Compound Selection Decisions: An Almost SURE Approach

Jiafeng Chen, Lihua Lei, Timothy Sudijono, Liyang Sun, and Tian Xie

ABSTRACT. This paper proposes methods for *compound selection decision problems* in a Gaussian sequence model. Inspired by Stein's unbiased risk estimate (SURE), we introduce ASSURE, a family of estimators for *welfare*, defined as the expected utility of a selection decision rule. ASSURE enables robust evaluation of selection decisions. For empirical Bayes decisions derived from random-effects models of unknown parameters, ASSURE estimates their welfare even when the models are misspecified. Optimizing ASSURE-estimated welfare over a pre-specified class of decision rules further borrows strength across noisy estimates and yields decision rules with favorable regret properties. These regret properties are again robust to potential misspecification of random effects models, thereby robustifying empirical Bayes methods. We apply ASSURE to selecting Census tracts for economic mobility, identifying discriminating firms, and evaluating p-value decision rules in A/B testing.

Keywords: Compound decisions, empirical bayes, empirical welfare maximization, deconvolution
JEL codes: C14, C44, C52

Date: July 22, 2026. Chen: Department of Economics, Stanford University, jiafeng@stanford.edu; Lei: Stanford GSB, lihualei@stanford.edu; Sudijono: Department of Statistics, Stanford University, tsudijon@stanford.edu; Sun: UCL and CEMFI, liyang.sun@ucl.ac.uk; Xie: UCL, tian.xie.20@ucl.ac.uk.
The authors thank Isaiah Andrews, Xu Cheng, Nikos Ignatiadis, Guido Imbens, Patrick Kline, Charles Manski, and seminar and workshop participants at Berkeley, Duke, Georgetown, Harvard/MIT, Northwestern, NTU, Princeton, Stanford, SMU, UC Davis, UC Santa Cruz, the 2026 Interaction Conference, and NASMES 2026 for helpful comments. The authors thank Chris Walters and Toru Kitagawa in particular for discussing this paper. Jiafeng Chen is grateful for the support of the Stanford and AWS Marketing Science Lab. Lihua Lei is grateful for the support of the National Science Foundation grant DMS-2338464. Liyang Sun gratefully acknowledges support from the Economic and Social Research Council (new investigator grant UKRI607). OpenAI GPT provided valuable insights and superb research assistance. Refine.ink and Claude were used to check the paper for consistency and clarity.

1. Introduction

Applied researchers often face the following *compound selection* problem: using noisy estimates of parameters, *select* units with large underlying parameters. Many such problems can be modeled as follows: For each parallel setting $i = 1, \ldots, n$, a researcher observes a noisy estimate Y_i, a standard error σ_i, some covariates x_i, and costs k_i. The estimates (Y_i, σ_i) are for certain unknown parameters μ_i, which represent the unknown payoffs to selection decisions. Given $(Y_i, \sigma_i, x_i, k_i)$, the researcher chooses thresholds δ_i and selects all units with $Y_i > \delta_i$. They would like to set these thresholds to identify units with $\mu_i > k_i$ and avoid units with $\mu_i < k_i$.

As an example, Bergman et al. (2024) seek neighborhoods with high economic mobility to recommend to housing voucher holders. In their context, Y_i is a noisy estimate of economic mobility of Census tract i, μ_i is true economic mobility, σ_i is the estimated standard error of Y_i, x_i collects other contextual information for a tract, and k_i is the cost of selection.[1] For a fixed set of thresholds, the decision maker's utility is the average net payoff of the selection decisions:[2]

$$u(\delta_{1:n}, \mu_{1:n}) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(Y_i > \delta_i) \, (\mu_i - k_i) \, . \tag{1.1}$$

We would like to choose $\delta_{1:n}$ from the data to maximize expected utility. The statistical challenge is that expected utility depends on the unknown parameters $\mu_{1:n}$ and cannot be directly optimized.

To proceed, we impose a distributional assumption, motivated by central limit theorems for the procedure generating Y_i: Conditional on (μ_i, σ_i, x_i), the estimate Y_i is Gaussian with known variance

$$Y_i \mid \mu_i, \sigma_i, x_i \sim \mathcal{N}(\mu_i, \sigma_i^2). \tag{1.2}$$

Following the compound decision and empirical Bayes literatures (Robbins, 1951/1985; Efron, 2012; Walters, 2024)—for which (1.2) is a standard starting point—we call these statistical decision problems *compound selection problems*.

[1] We assume k_i is known—amounting to assuming that the decision maker knows their decision problem. Sometimes, when μ_i measures net payoffs, we can assume $k_i = 0$. In the case of Bergman et al. (2024), it may be reasonable to imagine that the costs k_i are equal to some value k that represents the economic mobility level for which a social planner is indifferent between incentivizing a low-income family to move to a tract with mobility k and not doing so.

[2] Compound selection problems appear in many economic applications such as the meta-analysis of experiments (Azevedo et al., 2020), teacher value-added (Chetty et al., 2014; Kwon, 2023; Cheng et al., 2025), identifying discrimination (Kline et al., 2022), and moving to opportunity (Chetty et al., 2026; Bergman et al., 2024). This problem also appears in the statistical literature under the name *empirical Bayes testing with linear loss* (Liang, 1988, 2000, 2004; Karunamuni, 1996).

Casting the problem as compound allows one to "borrow strength" by learning collective structure across $\mu_{1:n}$, thereby improving relative to conditional empirical success rules $\mathbf{1}(Y_i > k_i)$ that treat the decision problems separately (Manski, 2004; Stoye, 2012; Hirano and Porter, 2009). For instance, if $\mu_i - k_i$ is on average negative over i, decision makers should favor higher thresholds and thereby stricter selections, as a randomly drawn unit from the sample is likely to have $\mu_i < k_i$—a fundamental insight of Robbins (1951/1985).

Existing approaches for compound selection problems are either indirect for (1.1)—targeting optimal estimation of μ_i or testing hypotheses on μ_i—or they rely on potentially misspecified random-effects models for $\mu_{1:n}$. For instance, Bergman et al. (2024) select high-mobility Census tracts by thresholding shrinkage estimators for μ_i. These shrinkage estimators can be motivated by optimal estimation of $\mu_{1:n}$, à la James and Stein (1961), but it is not obvious that better prediction necessitates better selection (Mehta, 2019; Manski, 2021). Alternatively, these shrinkage estimators can be rationalized as empirical Bayes posterior means under a random-effects model in which μ_i is i.i.d. Gaussian. If the model is correct, then screening on these shrinkage estimators is optimal. However, this Gaussian random-effects model is often misspecified (Chen, 2026; Bonhomme and Denis, 2024; Müller and Watson, 2024).

This paper develops robust methods both for evaluating and for improving compound selection decisions. First, for *evaluating* selection decisions, we propose a novel estimator $\hat{W}$ of the expected utility—which we term *compound welfare*—of a given $\delta_{1:n} := (\delta(z_1), \ldots, \delta(z_n))$:

$$W(\delta_{1:n}, \mu_{1:n}) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)}[\mathbf{1}(Y_i > \delta(z_i)) \underbrace{(\mu_i - k_i)}_{\text{Net payoff in case selected}}] \quad z_i := (\sigma_i, x_i, k_i). \tag{1.3}$$

We show that $\hat{W}$ converges to W for fixed $\mu_{1:n}$ and $z_{1:n}$, where randomness solely comes from Y_i under (1.2). Thus, the validity of evaluating welfare using our estimators does not depend on correctly specifying any model for the unknown parameters $\mu_{1:n}$. This allows practitioners to robustly and accurately assess the performance of selection decisions possibly derived from restrictive models for μ.

Second, this welfare estimator also enables practitioners to *improve* selection decisions by optimizing within a user-chosen class of decision rules. So long as our estimate of welfare W is accurate, optimizing the estimated welfare leads to approximately optimal decisions within the user's chosen class. Such *regret* guarantees are again robust to misspecification, in the sense that they are valid for fixed $\mu_{1:n}$ rather than relying on a model of their distribution.

Our strategy is inspired by Stein's unbiased risk estimate (SURE, Stein, 1981). In estimation contexts, SURE is an unbiased estimator of the squared error risk of given predictions $\delta(Y_i, z_i)$ for μ_i. Since it is unbiased when integrating solely over the random draws of $Y_{1:n}$, SURE robustly evaluates prediction decisions. SURE also generates good decisions by optimizing the estimated objective over a parametrized class of predictions (Xie et al., 2012; Kwon, 2023; Cheng et al., 2025; Ghosh et al., 2025). Yet, SURE targets mean-squared error prediction and is not obviously suitable for compound selection problems. Therefore, extending SURE to selection contexts is both important and nontrivial.

Our first contribution is a new estimator $\hat{W}$ of (1.3) that extends SURE to compound selection. We propose the following class of estimators, which we term ASSURE (**A**lmos**t** **SURE**): For a kernel $\mathsf{k}_h(y) := \frac{1}{h}\mathsf{k}\left(\frac{y}{h}\right)$, define $F_h(y) := \int_{-\infty}^{y/h} \mathsf{k}(t)\,dt$ so that $F_h\left(\frac{y-\delta}{\sigma}\right)$ is a smooth approximation of $\mathbf{1}(y > \delta)$. We then define

$$\hat{W}_{\text{ASSURE}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - k_i)F_h\left(\frac{Y_i - \delta_i}{\sigma_i}\right) - \sigma_i \mathsf{k}_h\left(\frac{Y_i - \delta_i}{\sigma_i}\right). \tag{1.4}$$

Like SURE, ASSURE is constructed from Stein's identity (Stein, 1981). Unlike SURE, we apply Stein's identity by first smoothing $\mathbf{1}(Y > \delta)$, which results in bias for ASSURE. This smoothing is necessary, however, since an exactly unbiased estimator of welfare does not exist (Stefanski, 1989).

Different smoothing kernels produce different bias properties for the welfare estimator. We focus on two choices. First, our recommended choice of $\mathsf{k}(\cdot)$ is the sinc kernel $\mathsf{k}(x) := \frac{\sin x}{\pi x}$ (Stefanski and Carroll, 1990). We call the resulting estimator ASSURE*. ASSURE* has bias exponentially vanishing in the tuning parameter h. Under bandwidth $h = 1/\sqrt{2\log n}$, the squared bias and variance are both $O(1/n)$, up to log factors; thus the ASSURE* estimator converges at near-parametric rates.

Second, an alternative smoothing uses the Gaussian kernel, which results in a welfare estimator that has a natural sample-splitting interpretation: It is equivalent to a Rao–Blackwellized version of *coupled bootstrap* (Oliveira et al., 2024; Chen, 2026; Ignatiadis and Sun, 2025).[3] We thus call this estimator ASSURE-CB. Because the Gaussian kernel is only a second-order kernel, it does not fully exploit the smoothness of the estimand, resulting in slower convergence for ASSURE-CB compared to ASSURE*.

[3]Coupled bootstrap (Oliveira et al., 2024) is an instance of data thinning (Neufeld et al., 2024) or data fission (Leiner et al., 2023; Lei, 2025), where independent variates are constructed by adding noise to data. With Gaussian Y, we can let $Y_{i1} = Y_i + \epsilon Q_i$ and $Y_{i2} = Y_i - \frac{1}{\epsilon}Q_i$ for independent $Q_i \sim \mathcal{N}(0, \sigma_i^2)$. Then $Y_{i1} \perp\!\!\!\perp Y_{i2} \mid \mu_i, \sigma_i^2$. When Y_i is the sample mean of micro-data Y_{ij}, the constructed Y_{i1}, Y_{i2} can be viewed as sample means on disjoint subsamples of $\{Y_{ij}\}_j$. Proposition 1 in Chen (2026) uses this *coupled bootstrap* idea to estimate (1.1) for selection decisions that depend on Y_{i1}.

Evaluation of decision rules naturally leads to optimizing the estimated welfare within user-selected class of decision rules $\mathcal{D}$. This workflow operates with arbitrary $\mathcal{D}$ and is thus versatile. In practice, these decisions classes can come from external preference for simple decisions, such as in Kitagawa and Tetenov (2018); Sudijono et al. (2024); Crippa (2025). They may also come from a *working* (empirical) Bayesian model on the parameters $\mu_{1:n}$—similar to Kwon (2023) and Cheng et al. (2025). A flexible class, which we call $\mathcal{D}_{\text{CLOSE-GAUSS}}$ following Chen (2026), is motivated by a correlated Gaussian random effects model for $\mu_{1:n}$ (Weinstein et al., 2018), which nests many linear shrinkage estimators.

Our second contribution derives statistical guarantees for ASSURE*, in the regime where the number of parallel settings diverges ($n \to \infty$), but each setting contains a noisy estimate ($\sigma_i^2 \not\to 0$). We show that the performance gap between the estimated optimal decision and the population optimizer of (1.3) within $\mathcal{D}$—i.e., *regret*—converges to zero at minimax optimal rates, up to log factors. Applied to $\mathcal{D}_{\text{CLOSE-GAUSS}}$, decisions tuned by ASSURE* would asymptotically match the performance of the best member of $\mathcal{D}_{\text{CLOSE-GAUSS}}$, and thus outperform alternative methods for selecting from $\mathcal{D}_{\text{CLOSE-GAUSS}}$, including plug-in empirical Bayes.

These favorable properties rely on the Gaussian distributional assumption (1.2), which, though standard, assumes the central limit theorem provides a perfect approximation. To assess robustness to the approximation quality, we analyze ASSURE* as an estimator of welfare in a regime where Y_i are sample means of m i.i.d. non-Gaussian random variables (Hirano and Porter, 2009, 2016). We consider a local asymptotic regime, where both the number of parallel estimates n and the sample sizes for each estimate m diverge to infinity, but the underlying parameters are within some $1/\sqrt{m}$ neighborhood and thus have meaningful statistical ambiguity. We find that the performance of ASSURE* degrades gracefully when Gaussianity becomes approximate. Moreover, because ASSURE* exploits the Gaussianity of Y_i, it estimates welfare more accurately than competing approaches that only exploit the consistency of the noisy estimates. This analysis partially addresses questions raised by Hirano (2023) for EB and compound decision approaches.

This paper is closely related to the empirical Bayes literature (among many others, Jiang, 2020; Gu and Koenker, 2023; Walters, 2024; Kline et al., 2024; Chen, 2026). EB analyses treat $\mu_{1:n}$ as random effects and integrate over their distribution, evaluating decisions through the Bayes risk instead:

$$W_{\text{EB}}(\delta_{1:n}) = \mathbb{E}_{\mu_{1:n} \sim P}[W(\delta_{1:n}, \mu_{1:n})] \text{ for } \mu_{1:n} \mid \sigma_{1:n}, k_{1:n}, x_{1:n} \sim P.$$

EB is versatile for many decision problems—simply requiring estimating the distribution P and computing Bayes-optimal decisions relative to the estimated P. So long as this estimated prior is accurate for P, the resulting decision rule is approximately optimal over *all* decisions.

