit of the outcome, our results imply that they should not interpret the resulting ATT as an average percentage effect, since

[22] Even in the case where Y is strictly positive and one first estimates the ATE in logs, this approach will only recover the ATE in levels under certain homogeneity assumptions, e.g. constant proportional effects. See Wooldridge (1992) for related discussion.

that ATT is dependent on the units in which the outcome is measured (Theorem 3.2.1).

In what follows, we consider alternative parameters that may be of interest when the marginal distributions of the potential outcomes are identified for some population of interest. Such identification is obtained in RCTs or under conditional unconfoundedness (for the full population), as well in instrumental variables settings (for the population of compliers), as these designs do not rely on functional form assumptions for identification. If the original identification strategy relies on a functional form assumption (e.g. parallel trends), then obtaining identification of the alternative parameters discussed below may require different identifying assumptions. We discuss these issues in detail in Section 3.5.2, where we revisit the difference-in-differences application in Sequeira (2016).

3.4.1 When the goal is interpretable units

We first consider the case where the researcher's primary goal is to obtain a treatment effect parameter with easily interpretable units, such as percentages.

Normalizing the ATE in levels. One possibility is to target the parameter

$$\theta_{\text{ATE\%}} = \frac{E[Y(1) - Y(0)]}{E[Y(0)]},$$

which is the ATE *in levels* expressed as a *percentage of the control mean*. For example, if a researcher is studying a program D meant to reduce healthcare spending Y, then $\theta_{\text{ATE\%}}$ is the percentage reduction in costs from implementing the program. This parameter is point-identified and scale-invariant, and thus has an intuitive percentage interpretation. Importantly, however, $\theta_{\text{ATE\%}}$ is the percentage change in the average outcome between treatment and control, but is *not* an average of individual-level percentage changes.[23] That is, $\theta_{\text{ATE\%}}$ does not take the form $E_P[g(Y(1), Y(0))]$, thus avoiding the trilemma in Theorem 3.3.3.

We note that $\theta_{\text{ATE\%}}$ is consistently estimable by Poisson regression (see Gourieroux *et al.* (1984); Santos Silva and Tenreyro (2006); Wooldridge (2010, Chapter 18.2)) under an appropriate identifying assumption. With a randomly assigned D, for example, estimation of $Y = \exp(\alpha + \beta D)U$ by Poisson

[23] This is roughly analogous to how quantile treatment effects show changes in the quantiles of the potential outcomes distributions, but *not* the quantiles of the treatment effects (without further assumptions).

quasi-maximum likelihood (QMLE) consistently estimates the population coefficient β, which satisfies $e^{\beta} - 1 = E[Y(1)]/E[Y(0)] - 1 = \theta_{\text{ATE\%}}$. In Section 3.5 below, we illustrate how $\theta_{\text{ATE\%}}$ can be estimated by Poisson regression in practice in several empirical examples, including both an RCT and DiD setting.

We also emphasize that $\theta_{\text{ATE\%}}$ is influenced by treatment effects along both the intensive and extensive margins. In particular, the numerator of $\theta_{\text{ATE\%}}$ is the ATE in levels. Thus, if an individual has a treatment effect of say 1, that contributes the same to $\theta_{\text{ATE\%}}$ regardless of whether their outcome changes from 0 to 1 (an extensive margin change) or 1 to 2 (an intensive margin change). The parameter $\theta_{\text{ATE\%}}$ may therefore be attractive in settings where the researcher does not want to distinguish between the intensive and extensive margins. For example, if Y is a count of publications by a researcher in a particular year, and publications are sometimes zero owing to the idiosyncracies of the publication process, then it may be reasonable to view a change between 0 and 1 as similar to a change between 1 and 2. On the other hand, in settings where a zero corresponds to a distinct economic choice, such as not participating in the labor market, then it may be of interest to separate the effects along the intensive and extensive margin, as we discuss in more detail in Section 3.4.3 below.

It is also worth noting that if the researcher has determined that the ATE in levels is not of economic interest, then similar issues will likely arise for $\theta_{\text{ATE\%}}$, since $\theta_{\text{ATE\%}}$ is just a re-scaling of the ATE in levels. For one, the ATE in levels (and hence $\theta_{\text{ATE\%}}$) imposes no diminishing returns, and thus might be dominated by individuals in the tail of the outcome distribution, particularly when the outcome is skewed. Whether this is warranted will depend on the economic question: if the policy-maker's goal is to reduce healthcare spending, it may not matter whether the savings are produced mainly by reducing spending for a small fraction of individuals with catastrophic medical spending. On the other hand, a policy that increases every American's income by $100 and one that increases Elon Musk's income by $35 billion and has no effect on anyone else would have approximately the same value of $\theta_{\text{ATE\%}}$, yet the former may be vastly preferred by an inequality-minded policy-maker. We therefore next turn to alternative approaches that place less weight on the tails of the outcome distribution.

Normalizing other functionals. While $\theta_{\text{ATE}\%}$ normalizes the ATE by the control mean, one can obtain scale-invariance by normalizing other functionals of the potential outcomes distributions.[24] For example,

$$\theta_{\text{Median}\%} = \frac{\text{Median}(Y(1)) - \text{Median}(Y(0))}{\text{Median}(Y(0))},$$

is the quantile treatment effect at the median normalized by the median of $Y(0)$.[25] Put otherwise, it captures the percentage change in the median between the treated and control distributions. ($\theta_{\text{Median}\%}$ thus may be particularly relevant for politicians interested in maximizing the happiness of the median voter!) As is typically the case with quantile treatment effects, however, the numerator of $\theta_{\text{Median}\%}$ need not correspond to the median of individual-level treatment effects. Moreover, in many settings, decision-makers may care about treatment effects throughout the distribution, not just at the median, in which case $\theta_{\text{Median}\%}$ may not be the most economically-relevant parameter.

Normalizing the outcome. A related approach to obtaining a treatment effect with more intuitive units is to estimate the ATE for a transformed outcome that has a percentage interpretation. One example is to consider an outcome of the form $\tilde{Y} = Y/X$, where Y is the original outcome and X is some pre-determined characteristic. For example, suppose Y is employment in a particular area. The treatment effect in levels for Y may be difficult to interpret, since a change in employment of 1,000 means something very different in New York City versus a small rural town. However, if X is the area's population, then $\tilde{Y}$ is the employment-to-population ratio, which may be more comparable across places, and is already in percentage (i.e. per capita) units. We note that the ATE for $\tilde{Y}$ is a scale-invariant, point-identified parameter of the form $\theta = E_P[g(Y(1), Y(0), X)]$, and thus escapes the trilemma in Theorem 3.3.3 by avoiding property (a).[26] The viability of this approach, of course, depends on having a variable X such that the normalized outcome $\tilde{Y}$ is of economic interest. We suspect that in many contexts, reasonable options will be available, including pre-treatment observations of the outcome (assuming these are positive), or the *predicted* control

[24] Indeed, any functional $\phi(P)$ is homogeneous of degree zero if and only if it can be written as the ratio of two homogeneous of degree one functionals.

[25] Note that $\theta_{\text{Median}\%}$ is well-defined only if $\text{Median}(Y(0)) > 0$.

[26] It is scale-invariant in the sense that $g(y_1, y_0, x) = g(ay_1, ay_0, ax)$.

outcome given some observable characteristics (i.e., $X = E[Y(0) \mid W]$, for observable characteristics W).

A second example is to use $\tilde{Y} = F_{Y^*}(Y)$, where F_{Y^*} is the cumulative distribution function (CDF) of some reference random variable Y^*, as suggested in Delius and Sterck (2020). The transformed outcome $\tilde{Y}$ then corresponds to the rank (i.e. percentile) of an individual in the reference distribution, and the ATE for $\tilde{Y}$ can be interpreted as the average change in rank caused by the treatment. The ATE for $\tilde{Y}$ is unit-invariant so long as Y and Y^* and measured in the same units. Outcomes of this form have become increasingly popular in the literature on intergenerational mobility, where $\tilde{Y}$ corresponds to a child's rank in the national income distribution. This approach has been found to yield more stable estimates than approaches using $\log(c + Y)$, which Chetty *et al.* (2014b) show are sensitive to the choice of c.[27]

Finally, the researcher might report treatment effects on transformed outcomes of the form $1[Y \geq y]$ for different values of y. For example, the researcher might report the impact of the treatment on the probability that an individual earns at least \$50,000, \$60,000, etc., and interpret it as the treatment effect on the probability of obtaining a "well-paying job."[28] Such treatment effects have interpretable units as percentage points (i.e. changes in probabilities). We note that treatment effects for outcomes of this form combine the effect of the treatment along the intensive and extensive margin, since for example, a worker who has $Y(1) > \$50,000 > Y(0)$ could either not work under control ($Y(0) = 0$) or work under control but have earnings below \$50,000.

3.4.2 When the goal is to capture decreasing returns

We next consider the case where the researcher wants to capture some form of decreasing marginal utility over the outcome. For example, when Y is strictly positively valued, the ATE in logs corresponds to the change in utility from implementing the treatment for a utilitarian social planner with log utility over the outcome, $U = E[\log(Y)]$. Intuitively, this social welfare function captures the fact that the

[27] Similar to the discussion in Footnote 22, the treatment effect in ranks cannot be converted back to obtain the ATE in levels without additional assumptions.

[28] The researcher could also report the implied CDF of $Y(1)$ and $Y(0)$, from which one can infer the treatment effect on outcomes of this form for all y.

planner values a percentage point of change in the outcome equally for all individuals, regardless of their initial level of the outcome.

Of course, log utility is not well-defined when there is an extensive margin: A coherent utility function defined with zero-valued outcomes must take a stand on the relative importance of the intensive versus extensive margins. Recall from Section 3.2.1 that when using transformations like $\log(1+y)$ or $\operatorname{arcsinh}(y)$, the scaling of the outcome implicitly determines the weights placed on these margins.

Instead of implicitly weighting the margins via the scaling of Y, a more transparent approach is to explicitly take a stand on how much one values the two margins of treatment. Of course, if one knows that their utility is captured by $U = E[m(Y)]$ (for a particular unit of Y, say earnings in dollars), then the ATE for $m(Y)$ is appropriate. If one is unsure exactly of their utility function, then a rough calibration is to specify how much one values a change in earnings from 0 to 1 relative to a percentage change in earnings for those with non-zero earnings. If, for example, one values the extensive margin effect of moving from 0 to 1 the same as a $100x$ percent increase in earnings, then one might consider setting $m(y) = \log(y)$ for $y > 0$ and $m(0) = -x$. The ATE for this transformation can be interpreted as an approximate percentage (log point) effect, where an increase from 0 to 1 is valued at $100x$ log points.[29]

We emphasize that for a fixed value of x, this approach necessarily depends on the scaling of the outcome (thus avoiding the trilemma in Theorem 3.3.3). However, this may not be so concerning since the appropriate choice of x also depends on the units of the outcome—e.g., saying a change from 0 to 1 is worth $100x$ percent means something very different if 1 corresponds with one dollar versus a million dollars. In other words, ATEs for transformations such as $\operatorname{arcsinh}(Y)$ may be difficult to interpret because the scaling of the outcome implicitly determines the relative importance of the intensive and extensive margins; this approach avoids that difficulty by *explicitly* taking a stand on the tradeoff between these two margins. Nevertheless, a challenge with this approach is that researchers

[29] Note that this transformation will generally only be sensible if the support of Y excludes $(0, e^{-x})$, since otherwise the function $m(y)$ is not monotone in y over the support of Y. It is common, however, to have a lower-bound on non-zero values of the outcome; e.g., a firm cannot have between 0 and 1 employees. In our application to Sequeira (2016) below, we normalize the minimum non-zero value of Y to 1 when applying this approach.

may have differing opinions over the appropriate choice of x (or more generally, over the appropriate utility function).

3.4.3 When the goal is to understand intensive and extensive margins

Finally, we consider the case where the researcher is interested in understanding the intensive and extensive margin effects separately. A common question in the literature on job training programs (Card *et al.*, 2010), for instance, is whether a program raises participants' earnings by helping them find a job—which would be expected only to have an extensive-margin effect—or by increasing human capital, which would be expected to also affect the intensive margin. In such settings, it is natural to target separate parameters for the intensive and extensive margins.

For example, the parameter

$$\theta_{\text{Intensive}} = E[\log(Y(1)) - \log(Y(0)) \mid Y(1) > 0, Y(0) > 0]$$

captures the ATE in logs for those who would have a positive outcome regardless of their treatment status. The parameter $\theta_{\text{Intensive}}$ is scale-invariant but is not point-identified from the marginal distributions of the potential outcomes (thus avoiding the trilemma in Theorem 3.3.3), and therefore cannot be consistently estimated without further assumptions.[30] However, Lee (2009) popularized a method for obtaining bounds on $\theta_{\text{Intensive}}$ under the monotonicity assumption that, for example, everyone with positive earnings without receiving a training would also have positive earnings when receiving the training.[31] Bounds on $\theta_{\text{Intensive}}$ can be reported alongside measures of the extensive margin effect, such as the change in the probability of having a non-zero outcome, $P(Y(1) > 0) - P(Y(0) > 0)$. One can also potentially tighten the bounds (or restore point-identification) by imposing additional assumptions on the joint distribution of the potential outcomes—we provide an example of this in our

[30] $\theta_{\text{Intensive}}$ also does not take the form $E_P[g(Y(1), Y(0))]$, although it can be written as

$$\frac{E_P\left[\mathbb{1}[Y(1) > 0, Y(0) > 0]\log(Y(1)/Y(0))\right]}{E_P[\mathbb{1}[Y(1) > 0, Y(0) > 0]]},$$

where both the numerator and denominator take this form.

[31] See, also, Zhang and Rubin (2003) for related results, including bounds without the monotonicity assumption.

application to Carranza *et al.* (2022) below; see Zhang *et al.* (2008, 2009) for related approaches.[32]

We note that the parameter $\theta_{\text{Intensive}}$ is generally distinct from the "intensive margin" marginal effects implied by two-part models (2PMs), which were recommended for scenarios with zero-valued outcomes by Mullahy and Norton (2023), among others. In Section C.4, we consider the causal interpretation of the marginal effects of 2PMs, building on the discussion in Angrist (2001). Our decomposition shows that the marginal effects from 2PMs yield the sum of a causal parameter similar to $\theta_{\text{Intensive}}$ as well as a "selection term" comparing potential outcomes for individuals for whom treatment only has an intensive margin effect to those with an extensive margin effect. It thus will generally be difficult to ascribe a causal interpretation to the marginal effects of 2PMs without assumptions about this selection.

3.5 Empirical applications

In this section, we focus on three concrete empirical applications to illustrate how the alternative parameters described in Section 3.4 can be estimated in practice. To illustrate a range of possible applications, we consider a randomized controlled trial, a difference-in-differences design, and an instrumental variables design.

3.5.1 An RCT setting: Carranza *et al.* (2022)

Carranza *et al.* (2022) conduct a randomized controlled trial (RCT) in South Africa. Individuals randomized to the treatment group are provided with certified test results that they can show to prospective employers to vouch for their skills. Individuals in the control group do not receive test results.[33] They then investigate how this treatment impacts labor market outcomes such as employment, hours worked, and earnings. We focus here on the effects on hours worked.

[32] We note that the Lee (2009) bounds will tend to be tight when the extensive margin effect is close to zero. As noted in Theorem 3.2.3, this is precisely the setting where ATEs for log-like transformations are relatively insensitive to finite changes in scale.

[33] Some individuals are also assigned to a "placebo" arm in which they are provided the test results but the form does not include the individual's name, and thus cannot credibly be shared with employers. We focus on the effect of the main treatment relative to the pure control group.

Original specification and sensitivity to units. Carranza *et al.* (2022) estimate the effect of their randomized treatment on the inverse hyperbolic sine of weekly hours worked. Formally, they estimate the OLS regression specification

$$\text{arcsinh}(Y_i) = \beta_0 + D_i\beta_1 + X_i'\gamma + u_i, \tag{3.3}$$

where Y_i is average weekly hours worked for unit i, D_i is an indicator for whether unit i was in the treatment group, and X_i is a vector of controls.[34] Their estimate of the ATE ($\hat{\beta}_1$) is 0.201 (see column (1) in Table 3.3). They interpret this as a 20% change in hours: "Certification increases average weekly hours worked, coded as zero for nonemployed candidates, by 20 percent" (p. 3560).

Table 3.3: Estimates using $\text{arcsinh}(Y)$ with different units of Y in Carranza *et al.* (2022)

	arcsinh(weekly hrs)	arcsinh(yearly hrs)	arcsinh(FTEs)
Treatment	0.201	0.417	0.031
	(0.052)	(0.096)	(0.012)
Units of outcome:	Weekly Hrs	Yearly Hrs	FTEs

Note: This table shows estimates of the average treatment effect in Carranza *et al.* (2022) on the inverse hyperbolic sine of hours worked, estimated using (3.3). In the first column, the outcome is the inverse hyperbolic sine of *weekly* hours, as in the original paper. The remaining columns use the inverse hyperbolic sine of annualized hours (weekly hours times 52) or the inverse hyperbolic sine of the number of full-time equivalents worked (weekly hours divided by 40). Standard errors are clustered at the assessment date (the unit of treatment assignment) as in the original paper.

However, the results in Section 3.2 suggest that the estimate of β_1 should not be interpreted as a percentage effect, since it depends on the units of the outcome. To illustrate this, in columns (2) and (3) we re-estimate specification (3.3) with Y_i redefined to be (a) yearly hours worked, i.e. weekly hours times 52, or (b) the number of full-time equivalents (FTE) worked, i.e. weekly hours divided by 40. The results change quite substantially depending on the units used, with an estimate of 0.417 using yearly hours and 0.031 using FTEs. We therefore turn next to alternative approaches with a percentage interpretation in this setting.

[34] Carranza *et al.* (2022) include individuals receiving the "placebo" treatment in the sample and add an indicator for receiving the placebo treatment in X_i. We follow the same practice, although the results are similar if units receiving the placebo treatment are dropped.

Percentage changes in the average. The average number of (weekly) hours worked was 9.84 in the treated group and 8.85 in the control group. A simple summary of the treatment effect is thus that average hours worked were 11% higher in the treated group ($9.84/8.85 = 1.11$). This is an estimate of the parameter $\theta_{\text{ATE\%}} = E[Y(1) - Y(0)]/E[Y(0)]$ discussed in Section 3.4.1 above. A numerically equivalent way to obtain this estimate of 11% is to use Poisson quasi-maximum likelihood estimation (Poisson QMLE) to estimate

$$Y_i = \exp(\beta_0 + \beta_1 D_i)U_i \tag{3.4}$$

and then calculate $\hat{\theta}_{\text{ATE\%}} = \exp(\hat{\beta}_1) - 1 = 0.11$ (see column (1) in Table 3.4).[35] This formulation in terms of Poisson QMLE is useful since it allows us to include covariates to potentially increase precision. Column (2) of Table 3.4 shows the estimate of $\hat{\theta}_{\text{ATE\%}}$ from estimating

$$Y_i = \exp(\beta_0 + \beta_1 D_i + X_i'\gamma)U_i \tag{3.5}$$

by Poisson QMLE, with smaller standard errors than in column (1) (0.069 vs. 0.081).

Table 3.4: Poisson Regression and Implied Proportional Effects in Carranza *et al.* (2022).

	(1)	(2)
β_0	2.180	0.150
	(0.058)	(0.311)
β_1	0.106	0.150
	(0.072)	(0.060)
Implied Prop. Effect	0.112	0.150
	(0.081)	(0.069)
Covariates	N	Y

Note: the first two rows of column (1) show the estimates of the coefficients β_0 and β_1 in (3.4), estimated using Poisson QMLE. The third row shows the implied estimate of the proportional effect, $E[Y(1) - Y(0)]/E[Y(0)]$, calculated as $\hat{\theta}_{\text{ATE\%}} = \exp(\hat{\beta}_1) - 1$. The second column shows analogous estimates using (3.5), which adds controls for pre-treatment covariates (we do not show the coefficients on the controls in the interest of brevity). Standard errors are clustered at the assessment date (the unit of treatment assignment) as in the original paper.

[35] This estimation is done in the sample of treated units and control units, discarding the placebo group. One could equivalently retain the units in the placebo group and add an indicator for the placebo group to (3.4).

Separate estimates for the extensive/intensive margins. As shown in Table 3.1, the treatment in Carranza *et al.* (2022) has an estimated extensive margin treatment effect of 0.055, meaning that it increases the fraction of people with positive hours worked by 5.5 percentage points. We may be interested in whether the overall 11% increase in hours worked is driven entirely by the extensive margin, or whether there is an intensive margin effect. That is, does the treatment increase hours only by bringing people into the labor force, or does it also allow people who would have worked anyway to find jobs with more hours (e.g. full-time instead of part-time)? To this end, we can use the method of Lee (2009) to compute bounds for the effect of the treatment for "always-takers" who would have positive hours worked regardless of treatment ($Y(1) > 0, Y(0) > 0$).[36] The Lee bounds approach requires the monotonicity assumption that anyone who would work positive hours without the treatment would also work positive hours when treated (i.e., $P(Y(1) = 0, Y(0) > 0) = 0$). This seems reasonable if workers only share the information provided by the treatment when it helps their job prospects. It could be violated, however, if workers mistakenly share their test score results when in fact employers view them negatively.

Column 1 of Table 3.5 reports bounds of $[-0.20, 0.28]$ for the effect of the treatment on log hours worked by the always-takers, while Column 2 shows bounds of $[-6.67, 2.77]$ for weekly hours (in levels). Unfortunately, in this setting the Lee bounds are fairly wide, including both a zero intensive-margin effect as well as fairly large intensive-margin effects (up to 28 log points). Thus, without further assumptions, the data is not particularly informative about the size of the intensive margin.

We can, however, say more if we are willing to impose some assumptions about how the always-takers, who would work regardless of treatment status ($Y(1) > 0, Y(0) > 0$), compare to the compliers ($Y(1) > 0, Y(0) = 0$), who only work positive hours when receiving the treatment. We might reasonably expect that the compliers are negatively selected relative to the always-takers and thus would work fewer hours when receiving treatment. We can formalize this by imposing that $E[Y(1) \mid \text{Complier}] = (1-c)E[Y(1) \mid \text{Always-taker}]$, i.e. that average hours worked for compliers under treatment is $100c\%$ lower than for always takers. Columns 3 through 5 of Table 3.5 report

[36] We again exclude units receiving the "placebo treatment."

estimates of the average effect on the always-takers, assuming $c = 0, 0.25$, and 0.5, respectively.[37] If we assume that always-takers and compliers work an equal number of hours under treatment ($c = 0$), then our point estimates suggest that there is actually a negative intensive-margin effect for the always-takers (-1.02 weekly hours). Under the assumption that compliers work 25% fewer hours ($c = 0.25$), the estimated effect for always-takers is near zero (-0.07 weekly hours), consistent with no important intensive margin. Finally, if we assume compliers work half as many hours as the always-takers ($c = 0.5$), then our estimates suggest a positive intensive margin effect (0.95 weekly hours). Our assessment of the importance of the intensive margin thus depends on how negatively-selected we think compliers are relative to always-takers.

Table 3.5: Bounds and point estimates for the intensive margin treatment effect in Carranza *et al.* (2022)

	(1)	(2)	(3)	(4)	(5)
Lower bound	-0.195	-6.665			
	(0.064)	(1.366)			
Upper bound	0.283	2.771			
	(0.114)	(2.067)			
Point estimate			-1.025	-0.069	0.954
			(1.182)	(1.349)	(1.588)
units	Log(Hours)	Hours	Hours	Hours	Hours
c			0	0.25	0.5

Note: This table shows bounds and point estimates of the intensive margin treatment effect in Carranza *et al.* (2022), i.e. the treatment effect on hours worked for "always-takers" who would work positive hours regardless of treatment status. The first first two columns of the table show Lee (2009) bounds for the effect of treatment on the always-takers when the outcome is $\log(\text{Hours})$ and weekly hours, respectively. Columns 3 through 5 show point estimates for the effect on weekly hours worked for always-takers under the assumption that average hours worked by "compliers" (who work only when treated) are $100c\%$ lower than for the always-takers. Standard errors are calculated via a non-parametric bootstrap using 1,000 draws, clustered at the assessment date level.

[37] Under the assumptions in Lee (2009), $E[Y(1) \mid Y(1) > 0] = \theta E[Y(1) \mid \text{Always-taker}] + (1-\theta)E[Y(1) \mid \text{Complier}]$, where $\theta = P(Y(0) > 0)/P(Y(1) > 0)$. Plugging in $E[Y(1) \mid \text{Complier}] = (1-c)E[Y(1) \mid \text{Always-taker}]$, it follows that $E[Y(1) \mid \text{Always-taker}] = 1/(\theta + (1-c)(1-\theta))E[Y(1) \mid Y(1) > 0]$. Further, $E[Y(0) \mid \text{Always-taker}] = E[Y(0) \mid Y(0) > 0]$. Our estimation plugs in sample analogs to these expressions to estimate $E[Y(1) - Y(0) \mid \text{Always-taker}]$.

3.5.2 A DiD setting: Sequeira (2016)

Sequeira (2016) studies a decrease in tariffs on trade between Mozambique and South Africa which occurred in 2008. She is interested in whether the reduction in tariffs reduced bribes paid to customs officers (among other outcomes). To study this question, she utilizes a difference-in-differences design comparing the change in bribes paid for products that were affected by the tariff change to that for a comparison group of products that did not experience a change in tariffs.

Original specification and sensitivity to units. Sequeira (2016) has repeated cross-sectional data with information on the bribe amount Y_{it} paid on shipment i in year t. She estimates the regression specification

$$\log(1+Y_{it}) = \beta_0 + D_i \times \text{Post}_t\,\beta_1 + D_i\,\beta_2 + \text{Post}_t\,\beta_3 + X_{it}'\beta_4 + \epsilon_{it}, \tag{3.6}$$

where D_i is an indicator for whether shipment i is for a product type affected by the tariff change in 2008, Post_t is an indicator for whether year t is after the tariff change, and X_{it} is a vector of covariates related to shipment i in period t. Sequeira (2016) estimates (3.6) with Y_{it} measured in 2007 Mozambican Metical (MZN) and obtains $\hat{\beta}_{1,(\text{MZN})} = -3.7$ (SE $= 1.1$). However, estimating the same specification with Y_{it} measured in thousands of U.S. dollars instead yields an estimate of $\hat{\beta}_{1,(\$1000)} = -0.11$ (SE $= 0.070$).[38] These results reinforce the conclusion from Section 3.2 that treatment effects for $m(y) = \log(1+y)$ should not be interpreted as approximating a percentage effect.

In what follows, we discuss a variety of alternative approaches that may be reasonable in this context. We note that in a non-experimental setting like this, different approaches may rely on different identifying assumptions. We therefore explicitly discuss the identifying assumptions needed by each of the methods we discuss.

[38] We use the conversion rate of 1 USD = 24.48 MZN as of January 1, 2007, as provided by fxtop.com.

Proportional treatment effects. One natural approach here is to target the average proportional treatment effect on the treated,

$$\theta_{\text{ATT\%}} = \frac{E[Y_{it}(1) \mid D_i = 1, \text{Post}_t = 1] - E[Y_{it}(0) \mid D_i = 1, \text{Post}_t = 1]}{E[Y_{it}(0) \mid D_i = 1, \text{Post}_t = 1]}.$$

This is the percentage change in the average outcome for the treated group in the post-treatment period.

Identification of $\theta_{\text{ATT\%}}$ requires us to infer the counterfactual post-treatment mean outcome for the treated group, $E[Y_{it}(0) \mid D_i = 1, \text{Post}_t = 1]$. Of course, one approach to obtain such identification would be to assume parallel trends in levels. However, given that the treated and control groups have different pre-treatment means (see the bottom panel of Table 3.6), it may be unreasonable to expect that time-varying factors (e.g. the macroeconomy) have equal level effects on the outcome. An alternative identifying assumption is to impose that, in the absence of treatment, the *percentage* changes in the mean would have been the same for the treated and control group. As in Wooldridge (2023), this can be formalized using a "ratio" version of the parallel trends assumption,

$$\frac{E[Y_{it}(0) \mid D_i = 1, \text{Post}_t = 1]}{E[Y_{it}(0) \mid D_i = 1, \text{Post}_t = 0]} = \frac{E[Y_{it}(0) \mid D_i = 0, \text{Post}_t = 1]}{E[Y_{it}(0) \mid D_i = 0, \text{Post}_t = 0]}. \tag{3.7}$$

Intuitively, (3.7) states that if the treatment had not occurred, the average percentage change in the mean outcome for the treated group would have been the same as the average percentage change in the mean outcome for the control group. Under (3.7), we can thus estimate the counterfactual percentage change in the mean outcome for the treated group using the observed percentage change for the control group.

Table 3.6 shows that the sample mean of the outcome for the untreated group decreased by 75% between the pre-treatment and post-treatment periods (from 4,742 to 1,172 (MZN)). Under the ratio parallel trends assumption (3.7), this suggests that the mean outcome for the treated group would also have decreased by 75% in the absence of treatment, thus implying an estimate of $2,602$ for the counterfactual mean outcome for the treated group. The actual post-treatment mean for the treated group is 465, which is 82% below this implied counterfactual. This implies that the tariff reduction reduced the average bribe in the post-treatment period by 82%, i.e. $\hat{\theta}_{\text{ATT\%}} = -0.82$. Conveniently, this estimate can also be obtained using Poisson QMLE to estimate

$$Y_{it} = \exp(\beta_0 + D_i \times \text{Post}_t\,\beta_1 + D_i\,\beta_2 + \text{Post}_t\,\beta_3)\epsilon_{it} \tag{3.8}$$

and then computing $\hat{\theta}_{\text{ATT}\%} = \exp(\hat{\beta}_1) - 1 = -0.82$, as shown in column (1) of Table 3.6.

Table 3.6: Poisson regression estimates of $\theta_{\text{ATT}\%}$

	(1)	(2)
Post × Treatment	-1.722	-1.272
	(0.632)	(0.606)
Prop. Effect	-0.821	-0.720
	(0.113)	(0.170)
Covariates	N	Y
Treated Group Means (Pre, Post):	10527	465
Untreated Group Means (Pre, Post):	4742	1172

Note: this table shows Poisson regression estimates of (3.8) and (3.9) in columns (1) and (2), respectively. The first row of the table shows the estimate $\hat{\beta}_1$. The second row shows $\exp(\hat{\beta}_1) - 1$, which is the implied estimate of the proportional treatment effect $\theta_{\text{ATT}\%}$. The coefficients on control variables are omitted for brevity. Standard errors are clustered at the four-digit product code as in the original paper. The mean bribe amounts (in MZN) by treatment group and time period are displayed in the bottom panel. The pre-period refers to the year 2007, whereas the post-treatment period is an average over the years 2008, 2011, and 2012 (the three post-treatment years for which data is available).

We can also re-incorporate the covariates X_{it} by estimating

$$Y_{it} = \exp(\beta_0 + D_i \times \text{Post}_t\,\beta_1 + D_i\,\beta_2 + \text{Post}_t\,\beta_3 + \beta_4' X_{it})\epsilon_{it}, \tag{3.9}$$

which yields an estimate of $\theta_{\text{ATT}\%}$ of -0.72, as shown in the second column of Table 3.6. As formalized in Wooldridge (2023), this estimate will be a consistent estimate of $\theta_{\text{ATT}\%}$ if (3.7) holds conditional on X_{it}, and the conditional expectation of Y_{it} takes the functional form implied by (3.9) (assuming ϵ_{it} has mean 1 conditional on the covariates). The approach with covariates thus suggests that the tariff change reduced the average bribe for treated products by 72% in the post-treatment period.

Sequeira (2016)'s data only contains information on one year prior to treatment (2007), and so in this context it is not possible to evaluate the plausibility of (3.7) using periods prior to the policy change of interest. If multiple pre-treatment periods were available, however, one could estimate a Poisson QMLE event-study of the form

$$Y_{it} = \exp\left(\lambda_t + D_i\,\beta_2 + \sum_{r\neq -1} D_i \times [\text{RelativeTime}_t = r]\,\beta_r^{ES}\right)\epsilon_{it}, \tag{3.10}$$

where $\text{RelativeTime}_t = t - 2008$ is the time relative to the treatment date. The event-study coefficients β_r^{ES} for $r < 0$ are analogous to "pre-trends" coefficients in typical difference-in-differences event-studies, and are informative about whether the pre-treatment analogue to (3.7) holds.[39]

Log effects with calibrated extensive margin value. The analysis above presented estimates of $\theta_{\text{ATT}\%}$, the proportional change in the *average* bribe caused by the treatment. It is well-known that averages can be heavily influenced by observations in the tail, especially when the outcome has a skewed distribution, as is the case here (see Figure 3.2). One might argue that a world in which most products receive medium-sized bribes is more corrupt than one in which a very small fraction of products receive large bribes—even if they both produce the same average bribe amount. This motivates studying the treatment effect on a concave transformation of the outcome that is less heavily influenced by outcomes in the tail of the distribution. As an illustration of this, we first normalize the outcome so that 1 corresponds to the value of the minimum non-zero bribe in the data (that is, we divide by $y_{\min} = \min_{Y_{it}>0} Y_{it} = 15.68$ MZN). We then estimate the treatment effect for the transformed outcome $m(Y)$, where $m(y) = \log(y)$ for $y > 0$ and $m(0) = -x$ for some choice of x, as described in Section 3.4.2. If x is set to 0, then this estimates the treatment effect in logs where all zero bribes are set to equal the smallest positive bribe in the data; this specification thus "shuts off" the extensive margin change between 0 and $y_{\min}$. If instead x is set to 0.1, for example, then a change

[39] More precisely, the exponentiated coefficients $\exp(\hat{\beta}_r) - 1$ correspond to the implied "placebo" proportional treatment effects for periods before treatment. We recommend plotting the exponentiated coefficients in event-studies, although we note that $\exp(\beta) - 1 \approx \beta$ for $\beta \approx 0$. As with typical tests for pre-trends, one should be cautious that a failure to reject the null that the pre-treatment coefficients equal zero does not necessarily imply that the identifying assumption is satisfied (Kahn-Lang and Lang, 2020; Roth, 2022). One can (partially) address these issues by applying sensitivity analysis tools for event-studies (e.g. Rambachan and Roth, 2023) to estimates of (3.10) to further gauge the robustness of the findings to violations of the identifying assumptions. We also refer the reader to Wooldridge (2023) for extensions of the Poisson regression approach to settings with staggered treatment timing.

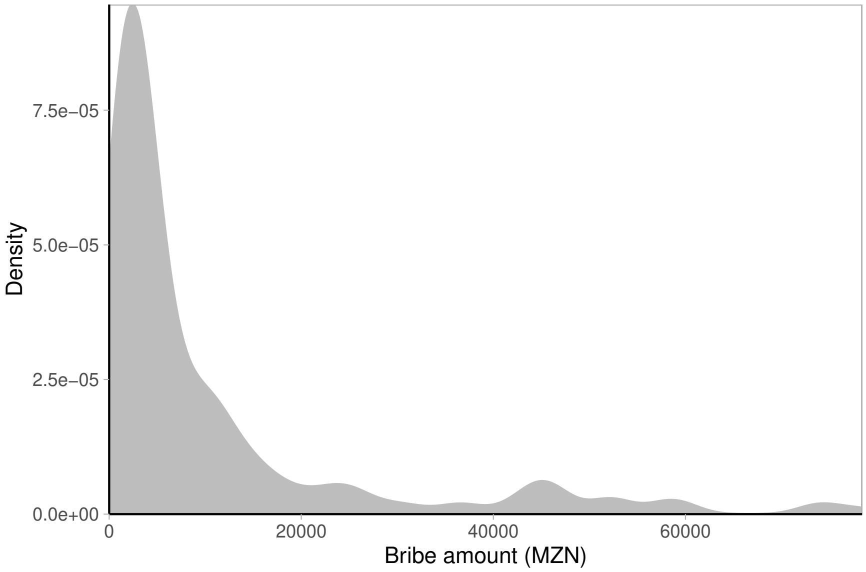


Figure 3.2: Density of bribe amount in Sequeira (2016)

Note: this figure shows a kernel density estimate of the bribe amount in Sequeira (2016), pooling across all observations with a positive bribe. The kernel density estimates are constructed using the default settings of the `stat_density` function in R.

between 0 and $y_{\min}$ is valued as the equivalent of a 10 log point change along the intensive margin.

We estimate the treatment effect for these transformations using the analogue to (3.6) that replaces $\log(1 + Y_{it})$ with $m(Y_{it})$ on the left-hand side.[40] The results for $x \in \{0, 0.1, 1, 3\}$ are shown in Table 3.7. Column (1) shows an effect of 249 log points ($\hat{\beta}_1 = -2.49$) when we treat zero bribes as if they were equal to $y_{\min}$ (i.e. setting $x = 0$). The estimated treatment effect grows in magnitude as we place more value on the extensive margin by increasing x. Interestingly, the original estimate in Sequeira (2016) of -3.748 using $\log(1 + Y)$ is similar to what we obtain when we value a change from 0 to $y_{\min}$ at 300 log points ($x = 3$). The original specification can thus be viewed as placing a rather large weight on the extensive margin.

[40] As usual, identification of the treatment effect for $m(Y)$ using difference-in-differences requires parallel trends for $m(Y(0))$. The identifying assumption thus varies depending on the choice of x. The results in Roth and Sant'Anna (2023) imply that parallel trends will hold for all values of x when a parallel trends assumption is satisfied for the distribution of $Y(0)$. If more pre-treatment periods were available, these identifying assumptions could be partially evaluated using pre-trends tests. See Theorem 3.4.2 for additional discussion of identification.

Table 3.7: Explicit calibration of the extensive margin in Sequeira (2016)

	(1)	(2)	(3)	(4)
Post × Treatment	-2.493	-2.538	-2.949	-3.860
	(0.740)	(0.752)	(0.861)	(1.106)
Extensive margin value (x):	0.000	0.100	1.000	3.000

Note: this table shows estimates of the treatment effect on the treated using $m(Y)$ as the outcome in Sequeira (2016), where $m(y)$ is defined to equal $\log(y)$ for $y > 0$ and $-x$ for $y = 0$. The outcome is normalized so that $Y = 1$ corresponds to the minimum non-zero value of the outcome. Thus, the treatment effect assigns a value of $100x$ log points to an extensive margin change between 0 and the minimum non-zero value of Y. The treatment effects are estimated using (3.6), except replacing $\log(1 + Y_{it})$ with $m(Y_{it})$. Standard errors are clustered at the four-digit product code as in the original paper.

3.5.3 An IV setting: Berkouwer and Dean (2022)

Berkouwer and Dean (2022) conduct an RCT in Nairobi in which they randomize the price for energy-efficient stoves. They use the randomized price (p_i) as an instrument for whether an individual i buys an energy-efficient stove (D_i). They use this instrument to estimate the effects of stove-adoption on outcomes such as charcoal spending (Y_i).

Original specification and sensitivity to scale. Let X_i be a vector of control variables (including a constant). Berkouwer and Dean (2022) estimate

$$\text{arcsinh}(Y_i) = D_i\beta + X_i'\gamma + \epsilon_i \tag{3.11}$$

by two-stage least squares (TSLS), using p_i as an instrument for D_i.[41] (They also report results where spending is measured in levels.) The estimated coefficient $\hat{\beta}$ is an estimate of the LATE of stove adoption on the arcsinh of charcoal spending for instrument-compliers whose decision of whether to purchase the stove depends on the price offered in the experiment.[42] In Berkouwer and Dean (2022),

[41] More precisely, each observation i is an individual-by-week pair, and some (but not all) individuals are surveyed on multiple weeks. Standard errors are clustered at the respondent level.

[42] We use the phrase "instrument-compliers" to distinguish compliers for the instrument, whose value of $D(z)$ depends on z, from "compliers" discussed earlier who have $Y(1) > 0, Y(0) = 0$. Since the instrument takes on multiple values (i.e. multiple price offers), β corresponds to a weighted average of treatment effects across instrument-compliers for different

Y_i is measured as weekly charcoal spending in dollars. They obtain a coefficient of $\hat{\beta} = -0.50$ and write "[t]he 50 log point reduction corresponds to a 39 percent decrease in charcoal consumption [since $\exp(-0.50) = 1 - 0.39$]" (p. 3306).

However, if we change the units of the outcome to annual charcoal spending in Kenyan shillings, the original currency in which charcoal spending was measured, the same specification yields an estimate of -0.44. Relative to our previous applications, the change in the treatment effect estimates is fairly small for these choices of units, due to a small estimated extensive margin of 0.01 (see Table 3.1).[43] Nevertheless, the fact that the treatment effects using an $\operatorname{arcsinh}$-transformed outcome depend on the units should give us pause in interpreting them as percentages. Indeed, a percentage effect is not well-defined for someone who has non-zero spending under treatment and zero spending under the control, so an average individual-level percentage effect does not make sense if the treatment can affect whether one has any charcoal spending.

Berkouwer and Dean (2022) first discuss the LATE in levels, and then immediately afterwards state that the treatment effect for the $\operatorname{arcsinh}$-transformed outcome "corresponds to a 39 percent decrease in charcoal consumption" (p. 3306). The main goal of taking the $\operatorname{arcsinh}$ transformation here thus appears to be to obtain a treatment effect with a percentage interpretation. We therefore next implement two approaches with an (approximate) percentage interpretation in this context.

Proportional LATE. One natural approach in this context is to estimate the proportional change in the average outcome for instrument-compliers, i.e. to estimate $\theta_{\text{ATE}\%}$ among the population of instrument-compliers. Put otherwise, we can express the LATE in levels as a percentage of the control mean for instrument-compliers. An estimate of the LATE in levels is naturally obtained using TSLS specification (3.11) with Y_i as the outcome, which yields an estimate of -2.46. As described in Abadie (2002), we can likewise obtain an estimate of the control instrument-complier mean by using TSLS with $-(D_i - 1) \cdot Y_i$ as the outcome, which yields an estimate of 5.86. Putting these together, we obtain an estimate of $\theta_{\text{ATE}\%}$ for instrument-compliers of $-2.46/5.86 = -0.42$ (SE = 0.046), which

values of the instrument (Angrist *et al.*, 2000).

[43] We note, however, that the t-statistic for the effect on $\operatorname{arcsinh}(Y_i)$ is rather sensitive here, changing from approximately 7 to 3 depending on the units.

suggests that average charcoal spending is 42% lower for instrument-compliers under treatment than under control.[44] If pollution is proportional to charcoal spending, then this parameter is economically relevant as it corresponds to the percentage reduction in pollution for instrument-compliers from gaining access to the efficient stove.

Lee bounds. Berkouwer and Dean (2022) benchmark their treatment effect estimates relative to engineering estimates of the efficiency gains of using an efficient stove relative to a non-efficient one. For this benchmarking exercise, it seems sensible to focus on the intensive-margin effect of the treatment—i.e., the treatment effect for instrument-compliers who would use a non-efficient stove if offered a high price and an efficient one if offered a low price. To do so, we can form Lee (2009)-type bounds for the average treatment effect in logs for instrument-compliers who would have positive charcoal spending regardless of treatment status.[45]

The bounds on $\theta_{\text{Intensive}}$ for instrument-compliers are $[-0.565, -0.538]$ (with SEs for the lower and upper bounds of 0.072 and 0.075).[46] This implies that for the instrument-compliers who would spend on charcoal regardless of treatment status, spending decreases by 54 to 56 log points. We note that the Lee bounds are fairly tight in this case, as tends to be the case when the extensive margin is small. It is also worth noting that in this example, the estimated treatment effects using $\text{arcsinh}(Y_i)$—both in terms of weekly spending in dollars and in terms of annual spending in Kenyan shillings—fall outside of the Lee bounds, although they are fairly close to the upper bound when using weekly spending in dollars.

[44] The standard error was calculated via a non-parametric bootstrap with 1,000 draws, clustered at the respondent level. We note that with a binary instrument, an estimate of $\theta_{\text{Intensive}}$ for instrument-compliers can be obtained using Poisson IV regression (e.g. the ivpoisson command in Stata); see Angrist (2001). However, we are not aware of a LATE interpretation of Poisson IV regression with a multi-valued instrument, and thus do not pursue it here. Whether Poisson IV regression has such an interpretation with a continuous IV strikes us an interesting topic for future work.

[45] The validity of the Lee (2009)-type bounds requires the "monotonicity" assumption that all instrument-compliers who would have some charcoal consumption when not buying an efficient stove would also have some charcoal consumption when buying an efficient stove, which seems reasonable. Note that this is a distinct assumption from the instrument monotonicity assumption needed for a LATE interpretation for instrumental variables (Imbens and Angrist, 1994), which in this context states that anyone who would buy a stove at a higher price would also buy at a lower price.

[46] We obtain these estimates using the procedure in Abadie (2002), as described in detail in Section C.5.

3.6 Conclusion

It is common in empirical work to estimate ATEs for transformations such as $\log(1+Y)$ or $\text{arcsinh}(Y)$ which are well-defined at zero and behave like $\log(Y)$ for large values of Y. We show that the ATEs for such transformations should not be interpreted as percentages, since they depend arbitrarily on the units of the outcome when there is an extensive margin. Further, we show that any parameter that is an average of individual-level treatment effects of the form $E_P[g(Y(1), Y(0))]$ must be scale-dependent if it is point-identified and well-defined at zero. We discuss several alternative approaches, including estimating scale-invariant normalized parameters (e.g. via Poisson regression), explicitly calibrating the value placed on the intensive versus extensive margins, and separately estimating effects for the intensive and extensive margins (e.g. using Lee bounds). We illustrate how these approaches can be applied in practice in three empirical applications.

References

ABADIE, A. (2002). Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models. *Journal of the American Statistical Association*, **97** (457), 284–292, publisher: [American Statistical Association, Taylor & Francis, Ltd.].

— (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, **59** (2), 391–425.

—, AGARWAL, A., IMBENS, G., JIA, S., MCQUEEN, J. and STEPANIANTS, S. (2023). Estimating the value of evidence-based decision making. *arXiv preprint arXiv:2306.13681*.

— and CATTANEO, M. D. (2021). Introduction to the special section on synthetic control methods. *Journal of the American Statistical Association*, **116** (536), 1713–1715.

—, DIAMOND, A. and HAINMUELLER, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, **105** (490), 493–505.

—, — and — (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, **59** (2), 495–510.

— and GARDEAZABAL, J. (2003). The economic costs of conflict: A case study of the basque country. *American economic review*, **93** (1), 113–132.

— and KASY, M. (2019). Choosing among regularized estimators in empirical economics: The risk of machine learning. *Review of Economics and Statistics*, **101** (5), 743–762.

— and VIVES-I-BASTIDA, J. (2021). *Synthetic Controls in Action*. Tech. rep.

ABALUCK, J., CACERES BRAVO, M., HULL, P. and STARC, A. (2021). Mortality effects and choice across private health insurance plans. *The quarterly journal of economics*, **136** (3), 1557–1610.

ABERNETHY, J., AGARWAL, A., BARTLETT, P. L. and RAKHLIN, A. (2009). A stochastic view of optimal regret through minimax duality. *arXiv preprint arXiv:0903.5328*.

ACZÉL, J. (1966). *Lectures on Functional Equations and Their Applications*. Academic Press, google-Books-ID: n7vckU_1tY4C.

AGAN, A. Y. and STARR, S. B. (2020). *Employer neighborhoods and racial discrimination*. Tech. rep., National Bureau of Economic Research.

AGER, P., BOUSTAN, L. and ERIKSSON, K. (2021). The intergenerational effects of a large wealth shock: White southerners after the civil war. *American Economic Review*, **111** (11), 3767–94.

AIHOUNTON, G. B. D. and HENNINGSEN, A. (2021). Units of measurement and the inverse hyperbolic sine transformation. *The Econometrics Journal*, **24** (2), 334–351.

ALESINA, A. F., GLAESER, E. L. and SACERDOTE, B. (2001). Why doesn't the us have a european-style welfare system?

ANDREWS, I. and KASY, M. (2019). Identification of and correction for publication bias. *American Economic Review*, **109** (8), 2766–94.

—, KITAGAWA, T. and MCCLOSKEY, A. (2023). Inference on winners.

— and MILLER, C. (2013). Optimal Social Insurance with Heterogeneity. p. 26.

ANGRIST, J. D. (2001). Estimation of Limited Dependent Variable Models With Dummy Endogenous Regressors. *Journal of Business & Economic Statistics*, **19** (1), 2–28, publisher: Taylor & Francis _eprint: https://doi.org/10.1198/07350010152472571.

—, GRADDY, K. and IMBENS, G. W. (2000). The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish. *The Review of Economic Studies*, **67** (3), 499–527, publisher: [Oxford University Press, Review of Economic Studies, Ltd.].

—, HULL, P. D., PATHAK, P. A. and WALTERS, C. R. (2017). Leveraging lotteries for school value-added: Testing and estimation. *The Quarterly Journal of Economics*, **132** (2), 871–919.

— and PISCHKE, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

ARELLANO, M. and BONHOMME, S. (2009). Robust priors in nonlinear panel data models. *Econometrica*, **77** (2), 489–536.

ARKHANGELSKY, D., ATHEY, S., HIRSHBERG, D. A., IMBENS, G. W. and WAGER, S. (2021). Synthetic difference-in-differences. *American Economic Review*, **111** (12), 4088–4118.