That versatility, however, depends critically on modeling the distribution of $\mu_{1:n}$ well. Failing to do so—either misspecifying how μ_i correlates with contextual information (Chen, 2026), how μ_i correlates with μ_j (Bonhomme and Denis, 2024), or whether μ_i is distributed parametrically (Gilraine et al., 2020)—could harm the selection decision. The full *joint* distribution of $\mu_{1:n}$ conditional on *all* covariates is exceedingly difficult to faithfully model and flexibly estimate; feasible empirical Bayes models are thus usually at risk of misspecification.[4]

Instead, ASSURE enables using empirical Bayes models as *working approximations*. A (parametric) empirical Bayes model generates a decision rule class $\mathcal{D}$ through its posterior.[5] Instead of plugging in an estimate $\hat{P}$, one could choose within this $\mathcal{D}$ by ASSURE. Since ASSURE's validity does not depend on whether the EB model is correctly specified, doing so asymptotically recovers the best decision within $\mathcal{D}$. Under correct specification of the EB model, ASSURE-based decisions do little harm. Under misspecification, if $\mathcal{D}$ includes simple decision rules like $\delta_i = k_i$, selection via ASSURE ensures that decisions are no worse than the simple benchmarks. In this way, ASSURE robustifies empirical Bayes by prioritizing features of P that are most relevant for the selection problem under a given $\mu_{1:n}$. We present Monte Carlo exercises in Sections 2.3 and 5 and three empirical applications in Section 6 that illustrate this point.

This paper is structured as follows. Section 2 defines the ASSURE estimators and discusses their bias and variance. Section 3 presents several theoretical guarantees on selecting a decision using ASSURE*, showing upper and lower bounds on regret. Section 4 presents two extensions on complicated decision procedures and on settings where Gaussianity holds approximately. Simulation results and empirical applications are discussed in Section 5 and Section 6. Proofs are found in the Appendix alongside additional computational details.

Notation. We use $u \star v$ to denote convolution. We use $\tilde{O}(\cdot), \tilde{O}_P(\cdot)$ as the analogues of big-O notation that ignores logarithmic factors. $\mathrm{polylog}(n)$ denotes logarithmic factors that are bounded by $(\log n)^C$ for some fixed $C > 0$. We say $f = \Omega(g)$ if $g = O(f)$. $\lesssim, \gtrsim$ record inequalities up to an absolute constant. We use $[n] := \{1, \ldots, n\}$.

[4] A similar robustness concern motivates the fixed effects panel literature (Dano et al., 2025), relative to correlated random effects.

[5] Empirical Bayesians solve selection problems by thresholding on the (estimated) posterior mean $\mathbb{E}_P[\mu_i \mid Y_i, x_i, \sigma_i] > k_i$. For Gaussian Y the posterior mean is monotone in Y_i, so the rule is equivalent to $\mathbf{1}(Y_i > \delta(z_i; P))$. If P is parametric, then this yields a parametric class $\mathcal{D}$ of decision rules.

2. Setup and Methodology

2.1. **Preliminaries and decision rules.** Let $Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2), i \in [n]$ with known σ_i, known costs k_i, with known covariates x_i. We treat both the contextual information $z_i := (x_i, k_i, \sigma_i)$ and the parameters μ_i as fixed, so all probability statements are taken over the Gaussian draws of Y_i. For simplicity, we assume that Y_i are mutually independent across i.[6] We also restrict attention to *separable* thresholds of the form $\delta_i(Y_{-i}) = \delta(z_i; \beta)$, so unit i's threshold depends on other units only through the estimated parameter β. With this parametrization, (1.1) and (1.3) can be rewritten as follows:

$$u(\beta) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(Y_i > \delta(z_i;\beta)) \cdot (\mu_i - k_i), \tag{2.1}$$

$$W(\beta) := \frac{1}{n}\sum_{i=1}^{n} (\mu_i - k_i)\,\mathbb{P}\left(Y_i \geq \delta(z_i;\beta)\right) = \frac{1}{n}\sum_{i=1}^{n} (\mu_i - k_i)\Phi\left(\frac{\mu_i - \delta(z_i;\beta)}{\sigma_i}\right). \tag{2.2}$$

These quantities are defined with respect to some parametrization of decision rules $\delta(\cdot;\beta)$, which we collect in a class $\mathcal{D} = \left\{\delta(\cdot;\beta) : \beta \in \mathcal{B} \subset \mathbb{R}^d\right\}$. Below, we enumerate several practically relevant classes that fall within this framework:

Example 1. (1) Simple truncation rules: $\mathcal{D}_{\text{threshold}} = \{\delta(z;\beta) = k + \beta : \beta \in [-M, M]\}$.

(2) Thresholding t-statistics: $\mathcal{D}_{t\text{-stat}} = \{\delta(z;\beta) = k + \beta\sigma : \beta \in [-M, M]\}$. Such rules nest selection via one-sided p-values (Mogstad et al., 2024), which is common in digital experimentation (Kohavi et al., 2020) and medicine (Manski, 2019).

(3) Finite-dimensional thresholds: $\mathcal{D}_{\text{finite}} = \text{conv}(\left\{\delta^{(1)}, \ldots, \delta^{(p)}\right\})$

(4) Gaussian empirical Bayes models: Consider Gaussian priors on $\mu_i \mid z_i \sim \mathcal{N}(m_0(z_i;\beta), s_0^2(z_i;\beta))$ and generate corresponding Bayes decision rules:

$$\mathcal{D}_{\text{CLOSE-GAUSS}} = \left\{\delta(z_i;\beta) = k_i + \frac{\sigma_i^2}{s_0^2(z_i;\beta)}(k_i - m_0(z_i;\beta)) : \beta \in \mathcal{B} \subset \mathbb{R}^d\right\}. \tag{2.3}$$

This produces a family of decision rule classes by parametrizing s_0, m_0 as different functions of β. Table 1 names some of them.

Name	m_0 specification	s_0^2 specification
Linear shrinkage	μ_0	τ^2
Fay-Herriot (Fay III and Herriot, 1979)	$x_i'\beta$	constant A
Parametric CLOSE-GAUSS (Chen, 2026; Walters, 2024)	$a_1 + a_2 \log \sigma_i$	$\exp(b_1 + b_2 \log \sigma_i)$

TABLE 1. Examples of Gaussian empirical Bayes models

[6] If $(Y_1, \ldots, Y_n) \sim \mathcal{N}((\mu_1, \ldots, \mu_n)', \Sigma)$ for some known Σ, then our arguments are generalizable by studying the conditional distributions $Y_i \mid Y_{-i}$ (Fithian and Lei, 2022; Luo et al., 2025, 2024).

2.2. **ASSURE-based welfare estimators.** We now turn to estimators $\hat{W}_n(\beta)$ for $W(\beta)$ and study their estimation error $\hat{W}_n - W$. Naturally, an estimator $\hat{W}_n$ with small estimation error is valuable for evaluating a fixed decision rule. Moreover, because $\hat{W}_n$ provides an empirical objective, practitioners can also choose β by optimizing $\hat{W}_n(\cdot)$:

$$\hat{\beta} \in \operatorname*{argmax}_{\beta \in \mathcal{B}} \hat{W}_n(\beta). \tag{2.4}$$

Accurate estimation of W also enables the estimated decision rule $\hat{\beta}$ to perform well—using bounds on $\hat{W} - W$, our subsequent theoretical results control the difference, or *regret*, between the welfare of the best decision rule in $\mathcal{D}$ and the expected utility of $\hat{\beta}$:

$$\mathsf{Regret}_n = \sup_{\beta \in \mathcal{B}} W(\beta) - \mathbb{E}_{\mu_{1:n}}[u(\hat{\beta})]. \tag{2.5}$$

For these two reasons, a good welfare estimator for the compound selection problem is the focus of this paper.

The ASSURE functional form (1.4) as kernel smoothing Stein's identity. Let $F(\cdot)$ be differentiable with derivative $f(\cdot)$. Fix some $k, \delta, \mu \in \mathbb{R}$ and $\sigma^2 > 0$, and consider $Y \sim \mathcal{N}(\mu, \sigma^2)$. Simple manipulations of Stein's identity reveal that $(Y-k)F(Y-\delta)-\sigma^2 f(Y-\delta)$ is an unbiased estimator for $(\mu - k)\mathbb{E}[F(Y - \delta)]$:

$$\mathbb{E}[(Y - k)F(Y - \delta) - \sigma^2 f(Y - \delta)] = (\mu - k)\mathbb{E}[F(Y - \delta)]. \qquad \text{(Stein's identity)}$$

On the other hand, with (2.2), estimating $W(\beta)$ reduces to estimation of a particular function of μ, which can be written as $(\mu - k)\mathbb{E}[H(Y - \delta)]$ for the *non-differentiable* $H(t) := \mathbf{1}(t > 0)$:

$$(\mu - k)\Phi\left(\frac{\mu - \delta}{\sigma}\right) = (\mu - k)\mathbb{E}[\mathbf{1}(Y > \delta)].$$

Accordingly, we seek a differentiable F whose expectation approximates $\Phi((\mu - \delta)/\sigma)$:

$$\mathbb{E}[F(Y - \delta)] \approx \Phi\left(\frac{\mu - \delta}{\sigma}\right) = \mathbb{E}\left[\mathbf{1}\left(\frac{Y - \delta}{\sigma} > 0\right)\right]. \tag{2.6}$$

For finding a suitable F, a natural idea is to smooth the step function $H(y) := \mathbf{1}(y > 0)$ with some kernel $\mathsf{k}_h(y) = \frac{1}{h}\mathsf{k}\left(\frac{y}{h}\right)$ so that the resulting function is differentiable:

$$F_h(y) := (H \star \mathsf{k}_h)(y) = \int_{-\infty}^{y/h} \mathsf{k}(t)\, dt.$$

This leads to the proposed class of estimators in (1.4), since $\sigma^2 \frac{d}{dy} F_h\left(\frac{y-\delta}{\sigma}\right) = \sigma \mathsf{k}_h\left(\frac{y-\delta}{\sigma}\right)$.

Choice of kernel and bandwidth. Which kernel should we choose? It turns out that the *sinc kernel* for (1.4) leads to optimal estimation and regret properties, building on the deconvolution literature (Kolmogorov, 1950; Tate, 1959; Liang, 2000; Pensky, 2017; Zhou and Li, 2019; Stefanski and Carroll, 1990). We call this variant ASSURE*.

Choosing bandwidth $h = 1/\sqrt{2\log n}$, ASSURE* is the following estimator:

$$\hat{W}_n(\beta) := \frac{1}{n}\sum_{i=1}^{n} w_{1/\sqrt{2\log n}}(Y_i; z_i, \beta) \tag{2.7}$$

$$\text{for } w_h(Y_i; z_i, \beta) := (Y_i - k_i)\,\mathrm{Csinc}\left(\frac{Y_i - \delta(z_i;\beta)}{\sigma_i h}\right) - \frac{\sigma_i}{h}\,\mathrm{sinc}\left(\frac{Y_i - \delta(z_i;\beta)}{\sigma_i h}\right). \tag{2.8}$$

where $\mathrm{sinc}(x) = \frac{\sin x}{\pi x}$, $\mathrm{Si}(x) = \int_0^x \frac{\sin t}{t} dt$, and $\mathrm{Csinc}(x) = \int_{-\infty}^x \mathrm{sinc}(t)\,dt = \frac{1}{2} + \frac{1}{\pi}\mathrm{Si}(x)$.

The next proposition shows that, pointwise in β, ASSURE* achieves squared error of order $O(\sqrt{\log n}/n)$.

Proposition 1. *(1) Fix μ, k, z where $\delta(z;\cdot)$ is defined. Let $Y \sim \mathcal{N}(\mu, \sigma^2)$. The bias of* ASSURE* *is exponentially decreasing in $1/h$: For all $h > 0$,*

$$\left|\mathbb{E}_\mu w_h(Y; z, \beta) - (\mu - k)\Phi\left(\frac{\mu - \delta(z;\beta)}{\sigma}\right)\right| \le |\mu - k| h^2 e^{-\frac{1}{2h^2}}.$$

The right-hand side converges to zero as $h \to 0$.

(2) For $h = 1/\sqrt{2\log n}$, for fixed β, the squared error of ASSURE* *satisfies*

$$\mathbb{E}\left[\left(\hat{W}_n(\beta) - W(\beta)\right)^2\right] \lesssim \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n}(\mu_i - k_i)^2 + \frac{\sqrt{\log n}}{n}\sum_{i=1}^{n}\sigma_i^2\right). \tag{2.9}$$

In particular, under uniformly bounded second moments, the pointwise mean-squared error is $O(\sqrt{\log n}/n)$.

The choice of the sinc kernel is motivated by its bias properties. Since the kernels are scaled so that the variances of $\hat{W}$ are $O((nh)^{-1})$, bias properties of different kernels determine estimation rates for W. Figure 1 illustrates ASSURE*'s rapidly vanishing bias as a function of h. This bias performance leads to attractive mean-squared error for ASSURE*.[7]

Unlike SURE, ASSURE* and the broader ASSURE family are not exactly unbiased for welfare. This lack of exact unbiasedness is unavoidable. We formalize in Proposition B.1 an argument in Stefanski (1989) that shows unbiased estimators (that grow slower than $e^{y^2/(2\sigma_i^2+\epsilon^2)}$ as functions of Y_i) do not exist. Short of an unbiased estimator, ASSURE* implements the next-best through the sinc kernel (Davis, 1975; Liang, 2000; Tsybakov, 2009).

[7] Theorem C.7 in a working paper draft shows the rate (2.9) is exactly minimax optimal for estimating $W(\beta)$ pointwise (Chen et al., 2026a).

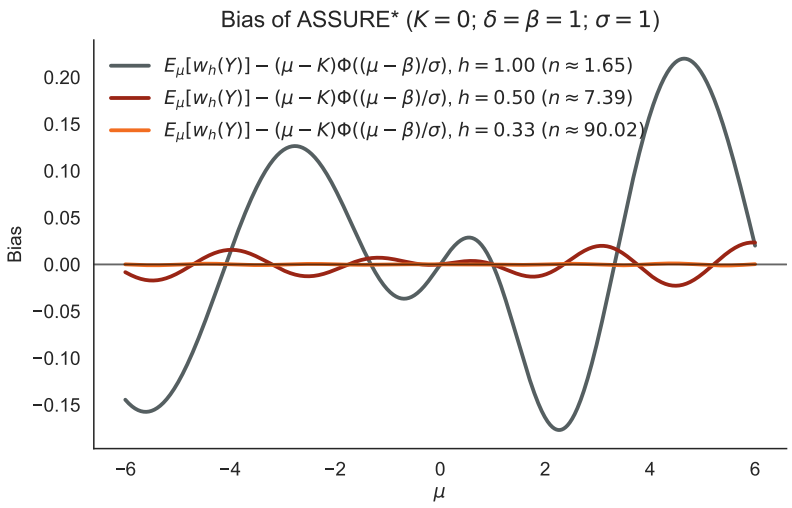


FIGURE 1. Illustration for the bias of ASSURE*. As h decreases from 1 to 0.33, the bias of ASSURE* rapidly vanishes.