ARMSTRONG, T. B., KOLESÁR, M. and PLAGBORG-MØLLER, M. (2022). Robust empirical bayes confidence intervals. *Econometrica*, **90** (6), 2567–2602.

ARNOLD, D., DOBBIE, W. and HULL, P. (2022). Measuring racial discrimination in bail decisions. *American Economic Review*, **112** (9), 2992–3038.

ARORA, A., BELENZON, S. and SHEER, L. (2021). Knowledge spillovers and corporate investment in scientific research. *American Economic Review*, **111** (3), 871–98.

ATHEY, S., BAYATI, M., DOUDCHENKO, N., IMBENS, G. and KHOSRAVI, K. (2021a). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, **116** (536), 1716–1730.

—, BICKEL, P. J., CHEN, A., IMBENS, G. and POLLMANN, M. (2021b). *Semiparametric estimation of treatment effects in randomized experiments*. Tech. rep., National Bureau of Economic Research.

— and WAGER, S. (2021). Policy learning with observational data. *Econometrica*, **89** (1), 133–161.

AUDIBERT, J.-Y. and TSYBAKOV, A. B. (2007). Fast learning rates for plug-in classifiers. *Annals of Statistics*.

AZEVEDO, E. M., DENG, A., MONTIEL OLEA, J. L., RAO, J. and WEYL, E. G. (2020). A/b testing with fat tails. *Journal of Political Economy*, **128** (12), 4614–000.

AZOULAY, P., FONS-ROSEN, C. and GRAFF ZIVIN, J. S. (2019). Does science advance one funeral at a time? *American Economic Review*, **109** (8), 2889–2920.

BAI, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, **71** (1), 135–171.

— and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, **70** (1), 191–221.

BAILY, N. (1978). Some Aspects of Optimal Unemployment Insurance. *Journal of Public Economics*, pp. 379–402.

BARRETT, N. and TOMA, E. F. (2013). Reward or punishment? class size and teacher quality. *Economics of education review*, **35**, 41–52.

BARTLETT, M. S. (1947). The Use of Transformations. *Biometrics*, **3** (1), 39–52, publisher: [Wiley, International Biometric Society].

BARTLETT, P. L., KOOLEN, W. M., MALEK, A., TAKIMOTO, E. and WARMUTH, M. K. (2015). Minimax fixed-design linear regression. In *Conference on Learning Theory*, PMLR, pp. 226–239.

BASTOS, P., SILVA, J. and VERHOOGEN, E. (2018). Export destinations and input prices. *American Economic Review*, **108** (2), 353–92.

BAUM-SNOW, N. and HAN, L. (2019). The microgeography of housing supply. *Work in progress, University of Toronto*.

BECKER, G. S., KOMINERS, S. D., MURPHY, K. M. and SPENKUCH, J. L. (2018). A theory of intergenerational mobility. *Journal of Political Economy*, **126** (S1), S7–S25.

BEERLI, A., RUFFNER, J., SIEGENTHALER, M. and PERI, G. (2021). The abolition of immigration restrictions and the performance of firms and workers: Evidence from switzerland. *American Economic Review*, **111** (3), 976–1012.

BELLEMARE, M. F. and WICHMAN, C. J. (2020). Elasticities and the Inverse Hyperbolic Sine Transformation. *Oxford Bulletin of Economics and Statistics*, **82** (1), 50–61, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/obes.12325.

BELLÉGO, C., BENATIA, D. and PAPE, L. (2022). Dealing with Logs and Zeros in Regression Models. ArXiv:2203.11820 [econ, stat].

BELOTTI, F., DEB, P., MANNING, W. G. and NORTON, E. C. (2015). Twopm: Two-Part Models. *The Stata Journal*, **15** (1), 3–20, publisher: SAGE Publications.

BEN-MICHAEL, E., FELLER, A. and ROTHSTEIN, J. (2019). Synthetic controls and weighted event studies with staggered adoption. *arXiv preprint arXiv:1912.03290*.

—, — and — (2021). The augmented synthetic control method. *Journal of the American Statistical Association*, **116** (536), 1789–1803.

BERGMAN, P., CHETTY, R., DELUCA, S., HENDREN, N., KATZ, L. F. and PALMER, C. (2023). *Creating moves to opportunity: Experimental evidence on barriers to neighborhood choice*. Tech. rep., National Bureau of Economic Research.

BERKOUWER, S. B. and DEAN, J. T. (2022). Credit, attention, and externalities in the adoption of energy efficient technologies by low-income households. *American Economic Review*, **112** (10), 3291–3330.

BLACKWELL, D. (1956). An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, **6** (1), 1–8.

BONHOMME, S., HOLZHEU, K., LAMADON, T., MANRESA, E., MOGSTAD, M. and SETZLER, B. (2020). How much should we trust estimates of firm effects and worker sorting?

— and MANRESA, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, **83** (3), 1147–1184.

BONVINI, M., KENNEDY, E. H. and KEELE, L. J. (2023). Minimax optimal subgroup identification. *arXiv preprint arXiv:2306.17464*.

BOTTMER, L., IMBENS, G., SPIESS, J. and WARNICK, M. (2021). A design-based perspective on synthetic control methods. *arXiv preprint arXiv:2101.09398*.

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

BOYD, S. P. and VANDENBERGHE, L. (2004). *Convex optimization*. Cambridge university press.

BROWN, G. W. (1951). Iterative solution of games by fictitious play. *Activity analysis of production and allocation*, **13** (1), 374–376.

BROWN, L. D. (2008). In-season prediction of batting averages: A field test of empirical bayes and bayes methodologies.

— and GREENSHTEIN, E. (2009). Nonparametric empirical bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics*, pp. 1685–1704.

BRUHN, J., IMBERMAN, S. and WINTERS, M. (2022). Regulatory arbitrage in teacher hiring and retention: Evidence from massachusetts charter schools. *Journal of Public Economics*, **215**, 104750.

BUBECK, S. and CESA-BIANCHI, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, **5** (1), 1–122.

BURBIDGE, J. B., MAGEE, L. and ROBB, A. L. (1988). Alternative Transformations to Handle Extreme Values of the Dependent Variable. *Journal of the American Statistical Association*, **83** (401), 123–127, publisher: [American Statistical Association, Taylor & Francis, Ltd.].

CABRAL, M., CUI, C. and DWORSKY, M. (2022). The demand for insurance and rationale for a mandate: Evidence from workers' compensation insurance. *American Economic Review*, **112** (5), 1621–68.

CAI, T. T. and LOW, M. G. (2011). Testing composite hypotheses, hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, pp. 1012–1041.

CALONICO, S., CATTANEO, M. D. and FARRELL, M. H. (2019). nprobust: Nonparametric kernel-based estimation and robust bias-corrected inference. *arXiv preprint arXiv:1906.00198*.

CAO, Y. and CHEN, S. (2022). Rebel on the canal: Disrupted trade access and social conflict in china, 1650–1911. *American Economic Review*, **112** (5), 1555–90.

CARD, D., KLUVE, J. and WEBER, A. (2010). Active Labour Market Policy Evaluations: A Meta-Analysis. *The Economic Journal*, **120** (548), F452–F477.

CARRANZA, E., GARLICK, R., ORKIN, K. and RANKIN, N. (2022). Job search and hiring with limited information about workseekers' skills. *American Economic Review*, **112** (11), 3547–83.

CESA-BIANCHI, N. and LUGOSI, G. (2006). *Prediction, learning, and games*. Cambridge university press.

CHAKRABORTY, B. and MURPHY, S. A. (2014). Dynamic treatment regimes. *Annual review of statistics and its application*, **1**, 447–464.

CHAMBERLAIN, G. (1984). Panel data. *Handbook of econometrics*, **2**, 1247–1318.

CHANDRA, A., FINKELSTEIN, A., SACARNY, A. and SYVERSON, C. (2016). Health care exceptionalism? performance and allocation in the us health care sector. *American Economic Review*, **106** (8), 2110–44.

CHEN, H.-B. and NILES-WEED, J. (2022). Asymptotics of smoothed wasserstein distances. *Potential Analysis*, **56** (4), 571–595.

CHEN, J. (2023). Mean-variance constrained priors have finite maximum bayes risk in the normal location model. *arXiv preprint arXiv:2303.08653*.

CHERNOZHUKOV, V., WÜTHRICH, K. and ZHU, Y. (2021). An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, **116** (536), 1849–1864.

CHETTY, R. (2006). A General Formula for the Optimal Level of Social Insurance. *Journal of Public Economics*, **90**, 1879–1901.

—, FRIEDMAN, J. N., HENDREN, N., JONES, M. R. and PORTER, S. R. (2020). *The opportunity atlas: Mapping the childhood roots of social mobility*. Tech. rep.

—, — and ROCKOFF, J. E. (2014a). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American economic review*, **104** (9), 2593–2632.

— and HENDREN, N. (2018). The impacts of neighborhoods on intergenerational mobility i: Childhood exposure effects. *The Quarterly Journal of Economics*, **133** (3), 1107–1162.

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

—, —, KLINE, P. and SAEZ, E. (2014b). Where is the land of opportunity? The geography of intergenerational mobility in the United States. *The Quarterly Journal of Economics*, **129** (4), 1553–1623, publisher: Oxford University Press.

CHYN, E. and KATZ, L. F. (2021). Neighborhoods matter: Assessing the evidence for place effects. *Journal of Economic Perspectives*, **35** (4), 197–222.

COEY, D. and HUNG, K. (2022). Empirical bayesian selection for value maximization. *arXiv preprint arXiv:2210.03905*.

COHN, J. B., LIU, Z. and WARDLAW, M. I. (2022). Count (and count-like) data in finance. *Journal of Financial Economics*, **146** (2), 529–551.

COLMER, J., HARDMAN, I., SHIMSHACK, J. and VOORHEIS, J. (2020). Disparities in pm2. 5 air pollution in the united states. *Science*, **369** (6503), 575–578.

CURRIE, J., KLEVEN, H. and ZWIERS, E. (2020). Technology and big data are changing economics: Mining text to track methods. In *AEA Papers and Proceedings*, vol. 110, pp. 42–48.

CUTLER, D. M., GLAESER, E. L. and VIGDOR, J. L. (1999). The rise and decline of the american ghetto. *Journal of political economy*, **107** (3), 455–506.

DE BRAUW, A. and HERSKOWITZ, S. (2021). Income Variability, Evolving Diets, and Elasticity Estimation of Demand for Processed Foods in Nigeria. *American Journal of Agricultural Economics*, **103** (4), 1294–1313, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajae.12139.

DEDECKER, J. and MICHEL, B. (2013). Minimax rates of convergence for wasserstein deconvolution with supersmooth errors in any dimension. *Journal of Multivariate Analysis*, **122**, 278–291.

DELIUS, A. and STERCK, O. (2020). Cash Transfers and Micro-Enterprise Performance: Theory and Quasi-Experimental Evidence from Kenya. *SSRN Electronic Journal*.

DELLAVIGNA, S. and LINOS, E. (2022). Rcts to scale: Comprehensive evidence from two nudge units. *Econometrica*, **90** (1), 81–116.

DERENONCOURT, E. (2022). Can you move to opportunity? evidence from the great migration. *American Economic Review*, **112** (2), 369–408.

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

DIMICK, J. B., STAIGER, D. O. and BIRKMEYER, J. D. (2010). Ranking hospitals on surgical mortality: the importance of reliability adjustment. *Health services research*, **45** (6p1), 1614–1629.

DOUDCHENKO, N. and IMBENS, G. W. (2016). *Balancing, regression, difference-in-differences and synthetic control methods: A synthesis*. Tech. rep., National Bureau of Economic Research.

DOYLE, J. J., GRAVES, J. A., GRUBER, J. *et al.* (2017). *Evaluating measures of hospital quality*. Tech. rep., National Bureau of Economic Research.

DURLAUF, S. N., KOURTELLOS, A. and TAN, C. M. (2022). The great gatsby curve. *Annual Review of Economics*, **14**, 571–605.

EFRON, B. (2014). Two modeling strategies for empirical bayes estimation. *Statistical science: a review journal of the Institute of Mathematical Statistics*, **29** (2), 285.

— (2016). Empirical bayes deconvolution estimates. *Biometrika*, **103** (1), 1–20.

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

— and MORRIS, C. (1975). Data analysis using stein's estimator and its generalizations. *Journal of the American Statistical Association*, **70** (350), 311–319.

EGAN, M., MATVOS, G. and SERU, A. (2022). When harry fired sally: The double standard in punishing misconduct. *Journal of Political Economy*, **130** (5), 1184–1248.

EINAV, L., FINKELSTEIN, A. and MAHONEY, N. (2022). *Producing Health: Measuring Value Added of Nursing Homes*. Tech. rep., National Bureau of Economic Research.

ELLIOTT, G., KUDRIN, N. and WÜTHRICH, K. (2022). Detecting p-hacking. *Econometrica*, **90** (2), 887–906.

FABER, B. and GAUBERT, C. (2019). Tourism and economic development: Evidence from mexico's coastline. *American Economic Review*, **109** (6), 2245–93.

FAN, Y., GUERRE, E. and ZHU, D. (2017). Partial identification of functionals of the joint distribution of "potential outcomes". *Journal of Econometrics*, **197** (1), 42–59.

FAY, R. E. and HERRIOT, R. A. (1979). Estimates of income for small places: an application of james-stein procedures to census data. *Journal of the American Statistical Association*, **74** (366a), 269–277.

FERMAN, B. and PINTO, C. (2021). Synthetic controls with imperfect pretreatment fit. *Quantitative Economics*, **12** (4), 1197–1221.

FETZER, T., SOUZA, P. C., VANDEN EYNDE, O. and WRIGHT, A. L. (2021). Security transitions. *American Economic Review*, **111** (7), 2275–2308.

FINKELSTEIN, A., GENTZKOW, M. and WILLIAMS, H. (2021). Place-based drivers of mortality: Evidence from migration. *American Economic Review*, **111** (8), 2697–2735.

FLASPOHLER, G. E., ORABONA, F., COHEN, J., MOUATADID, S., OPRESCU, M., ORENSTEIN, P. and MACKEY, L. (2021). Online learning with optimism and delay. In *International Conference on Machine Learning*, PMLR, pp. 3363–3373.

FU, L. J., JAMES, G. M. and SUN, W. (2020). Nonparametric empirical bayes estimation on heterogeneous data. *arXiv preprint arXiv:2002.12586*.

GANDHI, A., LU, Z. and SHI, X. (2023). Estimating demand for differentiated products with zeroes in market share data. *Quantitative Economics*, **14** (2), 381–418, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/QE1593.

George, E. I., Ročková, V., Rosenbaum, P. R., Satopää, V. A. and Silber, J. H. (2017). Mortality rate estimation and standardization for public reporting: Medicare's hospital compare. *Journal of the American Statistical Association*, **112** (519), 933–947.

Giacomini, R., Lee, S. and Sarpietro, S. (2023). A robust method for microforecasting and estimation of random effects.

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

Gourieroux, C., Monfort, A. and Trognon, A. (1984). Pseudo maximum likelihood methods: Applications to poisson models. *Econometrica: Journal of the Econometric Society*, pp. 701–720.

Gu, J. and Koenker, R. (2017). Empirical bayesball remixed: Empirical bayes methods for longitudinal data. *Journal of Applied Econometrics*, **32** (3), 575–599.

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

— and Walters, C. (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.

Hannan, J. (1958). Approximation to bayes risk in repeated play. In *Contributions to the Theory of Games (AM-39), Volume III*, Princeton University Press, pp. 97–140.

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

Hazan, E. (2019). Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*.

—, Agarwal, A. and Kale, S. (2007). Logarithmic regret algorithms for online convex optimization. *Machine Learning*, **69** (2-3), 169–192.

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

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, **47** (1), 153–161, publisher: [Wiley, Econometric Society].

Hirshberg, D. A. (2021). Least squares with error in variables. *arXiv preprint arXiv:2104.08931*.

Hjort, J. and Poulsen, J. (2019). The arrival of fast internet and employment in africa. *American Economic Review*, **109** (3), 1032–79.

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

Ignatiadis, N. and Wager, S. (2019). Covariate-powered empirical bayes estimation. *Advances in Neural Information Processing Systems*, **32**.

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

IMBENS, G. W. and ANGRIST, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, pp. 467–475.

JACKSON, C. K. and MACKEVICIUS, C. (2021). *The distribution of school spending impacts*. Tech. rep., National Bureau of Economic Research.

JIANG, W. (2020). On general maximum likelihood empirical bayes estimation of heteroscedastic iid normal means. *Electronic Journal of Statistics*, **14** (1), 2272–2297.

— and ZHANG, C.-H. (2009). General maximum likelihood empirical bayes estimation of normal means. *The Annals of Statistics*, **37** (4), 1647–1684.

— and — (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.

JING, B.-Y., LI, Z., PAN, G. and ZHOU, W. (2016). On sure-type double shrinkage estimation. *Journal of the American Statistical Association*, **111** (516), 1696–1704.

JOHNSON, M. S. (2020). Regulation by shaming: Deterrence effects of publicizing violations of workplace safety and health laws. *American Economic Review*, **110** (6), 1866–1904.

KAHN-LANG, A. and LANG, K. (2020). The Promise and Pitfalls of Differences-in-Differences: Reflections on 16 and Pregnant and Other Applications. *Journal of Business & Economic Statistics*, **38** (3), 613–620.

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

KALAI, A. and VEMPALA, S. (2005). Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, **71** (3), 291–307.

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

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pp. 887–906.

KITAGAWA, T. and TETENOV, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, **86** (2), 591–616.

KLINE, P., ROSE, E. K. and WALTERS, C. (2023). A discrimination report card. *arXiv preprint arXiv:2306.13005*.

—, — and WALTERS, C. R. (2022). Systemic discrimination among large us employers. *The Quarterly Journal of Economics*, **137** (4), 1963–2036.

KOENKER, R. and GU, J. (2017). Rebayes: an r package for empirical bayes mixture methods. *Journal of Statistical Software*, **82**, 1–26.

— and — (2019). Comment: Minimalist g-modeling.

— and MIZERA, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association*, **109** (506), 674–685.

KOROTIN, A., V'YUGIN, V. and BURNAEV, E. (2018). Aggregating strategies for long-term forecasting. In *Conformal and Probabilistic Prediction and Applications*, PMLR, pp. 63–82.

KUBOKAWA, T., ROBERT, C. and SALEH, A. M. E. (1993). Estimation of noncentrality parameters. *Canadian Journal of Statistics*, **21** (1), 45–57.

KWON, S. (2021). Optimal shrinkage estimation of fixed effects in linear panel data models. *EliScholar–A Digital Platform for Scholarly Publishing at Yal e*, p. 1.

LALIBERTÉ, J.-W. (2021). Long-term contextual effects in education: Schools and neighborhoods. *American Economic Journal: Economic Policy*, **13** (2), 336–377.

LAZEAR, E. P. (2001). Educational production. *The Quarterly Journal of Economics*, **116** (3), 777–803.

LEE, D. S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *The Review of Economic Studies*, **76** (3), 1071–1102.

LEHMANN, E. L. and CASELLA, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

LEI, J., G'SELL, M., RINALDO, A., TIBSHIRANI, R. J. and WASSERMAN, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, **113** (523), 1094–1111.

LIANG, T. (2000). On an empirical bayes test for a normal mean. *Annals of statistics*, pp. 648–655.

LINDSAY, B. G. (1995). Mixture models: theory, geometry, and applications. Ims.

LIU, L., MOON, H. R. and SCHORFHEIDE, F. (2020). Forecasting with dynamic panel data models. *Econometrica*, **88** (1), 171–201.

LUO, J., BANERJEE, T., MUKHERJEE, G. and SUN, W. (2023). Empirical bayes estimation with side information: A nonparametric integrative tweedie approach.

MACKINNON, J. G. and MAGEE, L. (1990). Transforming the Dependent Variable in Regression Models. *International Economic Review*, **31** (2), 315–339, publisher: [Economics Department of the University of Pennsylvania, Wiley, Institute of Social and Economic Research, Osaka University].

MANSKI, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica*, **72** (4), 1221–1246.

— and PEPPER, J. V. (2018). How do right-to-carry laws affect crime rates? coping with ambiguity using bounded-variation assumptions. *Review of Economics and Statistics*, **100** (2), 232–244.

MEAGER, R. (2022). Aggregating distributional treatment effects: A bayesian hierarchical analysis of the microcredit literature. *American Economic Review*, **112** (6), 1818–47.

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

MIRENDA, L., MOCETTI, S. and RIZZICA, L. (2022). The economic effects of mafia: firm level evidence. *American Economic Review*, **112** (8), 2748–73.

MOGSTAD, M., ROMANO, J. P., SHAIKH, A. and WILHELM, D. (2020). *Inference for ranks with applications to mobility across neighborhoods and academic achievement across countries*. Tech. rep., National Bureau of Economic Research.

—, —, SHAIKH, A. M. and WILHELM, D. (2023). A comment on:"invidious comparisons: Ranking and selection as compound decisions" by jiaying gu and roger koenker. *Econometrica*, **91** (1), 53–60.

MONTIEL OLEA, J. L., O'FLAHERTY, B. and SETHI, R. (2021). Empirical bayes counterfactuals in poisson regression with an application to police use of deadly force. *Available at SSRN 3857213*.

MORETTI, E. (2021). The effect of high-tech clusters on the productivity of top inventors. *American Economic Review*, **111** (10), 3328–75.

MORRIS, C. N. (1983). Parametric empirical bayes inference: theory and applications. *Journal of the American statistical Association*, **78** (381), 47–55.

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

MULLAHY, J. (2001). Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice: Comment. *Journal of Business & Economic Statistics*, **19** (1), 23–25, publisher: American Statistical Association, Taylor & Francis, Ltd.

— and NORTON, E. (2023). Why transform y? the pitfalls of transformed regressions with a mass at zero. *Oxford Bulletin of Economics and Statistics*, **Forthcoming**.

MÜLLER, U. K. and WATSON, M. W. (2022). Spatial unit roots. *manuscript, Princeton University*.

MÜLLER, U. K. (2023). A More Robust t-Test. *The Review of Economics and Statistics*, pp. 1–46.

NORRIS, S., PECENCO, M. and WEAVER, J. (2021). The effects of parental and sibling incarceration: Evidence from ohio. *American Economic Review*, **111** (9), 2926–63.

OLIVEIRA, N. L., LEI, J. and TIBSHIRANI, R. J. (2021). Unbiased risk estimation in the normal means problem via coupled bootstrap techniques. *arXiv preprint arXiv:2111.09447*.

OLLEY, G. S. and PAKES, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, **64** (6), 1263–1297.

ORABONA, F. (2019). A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*.

PENSKY, M. (2017). Minimax theory of estimation of linear functionals of the deconvolution density with or without sparsity.

POLYANSKIY, Y. and WU, Y. (2020). Self-regularizing property of nonparametric maximum likelihood estimator in mixture models. *arXiv preprint arXiv:2008.08244*.

— and — (2021). Sharp regret bounds for empirical bayes and compound decision problems. *arXiv preprint arXiv:2109.03943*.

RAKHLIN, A., SRIDHARAN, K. and TEWARI, A. (2011). Online learning: Stochastic, constrained, and smoothed adversaries. *Advances in neural information processing systems*, **24**.

RAMBACHAN, A. (2021). Identifying prediction mistakes in observational data.

— and ROTH, J. (2023). A More Credible Approach to Parallel Trends. *The Review of Economic Studies*, **90** (5), 2555–2591.

ROBBINS, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, **58** (5), 527–535.

— (1956). An empirical bayes approach to statistics.

ROBBINS, M. W., SAUNDERS, J. and KILMER, B. (2017). A framework for synthetic control methods with high-dimensional, micro-level data: evaluating a neighborhood-specific crime intervention. *Journal of the American Statistical Association*, **112** (517), 109–126.

ROGALL, T. (2021). Mobilizing the masses for genocide. *American Economic Review*, **111** (1), 41–72.

ROTH, J. (2022). Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends. *American Economic Review: Insights*, **4** (3), 305–322.

— and SANT'ANNA, P. H. (2023). When is parallel trends sensitive to functional form? *Econometrica*, **91** (2), 737–747.

SAHA, S. and GUNTUBOYINA, A. (2020). On the nonparametric maximum likelihood estimator for gaussian location mixture densities with application to gaussian denoising. *The Annals of Statistics*, **48** (2), 738–762.

SANTOS SILVA, J. M. C. and TENREYRO, S. (2006). The Log of Gravity. *The Review of Economics and Statistics*, **88** (4), 641–658.

SEN, B. (2018). A gentle introduction to empirical process theory and applications. *Lecture Notes, Columbia University*, **11**, 28–29.

SEQUEIRA, S. (2016). Corruption, Trade Costs, and Gains from Tariff Liberalization: Evidence from Southern Africa. *American Economic Review*, **106** (10), 3029–3063.

SHALEV-SHWARTZ, S. (2011). Online learning and online convex optimization. *Foundations and trends in Machine Learning*, **4** (2), 107–194.

SHEN, Y. and WU, Y. (2022). Empirical bayes estimation: When does g-modeling beat f-modeling in theory (and in practice)? *arXiv preprint arXiv:2211.12692*.

SHI, C., SRIDHAR, D., MISRA, V. and BLEI, D. (2022). On the assumptions of synthetic control methods. In *International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 7163–7175.

SOLOFF, J. A., GUNTUBOYINA, A. and SEN, B. (2021). Multivariate, heteroscedastic empirical bayes via nonparametric maximum likelihood. *arXiv preprint arXiv:2109.03466*.

STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The annals of Statistics*, pp. 1348–1360.

STOYE, J. (2009). Minimax regret treatment choice with finite samples. *Journal of Econometrics*, **151** (1), 70–81.

THAKRAL, N. and TÔ, L. T. (2023). When are estimates independent of measurement units?

TOBIN, J. (1958). Estimation of Relationships for Limited Dependent Variables. *Econometrica*, **26** (1), 24–36, publisher: [Wiley, Econometric Society].

TSYBAKOV, A. (2008). *Introduction to Nonparametric Estimation*. Springer Series in Statistics, Springer New York.

VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2009). Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak convergence*. Springer.

VERSHYNIN, R. (2018). *High-dimensional probability: An introduction with applications in data science*, vol. 47. Cambridge university press.

VIVIANO, D. and BRADIC, J. (2019). Synthetic learner: model-free inference on treatments over time. *arXiv preprint arXiv:1904.01490*.

VOINOV, V. G. and NIKULIN, M. S. (2012). *Unbiased estimators and their applications: volume 1: univariate case*, vol. 263. Springer Science & Business Media.

WEINBERGER, M. J. and ORDENTLICH, E. (2002). On delayed prediction of individual sequences. *IEEE Transactions on Information Theory*, **48** (7), 1959–1976.

WEINSTEIN, A., MA, Z., BROWN, L. D. and ZHANG, C.-H. (2018). Group-linear empirical bayes estimates for a heteroscedastic normal mean. *Journal of the American Statistical Association*, **113** (522), 698–710.

WERNERFELT, N., TUCHMAN, A., SHAPIRO, B. and MOAKLER, R. (2022). Estimating the value of offsite data to advertisers on meta. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (114).

WILLIAMS, C. B. (1937). The Use of Logarithms in the Interpretation of Certain Entomological Problems. *Annals of Applied Biology*, **24** (2), 404–414, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1744-7348.1937.tb05042.x.

WOOLDRIDGE, J. M. (1992). Some Alternatives to the Box-Cox Regression Model. *International Economic Review*, **33** (4), 935–955, publisher: [Economics Department of the University of Pennsylvania, Wiley, Institute of Social and Economic Research, Osaka University].

— (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

— (2023). Simple approaches to nonlinear difference-in-differences with panel data. *The Econometrics Journal*, **26** (3), C31–C66.

XIE, X., KOU, S. and BROWN, L. D. (2012). Sure estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association*, **107** (500), 1465–1479.

XU, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, **25** (1), 57–76.

ZHANG, C.-H. (1997). Empirical bayes and compound estimation of normal means. *Statistica Sinica*, **7** (1), 181–193.

ZHANG, J. L. and RUBIN, D. B. (2003). Estimation of Causal Effects via Principal Stratification When Some Outcomes are Truncated by "Death". *Journal of Educational and Behavioral Statistics*, **28** (4), 353–368, publisher: American Educational Research Association.

—, — and MEALLI, F. (2008). Evaluating the effects of job training programs on wages through principal stratification. In T. Fomby, R. Carter Hill, D. L. Millimet, J. A. Smith and E. J. Vytlacil (eds.), *Modelling and Evaluating Treatment Effects in Econometrics*, *Advances in Econometrics*, vol. 21, Emerald Group Publishing Limited, pp. 117–145.

—, — and — (2009). Likelihood-Based Analysis of Causal Effects of Job-Training Programs Using Principal Stratification. *Journal of the American Statistical Association*, **104** (485), 166–176.

Appendix A

Appendix to Chapter 1

A.1 Proofs and additional results

A.1.1 Proofs of Theorems 1.2.2, 1.3.1 and 1.3.2

We reproduce Theorem 5 of Hazan *et al.* (2007) in our notation.

Theorem A.1.1 (Theorem 5, Hazan *et al.* (2007))**.** *Assume that for all t, the function $\ell_t : \Theta \to \mathbb{R}$ can be written as*

$$\ell_t(\theta) = g_t(v_t'\theta)$$

for a univariate convex function $g_t : \mathbb{R} \to \mathbb{R}$ and some vector $v_t \in \mathbb{R}^n$. Assume that for some $R, a, b > 0$, we have $\|v_t\|_2 \leq R$ and for all $\theta \in \Theta$, we have $|g_t'(v_t'\theta)| \leq b$ and $g_t''(v_t'\theta) \geq a$, for all t. Then FTL on ℓ_t satisfies the following regret bound:

$$\mathrm{Regret}_T \leq \frac{2nb^2}{a}\left[\log\left(\frac{DRaT}{b}\right) + 1\right]$$

where $D = \max_{x,y\in\Theta}\|x - y\|_2$ is the diameter of Θ.

Proof of Theorem 1.2.2. Theorem 1.2.2 follows immediately from Theorem 5 in Hazan *et al.* (2007), reproduced in our notation as Theorem A.1.1. The proof of this theorem relies solely on optimality of θ_t (and the associated first-order condition); thus, in the case of multiple minima when minimizing $\sum_{t=1}^{s} \ell_t(\theta)$, any particular sequence of minima $\{\theta_t\}$ satisfies the guarantee.

Since Θ is the simplex, we know

$$D = \max_{\theta_1, \theta_2 \in \Theta} \|\theta_1 - \theta_2\|_2 \leq \max_{\theta_1, \theta_2 \in \Theta} \|\theta_1 - \theta_2\|_1 \leq \max_{\theta_1, \theta_2 \in \Theta} \|\theta_1\|_1 + \|\theta_2\|_1 = 2.$$

We choose $g_t(x) = \frac{1}{2}(y_{0t} - x)^2$ with $g_t'(x) = x - y_{0t}$ and $g_t''(x) = 1$. (The scaling by $1/2$ means that we obtain a bound on $1/2$ times the regret.) The vectors $v_t = \mathbf{y}_t$, whose dimensions are $n = N$ and whose 2-norms are bounded by $R = \sqrt{N}$. Note that $|\mathbf{y}_t'\theta| = |v_t'\theta| \leq \|v_t\|_\infty \|\theta\|_1 \leq 1$. Hence, $|g_t'(v_t'\theta)| = |\mathbf{y}_t'\theta - y_{0t}| \leq |\mathbf{y}_t'\theta| + |y_{0t}| \leq 2 \equiv b$ and $g_t''(x) \geq 1 \equiv a$. To summarize, we have $R = \sqrt{N}, a = 1, b = 2, D = 2$, and $n = N$.

Plugging in, we have

$$\frac{1}{2}\text{Regret}_T \leq 8N(\log(\sqrt{N}T) + 1),$$

which rearranges into the claim. □

Proof of Theorem 1.3.1 . The proof for Theorem 1.3.1 follows similarly, now with

$$g_t(x) = \frac{T}{2}\pi_t(y_{0t} - x)^2 \qquad g_t'(x) = T\pi_t(x - y_{0t}) \qquad g_t''(x) = T\pi_t.$$

Note that, since $\frac{1}{CT} \leq \pi_t \leq \frac{C}{T}$, we can take $a = 1/C$ and $b = 2C$. Doing so yields the expression in Theorem 1.3.1. □

Proof of Theorem 1.3.2. For Theorem 1.3.2, and in particular (1.12), by $(1, \infty)$-Hölder's inequality,

$$\sum_{t=1}^T \pi_t(y_{0t} - \hat{\theta}_t'\mathbf{y}_t)^2 \leq \left(\max_t \pi_t\right) \sum_{t=1}^T (y_{0t} - \hat{\theta}_t'\mathbf{y}_t)^2 \leq \frac{C}{T}\sum_{t=1}^T (y_{0t} - \hat{\theta}_t'\mathbf{y}_t)^2.$$

We then apply Theorem 1.2.2 to bound $\sum_{t=1}^T (y_{0t} - \hat{\theta}_t'\mathbf{y}_t)^2 = \min_{\theta \in \Theta} \sum_{t=1}^T (y_{0t} - \theta'\mathbf{y}_t)^2 + \text{Regret}_T$.

(1.13) follows immediately from (1.12) by taking the expectation E_Q, noting that

$$\begin{aligned} E_Q[(y_{0S} - \hat{\theta}_S' y_S)^2] &= E_Q\left[\sum_{t=1}^T \mathbb{1}(S = t)(y_{0t} - \hat{\theta}_t' y_t)^2\right] \\ &= E\left[E\left[\sum_{t=1}^T \mathbb{1}(S = t)(y_{0t} - \hat{\theta}_t' y_t)^2 \mid \mathbf{Y}\right]\right] \\ &= E\left[\sum_{t=1}^T Q(S = t \mid \mathbf{Y})(y_{0t} - \hat{\theta}_t' y_t)^2\right] \end{aligned}$$

We then apply (1.12) to complete the proof. □

A.1.2 Lack of regret control for fixed strategies

Lemma A.1.2. *In the online convex optimization setup, suppose the class of loss functions available to the adversary satisfies the following property: There exists $\epsilon > 0$ such that for any $\theta \in \Theta$, there exists $\tilde{\theta}$ and $\ell_1, \ldots, \ell_T$, for which $\ell_t(\tilde{\theta}) \le \ell_t(\theta) - \epsilon$. Then, the regret of any fixed strategy that outputs $\theta_t = \theta$ for every period is at least ϵT.*

Proof. Let $\ell_t, \tilde{\theta}$ be the sequence of loss functions and alternative satisfying the required property on the class of loss functions. Then $\operatorname{Regret}_T(\theta) \ge \sum_t \ell_t(\theta) - \sum_t \ell_t(\tilde{\theta}) = \epsilon T$. □

It is easy to see that the loss functions in the panel prediction problem are rich enough to satisfy the property in Theorem A.1.2. Fix, say, $\epsilon < 0.0001$. For any θ, we can find $\tilde{\theta} \in \Theta$ where $\|\tilde{\theta} - \theta\|_1 \ge \sqrt{\epsilon}$. Then, there exists some $\mathbf{y}$, $\|\mathbf{y}\|_\infty \le 1$ where

$$|(\tilde{\theta} - \theta)'\mathbf{y}| = \max_{\|\mathbf{y}\|_\infty \le 1} |(\tilde{\theta} - \theta)'\mathbf{y}| = \|\tilde{\theta} - \theta\|_1 \ge \sqrt{\epsilon}$$

since $\|\cdot\|_1$ is the dual norm to $\|\cdot\|_\infty$. The adversary chooses $\mathbf{y}_t = \mathbf{y}$ for all $t \in [T]$ and $y_{0t} = \tilde{\theta}'\mathbf{y}_t$. Then $\ell_t(\tilde{\theta}) = 0$ but $\ell_t(\theta) \ge (\sqrt{\epsilon})^2 = \epsilon$.

A.1.3 Static DID regret control

We could consider affine predictors with bounded intercepts

$$f(\mathbf{y}_t; \theta_0, \theta_1) = \theta_0 + \theta_1'\mathbf{y}_t \quad \Theta = [-2, 2] \times \Delta^{N-1}.$$

This choice corresponds to variations of synthetic control proposed by Doudchenko and Imbens (2016) and Ferman and Pinto (2021) in efforts to mimic behavior of DID estimators.[1] Our regret bound from Theorem 1.2.2 generalizes immediately to the affine predictions, where the benchmark oracle the

[1] Synthetic control with an intercept is equivalent to synthetic control with demeaned data $\{y_s - \frac{1}{t}\sum_{k \le t} y_k : s = 1, \ldots, t\}$ (Ferman and Pinto, 2021), since the constraint that $\theta_0 \in [-2, 2]$ does not bind.

regret measures against is

$$\min_{(\theta_0,\theta_1)\in\Theta} \sum_{t=1}^{T} (y_{0t} - \theta_0 - \theta_1' \mathbf{y}_t)^2. \tag{A.1}$$

(A.1) simultaneously chooses the best intercept and the best set of convex weights in hindsight. Because (A.1) is limited to using the same intercept for prediction in each period, it is, in some sense, a *static* DID estimator.

Theorem 1.2.2 can be adapted to show that synthetic control with an intercept is competitive against static DID.

Proposition A.1.3. *Consider demeaned synthetic control, where the analyst outputs the prediction* $\hat{y}_t = \hat{\theta}_{0t} + \hat{\theta}_t' \mathbf{y}_t$ *by solving the least-squares problem*

$$\hat{\theta}_{0t}, \hat{\theta}_t = \operatorname*{arg\,min}_{\theta_0,\theta\in[-2,2]\times\Delta^{N-1}} \sum_{s<t} (y_{0s} - \theta_0 - \theta' \mathbf{y}_s)^2.$$

Then, under bounded data $\|\mathbf{Y}\|_\infty \le 1$*, we have the following regret bound:*

$$\sum_{t=1}^{T} (y_{0t} - \hat{y}_t)^2 - \min_{\theta_0,\theta\in[-2,2]\times\Delta^{N-1}} \sum_{t=1}^{T} (y_{0s} - \theta_0 - \theta' \mathbf{y}_s)^2 \le CN \log T$$

for some constant C*.*

Proof. We define the loss as $\frac{1}{2}(x-y)^2$, which only affects the regret up to a factor of 2. Theorem A.1.3 can be proved with Theorem A.1.1. Note that the diameter of the parameter space $[-2,2]\times\Delta^{N-1}$ can be bounded by $D = 2\cdot\sqrt{2^2+1} = 2\sqrt{5}$. The 2-norm of the vector $v_t = [1, \mathbf{y}_t']'$ is now bounded by $R = \sqrt{N+1}$. The 1-norm of the parameter vector $\vartheta = [\theta_0, \theta']'$ is now bounded by $2+1 = 3$. Hence, $|v_t'\vartheta| \le 3$. Hence, we may take $b = 3+1 = 4$ and $a = 1$. Plugging in, we obtain

$$\text{Regret}_T \le 64N \left[\log\left(\frac{\sqrt{5}}{2} \sqrt{N+1} T \right) + 1 \right] < CN \log T$$

for some C. □

A.1.4 Proof of Theorem 1.3.3

Similarly to the proof of Theorem A.1.3, suppose the adversary picks the differences $|\tilde{y}_{it}| \le 2$, *without* the constraint that the resulting levels obey the restriction $\|\mathbf{Y}\|_\infty \le 1$. An application of Theorem A.1.1

shows that

$$\sum_{t=1}^{T}(\tilde{y}_{0t} - \hat{\theta}_t'\tilde{\mathbf{y}}_t)^2 - \min_{\theta\in\Theta}\sum_{t=1}^{T}(\tilde{y}_{0t} - \theta'\tilde{\mathbf{y}}_t)^2 \le CN\log T$$

for some C, uniformly over $|\tilde{y}_{it}| \le 2$, where $\hat{\theta}_t$ is the FTL strategy on the data $\tilde{y}_{it}$, which is exactly the synthetic control on the differenced data when $\mathbf{Y}$ is chosen by the adversary.

Now, given any $\|\mathbf{Y}\|_\infty \le 1$, we have that the corresponding differences $\tilde{y}_{it}$ obey the above regret bound, since they are bounded by 2. Moreover, for both synthetic control ($\theta_t = \hat{\theta}_t$) and the oracle σ_{TWFE} ($\theta_t = \theta$), the prediction error of the data y_{0t} is equal to the prediction error on the differences:

$$y_{0t} - \hat{y}_t = \frac{1}{t-1}\sum_{s<t} y_{0s} + \tilde{y}_{0t} - \left(\frac{1}{t-1}\sum_{s<t} y_{0s} + \theta_t'\tilde{\mathbf{y}}_t\right) = \tilde{y}_{0t} - \theta_t'\tilde{\mathbf{y}}_t.$$

Hence, we may rewrite the above regret bound as the bound

$$\sum_{t=1}^{T}(y_{0t} - \hat{y}_t)^2 - \min_{\theta\in\Theta}\sum_{t=1}^{T}(y_{0t} - \hat{y}_t(\sigma_{\mathrm{TWFE}}(\theta)))^2 \le CN\log T.$$

A.1.5 Proof of Theorem 1.3.5

Theorem A.1.4. *Assume that*

1. $\ell_t(\theta) \equiv \ell(\theta'\mathbf{y}_t, y_{0t})$ *is convex in* θ *for any* $\mathbf{Y}$.
2. *The regularizer* $\Phi(\theta)$ *is* 1*-strongly convex in some norm* $\|\cdot\|$. *Normalize* Φ *such that its minimum over* Θ *is zero and maximum is* $K < \infty$.
3. *All subgradients* $\nabla_\theta \ell_t(\theta)$ *are bounded in the dual norm* $\|\cdot\|_*$, *uniformly over* $\Theta, \mathbf{Y}$:

$$\|\nabla_\theta \ell_t(\theta)\|_*^2 \le G.$$

Then FTRL attains the regret bound

$$\mathrm{Regret}_T \le \frac{K}{\eta} + \frac{\eta T G}{2}.$$

We first reproduce Corollary 7.9 from Orabona (2019) in our notation. Consider an FTRL algorithm that regularizes according to

$$\theta_t \in \arg\min_{\theta}\sum_{s\le t}\ell_s(\theta) + \frac{1}{\eta}\Phi(\theta).$$

This corresponds to choosing $\eta_t = \eta$, $\psi(x) = \Phi(x)$, and $\min_\theta \Phi(\theta) = 0$ in Orabona (2019).

Theorem A.1.5 (Corollary 7.9, Orabona (2019)). *Let ℓ_t be a sequence of convex loss functions. Let $\Phi : \Theta \to \mathbb{R}$ be μ-strongly convex with respect to the norm $\|\cdot\|$. Then, FTRL guarantees*

$$\sum_{t=1}^{T} \ell_t(\theta_t) - \sum_{t=1}^{T} \ell_t(\theta) \le \frac{\Phi(\theta)}{\eta} + \frac{\eta}{2\mu} \sum_{t=1}^{T} \|g_t\|_*^2$$

for all subgradients $g_t \in \partial \ell_t(\theta_t)$ and all $\theta \in \Theta$, where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

Proof of Theorem A.1.4. Theorem A.1.4 then follows immediately where $\|g_t\|_*^2 \le G$, $\Phi(\theta) \le K$, and $\mu = 1$. □

Proof of Theorem 1.3.5. For both squared and absolute losses, we can bound the gradient of the loss function in terms of

$$\|\nabla_\theta \ell_t(\theta)\|_* = \|\nabla q(y - \hat{y}) \cdot \mathbf{y}_t\|_* = |\nabla q(y - \hat{y})| \|\mathbf{y}_t\|_* \le 2 \sup_{\|\mathbf{y}\|_\infty \le 1} \|\mathbf{y}\|_*$$

under any norm, where $q(t) = t^2/2$ or $q(t) = |t|$. This is because (i) for squared loss, the gradient $|\nabla f| = |y - \hat{y}|$ is bounded by 2 and (ii) for absolute loss, the subgradients $|\nabla f|$ are bounded by 1 and hence by 2. Hence, we should pick G to be $4 \sup_{\|\mathbf{y}\|_\infty \le 1} \|\mathbf{y}\|_*^2$.

For the quadratic penalty assumed, it is 1-strongly convex with respect to $\|\cdot\|_2$ by the assumption that the minimum eigenvalue of its Hessian is 1. Thus the dual norm $\|\cdot\|_*$ is also the Euclidean norm, and we may take $G = 4N$. This yields the bound by Theorem A.1.4, since

$$\frac{K}{\sqrt{K(2TN)^{-1}}} + \frac{4NT}{2}\sqrt{\frac{K}{2TN}} = 2\sqrt{2}\sqrt{NTK}.$$

Setting $K = 1/2$ yields the ridge penalty result.

The entropy penalty is 1-strongly convex with respect to $\|\cdot\|_1$.[2] Thus we may take $G = 4\|\mathbf{y}_t\|_\infty^2 =$

[2] This is a well-known result in online convex optimization. To prove it, we first note that

$$\Phi(y) = \Phi(x) + \nabla\Phi(x)'(y - x) + D_{\mathrm{KL}}(y\|x),$$

where $\Phi(x) = \sum_i x_i \log x_i + C$, $D_{\mathrm{KL}}(y\|x) = \sum_i y_i \log(y_i/x_i)$, and x, y lie in the interior of the simplex. Pinsker's inequality then implies

$$\Phi(y) \ge \Phi(x) + \nabla\Phi(x)'(y - x) + \frac{1}{2}\|x - y\|_1^2.$$

This is exactly the definition of 1-strong convexity with respect to $\|\cdot\|_1$.

4. The maximum of entropy (shifted so that its minimum is zero) can take $K = \log N$. This yields the bound via Theorem A.1.4. □

A.1.6 Two-way fixed effect calculation

Consider the TWFE regression with known, nonnegative weights $\sum_{i=1}^N w_i = 1$ and the normalization $w_0 = 1$:

$$\arg\min_{\mu_i, \alpha_t} \sum_{\substack{i,t:(i,t)\neq(0,S)\\ i\in\{0,\ldots,N\}\\ t\in[S]}} w_i(y_{it} - \mu_i - \alpha_t)^2.$$

We may eliminate $(i,t) = (0,S)$ from the sum since $\lambda \mathbb{1}(i = 0, S = t)$ in (1.14) absorbs that term, leaving μ_i, α_t unaffected. Consider forecasting y_{0S} with $\mu_0 + \alpha_S$ that solves the above program. As a reminder, in this subsection, we show that the estimated $\mu_0 + \alpha_S$ takes the form of forecasting with weighted average on differenced data.

The first-order condition for μ_i takes the form

$$\sum_{t=1}^{S-1} y_{it} - \mu_i - \alpha_t + \mathbb{1}(i \neq 0)(y_{iS} - \mu_i - \alpha_t) = 0.$$

Hence,

$$\mu_i = \begin{cases} \bar{y}_i - \bar{\alpha} & i \neq 0 \\ \bar{y}_0 - \frac{S}{S-1}\bar{\alpha} + \frac{1}{S-1}\alpha_S & i = 0 \end{cases}$$

where $\bar{\alpha} = \frac{1}{S}\sum_{t=1}^S \alpha_t$ and $\bar{y}_i$ is the sample mean of observations for unit i over time $1, \ldots, S$, with the understanding that y_{0S} is not included for $\bar{y}_0$. Hence, the forecast is $\mu_0 + \alpha_S = \bar{y}_0 + \frac{S}{S-1}\left(\alpha_S - \bar{\alpha}\right).$

Let us inspect the first-order condition for α_S:

$$\sum_{i=1}^N w_i(y_{iS} - \mu_i - \alpha_S) = \sum_{i=1}^N w_i(y_{iS} - \bar{y}_i + \bar{\alpha} - \alpha_S) = 0.$$

Rearrange to obtain that $\alpha_S - \bar{\alpha} = \sum_{i=1}^N w_i \left(\frac{S-1}{S} y_{iS} - \frac{1}{S}\sum_{t=1}^{S-1} y_{it}\right)$. Therefore, $\frac{S}{S-1}(\alpha_S - \bar{\alpha}) = \sum_{i=1}^N w_i \left(y_{iS} - \frac{1}{S-1}\sum_{t=1}^{S-1} y_{it}\right)$. Thus the forecast is

$$\mu_0 + \alpha_S = \frac{1}{S-1}\sum_{t=1}^{S-1} y_{0t} + \sum_{i=1}^N w_i \left(y_{iS} - \frac{1}{S-1}\sum_{t=1}^{S-1} y_{it}\right).$$

Note that arriving at this result does not use the fact that $w_0 = 1$. Hence, w_0 does not matter for $\mu_0 + \alpha_S$.

A.2 Further extensions

A.2.1 Adaptive regret

The online learning literature also has results for controlling the *adaptive regret*:

$$\text{AdaptiveRegret}_T = \sup_{1 \le r < s \le T} \sum_{t=r}^{s} \left\{ \ell_t(\theta_t) - \min_{\theta_{r,s}} \sum_{t=r}^{s} \ell_t(\theta_{r,s}) \right\}, \tag{A.2}$$

which is the worst regret over any subinterval of $[T]$. An upper bound of adaptive regret serves as an upper bound of the regret over any subperiod indexed by $r < s$. In particular, suppose we obtain a $O(\log T)$ upper bound on adaptive regret, then we obtain meaningful *average* regret upper bounds for all subperiods significantly longer than $O(\log T)$.