To see why the sinc kernel results in low bias, return to (2.6) and consider the case with $\delta = 0, \sigma = 1$ for simplicity. Our kernel-smoothed F_h has expectation

$$\mathbb{E}_{Y\sim\mathcal{N}(\mu,1)}[F_h(Y)] = (H \star \mathsf{k}_h \star \varphi)(\mu).$$

On the other hand, the target parameter is a convolution without k_h, $\Phi(\mu) = (H \star \varphi)(\mu)$. Thus, a low-bias kernel is one that barely smoothes the Gaussian density: $\mathsf{k}_h \star \varphi \approx \varphi$.

The sinc kernel is exactly motivated by this approximation in frequency space:[8] By taking the Fourier transform of both sides and recalling $(\mathcal{F}\varphi)(\omega) = e^{-\omega^2/2}$, we would like a kernel such that

$$(\mathcal{F}\{\mathsf{k}_h\})e^{-\omega^2/2} \approx e^{-\omega^2/2}, \text{ for } \mathcal{F}\{\mathsf{k}_h\} := \int_{\mathbb{R}} \mathsf{k}_h(t)e^{-it\omega}\,dt.$$

The sinc kernel has Fourier transform $\mathcal{F}\{\mathsf{k}_h\}(\omega) = \mathbf{1}(|\omega| < 1/h)$, which only truncates the high frequency signals in φ. Since $\mathcal{F}\{\varphi\}$ has Gaussian tails, this truncation leaves it essentially unchanged—allowing the bias to decay exponentially in $1/h$.

Remark 1 (ASSURE-CB)**.** Another natural kernel—though with worse bias behavior than ASSURE*—is the Gaussian kernel $\mathsf{k}(\cdot) = \varphi(\cdot)$. Interestingly, the resulting ASSURE-CB *estimator* has an interpretation as *coupled bootstrap* (Oliveira et al., 2024; Leiner et al., 2023; Ignatiadis and Sun, 2025; Chen, 2026). For $Y \sim \mathcal{N}(\mu, \sigma^2)$, we can construct two

[8] A large literature on Gaussian estimation and deconvolution follows similar ideas (Kolmogorov, 1950; Zhou and Li, 2019; Pensky, 2017; Tate, 1959; Stefanski, 1989).

independent Gaussians by adding and subtracting an independent Gaussian noise Q

$$Y_1 = Y + \sigma h Q \quad Y_2 = Y - \frac{\sigma}{h} Q \quad Q \sim \mathcal{N}(0,1) \implies Y_1 \perp\!\!\!\perp Y_2. \tag{2.10}$$

This *coupled bootstrap* procedure is a synthetic version of sample-splitting.[9]

With (2.10), the welfare of selection decisions based on $Y_1 > \delta$ can be unbiasedly estimated by $(Y_2 - k)\mathbf{1}(Y_1 > \delta)$, since Y_2 acts as fresh "testing data" that is independent of the "training data" Y_1. Because Y_1 is slightly noisier than the original estimate Y, this estimator is biased for the welfare of $\mathbf{1}(Y > \delta)$.

The estimator ASSURE-CB—choosing $\mathsf{k}(\cdot) = \varphi(\cdot)$ in (1.4)—is exactly the Rao–Blackwellization of coupled bootstrap:

$$\mathbb{E}[(Y_2 - k)\mathbf{1}(Y_1 > \delta) \mid Y] = (Y - k)\Phi\left(\frac{Y - \delta}{\sigma h}\right) - \frac{\sigma}{h}\varphi\left(\frac{Y - \delta}{\sigma h}\right).$$

Because the Gaussian kernel is only a second-order kernel, ASSURE-CB has $O(h^2)$ bias and therefore converges more slowly than ASSURE*.[10]

2.3. **Compound decisions, policy learning, and empirical Bayes.** We now contextualize our approach in the literature and compare it against a few alternatives. We also disambiguate our regret notion (2.5) from the regret notions in these literatures. We illustrate these comparisons using a classic example from Robbins (1951/1985).

Example 2 (Robbins (1951/1985))**.** The parameters are binary, $\mu_{1:n} \in \{-1, +1\}^n$, of which ρn have $\mu_i = +1$ and the remaining units have $\mu_i = -1$. For each unit i, suppose we observe $Y_i \sim \mathcal{N}(\mu_i, 1)$ and set $k_i = 0$ for simplicity. Restrict decision rules to the class $\delta_i(\beta) = \beta$ in the decision problem (1.1). The welfare in this setting is equal to $W(\beta) = \rho\Phi(1 - \beta) - (1 - \rho)\Phi(-1 - \beta)$. If $\mu_{1:n}$—or equivalently ρ—is known, then the optimal decision for (1.3) sets $\beta^* = \frac{1}{2}\log\frac{1-\rho}{\rho}$.

Treatment choice. Example 2 can be thought of as a compound version of the "testing an innovation" problem in Manski (2009), where μ_i could represent the average treatment effect of a new treatment relative to a known status quo. Modeling this problem as *compound* changes our statistical perspective relative to the treatment choice literature. That literature often seeks decision rules that optimize for worst-case FreqRegret (e.g., in Ishihara and

[9] If Y is a sample mean of Gaussian micro-data, $Y = \frac{1}{m}\sum_{j=1}^{m} Y_{(j)}$ for $Y_{(j)} \sim \mathcal{N}(\mu, m\sigma^2)$, then sample-split means over $Y_{(1)}, \ldots, Y_{(m)}$ can be represented in the form (2.10).

[10] ASSURE* is $\tilde{O}(n^{-1/2})$-consistent for welfare, whereas ASSURE-CB is only $\tilde{O}(n^{-2/5})$-consistent. Appendix D in a working paper draft of this paper provides a more formal comparison (Chen et al., 2026a).

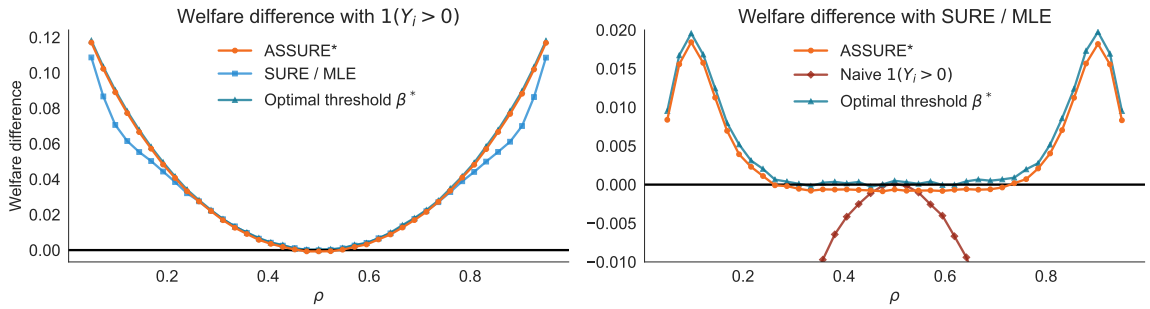


FIGURE 2. Welfare comparisons of different methods for fitting decision rules for Example 2, as a function of the proportion ρ of $\mu_i = +1$.

Notes. We consider decision rules—ASSURE*, MLE/SURE, oracle $\beta^* = \frac{1}{2}\log\frac{1-\rho}{\rho}$, and $\mathbf{1}(Y_i > 0)$—and compare their performances. The left panel shows differences against $\mathbf{1}(Y_i > 0)$ and the right panel shows differences against fitting β with SURE/MLE. In particular, SURE/MLE rationalizes the selection $\mathbf{1}(Y_i > \beta)$ with a Gaussian random effects model $\mu_i \sim \mathcal{N}(m_0, s_0^2)$ and fits (m_0, s_0^2) through MLE or through SURE; they imply identical selection decisions. □

Kurisu, 2022), where

$$\mathsf{FreqRegret}_n := \mathbb{E}_{\mu_{1:n}}\left[\frac{1}{n}\sum_{i=1}^{n}\max(0,\mu_i - k_i) - u(\hat{\beta})\right] \overset{\text{Example 2}}{=} \rho - \mathbb{E}[u(\hat{\beta})].$$

By contrast, our goal is asymptotically vanishing regret in (2.5), uniformly over suitable sequences $(\mu_i, z_i)_{i=1}^n$. Specializing to Example 2, (2.5) is instead

$$\mathsf{Regret}_n = \rho\Phi\left(1-\beta^*\right) - (1-\rho)\Phi\left(-1-\beta^*\right) - \mathbb{E}[u(\hat{\beta})].$$

These desiderata are distinct. Our later results show that ASSURE* ensures vanishing Regret_n, but its worst-case FreqRegret for any finite n may not be optimal. Conversely, the empirical success rule $\mathbf{1}(Y_i > 0)$ achieves minimax-FreqRegret (Proposition 1, Stoye, 2012), but its regret (2.5) does not vanish: $\sup_{\mu_{1:n}} \mathsf{Regret}_n = \Omega(1)$. Thus, along many sequences of $\mu_{1:n}$, eventually ASSURE* achieves higher welfare than the empirical success rule. Figure 2 illustrates this with Example 2: The welfare of ASSURE* closely tracks that of the oracle β^* and exceeds that of the empirical success rule for all $\rho \not\approx 0.5$. ASSURE* slightly underperforms the empirical success rule when ρ is close to 0.5.

This improvement comes from a key insight of Robbins (1951/1985). Intuitively, if n is large, certain aggregate quantities such as $\frac{1}{n}\sum_{i=1}^n(\mu_i - k_i)^q$ are precisely estimable from the data and carry information about the configuration of $\mu_{1:n}$. For instance, the average $\frac{1}{n}\sum_{i=1}^n(\mu_i - k_i)$ can be informative: A positive estimate provides evidence that selection is net-beneficial on average. If we are confident that $\frac{1}{n}\sum_{i=1}^n(\mu_i - k_i) > 0$, more lenient

decision rules such as $\mathbf{1}(Y_i \geq -\epsilon)$ can improve compound welfare, since any mildly negative realizations of Y_i are then more likely due to chance. This is especially stark in Example 2, where $\frac{1}{n}\sum_i Y_i \approx \frac{1}{n}\sum_{i=1}^n (\mu_i - k_i) = 2\rho - 1$ is directly informative of the key quantity ρ. ASSURE* precisely exploits this kind of information—even in settings without simple statistics like ρ—through estimating the objective function.

Empirical Bayes (G-modeling). A popular framework for compound decision problems is *empirical Bayes*, which models μ_i as random effects (Efron, 2012). This additional distributional structure implies that optimal decision rules are posterior quantities involving the estimable distribution of μ_i. Formally, EB methods assume that $\mu_{1:n} \mid z_{1:n} \sim P$. If the distribution P were known, then the optimal decision with respect to expected welfare compares the posterior mean of μ_i to k_i: $\mathbb{E}_P[\mu_i \mid Y_i, z_i] > k_i$. Thus, EB methods propose estimating P and plug $\hat{P}$ into posterior decision rules.

However, this recipe hinges on modeling and estimating the *entire joint distribution* of $\mu_{1:n} \mid z_{1:n}$ well. In full generality, $\mu_{1:n} \mid z_{1:n}$ can be highly complex: μ_i, μ_j may be correlated (Bonhomme and Denis, 2024; Müller and Watson, 2024); they may be flexibly predicted by rich covariates (Chen, 2026); or μ_i's randomness may simply be questionable if we observed the entire population. Even nonparametric empirical Bayes methods necessarily make some compromises,[11] and one may be concerned that these compromises harm resulting decisions. Moreover, it is unclear whether estimating *all* features of P well is necessary for a given decision problem; perhaps different decision problems call for different ways of estimating P. Some features of P, such as $\mathrm{Var}(\mu_i)$ in Gaussian EB, might be awkward to estimate. Unbiased estimates for $\mathrm{Var}(\mu_i)$ might be negative with positive probability, often motivating ad hoc methods for truncation or weighting (Armstrong et al., 2022; Chetty et al., 2026).

Our procedure is thus complementary to empirical Bayes. EB-style modeling for μ_i is apt for generating reasonable decision rule classes $\mathcal{D}$—indeed optimal ones if the model on μ_i happens to be correct. (A leading family of such EB-derived decision rules is the CLOSE-GAUSS class in Example 1.) Choosing a member of $\mathcal{D}$ with ASSURE attains guarantees for the compound selection problem directly—thus not requiring practitioners to take the random effects model fully seriously. When the random-effects model is correct, tuning via ASSURE does little harm. These advantages are attractive features of SURE (Ghosh et al., 2025; Cheng et al., 2025) that ASSURE inherits.

In terms of theoretical analysis, the EB literature studies a different regret criterion that averages over draws of $\mu_{1:n}$ and benchmarks against the oracle Bayes rule (Jiang, 2020;

[11] For instance, Chen (2026) assumes $\mu_{1:n}$ are mutually independent and restricts dependence on covariates to a location-scale family.

Soloff et al., 2024; Chen, 2026):

$$\mathsf{EBRegret}_n = \mathbb{E}_{\substack{\mu_{1:n}\sim P \\ Y_i|\mu_i,z_i\sim\mathcal{N}(\mu_i,\sigma_i^2)}}\left[\frac{1}{n}\sum_{i=1}^n \max(0, \mathbb{E}_P[\mu_i \mid Y_i, z_i] - k_i) - u(\hat{P})\right],$$

where $u(\hat{P}) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(\mathbb{E}_{\hat{P}}[\mu_i \mid Y_i, z_i] > k_i)(\mu_i - k_i)$. Compared to the regret (2.5), $\mathsf{EBRegret}_n$ additionally integrates over $\mu_{1:n} \mid z_{1:n} \sim P$ and chooses the benchmark as $\mathbf{1}(\mathbb{E}_P[\mu_i \mid Y_i, z_i] > k_i)$. When the class of decision rules $\mathcal{D}$ is exactly the class of Bayes decisions, controlling (2.5) automatically controls $\mathsf{EBRegret}_n$:

$$\mathsf{EBRegret}_n = \mathbb{E}_{\substack{\mu_{1:n}\sim P \\ Y_i|\mu_i,z_i\sim\mathcal{N}(\mu_i,\sigma_i^2)}}\left[W(P) - u(\hat{P})\right] \le \mathbb{E}_{\mu_{1:n}\sim P}\left[\mathsf{Regret}_n\right].$$

When the prior is misspecified, however, $\mathsf{EBRegret}$ ceases to be relevant. This is exactly where ASSURE* provides value. To illustrate, in Example 2 and Figure 2, we consider using a misspecified Gaussian prior under which $\mu_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(m_0, s_0^2)$. The implied selection rule, given by $\mathbf{1}\left(\mathbb{E}[\mu_i \mid Y_i] = \frac{s_0^2}{s_0^2+1}Y_i + \frac{m_0}{s_0^2+1} > 0\right)$, is therefore also misspecified. An empirical Bayesian may estimate (m_0, s_0^2) by maximum likelihood.[12] Figure 2 shows that this EB-MLE/SURE approach underperforms ASSURE*. This is because the misspecification of the prior *misaligns* optimal selection and estimation decisions (Manski, 2021).[13]

3. Theoretical Assurances

This section presents statistical guarantees for ASSURE*. Our main results are upper and matching lower bounds for regret (2.5). We first present upper and lower bounds achieving $\tilde{O}(1/\sqrt{n})$ regret, which hold under weak assumptions. Additional assumptions on $\mu_{1:n}$ allow for a faster rate by exploiting an analogue of the margin condition (Audibert and Tsybakov, 2005).