A simple meta-algorithm called *Follow The Leading History* (FLH) (Algorithm 31 in Hazan, 2019) serves as a wrapper for an online learning algorithm σ, such that

$$\text{AdaptiveRegret}_T(\text{FLH}(\sigma)) \le \text{Regret}_T(\sigma) + O(\log T). \tag{A.3}$$

When applied to synthetic control, FLH takes the following form. We initialize $p_1^1 = 1$ and set $\alpha = \frac{1}{4}$. At each time t, when prompted to make a prediction about y_{0t}:

1. Consider the synthetic control estimated weights $\theta_t^1, \ldots, \theta_t^t$, where θ_t^j is the synthetic control weights estimated based on data from *time horizons* $j, \ldots, t-1$.

2. Output the weighted average $\theta_t = \sum_{j=1}^t p_t^j \theta_t^j$.

3. After receiving $\mathbf{y}_t, y_{0t}$ (and hence receiving $\ell_t(\theta) = \frac{1}{2}(y_{0t} - \theta' \mathbf{y}_t)^2$), instantiate

$$p_{t+1}^i \leftarrow \frac{p_t^i e^{-\alpha \ell_t(\theta_t^i)}}{\sum_{j=1}^t p_t^j e^{-\alpha \ell_t(\theta_t^j)}} \quad 1 \le i \le t.$$

4. Set $p_{t+1}^{t+1} = \frac{1}{t+1}$ and further update

$$p_{t+1}^i \leftarrow \left(1 - \frac{1}{t+1}\right) p_{t+1}^i \quad 1 \le i \le t.$$

At each step, FLH applied to synthetic control continues to output a convex weighted average of control unit outcomes, making it a type of synthetic control algorithm. Theorem 10.5 in Hazan (2019) then implies the bound (A.3) for the above algorithm.[3] In a nutshell, FLH treats synthetic control predictions from different horizons as *expert predictions*, and applies a no-regret online learning algorithm to aggregate these expert predictions. We direct readers to Hazan (2019) for further intuitions about the algorithm.

Combined with Theorem 1.2.2 for synthetic control, we find that the adaptive regret of FLH-synthetic control is of the same order $O(N \log T + N \log N)$. This means that the average regret over any subperiod of length T' is $O\left(\frac{N \log T + N \log N}{T'}\right)$, a meaningful bound for long subperiods $T' \gg N \log T$. In other words, in a protocol where the adversary *additionally* picks a subperiod of length T', and nature subsequently samples a treatment timing uniformly randomly over the subperiod, FLH-synthetic control achieves expected regret bound of $O\left(\frac{N \log T + N \log N}{T'}\right)$. The adaptive regret bound thus partially relaxes the requirement for uniform treatment timing, and allows for expected regret control over random treatment timing on any subperiod.

A.2.2 A note on inference

Under the treatment assignment model $S \sim \text{Unif}[T]$, we may test the sharp null $H_0 : \mathbf{y}(1) = \mathbf{y}(0)$, leveraging symmetries arising from treatment assignment. This is similar in spirit to Bottmer *et al.* (2021), who consider design-based inference under random assignment of the treated unit. They compute the variance of the estimated treatment effect (for treated unit $M \sim \text{Unif}[N]$ at some fixed time S) under random assignment, holding the outcomes fixed, and propose an unbiased estimator.

[3] The proof follows immediately since $\frac{1}{2}(y_{0t} - \theta' \mathbf{y}_t)^2$ is $\frac{1}{4}$-exp-concave. That is,

$$\theta \mapsto \exp\left(-\frac{1}{4} \cdot \frac{1}{2}(y_{0t} - \theta' \mathbf{y}_t)^2\right)$$

is concave. This is because $-2 \le y_{0t} - \theta' \mathbf{y}_t \le 2$, and $g(x) = \exp\left(-\frac{1}{4} \cdot \frac{1}{2} x^2\right)$ is concave on $x \in [-2, 2]$. The Hessian of $\exp\left(-\frac{1}{4} \cdot \frac{1}{2}(y_{0t} - \theta' \mathbf{y}_t)^2\right)$ in θ is then $g''(y_{0t} - \theta' \mathbf{y}_t)\mathbf{y}_t \mathbf{y}_t'$, which is negative semidefinite.

This is also similar in spirit to unit-randomization-based placebo tests (Abadie *et al.*, 2010).

Let $y_t = y_{0t}$ for $t < S$ and let $y_t = \mathbf{y}_t(1)$ for $t \ge S$ be the observed time series of the treated unit. For any prediction $\hat{y}_t$ that does not depend on S—not limited to synthetic control predictions—we may form the residuals $r_t = |y_t - \hat{y}_t|$. One (finite-sample) test of the sharp null rejects when r_S is at least the $\lceil T(1-\alpha) \rceil^{\text{th}}$ order statistic of the sample $\{r_1, \dots, r_T\}$. Since, under the null, r_S is equally likely to equal any of $\{r_1, \dots, r_T\}$, the probability of it being the among largest $100\alpha\%$ is bounded by α. Similarly, if $S \sim \pi$ where $\pi_t \le C/T$, a least-favorable test may be constructed by rejecting when $r_t \ge r_{(T - \lfloor T\alpha/C \rfloor)}$. Informally speaking, this test is more powerful when the predictions $\hat{y}_t$ are better, and our regret guarantees are in this sense informative for inference. Moreover, note that this procedure is very similar to conformal inference (Lei *et al.*, 2018; Chernozhukov *et al.*, 2021). Conformal intervals rely on the assumption that the data is exchangeable in the underlying sampling process. This symmetry is true here by virtue of assuming $S \sim \text{Unif}[T]$, since the treated period is equally likely to be any one.

The argument above does not use the regret result. From Markov's inequality, we can control the probability for the prediction error to deviate far relative to its expectation

$$P_{S \sim \text{Unif}[T]}\left[(y_{0S} - \hat{y}_S)^2 > c\right] \le \frac{E_S[\ell_S(\theta_S)]}{c} \le \frac{1}{c}\left(\min_{\theta \in \Theta} \frac{1}{T}\sum_{i=1}^T \ell_t(\theta) + \frac{1}{T}\text{Regret}_T\right).$$

Under assumptions where the pre-treatment loss $\min_\theta \frac{1}{S-1}\sum_{t<S} \ell_t(\theta)$ is a consistent estimator for the oracle performance $\min_\theta \frac{1}{T}\sum_{i=1}^T \ell_t(\theta)$, the above observation allows for predictive confidence intervals for the untreated outcome and confidence intervals of the treatment effect, which are valid over random treatment timing.

A.2.3 Risk interpretation under idiosyncratic errors

We consider another interpretation of (1.9). In many data-generating processes,

$$E_P\left[\min_\theta \text{Risk}(\theta, \mathbf{Y}, \mathbf{y}(1))\right]$$

may not be small, because the realized data $\mathbf{Y}$ may contain certain unforecastable components. The purpose of this section is to leverage the decomposition

$$E_P[(\hat{y}_{0t} - y_{0t})^2] = E_P[\epsilon_t^2] + E_P[(\hat{y}_{0t} - \mu_t)^2],$$

where $\epsilon_t = y_{0t} - \mu_t$ is some unforecastable component satisfying $E_P[\epsilon_t \hat{y}_{0t}] = 0$. This decomposition breaks prediction errors into forecastable and unforecastable components. Because of this additive decomposition, under certain conditions on ϵ_t, we can interpret risk differences as regret on estimating the forecastable component μ_t (since $E_P[\epsilon_t^2]$ cancels in the difference). We can also decompose risk into the oracle error on estimating μ_t, the regret against the oracle on estimating μ_t, and the variance of the unforecastable errors ϵ_t.

For a fixed θ, under uniform treatment timing we have that

$$E_P[\text{Risk}(\theta, \mathbf{Y}, \mathbf{y}(1))] = E_P[E_S(y_{0S} - \mu_S)^2] + E_P[E_S(\theta' \mathbf{y}_S - \mu_S)^2]$$

for *some* mean component μ_t, possibly random, of the outcome process y_{0t}. For instance, we may take $\mu_t = E_P[y_{0t} \mid \mathbf{Y}_{1:t-1}, \mathbf{y}_t]$. For this μ_t, we can also write

$$E_P[\text{Risk}(\sigma, \mathbf{Y}, \mathbf{y}(1))] = E_P[E_S(y_{0S} - \mu_S)^2] + E_P[E_S(\hat{\theta}_t' \mathbf{y}_S - \mu_S)^2],$$

since $\hat{\theta}_t' \mathbf{y}_t$ depends solely on $\mathbf{Y}_{1:t-1}, \mathbf{y}_t$. We thus have the following implication of (1.9)

$$E_P[E_S(\hat{\theta}_t' \mathbf{y}_S - \mu_S)^2] - \min_{\theta \in \Theta} E_P[E_S(\theta' \mathbf{y}_S - \mu_S)^2] \le \frac{1}{T} \sup_{\|\mathbf{Y}\|_\infty \le 1} \text{Regret}_T(\sigma; \mathbf{Y}),$$

which says that the risk difference of estimating the conditional mean μ_t—the forecastable component of the outcome process—is upper bounded by the regret. As a corollary, if $P = P_T$ is a sequence of data-generating processes where, as $T \to \infty$,

$$\min_{\theta \in \Theta} E_P[E_S(\theta' \mathbf{y}_S - \mu_S)^2] \to 0,$$

then we obtain a consistency result for synthetic control, in that

$$E_P[E_S(\hat{\theta}_t' \mathbf{y}_S - \mu_S)^2] \to 0$$

as well.

Shifting from risk differences to risks themselves, this means that the treatment effect estimation risk for synthetic control admits the following upper bound

$$\begin{aligned} E_P[\mathrm{Risk}(\sigma, \mathbf{Y}, \mathbf{y}(1))] \le \min_{\theta \in \Theta} E_P[E_S(\theta' \mathbf{y}_S - \mu_S)^2] + \frac{1}{T} \sup_{\|\mathbf{Y}\|_\infty \le 1} \mathrm{Regret}_T(\sigma; \mathbf{Y}) \\ + E_P[E_S(y_{0S} - \mu_S)^2], \end{aligned}$$

where the first term is the best possible error on the forecastable component μ_t, the second term is the average regret, and the third term is the variance of the unforecastable component that cannot be improved upon. We think the first two terms are likely small, and the last term is unavoidable.

This argument also extends to non-uniformly random treatment timing. Suppose we have a joint distribution Q of $(\mathbf{Y}, \mathbf{y}(1), S)$ such that $\pi_t(\mathbf{Y}) = Q(S = t \mid \mathbf{Y}) \le C/T$. Suppose further that $y_{0t} = \mu_t + \epsilon_t$, where $E_Q[\epsilon_t \mid \mu_t, \pi_t, \mathbf{Y}_{1:t-1}, \mathbf{y}_t] = 0$ for some mean component μ_t.[4] Then we have a similar decomposition of the risk of estimating the treatment effect at S:

$$\begin{aligned} E_Q[(y_{0S} - \hat{\theta}_S' \mathbf{y}_t)^2] &= \sum_{t=1}^T E_Q[\pi_t(\mathbf{Y})(y_{0t} - \hat{\theta}_S' \mathbf{y}_t)^2] \\ &= \sum_{t=1}^T E_Q\left[\pi_t(\mathbf{Y})(y_{0t} - \mu_t)^2\right] + E_Q[\pi_t(\mathbf{Y})(\mu_t - \hat{\theta}_t' \mathbf{y}_t)^2] \\ &\quad + 2E_Q[\pi_t \epsilon_t (\mu_t - \hat{\theta}_t' \mathbf{y}_t)] \\ &= E_Q[\epsilon_S^2] + E_Q[(\mu_S - \hat{\theta}_S' \mathbf{y}_S)^2] && \text{(Last term is zero)} \\ &\le E_Q[\epsilon_S^2] + \frac{C}{T} \sum_{t=1}^T E_Q[(\mu_t - \hat{\theta}_t' \mathbf{y}_t)^2] && \text{((1,}\infty\text{)-Hölder's inequality)} \\ &\le E_Q[\epsilon_S^2] + C \Bigg(\min_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^T E_Q[(\mu_t - \theta' \mathbf{y}_t)^2] \\ &\qquad\qquad + \frac{1}{T} \sup_{\|\mathbf{Y}\|_\infty \le 1} \mathrm{Regret}_T(\sigma; \mathbf{Y}) \Bigg). \end{aligned}$$

The last right-hand side is equal to the variance of the unforecastable component ϵ_S plus C times the oracle risk on estimating the mean component, as well as $O(NT^{-1} \log T)$ regret. If the oracle risk

[4] We can take $\mu_t = E[y_{0t} \mid \mathbf{y}_t, \mathbf{Y}_{1:t-1}]$ whenever $S \perp\!\!\!\perp \mathbf{Y}$ under Q.

for estimating the mean component is small, then synthetic control is close to optimal, and its risk on estimating the mean component $E_Q[(\mu_S - \hat{\theta}_S' \mathbf{y}_S)^2]$ is also small.[5]

[5] Note that the bound

$$E_Q[(y_{0S} - \hat{\theta}_S' \mathbf{y}_t)^2] \le C \left(E_Q[\epsilon_S^2] + \min_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T} E_Q[(\mu_t - \theta' \mathbf{y}_t)^2] + \frac{1}{T} \sup_{\|\mathbf{Y}\|_\infty \le 1} \mathrm{Regret}_T(\sigma; \mathbf{Y}) \right)$$

is immediate and allows for $\mu_t = E[y_{0t} \mid \mathbf{Y}_{1:t-1}, \mathbf{y}_t] = 0$, yet the scaled idiosyncratic risk $CE_Q[\epsilon_S^2]$ may be large.

Appendix B

Appendix to Chapter 2

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

B.1.1 A simple regret rate lower bound: proof of Theorem 2.3.5

In this section, we prove Theorem 2.3.5, restated below.

Theorem 2.3.5. *Fix a set of valid hyperparameters $\mathcal{H} = (\sigma_\ell, \sigma_u, s_\ell, s_u, A_0, A_1, \alpha, \beta_0, p)$ for Assumptions 2.3.2 to 2.3.4. Let $\mathcal{P}(\mathcal{H}, \sigma_{1:n})$ be the set of distributions P_0 on support points $\sigma_{1:n}$ which satisfy (2.7) and Assumptions 2.3.2 to 2.3.4 corresponding to $\mathcal{H}$. For a given P_0, let $\theta_i^* = E_{P_0}[\theta_i \mid Y_i, \sigma_i]$ denote the oracle posterior means. Then there exists a constant $c_{\mathcal{H}} > 0$ such that the worst-case Bayes regret of any estimator exceeds $c_{\mathcal{H}} n^{-\frac{2p}{2p+1}}$:*

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

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

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

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

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

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

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

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

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

Then

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

We have just shown that

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

where the supremum is over m_0 satisfying Assumption 2.3.4, and the infimum is over all randomized estimators of $m_0(\sigma_1), \ldots, m_0(\sigma_n)$ with data (V_i, σ_i). Note that the squared error loss on the right-hand side takes expectation over the fixed design points $\sigma_1, \ldots, \sigma_n$.

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

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

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

Now, note that

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

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 $p \ge 1$ in Assumption 2.3.4.

Therefore,

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

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

Remark B.1.1. For ease of interpretation, Theorem 2.3.5 is stated in the expected regret version, which is slightly disconnected from the upper bound Theorem 2.3.3, which conditions on a high-probability event. Observe that Theorem 2.3.3 immediately implies the in-probability upper bound on Regret:

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

Using the in-probability version of the minimax lower bound for nonparametric regression in Theorem 2.3.5 then implies an analogous lower bound (See, for instance, Theorems 2.4 and 2.5 in Tsybakov, 2008). ■

B.1.2 Relating other decision objects to squared-error loss

Theorem 2.3.7. *Suppose* (2.4) *holds, but* (2.7) *may or may not hold. Let $\hat{\delta}_i$ be the plug-in decisions with any vector of estimates $\hat{\theta}_i$, not necessarily from* CLOSE-NPMLE. *We have the following inequalities* *on the expected regret corresponding to the decision rules $\hat{\delta}_i$:*

1. *For* UTILITY MAXIMIZATION BY SELECTION,

$$E[\text{UMRegret}_n] \le \left(E\left[\frac{1}{n} \sum_{i=1}^n (\hat{\theta}_i - \theta_i^*)^2 \right] \right)^{1/2}. \tag{2.22}$$

2. *For* TOP-m SELECTION,

$$E[\text{TopRegret}_n^{(m)}] \le 2\sqrt{\frac{n}{m}} \left(E\left[\frac{1}{n} \sum_{i=1}^n (\hat{\theta}_i - \theta_i^*)^2 \right] \right)^{1/2}. \tag{2.23}$$

Proof. 1. We compute

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

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

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

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

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

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

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

[1] 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 $E[\theta_i \mid U, Y_i, \sigma_i] = \theta_i^*$.

and hence, by law of iterated expectations,

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

Observe that this can be controlled by applying Theorem B.1.2, 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,

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

Divide through by m/n to obtain the result.

□

Proposition B.1.2. *Suppose $\sigma(\cdot)$ is a permutation such that $\hat{\theta}_{\sigma(n)} \geq \cdots \geq \hat{\theta}_{\sigma(1)}$. Then*

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

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

Proof. We compute

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

The last step follows from the observation that

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

The left-hand side is the sorted difference between θ_i^* and $\hat{\theta}_i$. This is smaller than the unsorted difference by an application of the rearrangement inequality.[2] □

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

B.1.3 Worst-case risk

Theorem 2.3.10. *Under (2.4) but not (2.7), assume the conditional distribution $\theta_i \mid \sigma_i$ has mean $m_0(\sigma_i)$ and variance $s_0^2(\sigma_i)$. Denote the set of distributions of $\theta_{1:n} \mid \sigma_{1:n}$ which obey these restrictions as $\mathcal{P}(m_0, s_0)$. Let $\hat{\theta}_{i,G_0^*,\eta_0}$ denote the posterior mean estimates with some prior P^* under the location-scale model $P^*\left(\theta_i \leq t \mid \sigma_i\right) = G_0^*\left(\frac{t-m_0(\sigma_i)}{s_0(\sigma_i)}\right)$, for some fixed G_0^* with zero mean and unit variance. Let $\bar{\rho} = \max_i s_0^2(\sigma_i)/\sigma_i^2 < \infty$ be the maximal conditional signal-to-noise ratio and assume that it is bounded. Then, for some $C_{\bar{\rho}} < \infty$ that solely depends on $\bar{\rho}$,*

$$\sup_{P_0 \in \mathcal{P}(m_0,s_0)} E_{P_0}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_{i,G_0^*,\eta_0} - \theta_i)^2\right] \leq C_{\bar{\rho}} \cdot \inf_{\hat{\theta}_{1:n}} \sup_{P_0 \in \mathcal{P}(m_0,s_0)} E_{P_0}\left[\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2\right]. \quad (2.24)$$

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

Proof. Note that

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

and

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

Thus,

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

Chen (2023) shows that

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

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

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

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 (2.13)). Hence, the Bayes risk under this P_0 lower bounds the minimax risk

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

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

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

Hence

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

□

B.1.4 Unbiased loss estimation

Proposition B.1.3. *Suppose (Y_i, σ_i) obey the Gaussian heteroskedastic location model, assumed to be independent across i (2.4). 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 2.1 are unbiased for the corresponding loss:*

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

Moreover, their conditional variances are equal to those expressions displayed in the third column of Table 2.1.

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

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

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

The conditional variance statement holds by definition.

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

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

   The conditional variance statement follows from

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

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

□

B.1.5 A discrete choice model

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

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

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

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

where $N_{j,-i}$ is the number of other individuals choosing item j. Since individuals other than i are symmetric to i, the expected utility (conditional on what i observes) is[3]

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

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

Since individuals are ex-ante symmetric, we assume that

$$\pi_{-ij} = \pi_j = P(E_i V_j \geq E_i V_k \quad \forall k).$$

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

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

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

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

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

Lemma B.1.4. $N_j \mid (\alpha_{1:n}, \beta_{1:n}) \sim \text{Pois}\left(\lambda(\alpha_j + \beta_j)\right)$ *independently across* j.

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

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

[3] Note that the externality that enters the utility is different from the externality in θ. This is for analytical tractability purposes.

To prevent the utility component from becoming negative, we additionally assume that $\alpha_j + \beta_j > -\rho$ almost surely, which imposes that $\rho > -\mu$.

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

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

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

We can interpret various empirical observations through this model:

- For hospital value-added (Chandra *et al.*, 2016), N_j positively selects on hospital quality α_j. This is likely true for most value-added settings.
- For teacher value-added, it is possible (Lazear, 2001; Barrett and Toma, 2013; Mehta, 2019) that teachers may prefer smaller classes, and school administrators may reward good teachers by letting them teach smaller classes. In the lens of this model, N_j negatively selects on quality.[4]
- In integenerational mobility, N_j is the number of poor minority households. Higher N_j leads to oppressive institutions and residential segregation. We can interpret these pernicious effects as a negative ρ.

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

- In unbalanced panel data settings, the length of the observed period for a unit—which relates to the precision of the unit's estimated fixed effect—may be correlated with the underlying fixed effect. This observation dates at least to Olley and Pakes (1996), who note that in a firm panel, those firms with shorter observed period are probably less productive and have to shut down sooner. For value-added modeling of nursing homes, Einav *et al.* (2022) note that patients with shorter stays at nursing homes typically experience an adverse health event, including death.

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

Such events are presumably more likely for worse nursing homes, again inducing a correlation between nursing home qualities and the sample sizes used to estimate them. Similarly, for teacher value-added, Bruhn *et al.* (2022) find that teachers who have shorter observed spells in administrative datasets tend to be worse and have noisier value added estimates.

B.1.6 Interpretation of empirical Bayes sampling model

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

Interpretation when independence of units fails

We consider the interpretation of the sampling model (2.4) when it is misspecified. Recall that we assume $(Y_i, \theta_i, \sigma_i)$ are sampled independently across i, with $Y_i \mid \theta_i, \sigma_i \sim \mathcal{N}(\theta_i, \sigma_i)$. This sampling model can fail in two ways. First, it is possible that $Y_{1:n} \mid \theta_{1:n}, \sigma_{1:n}$ are correlated but still multivariate Gaussian. Second, it is possible that (θ_i, σ_i) are correlated across i. Here, we limit our discussion to Decision Problem 1.

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

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

where $\operatorname{diag}(\Sigma) = [\sigma_1^2, \ldots, \sigma_n^2]$ and the variance-covariance matrix Σ is known. Let Q_0 be the joint distribution of $\boldsymbol{\theta} \mid \Sigma$. Now, the oracle Bayesian—who knows Q_0—would use $E_{Q_0}[\theta_i \mid \boldsymbol{Y}, \Sigma]$ as their decision rule. The empirical Bayesian can similarly emulate that oracle Bayes decision rule by estimating Q_0. If the empirical Bayesian is willing to assume that the location-scale assumption (2.7) describes Q_0:

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

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

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

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

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

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

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

Interpretation with additional covariates X_i

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

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

Here, we will discuss two questions. First, how do we handle covariates? Second, what is the difference between using X_i and σ_i—is the standard error simply a covariate?[5]

On the first question, there are two ways of incorporating covariates, under similar but distinct assumptions. First, CLOSE-methods can be extended to incorporate covariates by augmenting (2.7) to incorporate covariates. That is, we can instead assume that

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

and estimate m_0, s_0 nonparametrically. Instead of being one-dimensional nonparametric regression problems, they are now $(d+1)$-dimensional nonparametric problems. Under the same Hölder-type smoothness conditions, the corresponding regret rate replaces $n^{-\frac{2p}{2p+1}}$ with $n^{-\frac{2p}{2p+1+d}}$. Second, as we do in the empirical exercises, one could consider a strategy of residualizing against X_i in some arbitrary way, performing empirical Bayes, and undoing the residualization. This strategy dates back to Fay and Herriot (1979). That is, with raw data $\tilde{Y}_i$ for parameter ϑ_i, we can consider forming the residuals $Y_i = \tilde{Y}_i - b(X_i)$ and $\theta_i = Y_i - b(X_i)$, and perform empirical Bayes methods on $(Y_i, \theta_i, \sigma_i)$. At a high level, we can rationalize this strategy as mimicking a constrained oracle Bayesian who solely has access to Y_i, σ_i, who knows the joint distribution of (θ_i, σ_i), but who does not have access to X_i. Note that this interpretation is coherent regardless of the transformation $b(X_i)$, allowing us to be more blasé about modeling X_i than the previous approach. In particular, choosing $b(X_i) = 0$ ignores the covariate entirely; the resulting empirical Bayes procedure mimics an oracle that does not have access to X_i. Of course, when we impose the location-scale assumption (2.7) on (θ_i, σ_i), different $b(X_i)$ gives rise to different—and possibly mutually exclusive—underlying models on $(\vartheta_i, \sigma_i, X_i)$.

On the second question, in an operational sense, σ_i is simply another covariate. σ_i is not particularly special in the assumption (B.1), and one interpretation of CLOSE is treating σ_i precisely as a covariate to be regressed out. However, σ_i does occupy a special place in the statistical structure of the problem. The *likelihood* of the data, $Y_i \mid \theta_i, \sigma_i$, depends on σ_i but not X_i. This special role of σ_i means that we

[5] Covariates are considered in Ignatiadis and Wager (2019). They assume a homoskedastic setting where the prior depends on some covariates X_i: i.e., in our notation, $\theta_i \mid X_i \sim \mathcal{N}(m(X_i), s_0^2)$ and $Y_i \mid \theta_i \sim \mathcal{N}(\theta_i, \sigma^2)$. Starting from our setting (2.7), to obtain theirs, one would (i) restrict to homoskedasticity $\sigma_i = \sigma$, (ii) consider some covariates X_i that predict θ_i, and model $\theta_i \mid X_i$ as a conditional location—but not scale—family, and (iii) restrict $G_0 \sim \mathcal{N}(0, 1)$.

Their minimax lower bound on the regret uses essentially the same argument as we do in Theorem 2.3.5.

must treat it with more care so that the resulting procedure has a coherent interpretation. If we wanted to ignore covariates X_i, we can imagine an oracle Bayesian who does not have access to X_i, and the resulting empirical procedure simply mimics that constrained oracle. This line of reasoning does not work with σ_i, since any oracle Bayesian—constrained or otherwise—must have access to σ_i. As a result, we cannot avoid the problem of modeling $\theta_i \mid \sigma_i$ as easily as we could have avoided modeling $\theta_i \mid X_i, \sigma_i$ by changing the goalpost.

B.1.7 Alternatives to CLOSE

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 B.1.

Table B.1: Properties of alternative methods

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

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

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 $h^\star(T_i) = \frac{E[\sigma_i\theta_i|T_i]}{E[\sigma_i^2|T_i]}$. However, $h^\star$ is not the posterior mean of θ_i/σ_i given the t-ratio T_i, unless $\sigma_i^2 \perp\!\!\!\perp \theta_i/\sigma_i$. On the other hand, the loss function that does rationalize the posterior mean $h(T_i) = E[\theta_i/\sigma_i \mid T_i]$ is the precision-weighted compound loss $L(\boldsymbol{\delta}, \theta_{1:n}) = \frac{1}{n}\sum_{i=1}^n \sigma_i^{-2}(\delta_i - \theta_i)^2$.

Thus, rescaling posterior means on t-ratios achieves optimality for a weighted objective among a restricted class of decision rules $\delta_{i,\text{t-stat}}$. ■

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

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

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

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

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

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

Both papers work under Gaussian assumptions on the micro-data. This parametric assumption[6] 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

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

distributions:

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

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

This discussion is not to say that CLOSE is necessarily preferable to these alternatives. It highlights that the possible dependence between θ_i and σ_i cannot be easily resolved. As summarized in Table B.1, existing alternatives compromise on optimality, use a different loss function, or implicitly assume θ_i is independent from components of σ_i^2 (e.g., n_i). Of course, depending on the empirical context, these may well be reasonable features.

In contrast, our approach models $\theta_i \mid \sigma_i$ directly via the location-scale assumption (2.7). 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.

Alternative models for $\theta_i \mid \sigma_i$

One alternative is simply treating the joint distribution of (θ_i, σ_i) fully nonparametrically. For instance, an f-modeling approach with Tweedie's formula[7] implies that an estimate of the conditional distribu-

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

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

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

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

tion $Y_i \mid \sigma_i$ 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).[8] This has the advantage of estimating a smooth prior at the cost of having tuning parameters. We are not aware of regret results for this approach.

If we commit to making some substantive restriction on the joint distribution of (θ_i, σ_i), it is fair to ask why the conditional location-scale restriction (2.7) 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 (Section B.1.7 gives a heuristic justification for this strategy). If we wish to preserve the Gaussian location model structure on the transformed data, then effectively we can only consider affine transformations (i.e., $Z = a(\sigma) + b(\sigma)Y$) (shown in Theorem B.1.5 below). If we further wish that Z obeys a Gaussian location model in which prior independence holds (i.e., $\tau \equiv a(\sigma) + b(\sigma)\theta$ is independent from $\nu \equiv b(\sigma)\sigma$)—so that we can apply NPMLE-based approaches assuming prior independence—then we have no other choice but to assume (2.7). 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 2.3.3.

Lemma B.1.5. *Let $Y \sim \mathcal{N}(\theta, \sigma^2)$ with known σ^2. Consider a strictly increasing and differentiable function $g(\cdot)$. Let $Z = h(Y)$. Then the corresponding family of distributions of Z is a natural exponential family if and only if $h(Y) = a + bY$.*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Model-free interpretation of CLOSE-NPMLE

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

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

and the distribution implied by the convolution $H \star \mathcal{N}(0, \nu_i^2)$:

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

where $f_{H,\nu}$ is the density of the convolution $H \star \mathcal{N}(0, \nu^2)$. As shown by Jiang and Zhang (2009) and Jiang (2020) (see Section B.3.3), the regret in mean-squared error under a misspecified prior $\tau_i \sim H$ is upper bounded by the average squared Hellinger distance between the distribution of the data and the distribution implied by H. The average Hellinger distance is further upper bounded by the average KL divergence:

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

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

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

B.2 Additional empirical exercises

B.2.1 Positivity of $s_0(\cdot)$ in the Opportunity Atlas data

In the Opportunity Atlas data, we often observe that the estimated conditional variance is negative: $\hat{s}_0^2 < 0$. To test if this is due to sampling variation or underdispersion of the Opportunity Atlas estimates relative to the estimated standard error, we consider the following upward-biased estimator of $s_0^2(\sigma_i)$. Without loss, let us sort the Y_i, σ_i by σ_i, where $\sigma_1 \le \cdots \le \sigma_n$. Let

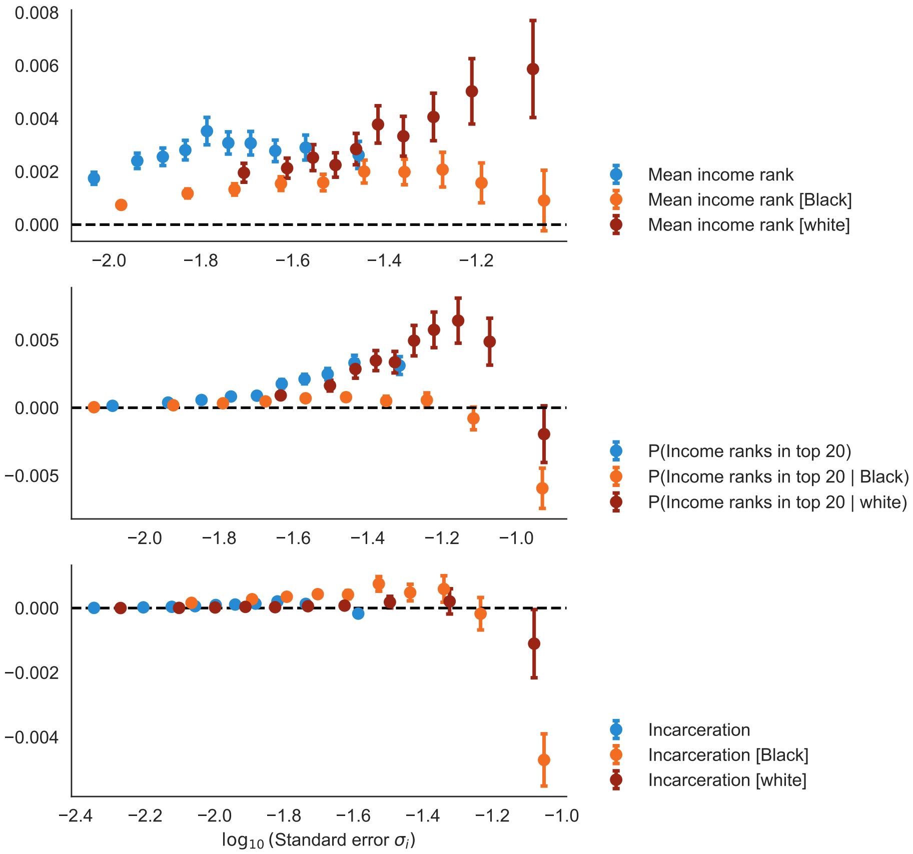
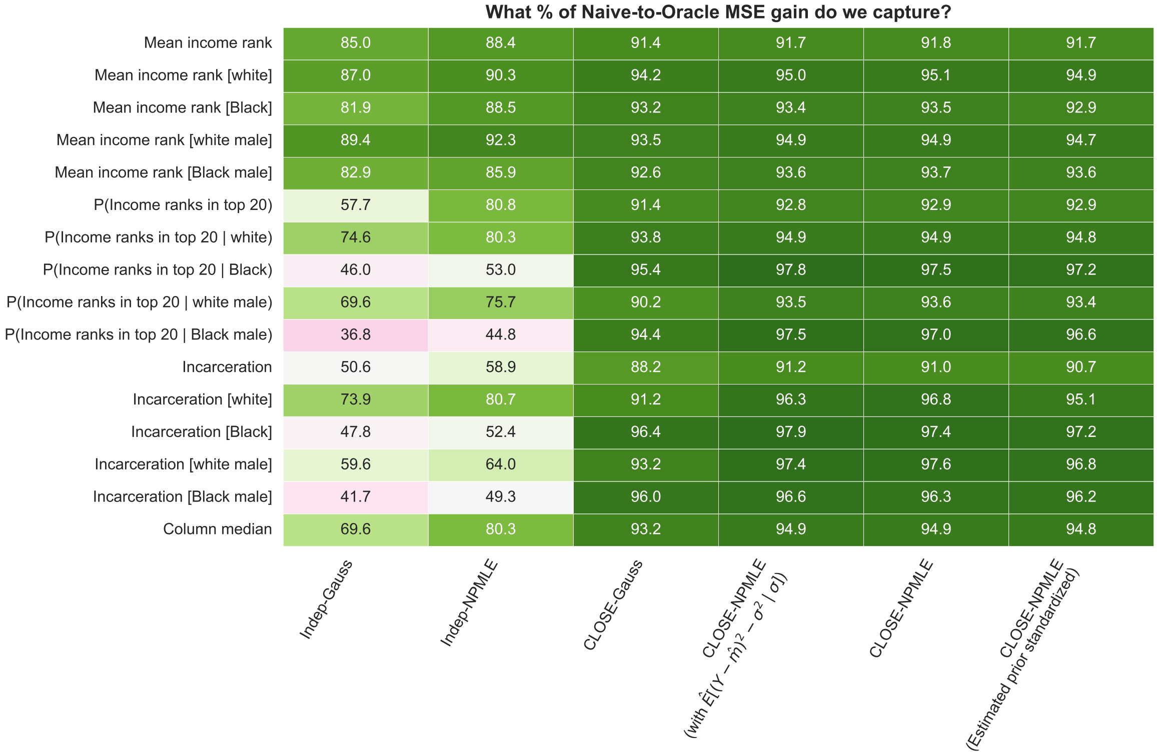


Figure B.1: Estimated conditional variance $s_0^2(\sigma)$, binned into deciles, with 95% uniform confidence intervals shown.

$S_i = \frac{1}{2}\left[(Y_{i+1} - Y_i)^2 - (\sigma_i^2 + \sigma_{i+1}^2)\right]$. Note that

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

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



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

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

Figure B.2: Additional CLOSE-NPMLE variants for the calibrated simulation in Section 2.5. Here the results average over 100 replications.

B.2.2 Robustness checks for the calibration exercise in Section 2.5

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

B.2.3 Simulation exercise setup

This section describes the details of the simulation exercise in Section 2.5. We restrict to the 10,109 tracts within the twenty largest Commuting Zones. Tracts with missing information are dropped for each measure of mobility. Specifically, the simulated data-generating process is as follows:

(Sim-1) Residualize $\tilde{Y}_i$ against some covariates X_i to obtain β and residuals Y_i. Estimate the conditional moments m_0, s_0 on (Y_i, σ_i) via local linear regression, described in Section B.7.

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

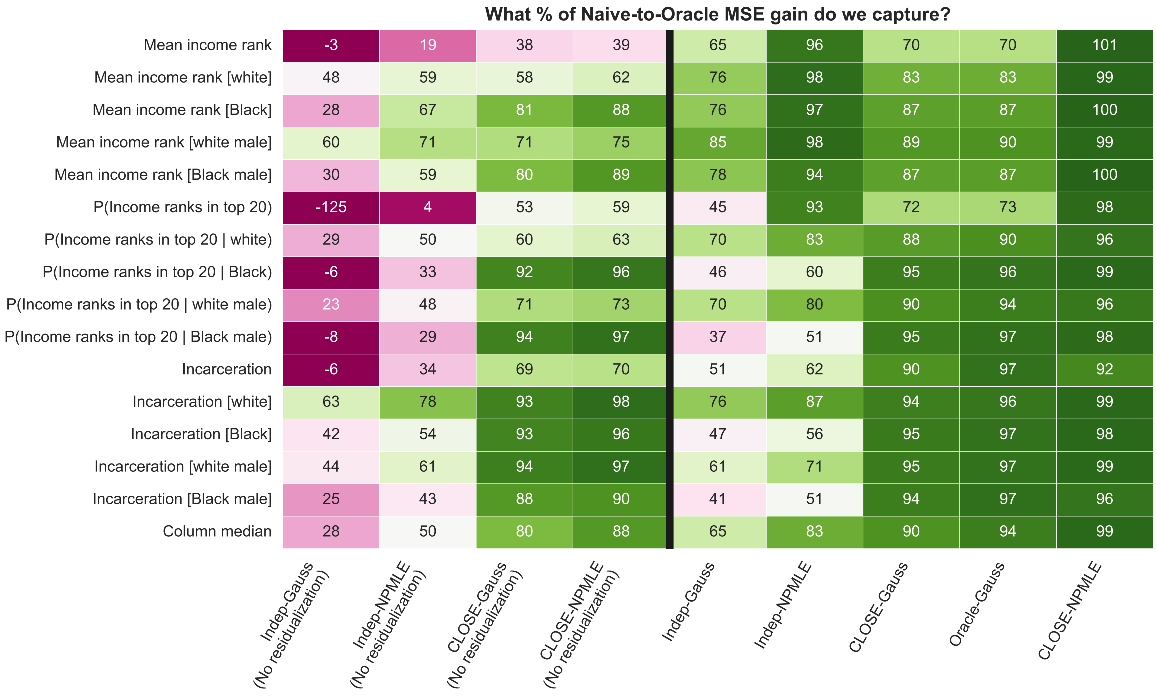
(Sim-3) Let $\vartheta_i^* = s_0(\sigma_i)\tau_i^* + m_0(\sigma_i) + \beta' X_i$ and let $\tilde{Y}_i^* \mid \theta_i^*, \sigma_i \sim \mathcal{N}(\theta_i^*, \sigma_i^2)$.

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

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

B.2.4 Different Monte Carlo setup

We have also conducted a Monte Carlo exercise where we replace (Sim-2) with the following step:


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	-3	19	38	39	65	96	70	70	101
Mean income rank [white]	48	59	58	62	76	98	83	83	99
Mean income rank [Black]	28	67	81	88	76	97	87	87	100
Mean income rank [white male]	60	71	71	75	85	98	89	90	99
Mean income rank [Black male]	30	59	80	89	78	94	87	87	100
P(Income ranks in top 20)	-125	4	53	59	45	93	72	73	98
P(Income ranks in top 20 | white)	29	50	60	63	70	83	88	90	96
P(Income ranks in top 20 | Black)	-6	33	92	96	46	60	95	96	99
P(Income ranks in top 20 | white male)	23	48	71	73	70	80	90	94	96
P(Income ranks in top 20 | Black male)	-8	29	94	97	37	51	95	97	98
Incarceration	-6	34	69	70	51	62	90	97	92
Incarceration [white]	63	78	93	98	76	87	94	96	99
Incarceration [Black]	42	54	93	96	47	56	95	97	98
Incarceration [white male]	44	61	94	97	61	71	95	97	99
Incarceration [Black male]	25	43	88	90	41	51	94	97	96
Column median	28	50	80	88	65	83	90	94	99

Figure B.3: Analogue of Figure 2.4 for the data-generating process in Section B.2.4. Here the results average over 100 replications.

- For each σ_i, let
$$\alpha_i = \frac{1}{2} + \frac{1}{2}\frac{m_0(\sigma_i) - \min_i m_0(\sigma_i)}{\max_i m_0(\sigma_i) - \min_i(\sigma_i)} \in [1/2, 1]$$
We sample $\tau_i^* \mid \sigma_i$ as a scaled and shifted Weibull distribution with shape α_i. The scaling and translation ensures that $\tau_i \mid \sigma_i$ has mean zero and variance one. Because we choose the Weibull distribution, the shape parameter α_i corresponds exactly to α in Assumption 2.3.2. Our choices of α_i implies that $\tau_i \mid \sigma_i$ has thicker tails than exponential and does not have a moment-generating function.

The Weibull distribution has thicker tails and is skewed, and as a result, NPMLE-based methods tend to greatly outperform methods based on assuming Gaussian priors. Figure B.3 show the analogue of Figure 2.4 for this data-generating process. Indeed, we see that INDEPENDENT-NPMLE improves over INDEPENDENT-GAUSS considerably, and similarly for CLOSE-NPMLE and ORACLE-GAUSS.

B.2.5 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 2.5. Since this is an empirical application on real, rather than synthetic, data, we no longer have access to oracle estimators. As a result, for the relative MSE performance, we normalize by a different benchmark. We can think of the performance gain of INDEPENDENT-GAUSS over NAIVE as the value of doing basic, standard empirical Bayes shrinkage. We normalize each method's estimated MSE improvement against NAIVE as a multiple of this "value of basic empirical Bayes." Figure B.4(a) shows the resulting relative performance. Since our notion of relative performance has changed, we use a different color scheme. A value of 1 means that a method does exactly as well as INDEPENDENT-GAUSS, and a value of 2 means that, relative to NAIVE, a method doubles the gain of basic empirical Bayes. Performance on a non-relative scale is shown in Figure B.4(b).

We find that our empirical patterns from the calibrated simulation Figure 2.4 mostly persists on real data. In particular, INDEPENDENT-NPMLE offers small improvements over INDEPENDENT-GAUSS. Nevertheless, CLOSE-NPMLE continues to dominate other methods. Across the definitions of ϑ_i, CLOSE-NPMLE generates a median of 180% the value of basic empirical Bayes. That is, on mean-squared error, moving from INDEPENDENT-GAUSS to CLOSE-NPMLE is about half as valuable as moving from NAIVE to INDEPENDENT-GAUSS. For our running example (TOP-20 PROBABILITY for Black individuals), moving from INDEPENDENT-GAUSS to CLOSE-NPMLE is more valuable than moving from NAIVE to INDEPENDENT-GAUSS. If practitioners find using the standard empirical Bayes method to be a worthwhile investment over using the raw estimates directly, then they may find using CLOSE-NPMLE over INDEPENDENT-GAUSS to be a similarly worthwhile investment.

B.2.6 Empirical Bayes pooling over all Commuting Zones in validation exercise

Here, we repeat the exercise in Figure 2.5, 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 2.5 in this setting, shown in Figure B.5. The results are extremely similar.

Separately, we consider the version of this exercise without covariates in Figure B.6. We see that covariates are extremely important for the performance of INDEPENDENT-GAUSS, as it frequently underperforms NAIVE without covariates.[9] By comparison, they are less important for the performance of CLOSE-NPMLE, as σ_i contains a lot of the signal in the tract-level covariates.

B.2.7 The tradeoff between accurate targeting and estimation precision

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

We investigate this question via coupled bootstrap in the Bergman *et al.* (2023) exercise. In particular, we let the subscript b (resp. w) denote quantities for Black (resp. white) individuals. We assume that $Y_{ib} \perp\!\!\!\perp Y_{iw} \mid \theta_{ib}, \theta_{iw}$. For each tract, we construct $\pi_i = n_{ib}/n_i$, where n_i (resp. n_i) is the number of (resp. Black) children under 18 living in the given tract with parents whose household income was below the national median.[10] Let $\theta_i = \pi_i\theta_{ib} + (1 - \pi_i)\theta_{iw}$ be a pooled measure, where

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

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

Here, we investigate whether screening tracts based on posterior mean estimates for θ_{iw} or θ_i generates better decisions in terms of θ_{ib}, owing to the precision in Y_{iw} and Y_i. Figure B.7 shows estimated performances of different empirical Bayes methods by different proxy variables that the screening targets. For each measure of economic mobility for Black individuals, dots on the thick

[9] This is in part since our implementation of INDEPENDENT-GAUSS uses weighted means for estimating the prior parameters, worsening the misspecification. See Footnote 55.

[10] 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.

black dashed line correspond to screening on the corresponding θ_{ib}. Dots on the red (resp. blue) dashed line correspond to screening on θ_{iw} (resp. θ_i). We see that for all three measures of economic mobility, using CLOSE-NPMLE to screen on the original parameter θ_{ib} performs best. In other words, the benefits of higher precision are insufficient to offset inaccurate targeting.

B.3 Regret control proofs: Setup, assumptions, and notation

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

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

and $(Y_i, \theta_i, \sigma_i)$ are mutually independent. Assume that the joint distribution for (θ_i, σ_i) takes the location-scale form (2.7)

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

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.

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

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

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

$\tau \sim G$. As a shorthand, we write

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

Let the average squared Hellinger distance be

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

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

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

Define

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

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

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

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

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

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

Explicitly, by Tweedie's formula,

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

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

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

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

$$\hat{\theta}_{i,G,\eta,\rho} = Y_i + s_i \hat{\nu}_i^2(\eta) \frac{f'_{G,\hat{\nu}_i(\eta)}(\hat{Z}_i(\eta))}{f_{G,\hat{\nu}_i(\eta)}(\hat{Z}_i(\eta)) \vee \frac{\rho}{\hat{\nu}_i(\eta)}} \tag{B.3}$$

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

Lastly, we will also define

$$\varphi_+(\rho) = \varphi^{-1}(\rho) = \sqrt{\log \frac{1}{2\pi\rho^2}} \quad \rho \in (0, (2\pi)^{-1/2}) \tag{B.4}$$

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

B.3.1 Assumptions

Recall the assumptions we stated in the main text.

Assumption 2.3.1. *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* (2.12)*, ignoring a constant factor* $1/\hat{\nu}_i$*. We assume that* $\hat{G}_n$ *satisfies*

$$\frac{1}{n}\sum_{i=1}^n \psi_i(Z_i, \hat{\eta}, \hat{G}_n) \geq \sup_{H \in \mathcal{P}(\mathbb{R})} \frac{1}{n}\sum_{i=1}^n \psi_i(Z_i, \hat{\eta}, H) - \kappa_n \tag{2.17}$$

for tolerance κ_n

$$\kappa_n = \frac{2}{n}\log\left(\frac{n}{\sqrt{2\pi e}}\right). \tag{2.18}$$

Moreover, we require that $\hat{G}_n$ *has support points within* $[\min_i \hat{Z}_i, \max_i \hat{Z}_i]$*. To ensure that* κ_n *is positive, we assume that* $n \geq 7 = \lceil \sqrt{2\pi e} \rceil$*.*[11]

Assumption 2.3.2. *The distribution* G_0 *is has zero mean, unit variance, and admits simultaneous moment control with parameter* $\alpha \in (0, 2]$*: There exists a constant* $A_0 > 0$ *such that for all* $p > 0$*,*

$$(E_{\tau \sim G_0}[|\tau|^p])^{1/p} \leq A_0 p^{1/\alpha}. \tag{2.19}$$

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

Assumption 2.3.3. *The variances* $(\sigma_{1:n}, s_0)$ *admit lower and upper bounds:*

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

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

Assumption 2.3.4. *Let* $C^p_{A_1}([\sigma_\ell, \sigma_u])$ *be the Hölder class of order* $p \geq 1$ *with maximal Hölder norm* $A_1 > 0$ *supported on* $[\sigma_\ell, \sigma_u]$*.*[12] *We assume that*

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

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

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

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

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

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

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

4. *The conditional variance estimator respects the conditional variance bounds in Assumption 2.3.3:* $P\left(\frac{s_{0\ell}}{2} < \hat{s} < 2 s_{0u}\right) = 1$*.*

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

Definition B.3.1. For some set $\mathcal{X} \subset \mathbb{R}$ and constant $A > 0, p > 0$, let $C^p_A(\mathcal{X})$ be the set of continuous functions $f : \mathcal{X} \to \mathbb{R}$ with $\|f\|_{(p)} \leq A$. The norm $\|\cdot\|_{(p)}$ is defined as follows. Let $\underline{p}$ be the greatest integer strictly smaller than p. Define

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

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

B.3.2 Regret control: result statement

Define the regret as the difference between the mean-squared error of some feasible posterior means $\hat{\theta}_{i,G,\eta}$ against the mean-squared error of the oracle posterior means

$$\begin{aligned}\operatorname{Regret}(G,\eta) &= \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_{i,G,\eta}-\theta_i)^2 - \frac{1}{n}\sum_{i=1}^{n}(\theta_i^*-\theta_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_{i,G,\eta}-\theta_i^*)^2 + \frac{2}{n}\sum_{i=1}^{n}(\theta_i^*-\theta_i)(\hat{\theta}_{i,G,\eta}-\theta_i^*) \end{aligned} \tag{B.5}$$

(B.5) decomposes the MSE regret into a mean term that equals the mean-squared distance between the feasible posterior means and the oracle posterior means, as well as a term that is mean zero conditional on the data $Y_1,\ldots,Y_n$, since $\theta_i^*-\theta_i$ represents irreducible noise.

Fix sequences $\Delta_n > 0$ and $M_n > 0$. Define the following "good" event which we use in Theorem B.6.2:

$$A_n = \left\{\|\hat{\eta}-\eta\|_\infty \equiv \max(\|\hat{m}-m_0\|_\infty, \|\hat{s}-s_0\|_\infty) \le \Delta_n,\ \bar{Z}_n \equiv \max_{i\in[n]}(|Z_i| \vee 1) \le M_n\right\}. \tag{B.6}$$

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

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

where $\mathbf{A}_n$ is the event in (2.20).

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

Theorem 2.3.3. *Assume Assumptions 2.3.1 to 2.3.4 hold. Then, for any $\delta \in (0, \frac{1}{2})$, there exists universal constants $C_{1,\mathcal{H},\delta} > 0$ and $C_{0,\mathcal{H},\delta} > 0$ such that (i) $P(\mathbf{A}_n(C_{1,\mathcal{H},\delta})) \ge 1-\delta$ and that (ii) the expected regret conditional on $\mathbf{A}_n(C_{1,\mathcal{H},\delta})$ is dominated by the rate function*

$$E\left[\operatorname{Regret}(\hat{G}_n,\hat{\eta}) \mid \mathbf{A}_n(C_{1,\mathcal{H},\delta})\right] \le C_{0,\mathcal{H},\delta} n^{-\frac{2p}{2p+1}} (\log n)^{\frac{2+\alpha}{\alpha}+3+2\beta_0}. \tag{2.21}$$

Proof. Immediately by Assumption 2.3.4(2–3), we can choose $C_{1,\mathcal{H}}$ so that $P(\mathbf{A}_n(C_{1,\mathcal{H}})) \ge 1-\delta$. Let $\Delta_n = C_{1,\mathcal{H}} n^{-\frac{p}{2p+1}}(\log n)^{\beta_0}$ and $M_n = C(\log n)^{1/\alpha}$ for some C to be chosen. Both $C_{1,\mathcal{H}}$ and C

may depend on δ. Moreover, we can decompose

$$\begin{aligned}
&E\left[\text{Regret}(\hat{G}_n,\hat{\eta}) \mid \mathbf{A}_n(C_{1,\mathcal{H}})\right] \\
&\le \frac{1}{1-\delta}\left\{E\left[\text{Regret}(\hat{G}_n,\hat{\eta})\mathbb{1}(A_n)\right] + E\left[\text{Regret}(\hat{G}_n,\hat{\eta})\,\underbrace{\mathbb{1}(\mathbf{A}_n(C_{1,\mathcal{H}}),\bar{Z}_n > M_n)}_{\mathbf{A}_n\setminus A_n}\right]\right\} \\
&\lesssim_{\mathcal{H}} n^{-\frac{2p}{2p+1}}(\log n)^{\frac{2+\alpha}{\alpha}+3+2\beta_0} + \frac{1}{n}(\log n)^{2/\alpha} \qquad \text{(Theorems B.6.1 and B.6.2)} \\
&\lesssim_{\mathcal{H}} n^{-\frac{2p}{2p+1}}(\log n)^{\frac{2+\alpha}{\alpha}+3+2\beta_0}
\end{aligned}$$

Note that the application of Theorems B.6.1 and B.6.2 implicitly picks some constant for $M_n = C(\log n)^{1/\alpha}$. This concludes the proof. □

Corollary B.3.2. *Assume the same setting as Theorem 2.3.3. Suppose, additionally, for all sufficiently large $C_{1,\mathcal{H}} > 0$, $P(\mathbf{A}_n(C_{1,\mathcal{H}})) \ge 1 - n^{-2}$. Then, there exists a constant $C_{0,\mathcal{H}} > 0$ such that the expected regret is dominated by the rate function*

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

Proof. Let Δ_n, M_n as in the proof of Theorem 2.3.3. Decompose

$$\begin{aligned}
E[\text{Regret}(\hat{G}_n,\hat{\eta})] &= E[\text{Regret}(\hat{G}_n,\hat{\eta})\mathbb{1}(A_n)] + E[\text{Regret}(\hat{G}_n,\hat{\eta})\mathbb{1}(A_n^{\mathrm{C}})] \\
&= E[\text{Regret}(\hat{G}_n,\hat{\eta})\mathbb{1}(A_n)] + E[\text{Regret}(\hat{G}_n,\hat{\eta})\mathbb{1}(\mathbf{A}_n^{\mathrm{C}} \cup \{Z_n > M_n\})] \\
&\le E[\text{Regret}(\hat{G}_n,\hat{\eta})\mathbb{1}(A_n)] + E[\text{Regret}(\hat{G}_n,\hat{\eta})\mathbb{1}(\mathbf{A}_n^{\mathrm{C}})] \\
&\quad + E[\text{Regret}(\hat{G}_n,\hat{\eta})\mathbb{1}(\bar{Z}_n > M_n)] \\
&\lesssim_{\mathcal{H}} n^{-\frac{2p}{2p+1}}(\log n)^{\frac{2+\alpha}{\alpha}+3+2\beta_0} + \frac{2}{n}(\log n)^{2/\alpha} \qquad \text{(Theorems B.6.1 and B.6.2)} \\
&\lesssim_{\mathcal{H}} n^{-\frac{2p}{2p+1}}(\log n)^{\frac{2+\alpha}{\alpha}+3+2\beta_0},
\end{aligned}$$

where our application of Theorem B.6.1 uses the assumption that $P(\mathbf{A}_n(C_{1,\mathcal{H}})^{\mathrm{C}}) = \mathbb{1}(\|\hat{\eta}-\eta\|_\infty > \Delta_n) \le \frac{1}{n^2}$. □

Remark B.3.3 (Relaxing Assumption 2.3.4(4))**.** Note that the event $\mathbf{A}_n(C)$ implies $s_{0\ell}/2 \le \hat{s} \le 2s_{0u}$ for all sufficiently large $n > N_{C,s_{0\ell},s_{0u},p,\beta_0}$. Since we condition on $\mathbf{A}_n(C)$ in Theorem 2.3.3, we can drop Assumption 2.3.4(3) by only requiring (2.21) to hold for all sufficiently large n. This is a minor modification since Theorem 2.3.3 is an upper bound on the convergence rate. On the other hand, dropping Assumption 2.3.4(4) does affect regret control on the event $\mathbf{A}_n^C(C_1)$ below. Our truncation rule for $\hat{s}(\cdot)$ in Section B.7 ensures that $\hat{s}(\cdot) \geq \frac{c}{n}$. We show in Section B.7 that this is sufficient for the conclusion of Theorem 2.3.4.[13] ■

B.3.3 Regret control: proof ideas

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

$$E\left[\operatorname{Regret}^\tau(\hat{G}_n, \eta_0)\right] \equiv E\left[\frac{1}{n}\sum_{i=1}^n (\hat{\tau}_{i,\hat{G}_n^*,\eta_0} - \tau_i^*)^2\right] \tag{B.7}$$

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

They do so by showing that, loosely speaking,

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

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

This is because distributions G that achieve high likelihood—which G_n^* does by construction—tend to have low average squared Hellinger distance with respect to G_0 (Theorem 6 in Soloff *et al.* (2021)). Roughly speaking, the rate function is linked to likelihood suboptimality (B.2):

$$\delta_n^2 \asymp \max\left(\operatorname{Sub}_n(\hat{G}_n^*), \frac{1}{n}(\log n)^{\frac{2\alpha}{2+\alpha}+1}\right). \tag{B.8}$$

(ii) For a *fixed* distribution G, the deviation from oracle between the *regularized* posterior means

[13] This lower bound on $\hat{s}$ also adds enough regularity to avoid writing "sufficiently large n" for the statement analogous to Theorem 2.3.3 as well. See Section B.7 for details.

(B.3) is bounded by the average squared Hellinger distance:

$$E[(\hat{\tau}_{i,G,\eta_0,\rho_n} - \tau^*_{i,\rho_n})^2] \lesssim (\log(1/\rho_n))^3 \bar{h}^2(f_{G,\cdot}, f_{G_0,\cdot}). \tag{B.9}$$

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

(iii) Additional arguments can handle the difference between (B.9) and (B.7).

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

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

- Section B.4 (Theorems B.4.1 and B.4.2) establishes that $\hat{G}_n$, estimated off $(\hat{Z}_i, \hat{\nu}_i)$, achieves high likelihood (i.e., low $\mathrm{Sub}_n(\hat{G}_n)$) on the data (Z_i, ν_i), with high probability. This is an oracle inequality in the sense that it bounds the performance degredation of $\hat{G}_n$ relative to a setting where η_0 is known.

- Section B.5 (Theorems B.5.1 and B.5.3) establishes that $\hat{G}_n$, with high probability, achieves low Hellinger distance. This is a result of independent interest, as it characterizes the quality of $f_{\hat{G}_n,\nu_i}$ as an estimate of the true density f_{G_0,ν_i}.

- Section B.6 (Theorem B.6.2) establishes that the regret of $\hat{\theta}_{i,\hat{G}_n,\hat{\eta}}$ is low, using the argument controlling (B.7).

Intuition for Section B.4

The argument in Section B.4 is our most novel theoretical contribution. Note that, by (B.8), to obtain a rate of the form $\delta_n^2 = n^{-\frac{2p}{2p+1}}(\log n)^\gamma$,[14] we would require that $\mathrm{Sub}_n(\hat{G}_n) \lesssim n^{-\frac{2p}{2p+1}}(\log n)^\gamma$. However, such a rate is not immediately attainable. To see this, note that a direct Taylor expansion in η

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

of the log-likelihood yields

$$\begin{aligned}
&\frac{1}{n}\sum_i \psi_i(Z_i, \hat{\eta}, \hat{G}_n) - \frac{1}{n}\sum_i \psi_i(Z_i, \eta_0, \hat{G}_n) \\
&\approx \frac{1}{n}\sum_i \left(\frac{\partial \psi_i}{\partial \eta_i}\right)'(\eta_i - \eta_{0i}) + \frac{1}{2n}\sum_i (\eta_i - \eta_{0i})' \frac{\partial^2 \psi_i}{\partial \eta_i^2}(\eta_i - \eta_{0i}). \qquad \text{(B.10)} \\
&\lessapprox (\log n)^\gamma \left\{ \frac{1}{n}\sum_i \frac{\partial \psi_i}{\partial \eta_i} O\left(n^{-\frac{p}{2p+1}}\right) + n^{-\frac{2p}{2p+1}} \sum_i \left\| \frac{\partial^2 \psi_i}{\partial \eta_i^2} \right\| \right\}
\end{aligned}$$

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

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

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

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

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

To summarize, through our calculation, the rate we obtain (Theorem B.4.2, (B.14)) for $\mathrm{Sub}_n(\hat{G}_n)$ is

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

A more detailed breakdown is presented in Section B.4.2.

Intuition for Section B.5

Since the rate for $\mathrm{Sub}_n(\hat{G}_n)$ from Section B.4 itself includes $\bar{h}$, it is necessary to adapt the argument in the literature on Hellinger rate control (See, e.g., Theorem 4 in Jiang, 2020).

Our argument proceeds by observing that, with high probability,

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

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

$$
\begin{aligned}
&P\left[C\lambda_n^{1-2^{-k}} \le \bar{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}) \le C\lambda_n^{1-2^{-k+1}}, \operatorname{Sub}_n(\hat{G}_n) \lesssim \gamma_n^2 + \bar{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot})\lambda_n\right] \\
&\le P\left[C\lambda_n^{1-2^{-k}} \le \bar{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}), \operatorname{Sub}_n(\hat{G}_n) \lesssim \gamma_n^2 + \lambda_n^{1-2^{-k+1}}\lambda_n\right]
\end{aligned} \tag{B.11}
$$

using standard arguments in the literature. This is now feasible since the event (B.11) comes with an upper bound for $\bar{h}$. Thus, by a union bound,

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

We can choose $K \to \infty$ appropriately slowly so as to obtain $\bar{h}^2 \lesssim \delta_n^2$ with high probability.

Intuition for Section B.6

All that is remaining before we can use the bound (B.7) directly is dealing with the difference between $\hat{\theta}_{i,\hat{G}_n,\hat{\eta}}$ and $\tau_{i,\hat{G}_n,\eta_0}$. In Section B.6.3, we can use a Taylor expansion to control the distance

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

Doing so requires bounding the second derivatives of ψ_i, which are posterior moments under $\hat{G}_n$ (Section B.4.10), and hence bounded due to assuming that $\hat{G}_n$ has supported bounded within the range of the data $\hat{Z}_i$ (Theorem B.4.18). We then immediately find that

$$
\left|\hat{\theta}_{i,\hat{G}_n,\eta_0} - \theta^*_{i,G_0,\eta_0}\right|
$$

is proportional to the difference in τ-space. Therefore, the existing argument for (B.7) controls the regret.

B.4 Regret control proofs: An oracle inequality for the likelihood

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

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

for some rate function ϵ_n. It is convenient to state a set of high-level assumptions on the rates Δ_n, M_n. These are satisfied for $\Delta_n \asymp n^{-p/(2p+1)}(\log n)^\beta$, $M_n \asymp (\log n)^{1/\alpha}$.

Assumption B.4.1. *Assume that*

1. $\frac{1}{\sqrt{n}} \lesssim_{\mathcal{H}} \Delta_n \lesssim_{\mathcal{H}} \frac{1}{M_n^3} \lesssim_{\mathcal{H}} 1$
2. $\sqrt{\log n} \lesssim_{\mathcal{H}} M_n$

Note that there exists ρ_n by Theorem B.4.13 that lower bounds the density $f_{\hat{G}_n,\nu_i}(z)$ for all Z_i. Then our main result is an oracle inequality.

Theorem B.4.1. *Let* $\|\hat{\eta} - \eta\|_\infty = \max(\|\hat{m} - m_0\|_\infty, \|\hat{s} - s_0\|_\infty)$ *and* $\bar{Z}_n = \max_{i\in[n]} |Z_i| \vee 1$. *Suppose* $\hat{G}_n$ *satisfies Assumption 2.3.1. Under Assumptions 2.3.2 to 2.3.4* and *B.4.1, there exists constants* $C_{1,\mathcal{H}}, C_{2,\mathcal{H}} > 0$ *such that the following tail bound holds: Let*

$$\epsilon_n = M_n\sqrt{\log n}\Delta_n \frac{1}{n}\sum_{i=1}^n h\left(f_{\hat{G}_n,\nu_i}, f_{G_0,\nu_i}\right) + \Delta_n M_n \sqrt{\log n} e^{-C_{2,\mathcal{H}}M_n^\alpha} + \Delta_n^2 M_n^2 \log n + M_n^2 \frac{\Delta_n^{1-\frac{1}{2p}}}{\sqrt{n}}. \tag{B.12}$$

Then,

$$P\left[\bar{Z}_n \le M_n, \|\hat{\eta} - \eta\|_\infty \le \Delta_n, \mathrm{Sub}_n(\hat{G}_n) > C_{1,\mathcal{H}}\epsilon_n\right] \le \frac{9}{n}.$$

The following corollary plugs in some concrete rates for Δ_n, M_n and verifies that they satisfy Assumption B.4.1.

Corollary B.4.2. *For* $\beta \ge 0$, *suppose*

$$\Delta_n = C_{\mathcal{H}} n^{-\frac{p}{2p+1}}(\log n)^\beta \text{ and } M_n = (C_{\mathcal{H}} + 1)(C_{2,\mathcal{H}}^{-1}\log n)^{1/\alpha}. \tag{B.13}$$

Then there exists a $C_{\mathcal{H}}^*$ *such that the following tail bound holds. Suppose* $\hat{G}_n$ *satisfies Assumption 2.3.1. Under Assumptions 2.3.2* to *2.3.4, define* ε_n *as:*

$$\varepsilon_n = n^{-\frac{p}{2p+1}}(\log n)^{\frac{2+\alpha}{2\alpha}+\beta}\bar{h}\left(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}\right) + n^{-\frac{2p}{2p+1}}(\log n)^{\frac{2+\alpha}{\alpha}+2\beta}, \tag{B.14}$$

we have that,

$$P\left[\bar{Z}_n \le M_n, \|\hat{\eta} - \eta\|_\infty \le \Delta_n, \mathrm{Sub}_n(\hat{G}_n) > C_{\mathcal{H}}^*\varepsilon_n\right] \le \frac{9}{n}.$$

The constant $C_{\mathcal{H}}$ in Δ_n, M_n affects the conclusion of the statement only through affecting the constant $C^*_{\mathcal{H}}$.

B.4.1 Proof of Theorem B.4.2

We first show that the specification of Δ_n and M_n means that the requirements of Assumption B.4.1 are satisfied. Among the requirements of Assumption B.4.1:

1. is satisfied since the polynomial part of Δ_n converges to zero slower than $n^{-1/2}$, but converges to zero faster than any logarithmic rate. M_n is a logarithmic rate.

2. is satisfied since $\alpha \leq 2$.

We also observe that by Jensen's inequality,

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

and so we can replace the corresponding factor in ϵ_n by $\bar{h}$. Now, we plug the rates Δ_n, M_n into ϵ_n. We find that the term

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

since $\log n > 1$ as $n > \sqrt{2\pi e}$ by Assumption 2.3.1. Plugging in the rates for the other terms, we find that

$$\epsilon_n \lesssim_{\mathcal{H}} \varepsilon_n.$$

Therefore, Theorem B.4.2 follows from Theorem B.4.1.

B.4.2 Proof of Theorem B.4.1

Decomposition of $\mathrm{Sub}_n(\hat{G}_n)$

Observe that, by definition of $\hat{G}_n$ in (2.17),

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

For random variables a_n, b_n such that almost surely

$$\left|\frac{1}{n}\sum_{i=1}^n \psi_i(Z_i, \hat\eta, \hat G_n) - \psi_i(Z_i, \eta_0, \hat G_n)\right| \le a_n$$
$$\left|\frac{1}{n}\sum_{i=1}^n \psi_i(Z_i, \hat\eta, G_0) - \psi_i(Z_i, \eta_0, G_0)\right| \le b_n$$

we have

$$\frac{1}{n}\sum_{i=1}^n \psi_i(Z_i, \eta_0, \hat G_n) - \frac{1}{n}\sum_{i=1}^n \psi_i(Z_i, \eta_0, G_0) \ge -a_n - b_n - \kappa_n$$

and

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

Therefore, it suffices to show large deviation results for a_n and b_n.

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 (B.6). Since $\psi_i(Z_i, \eta, G)$ is smooth in $(m_i, s_i) \in \mathbb{R} \times \mathbb{R}_{>0}$, we can take a second-order Taylor expansion:

$$\psi_i\left(Z_i, \hat\eta, \hat G_n\right) - \psi_i\left(Z_i, \eta_0, \hat G_n\right) = \frac{\partial\psi_i}{\partial m_i}\bigg|_{\eta_0, \hat G_n} \Delta_{mi} + \frac{\partial\psi_i}{\partial s_i}\bigg|_{\eta_0, \hat G_n} \Delta_{si} + \underbrace{\frac{1}{2}\Delta_i' H_i(\tilde\eta_i, \hat G_n)\Delta_i}_{R_{1i}} \tag{B.15}$$

where $H_i(\tilde\eta_i, \hat G_n)$ is the Hessian matrix $\frac{\partial^2\psi_i}{\partial\eta_i\partial\eta_i'}$ evaluated at some intermediate value $\tilde\eta_i$ lying on the line segment between $\hat\eta_i$ and η_{0i}.

We further decompose the first-order terms into an empirical process term and a mean-component term. By Theorem B.4.13, (B.37), and (B.39), for

$$\rho_n = \frac{1}{n^3} e^{-C_{\mathcal{H}} M_n^2 \Delta_n} \wedge \frac{1}{e\sqrt{2\pi}}, \tag{B.16}$$

we have that the numerators to the first derivatives can be truncated at ρ_n, as the truncation does not bind:

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

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

Let

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

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

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

Hence, we can decompose the first-order terms in a_n as

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

Let the second order term be $R_1 = \frac{1}{n} \sum_i R_{1i}$. We let $a_n = |R_1| + \sum_{k \in \{m,s\}} |U_{1k}| + |U_{2k}|$

Taylor expansion of $\psi_i(Z_i, \hat{\eta}, G_0) - \psi_i(Z_i, \eta_0, G_0)$

Like (B.15), we similarly decompose

$$\psi_i(Z_i, \hat{\eta}, G_0) - \psi_i(Z_i, \eta_0, G_0) = \left.\frac{\partial \psi_i}{\partial m_i}\right|_{\eta_0, G_0} \Delta_{mi} + \left.\frac{\partial \psi_i}{\partial s_i}\right|_{\eta_0, G_0} \Delta_{si} + \underbrace{\frac{1}{2} \Delta_i' H_i(\tilde{\eta}_i, G_0) \Delta_i}_{R_{2i}} \tag{B.17}$$

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

$$\equiv U_{3mi} + U_{3si} + R_{2i}. \tag{B.18}$$

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

Bounding each term individually

By our decomposition, we can write

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

The ensuing subsections bound each term individually. Here we give an overview of the main ideas:

1. We bound $\mathbb{1}(A_n)|U_{1m}|$ in almost sure terms in Theorem B.4.3 by observing that $|\bar{D}_{mi}|$ is small when $\hat{G}_n$ is close to G_0, since $\bar{D}_{mi}(G_0, \eta_0, 0) = 0$. To do so, we need to control the differences

$$\bar{D}_{mi}(\hat{G}_n, \eta_0, \rho_n) - \bar{D}_{mi}(G_0, \eta_0, \rho_n)$$

and

$$\bar{D}_{mi}(G_0, \eta_0, \rho_n) - \underbrace{\bar{D}_{mi}(G_0, \eta_0, 0)}_{=0} = \bar{D}_{mi}(G_0, \eta_0, \rho_n).$$

Controlling the first difference features the Hellinger distance, while controlling the second relies on the fact that $P_{X \sim f(X)}(f(X) \leq \rho)$ cannot be too large, by a Chebyshev's inequality argument in Theorem B.4.16. Similarly, we bound $\mathbb{1}(A_n)|U_{1s}|$ in Theorem B.4.4.

2. The empirical process terms U_{2m}, U_{2s} are bounded probabilistically in Theorems B.4.5 and B.4.6 with statements of the form

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

To do so, we upper bound $\mathbb{1}(A_n)U_{2k} \leq \bar{U}_{2k}$ in almost sure terms. The upper bound is obtained by projecting $\hat{G}_n$ onto a ω-net of $\mathcal{P}(\mathbb{R})$ in terms of some pseudo-metric d_{k,∞,M_n} induced by $\bar{D}_{k,i}$. The upper bound $\bar{U}_{2k}$ then takes the form

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

Large deviation of $\bar{U}_{2k}$ 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 Theorem B.4.11 and Theorem B.4.12, which are minor extensions to Lemma 4 and Theorem 7 in Jiang (2020). The covering number is of a manageable size since the induced distributions f_{G,ν_i} are very smooth.

3. Since $\bar{D}_{k,i}(G_0, \eta_0, 0) = 0$. U_{3m}, U_{3s} are effectively also empirical process terms, without the additional randomness in $\hat{G}_n$. Thus the ω-net argument above is unnecessary for U_{3m}, U_{3s}, whereas the bounding follows from the same Dudley's chaining argument. Theorem B.4.8 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 Theorem B.4.22, thanks to the simultaneous moment control for G_0. Hence $\mathbb{1}(A_n)R_1$ can be bounded in almost sure terms. We bound $\mathbb{1}(A_n)R_2$ probabilistically. The second derivatives are bounded in Theorems B.4.7 and B.4.9.

(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 Theorem B.4.21, we can simply add the rates. Doing so, we find that the first term in ϵ_n (B.12) comes from U_{1s}, which dominates U_{1m}. The second term comes from U_{2s}, which dominates U_{2m}. The third term comes from R_1, which dominates R_2. The fourth term comes from U_{3s}. The leading terms in ϵ_n dominate κ_n, recalling (2.18). This completes the proof.

Before we proceed to the individual lemmas, we highlight a few convenient facts:

- The support of $\hat{G}_n$ is within $[-\bar{M}_n, \bar{M}_n]$, where $\bar{M}_n = \max_i |\hat{Z}_i(\hat{\eta})| \vee 1$. Under Assumption B.4.1, $\mathbb{1}(A_n)\bar{M}_n \lesssim_{\mathcal{H}} M_n$ by Theorem B.4.15(3).
- As a result, moments of $\hat{G}_n$ and $f_{\hat{G}_n,\nu_i}$ is bounded by appropriate moments of M_n, up to constants, under A_n.

B.4.3 Bounding U_{1m}

Lemma B.4.3. *Under Assumptions 2.3.1 to 2.3.4, assume additionally that $\|\hat{\eta}-\eta\|_\infty \le \Delta_n, \bar{Z}_n \le M_n$. Assume that the rates satisfy Assumption B.4.1. Then*

$$|U_{1m}| \equiv \left| \frac{1}{n} \sum_{i=1}^n \bar{D}_{mi}(\hat{G}_n, \eta_0, \rho_n)\Delta_{mi} \right| \lesssim_{\mathcal{H}} \Delta_n \left[\frac{\log n}{n} \sum_{i=1}^n h(f_{G_0,\nu_i}, f_{\hat{G}_n,\nu_I}) + \frac{M_n^{1/3}}{n} \right]. \quad \text{(B.19)}$$

Proof. Note that

$$
\begin{aligned}
|\bar{D}_{m,i}(\hat{G}_n, \eta_0, \rho_n)| &\lesssim_{s_{0\ell}} \left| \int \frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}} f_{G_0,\nu_i}(z) dz \right| \\
&= \left| \int \frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}} [f_{G_0,\nu_i}(z) - f_{\hat{G}_n,\nu_i}(z) + f_{\hat{G}_n,\nu_i}(z)] dz \right| \\
&\le \left| \int \frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}} [f_{G_0,\nu_i}(z) - f_{\hat{G}_n,\nu_i}(z)]\, dz \right| && \text{(B.20)} \\
&\quad + \left| \int \frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}} f_{\hat{G}_n,\nu_i}(z) dz \right| && \text{(B.21)}
\end{aligned}
$$

By the bounds for (B.20) and (B.21) 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 B.4.1. □

Bounding (B.20)

Consider the first term (B.20):

$$
\begin{aligned}
&\left| \int \frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}} \left(f_{G_0,\nu_i}(z) - f_{\hat{G}_n,\nu_i}(z) \right) dz \right| \\
&= \left| \int \frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}} \left(\sqrt{f_{G_0,\nu_i}(z)} - \sqrt{f_{\hat{G}_n,\nu_i}(z)} \right) \left(\sqrt{f_{G_0,\nu_i}(z)} + \sqrt{f_{\hat{G}_n,\nu_i}(z)} \right) dz \right| \\
&\le \left\{ \underbrace{\int \left(\sqrt{f_{G_0,\nu_i}(z)} - \sqrt{f_{\hat{G}_n,\nu_i}(z)} \right)^2 dz}_{2h^2} \cdot \int \left(\frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}} \right)^2 \left(\sqrt{f_{G_0,\nu_i}(z)} + \sqrt{f_{\hat{G}_n,\nu_i}(z)} \right)^2 dz \right\}^{1/2} \\
& && \text{(Cauchy–Schwarz)} \\
&\lesssim h(f_{G_0,\nu_i}, f_{\hat{G}_n,\nu_i}) \left\{ \int \left(\frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}} \right)^2 (f_{G_0,\nu_i}(z) + f_{\hat{G}_n,\nu_i}(z))\, dz \right\}^{1/2} && \text{(B.22)}
\end{aligned}
$$

By Theorems B.4.13 and B.4.14,

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

Hence,

$$(\text{B.20}) \lesssim_{\mathcal{H}} h(f_{G_0,\nu_i}, f_{\hat{G}_n,\nu_i})\sqrt{\log n}$$

Bounding (B.21)

The second term (B.21) is

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

(Cauchy–Schwarz and (B.44))

By Jensen's inequality and law of iterated expectations, the first term is bounded by $\frac{1}{\nu_i}$. By Theorem B.4.16, the second term is bounded by $\rho_n^{1/3}\operatorname{Var}_{Z\sim f_{\hat{G}_n,\nu_i}}(Z)^{1/6}$. Now,

$$\operatorname{Var}_{Z\sim f_{\hat{G}_n,\nu_i}}(Z) \le \nu_i^2 + \mu_2^2(\hat{G}_n) \lesssim_{\mathcal{H}} M_n^2.$$

Hence,

$$\left|\int \frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee \frac{\rho_n}{\nu_i}} f_{\hat{G}_n,\nu_i}(z)\,dz\right| \lesssim_{\mathcal{H}} M_n^{1/3}\rho_n^{1/3} \lesssim_{\mathcal{H}} M_n^{1/3} n^{-1}. \qquad \text{(Theorem B.4.13)}$$

B.4.4 Bounding U_{1s}

Lemma B.4.4. *Under Assumptions 2.3.1 to 2.3.4 and B.4.1, if* $\|\hat{\eta} - \eta\|_\infty \le \Delta_n$, $\bar{Z}_n \le M_n$, *then*

$$|U_{1s}| \lesssim_{\mathcal{H}} \Delta_n \left[\frac{M_n\sqrt{\log n}}{n}\sum_{i=1}^n h(f_{\hat{G}_n,\nu_i}, f_{G_0,\nu_i}) + \frac{M^{4/3}}{n}\right]. \qquad \text{(B.23)}$$

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

$$|\bar{D}_{s,i}(\hat{G}_n, \eta_0, \rho_n)| \lesssim_{\sigma_\ell,\sigma_u,s_{0\ell},s_{0u}} \left| \int \frac{Q_i(z, \eta_0, \hat{G}_n)}{f_{\hat{G}_n,\nu_i}(z) \vee (\rho_n/\nu_i)} (f_{G_0,\nu_i}(z) - f_{\hat{G}_n,\nu_i}(z))\, dz \right| \tag{B.24}$$

$$+ \left| \int \frac{Q_i(z, \eta_0, \hat{G}_n)}{f_{\hat{G}_n,\nu_i}(z) \vee (\rho_n/\nu_i)} f_{\hat{G}_n,\nu_i}(z)\, dz \right|. \tag{B.25}$$

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

Bounding (B.24)

The first term (B.24) is bounded by

$$\left(\int \frac{Q_i(z, \eta_0, \hat{G}_n)}{f_{\hat{G}_n,\nu_i}(z) \vee (\rho_n/\nu_i)} \left[f_{G_0,\nu_i}(z) - f_{\hat{G}_n,\nu_i}(z) \right] \, dz \right)^2$$

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

similar to the computation in (B.22).

By Theorems B.4.13 and B.4.17,

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

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

$$\text{(B.24)} \lesssim_{\sigma_\ell,\sigma_u,s_{0\ell},s_{0u}} M_n \sqrt{\log n} h(f_{G_0,\nu_i}, f_{\hat{G}_n,\nu_i}). \tag{B.26}$$

Bounding (B.25)

Observe that

$$\text{(B.25)} = \left| \int \frac{Q_i(z, \eta_0, \hat{G}_n)}{f_{\hat{G}_n,\nu_i}(z)} \left(\frac{f_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee (\rho_n/\nu_i)} - 1 \right) f_{\hat{G}_n,\nu_i}(z)\, dz \right|$$

Similar to our argument for (B.21), by Cauchy–Schwarz,

$$\begin{aligned}(\text{B.25}) &\le \left(E_{f_{\hat{G}_n,\nu_i}(z)}\left[\left(\mathbf{E}_{\hat{G}_n,\nu_i}\left[(Z-\tau)\tau \mid Z\right]\right)^2\right]\right)^{1/2}\sqrt{P_{f_{\hat{G}_n,\nu_i}(z)}(f_{\hat{G}_n,\nu_i}(z) \le \rho_n/\nu_i)}\\ &\lesssim_{\mathcal{H}} M_n \cdot \rho_n^{1/3} M_n^{1/3} \lesssim_{\mathcal{H}} \frac{M_n^{4/3}}{n}.\end{aligned}$$

B.4.5 Bounding U_{2m}

Lemma B.4.5. *Under Assumptions 2.3.1 to 2.3.4 and B.4.1,*

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

Proof. We prove this claim by first showing that if $\|\hat{\eta}-\eta\|_\infty \le \Delta_n$ and $\bar{Z}_n \le M_n$, we can upper bound $|U_{2m}|$ by some stochastic quantity $\bar{U}_{2m}$. Now, observe that

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

Hence, a stochastic upper bound on $\bar{U}_{2m}$ would verify the claim.

We now construct $\bar{U}_{2m}$ assuming $\|\hat{\eta}-\eta\|_\infty \le \Delta_n$ and $\bar{Z}_n \le M_n$. 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

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

On the event $\bar{Z}_n \le M_n$, $D_{m,i,M_n} = D_{m,i}$. We recall that

$$\begin{aligned}|U_{2m}| &= \left|\frac{1}{n}\sum_{i=1}^n (D_{m,i} - \bar{D}_{m,i})\Delta_{mi}\right|\\ &\le \left|\frac{1}{n}\sum_{i=1}^n (D_{m,i,M_n} - \bar{D}_{m,i,M_n})\Delta_{mi}\right| + \left|\frac{1}{n}\sum_{i=1}^n (\bar{D}_{m,i} - \bar{D}_{m,i,M_n})\Delta_{mi}\right|.\end{aligned}$$

Note that

$$|\bar{D}_{m,i} - \bar{D}_{m,i,M_n}| \lesssim_{\sigma_\ell,\sigma_u,s_{0\ell},s_{0u}} \left| \int_{|z|>M_n} \underbrace{\frac{f'_{\hat{G}_n,\nu_i}(z)}{f_{\hat{G}_n,\nu_i}(z) \vee (\rho_n/\nu_i)}}_{\lesssim_{\mathcal{H}} \sqrt{\log n}, \quad \text{Theorems B.4.13 and B.4.14}} f_{G_0,\nu_i}(z)\, dz \right|$$

$$\lesssim_{\mathcal{H}} \sqrt{\log n} P_{G_0,\nu_i}(|Z_i| > M_n)$$

By Theorem B.4.20, $P_{G_0,\nu_i}(|Z_i| > M_n) \le \exp\left(-C_{\alpha,A_0,\nu_u} M_n^\alpha\right)$. Hence, the second term

$$|\frac{1}{n}\sum_{i=1}^n (\bar{D}_{m,i} - \bar{D}_{m,i,M_n})\Delta_{mi}|$$

is bounded above by $e^{-C_{\mathcal{H}} M_n^\alpha} \sqrt{\log n}\Delta_n$, up to constants.

Note that under our assumptions, $\max_i |\hat{Z}_i| \vee 1 \le C_{\mathcal{H}} M_n$. Let $\mathcal{L} = [-C_{\mathcal{H}} M_n, C_{\mathcal{H}} M_n] \equiv [-\bar{M}, \bar{M}]$. Define

$$S = \left\{(m,s) : \|m - m_0\| \le \Delta_n, \|s - s_0\| \le \Delta_n, (m,s) \in C^p_{A_1}([\sigma_\ell, \sigma_u])\right\}. \tag{B.27}$$

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

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

Let $G_1, \ldots, G_N$ be an ω-net of $\mathcal{P}(\mathcal{L})$ in terms of $d_{m,\infty,M_n}(G_1, G_2)$, where N is taken to be the covering number

$$N = N\left(\omega, \mathcal{P}(\mathcal{L}), d_{m,\infty,M_n}(\cdot,\cdot)\right).$$

Let G_{j^*} be a G_j where $d_{m,\infty,M_n}(\hat{G}_n, G_{j^*}) \le \omega$.

By construction, $|\bar{D}_{m,i,M_n}(\hat{G}_n, \hat{\eta}, \rho_n) - \bar{D}_{m,i,M_n}(G_{j^*}, \hat{\eta}, \rho_n)| \le \omega$ as well, since the integrand is bounded uniformly. Hence, by projecting $\hat{G}_n$ to G_{j^*}, we obtain

$$\left|\frac{1}{n}\sum_{i=1}^n (D_{m,i,M_n}(Z_i, \hat{G}_n, \eta_0, \rho_n) - \bar{D}_{m,i,M_n}(\hat{G}_n, \eta_0, \rho_n))(\hat{m}(\sigma_i) - m_0(\sigma_i))\right|$$

$$\le 2\omega\Delta_n + \max_{j\in[N]}\left|\frac{1}{n}\sum_{i=1}^n (D_{m,i,M_n}(Z_i, G_j, \eta_0, \rho_n) - \bar{D}_{m,i,M_n}(G_j, \eta_0, \rho_n))(\hat{m}(\sigma_i) - m_0(\sigma_i))\right| \tag{B.29}$$

Next, consider the process

$$\begin{aligned}\eta \mapsto & \frac{1}{n}\sum_{i=1}^{n}(D_{m,i,M_n}(Z_i, G_j, \eta_0, \rho_n) - \bar{D}_{m,i,M_n}(G_j, \eta_0, \rho_n))(m(\sigma_i) - m_0(\sigma_i)) \\ & \equiv \frac{1}{n}\sum_{i=1}^{n} v_{i,j}(\eta) \equiv V_{n,j}(\eta)\end{aligned}$$

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

$$(\text{B.29}) \lesssim \omega\Delta_n + \max_{j\in[N]}\sup_{\eta\in S}|V_{n,j}(\eta)|.$$

Thus, we can take

$$\bar{U}_{2m} = C_{\mathcal{H}}\left\{e^{-C_{\mathcal{H}}M_n^\alpha}\sqrt{\log n}\Delta_n + \omega\Delta_n + \max_{j\in[N]}\sup_{\eta\in S}|V_{n,j}(\eta)|\right\}$$

where we shall prove a stochastic upper bound and optimize ω shortly.

By the results in Section B.4.5 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 Section B.4.5, we can pick ω such that

$$\omega\Delta_n + \max_{j\in[N]}\sup_{\eta\in S}V_{nj}(\eta) \lesssim_{\mathcal{H}} \Delta_n\sqrt{\log n}\left(\frac{\log n}{\sqrt{n}} + \frac{1}{\sqrt{n\Delta_n^{1/p}}}\right) \tag{B.30}$$

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

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

This concludes the proof. □

Bounding $\max_{j\in[N]}\sup_{\eta\in S}|V_{n,j}(\eta)|$

Note that $Ev_{ij}(\eta) = 0$. Moreover, by Theorems B.4.13 and B.4.14,

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

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,[15]

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

Hence $V_{n,j}(\eta)$ is a mean-zero process with subgaussian increments[16] 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 $u > 0$,

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

Note that

$$\sqrt{\log N(\epsilon, S, \|\cdot\|_\infty)} \leq \sqrt{2\log N(\epsilon, C^p_{A_1}([-\sigma_\ell, \sigma_u]), \|\cdot\|_\infty)} \leq \sqrt{2\log N(\epsilon/A_1, C^p_1([-\sigma_\ell, \sigma_u]), \|\cdot\|_\infty)}$$

By Theorem 2.7.1 in van der Vaart and Wellner (1996),

$$\log N(\epsilon/A_1, C^p_1([-\sigma_\ell, \sigma_u]), \|\cdot\|_\infty) \lesssim_{p,\sigma_\ell,\sigma_u} \left(\frac{A_1}{\epsilon}\right)^{1/p} \lesssim_{\mathcal{H}} \left(\frac{1}{\epsilon}\right)^{1/p}.$$

Hence, plugging in these calculations, we obtain

$$P\left[\sup_{\eta\in S}|V_{n,j}(\eta)| \gtrsim_{\mathcal{H}} \frac{\sqrt{\log n}}{\sqrt{n}}\left\{\Delta_n^{1-\frac{1}{2p}} + u\Delta_n\right\}\right] \leq 2e^{-u^2}.$$

This implies that

$$\sup_{\eta\in S}|V_{n,j}(\eta)| \lesssim_{\mathcal{H}} \frac{\sqrt{\log n}}{\sqrt{n}}\Delta_n^{1-\frac{1}{2p}} + \tilde{V}_{n,j},$$

[15] See Definition 2.5.6 in Vershynin (2018) for a definition of the ψ_2-norm (subgaussian norm).

[16] See Definition 8.1.1 in Vershynin (2018).

for some random variable $\tilde{V}_{n,j} \geq 0$ and $\|\tilde{V}_{n,j}\|_{\psi_2} \lesssim_{\mathcal{H}} \frac{\Delta_n}{\sqrt{n}}\sqrt{\log n}$.[17] Thus,

$$(B.29) \lesssim_{\mathcal{H}} \Delta_n \left[\omega + \frac{\sqrt{\log n}}{\sqrt{n\Delta_n^{1/p}}}\right] + \max_{j\in[N]} \tilde{V}_{n,j}.$$

Finally, note that by Theorem B.4.19 with the choice $t = \sqrt{\log n}$,

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

Selecting ω

The rate function that involves ω and $\log N$ is of the form

$$\omega + \sqrt{\frac{\log N}{n}}\sqrt{\log n}$$

Reparametrizing $\omega = \delta \log(1/\delta)\frac{\sqrt{\log(1/\rho_n)}}{\rho_n}$, by Theorem B.4.12, 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 B.4.1,

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

[17] 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 $P(\tilde{V}_{n,j} \gtrsim_{\mathcal{H}} u\frac{\Delta_n}{\sqrt{n}}M_n) \leq 2e^{-u^2}$ implies the ψ_2-norm bound by expression (2.14) in Vershynin (2018).

B.4.6 Bounding U_{2s}

Lemma B.4.6. *Under Assumptions 2.3.1 to 2.3.4 and B.4.1,*

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

Proof. This proof operates much like the proof of Theorem B.4.5. We observe that we can come up with an upper bound $\bar{U}_{2s}$ of U_{2s} under the event $\|\hat{\eta}-\eta\|_\infty \le \Delta_n$ and $\bar{Z}_n \le M_n$. A stochastic upper bound on $\bar{U}_{2s}$ then implies the lemma.

Let us first assume $\|\hat{\eta}-\eta\|_\infty \le \Delta_n$ and $\bar{Z}_n \le M_n$. Define D_{s,i,M_n} and $\bar{D}_{s,i,M_n}$ analogously to D_{m,i,M_n} and $\bar{D}_{m,i,M_n}$. A similar decomposition shows

$$|U_{2s}| \le \left|\frac{1}{n}\sum_{i=1}^n (D_{s,i,M_n} - \bar{D}_{s,i,M_n})\Delta_{si}\right| + \left|\frac{1}{n}\sum_{i=1}^n (\bar{D}_{s,i} - \bar{D}_{s,i,M_n})\Delta_{si}\right|$$

Theorem B.4.17 is a uniform bound on the integrand in the second term. Hence, the second term is bounded by

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^n (\bar{D}_{s,i} - \bar{D}_{s,i,M_n})\Delta_{si}\right| \\
&\lesssim_{\mathcal{H}} \Delta_n\sqrt{\log(1/\rho_n)}\frac{1}{n}\sum_{i=1}^n\left(\int_{|Z_i|>M_n}|z| f_{G_0,\nu_i}(z)\,dz + \sqrt{\log(1/\rho_n)}\int_{|Z_i|>M_n} f_{G_0,\nu_i}(z)\,dz\right) \\
&\lesssim_{\mathcal{H}} \Delta_n\sqrt{\log n}\left\{e^{-\frac{C_{\mathcal{H}}}{2}M_n^\alpha}\max_{i\in[n]}\mu_2(f_{G_0,\nu_i}) + \sqrt{\log n}e^{-C_{\mathcal{H}}M_n^\alpha}\right\} \\
&\qquad\qquad\text{(Cauchy–Schwarz for the first term and apply Theorems B.4.13 and B.4.20)} \\
&\lesssim_{\mathcal{H}} \Delta_n(\log n)e^{-C_{\mathcal{H}}M_n^\alpha}.
\end{aligned}$$

Note that under our assumptions, $\max_i |\hat{Z}_i| \vee 1 \le C_{\mathcal{H}}M_n$. Let $\mathcal{L} = [-C_{\mathcal{H}}M_n, C_{\mathcal{H}}M_n] \equiv [-\bar{M}, \bar{M}]$. 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

$$d_{s,\infty,M_n}(G_1,G_2) = \max_{i\in[n]}\sup_{|z|\le M_n}|D_{s,i}(z,G_1,\eta_0,\rho_n) - D_{s,i}(z,G_2,\eta_0,\rho_n)| \tag{B.31}$$

Let $G_1, \dots, G_N$ be an ω-net of $\mathcal{P}(\mathcal{L})$ in terms of $d_{s,\infty,M_n}(G_1, G_2)$, where

$$N = N\left(\omega, \mathcal{P}(\mathcal{L}), d_{s,\infty,M_n}(\cdot,\cdot)\right).$$

Let G_{j^*} be a G_j where $d_{s,\infty,M_n}(\hat{G}_n, G_{j^*}) \le \omega$. By construction, $|\bar{D}_{s,i,M_n}(\hat{G}_n, \eta_0, \rho_n) - \bar{D}_{s,i,M_n}(G_{j^*}, \eta_0, \rho_n)| \le \omega$ as well, since the integrand is bounded uniformly.

Hence

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^n (D_{s,i,M_n}(Z_i, \hat{G}_n, \eta_0, \rho_n) - \bar{D}_{s,i,M_n}(\hat{G}_n, \eta_0, \rho_n))(\hat{s}(\sigma_i) - s_0(\sigma_i))\right| \\ &\le 2\omega\Delta_n + \max_{j\in[N]}\left|\frac{1}{n}\sum_{i=1}^n (D_{s,i,M_n}(Z_i, G_j, \eta_0, \rho_n) - \bar{D}_{s,i,M_n}(G_j, \eta_0, \rho_n))(\hat{s}(\sigma_i) - s_0(\sigma_i))\right| \quad \text{(B.32)}\end{aligned}$$

Next, consider the process

$$\eta \mapsto \frac{1}{n}\sum_{i=1}^n (D_{s,i,M_n}(Z_i, G_j, \eta_0, 0) - \bar{D}_{s,i,M_n}(G_j, \eta_0, 0))(s(\sigma_i) - s_0(\sigma_i)) \equiv \frac{1}{n}\sum_{i=1}^n v_{i,j}(\eta) \equiv V_{n,j}(\eta)$$

so that (B.32) $\lesssim \omega\Delta_n + \max_{j\in[N]} \sup_{\eta\in S} |V_{n,j}(\eta)|$. This again upper bounds $|U_{is}|$ with some $\bar{U}_{is}$ that does not depend on the event $\|\hat{\eta} - \eta\|_\infty \le \Delta_n, \bar{Z}_n \le M_n$, on the event $\|\hat{\eta} - \eta\|_\infty \le \Delta_n, \bar{Z}_n \le M_n$. Hence, we can choose

$$\bar{U}_{2s} = C_{\mathcal{H}}\left\{\omega\Delta_n + \max_{j\in[N]}\sup_{\eta\in S}|V_{n,j}(\eta)| + \Delta_n(\log n)e^{-C_{\mathcal{H}}M_n^\alpha}\right\}.$$

It remains to show a tail bound with an appropriate choice of ω for $\bar{U}_{2s}$.