3.1. Main regret bound.

Our main result is the following $(1/\sqrt{n})$-regret bound. To state the theorem, for $q \ge 1$, define $m_q := \|m\|_{q,n} := \left(\frac{1}{n}\sum_{i=1}^n |\mu_i - k_i|^q\right)^{1/q}$.

Theorem 1 ($\tilde{O}(1/\sqrt{n})$-regret for ASSURE*)**.** *Suppose that the class of decision rules* $\left\{\delta(z;\beta) : \beta \in \mathcal{B} \subset \mathbb{R}^d\right\}$ *is VC-subgraph with index* $V(\mathcal{D})$ *and envelope function* D*. Let* $\hat{\beta} \in \operatorname{argmax}_{\beta\in\mathcal{B}} \hat{W}(\beta)$ *be the optimal decision chosen by* ASSURE*. *If* $\sigma_{1:n}, \sigma_{1:n}^{-1}$ *and the envelope* D *are all uniformly bounded by an absolute constant, then for some further constant*

[12] Interestingly, in Example 2, this yields the same selection decisions as choosing (m_0, s_0^2) by SURE.

[13] To be sure, some of the misspecification can be remedied by nonparametric modeling (Jiang and Zhang, 2009). Even with a nonparametric empirical Bayes model, one could continue to use ASSURE to robustly evaluate and further fine-tune decisions. We discuss an implementation in Section 4.

C,

$$\mathsf{Regret}_n \lesssim CV(\mathcal{D})(m_2 + m_4^4)\frac{(\log n)^{1/4}\sqrt{\log\log n}}{\sqrt{n}}.$$

Thus, further assuming that $|\mu_i - k_i|, V(\mathcal{D})$ are uniformly bounded, ASSURE* achieves $\frac{(\log n)^{1/4}\sqrt{\log\log n}}{\sqrt{n}}$ regret. We check in Appendix D.1 that the leading decision rule classes enumerated in Example 1 are VC-subgraph.

Broadly speaking, Theorem 1 follows a standard empirical risk minimization argument. Regret can be controlled by controlling the empirical process $\sup_{\beta\in\mathcal{B}}|\hat{W}(\beta) - W(\beta)|$, which is manageable as $\mathcal{D}$ is assumed to be VC-subgraph. There are two complications in our setup that require care. First, standard empirical risk minimization (e.g., Theorem 8.4.4 in Vershynin (2009)) controls—in our notation—the gap $W(\beta^*) - \mathbb{E}[W(\hat{\beta})]$. This is a different quantity than (2.5) since $\mathbb{E}[W(\hat{\beta})] \neq \mathbb{E}[u(\hat{\beta})]$. Second, since (z_i, μ_i) are treated as constants, the random variables Y_i are independent but not identically distributed, complicating standard i.i.d. empirical process theory.

3.2. **Lower bound.** Theorem 1 is unimprovable in the worst case, up to log factors.

Theorem 2 (Lower bounds for regret)**.** *Fix $\sigma_i = 1$ and $k_i = 0$ for all i. Without loss of generality, assume $\delta(z_i;\beta) = \beta \in \mathbb{R}$. Let $M > 0$, then over* all *decision rules $a_i(Y_1,\ldots,Y_n) \in [0,1]$, for a constant $C = C_M$,*

$$\inf_{a(\cdot):a_i\in[0,1]}\sup_{\mu_{1:n}\in[-1,1]^n}\mathsf{Regret}_n(a_{1:n}) \geq \frac{C}{\sqrt{n}},$$

where

$$\mathsf{Regret}_n(a_{1:n}) := \sup_{|\beta|\leq M} W(\beta;\mu_{1:n}) - \frac{1}{n}\sum_{i=1}^n \mu_i \mathbb{E}_{\mu_{1:n}}[a_i(Y_1,\ldots,Y_n)]. \tag{3.1}$$

Here, the regret notion in Theorem 2 compares the welfare of an arbitrary randomized selection decisions $a_i(Y_1,\ldots,Y_n)$ to the welfare of the best *threshold decision* restricted to $|\beta| \leq M$. Since (3.1) removes constraints on the decision maker's actions but imposes constraints on the optimal benchmark, lower bounds of (3.1) lower bound regret.[14]

Theorem 2 is derived by an application of Le Cam's two-point method, considering cases in which, for some $h > 0$, either all $\mu_i = h/\sqrt{n}$ or all $\mu_i = -h/\sqrt{n}$.[15] In this case, the optimal β^* is either $+M$ or $-M$, meaning that $W(\cdot)$ is maximized on the boundary of the

[14]Since the setting in Theorem 2 does not have contextual information, the optimal separable decision rule does take the threshold form. An analogous regret notion was considered by Polyanskiy and Wu (2021) in their (12) for the mean square case.

[15]An analogous construction is applicable to an empirical Bayes setup, which partially resolves a conjecture regarding the tightness of Theorem 4 in Chen (2026).

parameter space. In fact, when β^* is a local maximum, the upper and lower bounds can be improved to $\tilde{O}(1/n)$. We turn to this faster rate upper bound next.

3.3. **Fast rates for ASSURE*.** As the lower bound hints, the performance for ASSURE* can be better than the $\tilde{O}(n^{-1/2})$ rate predicts, when the population optimal decision is interior. The additional assumptions needed are reminiscent of the *margin condition* for fast rates in classification and treatment choice (Audibert and Tsybakov, 2005, 2007; Kitagawa and Tetenov, 2018; Ponomarev and Semenova, 2024; Crippa, 2025).[16]

For simplicity, we will restrict to a case where the decision rules are parametrized by a scalar parameter ($\mathcal{B} \subset \mathbb{R}$). We leave the multidimensional extension to future work.

Assumption 1. *(1) (Boundedness). Fix an absolute constant $B > 1$. Assume that $\mu_i, k_i \in [-B, B]$ and $\sigma_i \in [B^{-1}, B]$. Furthermore, assume that $\delta(\cdot, z_i)$ are thrice-differentiable for each i and $\delta(\cdot, z_i)$ and its three derivatives are uniformly bounded by B.*

(2) (Unimodality). There exists a unique local and global maximizer β^* in the interior of $\mathcal{B}$ at which $W(\beta^*) > 0$.

(3) (Curvature). Fix some absolute $\kappa > 0$. $W''(\beta^*) < -\kappa$.

(4) (Well-separated maximum). Fix some absolute $\xi > 0$. Then $\sup_{|\beta - \beta^*| > \frac{\kappa}{B}} W(\beta) < W(\beta^*) - \xi$.

Theorem 3 (Fast Rates for ASSURE*)**.** *Suppose Assumption 1 holds. Lastly, suppose that δ and its derivatives $\delta(\beta, z), \delta'(\beta, z), \delta''(\beta, z)$ form VC-subgraph function classes $\mathcal{D}, \mathcal{D}', \mathcal{D}''$ with indices $V(\mathcal{D})$, $V(\mathcal{D}')$, and $V(\mathcal{D}'')$, respectively. Then the regret of ASSURE* satisfies*

$$\mathbb{E}\left[W(\beta^*) - u(\hat{\beta})\right] \leq C \frac{(\log n)^6}{n}.$$

for constant $C > 0$ that depends only on B, κ, ξ in Assumption 1 and $V(\mathcal{D}'), V(\mathcal{D}''), \mathcal{B}$.

Theorem 3 exploits the fact that $\hat{\beta}$ is a local maximum, which enables the approximation $W(\hat{\beta}) \approx W(\beta^*) + \frac{1}{2} W''(\beta^*)(\hat{\beta} - \beta^*)^2$. Analogously Taylor approximating $0 = \hat{W}'(\hat{\beta})$ shows that $\hat{\beta} - \beta^* = O_P(|\hat{W}' - W'|)$; therefore, regret is approximately $O_P(|\hat{W}' - W'|^2)$. Since it can be shown that ASSURE* also estimates W' at $\tilde{O}(n^{-1/2})$ rates, we obtain the rate in Theorem 3. Appendix B.2 illustrates a class of decision rules and lower-level assumptions on $\mu_{1:n}$ for which Theorem 3 applies. The log factors in Theorem 3 are crude.

Finally, we present a lower bound for the fast rate in Theorem 3.

[16] There, the margin condition bounds the density of difficult examples near the optimal classification boundary. Intuitively, with many units near the classification boundary, it is difficult to distinguish between different candidate parameter values, meaning that the objective function lacks a cleanly separated maximizer. This is exactly the failure mode that Theorem 2 exploits for the lower bound, which we rule out in Assumption 1.

Theorem 4 (Fast Regret Lower Bounds)**.** *Fix $\sigma_i = 1$ and $k_i = 0$. Consider decision rules $\delta(z_i; \beta) = \beta$ for $\beta \in \mathbb{R}$. Let $\Theta = \Theta_{n,\kappa,\xi} \subseteq \mathbb{R}^n$ be the space of $(\mu_1, \ldots, \mu_n)$ satisfying the assumptions Assumptions 1(1) to 1(4) for a fixed ξ, κ. Then there exists choices of ξ and κ such that*

$$\inf_{a(\cdot):a_i \in \{0,1\}} \sup_{\mu_{1:n} \in \Theta} \mathsf{Regret}_n(a_{1:n}) = \Omega(n^{-1}),$$

with $\mathsf{Regret}_n(a_{1:n})$ defined as in (3.1).

The proof of Theorem 4 lower bounds the compound regret by the Bayes regret for a well-chosen prior and then further lower bounds the Bayes regret by Le Cam's two point argument.[17]

4. Extensions

This section presents two extensions. We first consider robustness to Gaussian approximation. We then outline how ASSURE* can be applied to more complex decision rules via leave-one-out cross-validation.

4.1. Robustness to Gaussianity.

So far, we have operated under the Gaussian sequence model (1.2). Gaussianity of Y_i is motivated by applications where the noisy estimates are averages of micro-data, for which the central limit theorem justifies a Gaussian approximation. However, since ASSURE methods rely on the Gaussianity of Y_i,[18] a reasonable worry is whether our methods are sensitive to the quality of the Gaussian approximation.

This section thus considers Y_i explicitly as averages of i.i.d. random variables, and analyzes a feasible version of ASSURE* as an estimator of a rescaled version of welfare à la Hirano and Porter (2009). Our focus is on welfare estimation here, but we expect these results translate to regret control like Theorems 1 and 3. We find that the behavior of ASSURE* degrades smoothly in the quality of the Gaussian approximation. Moreover, because ASSURE* exploits approximate Gaussianity of Y_i, it outperforms simple estimators of welfare that do not exploit this information.

[17] Section 3.1 of Liang (2004) states a $\Omega(n^{-1})$ lower bound for a related but distinct regret quantity, where the decision rule is learned from data points $Y_1, \ldots, Y_n$ and evaluated on a newly drawn decision problem (μ_{n+1}, Y_{n+1}). We are concerned with an "in-sample" version of Liang (2004)'s regret, motivating a distinct proof strategy.

[18] The same would be true of using SURE to tune decision rules that are nonlinear in Y_i. If decision rules are restricted to be linear in Y_i, the quadratic nature of SURE implies that consistent estimation of certain moments is sufficient.

We consider a stylized data-generating process where the micro-data $Y_{ij} \overset{\text{i.i.d.}}{\sim} G_i^{(m)}$, for heterogeneous distributions $G_1^{(m)}, \ldots, G_n^{(m)}$ and $j = 1, \ldots, m$.[19] The decision-maker observes the sample mean of cell i, $\bar{Y}_i := \frac{1}{m}\sum_{j=1}^m Y_{ij}$. For simplicity, assume that costs are zero $k_i = 0$. We will consider a local asymptotic regime similar to Hirano and Porter (2009) and Chen et al. (2026b), where the micro-data distribution $G_i^{(m)}$ has mean $\mu_i^{(m)} := \mu_i/\sqrt{m}$ and standard deviation σ_i.

Let $F_i^{(m)}(t) := P_{G_i^{(m)}}((\bar{Y}_i - \mu_i^{(m)})/\sigma_i \le t)$ be the CDF of the null t-statistic and let $\delta_i = \sqrt{m}\delta_i^{(m)}$ be a rescaled version of the threshold. This rescaling allows us to rescale the welfare and write it as:

$$W(\delta) := \frac{1}{n}\sum_{i=1}^n \underbrace{(\sqrt{m}\cdot\mu_i^{(m)})}_{\text{scaling by } \sqrt{m}} \mathbb{E}_{G_i^{(m)}}[\mathbf{1}(\bar{Y}_i > \delta_i^{(m)})] = \frac{1}{n}\sum_{i=1}^n \mu_i \bar{F}_i^{(m)}\left(\frac{\delta_i - \mu_i}{\sigma_i}\right), \qquad (4.1)$$

$$\text{for } \bar{F}_i^{(m)}(t) := 1 - F_i^{(m)}(t)$$

The welfare (4.1) is the analogue of (1.1); here, the selection probability is calculated via the actual finite-sample distribution $G_i^{(m)}$ rather than its Gaussian approximation. Relative to the parameters $\mu_i^{(m)}$ in their native scale, (4.1) scales welfare by $\sqrt{m}$, accentuating differences at the $1/\sqrt{m}$ scale.

For fixed δ, we study the estimation performance of a feasible version of ASSURE* for (4.1) and compare it to reasonable alternatives. Let

$$\hat{\sigma}_i^2 = \frac{1}{m-1}\sum_{j=1}^m (Y_{ij} - \bar{Y}_i)^2$$

be the estimated sample variance and let $Y_i := \sqrt{m}\bar{Y}_i$. The feasible version of ASSURE* (2.7) is then, for $h = 1/\sqrt{2\log n}$,

$$\hat{W}(\delta) = \frac{1}{n}\sum_{i=1}^n \hat{w}_{h,i}(Y_i;\delta) := \frac{1}{n}\sum_{i=1}^n Y_i \,\mathrm{Csinc}\left(\frac{Y_i - \delta_i}{\hat{\sigma}_i h}\right) - \frac{\hat{\sigma}_i}{h}\,\mathrm{sinc}\left(\frac{Y_i - \delta_i}{\hat{\sigma}_i h}\right). \qquad (4.2)$$

Theorem 5 (ASSURE* bias with non-Gaussian micro-data)**.** *Under the preceding setup, fix $\delta_1, \ldots, \delta_n$. Suppose $G_i^{(m)}$ have mean $\mu_i/\sqrt{m}$, variance σ_i^2, and third and fourth absolute central moments $\mu_{3,i}, \mu_{4,i}$. The bias of* ASSURE* *satisfies*

$$\left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{G_i^{(m)}}\left[\hat{w}_{h,i}(\sqrt{m}\bar{Y}_i;\delta)\right] - W(\delta)\right| \le C\left(\frac{\log(n)^2}{\sqrt{m}} + \frac{1}{n\log n}\right),$$

where the constant C depends on $\min_i \sigma_i$ and averages of polynomials in $|\mu_i|, \sigma_i, \mu_{3,i}, \mu_{4,i}$.