By Theorem B.4.17, 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 Section B.4.5, with a different constant for the subgaussianity. Hence, by the same argument as in Section B.4.5, with probability at least $1 - 2/n$,

$$\max_{j\in[N]}\sup_{\eta\in S}|V_{n,j}(\eta)| \lesssim_{\mathcal{H}} \frac{\Delta_n M_n\sqrt{\log n}}{\sqrt{n}}\left[\Delta_n^{-1/(2p)} + \sqrt{\log N} + \sqrt{\log n}\right]$$

We turn to selecting ω. The relevant term for selecting ω is $\omega + \frac{M_n\sqrt{\log n}}{\sqrt{n}}\sqrt{\log N}$. Reparametrize $\omega = M_n\sqrt{\log(1/\rho_n)}\delta\log(1/\delta)/\rho_n$. Pick $\delta = \rho_n/\sqrt{n} < 1/e$. The same argument as in Section B.4.5

with Theorem B.4.12 shows that

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

Therefore, we can select ω such that, overall, with probability at least $1 - 2/n$, under Assumption B.4.1,

$$\bar{U}_{2s} \lesssim_{\mathcal{H}} \Delta_n \left\{ M_n\sqrt{\log n}\exp\left(-C_{\alpha,A_0,\nu_u}M_n^\alpha\right) + \frac{M_n(\log n)^{3/2}}{\sqrt{n}} + M_n\sqrt{\log n}\frac{1}{\sqrt{n\Delta_n^{1/p}}} + \frac{\sqrt{\log n}}{\sqrt{n}}M_n\sqrt{\log n} \right\}$$
$$\lesssim_{\mathcal{H}} \Delta_n M_n\sqrt{\log n}\left\{ e^{-C_{\mathcal{H}}M_n^\alpha} + \frac{\log n}{\sqrt{n}} + \frac{1}{\sqrt{n\Delta_n^{1/p}}} \right\}.$$

This concludes the proof. □

B.4.7 Bounding R_1

Lemma B.4.7. *Recall R_{1i} from (B.15). Then, under Assumptions 2.3.1 to 2.3.4 and B.4.1, if $\|\hat{\eta} - \eta\|_\infty \le \Delta_n$ and $\bar{Z}_n \le M_n$, then $R_{1i} \lesssim_{\mathcal{H}} \Delta_n^2 M_n^2 \log n$.*

Proof. Observe that $R_{1i} \lesssim_{\sigma_\ell,\sigma_u,s_{0\ell},s_{0u}} \max\left(\Delta_{mi}^2, \Delta_{si}^2\right) \cdot \|H_i(\tilde{\eta}_i, \hat{G}_n)\|_\infty$, where $\|\cdot\|_\infty$ takes the largest element from a matrix by magnitude. By assumption, the first term is bounded by Δ_n^2. By Theorem B.4.18, 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. □

B.4.8 Bounding U_{3m}, U_{3s}

Lemma B.4.8. *Under Assumptions 2.3.2 to 2.3.4 and B.4.1,*

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

Proof. The proof structure follows that of Theorems B.4.5 and B.4.6.

Recall that

$$\begin{aligned}
U_{3m} &= \frac{1}{n}\sum_{i=1}^{n} D_{m,i}(Z_i, G_0, \eta_0, 0)(\hat{m}_i - m_0). \\
&= \frac{1}{n}\sum_{i=1}^{n}(D_{m,i,M_n} - \bar{D}_{m,i,M_n})(\hat{m}_i - m_0) + \bar{D}_{m,i,M_n}(\hat{m}_i - m_0)
\end{aligned}$$

Note that

$$\begin{aligned}
|\bar{D}_{m,i,M_n}| &= \left| \int_{|z|\le M_n} \frac{f'_{G_0,\nu_i}(z)}{f_{G_0,\nu_i}(z)} f_{G_0,\nu_i}(z)\, dz \right| \\
&= \left| \int \mathbb{1}\left(|z| > M_n\right) \cdot \frac{f'_{G_0,\nu_i}(z)}{f_{G_0,\nu_i}(z)} f_{G_0,\nu_i}(z)\, dz \right| \\
&\lesssim_{\sigma_\ell,\sigma_u,s_{0\ell},s_{0u}} P(|z| > M_n)^{1/2} \\
&\qquad \text{(Cauchy–Schwarz, Jensen, and law of iterated expectations via (B.44))} \\
&\lesssim_{\mathcal{H}} e^{-C_{\mathcal{H}} M_n^\alpha}. \qquad \text{(B.33)}
\end{aligned}$$

Recall S in (B.27). Define the process $V_n(\eta) = \frac{1}{n}\sum_i v_{n,i}(\eta) \equiv \frac{1}{n}\sum_{i=1}^n (D_{m,i,M_n} - \bar{D}_{m,i,M_n})(\hat{m}_i - m_0)$. Therefore, if $\|\hat{\eta} - \eta\|_\infty \le \Delta_n$, $\bar{Z}_n \le M_n$,

$$|U_{3m}| \lesssim_{\mathcal{H}} \Delta_n e^{-C_{\mathcal{H}} M_n^\alpha} + \sup_{\eta \in S} |V_n(\eta)| \equiv \bar{U}_{3m}.$$

Therefore, to bound U_{3m} it suffices to show a tail bound for $\sup_{\eta\in S}|V_n(\eta)|$. Observe that

$$V_n(\eta_1) - V_n(\eta_2) = \frac{1}{n}\sum_i (D_{m,i,M_n} - \bar{D}_{m,i,M_n})(\eta_{1i} - \eta_{2i})$$

Now, by Lemma 2.6.8 in Vershynin (2018), since $|D_{m,i,M_n}| \lesssim_{\mathcal{H}} M_n$ by Theorem B.4.22,

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

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

$$\|V_n(\eta_1) - V_n(\eta_2)\|_{\psi_2} \lesssim_{\mathcal{H}} \frac{M_n}{\sqrt{n}} \|\eta_1 - \eta_2\|_\infty \qquad \text{(B.34)}$$

Hence, by the same Dudley's chaining calculation in Section B.4.5, with probability at least $1 - 2/n$,

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

This concludes the proof for U_{3m}.

The proof for U_{3s} is similar. We need to establish the analogue of (B.33) and (B.34). For the tail bound (analogue of (B.33)), we have the same bound

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

For the analogue of (B.34), since Theorem B.4.22 implies that $|D_{s,i,M_n}| \lesssim_{\mathcal{H}} Z_i^2 \mathbb{1}(Z_i \le M_n) \le M_n^2$,

$$\|V_n(\eta_1) - V_n(\eta_2)\|_{\psi_2} \lesssim_{\mathcal{H}} \frac{M_n^2}{\sqrt{n}} \|\eta_1 - \eta_2\|_\infty.$$

Hence, with probability at most $2/n$

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

□

B.4.9 Bounding R_2

Lemma B.4.9. *Under Assumptions 2.3.2 to 2.3.4 and B.4.1, then*

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

Proof. Recall that $\mathbb{1}(A_n) = \mathbb{1}(\|\hat{\eta} - \eta\|_\infty \le \Delta_n, \bar{Z}_n \le M_n)$. 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 Theorem B.4.22, under G_0, these posterior moments are bounded by above by corresponding moments of $\hat{Z}_i(\tilde{\eta}_i)$. By Theorem B.4.22, under G_0, these posterior moments are bounded by above by corresponding moments of $\hat{Z}_i(\tilde{\eta}_i)$. Hence,

$$\mathbb{1}(A_n)\|H_i\|_\infty \lesssim_{\mathcal{H}} \mathbb{1}(A_n) \left(\hat{Z}_i(\tilde{\eta}_i) \vee 1\right)^4 \lesssim_{\mathcal{H}} (Z_i \vee 1)^4. \tag{B.35}$$

Hence,

$$\mathbb{1}(A_n)|R_2| \lesssim_{\mathcal{H}} \Delta_n^2 \frac{1}{n}\sum_{i=1}^n (Z_i \vee 1)^4.$$

By Chebyshev's inequality,

$$P\left(\frac{1}{n}\sum_{i=1}^n (Z_i \vee 1)^4 > E[(Z_i \vee 1)^4] + t\right) \le \frac{1}{t^2}\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^n (Z_i \vee 1)^4\right) = \frac{\operatorname{Var}(Z_i^4 \vee 1)}{nt^2}.$$

Picking $t^2 = \operatorname{Var}(Z_i^4 \vee 1)$ yields that

$$P\left(\frac{1}{n}\sum_{i=1}^n (Z_i \vee 1)^4 \gtrsim_{\mathcal{H}} 1\right) \le \frac{1}{n}.$$

Hence,

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

□

B.4.10 Derivative computations

It is sometimes useful to relate the derivatives of ψ_i to $\mathbf{E}_{G,\eta}$.

We compute the following derivatives. Since they are all evaluated at G, η, we let $\hat{\nu} = \hat{\nu}_i(\eta)$ and

$\hat{z} = \hat{Z}_i(\eta)$ as a shorthand.

$$\left.\frac{\partial \psi_i}{\partial m_i}\right|_{\eta,G} = -\frac{1}{s_i}\frac{f'_{G,\hat{\nu}}(\hat{z})}{f_{G,\hat{\nu}}(\hat{z})} \tag{B.36}$$

$$= \frac{s_i}{\sigma_i^2}\mathbf{E}_{G,\hat{\nu}}[Z-\tau \mid \hat{z}] \tag{B.37}$$

$$\left.\frac{\partial \psi_i}{\partial s_i}\right|_{\eta,G} = \frac{1}{\sigma_i \hat{\nu}_i(\eta) f_{G,\hat{\nu}(\eta)}(\hat{Z}_i(\eta))} \underbrace{\int (\hat{Z}_i(\eta)-\tau)\tau\varphi\left(\frac{\hat{Z}_i(\eta)-\tau}{\hat{\nu}_i(\eta)}\right)\frac{1}{\hat{\nu}_i(\eta)}\, G(d\tau)}_{Q_i(Z_i,\eta,G)} \tag{B.38}$$

$$= \frac{1}{\sigma_i\hat{\nu}}\mathbf{E}_{G,\hat{\nu}}[(Z-\tau)\tau \mid \hat{z}] \tag{B.39}$$

$$\left.\frac{\partial^2 \psi_i}{\partial m_i^2}\right|_{\eta,G} = \frac{1}{s_i^2}\left[\frac{f''_{G,\hat{\nu}}(\hat{z})}{f_{G,\hat{\nu}}(\hat{z})} - \left(\frac{f'_{G,\hat{\nu}}(\hat{z})}{f_{G,\hat{\nu}}(\hat{z})}\right)^2\right] \tag{B.40}$$

$$= \frac{1}{s_i^2}\left[\frac{1}{\hat{\nu}^4}\mathbf{E}_{G,\hat{\nu}}[(\tau-Z)^2 \mid \hat{z}] - \frac{1}{\hat{\nu}^2} - \frac{1}{\hat{\nu}^4}\left(\mathbf{E}_{G,\hat{\nu}}[(\tau-Z) \mid \hat{z}]\right)^2\right] \tag{B.41}$$

$$\left.\frac{\partial^2 \psi_i}{\partial m_i \partial s_i}\right|_{\eta,G} = \left(\frac{1}{\sigma_i^2}\mathbf{E}_{G,\hat{\nu}}[(Z-\tau)\tau \mid \hat{z}] - \frac{1}{s_i^2}\right)\frac{1}{\hat{\nu}^2}\mathbf{E}_{G,\hat{\nu}}[(\tau-Z) \mid \hat{z}] + \frac{\mathbf{E}_{G,\hat{\nu}}[(\tau-Z)^2\tau \mid \hat{z}]}{\hat{\nu}\sigma_i s_i} \tag{B.42}$$

$$\left.\frac{\partial^2 \psi}{\partial s^2}\right|_{\eta,G} = \frac{1}{\sigma_i^2}\left\{\mathbf{E}_{G,\hat{\nu}}\left[\left(\frac{s_i^2}{\sigma_i}(Z-\tau)^2 - 1\right)\tau^2 \mid \hat{z}\right] - \frac{1}{\hat{\nu}^2}\left(\mathbf{E}_{G,\hat{\nu}}[(Z-\tau)\tau \mid \hat{z}]\right)^2\right\} \tag{B.43}$$

It is useful to note that

$$\frac{f'_{G,\nu}(z)}{f_{G,\nu}(z)} = \frac{1}{\nu^2}\mathbf{E}_{G,\nu}[(\tau-Z) \mid z] \tag{B.44}$$

$$\frac{f''_{G,\nu}(z)}{f_{G,\nu}(z)} = \frac{1}{\nu^4}\mathbf{E}_{G,\nu}[(\tau-Z)^2 \mid z] - \frac{1}{\nu^2} \tag{B.45}$$

B.4.11 Metric entropy of $\mathcal{P}(\mathbb{R})$ under moment-based distance

The following is a minor generalization of Lemma 4 and Theorem 7 in Jiang (2020). In particular, Jiang (2020)'s Lemma 4 reduces to the case $q = 0$ below, and Jiang (2020)'s Theorem 7 relies on the results below for $q = 0, 1$. The proof largely follows the proofs of these two results of Jiang (2020).

We first state the following fact readily verified by differentiation.

Lemma B.4.10. *For all integer $m \geq 0$:*

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

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

Proposition B.4.11. *Fix some $q \in \mathbb{N} \cup \{0\}$ and $M > 1$. Consider the pseudometric*

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

Let ν_ℓ, ν_u be the lower and upper bounds of ν_i. Then, for all $0 < \delta < \exp(-q/2) \wedge e^{-1}$,

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

Proof. The proof strategy is as follows. First, we discretize $[-M, M]$ into a union of small intervals I_j. Fix G. There exists a finitely supported distribution G_m that matches moments of G on every I_j. It turns out that such a G_m is close to G in terms of $\|\cdot\|_{q,\infty,M}$. Next, we discipline G_m by approximating G_m with $G_{m,\omega}$, a finitely supported distribution supported on the fixed grid $\{k\omega : k \in \mathbb{Z}\} \cap [-M, M]$. Finally, the set of all $G_{m,\omega}$'s may be approximated by a finite set of distributions, and we count the size of this finite set.

Approximating G with G_m

First, let us fix some $\omega < \varphi(\sqrt{q}) \wedge \varphi(1)$.

Let $a = \frac{\nu_u}{\nu_\ell} \varphi_+(\omega) \ge 1$. Let $I_j = [-M + (j-2)a\nu_\ell, -M + (j-1)a\nu_\ell]$ be such that

$$I = [-M - a\nu_\ell, +M + a\nu_\ell] \subset \bigcup_j I_j$$

where I_j is a width $a\nu_\ell$ interval. Let $j^* = \lceil \frac{2M}{a\nu_\ell} + 2 \rceil$ be the number of such intervals.

There exists by Carathéodory's theorem a distribution G_m with support on I and no more than

$$m = (2k^* + q + 1)j^* + 1$$

support points s.t. the moments match

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

for some k^* to be chosen later.

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

$$d_{q,i,M}(G, G_m) \leq \max_{0 \leq v \leq q} \left[\left| \int_{(I_{j-1} \cup I_j \cup I_{j+1})^{\mathrm{C}}} \left(\frac{u-x}{\nu_i} \right)^v \varphi \left(\frac{x-u}{\nu_i} \right) (G(du) - G_m(du)) \right| \right. \tag{B.46}$$

$$\left. + \left| \int_{I_{j-1} \cup I_j \cup I_{j+1}} \left(\frac{u-x}{\nu_i} \right)^v \varphi \left(\frac{x-u}{\nu_i} \right) (G(du) - G_m(du)) \right| \right] \tag{B.47}$$

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 (B.46) is bounded by $\varphi_+(\omega)^v \omega$, as $\frac{|u-x|}{\nu_i} \geq a\nu_\ell/\nu_u = \varphi_+(\omega)$:

$$\text{(B.46)} \leq 2 \max_{0 \leq v \leq q} \varphi_+(\omega)^v \omega = 2\varphi_+(\omega)^q \omega.$$

Note that

$$\varphi(t) = \sum_{k=0}^{\infty} \frac{(-t^2/2)^k}{\sqrt{2\pi} k!} = \sum_{k=0}^{k^*} \frac{(-t^2/2)^k}{\sqrt{2\pi} k!} + R(t)$$

Thus the second term (B.47) 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 (B.47) is equal to

$$\text{(B.47)} = \max_{0 \leq v \leq q} \left| \int_{(I_{j-1} \cup I_j \cup I_{j+1}^{\mathrm{C}})} \left(\frac{u-x}{\nu_i} \right)^v R \left(\frac{x-u}{\nu_i} \right) (G(du) - G_m(du)) \right|.$$

We know that since $\varphi(t)$ has alternating-signed Taylor expansion,

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

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

$$
\begin{aligned}
(\text{B.47}) &\le 2 \cdot (2a)^q \cdot \frac{\left((2a)^2/2\right)^{k^*+1}}{\sqrt{2\pi}(k^*+1)!} && ((2a)^v \le (2a)^q) \\
&\le \frac{2(2a)^q}{(2\pi)\sqrt{k^*+1}} \left(\frac{2a^2}{k^*+1} e\right)^{k^*+1} \\
&\qquad \text{(Recall Stirling's formula } (k^*+1)! \ge \sqrt{2\pi(k^*+1)} \left(\tfrac{k^*+1}{e}\right)^{k^*+1}\text{.)} \\
&\le \frac{(2a)^q}{\pi\sqrt{k^*+1}} \left(\frac{e}{3}\right)^{k^*+1} && (\text{Choosing } k^*+1 \ge 6a^2 \ge 6) \\
&\le \frac{(2a)^q}{\pi\sqrt{k^*+1}} \exp\left(-\frac{1}{2}\frac{k^*+1}{6}\right) && ((e/3)^6 \le e^{-1/2}) \\
&\le \frac{(2a)^q}{\sqrt{k^*+1}\sqrt{\pi/2}} \underbrace{\varphi(a\nu_\ell/\nu_u)}_{\varphi(\varphi_+(\omega))} && (k^*+1 \ge 6a^2 \ge 6(a\nu_\ell/\nu_u)^2) \\
&\le \frac{(2a)^q}{\sqrt{k^*+1}\sqrt{\pi/2}} \omega \\
&\le \frac{2^q}{\sqrt{3\pi}} \left(\frac{\nu_u}{\nu_\ell}\right)^{q-1} \varphi_+^{q-1}(\omega)\omega && (k^*+1 \ge 6a^2)
\end{aligned}
$$

This bounds (B.46) + (B.47) uniformly over $|x| \le M$. Therefore,

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

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 ω.

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

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

In the case of $g(x,u) = ((x-u)/\nu_i)^v \varphi((x-u)/\nu_i)$, this function is Lipschitz by Theorem B.4.10, we thus have that,

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

So far, we have shown that there exists a distribution with at most m support points, supported on the lattice points $\{j\omega : j \in \mathbb{Z}, |j\omega| \in I\}$, that approximates G up to

$$C_{q,\nu_u,\nu_\ell} \omega \log^{q/2}(1/\omega)$$

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

Covering the set of $G_{m,\omega}$

Let Δ^{m-1} be the $(m-1)$-simplex of probability vectors in m dimensions. Consider discrete distributions supported on the support points of $G_{m,\omega}$, which can be identified with a subset of Δ^{m-1}. Thus, there are at most $N(\omega, \Delta^{m-1}, \|\cdot\|_1)$ such distributions that form an ω-net in $\|\cdot\|_1$. Now, consider a distribution $G'_{m,\omega}$ where

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

Since $t^q \varphi(t)$ is bounded, we have that

$$\|G'_{m,\omega} - G_{m,\omega}\|_{q,i,M} \le \omega \max_{0 \le v \le q} v^{v/2} \varphi(\sqrt{v}) \lesssim_q \omega$$

by Theorem B.4.10.

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

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

Putting things together

In other words,

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

((6.24) in Jiang (2020))

Since $\omega < 1$ and $m \geq 2\frac{12a^2+3+q}{a\nu_\ell}(M+a\nu_\ell)$, the first term is of the form C^m:

$$
\frac{(\omega+2)(\omega+2(M+a\nu_\ell))e}{m} \leq \frac{3e}{m}(1+2(M+a\nu_\ell)) \lesssim \frac{a\nu_\ell}{12a^2+3+q} \lesssim \nu_\ell.
$$

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 $k^*+1 \geq 6a^2$, 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 $K \geq 1$. 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

$$
\omega^* = \frac{\delta}{c}\log(cK/\delta)^{q/2} \leq \delta\log(1/\delta)^{q/2}.
$$

In this case,

$$
\begin{aligned}
N\left(\delta \log(1/\delta)^{q/2}, \mathcal{P}(\mathbb{R}), \|\cdot\|_{q,\infty,M}\right) &\le N\left(\omega^* \log(1/\delta)^{q/2}, \mathcal{P}(\mathbb{R}), \|\cdot\|_{q,\infty,M}\right) \\
&\lesssim_{q,\nu_u,\nu_\ell} \log(1/\omega)^2 \max\left(\frac{M}{\sqrt{\log(1/\omega)}}, 1\right) \\
&\lesssim_{q,\nu_u,\nu_\ell} \log(1/\delta)^2 \max\left(\frac{M}{\sqrt{\log(1/\delta)}}, 1\right)
\end{aligned}
$$

This bound holds for all sufficiently small δ. Since $\delta \log(1/\delta)^{q/2}$ is increasing over $(0, e^{-q/2} \wedge e^{-1})$ and the right-hand side does not vanish over the interval, we can absorb larger δ's into the constant.

□

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

Proposition B.4.12. *Consider $d^{(q)}_{\infty,M}$ in Theorem B.4.11, $d_{s,\infty,M}$ in (B.31), and $d_{m,\infty,M}$ in (B.28) 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 $\delta \in (0, 1/e)$,

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

Proof. Recall that

$$
D_{s,i}(z_i, G, \eta_0, \rho_n) = \frac{s_i}{\sigma_i^2}\frac{Q_i(Z_i, \eta_0, G)}{f_{i,G} \vee \frac{\rho_n}{\nu_i}}.
$$

Hence

$$
\begin{aligned}
&|D_{s,i}(z, G_1, \eta_0, \rho_n) - D_{s,i}(z, G_2, \eta_0, \rho_n)| \\
&\lesssim_{\mathcal{H}} \frac{1}{f_{i,G_1} \vee \frac{\rho_n}{\nu_i}} |Q_i(Z_i, \eta_0, G_1) - Q_i(Z_i, \eta_0, G_2)| + |Q_i(Z_i, \eta_0, G_2)| \left| \frac{1}{f_{i,G_1} \vee \frac{\rho_n}{\nu_i}} - \frac{1}{f_{i,G_1} \vee \frac{\rho_n}{\nu_i}} \right| \\
&\lesssim_{\mathcal{H}} \frac{1}{\rho_n} |f_{i,G_1} \mathbf{E}_{G_1,\nu_i}[(Z-\tau)\tau \mid z] - f_{i,G_2} \mathbf{E}_{G_2,\nu_i}[(Z-\tau)\tau \mid z]| \\
&\qquad + \frac{M\sqrt{\log(1/\rho_n)} + \log(1/\rho_n)}{\rho_n} |f_{i,G_1} - f_{i,G_2}|
\end{aligned}
$$

where the last inequality follows from the definition of Q_i and Theorem B.4.17.

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 $d_{\infty,M}^{(2)}$,

$$
|\mathbf{E}_{G_1,\nu_i}[(Z-\tau)\tau \mid z] - \mathbf{E}_{G_2,\nu_i}[(Z-\tau)\tau \mid z]| \lesssim_{\mathcal{H}} \delta(1+M) \lesssim M\delta.
$$

Similarly, $|f_{i,G_1} - f_{i,G_2}| \lesssim_{\mathcal{H}} \delta$. Hence,

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

Similarly,

$$
D_{m,i}(z, G, \eta_0) = \frac{s_i}{\sigma_i^2} \frac{f_{i,G} \mathbf{E}_{G,\nu_i}[(Z-\tau) \mid z]}{f_{i,G} \vee \rho_n/\nu_i}.
$$

Therefore

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

by a similar calculation, involving Theorem B.4.14.

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

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

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

B.4.12 Auxiliary lemmas

Lemma B.4.13. *Suppose* $|\bar{Z}_n| = \max_{i\in[n]} |Z_i| \vee 1 \le M_n$, $\|\hat{s} - s_0\|_\infty \le \Delta_n$, *and* $\|\hat{m} - m_0\|_\infty \le \Delta_n$. *Let* $\hat{G}_n$ *satisfy Assumption 2.3.1. Then, under Assumption B.4.1,*

1. $|\hat{Z}_i \vee 1| \lesssim_{\mathcal{H}} M_n$
2. *There exists* $C_{\mathcal{H}}$ *such that with* $\rho_n = \frac{1}{n^3}\exp\left(-C_{\mathcal{H}} M_n^2 \Delta_n\right) \wedge \frac{1}{e\sqrt{2\pi}}$,
$$f_{\hat{G}_n,\nu_i}(Z_i) \ge \frac{\rho_n}{\nu_i}.$$
3. *The choice of* ρ_n *satisfies* $\log(1/\rho_n) \asymp_{\mathcal{H}} \log n$, $\varphi_+(\rho_n) \asymp_{\mathcal{H}} \sqrt{\log n}$, *and* $\rho_n \lesssim_{\mathcal{H}} n^{-3}$.

Proof. Observe that $|\hat{Z}_i| \vee 1 \lesssim_{\sigma_\ell,\sigma_u,s_{0\ell},s_{0u}} (1+\Delta_n)M_n + \Delta_n \lesssim (1+\Delta_n)M_n$ by Theorem B.4.15(3). Hence by Assumption B.4.1, $|\hat{Z}_i| \vee 1 \lesssim_{\mathcal{H}} M_n$.

For (2), we note by Theorem 5 in Jiang (2020),

$$f_{\hat{G}_n,\hat{\nu}_i}(\hat{Z}_i) \ge \frac{1}{n^3 \hat{\nu}_i}$$

thanks to κ_n in (2.18). That is,

$$\int \varphi\left(\frac{\hat{Z}_i - \tau}{\hat{\nu}_i}\right) \hat{G}_n(d\tau) \ge \frac{1}{n^3}.$$

Now, note that

$$\frac{\hat{Z}_i - \tau}{\hat{\nu}_i} = \frac{Z_i + \frac{m_{0i}-\hat{m}_i}{s_{0i}} + \left(1 - \frac{\hat{s}_i}{s_{0i}}\right)\tau - \tau}{\nu_i} = \frac{Z_i - \tau}{\nu_i} + \frac{m_{0i} - \hat{m}_i}{\sigma_i} + \frac{1}{\sigma_i}(s_i - s_{0i})\tau = \frac{Z_i - \tau}{\nu_i} + \xi(\tau) \tag{B.48}$$

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

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

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

Therefore,

$$\int \varphi\left(\frac{Z_i-\tau}{\nu_i}\right)\hat{G}_n(d\tau) \ge \frac{1}{n^3}e^{-C_{\mathcal{H}}\Delta_n M_n^2}.$$

Dividing by ν_i on both sides finishes the proof of (2). Claim (3) is immediate by calculating $\log(1/\rho_n) = \left(3\log n - C_{\mathcal{H}}M_n^2\Delta_n^2\right) \vee \log(e\sqrt{2\pi}) \lesssim_{\mathcal{H}} \log n$ and apply Assumption B.4.1(1) to obtain that $\Delta_n M_n^2 \lesssim_{\mathcal{H}} 1$. □

Lemma B.4.14 (Lemma 2 Jiang (2020)). *For all $x \in \mathbb{R}$ and all $\rho \in (0, 1/\sqrt{2\pi e})$,*

$$\left|\frac{\nu^2 f'_{H,\nu}(x)}{(\rho/\nu) \vee f_{H,\nu}(x)}\right| \le \nu\varphi_+(\rho).$$

Moreover, for all $x \in \mathbb{R}$ and all $\rho \in (0, e^{-1}/\sqrt{2\pi})$,

$$\left|\left(\frac{\nu^2 f''_{H,\nu}(x)}{f_{H,\nu}(z)} + 1\right)\left(\frac{\nu f_{H,\nu}(x)}{(\nu f_{G,\nu}(x)) \vee \rho}\right)\right| \le \varphi_+^2(\rho),$$

where we recall φ_+ from (B.4).

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

$$0 \le \frac{\nu^2 f''_{H,\nu}(x)}{f_{H,\nu}(z)} + 1 \le \underbrace{\log\frac{1}{2\pi\nu^2 f_{H,\nu}(z)^2}}_{\varphi_+^2(\nu f_{H,\nu}(z))}.$$

Case 1 ($\nu f_{H,\nu}(x) \le \rho < e^{-1}/\sqrt{2\pi}$): Observe that $t\log\frac{1}{2\pi t^2}$ is increasing over $t \in (0, e^{-1}(2\pi)^{-1/2})$. Hence,

$$\left(\frac{\nu^2 f''_{H,\nu}(x)}{f_{H,\nu}(z)} + 1\right)\nu f_{H,\nu}(x) \le \nu f_{H,\nu}\log\frac{1}{2\pi\nu^2 f_{H,\nu}(z)^2} \le \rho\log\frac{1}{2\pi\rho^2}.$$

Dividing by $(\nu f) \vee \rho = \rho$ confirms the bound for $\nu f < \rho$.

Case 2 ($\nu f > \rho$): Since $\log \frac{1}{2\pi t^2}$ is decreasing in t, we have that

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

□

Lemma B.4.15. *The following statements are true:*

1. *Under Assumption 2.3.4, $1/\hat{\nu}_i \lesssim_{s_{0u},\sigma_\ell} 1$ and $\hat{\nu}_i \lesssim_{s_{0\ell},\sigma_u} 1$*
2. *Under Assumption 2.3.4, $|1 - \frac{s_{0i}}{\hat{s}_i}| \lesssim_{s_{0\ell}} \|\hat{s} - s_0\|_\infty$*
3. *Under Assumption 2.3.4,*

$$\max_i |\hat{Z}_i| \lesssim_{\sigma_\ell,\sigma_u,s_{0\ell},s_{0u}} (1 + \|\hat{s} - s_0\|_\infty)\bar{Z}_n + \|\hat{m} - m_0\|_\infty$$

where $\bar{Z}_n$ is defined in (B.6).

Proof. 1. Immediate by $1/\hat{\nu}_i = \hat{s}_i/\sigma_i$ and $P[s_{0\ell} < \hat{s}_i < s_{0u}] = 1$.

2. Immediate by observing that $|1 - \frac{s_{0i}}{\hat{s}_i}| = |\frac{\hat{s}_i - s_{0i}}{\hat{s}_i}|$ and $P[s_{0\ell} < \hat{s}_i < s_{0u}] = 1$.

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

□

Lemma B.4.16 (Zhang (1997), p.186)**.** *Let f be a density and let $\sigma(f)$ be its standard deviation. Then, for any $M, t > 0$,*

$$\int_{-\infty}^{\infty} \mathbb{1}(f(z) \leq t) f(z)\, dz \leq \frac{\sigma(f)^2}{M^2} + 2Mt.$$

In particular, choosing $M = t^{-1/3}\sigma(f)^{2/3}$ gives

$$\int_{-\infty}^{\infty} \mathbb{1}(f(z) \leq t) f(z)\, dz \leq 3t^{2/3}\sigma^{2/3}.$$

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 $E_f[Z] = 0$.

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

□

Lemma B.4.17. *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 $\rho_n \in (0, e^{-1}/\sqrt{2\pi})$,*

$$\left| \frac{Q_i(z, \eta_0, G)}{f_{G,\nu_i}(z) \vee (\rho_n/\nu_i)} \right| \le \varphi_+(\rho_n)\left(\nu_i|z| + \nu_i\varphi_+(\rho_n)\right). \tag{B.49}$$

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 Theorem B.4.14,

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

and

$$\frac{f_{G,\nu_i}(z)}{f_{G,\nu_i}(z) \vee (\rho_n/\nu_i)} \mathbf{E}_{G,\nu_i}[(z-\tau)^2 \mid z] = \nu_i^2 \left(\frac{\nu_i^2 f''_{i,G}}{f_{i,G}} + 1 \right) \frac{f_{G,\nu_i}(z)}{f_{G,\nu_i}(z) \vee (\rho_n/\nu_i)} \le \nu_i^2 \varphi_+^2(\rho_n).$$

Therefore,

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

□

Lemma B.4.18. *Under the assumptions in Theorem B.4.13, suppose $\tilde{\eta}_i$ lies on the line segment between η_0 and $\hat{\eta}_i$ and define $\tilde{\nu}_i, \tilde{m}_i, \tilde{s}_i, \tilde{Z}_i$ accordingly. Then, the second derivatives* (B.40), (B.42), (B.43), *evaluated at* $\tilde{\eta}_i, \hat{G}_n, \tilde{Z}_i$, *satisfy*

$$|(\text{B.40})| \lesssim_{\mathcal{H}} \log n$$

$$|(\text{B.42})| \lesssim_{\mathcal{H}} M_n \log n$$

$$|(\text{B.43})| \lesssim_{\mathcal{H}} M_n^2 \log n.$$

Proof. First, we show that

$$|\log(f_{\hat{G}_n, \tilde{\nu}_i}(\tilde{Z}_i)\tilde{\nu}_i)| \lesssim_{\mathcal{H}} \log n. \tag{B.50}$$

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 $\|\tilde{s} - \hat{s}\|_\infty \leq \Delta_n$ and $\|\tilde{m} - \hat{m}\|_\infty \leq \Delta_n$. This also shows that $|\tilde{Z}_i| \lesssim_{\mathcal{H}} M_n$ under the assumptions.

Note that by the same argument in (B.48) in Theorem B.4.13, we have that

$$\varphi\left(\frac{\hat{Z}_i - \tau}{\hat{\nu}_i}\right) \leq \varphi\left(\frac{\tilde{Z}_i - \tau}{\tilde{\nu}_i}\right) e^{-C_{\mathcal{H}}\Delta_n M_n^2}.$$

Hence,

$$\tilde{\nu}_i f_{\hat{G}_{(i)}, \tilde{\nu}_i}(\tilde{Z}_i) \geq \frac{1}{n^3} e^{-C_{\mathcal{H}}\Delta_n M_n^2}.$$

This shows (B.50).

Now, observe that

$$\mathbf{E}_{\hat{G}_n, \tilde{\nu}}[(\tau - Z)^2 \mid \tilde{Z}_i] \lesssim_{\mathcal{H}} \log\left(\frac{1}{\tilde{\nu}_i f_{\hat{G}_{(i)}, \tilde{\nu}_i}(\tilde{Z}_i)}\right) \lesssim_{\mathcal{H}} \log n$$

and

$$\mathbf{E}_{\hat{G}_n, \tilde{\nu}}[|\tau - Z| \mid \tilde{Z}_i] \lesssim_{\mathcal{H}} \sqrt{\log\left(\frac{1}{\tilde{\nu}_i f_{\hat{G}_{(i)}, \tilde{\nu}_i}(\tilde{Z}_i)}\right)} \lesssim_{\mathcal{H}} \sqrt{\log n}$$

by Theorem B.4.14, 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 Theorem B.4.17, and plugging in $\rho = \tilde{\nu}_i f_{\hat{G}_{(i)}, \tilde{\nu}_i}(\tilde{Z}_i) \wedge \frac{1}{\sqrt{2\pi e}}$,

$$\left|\mathbf{E}_{\hat{G}_n, \tilde{\nu}}[(\tau - Z)Z \mid \tilde{Z}_i]\right| \lesssim_{\mathcal{H}} \sqrt{\log n}|\tilde{Z}_i| + \log n \lesssim_{\mathcal{H}} \sqrt{\log n} M_n.$$

Observe that

$$\left|\mathbf{E}_{\hat{G}_n,\tilde{\nu}_i}[(\tau - Z)^2\tau \mid \tilde{Z}_i]\right| \lesssim_{\mathcal{H}} M_n \mathbf{E}_{\hat{G}_n,\tilde{\nu}_i}[(\tau - Z)^2] \lesssim_{\mathcal{H}} M_n \log n.$$

since $|\tau| \lesssim_{\mathcal{H}} M_n$ under $\hat{G}_n$. Similarly,

$$\mathbf{E}_{\hat{G}_n,\tilde{\nu}_i}[(Z-\tau)^2\tau^2 \mid \tilde{Z}_i] \lesssim_{\mathcal{H}} M_n^2 \log n \quad \mathbf{E}_{\hat{G}_n,\tilde{\nu}_i}[\tau^2 \mid \tilde{Z}_i] \lesssim_{\mathcal{H}} M_n^2.$$

Plugging these intermediate results into (B.40), (B.42), (B.43) proves the claim. □

Lemma B.4.19. *Let $X_1, \ldots, X_J$ be subgaussian random variables with $K = \max_i \|X_i\|_{\psi_2}$, not necessarily independent. Then for some universal C, for all $t \geq 0$,*

$$P\left[\max_i |X_i| \geq CK\sqrt{\log J} + CKt\right] \leq 2e^{-t^2}.$$

Proof. By (2.14) in Vershynin (2018), $P(|X_i| > t) \leq 2e^{-ct^2/\|X_i\|_{\psi_2}^2} \leq 2e^{-ct^2/K}$ for some universal c. By a union bound,

$$P\left[\max_i |X_i| \geq Ku\right] \leq 2\exp\left(-cu^2 + \log J\right)$$

Choose $u = \frac{1}{\sqrt{c}}(\sqrt{\log J} + t)$ so that $cu^2 = \log J + t^2 + 2t\sqrt{\log J} \geq \log J + t^2$. Hence

$$2\exp\left(-cu^2 + \log J\right) \leq 2e^{-t^2}.$$

Implicitly, $C = 1/\sqrt{c}$. □

Lemma B.4.20. *Suppose Z has simultaneous moment control $E[|Z|^p]^{1/p} \leq Ap^{1/\alpha}$. Then*

$$P(|Z| > M) \leq \exp\left(-C_{A,\alpha} M^\alpha\right).$$

As a corollary, suppose $Z \sim f_{G_0,\nu_i}(\cdot)$ and G_0 obeys Assumption 2.3.2, then

$$P(|Z| > M) \leq \exp\left(-C_{A_0,\alpha,\nu_u} M^\alpha\right).$$

Proof. Observe that

$$P(|Z| > M) = P(|Z|^p > M^p) \leq \left\{\frac{Ap^{1/\alpha}}{M}\right\}^p. \qquad \text{(Markov)}$$

Choose $p = (M/(eA))^\alpha$ such that

$$\left\{\frac{Ap^{1/\alpha}}{M}\right\}^p = \exp(-p) = \exp\left(-\left(\frac{1}{eA}\right)^\alpha M^\alpha\right).$$

□

Lemma B.4.21. *Let E be some event and assume that*

$$P(E, A > a) \leq p_1 \quad P(E, B > b) \leq p_2$$

Then $P(E, A + B > a + b) \leq p_1 + p_2$

Proof. Note that $A + B > a + b$ implies that one of $A > a$ and $B > b$ occurs. Hence

$$P(E, A + B > a + b) \leq P(\{E, A > a\} \cup \{E, B > b\}) \leq p_1 + p_2$$

by union bound. □

Lemma B.4.22. *Let $\tau \sim G_0$ where G_0 satisfies Assumption 2.3.2. Let $Z \mid \tau \sim \mathcal{N}(\tau, \nu^2)$. Then the posterior moment is bounded by a power of $|z|$:*

$$E[|\tau|^p \mid Z = z] \lesssim_p (|z| \vee 1)^p$$

Proof. Let $M \geq |z| \vee 2$. We write

$$E[|\tau|^p \mid Z = z] = \frac{1}{f_{G_0,\nu}(z)} \int |\tau|^p \varphi\left(\frac{z-\tau}{\nu}\right) \frac{1}{\nu} G_0(d\tau).$$

Note that

$$\begin{aligned}\int |\tau|^p \varphi\left(\frac{z-\tau}{\nu}\right) \frac{1}{\nu} G_0(d\tau) &\leq (3M)^p f_{G_0,\nu}(z) + \int \mathbb{1}(|\tau| > 3M) |\tau|^p \varphi\left(\frac{z-\tau}{\nu}\right) \frac{1}{\nu} G_0(d\tau) \\ &\leq (3M)^p f_{G_0,\nu}(z) + \int_{|\tau|>3M} |\tau|^p G_0(d\tau) \cdot \frac{1}{\nu} \varphi(|2M|/\nu) \\ &\qquad (|z - \tau| \geq 2M \text{ when } |\tau| > 3M)\end{aligned}$$

Also note that

$$f_{G_0,\nu}(z) = \int \varphi\left(\frac{z-\tau}{\nu}\right)\frac{1}{\nu}G_0(d\tau) \geq \frac{1}{\nu}\varphi\left(|2M|/\nu\right)G_0([-M,M])$$

$$(|z-\tau| \leq 2M \text{ if } \tau \in [-M,M])$$

Hence,

$$E[|\tau|^p \mid Z=z] \leq (3M)^p + \frac{\int |\tau|^p G_0(d\tau)}{G_0([-M,M])}$$

Since G_0 is mean zero and variance 1, by Chebyshev's inequality, $G_0([-M,M]) \geq G_0([-2,2]) \geq 3/4$.

Hence

$$E[|\tau|^p \mid Z=z] \lesssim_p M^p \lesssim_p (|z| \vee 1)^p,$$

since we have simultaneous moment control by Assumption 2.3.2. □

B.5 Regret control proofs: A large-deviation inequality for the average Hellinger distance

Theorem B.5.1. *For some $n > \sqrt{2\pi}e$, let $\tau_1,\ldots,\tau_n \mid (\nu_1^2,\ldots,\nu_n^2) \overset{\text{i.i.d.}}{\sim} G_0$ where G_0 satisfies Assumption 2.3.2. Let $\nu_u = \max_i \nu_i$ and $\nu_\ell = \min_i \nu_i$. Assume $Z_i \mid \tau_i, \nu_i^2 \sim \mathcal{N}(\tau_i, \nu_i^2)$. Fix positive sequences $\gamma_n, \lambda_n \to 0$ with $\gamma_n, \lambda_n \leq 1$ and constant $\epsilon > 0$. Fix some positive constant C^*. Consider the set of distributions that approximately maximize the likelihood*

$$A(\gamma_n,\lambda_n) = \left\{H : \mathrm{Sub}_n(H) \leq C^*\left(\gamma_n^2 + \bar{h}(f_{H,\cdot}, f_{G_0,\cdot})\lambda_n\right)\right\}.$$

Also consider the set of distributions that are far from G_0 in $\bar{h}$:

$$B(t,\lambda_n,\epsilon) = \left\{H : \bar{h}(f_{H,\cdot}, f_{G_0,\cdot}) \geq tB\lambda_n^{1-\epsilon}\right\}$$

with some constant B to be chosen. Assume that for some C_λ,

$$\lambda_n^2 \geq \left(\frac{C_\lambda}{n}(\log n)^{1+\frac{\alpha+2}{2\alpha}}\right) \vee \gamma_n^2.$$

Then the probability that $A \cap B$ is nonempty is bounded for $t > 1$: There exists a choice of B that depends only on $\nu_\ell, \nu_u, C^*, C_\lambda$ such that

$$P\left[A(\gamma_n, \lambda_n) \cap B(t, \lambda_n, \epsilon) \neq \emptyset\right] \leq (\log_2(1/\epsilon) + 1)n^{-t^2}. \tag{B.51}$$

Corollary B.5.2. *Let $\lambda_n = n^{-\frac{p}{2p+1}}(\log n)^{\gamma_1} \wedge 1$ and $\gamma_n = n^{-\frac{p}{2p+1}}(\log n)^{\gamma_2} \wedge 1$ where $\gamma_1 \geq \gamma_2 > 0$. Fix some $C^*_{\mathcal{H}}$. Fix $\epsilon > 0$. Then there exists $B_{\mathcal{H}}$ that depends solely on $C^*_{\mathcal{H}}, p, \gamma_1, \gamma_2, \nu_\ell, \nu_u$ such that*

$$P\left[\textit{There exists } H\textit{: } \mathrm{Sub}_n(H) \leq C^*_{\mathcal{H}}(\gamma_n^2 + \bar{h}(f_{H,\cdot}, f_{G_0,\cdot})\lambda_n) \textit{ and } \bar{h}(f_{H,\cdot}, f_{G_0,\cdot}) \geq tB_{\mathcal{H}}n^{-\frac{p}{2p+1}}(\log n)^{\gamma_1}\right]$$
$$\leq \left(\frac{\log\log n}{\log 2} + 1\right)n^{-t^2}$$

Proof. First, note that $\lambda_n^2 \geq \gamma_n^2$ and $\lambda_n^2 \gtrsim \frac{(\log n)^{1+\frac{\alpha+2}{2\alpha}}}{n}$.

Note that $tB\lambda_n^{1-\epsilon} \leq tB_{\mathcal{H}}n^{-\frac{p}{2p+1}+\epsilon\frac{p}{2p+1}}(\log n)^{\gamma_1} \leq tB_{\mathcal{H}}n^{-\frac{p}{2p+1}+\epsilon}(\log n)^{\gamma_1}$. Therefore,

$$\left\{H : \bar{h}(f_{H,\cdot}, f_{G_0,\cdot}) \geq tB_{\mathcal{H}}n^{-\frac{p}{2p+1}+\epsilon}(\log n)^{\gamma_1}\right\} \subset \left\{H : \bar{h}(f_{H,\cdot}, f_{G_0,\cdot}) \geq tB\lambda_n^{1-\epsilon}\right\}.$$

As a result, the probability

$$P\left[\text{There exists } H\text{: } \mathrm{Sub}_n(H) \leq C^*_{\mathcal{H}}(\gamma_n^2 + \bar{h}(f_{H,\cdot}, f_{G_0,\cdot})\lambda_n) \text{ and } \bar{h}(f_{H,\cdot}, f_{G_0,\cdot}) \geq tB_{\mathcal{H}}n^{-\frac{p}{2p+1}+\epsilon}(\log n)^{\gamma_1}\right]$$

is upper bounded by

$$P\left[A(\gamma_n, \lambda_n) \cap B(t, \lambda_n, \epsilon) \neq \emptyset\right] \leq (\log_2(1/\epsilon) + 1)n^{-t^2}$$

via an application of Theorem B.5.1.

Finally, set $\epsilon = \frac{1}{\log n}$. Note that $n^\epsilon = n^{\frac{1}{\log n}} = \exp(\log n / \log n) = e$. Hence

$$tB_{\mathcal{H}}n^{-\frac{p}{2p+1}+\epsilon}(\log n)^{\gamma_1} = tB_{\mathcal{H}}en^{-\frac{p}{2p+1}}(\log n)^{\gamma_1}. \qquad \square$$

Corollary B.5.3. *Assume the conditions in Theorem B.4.2. That is,*

1. *The estimate $\hat{G}_n$ satisfies Assumption 2.3.1.*

2. *For $\beta \geq 0$, and suppose that Δ_n, M_n take the form (B.13).*

3. *Suppose Assumptions 2.3.2 to 2.3.4 hold.*

Define the rate function

$$\delta_n = n^{-p/(2p+1)}(\log n)^{\frac{2+\alpha}{2\alpha}+\beta}. \tag{B.52}$$

Then, there exists some constant $B_{\mathcal{H}}$, depending solely on $C^*_{\mathcal{H}}$ in Theorem B.4.2, β, and p, ν_ℓ, ν_u such that

$$P\left[\bar{Z}_n \leq M_n, \|\hat{\eta} - \eta\|_\infty \leq \Delta_n, h(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}) > B_{\mathcal{H}}\delta_n\right] \leq \left(\frac{\log\log n}{\log 2} + 10\right)\frac{1}{n}.$$

Proof. Let $\gamma = \frac{2+\alpha}{2\alpha} + \beta$. We first verify that, for ε_n in (B.14), we make the choices

$$\lambda_n = n^{-p/(2p+1)}(\log n)^{\frac{2+\alpha}{2\alpha}+\beta} \wedge 1 \quad \gamma_n = n^{-p/(2p+1)}(\log n)^{\frac{2+\alpha}{2\alpha}+\beta} \wedge 1$$

does satisfy $\lambda_n^2 \geq \gamma_n^2$, as required by Theorem B.4.2. Since $\varepsilon_n \lesssim \lambda_n \bar{h} + \gamma_n^2$, the truncation by 1 only affects our subsequent results by constant factors.

The event in question is a subset of the union of

$$\left\{\bar{Z}_n \leq M_n, \|\hat{\eta} - \eta\|_\infty \leq \Delta_n, \mathrm{Sub}_n(\hat{G}_n) > C^*_{\mathcal{H}}\varepsilon_n\right\}$$

and

$$\left\{\bar{Z}_n \leq M_n, \|\hat{\eta} - \eta\|_\infty \leq \Delta_n, \mathrm{Sub}_n(\hat{G}_n) \leq C^*_{\mathcal{H}}\varepsilon_n, \bar{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}) > B_{\mathcal{H}} n^{-p/(2p+1)}(\log n)^\gamma\right\}.$$

The first event has measure at most $9/n$ by Theorem B.4.2, and there exists a choice of $B_{\mathcal{H}}$ such that the second has measure at most $n^{-1}\left(\frac{\log\log n}{\log 2} + 1\right)$ by Theorem B.5.2. We conclude the proof by applying a union bound. □

B.5.1 Proof of Theorem B.5.1

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,

$$B_k = \left\{H : \bar{h}\left(f_{H,\cdot}, f_{G_0,\cdot}\right) \in \left(tB\lambda_n^{1-2^{-k}}, tB\lambda_n^{1-2^{-k+1}}\right]\right\}.$$

The relation $B(t, \lambda_n, \epsilon) \subset \bigcup_k B_k$ holds if we take $K = \lceil|\log_2(1/\epsilon)|\rceil$, since, in that case, $K \geq \log_2(1/\epsilon) \implies 2^{-K} \leq \epsilon \implies \lambda_n^{1-2^{-K}} \leq \lambda_n^{1-\epsilon}$.

We will bound

$$P(A(\gamma_n, \lambda_n) \cap B_k(t, \lambda_n) \neq \emptyset) \leq n^{-t^2}$$

which becomes the bound (B.51) by a union bound. For $k \in [K]$, define $\mu_{n,k} = B\lambda_n^{1-2^{-k+1}}$ such that $B_k = \left\{H : \bar{h}\left(f_{H,\cdot}, f_{G_0,\cdot}\right) \in (t\mu_{n,k+1}, t\mu_{n,k}]\right\}$. To that end, fix some k.

Construct a net for the set of densities f_G

Fix a positive constant M and define the seminorm

$$\|G\|_{\infty,M} = \max_{i\in[n]} \sup_{y\in[-M,M]} f_{G,\nu_i}(y).$$

Note that $\|G\|_{\infty,M}$ is proportional to $\|G\|_{0,\infty,M}$ defined in Theorem B.4.11. Fix $\omega = \frac{1}{n^2} > 0$ and consider an ω-net for the distribution $\mathcal{P}(\mathbb{R})$ under $\|\cdot\|_{\infty,M}$. Let $N = N(\omega, \mathcal{P}(\mathbb{R}), \|\cdot\|_{\infty,M})$ and the ω-net is the distributions $H_1, \ldots, H_N$. For each j, let $H_{k,j}$ be the distribution with

$$\bar{h}(f_{H_{k,j},\cdot}, f_{G_0,\cdot}) \geq \mu_{n,k+1}$$

if it exists, and let J_k collect the indices for which $H_{j,k}$ exists.

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 $\|H - H_j\|_{\infty,M} \leq \omega$. Moreover, H serves as a witness that $H_{k,j}$ exists, with $\|H - H_{k,j}\|_{\infty,M} \leq 2\omega$.

We can construct an upper bound for $f_{H,\nu_i}(z)$ via

$$f_{H,\nu_i}(z) \leq \begin{cases} f_{H_{k,j},\nu_i}(z) + 2\omega & |z| < M \\ \frac{1}{\sqrt{2\pi\nu_i}} & |z| \geq M \end{cases}.$$

Define

$$v(z) = \omega \mathbb{1}(|z| < M) + \frac{\omega M^2}{z^2} \mathbb{1}(|z| \geq M).$$

Observe that

$$f_{H,\nu_i}(z) \leq \frac{f_{H_{k,j},\nu_i}(z) + 2v(z)}{\sqrt{2\pi}\nu_i v(z)} \text{ if } |z| > M$$

$$f_{H,\nu_i}(z) \leq f_{H_{k,j},\nu_i}(z) + 2v(z) \text{ if } |z| \leq M.$$

Hence, the likelihood ratio between H and G_0 is upper bounded:

$$\begin{aligned}\prod_{i=1}^{n} \frac{f_{H,\nu_i}(Z_i)}{f_{G_0,\nu_i}(Z_i)} &\leq \prod_{i=1}^{n} \frac{f_{H_{k,j},\nu_i}(Z_i) + 2v(Z_i)}{f_{G_0,\nu_i}(Z_i)} \prod_{i:|Z_i|>M} \frac{1}{\sqrt{2\pi}\nu_i v(Z_i)} \\ &\leq \left(\max_{j \in J_k} \prod_{i=1}^{n} \frac{f_{H_{k,j},\nu_i}(Z_i) + 2v(Z_i)}{f_{G_0,\nu_i}(Z_i)}\right) \prod_{i:|Z_i|>M} \frac{1}{\sqrt{2\pi}\nu_i v(Z_i)}\end{aligned}$$

If $H \in A(t, \gamma_n, \lambda_n)$, then the likelihood ratio is lower bounded:

$$\prod_{i=1}^{n} \frac{f_{H,\nu_i}(Z_i)}{f_{G_0,\nu_i}(Z_i)} \geq \exp\left(-nC^*(\gamma_n^2 + \bar{h}\left(f_{H,\cdot}, f_{G_0,\cdot}\right)\lambda_n)\right) \geq \exp\left(-ntC^*(t\gamma_n^2 + \bar{h}\left(f_{H,\cdot}, f_{G_0,\cdot}\right)\lambda_n)\right). \tag{$t > 1$}$$

Hence,

$$\begin{aligned}&P\left[A(t, \gamma_n, \lambda_n) \cap B_k(t, \lambda_n) \neq \emptyset\right] \\ &\leq P\left\{\left(\max_{j \in J_k} \prod_{i=1}^{n} \frac{f_{H_{k,j},\nu_i}(Z_i) + 2v(Z_i)}{f_{G_0,\nu_i}(Z_i)}\right) \prod_{i:|Z_i|>M} \frac{1}{\sqrt{2\pi}\nu_i v(Z_i)} \geq \exp\left(-nt^2C^*(\gamma_n^2 + \mu_{n,k}\lambda_n)\right)\right\} \\ &\leq P\left[\max_{j \in J_k} \prod_{i=1}^{n} \frac{f_{H_{k,j},\nu_i} + 2v(Z_i)}{f_{G_0,\nu_i}(Z_i)} \geq e^{-nt^2aC^*(\gamma_n^2 + \mu_{n,k}\lambda_n)}\right] \qquad \text{(B.53)} \\ &\qquad + P\left[\prod_{i:|Z_i|>M} \frac{1}{\sqrt{2\pi}\nu_i v(Y_i)} \geq e^{nt^2(a-1)C^*(\gamma_n^2 + \mu_{n,k}\lambda_n)}\right] \qquad \text{(B.54)}\end{aligned}$$

The first inequality follows from plugging in $\bar{h} \leq t\mu_{n,k}$. The second inequality follows from choosing some $a > 1$ and applying union bound.

Bounding (B.53)

We consider bounding the first term (B.53) now:

$$\text{(B.53)} \leq \sum_{j \in J_k} P\left[\prod_{i=1}^{n} \frac{f_{H_{k,j},\nu_i} + 2v(Z_i)}{f_{G_0,\nu_i}(Z_i)} \geq e^{-nat^2 C^*(\gamma_n^2 + \mu_{n,k}\lambda_n)}\right] \qquad \text{(Union bound)}$$

$$\leq \sum_{j \in J_k} E\left[\prod_{i=1}^{n} \sqrt{\frac{f_{H_{k,j},\nu_i}(Z_i) + 2v(Z_i)}{f_{G_0,\nu_i}(Z_i)}}\right] e^{nat^2 C^*(\gamma_n^2 + \mu_{n,k}\lambda_n)/2}$$

(Take square root of both sides, then apply Markov's inequality)

$$= \sum_{j \in J_k} e^{nat^2 C^*(\gamma_n^2 + \mu_{n,k}\lambda_n)/2} \prod_{i=1}^{n} E\left[\sqrt{\frac{f_{H_{k,j},\nu_i}(Z_i) + 2v(Z_i)}{f_{G_0,\nu_i}(Z_i)}}\right] \qquad \text{(B.55)}$$

where the last step (B.55) is by independence over i. Note that

$$E\left[\sqrt{\frac{f_{H_{k,j},\nu_i}(Z_i) + 2v(Z_i)}{f_{G_0,\nu_i}(Y_i)}}\right] = \int_{-\infty}^{\infty} \sqrt{f_{H_{k,j},\nu_i}(x) + 2v(x)}\sqrt{f_{G_0,\nu_i}(x)}\, dx$$

$$\leq 1 - h^2(f_{H_{k,j},\nu_i}, f_{G_0,\nu_i}) + \int_{-\infty}^{\infty} \sqrt{2v(x) f_{G_0,\nu_i}(x)}\, dx$$

$$(\sqrt{a+b} \leq \sqrt{a} + \sqrt{b})$$

$$\leq 1 - h^2(f_{H_{k,j},\nu_i}, f_{G_0,\nu_i}) + \left(2\int_{-\infty}^{\infty} v(x)\, dx\right)^{1/2}$$

(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

$$\prod_i t_i = \exp \sum_i \log t_i \leq \exp\left(\sum_i (t_i - 1)\right).$$

and thus

$$\prod_{i=1}^{n} E\left[\sqrt{\frac{f_{H_{k,j},\nu_i} + 2v(Z_i)}{f_{G_0,\nu_i}(Z_i)}}\right] \leq \exp\left[-n\bar{h}^2(f_{H_{k,j},\cdot}, f_{G_0,\cdot}) + n\sqrt{8M\omega}\right].$$

Thus, we can further bound (B.55):

$$
\begin{aligned}
\text{(B.53)} \le \text{(B.55)} &= \sum_{j\in J_k} e^{n\alpha t^2(\gamma_n^2+\mu_{n,k}\lambda_n)/2} \prod_{i=1}^{n} E\left[\sqrt{\frac{f_{H_{k,j},\nu_i}+2v(Z_i)}{f_{G_0,\nu_i}(Z_i)}}\right] \\
&\le \sum_{j\in J_k} \exp\left\{\frac{nat^2C^*}{2}(\gamma_n^2+\mu_{n,k}\lambda_n) - n\bar{h}^2(f_{H_{k,j},\cdot}, f_{G,\cdot}) + n\sqrt{8M\omega}\right\} \\
&\le \sum_{j\in J_k} \exp\left\{\frac{nat^2C^*}{2}(\gamma_n^2+\mu_{n,k}\lambda_n) - nt^2\mu_{n,k+1}^2 + n\sqrt{8M\omega}\right\} \\
&\le \exp\left\{\frac{nat^2C^*}{2}(\gamma_n^2+\mu_{n,k}\lambda_n) - nt^2\mu_{n,k+1}^2 + n\sqrt{8M\omega} + \log N\right\} \quad (|J_k| \le N) \\
&\le \exp\left\{\frac{nat^2C^*}{2}(\gamma_n^2+\mu_{n,k}\lambda_n) - nt^2\mu_{n,k+1}^2 + n\sqrt{8M\omega} + C|\log\omega|^2 \max\left(\frac{M}{\sqrt{|\log\omega|}}, 1\right)\right\} \\
&\qquad\qquad (\text{Theorem B.4.11}, q = 0) \\
&= \exp\left\{\frac{nat^2C^*}{2}(\gamma_n^2+\mu_{n,k}\lambda_n) - nt^2\mu_{n,k+1}^2 + \sqrt{8M} + C(\log n)^2 \max\left(\frac{M}{\sqrt{\log n}}, 1\right)\right\}. \\
&\qquad\qquad (\text{Recall that } \omega = \tfrac{1}{n^2})
\end{aligned}
$$

Bounding (B.54)

We now consider bounding the second term (B.54). By Markov's inequality again (taking $x \mapsto x^{1/(2\log n)}$ on both sides, we can choose to bound

$$
\text{(B.54)} \le E\left[\prod_{i=1}^{n}\left(\frac{1}{(2\pi\nu_i^2)^{1/4}}\frac{Z_i}{M\sqrt{\omega}}\right)^{\frac{1}{\log n}\mathbb{1}(|Z_i|>M)}\right] \exp\left(-\frac{n(a-1)t^2C^*(\gamma_n^2+\mu_{n,k}\lambda_n)}{2\log n}\right)
$$

instead. Define

$$
a_i = \frac{1}{(2\pi\nu_i^2)^{1/4}M\sqrt{\omega}} \le \frac{C_{\nu_\ell}n}{M} \quad \lambda = \frac{1}{\log n}
$$

Apply Theorem B.5.4 to obtain the following. Note that to do so, we require

$$
M \ge \nu_u\sqrt{8\log n} \quad p \ge \frac{1}{\log n}
$$

Then,

$$\begin{aligned}
\log E\left[\prod_{i=1}^{n}\left(\frac{1}{(2\pi\nu_i^2)^{1/4}}\frac{Z_i}{M\sqrt{\omega}}\right)^{\frac{1}{\log n}\mathbb{1}(|Z_i|>M)}\right] &= \log E\left[\prod_{i}(a_iZ_i)^{\lambda\mathbb{1}(|Z_i|\geq M)}\right] \\
&\lesssim_{\nu_u} \sum_{i=1}^{n}(a_iM)^{\lambda}\left(\frac{1}{Mn}+\frac{2^p\mu_p^p(G_0)}{M^p}\right) \\
&\leq \sum_{i=1}^{n}(C_{\nu_\ell}n)^{\frac{1}{\log n}}\left(\frac{1}{Mn}+\frac{2^p\mu_p^p(G_0)}{M^p}\right) \\
&\lesssim_{\nu_u,\nu_\ell} \frac{1}{M}+\frac{2^pn\mu_p^p(G_0)}{M^p}
\end{aligned}$$

As a result,

$$\log[(\text{B.54})] \leq C_{\nu_u,\nu_\ell}\left(\frac{1}{M}+\frac{2^pn\mu_p^p(G_0)}{M^p}\right)-\frac{n(a-1)}{2\log n}t^2C^*\left(\gamma_n^2+B\lambda_n^{2(1-2^{-k})}\right). \tag{B.56}$$

Choosing p, M, a and verifying conditions

By Assumption 2.3.2, $\mu_p^p(G_0) \leq A_0^p p^{p/\alpha}$. Let $M = 2eA_0(c_m \log n)^{1/\alpha}$ and $p = (M/(2eA_0))^{1/\alpha}$ so that

$$2^p\mu_p^p(G_0)/M^p \leq \exp\left(-c_m \log n\right)$$

We choose $c_m \geq 2$ sufficiently large such that $M = 2eA_0(c_m \log n)^{1/\alpha} > \nu_u\sqrt{8\log n} \vee 1$ and $p \geq 1$ for all $n > 2$ to ensure that our application of Theorem B.5.4 is correct. Since $\alpha \leq 2$, such a choice is available. Hence,

$$\frac{2^pn\mu_p^p(G_0)}{M^p} \leq \frac{1}{n}.$$

Hence the first term in (B.56) is less than $2C_{\nu_u,\nu_\ell}$.

Choose $a = 1.5$ to obtain that

$$\begin{aligned}
\log[(\text{B.54})] &\leq 2C_{\nu_u,\nu_\ell}-\frac{n}{4\log n}t^2C^*\left(\gamma_n^2+B\lambda_n^{2(1-2^{-k})}\right) \\
&\leq t^2\left[2C_{\nu_u,\nu_\ell}-\frac{n}{4\log n}C^*B\lambda_n^2\right] && (t\geq 1, \gamma_n>0, \lambda_n<1) \\
&\leq t^2\left[2C_{\nu_u,\nu_\ell}-\frac{C^*BC_\lambda}{4}(\log n)\right] && (\lambda_n^2 \geq C_\lambda(\log n)^{1+\frac{\alpha+2}{2\alpha}}/n \geq C_\lambda(\log n)^2/n)
\end{aligned}$$

There exists a sufficiently large B dependent only on $C^*, C_\lambda, C_{\nu_u,\nu_\ell}$ where $2C_{\nu_u,\nu_\ell} - \frac{C^*BC_\lambda}{4}(\log n) \leq$

$-\log n$ for all $n \geq 2$. Hence, for all sufficiently large B,

$$\log[(\text{B.54})] \leq -t^2 \log n.$$

Similarly, under these choices,

$$\begin{aligned}
\log[(\text{B.53})] &\leq -nt^2 \left[-\frac{3}{4} C^* (\gamma_n^2 + B\lambda_n^{2(1-2^{-k})}) + B^2 \lambda_n^{2(1-2^{-k+1})} \right] + C(\log n)^{1+\frac{2+\alpha}{2\alpha}} \\
&\leq -nt^2 \left[-\frac{3}{4} C^* (\lambda_n^2 + B\lambda_n^{2(1-2^{-k})}) + B^2 \lambda_n^{2(1-2^{-k+1})} \right] + C(\log n)^{1+\frac{2+\alpha}{2\alpha}} t^2 \\
& \hfill (\gamma_n \leq \lambda_n, t \geq 1) \\
&\leq -t^2 \left[n\lambda_n^2 \left(-\frac{3}{4} C^* - \frac{3}{4} C^* B \left(\frac{1}{\lambda_n}\right)^{2^{-k+1}} + B^2 \left(\frac{1}{\lambda_n}\right)^{2^{-k+2}} \right) - C(\log n)^{1+\frac{2+\alpha}{2\alpha}} \right] \\
&\leq -t^2 \left[n\lambda_n^2 \left(\frac{1}{\lambda_n}\right)^{2^{-k+2}} \left(-\frac{3}{4} C^* - \frac{3}{4} C^* B + B^2 \right) - C(\log n)^{1+\frac{2+\alpha}{2\alpha}} \right] \\
& \hfill (\lambda_n \leq 1. \text{ Pick } B \text{ such that } -\tfrac{3}{4} C^* - \tfrac{3}{4} C^* B + B^2 > 0) \\
&\leq -t^2 \left[n\lambda_n^2 \left(-\frac{3}{4} C^* - \frac{3}{4} C^* B + B^2 \right) - C(\log n)^{1+\frac{2+\alpha}{2\alpha}} \right] \\
&\leq -t^2 (\log n)^{1+\frac{2+\alpha}{2\alpha}} \left[C_\lambda \left(-\frac{3}{4} C^* - \frac{3}{4} C^* B + B^2 \right) - C \right]
\end{aligned}$$

There exists choices of B, depending solely on $C^*, C, C_\lambda, C_{\nu_u,\nu_\ell}$ where

$$\left[C_\lambda \left(-\frac{3}{4} C^* - \frac{3}{4} C^* B + B^2 \right) - C \right] > 1$$

so that the above is at most $-t^2 \log n - \log 2$.

Putting the union bound together, we obtain that

$$(\text{B.53}) + (\text{B.54}) \leq n^{-t^2}.$$

This concludes the proof.

B.5.2 Auxiliary lemmas

Lemma B.5.4 (Lemma 5 in Jiang (2020))**.** *Suppose $Z_i \mid \tau_i \sim \mathcal{N}(\tau_i, \nu_i^2)$ where $\tau_i \mid \nu_i^2 \sim G_0$ independently across i. Let $0 < \nu_u, \nu_\ell < \infty$ be the upper and lower bounds for ν_i. Then, for all constants $M > 0, \lambda > 0, a_i > 0, p \in \mathbb{N}$ such that $M \geq \nu_u \sqrt{8 \log n}$, $\lambda \in (0, p \wedge 1)$, and*

$a_1, \dots, a_n > 0$:

$$E\left\{\prod_i |a_i Z_i|^{\lambda \mathbb{1}(|Z_i|\ge M)}\right\} \le \exp\left\{\sum_{i=1}^n (a_i M)^\lambda \left(\frac{4\nu_u}{Mn\sqrt{2\pi}} + \left(\frac{2\mu_p(G_0)}{M}\right)^p\right)\right\}.$$

B.6 Regret control proofs: An oracle inequality for the Bayes squared-error risk

Recall the definition of Regret in (B.5) and the event A_n in (B.6).

B.6.1 Controlling Regret on A_n^{C}

The first term is the regret when a bad event occurs, on which either the nuisance estimates are bad or the data has large values. The probability of this bad event is

$$P(A_n^{\mathrm{C}}) \le P(\|\hat\eta - \eta\|_\infty > \Delta_n) + P(\bar Z_n > M_n) \le P(\|\hat\eta - \eta\|_\infty > \Delta_n) + n^{-2}.$$

There exist choices of the constant in (B.13) for M_n such that $P(\bar Z_n > M_n) \le n^{-2}$, by Theorem B.6.8. Thus, at a minimum, the first term is $o(1)$ for appropriate choices of Δ_n, M_n such that $P(A_n^{\mathrm{C}}) \to 0$. We can also control the expected value of Regret on the bad event A_n^{C}.

Lemma B.6.1. *Under Assumptions 2.3.1 to 2.3.4. For $\beta \ge 0$, suppose $n > 3$ and suppose Δ_n, M_n satisfies (B.13) such that $P(\bar Z_n > M_n) \le n^{-2}$, we can decompose*

$$E[\mathrm{Regret}(\hat G_n, \hat\eta)\mathbb{1}(\|\hat\eta - \eta\|_\infty > \Delta_n)] \lesssim_{\mathcal{H}} P(\|\hat\eta - \eta\|_\infty > \Delta_n)^{1/2}(\log n)^{2/\alpha}$$
$$E[\mathrm{Regret}(\hat G_n, \hat\eta)\mathbb{1}(\bar Z_n > M_n)] \lesssim_{\mathcal{H}} \frac{1}{n}(\log n)^{2/\alpha}$$

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

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

by Cauchy–Schwarz.

A crude bound (Theorem B.6.7) shows that, almost surely,

$$\left\{\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_{i,\hat{G},\hat{\eta}} - \theta_i^*)^2\right\}^2 \lesssim_{\mathcal{H}} \bar{Z}_n^4.$$

Apply Theorem B.6.8 to find that $E[\bar{Z}_n^4] \lesssim_{\mathcal{H}} (\log n)^{4/\alpha}$. This proves both claims. □

B.6.2 Controlling Regret on A_n

Theorem B.6.2. *Assume the conditions in Theorem B.5.3. That is,*

1. *Suppose $\hat{G}_n$ satisfies Assumption 2.3.1.*
2. *For $\beta \geq 0$, suppose Δ_n, M_n satisfies (B.13).*
3. *Suppose Assumptions 2.3.2 to 2.3.4 hold.*

Then,

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

Proof. Let $C_{\mathcal{H}}^*$ be the constant in Theorem B.4.2 and $B_{\mathcal{H}}$ be the constant in Theorem B.5.3. Recall the Hellinger rate δ_n in (B.52).

Recall the decomposition (B.5) for Regret. Note that the term corresponding to the second term in the decomposition (B.5),

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

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

$$E\left[\frac{\mathbb{1}(A_n)}{n}\sum_{i=1}^n (\hat{\theta}_{i,\hat{G}_n,\hat{\eta}} - \theta_i^*)^2\right] \equiv \frac{1}{n}E[\mathbb{1}(A_n)\|\hat{\theta}_{\hat{G}_n,\hat{\eta}} - \theta^*\|^2], \tag{B.57}$$

where we let $\hat{\theta}_{\hat{G}_n,\hat{\eta}}$ denote the vector of estimated posterior means and let θ^* denote the corresponding vector of oracle posterior means. Let the subscript ρ_n denote a vector of regularized posterior means as in (B.3). Thus, we may further decompose,

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

Let

$$\xi_1 = \frac{\mathbb{1}(A_n)}{n} \|\hat{\theta}_{\hat{G}_n,\hat{\eta}} - \hat{\theta}_{\hat{G}_n,\eta_0}\|^2 \tag{B.58}$$

$$\xi_2 = \frac{\mathbb{1}(A_n)}{n} \|\hat{\theta}_{\hat{G}_n,\eta_0} - \hat{\theta}_{\hat{G}_n,\eta_0,\rho_n}\|^2 \tag{B.59}$$

$$\xi_3 = \frac{\mathbb{1}(A_n)}{n} \|\hat{\theta}_{\hat{G}_n,\eta_0,\rho_n} - \theta^*_{\rho_n}\|^2 \tag{B.60}$$

$$\xi_4 = \frac{\mathbb{1}(A_n)}{n} \|\theta^*_{\rho_n} - \theta^*\|^2 \tag{B.61}$$

corresponding to the square of each of the terms, such that

$$(\text{B.57}) \le 4(E\xi_1 + E\xi_2 + E\xi_3 + E\xi_4) = 4(E\xi_1 + E\xi_3 + E\xi_4).$$

Observe that $\xi_2 = 0$ by Theorem B.4.13, since the truncation by ρ_n does not bind when A_n occurs.

The ensuing subsections control $E\xi_1, E\xi_3, E\xi_4$ individually. Putting together the rates we obtain, we find that

$$\begin{aligned} \xi_1 \lesssim_{\mathcal{H}} M_n^6 \Delta_n^2 \implies E\xi_1 &\lesssim_{\mathcal{H}} M_n^2 (\log n)^2 \Delta_n^2 \\ E\xi_3 &\lesssim_{\mathcal{H}} (\log n)^3 \delta_n^2 \\ E\xi_4 &\lesssim_{\mathcal{H}} \frac{1}{n} \end{aligned}$$

Now, observe that $\delta_n \asymp_{\mathcal{H}} \Delta_n M_n^2 \gtrsim_{\mathcal{H}} \Delta_n M_n \log n$ and $\frac{1}{n} \lesssim_{\mathcal{H}} (\log n)^3 \delta_n^2$. Hence, the dominating rate is $(\log n)^3 \delta_n^2$. Plugging in δ_n^2 in (B.52) to obtain the rate

$$(\text{B.57}) \lesssim_{\mathcal{H}} n^{-\frac{2p}{2p+1}} (\log n)^{\frac{2+\alpha}{\alpha} + 3 + 2\beta_1}. \qquad \square$$

B.6.3 Controlling ξ_1

Lemma B.6.3. *Under the assumptions of Theorem B.6.2, in the proof of Theorem B.6.2,* $\xi_1 \lesssim_{\mathcal{H}} M_n^2 (\log n)^2 \Delta_n^2$.

Proof. Note that, by an application of Taylor's theorem,

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

where we use $\tilde{\eta}_i$ to denote some intermediate value lying on the line segment between $\hat{\eta}_i$ and η_{0i}. By Theorem B.4.18,

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

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

B.6.4 Controlling ξ_3

Lemma B.6.4. *Under the assumptions of Theorem B.6.2, in the proof of Theorem B.6.2,* $E\xi_3 \lesssim_{\mathcal{H}} (\log n)^3 \delta_n^2$.

Proof. Observe that

$$\left|\hat{\theta}_{i,\hat{G}_n,\eta_0,\rho_n} - \theta^*_{i,\rho_n}\right| = s_{0i} \left|\hat{\tau}_{i,\hat{G}_n,\eta_0,\rho_n} - \tau^*_{i,\rho_n}\right|$$

where $\hat{\tau}_{i,\hat{G}_n,\eta_0,\rho_n}$ is the regularized posterior with prior $\hat{G}_n$ at nuisance parameter η_0 and $\tau^*_{i,\rho_n} = \hat{\tau}_{i,G_0,\eta_0,\rho_n}$.

We shall focus on controlling

$$\mathbb{1}(A_n) \|\hat{\tau}_{\hat{G}_n,\eta_0,\rho_n} - \tau^*_{\rho_n}\|^2$$

Fix the rate function δ_n in (B.52) and the constant $B_{\mathcal{H}}$ in Theorem B.5.3 (which in turn depends on $C^*_{\mathcal{H}}$ in Theorem B.4.2). Let $B_n = \{\bar{h}(f_{\hat{G}_n,\cdot}, f_{G_0,\cdot}) < B_{\mathcal{H}} \delta_n\}$ be the event of a small average squared Hellinger distance. Let $G_1, \ldots, G_N$ be a finite set of prior distributions (chosen to be a net of $\mathcal{P}(\mathbb{R})$ in some distance), and let $\tau^{(j)}_{\rho_n}$ be the posterior mean vector corresponding to prior G_j with nuisance parameter η_0 and regularization ρ_n.

Then

$$\frac{\mathbb{1}(A_n)}{n}\|\hat{\tau}_{\hat{G}_n,\eta_0,\rho_n} - \tau^*_{\rho_n}\|^2 \leq \frac{4}{n}\left(\zeta_1^2 + \zeta_2^2 + \zeta_3^2 + \zeta_4^2\right)$$

where

$$\zeta_1^2 = \|\hat{\tau}_{\hat{G}_n,\eta_0,\rho_n} - \tau^*_{\rho_n}\|^2 \mathbb{1}\left(A_n \cap B_n^{\mathrm{C}}\right) \tag{B.62}$$

$$\zeta_2^2 = \left(\|\hat{\tau}_{\hat{G}_n,\eta_0,\rho_n} - \tau^*_{\rho_n}\| - \max_{j\in[N]}\|\tau^{(j)}_{\rho_n} - \tau^*_{\rho_n}\|\right)_+^2 \mathbb{1}(A_n \cap B_n) \tag{B.63}$$

$$\zeta_3^2 = \max_{j\in[N]}\left(\|\tau^{(j)}_{\rho_n} - \tau^*_{\rho_n}\| - E\left[\|\tau^{(j)}_{\rho_n} - \tau^*_{\rho_n}\|\right]\right)_+^2 \tag{B.64}$$

$$\zeta_4^2 = \max_{j\in[N]}\left(E\left[\|\tau^{(j)}_{\rho_n} - \tau^*_{\rho_n}\|\right]\right)^2 \tag{B.65}$$

The decomposition ζ_1 through ζ_4 is exactly analogous to Section C.3 in Soloff *et al.* (2021) and to the proof of Theorem 1 in Jiang (2020). In particular, ζ_1 is the gap on the "bad event" where the average squared Hellinger distance is large, which is manageable since $\mathbb{1}(A_n \cap B_n^{\mathrm{C}})$ has small probability by Theorem B.5.3. ζ_2 is the distance from the posterior means at $\hat{G}_n$ to the closest posterior mean generated from the net $G_1, \ldots, G_N$; ζ_2 is small if we make the net very fine. ζ_3 measures the distance between $\|\tau^{(j)}_{\rho_n} - \tau^*_{\rho_n}\|$ and its expectation; ζ_3 can be controlled by (i) a large-deviation inequality and (ii) controlling the metric entropy of the net (Theorem B.4.12). Lastly, ζ_4 measures the expected distance between $\tau^{(j)}_{\rho_n}$ and $\tau^*_{\rho_n}$; it is small since G_j are fixed priors with small average squared Hellinger distance.

However, our argument for ζ_3 is slightly different and avoids an argument in Jiang and Zhang (2009) which appears to not apply in the heteroskedastic setting. See Theorem B.6.5.

The subsequent subsections control ζ_1 through ζ_4, and find that $\zeta_4 \lesssim_{\mathcal{H}} (\log n)^3\delta_n^2$ is the dominating term. □

Controlling ζ_1

First, we note that

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

By Theorem B.5.3, $P(A_n \cap B_n^{\mathrm{C}}) \le \left(\frac{\log\log n}{\log 2} + 9\right)\frac{1}{n}$, and hence

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

Controlling ζ_2

Choose $G_1, \ldots, G_N$ to be a minimal ω-covering of $\left\{G : \bar{h}(f_{G,\cdot}, f_{G_0,\cdot}) \le \delta_n\right\}$ under the pseudometric

$$d_{M_n,\rho_n}(H_1, H_2) = \max_{i\in[n]} \sup_{z:|z|\le M_n} \left| \frac{\nu_i^2 f'_{H_1,\nu_i}(z)}{f_{H_1,\nu_i}(z) \vee \left(\frac{\rho_n}{\nu_i}\right)} - \frac{\nu_i^2 f'_{H_2,\nu_i}(z)}{f_{H_2,\nu_i}(z) \vee \left(\frac{\rho_n}{\nu_i}\right)} \right| \tag{B.66}$$

where $N \le N\left(\omega, \mathcal{P}(\mathbb{R}), d_{M_n,\rho_n}\right)$. We note that (B.66) and (B.28) are different only by constant factors. Therefore, Theorem B.4.12 implies that

$$\log N\left(\frac{\delta \log(1/\delta)}{\rho_n}\sqrt{\log(1/\rho_n)}, \mathcal{P}(\mathbb{R}), d_{M_n,\rho_n}\right) \lesssim_{\mathcal{H}} \log(1/\delta)^2 \max\left(1, \frac{M_n}{\sqrt{\log(1/\delta)}}\right) \tag{B.67}$$

for all sufficiently small $\delta > 0$.

Then

$$\begin{aligned}
\frac{1}{n}\zeta_2^2 &\le \mathbb{1}(A_n \cap B_n) \max_{j\in[N]} \|\hat{\tau}_{\hat{G}_n,\eta_0,\rho_n} - \tau_{\rho_n}^{(j)}\|^2 \quad \text{(Triangle inequality : } \|a-b\| - \|b-c\| \le \|a-c\|) \\
&= \mathbb{1}(A_n \cap B_n) \max_{j\in[N]} \sum_{i=1}^n \mathbb{1}\left(|Z_i| \le M_n\right) \left(\frac{\nu_i^2 f'_{\hat{G}_n,\nu_i}(Z_i)}{f_{\hat{G}_n,\nu_i}(Z_i) \vee \left(\frac{\rho_n}{\nu_i}\right)} - \frac{\nu_i^2 f'_{G_j,\nu_i}(Z_i)}{f_{G_j,\nu_i}(Z_i) \vee \left(\frac{\rho_n}{\nu_i}\right)}\right)^2 \\
&\le \omega^2 \\
&\le \frac{\delta^2 \log(1/\delta)^2}{\rho_n^2} \log(1/\rho_n). \qquad \text{(Reparametrize } \omega = \delta\log(1/\delta)\rho_n^{-1}\sqrt{\log(1/\rho_n)})
\end{aligned}$$

Controlling ζ_3

We first observe that $V_{ij} \equiv |\tau_{i,\rho_n}^{(j)} - \tau_{i,\rho_n}^*| \lesssim_{\mathcal{H}} \sqrt{\log n}$, by Theorem B.4.14. Let $V_j = (V_{1j}, \ldots, V_{nj})'$, we have that

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

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

Observe that

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

by Theorem B.6.9. By a union bound,

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

Therefore

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

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

$$\frac{1}{n}E[\zeta_2^2 + \zeta_3^2] \lesssim_{\mathcal{H}} \frac{(\log n)^3}{n}.$$

Remark B.6.5. For the analogous term in the homoskedastic setting, Jiang and Zhang (2009) (and, later on, Saha and Guntuboyina (2020)) observe that $\|\tau_{\rho_n}^{(j)} - \tau_{\rho_n}^*\|$ is a Lipschitz function of the noise component $Z_i - \tau_i$. As a result, a Gaussian isoperimetric inequality (Theorem 5.6 in Boucheron *et al.* (2013)) establishes that

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

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 $E\left[\|\tau_{\rho_n}^{(j)} - \tau_{\rho_n}^*\| \mid \tau_1, \ldots, \tau_n\right]$. In the homoskedastic setting where $\nu_i = \nu$,

$$E\left[\|\tau_{\rho_n}^{(j)} - \tau_{\rho_n}^*\| \mid \tau_1, \ldots, \tau_n\right] = E_{G_{0,n}}\left[\|\tau_{\rho_n}^{(j)} - \tau_{\rho_n}^*\|\right] \tag{B.68}$$

where $G_{0,n} = \frac{1}{n}\sum_i \delta_{\tau_i}$ is the empirical distribution of the τ's. However, (B.68) no longer holds in the heteroskedastic setting, and to adapt this argument, we need to additionally control the difference between $E\left[\|\tau_{\rho_n}^{(j)} - \tau_{\rho_n}^*\| \mid \tau_1, \ldots, \tau_n\right]$ and $E\left[\|\tau_{\rho_n}^{(j)} - \tau_{\rho_n}^*\|\right]$. The arguments of Jiang (2020) (p.2289) and Soloff *et al.* (2021) (Section C.3.3, arXiv:2109.03466v1) appear to use the Gaussian concentration of Lipschitz functions argument without the additional step.

Instead, we establish control of ζ_3 by observing that entries of $\tau_{\rho_n}^{(j)} - \tau_{\rho_n}^*$ are bounded and applying the convex Lipschitz concentration inequality. Since, like Soloff *et al.* (2021), we seek regret control in terms of mean-squared error, this argument applies to their setting as well. Jiang (2020), on the other hand, seeks regret control in terms of root-mean-squared error, and it is unclear if similar fixes apply.

■

Controlling ζ_4

Consider a change of variables where we let $w_i = z/\nu_i$ and $\lambda_i = \tau/\nu_i$. Let $G_{(i)}$ be the distribution of λ_i under G, where

$$G_{(i)}(d\lambda) = G(d\tau)$$

Then

$$f_{G,\nu_i}(z) = \int \frac{1}{\nu_i}\varphi\left(w_i - \lambda_i\right) G(d\tau) = \frac{1}{\nu_i}\int \varphi\left(w_i - \lambda_i\right) G_{(i)}(d\lambda_i) = \frac{1}{\nu_i} f_{G_{(i)},1}(w_i)$$

and

$$f'_{G,\nu_i}(z) = \frac{1}{\nu_i^2} f'_{G_{(i)},1}(w_i).$$

Hence,

$$\begin{aligned} E(\tau_{\rho_n}^{(j)} - \tau_{\rho_n}^*)^2 &= \nu_i^2 E\left(\frac{f'_{G_{ji},1}(w_i)}{f_{G_{ji},1}(w_i) \vee \rho_n} - \frac{f'_{G_{0i},1}(w_i)}{f_{G_{0i},1}(w_i) \vee \rho_n}\right)^2 \\ &\lesssim_{\mathcal{H}} \max\left((\log 1/\rho_n)^3, |\log h(f_{G_{ji},1}, f_{G_{0i},1})|\right) h^2(f_{G_{ji},1}, f_{G_{0i},1}) \\ &\qquad \text{(Theorems B.4.13 and B.6.10)} \\ &= \max\left((\log 1/\rho_n)^3, |\log h(f_{G_j,\nu_i}, f_{G_0,\nu_i})|\right) h^2(f_{G_j,\nu_i}, f_{G_0,\nu_i}) \\ &\qquad \text{(Hellinger distance is invariant to change-of-variables)} \end{aligned}$$

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

Hence,

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

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 B.4.1)

Therefore the first term dominates, and

$$\frac{1}{n}E[\zeta_4^2] \lesssim_{\mathcal{H}} (\log n)^3 \delta_n^2.$$

B.6.5 Controlling ξ_4

Lemma B.6.6. *Under the assumptions of Theorem B.6.2, in the proof of Theorem B.6.2,* $E\xi_4 \lesssim_{\mathcal{H}} \frac{1}{n}$.

Proof. Note that

$$\begin{aligned}E[(\theta^*_{i,\rho_n} - \theta_i^*)^2] &= \int \left(\nu_i^2 \frac{f'_{G_0,\nu_i}(z)}{f_{G_0,\nu_i}(z)}\right)^2 \left(1 - \frac{f_{G_0,\nu_i}}{f_{G_0,\nu_i} \vee \frac{\rho_n}{\nu_i}}\right)^2 f_{G_0,\nu_i}(z)\, dz \\ &\le E\left[\left(\nu_i^2 \frac{f'_{G_0,\nu_i}(z)}{f_{G_0,\nu_i}(z)}\right)^4\right]^{1/2} P\left[f_{G_0,\nu_i}(Z) < \rho_n/\nu_i\right]^{1/2} \qquad \text{(Cauchy–Schwarz)} \\ &\lesssim_{\mathcal{H}} \rho_n^{1/3}\,\mathrm{Var}(Z)^{1/6} \qquad \text{(Theorem B.4.16)} \\ &\lesssim_{\mathcal{H}} \frac{1}{n}.\end{aligned}$$

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

B.6.6 Auxiliary lemmas

Lemma B.6.7. *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 $[-\bar{M}_n, \bar{M}_n]$ where $\bar{M}_n = \max_i |\hat{Z}_i(\hat{\eta}) \vee 1|$. 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 2.3.2 and 2.3.3 holds.*
3. $\hat{s} \gtrsim_{\mathcal{H}} s_{\ell n}$ *for some fixed sequence* $s_{\ell n} > 0$.

Then

$$\left|\hat{\theta}_{i,\hat{G},\hat{\eta}} - \theta_i^*\right| \lesssim_{\mathcal{H}} \bar{s}_{\ell n}^{-2} \bar{Z}_n.$$

Moreover, the assumptions are satisfied by Assumptions 2.3.1 to 2.3.4 with $s_{\ell n} = s_{0\ell} \asymp 1$.

Proof. Observe that

$$\left|\hat{\theta}_{i,\hat{G}_n,\hat{\eta}} - \hat{\theta}_{i,G_0,\eta_0}\right| = \left|\frac{1}{\hat{s}_i}\frac{\hat{\nu}_i^2 f'_{\hat{G}_n,\hat{\nu}_i}(\hat{Z}_i)}{f_{\hat{G}_n,\hat{\nu}_i}(\hat{Z}_i)} - \frac{1}{s_{0i}}\frac{v_i^2 f'_{G_0\nu_i}(Z_i)}{f_{G_0,\nu_i}(Z_i)}\right|$$
$$\lesssim_{\mathcal{H}} s_{\ell n}^{-1}\bar{M}_n + \bar{Z}_n.$$

by the boundedness of $\hat{G}_n$ and Theorem B.4.22. 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,

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

□

Lemma B.6.8. *Let* $\bar{Z}_n = \max_i |Z_i| \vee 1$. *Under Assumption 2.3.2, for* $t > 1$

$$P(\bar{Z}_n > t) \le n\exp\left(-C_{A_0,\alpha,\nu_u} t^\alpha\right).$$

and

$$E[\bar{Z}_n^p] \lesssim_{p,\mathcal{H}} (\log n)^{p/\alpha}.$$

Moreover, if $M_n = (C_{\mathcal{H}} + 1)(C_{2,\mathcal{H}}^{-1}\log n)^{1/\alpha}$ *as in (B.13), then for all sufficiently large choices of* $C_{\mathcal{H}}$, $P(\bar{Z}_n > M_n) \le n^{-2}$.

Proof. The first claim is immediate under Theorem B.4.20 and a union bound.

The second claim follows from the observation that

$$E[\max_i(|Z_i| \vee 1)^p] \leq \left(\sum_i E[(|Z_i| \vee 1)^{pc}]\right)^{1/c} \leq n^{1/c} C_{\mathcal{H}}^p (pc)^{p/\alpha}.$$

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

$$P(Z_n > M_n) \leq \exp\left(\log n - C_{A_0,\alpha,\nu_u}(C_{\mathcal{H}}+1)^\alpha C_{2,\mathcal{H}}^{-1} \log n\right)$$

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 $P(Z_n > M_n) \leq e^{-2\log n} = n^{-2}$. □

Lemma B.6.9. *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* $t > 0$

$$P\left[\|W\| > E\|W\| + t\right] \leq e^{-t^2/2}.$$

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 separate convexity required in Theorem 6.10 of Boucheron *et al.* (2013). This checks all conditions and the claim follows by applying Theorem 6.10 of Boucheron *et al.* (2013). □

Lemma B.6.10. *Let $f_H = f_{H,1}$. Then, for $0 < \rho_n \le \frac{1}{\sqrt{2\pi e^2}}$,*

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

where we define the right-hand side to be zero if $H_1 = H_0$.

Proof. This claim is an intermediate step of Theorem 3 of Jiang and Zhang (2009). In (3.10) in Jiang and Zhang (2009), the left-hand side of this claim is defined as $r(f_{H_1}, \rho_n)$. Their subsequent calculation, which involves Lemma 1 of Jiang and Zhang (2009), proceeds to bound

$$r(f_{H_1}, \rho_n) \le 4e^2 h^2(f_{H_1}, f_{H_0}) \max\left(\varphi_+^6(\rho_n), 2a^2\right) + 2\varphi_+(\rho_n)\sqrt{2} h(f_{H_1}, f_{H_0}),$$

for $a^2 = \max\left(\varphi_+^2(\rho_n) + 1, |\log h^2\left(f_{H_1}, f_{H_0}\right)|\right)$. Collecting the powers on $h, \log h$ and using $\varphi_+(\rho_n) \lesssim \sqrt{\log(1/\rho_n)}$ proves the claim. □

B.7 Estimating η_0 by local linear regression

In this section, we verify that estimating η_0 by local linear regression satisfies the conditions we require for the nuisance estimators, when the true nuisance parameters belong to a Hölder class of order $p = 2$: $m_0(\sigma), s_0(\sigma) \in C^2_{A_1}([\sigma_\ell, \sigma_u])$.

In our empirical application, we estimate m_0, s_0 by nonparametrically regressing Y_i on $x_i \equiv \log_{10}(\sigma_i)$.[18] 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 $(E[Y \mid x], \mathrm{Var}(Y \mid x) - \sigma^2(x))$, 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 $m_0(x) = E[Y_i \mid x_i = x], s_0(x) = \mathrm{Var}(\theta_i \mid x_i = x)$. 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

[18] Correspondingly, let $\sigma(x) = 10^x$.

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:[19]

$$\hat{h}_{n,m} \leftarrow (\hat{h}_{n,m} \vee C_h^{-1} n^{-1/5}) \wedge C_h n^{-1/5}.$$

(LLR-3) Using $\hat{h}_{n,m}$, estimate m_0 with the local linear regression estimator $\hat{m}_{\text{raw}}$ under kernel $K(\cdot)$ and bandwidth $\hat{h}_{n,m}$.

(LLR-4) Project the resulting estimator $\hat{m}$ to the Hölder class $C^2_{A_3}([0,1])$:

$$\hat{m} \in \operatorname*{arg\,min}_{m \in C^2_{A_3}([0,1])} \|m - \hat{m}_{\text{raw}}\|_\infty.$$

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 $v(x) = E[R_i^2 \mid X = x]$ 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

$$p_n = \frac{1}{n} \sum_{i=1}^n \frac{1}{\sum_{j=1}^n \ell_i^2(x_j, \hat{h}_{n,s})}. \tag{B.69}$$

[19] We use the ← notation to reassign a variable so that we can reduce notation clutter.

(LLR-10) Truncate the estimated conditional standard deviation:

$$\hat{s}_{\text{raw}}(x) = \sqrt{\hat{v}(x) - \sigma^2(x)} \vee \sqrt{\frac{2}{p_n + 2}\hat{v}(x)}. \tag{B.70}$$

(LLR-11) Finally, project the resulting estimate to the Hölder class as in (LLR-4):

$$\hat{s}(x) \in \operatorname*{arg\,min}_{\substack{s \in C^2_{A_3}([0,1]) \\ s^2(\cdot) \geq \frac{2}{p_n+2} \min_i \sigma_i^2}} \|s - \hat{s}_{\text{raw}}\|_\infty.$$

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 2.3.4(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 (B.70). 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 B.7.1 (The truncation rule in (B.70))**.** The truncation rule in (B.70) is an ad hoc adjustment without affecting asymptotic performance.[20] It is based on a literature on the estimation of non-central χ^2 parameters (Kubokawa *et al.*, 1993). Specifically, let $U_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\lambda_i, 1)$ and let $V = \sum_{i=1}^p U_i^2$ be a noncentral χ^2 random variable with p degrees of freedom and noncentrality parameter $\lambda = \sum_{i=1}^p \lambda_i^2$. The UMVUE for λ is $V - p$, which is dominated by its positive part $(V-p)_+$. Kubokawa *et al.* (1993) derive a class of estimators of the form $V - \phi(V; p)$ that dominate $(V-p)_+$ in squared error risk. An estimator in this class is $(V-p) \vee \frac{2}{p+2}V$.[21]

[20] Indeed, since we already assumed that the true conditional variance $s_0(x) > s_\ell$, we can truncate by any vanishing sequence. Given any vanishing sequence, eventually it is lower than s_ℓ, and eventually $|\hat{s} - s_0|$ is small enough for the truncation to not bind. This is, in some sense, silly, since finite sample performance is likely affected if we truncate by, say, $\frac{1}{\log\log n}$, reflected in a large constant in the corresponding rate expression. Our following argument assumes that the truncation of order $O(n^{-4/5})$. Doing so is likely to achieve a smaller constant in the rate expression, despite not mattering asymptotically.

[21] Though, since neither $(V-p)_+$ and $(V-p) \vee \frac{2}{p+2}V$ is differentiable in V, they are not admissible.

This setting is loosely connected to ours. Suppose m_0 is known, and we were using a Nadaraya–Watson estimator with uniform kernel. Then, for a given evaluation point x_0, we would be averaging nearby R_i^2's. Each R_i is conditionally Gaussian, $R_i \mid (\theta_i, \sigma_i) \sim \mathcal{N}(\theta_i - m_0(\sigma_i), \sigma_i^2)$ with approximately equal variance $\sigma_i^2 \approx \sigma(x_0)^2$. If there happens to be p_0 R_i^2's that we are averaging, the Nadaraya–Watson estimator is of the form

$$\hat{v}(x_0) = \frac{\sigma(x_0)^2}{p_0} \sum_{i=1}^{p} \left(\frac{R_i}{\sigma(x_0)} \right)^2$$

Conditional on σ_i^2, θ_i, the quantity $\sum_{i=1}^{p} \left(\frac{R_i}{\sigma(x_0)} \right)^2$ is (approximately) noncentral χ^2 with p degrees of freedom and noncentrality parameter

$$\lambda = \sum_{i=1}^{p_0} \left(\frac{\theta_i - m_0(x_i)}{\sigma(x_0)} \right)^2$$

Therefore, correspondingly, applying the truncation rule from Kubokawa *et al.* (1993), an estimator for the sample variance of θ_i, $\frac{1}{p_0} \sum_{i=1}^{p_0} (\theta_i - m_0(x_i))^2$, is

$$\left(\hat{v}(x_0) - \sigma^2(x_0)\right) \vee \frac{2}{p_0 + 2} \hat{v}(x_0).$$

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

$$\operatorname{Var}\left(\sum_i \ell_i(x) V_i \right) = \sum_{i=1}^{n} \ell_i^2(x).$$

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 (B.69) and use it as a candidate for p. ■

The goal in this section is to control the following probability as a function of $t > 0$

$$P\left(\|\hat{\eta} - \eta_0\|_\infty > C_{\mathcal{H}} t n^{-2/5} (\log n)^\beta \right)$$

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 B.7.1. *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 B.7.1 is automatically satisfied by the projection steps (LLR-3) and (LLR-7).