[19] Extending this setup to heterogeneous cell-sizes m_i is notationally cumbersome but straightforward.

By independence and the bound $\mathrm{Var}(\hat{w}_{h,i}(Y_i;\delta)) = O((\log n)^2)$, we have that $\mathrm{Var}(\hat{W}(\delta)) = O\left(\frac{(\log n)^2}{n}\right)$. The mean-squared error of ASSURE* as an estimator of $W(\delta)$, for fixed decisions, is thus

$$\mathbb{E}[(\hat{W} - W)^2] = O\left(\frac{\log(n)^4}{m} + \frac{\log(n)^2}{n}\right).$$

The $(\log n)^4/m$ term reflects the fact that the Gaussianity of Y_i is only approximate, which disappears smoothly as $m \to \infty$ and Y_i becomes Gaussian. At this scaling, this term also reflects the benefit of exploiting the approximate Gaussianity. To see this, we can contrast with the mean-squared error of a simple estimator that does not use this information: Consider an estimator that simply replaces μ_i with $\bar{Y}_i$:[20]

$$\hat{W}_{\mathrm{CES}}(\delta) = \frac{1}{n}\sum_{i=1}^{n} \sqrt{m}\bar{Y}_i \mathbf{1}(\bar{Y}_i > \delta_i^{(m)}).$$

Under mild assumptions, it is not difficult to show that the bias at this scale does not vanish,[21] leaving the mean-squared error no better than $1 + 1/n$:

$$|\mathbb{E}[\hat{W}_{\mathrm{CES}}] - W| = \Omega(1) \quad \mathbb{E}[(\hat{W}_{\mathrm{CES}} - W)^2] = \Omega\left(1 + \frac{1}{n}\right).$$

Therefore, as an estimator for the welfare (4.1), ASSURE* strictly dominates these welfare estimators whenever $m \gg \log(n)^4$ and $(m, n) \to \infty$.

4.2. **Ensembling complex decision rules.** In our second extension, we discuss how ASSURE can enable practitioners to *ensemble* multiple decision rules. For instance, one decision rule may come from a nonparametric EB model without modeling z-dependence—in which $\mu_i \mid z_i \overset{\text{i.i.d.}}{\sim} G_1$ for some G_1 modeled nonparametrically, perhaps estimated through nonparametric maximum likelihood. Another may come from a parametric EB model that considers contextual information—$\mu_i \mid z_i \sim G_2(\cdot \mid z_i)$, where $G_2(\cdot)$ is parametric CLOSE-GAUSS and estimated via plug-in methods. Yet another method may simply try to estimate $\mathbb{E}[Y_i \mid z_i] = \mathbb{E}[\mu_i \mid z_i]$ flexibly.

This is made tractable with leave-one-out cross-fitting. For the example above, for each $i = 1, \ldots, n$, let $m_{\hat{G}_1}^{(-i)}(y, z)$ be a posterior mean function fitted on the data (Y_{-i}, z_{-i}). Let $m_{\hat{G}_2}^{(-i)}(y, z)$ denote the analogous leave-one-out posterior mean for a second empirical Bayes method. Finally, let $\hat{f}^{(-i)}$ denote a regression model trained using the data (Y_{-i}, z_{-i}). We may consider the class of decision rules which selects a unit whenever the ensembled

[20] Maximizing $\hat{W}_{\mathrm{CES}}$ over fixed thresholds $\delta_i = \beta$ results in the empirical success rule $\mathbf{1}(\bar{Y}_i > 0)$ (Manski, 2004).

[21] Similar bias and mean-squared error properties apply to $\hat{W}_{\text{plug-in}}(\delta) := \frac{1}{n}\sum_{i=1}^{n} Y_i \Phi\left(\frac{Y_i - \delta}{\hat{\sigma}_i}\right)$.

prediction is greater than the implementation cost:

$$b_1 m_{\hat{G}_1}^{(-i)}(Y_i, z_i) + b_2 m_{\hat{G}_2}^{(-i)}(Y_i, z_i) + b_3 \hat{f}^{(-i)}(z_i) \geq k_i, \tag{4.3}$$

where the parameters $\beta = (b_1, b_2, b_3)$ are restricted to be in the unit simplex. By the monotonicity of the posterior mean function for Gaussian observations, the decision (4.3) is equivalent to $Y_i \geq \delta_{\text{ensemble}}(z_{1:n}, Y_{-i}; \beta)$ for some function δ_{ensemble}. ASSURE can be used to tune the parameter β.

To that end, let $\delta(z_{1:n}, Y_{-i}; \beta)$ be some parametrized threshold that depends on all $z_{1:n}$ and other units Y_{-i}. Following (2.7), let

$$w_h(Y_i; \beta, z_i, Y_{-i}) := (Y_i - k_i)\,\text{Csinc}\left(\frac{Y_i - \delta(z_{1:n}, Y_{-i}; \beta)}{\sigma_i h}\right) - \frac{\sigma_i}{h}\,\text{sinc}\left(\frac{Y_i - \delta(z_{1:n}, Y_{-i}; \beta)}{\sigma_i h}\right).$$

ASSURE* then estimates welfare by averaging $w_h(Y_i; \beta, z_i, Y_{-i})$ across i. Then, by the law of iterated expectations conditioning on Y_{-i}, we have that ASSURE* remains a low-bias estimator for the welfare of this ensemble decision:

$$\begin{aligned}\mathbb{E}\hat{W}(\beta) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E} w_h(Y_i; \beta, z_i, Y_{-i}) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}\left[w_h(Y_i; \beta, z_i, Y_{-i}) \mid Y_{-i}\right]\right] \\ &\approx \frac{1}{n}\sum_{i=1}^{n} (\mu_i - k_i)\mathbb{E}\left[\Phi\left(\frac{\mu_i - \delta(z_{1:n}, Y_{-i}; \beta)}{\sigma_i}\right)\right]. \\ &\qquad \text{(Proposition 1)}\end{aligned}$$

The approximation in the last line is suggested by Proposition 1.

Under mild stability assumptions, ASSURE is expected to perform no worse asymptotically than the constituent models since the decision class nests each model individually. This ensemble class thus enables ASSURE to improve on or robustify any set of methods. We defer a detailed theoretical analysis of the ensembling method for future work.

5. Simulation

To illustrate ASSURE methods as robustifying empirical Bayes, we consider a setup where $\mu_{1:n}$ do come from some random effect distribution, but we misspecify this distribution. In particular, we consider[22]

$$\mu_i \mid \sigma_i \sim \mathcal{N}(m_0(\log \sigma_i), 0.2) \quad m_0(x) = -\frac{x}{2} + 3e^{-\frac{1}{2}\left(\frac{x-1}{0.12}\right)^2} \quad \log \sigma_i \in [-1.15, 1.15]. \tag{5.1}$$

Assume costs k_i are zero. The class of decision rules we consider is

$$\mathcal{D} = \left\{\delta(\sigma_i; \beta) = -\sigma_i^2(a + b \log \sigma_i) : (a, b) \in \mathbb{R}^2\right\}.$$

[22] The marginal distribution of $\log \sigma_i$ is drawn from $\mathcal{N}(0.5, 0.35^2)$, truncated to $[-1.15, 1.15]$.

$\mathcal{D}$ is motivated by the misspecified model where we model m_0 as linear:

$$\mu_i \mid \sigma_i \sim \mathcal{N}(\tilde{a} + \tilde{b} \log \sigma_i, s_0^2), \tag{5.2}$$

for parameters $(\tilde{a}, \tilde{b}, s_0^2)$. In particular, under this empirical Bayes model, since costs are zero, the posterior decision rules are $Y_i > -\sigma_i^2 \frac{\tilde{a}+\tilde{b}\log\sigma_i}{s_0^2}$, and only $a = \tilde{a}/s_0^2$ and $b = \tilde{b}/s_0^2$ are relevant for parametrizing thresholds.

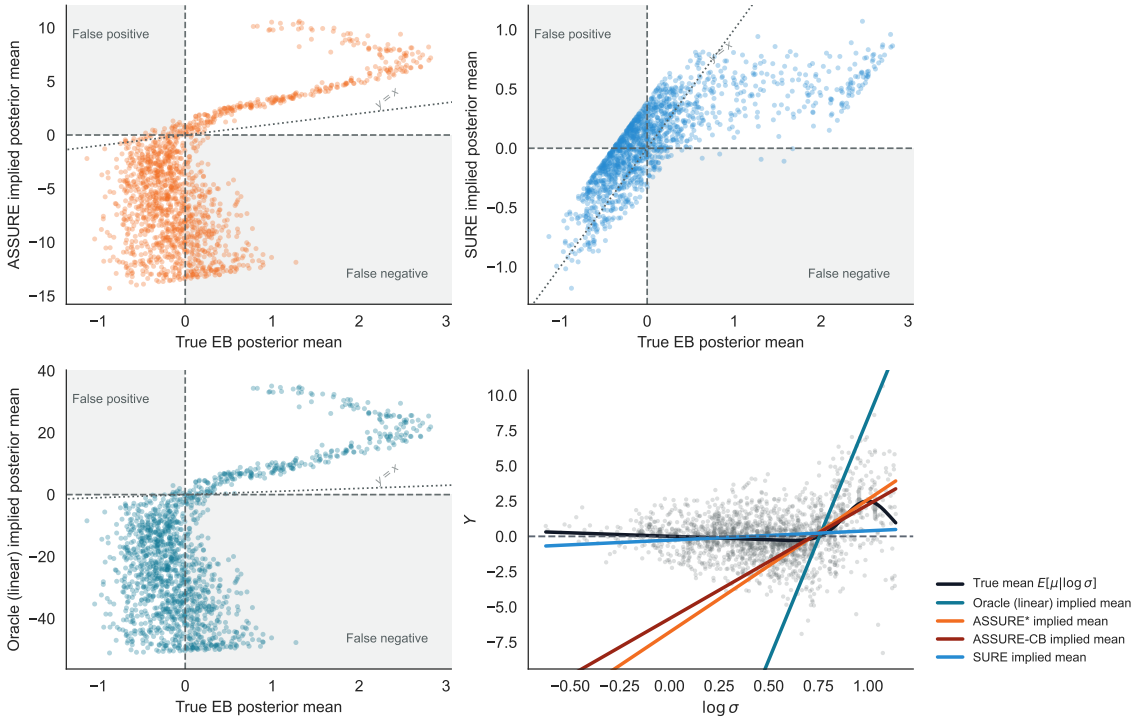


FIGURE 3. Results from the DGP (5.1) with a misspecified prior.

Notes. We show a single draw from (5.1) with $n = 1500$. The lower-right panel plots Y_i against $\log \sigma_i$ for one sample with $n = 1500$, together with the true conditional mean $m_0(\log \sigma_i)$ and the linear empirical-Bayes mean functions implied by SURE, ASSURE*, ASSURE-CB, and the oracle welfare (2.2) maximizer over $\mathcal{D}$. The other three panels compare each method's implied posterior mean with the true posterior mean under the data-generating process; the dashed line indicates the 45 degree line. ASSURE-CB produces very similar posterior means as ASSURE* and is thus omitted. □

For a draw of Y_i, we fit decision rules over $\mathcal{D}$ with ASSURE*, ASSURE-CB, and by maximizing the oracle welfare function (2.2). We likewise fit $\tilde{a}, \tilde{b}, s_0^2$ in (5.2) through SURE or plug-in empirical Bayes.[23] Table 2 shows the average utility across 1000 draws. We see from

[23]Plug-in estimation for (5.2) estimates $\tilde{a}, \tilde{b}$ via OLS of Y_i on a constant and $\log \sigma_i$. It then estimates s_0^2 by taking the difference in the residual variance of this regression and the average σ_i^2.

	True posterior mean	Oracle over $\mathcal{D}$	ASSURE*	ASSURE-CB	SURE	Plug-in empirical Bayes	$\mathbf{1}(Y_i > 0)$
Average utility	0.295	0.278	0.267	0.2668	0.227	0.228	0.201

TABLE 2. Average utility across 1000 draws from (5.1) with $n = 1500$.

Notes. (μ, σ, Y_i) are redrawn. Average utility is (1.1) averaged over 1000 draws. Plug-in empirical Bayes fits $\tilde{a}, \tilde{b}$ with ordinary least-squares of Y_i on $\log \sigma_i$. It then fits s_0^2 by averaging the squared OLS residuals and subtracting by the average of σ_i^2. □

Table 2 that ASSURE methods substantially improve on SURE or plug-in empirical Bayes. Normalizing the performance of $\mathbf{1}(Y_i > 0)$ to zero and oracle over $\mathcal{D}$ to one, ASSURE* has relative performance 0.857 while SURE and plug-in have relative performances only about 0.35.

To unpack why ASSURE* has good performance, it is helpful to compare results in terms of implied EB posterior means. We thus convert the threshold estimates (a, b) into implied $\tilde{a}, \tilde{b}$ by fixing the prior variance at its true value $s_0^2 = 0.2$. Under this normalization, the corresponding EB coefficients are $\tilde{a} = 0.2a$, $\tilde{b} = 0.2b$. Figure 3 visualizes different methods on this empirical Bayes scale for a single draw of the data.

The bottom-right of Figure 3 shows a plot of Y_i against $\log \sigma_i$ with the implied prior mean estimates of various methods. SURE effectively fits a weighted squared-error approximation to the true $m_0(\log \sigma_i)$. On the other hand, ASSURE and oracle directly optimize for threshold rules, thereby resulting in implied $\mathbb{E}[\mu \mid \log \sigma]$ estimates that have high fidelity to where the conditional mean function $m_0(\log \sigma)$ crosses zero.

We can visualize the gain of ASSURE by plotting the implied posterior mean estimates against the true posterior mean under (5.1) in the other three panels of Figure 3. Indeed, SURE results in a good squared error approximation of the true posterior mean, aligning well with the 45-degree line. However, because it does not tune parameters for the selection objective, it incurs many false positives—where the SURE estimates of μ_i are positive but true posterior mean is negative. On the other hand, tuning decision rules implied by the linear model through the oracle objective (2.2) trades off false positives and false negatives more efficiently, resulting in higher welfare than SURE. ASSURE mimics the behavior of the oracle—its implied posterior means are inaccurate predictions of μ_i, but lead to better selections.

Lastly, we can examine whether the ASSURE* objective function is made ill-behaved by the oscillations in the sinc kernel. Figure 4 shows the objective function for ASSURE*, ASSURE-CB, and the true welfare function (1.3) for one draw of data. Both ASSURE methods do result in nonconvex objective function surfaces, possibly motivating sampling-based

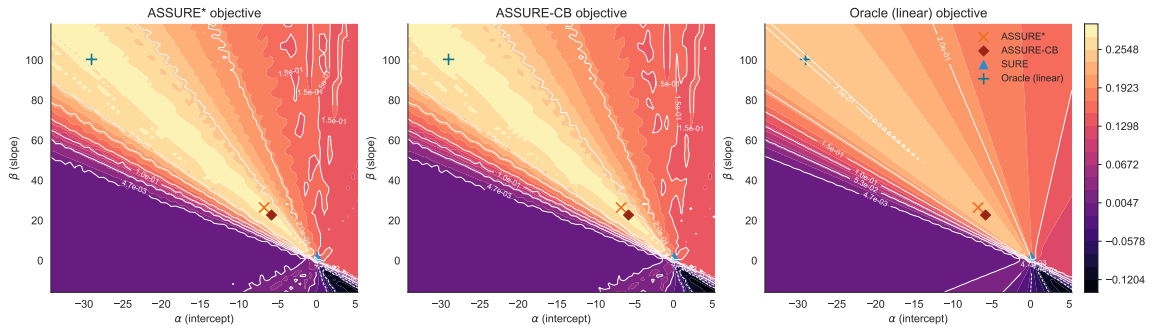


FIGURE 4. Objective function landscape of (5.1).

methods à la Chernozhukov and Hong (2003). However, it does not appear that the higher-order sinc kernel *per se* results in worse optimization landscape than the Gaussian kernel.

6. Empirical Applications

We illustrate ASSURE in three empirical applications in economics: selecting census tracts to maximize economic mobility, identifying discrimination among large firms, and selecting innovations in experimentation programs. The first two applications show that plug-in linear-shrinkage empirical Bayes methods can be suboptimal for welfare when the underlying model for P is misspecified, and that ASSURE improves robustness in those settings. The last application instead uses ASSURE to assess hypothesis-testing rules for A/B testing.

6.1. The Opportunity Atlas.

We begin with the setting studied by Bergman et al. (2024), which can be cast naturally as a compound selection problem. Their goal is to identify high-mobility Census tracts for recommending to low-income households, using noisy estimates from Chetty et al. (2026). Our implementation follows Chen (2026), who studies a national version of this application using nonparametric empirical Bayes methods.

To cast this decision problem in our framework, let Y_i denote the estimated economic mobility of a Census tract i, defined either as mean income rank or as probability of high income for Black individuals growing up in relatively poor households in tract i. Denote true economic mobility by μ_i, and the estimated standard error of Y_i in Chetty et al. (2026) as σ_i.

Bergman et al. (2024) are interested in selecting the top third of census tracts. We interpret this targeting rule as implicitly defining a constant implementation cost $k_i = k$. To calibrate k, we compute posterior means under Fay–Herriot and let $k_{\text{calibrated}}$ be their 66th percentile. Under this calibration, an empirical Bayesian is indifferent between selecting and not selecting the marginal tract in the top third. We fix this calibrated cost throughout.

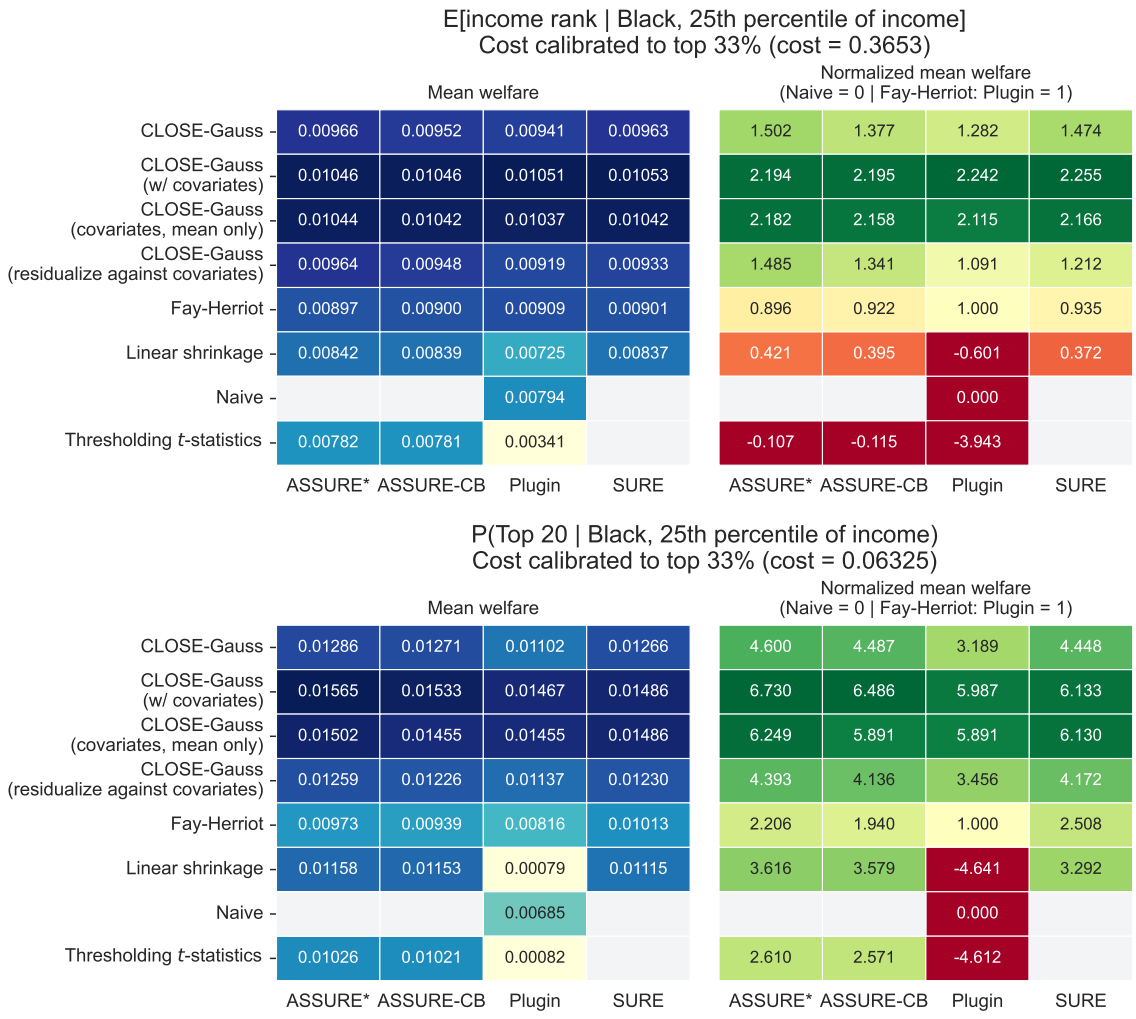


FIGURE 5. Absolute and relative welfare for two different economic mobility measures in Bergman et al. (2024) application.

Notes. Results are averaged over 1000 coupled-bootstrap draws. "Naive" is the decision rule that simply selects $\mathbf{1}(Y_i > k)$. CLOSE-GAUSS ignores covariates and models $\mu \mid \sigma_i$. In addition, we consider three variants of CLOSE-GAUSS with covariates. The first variant (CLOSE-GAUSS w/ covariates) includes covariates in both prior mean and prior variance symmetrically as with $\log \sigma_i$. The second variant (CLOSE-GAUSS w/ covariates, mean only) includes covariates in the prior mean, but models the prior variance as completely constant. The third variant (CLOSE-GAUSS, residualize against covariates) residualizes Y_i linearly against the covariates, and runs CLOSE-GAUSS on the residuals. In relative comparisons, we normalize the performance of naive to 0 and that of plug-in Fay-Herriot to 1. □

To evaluate the performance of different selection decisions, following Chen (2026), we employ coupled bootstrap (2.10) to mimic a $90/10$ split by synthetically constructing

a holdout sample that provides an unbiased estimate of performance on the compound decision problem with inflated variances. We consider a few instances of decision rules discussed in Table 1 and Example 1: (i) Linear shrinkage, (ii) Fay–Herriot, (iii) Parametric CLOSE-GAUSS, (iv) Parametric CLOSE-GAUSS with covariates and its variants, and (v) Thresholding t-statistics. See Appendix E for implementation details.

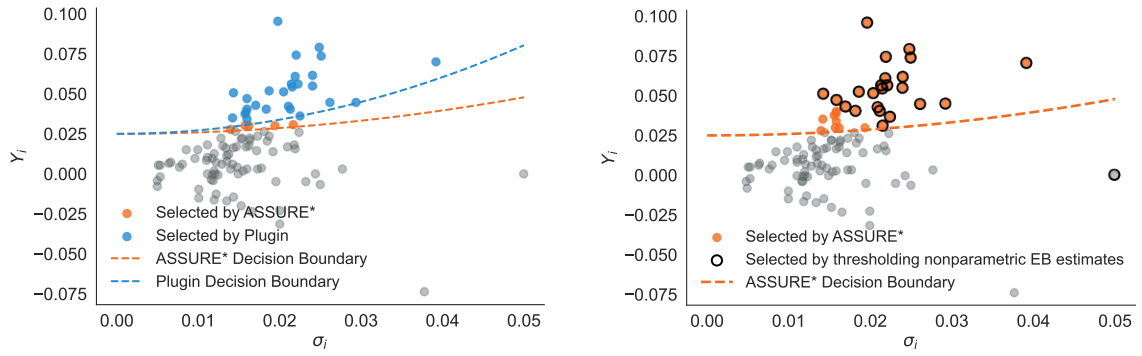
For each class of decision rules, we then consider fitting their parameters with ASSURE*, ASSURE-CB, SURE, and plug-in estimation. Plug-in estimation for CLOSE-GAUSS methods involves method-of-moment estimation of the prior parameters. Plug-in estimation for thresholding t-statistics simply chooses the conventional threshold $\beta = 1.96$.

Figure 5 summarizes results. The left panels report absolute welfare and the right panels report welfare normalized so that the performance of $\mathbf{1}(Y_i > k)$ achieves zero relative welfare and the plug-in Fay–Herriot decision achieves relative welfare equal to 1. For flexible empirical Bayes models that are arguably better specified, all methods perform similarly: ASSURE methods do not generate large improvements, though they do not appear to hurt. In those cases, ASSURE certifies that plug-in empirical Bayes methods, though they optimize on average across $\mu_{1:n} \sim P$, do well on a particular instance of the parameters. On the other hand, for less flexible decision classes, ASSURE* and ASSURE-CB noticeably improve over plug-in and SURE tuning—again illustrating the robustness gains from targeting welfare directly.

6.2. **Discrimination in large firms.** Kline et al. (2022) conducted a large correspondence experiment to measure the contact gap μ_i between white and Black job applicants across $n = 108$ large U.S. firms. Specifically, fictitious applications were sent to approximately 125 job listings per firm. The difference in callback rates between races was then averaged across jobs within a firm to produce the estimate Y_i.

Using a procedure that controls the false discovery rate at the 5% level, Kline et al. (2022) identify 23 firms that discriminate against Black applicants. Under the "intensive-margin" preference in their paper, using these estimates to select firms for further investigation maps directly into utility (1.1), where k is the cost of an investigation. If we calibrate k using the nonparametric empirical Bayes benchmark from Kline et al. (2022), adapted from Efron (2016), the implied value is $k = 0.025$, which rationalizes the selection of 23 firms.

To illustrate the performance of ASSURE*, consider the class of linear shrinkage decision rules (Table 1). With constant k, linear shrinkage implied thresholds simplify to $\delta(\sigma_i, \mu_0, s_0^2) = k + \sigma_i^2 \frac{k - m_0}{s_0^2} =: k + \sigma_i^2 \beta$, and thus we can let $\beta := (k - m_0)/s_0^2$ be a scalar parameter. Figure 6(a) shows that directly plugging in the MLE estimates for (m_0, s_0) would give a higher threshold and select fewer firms than ASSURE*.



(a) ASSURE* vs. plug-in linear shrinkage

(b) ASSURE* vs nonparametric EB

FIGURE 6. Selections from $n = 108$ firms studied by Kline et al. (2022)

While Kline et al. (2022) reported linear shrinkage estimates, their preferred model is a nonparametric empirical Bayes model as shown in Figure 6(b). This is precisely due to concerns that the parametric Gaussian prior corresponding to the linear shrinkage rule is misspecified. For instance, the underlying distribution of true contact gaps may be skewed. As a result, the posterior mean tends to underestimate large μ_i, making the shrinkage rule conservative. ASSURE* can partly correct for this and partially salvage linear shrinkage rules. In a cross-validation exercise over ten splits (three waves to estimate the threshold and two to estimate the resulting welfare), the plug-in linear shrinkage rule achieves welfare 39% below that of thresholding based on the nonparametric EB estimates, whereas ASSURE* is only 25% below.

6.3. **Experimentation programs.** In the technology industry, *experimentation programs* are portfolios of A/B tests that evaluate interventions affecting a business metric of interest (Azevedo et al., 2020; Sudijono et al., 2024; Chou et al., 2025). For experiment i, the average treatment effect estimate Y_i is plausibly normally distributed around the true average treatment effect μ_i, measured in terms of the business metric. We use an anonymized Netflix dataset from (Sudijono et al., 2024), containing treatment-effect estimates and standard deviations for 331 feature experiments. Figure 7 summarizes the data.

We consider selecting for interventions to implement in the platform as a compound selection problem. A common ad hoc decision procedure selects via the t-statistic decision class in Example 1, $\mathcal{D} = \{\delta(z;\beta) = k + \sigma\beta\}$. Conventional practice chooses the cutoff to be $t = \beta = 1.96$, equivalently a two-sided p-value of 0.05. We instead use ASSURE* to choose the threshold β that maximizes the expected welfare of the decision, taking $k_i = 0$. Figure 8 shows the ASSURE* estimate $\hat{W}(\beta)$. By maximizing the curve $\hat{W}(\beta)$, ASSURE* suggests using a decision corresponding to $\hat{\beta} \approx 0$.