Assumption B.7.2. *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 $n \ge n_0$, 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 $\lambda_{\min}(B_{nx}(\tilde{h})) \ge \lambda_0$.*

2. *There exists a real number $a_0 > 0$ such that for any interval $I \subset [0, 1]$ and all $n \ge 1$,*

$$\frac{1}{n} \sum_{i=1}^n \mathbb{1}(x_i \in I) \le a_0 \left(\lambda(I) \vee \frac{1}{n}\right)$$

   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 (B.69) 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 B.7.2(1–3) is nearly the same as Assumption (LP) in Tsybakov (2008). The only difference is that Assumption B.7.2(1) requires the lower bound λ_0 to hold uniformly over a range of bandwidth choices, relative to LP-1 in Tsybakov (2008), which requires λ_0 to hold for some deterministic sequence h_n. This is a mild strengthening of LP-1: Note that if x_i are drawn from a Lipschitz-continuous, everywhere-positive density $f(x)$, then for $h \to 0, nh \to \infty$,

$$B_{nx}(h) \approx \int K(t)u(t)u(t)' f(x)\, dt \succeq \int K(t)u(t)u(t)'\, dt \left(\min_{x \in [0,1]} f(x) \right)$$

where $\succ$ denotes the positive-definite matrix order. Thus the minimum eigenvalue of $B_{nx}(h)$ should be positive irrespective of x and h. See, also, Lemma 1.5 in Tsybakov (2008).

Assumption B.7.2(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 B.7.2 is satisfied by regular design points $x_i = i/n$.

Assumption B.7.3. *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 B.7.3 is satisfied for a wide range of kernels, e.g. the Epanechnikov kernel. By Lemma 7.22 in Sen (2018), reproduced as Theorem B.7.6 below, so long as the function $t \mapsto t^{k-1} K(t)$

is bounded (assumed in Assumption B.7.2(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 Theorem B.7.4 on the bias and variance of local linear regression estimators. Theorem B.7.4 is uniform in both the evaluation point x and the bandwidth h, as long as the latter converges at the optimal rate.

Theorem B.7.2. *Suppose the conditional distribution $\theta_i \mid \sigma_i$ and the design points $\sigma_{1:n}$ satisfy Assumptions 2.3.2, 2.3.3 and B.7.2. Moreover, suppose m_0, s_0 satisfies Assumption 2.3.4(1) with $p = 2$. Suppose the kernel $K(\cdot)$ satisfies Assumption B.7.3. Let $\hat{m}, \hat{s}$ denote the estimators computed by (LLR-1) through (LLR-11). Then:*

1. *$P\left(\hat{m}, \hat{s} \in C^2_{A_3}([0,1])\right) = 1$*

2. *For some C depending only on the parameters in the assumptions, for all $n \geq 7$ and $t > 1$,*

$$P\left(\max\left(\|\hat{m} - m_0\|_\infty, \|\hat{s} - s_0\|\right) \geq Ctn^{-\frac{2}{5}}(\log n)^{1+2/\alpha}\right) \leq \frac{1}{n^{10}t^2}. \tag{B.71}$$

3. *For some c depending only on the parameters in the assumptions, for all $n \geq 7$,*

$$P\left(\frac{c}{n} \leq \hat{s}\right) = 1.$$

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

$$P\left(\max\left(\|\hat{m}_{\text{raw}} - m_0\|_\infty, \|\hat{s}_{\text{raw}} - s_0\|_\infty\right) > Ctn^{-2/5}(\log n)^{\beta}\right) \leq \frac{1}{n^{10}t^2}$$

for some C and $\beta = 1 + 2/\alpha$.

Let $Y_i = m_0(x_i) + \xi_i$ where $\xi_i = \theta_i - m_0(x_i) + (Y_i - \theta_i)$. Note that we have simultaneous moment control for ξ_i:

$$\max_i E[|\xi_i|^p]^{1/p} \lesssim p^{1/\alpha}$$

where α is the constant in Assumption 2.3.2. Therefore, we can apply Theorem B.7.4 to obtain

$$P\left(\|\hat{m}_{\text{raw}} - m_0\|_\infty > Ctn^{-2/5}(\log n)^{1+1/\alpha}\right) \leq \frac{1}{2n^{10}t^2}$$

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

$$|\hat{s}^2_{\text{raw}} - s_0^2| = |\hat{s}_{\text{raw}} - s_0|(\hat{s}_{\text{raw}} + s_0) \geq s_{0\ell}|\hat{s}_{\text{raw}} - s_0|.$$

Also observe that for a positive f_0,

$$|\hat{f} \vee g - f_0| \leq |\hat{f} - f_0| \vee |g|.$$

As a result, it suffices to control the upper bound in

$$\begin{aligned}
\|\hat{s}_{\text{raw}} - s_0\|_\infty &\leq \frac{1}{s_{0\ell}}\left(\|\hat{v} - v_0\|_\infty \vee \left(\frac{2}{2+p_n}\hat{v}\right)\right) && (v_0(x) \equiv \operatorname{Var}(Y_i \mid x_i = x)) \\
&\lesssim \|\hat{v} - v_0\|_\infty \vee \frac{\|\hat{v} - v_0\|_\infty + \|v_0\|_\infty}{2 + n^{4/5}} && (\text{Assumption B.7.2}) \\
&\lesssim \|\hat{v} - v_0\|_\infty && \text{(B.72)}
\end{aligned}$$

Now, observe that $\hat{R}_i^2 = R_i^2 + (m_0 - \hat{m})^2 - 2(m_0 - \hat{m})\xi_i$. Hence,

$$\begin{aligned}
|\hat{v}(x) - v_0(x)| &\leq \left|\sum_{i=1}^n \ell_i(x, \hat{h}_{n,s})R_i^2 - v_0(x)\right| \\
&\quad + \left\{\|m_0 - \hat{m}\|_\infty^2 + 2\|m_0 - \hat{m}\|_\infty \left(\max_{i\in[n]} |\xi_i|\right)\right\} \sum_{i=1}^n |\ell_i(x, \hat{h}_{n,s})| \\
&\leq \left|\sum_{i=1}^n \ell_i(x, \hat{h}_{n,s})R_i^2 - v_0(x)\right| + C\left\{\|m_0 - \hat{m}\|_\infty^2 + 2\|m_0 - \hat{m}\|_\infty \left(\max_{i\in[n]} |\xi_i|\right)\right\}.
\end{aligned} \tag{B.73}$$

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 ($\tilde{\alpha} = \alpha/2$):

$$\max_i (E|\tilde{\xi}_i|^p)^{1/p} \lesssim p^{2/\alpha}.$$

Thus, applying Theorem B.7.4 and taking care to plug in $\tilde{\xi}, \tilde{\alpha}$, we can bound the first term in (B.73)

$$P\left(\left\| \sum_{i=1}^n \ell_i(x, \hat{h}_{n,s}) R_i^2 - v_0(x) \right\|_\infty \ge Ctn^{-2/5}(\log n)^{1+2/\alpha} \right) \le \frac{1}{4n^{10}t^2}.$$

Note that by an application of Theorem B.6.8, for any $a, b > 0$, we have that

$$P\left(\max_i |\xi_i| > C(a,b)t(\log n)^{1/\alpha} \right) < an^{-b}e^{-t^2}$$

As a result, the second term in (B.73) admits

$$P\left(\|m_0 - \hat{m}\|_\infty^2 + 2\|m_0 - \hat{m}\|_\infty \left(\max_{i \in [n]} |\xi_i| \right) > Ctn^{-2/5}(\log n)^{1+2/\alpha} \right) \le \frac{1}{4n^{10}t^2}$$

Finally, putting these bounds together, we have that

$$P\left(\|\hat{v} - v_0\|_\infty > Ctn^{-2/5}(\log n)^{1+2/\alpha} \right) \le \frac{1}{2n^{10}t^2},$$

where the same bound (with a different constant) holds for $\hat{s}_{\text{raw}}$ by (B.72).

Combining the bounds for $\hat{m}$ and $\hat{s}$, we obtain (B.71). This concludes the proof. □

Theorem B.7.3. *Under the assumptions of Theorem B.7.2, let $\hat{\eta} = (\hat{m}, \hat{s})$ denote estimators computed by (LLR-1) through (LLR-11). Then,*

$$E\left[\text{Regret}(\hat{G}_n, \hat{\eta}) \right] \lesssim n^{-2/5}(\log n)^{1+2/\alpha}.$$

Proof. Recall the event A_n in (B.6) 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

$$E\left[\text{Regret}(\hat{G}_n, \hat{\eta}) \right] = E\left[\text{Regret}(\hat{G}_n, \hat{\eta}) \mathbb{1}(\tilde{A}_n) \right] + E\left[\text{Regret}(\hat{G}_n, \hat{\eta}) \mathbb{1}(\tilde{A}_n^{\mathrm{C}}) \right].$$

Note that, for all sufficiently large $n > N$, such that N depends only on C_1, β, s_ℓ, s_u, the event

A_n implies $\{s_{0\ell}/2 \leq \hat{s} \leq 2s_{0u}\}$ and hence $A_n = \tilde{A}_n$. Thus, by Theorem B.7.2, for all sufficiently large n, on the event A_n, statements analogous to Assumption 2.3.4(2–4) hold for the estimator $\hat{\eta}$. As a result, we may apply Theorem B.6.2, *mutatis mutandis*, to obtain that

$$E\left[\operatorname{Regret}(\hat{G}_n, \hat{\eta})\mathbb{1}(\tilde{A}_n)\right] \lesssim n^{-4/5}(\log n)^{\frac{2+\alpha}{\alpha}+3+2\beta}$$

for all sufficiently large choices of C_1, C_2.

To control $E\left[\operatorname{Regret}(\hat{G}_n, \hat{\eta})\mathbb{1}(\tilde{A}_n^{\mathrm{C}})\right]$, we observe that under Theorem B.6.7 and Theorem B.7.2(1 and 3), we have that almost surely,

$$\operatorname{Regret}(\hat{G}_n, \hat{\eta}) \lesssim n^4 \bar{Z}_n^2.$$

Hence, by Cauchy–Schwarz as in Theorem B.6.1,

$$E\left[\operatorname{Regret}(\hat{G}_n, \hat{\eta})\mathbb{1}(\tilde{A}_n^{\mathrm{C}})\right] \lesssim P(\tilde{A}_n^{\mathrm{C}})^{1/2} n^4 (\log n)^{2/\alpha},$$

where we apply Theorem B.6.8 to bound $E[\bar{Z}_n^4]$.

For all sufficiently large $n > N$,

$$P(A_n^{\mathrm{C}}) = P(\tilde{A}_n^{\mathrm{C}}) \leq P(\bar{Z}_n > M_n) + P(\|\hat{\eta} - \eta_0\|_\infty > \Delta_n).$$

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 B.7.2 to bound $\|\hat{\eta} - \eta_0\|_\infty$. We can apply Theorem B.6.8 to bound $P(\bar{Z}_n > M_n)$.

As a result, we would obtain

$$E\left[\operatorname{Regret}(\hat{G}_n, \hat{\eta})\mathbb{1}(\tilde{A}_n^{\mathrm{C}})\right] \lesssim \frac{1}{n}(\log n)^{2/\alpha}$$

for all sufficiently large n.

Since $E[\operatorname{Regret}(\hat{G}_n, \hat{\eta})] \lesssim n^4(\log n)^{2/\alpha}$ is finite for all n, at the cost of enlarging the implicit constant, we have the result of the theorem holding for all n. □

B.7.1 Auxiliary lemmas

Proposition B.7.4. *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: $f_0 \in C^2_L([0,1])$ for some $L > 0$. Suppose that the design points satisfy Assumption B.7.2 and the (possibly data-driven) bandwidths h_n satisfy Assumption B.7.1. Assume the kernel additionally satisfies Assumption B.7.3.*

Assume that the residuals ξ_i are mean zero, and there exists a constant $A_\xi > 0, \alpha > 0$ such that

$$\max_{i=1,\ldots,n} (E[|\xi_i|^p])^{1/p} \leq A_\xi p^{1/\alpha}$$

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 B.7.1. 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 $t \geq 1$*

$$P\left(\sup_{x\in[0,1],h\in H_n} |v(x,h)| > C(a,b)\cdot t \cdot (\log n)^{1+1/\alpha} n^{-2/5}\right) \leq an^{-b}\frac{1}{t^2}.$$

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 $t \geq 1$,*

$$P\left(\|\hat{f} - f_0\|_\infty > C(a,b)t(\log n)^{1+1/\alpha} n^{-2/5}\right) \leq an^{-b}\frac{1}{t^2}.$$

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

$$\sup_{x\in[0,1]} |b(x,h_n)| \le Ch_n^2,$$

which is uniformly bounded by $Cn^{-2/5}$ by Assumption B.7.1. Hence

$$\sup_{x\in[0,1],h\in H_n} |b(x,h)| \lesssim n^{-2/5}.$$

2. Let M be a truncation point to be defined. Let

$$\xi_{i,<M} = \xi_i \mathbb{1}(|\xi_i| \le M) - E[\xi_i \mathbb{1}(|\xi_i| \le M)] \quad \xi_{i,>M} = \xi_i \mathbb{1}(|\xi_i| > M) - E[\xi_i \mathbb{1}(|\xi_i| > M)]$$

be truncated and demeaned variables. Note that

$$\xi_i = \xi_{i,<M} + \xi_{i,>M}.$$

First, let $V_{1n}(x,h_n) = \sum_{i=1}^n \ell_i(x,h_n)\xi_{i,>M}$. Note that by Cauchy–Schwarz, uniformly over x, h_n,

$$\begin{aligned} V_{1n}^2 &\le \sum_{i=1}^n \ell_i(x,h_n)^2 \sum_{i=1}^n \xi_{i,>M}^2 \\ &\lesssim \frac{1}{h_n^2}\frac{1}{n}\sum_{i=1}^n \xi_{i,>M}^2 && \text{(Lemma 1.3(i) in Tsybakov (2008) shows that } |\ell_i(x,h_n)| \le \tfrac{C}{nh_n}) \\ &\lesssim n^{2/5}\frac{1}{n}\sum_{i=1}^n \xi_{i,>M}^2 \end{aligned}$$

Now, for some C related to the implicit constant in the above display,

$$P\left(\sup_{x\in[0,1],h_n\in H_n} V_{1n}^2(x,h_n) > Ct^2\right) \le P\left(\frac{1}{n}\sum_{i=1}^n \xi_{i,>M}^2 > t^2 n^{-2/5}\right) \le \frac{\max_i E\xi_{i,>M}^2}{t^2} n^{2/5}.$$

(Markov's inequality)

We note that by Cauchy–Schwarz,

$$E[\xi_{i,>M}^2] \le \sqrt{E[\xi_i^4]}\sqrt{P(|\xi_i| > M)} \lesssim \sqrt{P(|\xi_i| > M)} \le \exp(-cM^\alpha) \qquad \text{(Theorem B.4.20)}$$

where c depends on A_ξ. Hence, for a potentially different constant C,

$$P\left(\sup_{x\in[0,1],h_n\in H_n}|V_{1n}(x,h_n)|>Ct\right)\leq\exp\left(-cM^\alpha-2\log t+\frac{2}{5}\log n\right). \tag{B.74}$$

Next, consider the process

$$\begin{aligned}V_{2n}(x,h_n)&=\sum_{i=1}^n\ell_i(x,h_n)\xi_{i,<M}\\&=\frac{1}{nh_n}\sum_{i=1}^n\underbrace{u(0)'B_{nx}^{-1}\begin{bmatrix}1\\0\end{bmatrix}}_{A_1(x,h_n)}K\left(\frac{x_i-x}{h_n}\right)\xi_{i,<M}\\&\quad+\frac{1}{nh_n}\sum_{i=1}^n\underbrace{u(0)'B_{nx}^{-1}\begin{bmatrix}0\\1\end{bmatrix}}_{A_2(x,h_n)}K\left(\frac{x_i-x}{h_n}\right)\left(\frac{x_i-x}{h_n}\right)\xi_{i,<M}\\&\equiv\frac{A_1(x,h_n)}{h_n}\frac{1}{n}\sum_{i=1}^nK\left(\frac{x_i-x}{h_n}\right)\xi_{i,<M}+\frac{A_2(x,h_n)}{h_n}\frac{1}{n}\sum_{i=1}^nK\left(\frac{x_i-x}{h_n}\right)\left(\frac{x_i-x}{h_n}\right)\xi_{i,<M}.\end{aligned}$$

Note that, by Assumption B.7.2(1), uniformly over $x\in[0,1]$ and $h_n\in H_n$,

$$|A_k(x,h_n)|\leq\|u(0)'B_{nx}^{-1}\|\leq\frac{1}{\lambda_0}.$$

By triangle inequality,

$$\begin{aligned}V_{2n}(x,h_n)&\lesssim\frac{1}{h_n}\left|\frac{1}{n}\sum_{i=1}^nK\left(\frac{x_i-x}{h_n}\right)\xi_{i,<M}\right|+\frac{1}{h_n}\left|\frac{1}{n}\sum_{i=1}^nK\left(\frac{x_i-x}{h_n}\right)\left(\frac{x_i-x}{h_n}\right)\xi_{i,<M}\right|\\&\equiv\frac{1}{\sqrt{n}h_n}V_{2n,1}(x,h_n)+\frac{1}{\sqrt{n}h_n}V_{2n,2}(x,h_n).\end{aligned}$$

We will aim to control the ψ_2-norm of the left-hand side. Note that it suffices to control the ψ_2-norm of both terms on the right-hand side:

$$\left\|\sup_{x\in[0,1],h_n\in H_n}|V_{2n}(x,h_n)|\right\|_{\psi_2}\lesssim\frac{1}{\sqrt{n}h_n}\max_{k=1,2}\left(\left\|\sup_{x\in[0,1],h_n\in H_n}|V_{2n,k}(x,h_n)|\right\|_{\psi_2}\right).$$

The above display follows from replacing the sum with two times the maximum and Lemma 2.2.2 in van der Vaart and Wellner (1996).

We will do so by applying Theorem B.7.5. The analogue of f in Theorem B.7.5 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

$$f(t; x, h) \leq \mathbb{1}(|t - x| \leq h) K_0$$

and thus the diameter of $\mathcal{F}_k$ is at most

$$\sup_{A \subset [0,1] : \lambda(A) \leq 4C_h n^{-1/5}} K_0 \sqrt{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i \in A)} \lesssim n^{-1/10}$$

by Assumption B.7.2(2). Therefore, by Assumption B.7.3, we apply Theorem B.7.5 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

$$\left\| \sup_{x \in [0,1], h \in H_n} |V_{2n}(x,h)| \right\|_{\psi_2} \lesssim \frac{1}{\sqrt{n} h_n n^{1/10}} M \sqrt{\log n} \lesssim n^{-2/5} M \sqrt{\log n}. \tag{B.75}$$

Putting things together, we can choose $M = (c_m \log n)^{1/\alpha}$ for sufficiently large c_m so that by (B.74),

$$P\left(\sup_{x \in [0,1], h \in H_n} |V_{1n}(x,h)| > Ctn^{-2/5} \right) \leq \frac{a}{2} n^{-b} \frac{1}{t^2},$$

where c_m depends on a, b. The bound (B.75) in turns shows that

$$P\left(\sup_{x \in [0,1], h_n \in H_n} |V_{2n}(x,h_n)| > C(a,b) t (\log n)^{\frac{2+\alpha}{2\alpha}} n^{-2/5} \right) \leq 2e^{-t^2}$$

Taking $t = \sqrt{b \log n + \log(a/4)s}$ gives

$$P\left(\sup_{x \in [0,1], h_n \in H_n} |V_{2n}(x,h_n)| > C(a,b) s (\log n)^{1+1/\alpha} n^{-2/5} e^{-s^2} \right) \leq \frac{a}{2} n^{-b} e^{-s^2} < \frac{a}{2} n^{-b} \frac{1}{s^2}$$

for all $s > 1$.

Therefore, combining the two bounds,

$$P\left(\sup_{x\in[0,1],h_n\in H_n}|v(x,h_x)| > C(a,b)t(\log n)^{1+1/\alpha}n^{-2/5}\right) \le an^{-b}\frac{1}{t^2}.$$

□

Lemma B.7.5. *Suppose ξ_i are bounded by $M \ge 1$ and mean zero. Consider the process*

$$V_n(f) = \frac{1}{\sqrt{n}}\sum_{i=1}^n f(x_i)\xi_i$$

over a class of real-valued functions $f \in \mathcal{F}$ and evaluation points $x_1, \ldots, x_n \in [0,1]$. Define the seminorm $\|\cdot\|_n$ relative to $x_1, \ldots, x_n$ by

$$\|f\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^n f(x_i)^2}.$$

Suppose $0 \in \mathcal{F}$ and $\mathcal{F}$ has polynomial covering numbers:

$$N(\epsilon, \mathcal{F}, \|\cdot\|_n) \le C(1/\epsilon)^V \quad \epsilon \in [0,1]$$

where $C, V > 0$ depend solely on $\mathcal{F}$. Then

$$\left\|\sup_{f\in\mathcal{F}}|V_n(f)|\right\|_{\psi_2} \lesssim M\mathrm{diam}(\mathcal{F})\sqrt{\log(1/\mathrm{diam}(\mathcal{F}))},$$

where $\mathrm{diam}(\mathcal{F}) = \sup_{f_1,f_2\in\mathcal{F}}\|f_1 - f_2\|_n$.

Proof. The process $V_n(f)$ has subgaussian increments with respect to $\|\cdot\|_n$:

$$\|V_n(f_1) - V_n(f_2)\|_{\psi_2} \lesssim M\|f_1 - f_2\|_n.$$

Hence, by Dudley's chaining argument (e.g. Corollary 2.2.5 in van der Vaart and Wellner (1996)), for some fixed $f_0 \in \mathcal{F}$,

$$\left\|\sup_f V_n(f)\right\|_{\psi_2} \le \|V_n(f_0)\|_{\psi_2} + CM\int_0^{\mathrm{diam}(\mathcal{F})}\sqrt{\log N(\delta, \mathcal{F}, \|\cdot\|_n)}\,d\delta.$$

Note that (i) the metric entropy integral is bounded by $C\mathrm{diam}(\mathcal{F})\sqrt{\log(1/\mathrm{diam}(\mathcal{F}))}$, and (ii) for a fixed f_0, $\|V_n(f_0)\|_{\psi_2} \lesssim \|f_0\|_n M \leq \mathrm{diam}(\mathcal{F})M$ since $0 \in \mathcal{F}$. Therefore,

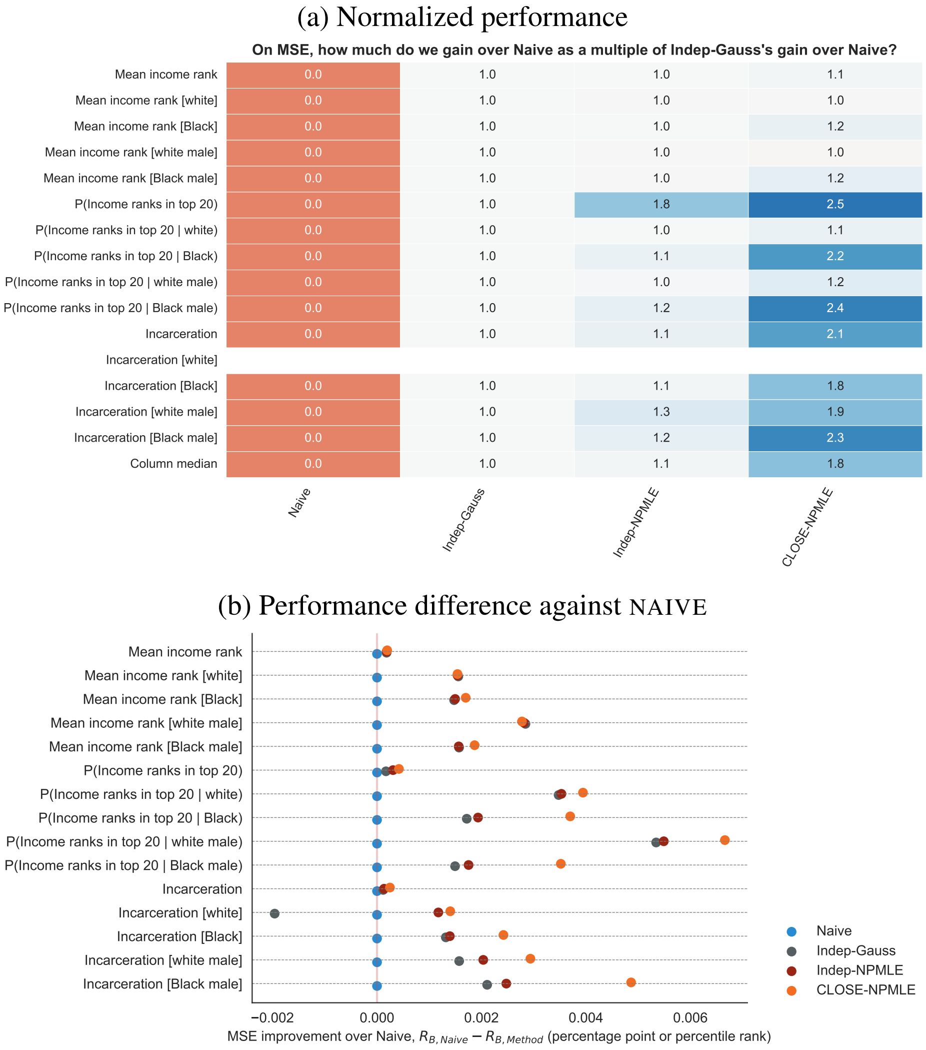
$$\left\|\sup_f V_n(f)\right\|_{\psi_2} \lesssim M\mathrm{diam}(\mathcal{F})\sqrt{\log(1/\mathrm{diam}(\mathcal{F}))}.$$

□

Lemma B.7.6 (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.



(a) Normalized performance

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

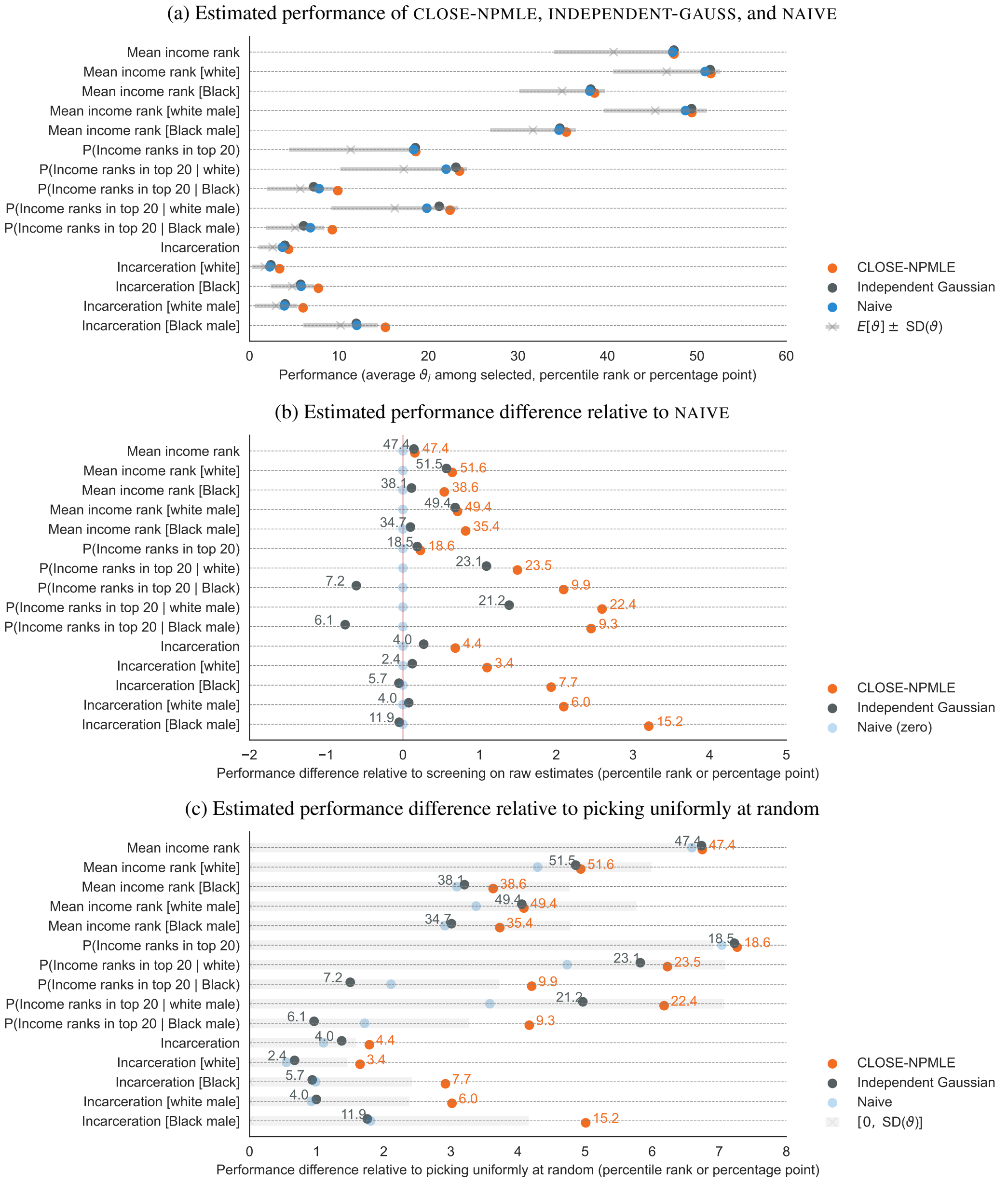
	Naive	Indep-Gauss	Indep-NPMLE	CLOSE-NPMLE
Mean income rank	0.0	1.0	1.0	1.1
Mean income rank [white]	0.0	1.0	1.0	1.0
Mean income rank [Black]	0.0	1.0	1.0	1.2
Mean income rank [white male]	0.0	1.0	1.0	1.0
Mean income rank [Black male]	0.0	1.0	1.0	1.2
P(Income ranks in top 20)	0.0	1.0	1.8	2.5
P(Income ranks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

(b) Performance difference against NAIVE

Mean income rank
Mean income rank [white]
Mean income rank [Black]
Mean income rank [white male]
Mean income rank [Black male]
P(Income ranks in top 20)
P(Income ranks in top 20 | white)
P(Income ranks in top 20 | Black)
P(Income ranks in top 20 | white male)
P(Income ranks in top 20 | Black male)
Incarceration
Incarceration [white]
Incarceration [Black]
Incarceration [white male]
Incarceration [Black male]
−0.002
0.000
0.002
0.004
0.006
MSE improvement over Naive, $R_{B,Naive} - R_{B,Method}$ (percentage point or percentile rank)
Naive
Indep-Gauss
Indep-NPMLE
CLOSE-NPMLE

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

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

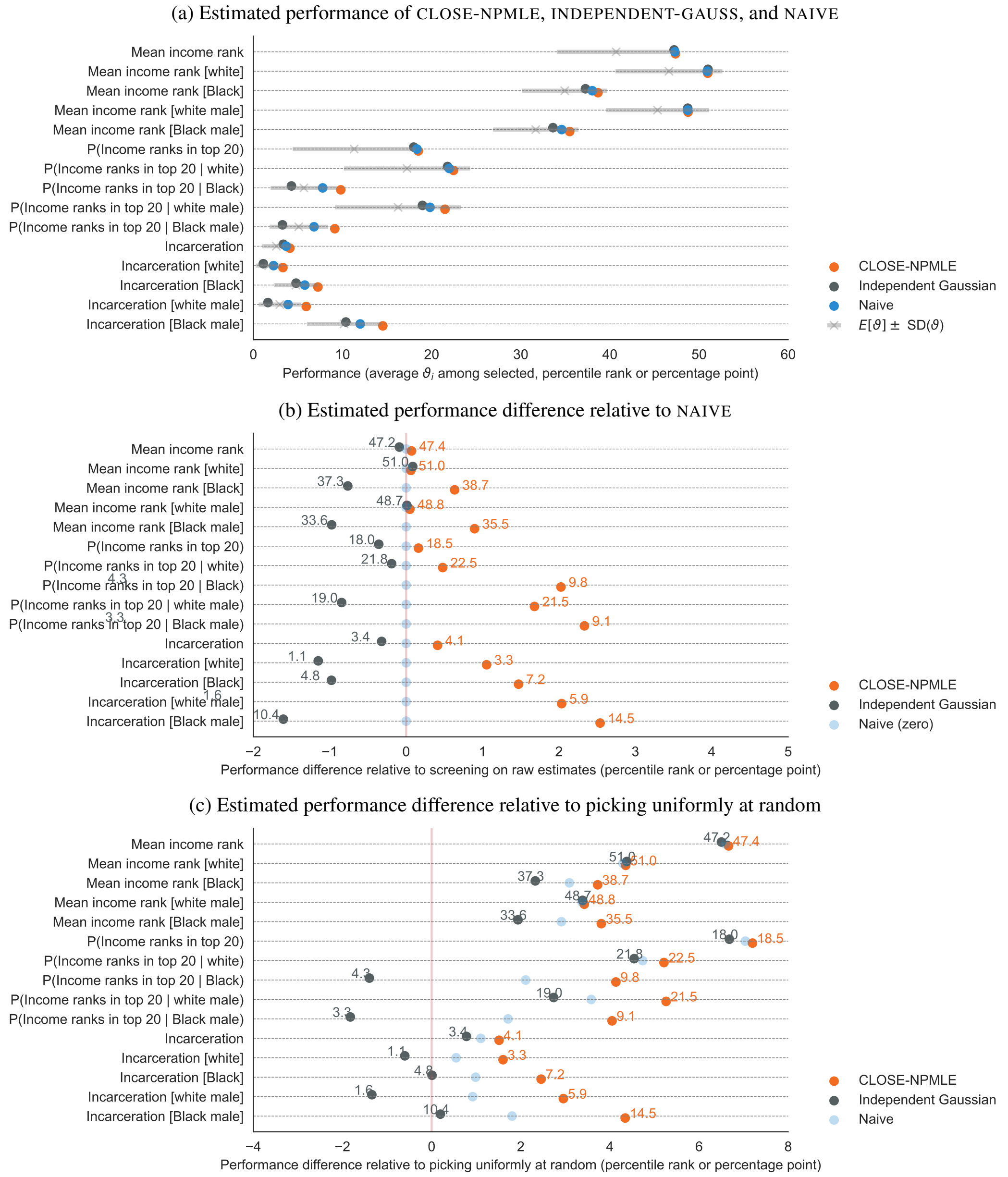

(a) Estimated performance of CLOSE-NPMLE, INDEPENDENT-GAUSS, and NAIVE

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

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

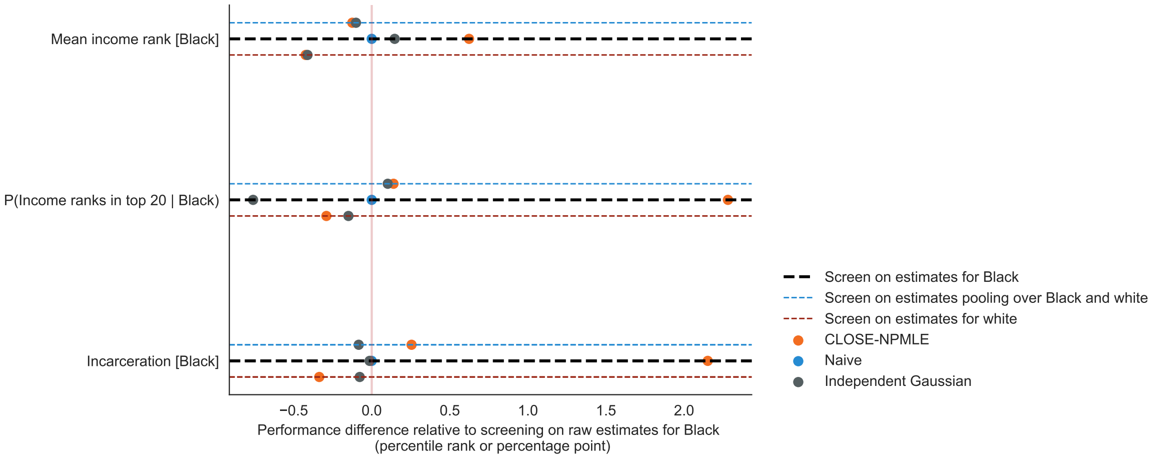

(a) Estimated performance of CLOSE-NPMLE, INDEPENDENT-GAUSS, and NAIVE

(b) Estimated performance difference relative to NAIVE

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

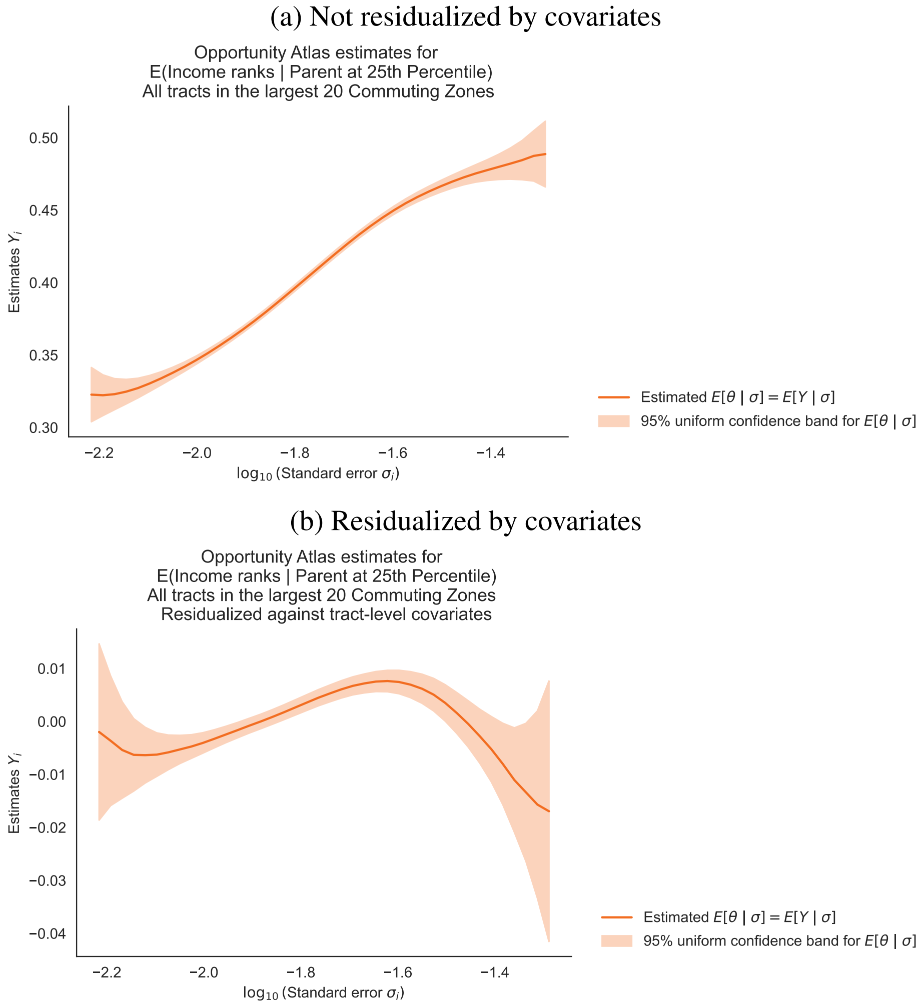
Notes. These figures show the estimated performance of various decision rules over 100 coupled bootstrap draws. There are no covariates to residualize against. Performance is measured as the mean ϑ_i among selected Census tracts. All decision rules select the top third of Census tracts within each Commuting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

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



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

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


(a) Not residualized by covariates

(b) Residualized by covariates

Notes. This figure shows the estimated $E[\theta \mid \sigma]$ for mean income rank, pooling over all demographic groups. This is the measure of economic mobility used by Bergman *et al.* (2023). The estimation and the confidence band procedures are the same as those in Figure 2.1. In panel (a), θ_i, Y_i are defined as unresidualized measures of mean income rank. In panel (b), we treat θ_i, Y_i as residualized against a vector of tract-level covariates as specified in Section B.2.3. □

Figure B.8: Estimated $E[\theta \mid \sigma]$ for mean income rank among those with parents at the 25th percentile

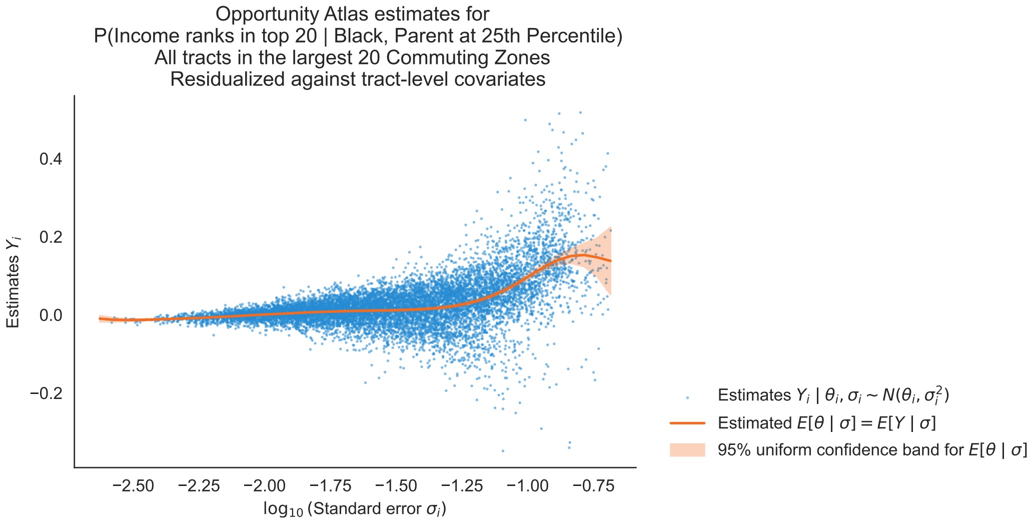


Figure B.9: The analogue of Figure 2.1 where Y_i, θ_i are treated as residualized against a vector of covariates as specified in Section B.2.3.

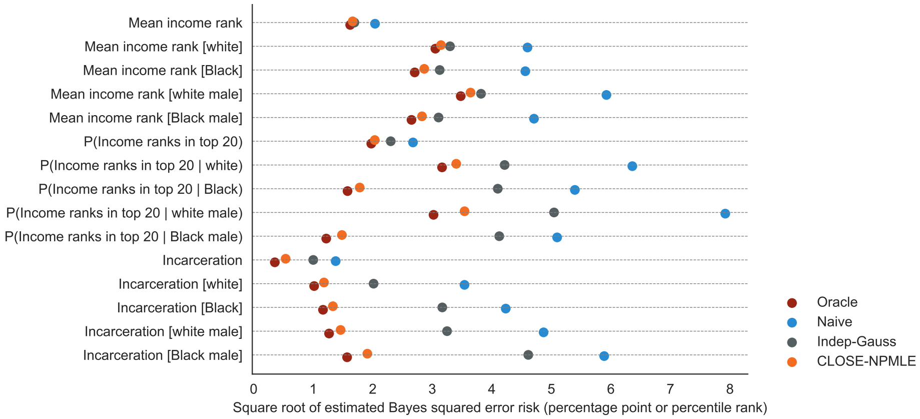


Figure B.10: Absolute mean-squared error risk of key methods for the calibrated simulation in Figure 2.4.

Appendix C

Appendix to Chapter 3

C.1 Proofs of results in the main text

C.1.1 Proof of Theorem 3.2.1

Proposition C.1.1. *Suppose that:*

1. *(The function m is continuous and increasing) $m : [0, \infty) \to \mathbb{R}$ is a continuous, weakly increasing function.*

2. *(The function m behaves like $\log$ for large values) $m(y)/\log(y) \to 1$ as $y \to \infty$.*

3. *(Treatment affects the extensive margin) $P(Y(1) = 0) \neq P(Y(0) = 0)$.*

4. *(Finite expectations) $E_{P_{Y(d)}}[|\log(Y(d))| \mid Y(d) > 0] < \infty$ for $d = 0, 1$.*[1]

Then, for every $\theta^* \in (0, \infty)$, there exists an $a > 0$ such that $|\theta(a)| = \theta^*$. In particular, $\theta(a)$ is continuous with $\theta(a) \to 0$ as $a \to 0$ and $|\theta(a)| \to \infty$ as $a \to \infty$.

Proof. Note that $\theta(0) = E_P[m(0)] - E_P[m(0)] = 0$. Additionally, Theorem C.2.1 below implies that $|\theta(a)| \to \infty$ as $a \to \infty$. To establish the proof, it thus suffices to show that $\theta(a)$ is continuous on $[0, \infty)$. The desired result is then immediate from the intermediate value theorem.

[1] This assumption simply ensures that $E_{P_{Y(d)}}[|m(aY(d))| \mid Y > 0]$ exists for all values of $a > 0$.

To establish continuity, fix some $a \in [0, \infty)$ and consider a sequence $a_n \to a$. Without loss of generality, assume $a_n < a + 1$ for all n. Let $m_{a_n}(y) = m(a_n y)$. Since m is continuous, $m_{a_n}(y) \to m_a(y)$ pointwise. We are done if we can apply the dominated convergence theorem to show that therefore $E[m_{a_n}(Y)] \to E[m_a(Y)]$.

Since $m(y)/\log(y) \to 1$ as $y \to \infty$, there exists $\bar{y}$ such that $m(y) < 2\log(y)$ for all $y \geq \bar{y}$. From the monotonicity of m, it follows that

$$\begin{aligned} m(0) \leq m(y) &\leq \mathbb{1}[y \leq \bar{y}]m(\bar{y}) + \mathbb{1}[y > \bar{y}]2\log(y) \\ &\leq \eta + 2 \cdot \mathbb{1}[y > \bar{y}]\log(y), \end{aligned} \tag{C.1}$$

where $\eta = |m(\bar{y})|$, and hence

$$\begin{aligned} m(0) \leq m_{a_n}(y) &\leq \eta + 2 \cdot \mathbb{1}[a_n y > \bar{y}]\log(a_n y) \\ &\leq \eta + 2 \cdot \mathbb{1}[y > 0] \cdot (|\log(a+1)| + |\log(y)|) =: \bar{m}(y). \end{aligned}$$

for all n. Hence, we have that $|m_{a_n}(y)| \leq |m(0)| + \bar{m}(y)$ for all n, and the bounding function is integrable for $Y(d)$ for $d = 0, 1$ by the fourth assumption of the proposition. It follows from the dominated convergence theorem that $E_P[m_{a_n}(Y(d))] \to E_P[m_a(Y(d))]$ for $d = 0, 1$, and thus $\theta(a_n) \to \theta(a)$, as we wished to show. □

C.1.2 Proof of Theorem 3.3.3

Proposition C.1.2 (A trilemma)**.** *The following three properties cannot hold simultaneously:*

(a) $\theta_g = E_P[g(Y(1), Y(0))]$ *for a non-constant function* $g : [0, \infty)^2 \to \mathbb{R}$ *that is weakly increasing in its first argument.*

(b) The function g *is scale-invariant.*

(c) θ_g *is point-identified over* $\mathcal{P}_+$.[2]

[2] A minor technical complication arises from the fact that $E_P[g(Y(1), Y(0)]$ could be infinite for some P. For the purposes of our result, it suffices to trivially define θ_g to be identified in this case. Alternatively, the same result holds if part (c) is modified to impose only that θ_g is point-identified over all distributions in $\mathcal{P}_+$ with finite support, thus avoiding issues related to undefined expectations.

Proof. To establish the proof of Theorem 3.3.3, we rely on Theorem C.1.3, which shows that the only scale-invariant parameter of the form $E_P[g(Y(1), Y(0))]$ that is identified over distributions on the positive reals is the ATE in logs (up to an affine transformation).

Given Theorem C.1.3, note that if $g : [0, \infty)^2 \to \mathbb{R}$ is increasing in y_1, then it cannot be equal to $c \log(y_1/y_0) + d$ for $c > 0$ everywhere on $(0, \infty)^2$, since this would imply that $\lim_{y_1 \to 0} g(y_1, 1) = -\infty < g(0, 1)$. Theorem 3.3.3 is then immediate from Theorem C.1.3, which shows that if properties (a) and (b) are satisfied, and θ_g is point-identified over $\mathcal{P}_{++} \subset \mathcal{P}_+$, then $g = c \log(y_1/y_0) + d$ on $(0, \infty)^2$. Thus, there does not exist such a g. □

Proposition C.1.3. *Let $\mathcal{P}_{++}$ denote the set of distributions over compact subsets of $(0, \infty)^2$. Suppose $g : (0, \infty)^2 \to \mathbb{R}$ is weakly increasing in y_1 and scale-invariant. Then θ_g is point-identified over $\mathcal{P}_{++}$ if and only if $g(y_1, y_0) = c \cdot (\log(y_1) - \log(y_0)) + d$, for constants $c \geq 0$ and $d \in \mathbb{R}$.*

Proof. We first show that point-identification over $\mathcal{P}_{++}$ implies that $g(\cdot, \cdot)$ must be additively separable. We do so by considering the points $\{y_0, y_0 + b\} \times \{y_1, y_1 + a\}$ on a rectangular grid. If $g(\cdot, \cdot)$ is not additively separable, then its expectation with respect to distributions supported on the rectangular grid depends on the correlation. Similar arguments appear in, e.g., Fan *et al.* (2017).