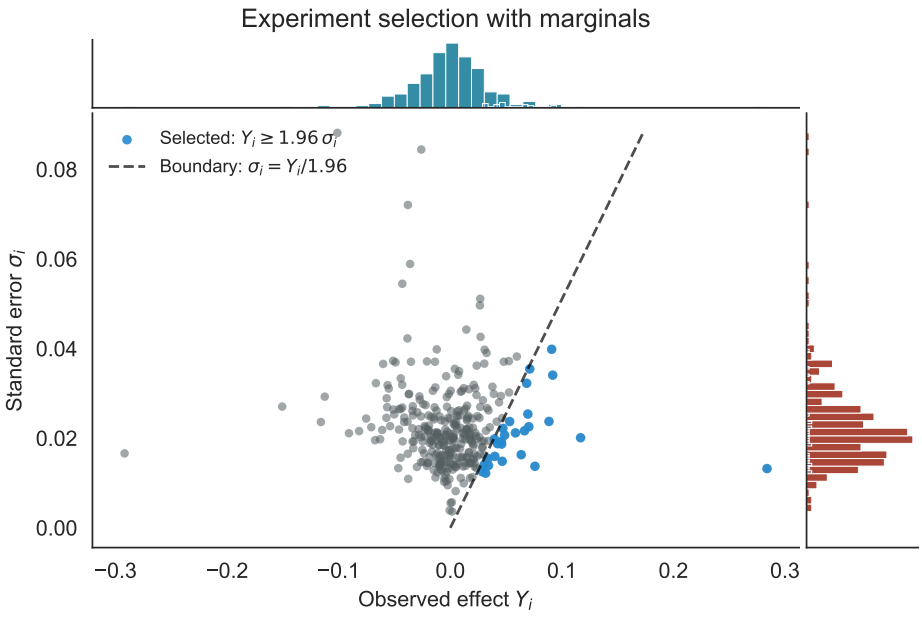


FIGURE 7. Experimentation-program data with the conventional significance threshold

Notes. The center panel plots observed treatment effects against their standard errors for roughly 330 feature experiments. Blue points satisfy the conventional rule $Y_i \geq 1.96\sigma_i$, and the dashed line traces this decision boundary. The top and right margins show the empirical distributions of observed effects and standard errors. □

Because the data are anonymized (both the treatment-effect estimates and their standard deviations are obfuscated by random multiplicative factors), the optimal decision cannot be interpreted directly. Even so, the shape of the estimated welfare curve is informative: If implementation is nearly costless, the preferred threshold is much more lenient than the conventional ($t = 1.96$) rule, echoing Azevedo et al. (2020); Berman and Van den Bulte (2022); Sudijono et al. (2024).

The ASSURE welfare estimate can also be used in reverse: It lets practitioners infer the cost level under which the conventional $t = 1.96$ rule would be approximately optimal. To illustrate this, Figure 9 scales the implementation cost to be κ times the median estimated treatment-effect size in the dataset. The left panel plots $\hat{W}(\beta)$ for different values of κ, and the right panel plots the corresponding ASSURE-optimal threshold. As κ rises, the preferred rule becomes more conservative. The figure suggests that the conventional $t = 1.96$ benchmark is consistent with implementation costs roughly twice the median treatment effect.

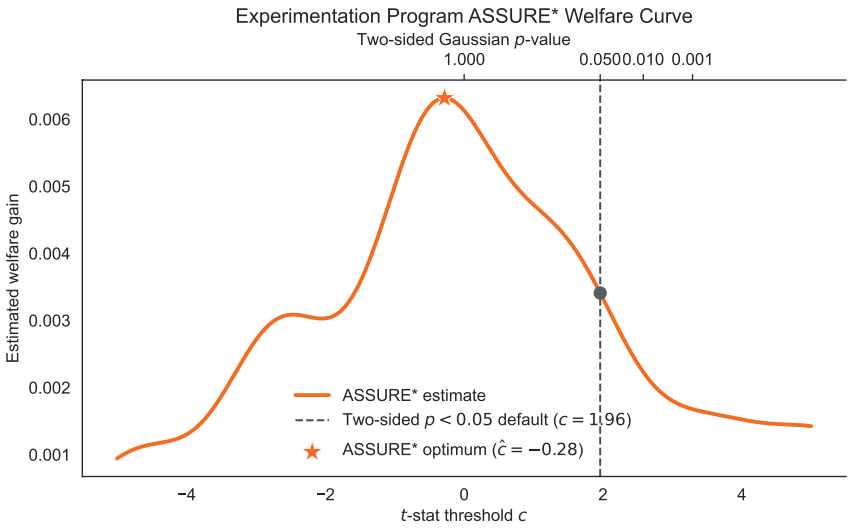


FIGURE 8. ASSURE* welfare curve for t-statistic thresholds.

Notes. The solid curve plots the estimated welfare $\hat{W}(c)$ for decision rules of the form $Y_i \geq c\sigma_i$. The star marks the ASSURE*-optimal threshold, while the dashed vertical line marks the conventional benchmark $c = 1.96$. The top axis reports the corresponding two-sided Gaussian p-values. □

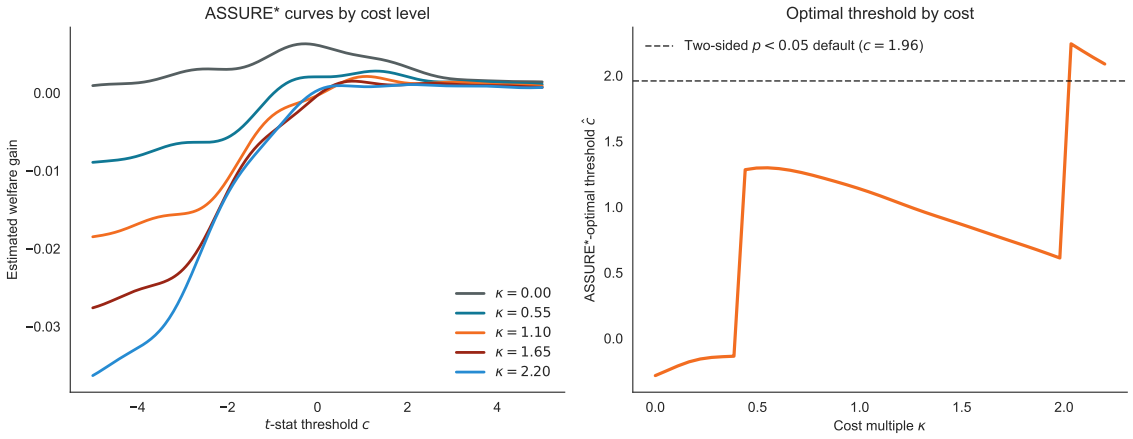


FIGURE 9. ASSURE* welfare and optimal thresholds by cost level.

Notes: The left panel shows estimated welfare curves for decision rules $Y_i \geq k + c\sigma_i$ under several constant implementation costs k. The right panel plots the resulting ASSURE*-optimal threshold c as costs vary, where costs are normalized as multiples of the median estimated treatment-effect magnitude in the dataset. The dashed horizontal line marks the conventional benchmark $c = 1.96$.

7. Conclusion

This paper introduces ASSURE, a data-driven approach for compound selection decisions, closely related to SURE. In particular, ASSURE* is near-unbiased for the expected utility of selection problems for fixed $\mu_{1:n}$, under weak assumptions, producing robust evaluations of the quality of selection decisions. Because it accurately estimates welfare, optimizing ASSURE* also leads to attractive selection decisions. We present both theoretical guarantees and practical applications illustrating these advantages.

References

Armstrong, Timothy B, Michal Kolesár, and Mikkel Plagborg-Møller, "Robust empirical Bayes confidence intervals," *Econometrica*, 2022, *90* (6), 2567–2602. 13

Aschenbrenner, Matthias, Alf Dolich, Deirdre Haskell, Dugald Macpherson, and Sergei Starchenko, "Vapnik–Chervonenkis density in some theories without the independence property, II," 2013. OA-36

Audibert, Jean-Yves and Alexandre B Tsybakov, "Fast learning rates for plug-in classifiers under the margin condition," *arXiv preprint math/0507180*, 2005. 14, 16

_____ **and** _____, "Fast learning rates for plug-in classifiers," 2007. 16

Azevedo, Eduardo M, Alex Deng, José Luis Montiel Olea, Justin Rao, and E Glen Weyl, "A/b testing with fat tails," *Journal of Political Economy*, 2020, *128* (12), 4614–4672. 2, 26, 27, 39

Bergman, Peter, Raj Chetty, Stefanie DeLuca, Nathaniel Hendren, Lawrence F Katz, and Christopher Palmer, "Creating moves to opportunity: Experimental evidence on barriers to neighborhood choice," *American Economic Review*, 2024, *114* (5), 1281–1337. 2, 3, 23, 24

Berman, Ron and Christophe Van den Bulte, "False discovery in A/B testing," *Management Science*, 2022, *68* (9), 6762–6782. 27

Bonhomme, Stéphane and Angela Denis, "Estimating heterogeneous effects: applications to labor economics," *Labour Economics*, 2024, *91*, 102638. 3, 6, 13

Chen, Jiafeng, "Empirical Bayes when estimation precision predicts parameters," *Econometrica*, 2026, *94* (2), 305–340. 3, 4, 5, 6, 7, 10, 13, 14, 15, 23, 24, OA-36

_____**, Lihua Lei, Timothy Sudijono, Liyang Sun, and Tian Xie**, "Compound Selection Decisions: An Almost SURE Approach," *arXiv preprint arXiv:2511.11862v2*, 2026. 9, 11

_____**, Nabarun Deb, and Nikolaos Ignatiadis**, "Normal approximations in nonparametric empirical Bayes," *arXiv preprint arXiv:2605.31599*, 2026. 18

Cheng, Xu, Sheng Chao Ho, and Frank Schorfheide, "Optimal estimation of two-way effects under limited mobility," *arXiv preprint arXiv:2506.21987*, 2025. 2, 4, 5, 13

Chernozhukov, Victor and Han Hong, "An MCMC approach to classical estimation," *Journal of econometrics*, 2003, *115* (2), 293–346. 23

Chetty, Raj, John N Friedman, and Jonah E Rockoff, "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates," *American economic review*, 2014, *104* (9), 2593–2632. 2

_____, _____, **Nathaniel Hendren, Maggie R Jones, and Sonya R Porter**, "The opportunity atlas: Mapping the childhood roots of social mobility," Technical Report 1 2026. 2, 13, 23, OA-36

Chou, Winston, Colin Gray, Nathan Kallus, Aurélien Bibaut, and Simon Ejdemyr, "Evaluating Decision Rules Across Many Weak Experiments," in "Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2" 2025, pp. 4365–4374. 26

Crippa, Federico, "Regret analysis in threshold policy design," *Journal of Econometrics*, 2025, *249*, 105998. 5, 16

Dano, Kevin, Bo E Honoré, and Martin Weidner, "Binary choice logit models with general fixed effects for panel and network data," *arXiv preprint arXiv:2508.11556*, 2025. 6

Davis, Kathryn Bullock, "Mean square error properties of density estimates," *The Annals of Statistics*, 1975, *3* (4), 1025–1030. 9

den Dries, Lou Van, *Tame topology and o-minimal structures*, Vol. 248, Cambridge university press, 1998. OA-36

Efron, Bradley, *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, Vol. 1, Cambridge University Press, 2012. 2, 13

_____, "Empirical Bayes deconvolution estimates," *Biometrika*, 2016, *103* (1), 1–20. 25

Fay III, Robert E. and Roger A. Herriot, "Estimates of income for small places: an application of James-Stein procedures to census data," *Journal of the American Statistical Association*, 1979, *74* (366a), 269–277. 7

Fithian, William and Lihua Lei, "Conditional calibration for false discovery rate control under dependence," *The Annals of Statistics*, 2022, *50* (6), 3091–3118. 7

Ghosh, Sulagna, Nikolaos Ignatiadis, Frederic Koehler, and Amber Lee, "Stein's unbiased risk estimate and Hyvarinen's score matching," *arXiv preprint arXiv:2502.20123*, 2025. 4, 13

Gilraine, Michael, Jiaying Gu, and Robert McMillan, "A new method for estimating

teacher value-added," Technical Report, National Bureau of Economic Research 2020. 6

Gu, Jiaying and Roger Koenker, "Invidious comparisons: Ranking and selection as compound decisions," *Econometrica*, 2023, *91* (1), 1–41. 5

Hirano, Keisuke, "A Comment on: "Invidious Comparisons: Ranking and Selection as Compound Decisions" by Jiaying Gu and Roger Koenker," *Econometrica*, 2023, *91* (1), 43–46. 5

_____ **and Jack R Porter**, "Asymptotics for statistical treatment rules," *Econometrica*, 2009, *77* (5), 1683–1701. 3, 5, 17, 18

_____ **and** _____, "Panel asymptotics and statistical decision theory," *The Japanese Economic Review*, 2016, *67* (1), 33–49. 5

Ignatiadis, Nikolaos and Dennis L Sun, "Empirical Bayes estimation via data fission," *Journal of the American Statistical Association*, 2025, *120* (549), 165–166. 4, 10

Ishihara, Takuya and Daisuke Kurisu, "Shrinkage Methods for Treatment Choice," *arXiv preprint arXiv:2210.17063*, 2022. 11

James, William and Charles Stein, "Estimation with quadratic loss," in "Proceedings of the fourth Berkeley symposium on mathematical statistics and probability," Vol. 1 University of California Press 1961, pp. 361–379. 3

Jiang, Wenhua, "On general maximum likelihood empirical Bayes estimation of heteroscedastic IID normal means," 2020. 5, 13

_____ **and Cun-Hui Zhang**, "General maximum likelihood empirical Bayes estimation of normal means," 2009. 14

Karunamuni, Rohana J, "Optimal rates of convergence of empirical Bayes tests for the continuous one-parameter exponential family," *The Annals of Statistics*, 1996, pp. 212–231. 2

Kitagawa, Toru and Aleksey Tetenov, "Who should be treated? empirical welfare maximization methods for treatment choice," *Econometrica*, 2018, *86* (2), 591–616. 5, 16

Kline, Patrick, Evan K Rose, and Christopher Walters, "Systemic discrimination among large US employers," *The Quarterly Journal of Economics*, 2022, *137* (4), 1963–2036. 2, 25, 26

Kline, Patrick M, Evan K Rose, and Christopher Walters, "Discrimination Report Cards: An Empirical Bayes Ranking Approach," 2024. 5

Kohavi, Ron, Diane Tang, and Ya Xu, *Trustworthy online controlled experiments: A practical guide to a/b testing*, Cambridge University Press, 2020. 7

Kolmogorov, Andrei Nikolaevich, "Unbiased estimates," *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya*, 1950, *14* (4), 303–326. 9, 10

Kwon, Soonwoo, "Optimal shrinkage estimation of fixed effects in linear panel data models," *arXiv preprint arXiv:2308.12485*, 2023. 2, 4, 5

Lei, Lihua, "Discussion of "Data Fission: Splitting a Single Data Point"–Some Asymptotic Results for Data Fission," *Journal of the American Statistical Association*, 2025, *120* (549), 147–150. 4

Leiner, James, Boyan Duan, Larry Wasserman, and Aaditya Ramdas, "Data fission: splitting a single data point," *Journal of the American Statistical Association*, 2023, pp. 1–12. 4, 10

Liang, Ta Chen, "On optimal convergence rate of empirical Bayes tests," *Statistics & probability letters*, 2004, *68* (2), 189–198. 2, 17

Liang, TaChen, "On the convergence rates of empirical Bayes rules for two-action problems: discrete case," *The Annals of Statistics*, 1988, pp. 1635–1642. 2

_____, "On an empirical Bayes test for a normal mean," *Annals of statistics*, 2000, pp. 648–655. 2, 9

Luo, Yixiang, William Fithian, and Lihua Lei, "Estimating the FDR of variable selection," *arXiv preprint arXiv:2408.07231*, 2024. 7

_____**,** _____**, and** _____, "Improving knockoffs with conditional calibration," *The Annals of Statistics*, 2025, *53* (5), 2283–2302. 7