Formally, suppose that there there exist positive values $y_1, y_0, a, b > 0$ such that

$$g(y_1, y_0) + g(y_1 + a, y_0 + b) \neq g(y_1 + a, y_0) + g(y_1, y_0 + b).$$

Now, consider the marginal distributions $P_{Y(d)}$ such that $P(Y(1) = y_1) = \frac{1}{2} = P(Y(1) = y_1 + a)$ and $P(Y(0) = y_0) = \frac{1}{2} = P(Y(0) = y_0 + b)$. Let P_1 and P_2 denote the joint distributions corresponding with these marginals and perfect positive and negative correlation of the potential outcomes, respectively. Then we have that

$$\begin{aligned} E_{P_1}(g(Y(1), Y(0))) &= \frac{1}{2}(g(y_1, y_0) + g(y_1 + a, y_0 + b)) \\ &\neq \frac{1}{2}(g(y_1 + a, y_0) + g(y_1, y_0 + b)) \\ &= E_{P_2}(g(Y(1), Y(0))), \end{aligned}$$

and thus θ_g is not point-identified from the marginals at P_1. Hence, if θ_g is identified over $\mathcal{P}_{++}$, then it must be that

$$g(y_1, y_0) + g(y_1 + a, y_0 + b) = g(y_1 + a, y_0) + g(y_1, y_0 + b) \text{ for all } y_1, y_0, a, b > 0,$$

and hence

$$g(y_1 + a, y_0) - g(y_1, y_0) = g(y_1 + a, y_0 + b) - g(y_1, y_0 + b) \text{ for all } y_1, y_0, a, b > 0.$$

It follows that we can write $g(y_1, y_0) = r(y_1) + q(\frac{1}{y_0})$, where $r(y_1) = g(y_1, 1) - g(1, 1)$ and $q(\frac{1}{y_0}) = g(1, y_0)$.

Second, we show that homogeneity of degree zero, combined with monotonicity, implies that g must be a difference in logarithms. Observe that since g is scale-invariant,

$$g(y_1, y_0) = g\left(\frac{y_1}{y_0}, \frac{y_0}{y_0}\right) = g\left(\frac{y_1}{y_0}, 1\right) =: h\left(\frac{y_1}{y_0}\right),$$

where h is an increasing function. We thus have that for any $a, b > 0$,

$$\begin{aligned} g(1,1) &= h(1) = r(1) + q(1) \\ g(a,1) &= h(a) = r(a) + q(1) \\ g\left(1, \frac{1}{b}\right) &= h(b) = r(1) + q(b) \\ g\left(a, \frac{1}{b}\right) &= h(ab) = r(a) + q(b) \end{aligned}$$

and hence $h(ab) = h(a) + h(b) - h(1)$. It follows that $\tilde{h}(x) = h(x) - h(1)$ is an increasing function such that $\tilde{h}(ab) = \tilde{h}(a) + \tilde{h}(b)$ for all $a, b \in \mathbb{R}$, i.e. an increasing function satisfying Cauchy's logarithmic function equation: $\phi(ab) = \phi(a) + \phi(b)$ for all positive reals a, b. Recall that if a function is increasing, then it has countably many discontinuity points, and thus is continuous somewhere. It is a well-known result in functional equations that the only solutions to Cauchy's logarithmic equation are of the form $\phi(t) = c\log(t)$, if we require that these solutions are continuous at some point; see Aczél (1966), Theorem 2 in Section 2.1.2.[3] Since we require monotonicity, the constant $c \geq 0$. Thus, $g(y_1, y_0) = h(y_1/y_0) = \tilde{h}(y_1/y_0) + \tilde{h}(1) = c\log(y_1) - c\log(y_0) + \tilde{h}(1)$. Letting $d = \tilde{h}(1)$ completes

[3] Correspondingly, non-trivial solutions to Cauchy's logarithmic equations are highly ill-behaved.

the proof of Theorem C.1.3. □

C.2 Extensions

C.2.1 Sensitivity to finite changes in scale

The following result formalizes the discussion in Theorem 3.2.3 about how the ATE for $m(Y)$ changes with finite changes in the scale of Y.

Proposition C.2.1. *Suppose that:*

1. $m : [0, \infty) \to \mathbb{R}$ *is a weakly increasing function.*

2. $m(y)/\log(y) \to 1$ *as* $y \to \infty$.

3. $E_{P_{Y(d)}}[|\log Y(d)| \mid Y(d) > 0] < \infty$ *for* $d = 0, 1$.

Then, as $a \to \infty$,

$$E_P[m(a \cdot Y(1)) - m(a \cdot Y(0))] = (P(Y(1) > 0) - P(Y(0) > 0)) \cdot \log(a) + o(\log(a)).$$

Proof. Fix a sequence $a_n \to \infty$, and without loss of generality, assume $a_n > e$. We will show that

$$\frac{1}{\log a_n} E_P[m(a_n Y(1)) - m(a_n Y(0))] \to P(Y(1) = 0) - P(Y(0) = 0). \tag{C.2}$$

Define $f_n(y) = m(a_n y)/\log(a_n)$. Note that $f_n(y) \to \mathbb{1}[y > 0]$ pointwise, since $f_n(0) = m(0)/\log(a_n) \to 0$, while for $y > 0$,

$$f_n(y) = \frac{m(a_n y)}{\log(a_n)} = \frac{m(a_n y)}{\log(a_n y)} \frac{\log(a_n) + \log(y)}{\log(a_n)} \to 1,$$

where we use the fact that $m(y)/\log(y) \to 1$ as $y \to \infty$ by assumption. We apply the dominated convergence theorem to show that $E_P[f_n(Y(d))] \to P(Y(d) > 0)$.

We showed in the proof to Theorem 3.2.1 that

$$|m(y)| \leq \kappa + 2 \cdot \mathbb{1}[y > 0] \cdot |\log(y)|$$

where κ is a constant not depending on y.[4] It follows that f_n is similarly dominated:

$$|f_n(y)| = \frac{|m(a_n y)|}{\log(a_n)} \leq \kappa + 2 \cdot \mathbb{1}[y > 0] \cdot (1 + |\log(y)|).$$

Further, since $E_P[|\log(Y(d))| \mid Y(d) > 0]$ is finite by assumption, the upper bound is integrable for $y = Y(d)$ for $d = 0, 1$. It follows from the dominated convergence theorem that

$$E_P[f_n(Y(d))] = E_P\left[\frac{m(a_n Y(d))}{\log(a_n)}\right] \to E_P[\mathbb{1}[Y(d) > 0]] = P(Y(d) > 0).$$

Equation (C.2) then follows from applying this result for $d = 0, 1$ and taking the difference of the limits. □

C.2.2 Extension to continuous treatments

Although we focus on binary treatment in the main text for simplicity, similar issues arise with continuously distributed D. Suppose now that D can take a continuum of values on some set $\mathcal{D} \subseteq \mathbb{R}$. Let $Y(d)$ denote the potential outcome at the dose d, and P the distribution of $Y(\cdot)$. Consider the parameter

$$\theta(a) = \int_{\mathcal{D}} \omega(d) E_P[m(aY(d))],$$

which is a weighted sum of the average values of $m(aY(d))$ across different values of d with weights $\omega(d)$. For example, in an RCT with a continuous treatment, a regression of $m(aY)$ on D yields a parameter of the form $\theta(a)$ where, by the Frisch–Waugh–Lovell theorem, the weights are proportional to $(d - E[D])p(d)$ and integrate to 0.[5]

We now show that $\theta(a)$ can be made to have arbitrary magnitude via the choice of a when there is an extensive margin effect. In particular, by an extensive margin effect we mean that $\int \omega(d) P(Y(d) > 0) \neq 0$, i.e. when there is an average effect on the probability of a zero outcome, using the same weights $\omega(d)$ that are used for $\theta(a)$. When $\theta(a)$ is the regression of $m(aY)$ on D in an RCT, for example, $\int \omega(d) P(Y(d) > 0) \neq 0$ if the regression of $\mathbb{1}[Y > 0]$ on D yields a non-zero coefficient.

Proposition C.2.2. *Suppose that:*

[4] In particular, (C.1) implies the inequality for $\kappa = \eta + |m(0)|$.

[5] Here, $p(d)$ denotes the density of D at d over the randomization distribution.

1. *The function m satisfies parts 1 and 2 of Theorem 3.2.1.*

2. *(Extensive margin effect) $\int_{\mathcal{D}} \omega(d)P(Y(d) > 0) \neq 0$.*

3. *(Bounded expectations) For all d, $E_P[|\log(Y(d))| \mid Y(d) > 0] < \infty$.*

4. *(Regularity for weights) The weights $\omega(d)$ satisfy $\int_{\mathcal{D}} \omega(d) = 0$, $\int_{\mathcal{D}} |\omega(d)| < \infty$ and $\int_{\mathcal{D}} |\omega(d)| \cdot E_P[|\log(Y(d))| \mid Y(d) > 0] < \infty$.*

Then for every $\theta^* \in (0, \infty)$, there exists $a > 0$ such $|\theta(a)| = \theta^*$. In particular, $\theta(a)$ is continuous and $\theta(a) \to 0$ as $a \to 0$ and $|\theta(a)| \to \infty$ as $a \to \infty$.

Proof. Note that $\theta(0) = \int \omega(d)m(0) = 0$. It thus suffices to show that $\theta(a)$ is continuous for $a \in [0, \infty)$ and that $|\theta(a)| \to \infty$ as $a \to \infty$. The result then follows from the intermediate value theorem.

We first show continuity. Fix $a \in [0, \infty)$ and a sequence $a_n \to a$. Let $f_n(d) = \omega(d)E_P[m(a_n Y(d))]$. We showed in the proof to Theorem 3.2.1 that $E_P[m(a_n Y(d))] \to E_P[m(aY(d))]$, and thus $f_n(d) \to \omega(d)E_P[m(aY(d))]$ pointwise. We also showed in the proof to Theorem 3.2.1 that for a_n sufficiently close to a,

$$|m(a_n Y)| \leq \kappa + 2 \cdot \mathbb{1}[y > 0] \cdot |\log(y)|,$$

for a constant κ not depending on n. It follows that

$$|f_n(d)| \leq |\omega(d)| \cdot |\kappa| + 2|\omega(d)| \cdot E_P[|\log(Y(d))| \mid Y(d) > 0],$$

and the upper bound is integrable by part 4 of the Proposition. Hence, by the dominated convergence theorem, we have that $\theta(a_n) = \int_{\mathcal{D}} f_n(d) \to \int_{\mathcal{D}} \omega(d)E_P[m(aY(d)] = \theta(a)$, as needed.

To show that $|\theta(a)| \to \infty$ as $a \to \infty$, we will show that

$$\frac{\theta(a)}{\log(a)} \to \int_{\mathcal{D}} \omega(d)P[Y(d) > 0]$$

as $a \to \infty$. Consider $a_n \to \infty$, and suppose without loss of generality that $a_n > e$. Observe that

$$\frac{\theta(a_n)}{\log(a_n)} = \int_{\mathcal{D}} \omega(d)\frac{E_P[m(a_n Y(d))]}{\log(a_n)}.$$

We showed in the proof to Theorem C.2.1 that for each d,

$$\frac{E_P[m(a_n Y(d))]}{\log(a_n)} \to P(Y(d) > 0).$$

Letting $f_n(d) = \omega(d) \dfrac{E_P[m(a_n Y(d))]}{\log(a_n)}$, we thus have that $f_n(d) \to \omega(d) P(Y(d) > 0)$ pointwise. Moreover, we showed in the proof to Theorem 3.2.1 that

$$|m(y)| \leq \kappa + 2 \cdot \mathbb{1}[y > 0] \cdot |\log(y)|$$

where κ is a constant not depending on y. It follows that

$$\frac{|m(a_n y)|}{\log(a_n)} \leq \kappa + 2 \cdot \mathbb{1}[y > 0] \cdot (1 + |\log(y)|)$$

and thus that

$$|f_n(d)| \leq |\omega(d)| \cdot (\kappa + 2 + 2E_P[|\log(Y(d)| \mid Y(d) > 0])$$

where the upper bound is integrable by the fourth part of the proposition. The result then follows from dominated convergence.

□

C.2.3 Extension to OLS estimands and standard errors

As noted in Theorem 3.2.5, our results imply that any consistent estimator of the ATE for an outcome of the form $m(aY)$ will be (asymptotically) sensitive to scaling when there is an extensive margin effect. Our results thus cover the OLS estimator when it is consistent for the ATE for some (sub)-population P (e.g. in an RCT or under unconfoundedness). Given the prominence of OLS in applied work—and the fact that it is sometimes used for non-causal analyses—we now provide a direct result on the sensitivity to scaling of the estimand of an OLS regression of an outcome of the form $m(aY)$ on an arbitrary random variable X.

Specifically, suppose that $(X, Y) \sim Q$, for $Y \in [0, \infty)$ and $X \in \mathbb{R}^J$, where the first element of X is a constant. Consider the OLS estimand

$$\beta(a) = E_Q[XX']^{-1} E_Q[X m(aY)],$$

i.e. the population coefficient from a regression of $m(aY)$ on X. We assume that $E_Q[XX']$ is full-rank so that $\beta(a)$ is well-defined. Letting $\beta_j(a) = e_j'\beta(a)$ be the j^{th} element of $\beta(a)$, we will show that $\beta_j(a)$ can be made to have arbitrary magnitude via the choice of a if $\gamma_j \neq 0$, where

$$\gamma = E_Q[XX']^{-1}E_Q[X\mathbb{1}[Y > 0]]$$

is the coefficient from a regression of $\mathbb{1}[Y > 0]$ on X.

Proposition C.2.3. *Suppose that*

1. *The function m satisfies parts 1 and 2 of Theorem 3.2.1.*

2. *(Finite expectations) $E_Q[\|X\|] < \infty$ and $E_Q[\|X\log(Y)\| \mid Y > 0] < \infty$.*

Then for every $j \in \{2, ..., J\}$, $\beta_j(a)/\log(a) \to \gamma_j$ as $a \to \infty$. Moreover, if $\gamma_j \neq 0$ for some $j \in \{2, ..., J\}$, then for every $\beta_j^* \in (0, \infty)$, there exists $a > 0$ such that $|\beta_j(a)| = \beta_j^*$. In particular $\beta_j(a)$ is continuous with $\beta_j(a) \to 0$ as $a \to 0$ and $|\beta_j(a)| \to \infty$ as $a \to \infty$.

We note that Theorem C.2.3 implies that the OLS estimator for the j^{th} coefficient, $\hat{\beta}_j(a)$, will be arbitrarily sensitive to the choice of a when the corresponding extensive margin OLS estimator $\hat{\gamma}_j$, is non-zero. This follows immediately from setting Q to be the empirical distribution of $(Y_i, X_i)_{i=1}^N$ and applying Theorem C.2.3 (note that part 2 of the Proposition is trivially satisfied for the empirical distribution, since X and Y are both bounded over the empirical distribution).

OLS Standard Errors. We also show that as $a \to \infty$, the t-statistic for the OLS estimate $\hat{\beta}_j$ constructed using heteroskedasticity-robust standard errors converges to the t-statistic for $\hat{\gamma}_j$ (again using heteroskedasticity-robust standard errors). Formally, let

$$\hat{\Omega}_\beta(a) = \left(\frac{1}{N}\sum_i X_i X_i'\right)^{-1} \left(\frac{1}{N}\sum_i X_i X_i' \hat{\epsilon}_i(a)^2\right) \left(\frac{1}{N}\sum_i X_i X_i'\right)^{-1}$$

denote the estimator of the heteroskedasticity-robust variance matrix for $\hat{\beta}(a)$, where $\hat{\epsilon}_i(a) = m(aY_i) - X_i'\hat{\beta}(a)$, and $\hat{\beta}(a)$ is the OLS estimate of $\beta(a)$. The t-statistic for $\hat{\beta}_j(a)$ is then $\hat{t}_{\beta_j}(a) = \hat{\beta}_j(a)/\hat{\sigma}_{\beta_j}(a)$, where $\hat{\sigma}_{\beta_j}(a) = \sqrt{e_j'\hat{\Omega}_\beta(a)e_j}/\sqrt{N}$. Analogously, let

$$\hat{\Omega}_\gamma = \left(\frac{1}{N}\sum_i X_iX_i'\right)^{-1}\left(\frac{1}{N}\sum_i X_iX_i'\hat{u}_i^2\right)\left(\frac{1}{N}\sum_i X_iX_i'\right)^{-1}$$

be the heteroskedasticity-robust variance estimator for $\hat{\gamma}$, the OLS estimate of γ, where $u_i = \mathbb{1}[Y_i > 0] - X_i'\hat{\gamma}$. The t-statistic for $\hat{\gamma}_j$ is then $\hat{t}_{\gamma_j} = \hat{\gamma}_j/\hat{\sigma}_{\gamma_j}$, where $\hat{\sigma}_{\gamma_j} = \sqrt{e_j'\hat{\Omega}_\gamma e_j}/\sqrt{N}$.

Proposition C.2.4. *Suppose that* $\left(\frac{1}{N}\sum_i X_iX_i'\right)$ *is full-rank and that* $\hat{\sigma}_{\gamma_j} > 0$. *If the function* m *satisfies parts 1 and 2 of Theorem 3.2.1 and* $\hat{\gamma}_j > 0$, *then* $\hat{t}_{\beta_j}(a) \to \hat{t}_{\gamma_j}$ *as* $a \to \infty$.

It follows that when the units of Y are made large, the t-statistic for a treatment effect estimate for $m(Y)$ estimated using OLS will converge to the t-statistic for the OLS estimate of the extensive margin. Figure C.1 shows that, indeed, the t-statistics for estimates using $\operatorname{arcsinh}(Y)$ in the *AER* tend to be close to the t-statistics for the extensive margin, and tend to become even closer after rescaling the units by a factor of 100.

Proof of Theorem C.2.3. Fix $j \in \{2, ..., J\}$. Note that $\beta(0) = E_Q[XX']^{-1}E[Xm(0)]$, is the coefficient from a regression of a constant outcome $m(0)$ on X, and thus $\beta_1(0) = m(0)$ while $\beta_k(0) = 0$ for $k \geq 2$. Thus $\beta_j(0) = 0$. To complete the proof, we will first show that $\beta_j(a_n) = \gamma_j \log(a_n) + o(\log(a_n))$. Hence, if $\gamma_j > 0$, then $|\beta_j(a)| \to \infty$ as $a \to \infty$. We will then establish that $\beta_j(a)$ is continuous for $a \in [0, \infty)$. The fact that one can obtain any positive value for $|\beta_j(a)|$ then follows from the intermediate value theorem.

For ease of notation, let $\nu' = e_j'E_Q[XX']^{-1}$, so that $\beta_j(a) = E_Q[\nu'Xm(aY)]$.

We first show that $\beta_j(a_n) = \gamma_j \log(a_n) + o(\log(a_n))$. Consider a sequence $a_n \to \infty$, and assume without loss of generality that $a_n > e$. Let $f_n(x, y) = \nu'x \cdot m(a_ny)/\log(a_n)$. Observe that $f_n(x, y) \to \nu'x \cdot \mathbb{1}[y > 0]$ pointwise, since $f_n(x, 0) = \nu'x \cdot m(0)/\log(a_n) \to 0$, while for $y > 0$,

$$f_n(x, y) = \nu'x \cdot \frac{m(a_ny)}{\log(a_n)} = \nu'x \cdot \frac{m(a_ny)}{\log(a_ny)}\frac{\log(a_n) + \log(y)}{\log(a_n)} \to \nu'x,$$

where we use the fact that $m(y)/\log(y) \to 1$ as $y \to \infty$. We showed in the proof to Theorem C.2.1 that

$$\frac{|m(a_ny)|}{\log(a_n)} \leq \kappa + 2 \cdot \mathbb{1}[y > 0] \cdot (1 + |\log(y)|),$$

which implies that

$$|f_n(x,y)| \leq |\nu' x \cdot (\kappa + 2 \cdot \mathbb{1}[y>0] \cdot (1 + |\log(y)|))| =: \bar{f}(x,y).$$

Moreover, part 2 of the proposition implies that $\bar{f}(X,Y)$ is integrable. From the dominated convergence theorem, it follows that

$$\frac{\beta_j(a_n)}{\log(a_n)} = E_Q[f_n(X,Y)] \to E_Q[\nu' X \mathbb{1}[Y>0]] = \gamma_j.$$

Hence, we see that $\beta_j(a_n) = \gamma_j \log(a_n) + o(\log(a_n))$. It follows that $|\beta_j(a_n)| \to \infty$ when $\gamma_j \neq 0$.

To complete the proof, we show continuity of $\beta_j(a)$. Fix $a \in [0,\infty)$, and consider a sequence $a_n \to a$. Assume without loss of generality that $a_n < a+1$ for all n. Let $f_n(x,y) = \nu' x \cdot m(a_n y)$. From the continuity of m, we have that $f_n(x,y) \to \nu' x \cdot m(ay)$ pointwise. We showed in the proof to Theorem 3.2.1 that there exists some κ (not depending on n) such that

$$|m(a_n y)| \leq \kappa + 2\mathbb{1}[y>0] \cdot |\log(y)|.$$

Hence,

$$|f_n(x,y)| \leq |\nu' x \cdot (\kappa + 2\mathbb{1}[y>0]|\log(y)|)|.$$

Moreover, the bounding function is integrable over the distribution of (X,Y) by part 2 of the proposition. Applying the dominated convergence theorem again, we obtain that

$$\beta_j(a_n) = E_Q[f_n(X,Y)] \to E_Q[\nu' X \cdot m(aY)] = \beta_j(a),$$

as needed. □

Proof of Theorem C.2.4. Consider $a_n \to \infty$. Applying Theorem C.2.3 to the empirical distribution, we have that $\hat{\beta}(a_n)/\log(a_n) = \hat{\gamma} + o(1)$. It follows that

$$\frac{1}{\log(a_n)}\hat{\epsilon}_i(a_n) = \frac{m(a_n Y_i)}{\log(a_n)} - \frac{\hat{\beta}(a_n)' X_i}{\log(a_n)} = \mathbb{1}[Y_i > 0] - \hat{\gamma}' X_i + o(1) = \hat{u}_i + o(1).$$

Since $\hat{\Omega}_n(a_n)$ is a continuous function of the $\hat{\epsilon}_i(a_n)^2$, we obtain that $\log(a_n)^{-2}\hat{\Omega}_\beta(a_n) \to \hat{\Omega}_\gamma$, and thus that $\log(a_n)^{-1}\hat{\sigma}_{\beta_j}(a_n) = \hat{\sigma}_{\gamma_j} + o(1)$. It follows that

$$\hat{t}_{\beta_j}(a_n) = \frac{\hat{\beta}_j(a_n)/\log(a_n)}{\hat{\sigma}_{\beta_j}(a_n)/\log(a_n)} = \frac{\hat{\gamma}_j + o(1)}{\hat{\sigma}_{\gamma_j} + o(1)} \rightarrow \frac{\hat{\gamma}_j}{\hat{\sigma}_{\gamma_j}} = \hat{t}_{\gamma_j},$$

as needed. □

C.3 Connection to structural equations models

Previous work has considered a variety of estimators for settings with zero-valued outcomes beginning with structural equations models rather than the potential outcomes model that we consider. These papers have reached different results, with some concluding that regressions with $\text{arcsinh}(Y)$ have the interpretation of an elasticity, and others showing that they are inconsistent and advocating for other methods (e.g. Poisson regression) instead. In this section, we interpret the results in those papers from the perspective of the potential outcomes model, and show that these diverging conclusions stem from different implicit assumptions about the potential outcomes, as well as a focus on different causal parameters.

Before discussing specific papers, we first note that, broadly speaking, structural equation models can be viewed as constraining the joint distribution of potential outcomes. Observe that, for any pair of potential outcomes $(Y(1), Y(0))$, we can represent them as $(Y(1, U), Y(0, U))$ for some function $Y(d, u)$ and individual-level unobservable (or "structural error") U. The potential outcomes framework we work with in this paper does not impose any functional form assumptions on $Y(d, u)$. Structural equation models, on the other hand, tend to specify explicit functional forms for $Y(d, u)$. In what follows, we consider the implicit restrictions placed on the potential outcomes as well as the target estimand in work related work that starts with a structural equations model.

C.3.1 Bellemare and Wichman (2020) and Thakral and Tô (2023)

Bellemare and Wichman (2020) consider OLS regressions of the form[6]

$$\text{arcsinh}(Y) = \beta_0 + D\beta_1 + U. \tag{C.3}$$

[6] They also consider specifications with additional covariates on the right-hand side, although we abstract away from this for expositional simplicity.

Note that when D is binary and randomly assigned, $D \perp\!\!\!\perp (Y(1), Y(0))$, then from the perspective of the potential outcomes model, the population coefficient β_1 is the ATE for $\text{arcsinh}(Y)$. Bellemare and Wichman (2020) instead consider the interpretation of β_1 when (C.3) is treated as structural, i.e. if there are constant treatment effects of D on $\text{arcsinh}(Y)$. From the perspective of the potential outcomes model, this amounts to imposing that the potential outcomes $Y(d) := Y(d, U)$ take the form

$$\text{arcsinh}(Y(d, U)) = \beta_0 + d\beta_1 + U, \tag{C.4}$$

where the individual-level random variable U takes the same value for all values of d. Under (C.4), we have that

$$\beta_1 = \text{arcsinh}(Y(1, U)) - \text{arcsinh}(Y(0, U)).$$

Since $\text{arcsinh}(y) \approx \log(2y)$ for y large, it follows that $\beta_1 \approx \log(Y(1, U)/Y(0, U))$ when $Y(1, U)$ and $Y(0, U)$ are large. Thus, Bellemare and Wichman (2020) argue that β_1 approximates the semi-elasticity of the outcome with respect to d when the outcome is large. They likewise provide similar results for the elasticity of $Y(d, U)$ with respect to treatment when treatment is continuous. Their results thus imply that the ATE for $\text{arcsinh}(Y)$ has a sensible interpretation as a (semi-)elasticity when the structural equation for the potential outcomes given in (C.4) holds.

It is worth emphasizing, however, that (C.4) will generally be incompatible with the data when both $Y(1)$ and $Y(0)$ have point-mass at zero, and $\beta_1 \neq 0$. Specifically, note that (C.4) implies that for all values of U,

$$\text{arcsinh}(Y(1, U)) - \text{arcsinh}(Y(0, U)) = \beta_1.$$

If $\beta_1 > 0$, for example, this implies that $\text{arcsinh}(Y(1, U)) > \text{arcsinh}(Y(0, U))$, and hence $Y(1, U) > Y(0, U)$, since the $\text{arcsinh}(y)$ function is strictly increasing for $y \geq 0$. However, since $Y(0, U) \geq 0$ by assumption, this implies that $Y(1, U) > 0$ with probability 1. Thus, the model in (C.4) is incompatible with $P(Y(1) = 0) > 0$ if $\beta_1 > 0$. By similar logic, the model is also incompatible with $P(Y(0) = 0) > 0$ if $\beta_1 < 0$. In settings where there is point-mass at zero, the model that Bellemare and Wichman (2020) show gives β_1 an interpretation as a semi-elasticity will therefore typically be rejected by the data. It is also worth noting that even if there are no zeros in the data, the model in (C.4) will generally be sensitive to units, in the sense that if (C.4) holds for Y measured in dollars, it will generally not hold when Y is measured in cents. The validity of the interpretation of β_1 as an elasticity thus depends on having chosen the "correct" scaling of the outcome such that (C.4) holds.

Similar issues apply if we consider alternative transformations on the left-hand side of (C.3). For example, Thakral and Tô (2023) consider versions of (C.3) that replaces $\text{arcsinh}(Y)$ with the power function Y^k. They then consider the implied "semi-elasticities" of the form $\eta(y_0) = \beta_1/(ky_0^k)$. The parameter $\eta(y_0)$ has the interpretation as a structural semi-elasticity when d has a contant effect on Y^k. Specifically, if D is continuous and the structural equation

$$Y(d,U)^k = \beta_0 + d\beta_1 + U, \tag{C.5}$$

holds, then $\eta(y_0) = \left(\frac{\partial}{\partial d} Y(d,U)\right) / Y(d,U)$ evaluated at $Y(d,U) = y_0$, so $\eta(y_0)$ corresponds to the semi-elasticity of $Y(d,U)$ with respect to d. However, as with (C.4), (C.5) is generally incompatible with settings in which $P(Y(d,U) = 0)$ for multiple values of d. For example, if $\beta_1 > 0$, then $Y(0,U) \geq 0$ implies that $Y(1,U) > 0$. Equation (C.5), which gives a causal interpretation to $\eta(\beta_0)$ as a semi-elasticity, will thus generally be incompatible with settings in which some units have $Y = 0$ under multiple treatment statuses.

C.3.2 Cohn *et al.* (2022)

Cohn *et al.* (2022) consider structural equations of the form

$$Y = \exp(\alpha + D\beta)U. \tag{C.6}$$

When $E[U \mid D] = 1$, they show that Poisson regression is consistent for β, whereas regressions of $\log(1+Y)$ or $\log(Y)$ on D may be inconsistent for β.[7] Although Cohn *et al.* (2022) do not consider a potential outcomes interpretation of β, we can give β a causal interpreation if we impose that the potential outcomes take the form

$$Y(d,U) = \exp(\alpha + d\beta)U(d), \tag{C.7}$$

[7] We thank Kirill Borusyak for an insightful discussion on this topic. Relatedly, in an influential paper, Santos Silva and Tenreyro (2006) consider the structural equations model $Y_i = \exp(X_i'\beta)U_i$ where $E[U_i \mid X_i] = 1$, and show that Poisson regression consistently estimates β while a regression using log on the left-hand side does not, although they do not provide any formal results on log-like transformations.

where $E[U(d)] = 1$. Under (C.7), it follows that $\exp(\beta) = E[Y(1)]/E[Y(0)]$, i.e. the parameter $\theta_{\text{ATE\%}}$ considered in Section 3.4.1.[8]

We note, however, that if one were instead to impose (C.6) with the assumption that $E[\log(U(d)) \mid D] = 0$, then the regression of $\log(Y)$ on D would be consistent for β, whereas Poisson regression would generally be inconsistent for β. Indeed, under the potential outcomes model in (C.7) with the assumption that $E[\log(U(d))] = 0$, we have that $\beta = E[\log(Y(1)) - \log(Y(0))]$, the ATE in logs.[9]

This discussion highlights that whether or not an estimator is consistent depends on the specification of the *target parameter*. Our results help to illuminate what parameters can be consistently estimated by enumerating the properties that identified causal parameters can (or cannot) have.

C.3.3 Tobit models

An alternative structural approach is to explicitly model the extensive margin, a classic example of which is the Tobit model (Tobin, 1958). Following the discussion of Tobit models in Angrist and Pischke (2009), suppose there exist latent potential outcomes $Y^*(d) = \mu_d + U$, where $U \sim \mathcal{N}(0, \sigma^2)$ and $D \perp\!\!\!\perp U$. The observed potential outcome $Y(d)$ is then the latent potential outcome truncated at zero, $Y(d) = \max(Y^*(d), 0)$. We note that in this model, the treatment has a constant additive effect of $\mu_1 - \mu_0$ on the latent outcome, and the latent potential outcomes are assumed to be normally distributed.

Thanks to the parametric assumptions, the unknown parameters μ_1, μ_0, σ^2 are identified and estimable via, e.g., maximum likelihood. As a result, the entire joint distribution of potential outcomes is identified, since this depends only on (μ_1, μ_0, σ). This implies, in turn, that all of the possible target parameters considered in Section 3.4 are point-identified. For example, under this model

$$E[\log Y(d) \mid Y(1) > 0, Y(0) > 0] = E\left[\log\left(\mu_d + U\right) \mid U > -\mu_1, U > -\mu_0\right],$$

where the right-hand side can be computed numerically since $U \sim \mathcal{N}(0, \sigma^2)$. Thus, the intensive

[8] Bellégo *et al.* (2022) also consider (C.6), but consider the more general class of identifying restrictions of the form $E[D\log(U + \delta)] = 0$, where δ is a tuning parameter.

[9] Note that the assumption that $E[\log(U)] = 0$ implicitly implies that $U > 0$, and thus $Y > 0$.

margin treatment effect in logs, $\theta_{\text{Intensive}}$, is actually point-identified under the Tobit model.[10]

It is worth nothing that unlike some of the models considered above, the Tobit model is consistent with a nonzero extensive margin. However, the assumptions of normal errors and constant treatment effects on the latent index are restrictive. As discussed in Section 3.4, imposing these assumptions is not necessary for identification if one is ultimately interested in, say, $E[Y(1) - Y(0)]/E[Y(0)]$, and one can obtain bounds on the intensive margin effect without imposing these assumptions.[11] Moreover, as Angrist and Pischke (2009) and Angrist (2001) point out, it is often not clear what the economic meaning of the latent potential outcome $Y^*(d)$ is—if $Y(d)$ is earnings, for example, what is the meaning of having negative latent earnings ($Y^*(d) < 0$)?

C.4 Connection to two-part models

One approach recommended for settings with weakly-positive outcomes is to estimate a two-part model (Mullahy and Norton, 2023). In this section, we briefly review two-part models, and show that the marginal effects implied by these models do not correspond with ATEs for the intensive margin without further restrictions on the potential outcomes. Thus, while two-part models strike us as a reasonable approach if the goal is to model the conditional expectation function of observed outcomes Y given treatment D (as in Mullahy and Norton (2023)), they will often not be appropriate if instead the goal is to learn about a causal effect along the intensive margin.[12]

The idea of a two-part model is to separately model the conditional distribution $Y \mid D$ using (a) a first model for the probability that Y is positive given D, $P(Y > 0 \mid D)$ (b) a second model for the conditional expectation of Y given that it is positive, $E[Y \mid D, Y > 0]$. Common specifications include logit or probit for part (a), and a linear regression of the positive values of Y on D for part b); see, e.g., Belotti *et al.* (2015). After obtaining estimates of the two-part model, it is common to

[10] Likewise, the intensive margin treatment effect in levels, $E[Y(1) - Y(0) \mid Y(1) > 0, Y(0) > 0]$ is simply $\mu_1 - \mu_0$.

[11] We note that the assumptions of the Tobit model imply (but are strictly stronger than) the assumption of rank preservation of the potential outcomes. However, rank preservation alone suffices to point identify $E[\log Y(1) - \log Y(0) \mid Y(1) > 0, Y(0) > 0]$.

[12] We are particularly grateful to John Mullahy for an enlightening discussion of this topic.

evaluate the marginal effects of D on both parts, i.e. the implied values of

$$\tau_a = P(Y > 0 \mid D = 1) - P(Y > 0 \mid D = 0)$$
$$\tau_b = E[Y \mid Y > 0, D = 1] - E[Y \mid Y > 0, D = 0].$$

We now consider how the parameters of the two-part model relate to causal effects in the potential outcomes model. Suppose, for simplicity, that the two-part model is well-specified, so that it correctly models $P(Y > 0 \mid D)$ and $E[Y \mid Y > 0, D]$. Suppose further that D is randomly assigned, $D \perp\!\!\!\perp Y(1), Y(0)$. In this case, we have that

$$\tau_a = P(Y(1) > 0) - P(Y(0) > 0)$$
$$\tau_b = E[Y(1) \mid Y(1) > 0] - E[Y(0) \mid Y(0) > 0].$$

From the previous display, we see that the marginal effect on the first margin, τ_a, has a causal interpretation: it is the treatment's effect on the probability that the outcome is positive.

The interpretation of the marginal effect on the second margin, τ_b, is more complicated, however. For simplicity, suppose are willing to impose the "monotonicity" assumption discussed in Section 3.4, $P(Y(1) = 0, Y(0) > 0) = 0$, so that anyone with a zero outcome under treatment also has a zero outcome under control. Then, letting $\alpha = P(Y(0) = 0 \mid Y(1) > 0)$, we can write τ_b as

$$\begin{aligned}\tau_b &= (1-\alpha)E[Y(1) \mid Y(1) > 0, Y(0) > 0] \\ &\quad + \alpha E[Y(1) \mid Y(1) > 0, Y(0) = 0] - E[Y(0) \mid Y(1) > 0, Y(0) > 0] \\ &= \underbrace{E[Y(1) - Y(0) \mid Y(1) > 0, Y(0) > 0]}_{\text{Intensive margin effect}} \\ &\quad + \alpha \underbrace{\left(E[Y(1) \mid Y(1) > 0, Y(0) = 0] - E[Y(1) \mid Y(1) > 0, Y(0) > 0]\right)}_{\text{Selection term}},\end{aligned}$$

where the first equality uses iterated expectations, and the second re-arranges terms.

The previous display shows that τ_b is the sum of two terms. The first is the ATE for individuals who would have a positive outcome regardless of treatment status (similar to $\theta_{\text{Intensive}}$ in Section 3.4, except using Y instead of $\log(Y)$). The second term is not a causal effect, but rather represents a selection term: it is proportional to the difference in the average value of $Y(1)$ for "compliers" who would have positive outcomes only under treatment versus "always-takers" who would have positive outcomes regardless of treatment status. In many economic contexts, we may expect this selection effect to be negative. For example, we may suspect that individuals who would only get a job if they receive a particular training have lower ability, and hence lower values of $Y(1)$, than individuals who would have a job regardless of training status. The marginal effect τ_b thus only has an interpretation as an ATE along the intensive margin if either (a) there is no extensive margin effect ($\alpha = 0$) or (b) we are willing to assume that the selection term is zero. Angrist (2001) provided a similar decomposition (without imposing monotonicity), concluding that the two-part model "seems ill-suited for causal inference," at least without further restrictions on the potential outcomes. See, also, Mullahy (2001) for additional discussion.

C.5 Details on Lee bounds using IV in Berkouwer and Dean (2022)

We now describe in detail our approach for constructing Lee (2009)-type bounds in the IV setting of Berkouwer and Dean (2022).

Estimating the instrument-complier distributions. The first step is to estimate the distribution of $Y(0)$ and $Y(1)$ for instrument-compliers. As shown in Abadie (2002), the CDF for $Y(1)$ for instrument-compliers at a point y can be consistently estimated by using two-stage least squares to estimate the effect of treatment on the outcome $D_i 1[Y_i \leq y]$. The CDF for $Y(0)$ for instrument-compliers can analogously be obtained using the outcome $(D_i - 1)1[Y_i \leq y]$. We estimate these TSLS regressions using analogues to (3.11) (except replacing $\text{arcsinh}(Y_i)$ with the outcomes just described) for all values of y contained in the data. We thus obtain empirical estimates of the CDFs for instrument-compliers, $\hat{F}_{Y(d)}(y)$ for $d = 0, 1$.

Constructing bounds. Note that if $U \sim U[0,1]$, then $Y(d) \sim F^{-1}_{Y(d)}(U)$, where $F^{-1}_{Y(d)}(u) := \inf\{y \mid F_{Y(d)}(y) \geq u\}$. With this formulation in mind, Lee (2009)'s bounds for $E[\log(Y(1)) \mid$

$Y(1) > 0, Y(0) > 0]$ can be written as

$$E[\log(F^{-1}_{Y(1)}(U)) \mid U \in [\theta_{NT}, \theta_{NT} + \theta_{AT}]] \leq E[\log(Y(1)) \mid Y(1) > 0, Y(0) > 0] \\ \leq E[\log(F^{-1}_{Y(1)}(U)) \mid U \in [1 - \theta_{AT}, 1]], \qquad \text{(C.8)}$$

where $\theta_{AT} = P(Y(1) > 0, Y(0) > 0)$, $\theta_{NT} = P(Y(1) = 0, Y(0) = 0)$, and $\theta_C = P(Y(1) > 0, Y(0) = 0)$. We estimate the bounds in (C.8) by plugging in the estimated CDFs for instrument-compliers described above, as well as the values of $\theta_{AT}, \theta_{NT}, \theta_C$ implied by the estimated CDFs. We approximate the expectation over U by taking the average over 100,000 uniform draws.[13] Finally, to compute the bounds on the treatment effect, we must estimate $E[Y(0) \mid Y(0) > 0]$. To do this, we use the fact that

$$E[Y(0) \mid Y(0) > 0] = E[F^{-1}_{Y(0)}(U) \mid U \in [\theta_{NT} + \theta_C, 1]].$$

As before, we then estimate the right-hand-side in the previous display by plugging-in the estimated CDF for instrument-compliers, and simulating over 100,000 uniform draws. The Lee bounds for $\theta_{\text{Intensive}}$ are then obtained by subtracting the estimate of $E[Y(0) \mid Y(0) > 0]$ from the estimates of the lower and upper bounds in (C.8). We estimate standard errors for the bounds using 1,000 draws from a non-parametric clustered bootstrap.[14]

C.6 Appendix tables and figures

- Table C.1 contains information on the *AER* papers discussed.
- Figure C.1 shows how t-statistics change in the replication exercise.
- Table C.2 shows the analogue of Table 3.1 for $\log(1 + Y)$.

[13] We note that in finite samples, the estimated CDF $\hat{F}_{Y(d)}(y)$ may be non-monotonic. Nevertheless, the inverse $\hat{F}^{-1}_{Y(d)}(u) := \inf\{y \mid \hat{F}_{Y(d)}(y) \geq u\}$ remains well defined.

[14] One complication that arises is that for some draws from the bootstrap distribution, the sign of the extensive margin can be the opposite of that in the original data. In our bootstrap procedure, we construct Lee-type bounds assuming monotonicity in whichever direction matches the bootstrapped data. The resulting bootstrap estimates of the bounds appear to be approximately normally distributed, but we think a formal theoretical evaluation of the bootstrap in this setting is an interesting topic for future work.

Table C.1: Papers in the *AER* estimating effects for $\operatorname{arcsinh}(Y)$ with selected quotes

Paper	Interprets Units as %	Original Units	Quote About Percents / Notes
Azoulay et al (2019)	Yes	Publications (yearly)	“In this case, coefficient estimates can be interpreted as elasticities, as an approximation.”
Beerli et al (2021)	Yes	Patent applications (yearly)	“The estimates thus reflect an approximate percentage increase.”
Berkouwer and Dean (2022)	Yes	Weekly expenditure (dollars)	“A 0.50 IHS reduction corresponds to a 39 percent reduction relative to the control group.”
Cabral et al (2022)	Yes	Costs (dollar) per $10K risk-adjusted covered payroll	Refers to estimates as “the elasticities reported in panel A”
Carranza et al (2022)	Yes	Hours worked (weekly)	“Weekly earnings increase by 34% (Table 1, column 3)”
Faber & Gauber (2019)	Yes	Municipality GDP (1000s of Pesos)	“A one standard deviation increase in tourism attractiveness increases local manufacturing GDP by about 40 percent.”
Hjort and Poulsen (2019)	Yes	KB per second	“We find that cable arrival increases measured speed in connected locations, relative to unconnected locations, by around 35 percent”
Johnson (2020)	Yes	Violations (monthly)	“[T]he regression coefficient estimates the ITT effect of a press release on the percent change in the number of violations. The point estimate (-0.18) is identical to the baseline estimate in percent terms -0.40/2.29 = 17.5%).”
Mirenda et al (2022)	Yes	Contract size (euros)	“The amount of public funds awarded raises by 3.4 percent.”
Norris et al (2021)	Yes	Criminal charges	“We measure both the extensive margin (using a binary indicator for the outcome ever occurring) and the intensive margin (taking the inverse hyperbolic sine, IHS, of the number of times the outcome occurred, so the coefficient is interpreted as a percent change)”
Ager et al (2021)	No interpretation	Wealth (1870 dollars)	
Arora et al (2021)	No interpretation	Publications (yearly)	
Bastos et al (2018)	No interpretation	Sales (yearly, euros)	
Fetzer et al (2021)	No interpretation	Incidents (quarterly)	
Moretti (2021)	No interpretation	Patents (yearly)	
Rogall (2021)	No interpretation	Perpetrators	
Cao and Chen (2022)	No	Rebellions per million population in 1600	They compute $\exp(\hat{\beta})-1$ and multiply by the baseline mean, then interpret this as the effect in levels

Note: this table lists papers in the *AER* estimating treatment effects for $\operatorname{arcsinh}(Y)$. The second column classifies papers by whether they interpret the units of the treatment effect as a percent/elasticity, with categories “yes”, “no”, or “no interpretation given.” The third column describes the units of the outcome before applying the $\operatorname{arcsinh}$ transformation, and the final column provides selected quotes and notes about the interpretation of the estimates. See Section 3.2.3 for details.

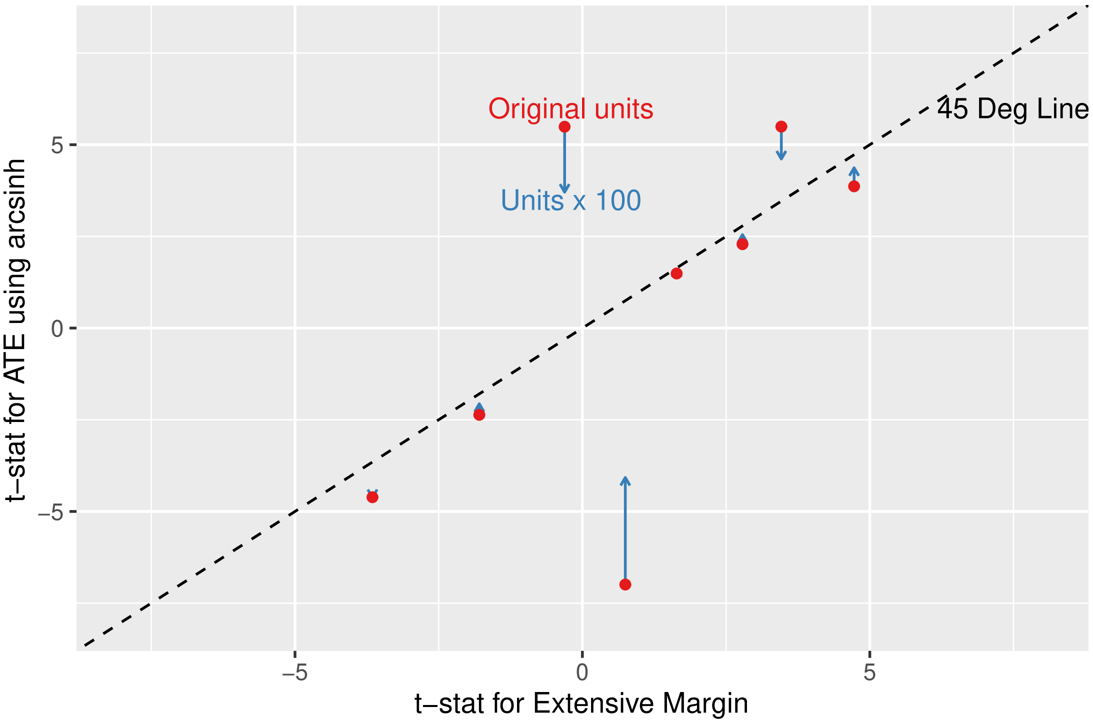


Figure C.1: t-statistics for effect on $\text{arcsinh}(Y)$, versus extensive margin t-statistic

Note: this table shows the t-statistic for the extensive margin effect on the x-axis, and the t-statistic for the treatment effect using $\text{arcsinh}(Y)$ on the y-axis. The circle shows the t-statistic using the original units, whereas the arrow shows the change if we first multiply the units by 100 before applying the arcsinh transformation. We omit two papers where there is no extensive margin. The plot shows that the t-statistics are close to the 45 degree line when the extensive margin is not close to zero, and tend to become closer when the units are made larger.

Table C.2: Change in estimated treatment effects using $\log(1+Y)$ from re-scaling the outcome by a factor of 100 in papers published in the *AER*

	Treatment Effect Using:			Change from rescaling units:	
Paper	$\log(1+Y)$	$\log(1+100\cdot Y)$	Ext. Margin	Raw	%
Azoulay et al (2019)	0.002	0.015	0.003	0.012	529
Fetzer et al (2021)	-0.138	-0.410	-0.059	-0.272	197
Johnson (2020)	-0.139	-0.408	-0.057	-0.269	194
Carranza et al (2022)	0.166	0.415	0.055	0.249	149
Cao and Chen (2022)	0.032	0.076	0.010	0.044	136
Rogall (2021)	1.109	2.015	0.195	0.906	82
Moretti (2021)	0.041	0.067	0.000	0.026	64
Berkouwer and Dean (2022)	-0.412	-0.484	0.010	-0.072	17
Arora et al (2021)	0.110	0.111	-0.001	0.001	1
Hjort and Poulsen (2019)	0.354	0.354	0.000	0.001	0

Note: this table repeats the exercise in Table 3.1 but replacing $\text{arcsinh}(Y)$ with $\log(1+Y)$ as the outcome in the second column, and $\text{arcsinh}(100Y)$ with $\log(1+100Y)$ in the third column. The fourth column shows the estimated extensive margin effect, which is identical to the fourth column of Table 3.1. The final two columns show the raw difference and percentage difference between the second and third columns. The rows are sorted based on the percentage differences. Among the papers surveyed, which by construction report at least one specification using $\text{arcsinh}(Y)$, Arora *et al.* (2021); Fetzer *et al.* (2021); Moretti (2021); Rogall (2021) also report specifications that contain $\log(1+Y)$ on the left-hand side, and Johnson (2020) reports a specification with $\log(c+Y)$ on the left-hand side, where c is the first nonzero percentile of the distribution of the observed outcome variable.