Manski, Charles F, "Statistical treatment rules for heterogeneous populations," *Econometrica*, 2004, *72* (4), 1221–1246. 3, 19

_____, *Identification for prediction and decision*, Harvard University Press, 2009. 11

_____, "Treatment choice with trial data: Statistical decision theory should supplant hypothesis testing," *The American Statistician*, 2019, *73* (sup1), 296–304. 7

_____, "Econometrics for decision making: Building foundations sketched by Haavelmo and Wald," *Econometrica*, 2021, *89* (6), 2827–2853. 3, 14

Mehta, Nirav, "Measuring quality for use in incentive schemes: The case of "shrinkage" estimators," *Quantitative Economics*, 2019, *10* (4), 1537–1577. 3

Mogstad, Magne, Joseph P Romano, Azeem M Shaikh, and Daniel Wilhelm, "Inference for ranks with applications to mobility across neighbourhoods and academic achievement across countries," *Review of Economic Studies*, 2024, *91* (1), 476–518. 7

Müller, Ulrich K and Mark W Watson, "Spatial unit roots and spurious regression," *Econometrica*, 2024, *92* (5), 1661–1695. 3, 13

Neufeld, Anna, Ameer Dharamshi, Lucy L Gao, and Daniela Witten, "Data thinning for convolution-closed distributions," *Journal of Machine Learning Research*, 2024, *25* (57), 1–35. 4

Oliveira, Natalia L, Jing Lei, and Ryan J Tibshirani, "Unbiased risk estimation in the normal means problem via coupled bootstrap techniques," *Electronic Journal of Statistics*, 2024, *18* (2), 5405–5448. 4, 10

Onshuus, Alf and Adolfo J Quiroz, "Metric entropy estimation using o-minimality theory," *arXiv preprint arXiv:1511.07098*, 2015. OA-36

Pensky, Marianna, "Minimax theory of estimation of linear functionals of the deconvolution density with or without sparsity," 2017. 9, 10

Polyanskiy, Yury and Yihong Wu, "Sharp regret bounds for empirical Bayes and compound decision problems," *arXiv preprint arXiv:2109.03943*, 2021. 15

Ponomarev, Kirill and Vira Semenova, "On the lower confidence band for the optimal welfare," *arXiv preprint arXiv:2410.07443*, 2024. 16

Robbins, Herbert, "Asymptotically subminimax solutions of compound statistical decision problems," *Herbert Robbins Selected Papers*, 1951/1985, pp. 7–24. 2, 3, 11, 12

Schönberg, Isaac Jacob, "On variation-diminishing integral operators of the convolution type," *Proceedings of the National Academy of Sciences*, 1948, *34* (4), 164–169. 43

Soloff, Jake A, Adityanand Guntuboyina, and Bodhisattva Sen, "Multivariate, heteroscedastic empirical Bayes via nonparametric maximum likelihood," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2024, p. qkae040. 14

Stefanski, Leonard A, "Unbiased estimation of a nonlinear function a normal mean with application to measurement error models," *Communications in Statistics-Theory and Methods*, 1989, *18* (12), 4335–4358. 4, 9, 10, 42, 43

_____ **and Raymond J Carroll**, "Deconvolving kernel density estimators," *Statistics*, 1990, *21* (2), 169–184. 4, 9

Stein, Charles M, "Estimation of the mean of a multivariate normal distribution," *The annals of Statistics*, 1981, pp. 1135–1151. 4

Stoye, Jörg, "Minimax regret treatment choice with covariates or with limited validity of experiments," *Journal of Econometrics*, 2012, *166* (1), 138–156. 3, 12

Sudijono, Timothy, Simon Ejdemyr, Apoorva Lal, and Martin Tingley, "Optimizing returns from experimentation programs," *arXiv preprint arXiv:2412.05508*, 2024. 5, 26, 27

Tate, Robert F, "Unbiased estimation: functions of location and scale parameters," *The Annals of Mathematical Statistics*, 1959, pp. 341–366. 9, 10

Tsybakov, Alexandre B., *Introduction to Nonparametric Estimation* Springer Series in Statistics, Springer, 2009. 9

van der Vaart, AW and Jon A Wellner, "Empirical processes," in "Weak Convergence

and Empirical Processes: With Applications to Statistics," Springer, 2023, pp. 127–384. OA-10, OA-20, OA-31, OA-32, OA-33, OA-34, OA-35, OA-36

Vershynin, Roman, "High-dimensional probability," 2009. 15

Walters, Christopher, "Empirical Bayes methods in labor economics," in "Handbook of Labor Economics," Vol. 5, Elsevier, 2024, pp. 183–260. 2, 5, 7

Weinstein, Asaf, Zhuang Ma, Lawrence D Brown, and Cun-Hui Zhang, "Group-linear empirical Bayes estimates for a heteroscedastic normal mean," *Journal of the American Statistical Association*, 2018, *113* (522), 698–710. 5

Wilkie, Alex J, "Model completeness results for expansions of the ordered field of real numbers by restricted Pfaffian functions and the exponential function," *Journal of the American Mathematical Society*, 1996, *9* (4), 1051–1094. OA-36

Xie, Xianchao, SC Kou, and Lawrence D Brown, "SURE estimates for a heteroscedastic hierarchical model," *Journal of the American Statistical Association*, 2012, *107* (500), 1465–1479. 4

Zhou, Fan and Ping Li, "A Fourier Analytical Approach to Estimation of Smooth Functions in Gaussian Shift Model," *arXiv preprint arXiv:1911.02010*, 2019. 9, 10

Appendix A. Proofs

Throughout, define

$$\begin{aligned}\Psi_h(Y_i; z_i, C) &:= (Y_i - k_i)\,\mathrm{Csinc}\left(\frac{Y_i - C}{\sigma_i h}\right) - \frac{\sigma_i}{h}\,\mathrm{sinc}\left(\frac{Y_i - C}{\sigma_i h}\right) \\ &= (Y_i - k_i)\left\{\frac{1}{2} + \frac{1}{\pi}\,\mathrm{Si}\left(\frac{Y_i - C}{\sigma_i h}\right)\right\} - \frac{\sigma_i}{h}\,\mathrm{sinc}\left(\frac{Y_i - C}{\sigma_i h}\right). \qquad \text{(A.1)}\end{aligned}$$

Proof of Proposition 1. We first invoke Lemma C.2, where a calculation with the well-known Dirichlet integral

$$\int_0^\infty \frac{\sin ax}{x} dx = \lim_{h \downarrow 0} \int_0^{1/h} \frac{\sin ax}{x} dx$$

shows that

$$\begin{aligned}(\mu - k)\Phi\left(\frac{\mu - C}{\sigma}\right) &= \frac{1}{2}(\mu - k) + \frac{1}{\pi}(\mu - k)\int_0^\infty \frac{1}{\omega} e^{-\frac{1}{2}\omega^2} \sin(\omega(\mu - C)/\sigma) d\omega \\ \mathbb{E}\Psi_h(Y, z, C) &= \frac{1}{2}(\mu - k) + \frac{1}{\pi}(\mu - k)\int_0^{1/h} \frac{1}{\omega} e^{-\frac{1}{2}\omega^2} \sin(\omega(\mu - C)/\sigma) d\omega.\end{aligned}$$

Taking the difference,

$$(\mu - k)\Phi\left(\frac{\mu - C}{\sigma}\right) - \mathbb{E}\Psi_h(Y, z, C) = \frac{1}{\pi}\int_{1/h}^\infty \frac{1}{\omega} e^{-\frac{1}{2}\omega^2} (\mu - k) \sin(\omega(\mu - C)/\sigma) d\omega,$$

which has an upper bound given by

$$\frac{|\mu - k|}{\pi} \int_{1/h}^{\infty} \frac{1}{\omega} e^{-\frac{1}{2}\omega^2} d\omega \leq \frac{h|\mu - k|}{\pi} \int_{1/h}^{\infty} e^{-\frac{1}{2}\omega^2} d\omega \leq h^2 |\mu - k| e^{-1/(2h^2)}.$$

The last line follows from the Mills ratio bound, which holds uniformly over C, σ. As a result, the bias for $\hat{W}_h$ is bounded above by $\left(\frac{1}{n}\sum_{i=1}^n |\mu_i - k_i|\right) h^2 e^{-1/(2h^2)}$, by letting $C_i = \delta(z_i; \beta)$.

Next, we establish the second claim pointwise in β. Set $\lambda_n := 1/h = \sqrt{2 \log n}$ and $C_i := \delta(z_i; \beta)$. Using $\mathrm{Csinc}(x) = 1/2 + \pi^{-1} \mathrm{Si}(x)$, write the ith summand as

$$w_h(Y_i; z_i, \beta) = \frac{1}{2}(Y_i - k_i) + \frac{1}{\pi}(Y_i - k_i) \,\mathrm{Si}\left(\lambda_n \frac{Y_i - C_i}{\sigma_i}\right) - \sigma_i \lambda_n \,\mathrm{sinc}\left(\lambda_n \frac{Y_i - C_i}{\sigma_i}\right).$$

Since Si is bounded, the second moment of the first two terms is bounded by a constant multiple of $\mathbb{E}(Y_i - k_i)^2 = \sigma_i^2 + (\mu_i - k_i)^2$. For the last term, Lemma C.3 calculates

$$\mathbb{E}\left[\left(\sigma_i \lambda_n \,\mathrm{sinc}\left(\lambda_n \frac{Y_i - C_i}{\sigma_i}\right)\right)^2\right] \lesssim \lambda_n \sigma_i^2.$$

Therefore,

$$\mathrm{Var}(w_h(Y_i; z_i, \beta)) \lesssim (\mu_i - k_i)^2 + \lambda_n \sigma_i^2,$$

where we used $\lambda_n \geq 1$ to absorb the remaining σ_i^2 term. By independence,

$$\mathrm{Var}(\hat{W}_n(\beta)) \lesssim \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^n (\mu_i - k_i)^2 + \lambda_n \frac{1}{n}\sum_{i=1}^n \sigma_i^2\right).$$

The bias bound above, evaluated at $h = 1/\sqrt{2 \log n}$, gives

$$\left|\mathbb{E}\hat{W}_n(\beta) - W(\beta)\right| \lesssim \frac{1}{n \log n}\left(\frac{1}{n}\sum_{i=1}^n |\mu_i - k_i|\right).$$

By Jensen's inequality, the square of this bias bound is dominated by $n^{-1}(n^{-1}\sum_{i=1}^n (\mu_i - k_i)^2)$. Combining these displays with the bias-variance decomposition proves the claimed pointwise mean-squared error bound. □

Proof of Theorem 1. Let $\beta^* \in \mathrm{argmax}_{\beta \in \mathcal{B}} W(\beta)$. The optimality of $\hat{\beta}$ for $\hat{W}_n$ implies $\hat{W}_n(\hat{\beta}) \geq \hat{W}_n(\beta^*)$. Recalling $u(\beta)$ from (2.1), we have $\mathbb{E}[u(\beta^*)] = W(\beta^*)$ and

$$\begin{aligned}
\mathsf{Regret}_n &= W(\beta^*) - \mathbb{E}[u(\hat{\beta})] = \mathbb{E}[u(\beta^*) - u(\hat{\beta})] \\
&\leq \mathbb{E}\left[u(\beta^*) - \hat{W}_n(\beta^*) + \hat{W}_n(\hat{\beta}) - u(\hat{\beta})\right] \leq 2\mathbb{E}\left[\sup_{\beta \in \mathcal{B}} |u(\beta) - \hat{W}_n(\beta)|\right] \\
&\leq 2\mathbb{E}\left[\sup_{\beta \in \mathcal{B}} |\hat{W}(\beta) - W(\beta)|\right] + 2\mathbb{E}\left[\sup_{\beta \in \mathcal{B}} |u(\beta) - W(\beta)|\right]. \qquad \text{(A.2)}
\end{aligned}$$

Starting from (A.2), we use the bias bound in Proposition 1(1) to obtain:

$$(\text{A.2}) \lesssim \mathbb{E}\left[\sup_{\beta\in\mathcal{B}}|u(\beta)-W(\beta)|\right] + \mathbb{E}\left[\sup_{\beta\in\mathcal{B}}|\hat{W}(\beta)-\mathbb{E}\hat{W}(\beta)|\right] + \underbrace{\frac{m_1}{n\log n}}_{\text{by Proposition 1}}.$$

Letting $A_1(\beta) := u(\beta) - W(\beta)$, $A_2(\beta) := \hat{W}(\beta) - \mathbb{E}\hat{W}(\beta)$, Propositions C.1 and C.2 prove if $\sigma_{1:n}, \sigma_{1:n}^{-1}$ and the envelope D are all uniformly bounded by an absolute constant, that

$$\mathbb{E}\sup_{\beta\in\mathcal{B}}|A_1(\beta)| \lesssim m_2\frac{\sqrt{V(\mathcal{D})}}{\sqrt{n}}$$
$$\mathbb{E}\sup_{\beta\in\mathcal{B}}|A_2(\beta)| \lesssim CV(\mathcal{D})(m_2+m_4^4)\frac{(\log n)^{1/4}\sqrt{\log\log n}}{\sqrt{n}},$$

for some constant C that depends only on $\sigma_{1:n}$ and D. Therefore

$$\mathsf{Regret}_n \le (\text{A.2}) \lesssim CV(\mathcal{D})(m_2+m_4^4)\frac{(\log n)^{1/4}\sqrt{\log\log n}}{\sqrt{n}}. \qquad \square$$

Proof of Theorem 2. Fix some $h > 0$ and let $\mu_+ = (h/\sqrt{n}, \ldots, h/\sqrt{n})'$ and $\mu_- = -\mu_+$. For any selection decision $a_1, \ldots, a_n$, let $g_+ = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mu_+}[a_i(Y_1,\ldots,Y_n)]$ and define g_- analogously. Then the regret at μ_+ is

$$\sup_{|\beta|<M} W(\beta;\mu_+) - \frac{1}{n}\sum_{i=1}^n \mu_i \mathbb{E}_{Y_{1:n}\sim\mathcal{N}(\mu_+,I_n)}[a_i(Y_1,\ldots,Y_n)] = \frac{h}{\sqrt{n}}\left(\Phi\left(\frac{h}{\sqrt{n}}+M\right) - g_+\right)$$

Analogously, the regret at μ_- is

$$-\frac{h}{\sqrt{n}}\left(\Phi\left(-\frac{h}{\sqrt{n}}-M\right) - g_-\right) = \frac{h}{\sqrt{n}}\left(\Phi\left(\frac{h}{\sqrt{n}}+M\right) - 1 + g_-\right)$$

Thus, the maximum regret over μ_+ and μ_- is

$$\frac{h}{\sqrt{n}}\left(\Phi\left(\frac{h}{\sqrt{n}}+M\right) - 1 + \max(1-g_+, g_-)\right) \tag{A.3}$$

Thus, for any choice of selection decisions $a_{1:n}$,

$$\sup_{\mu_{1:n}\in[-1,1]^n}\left\{\sup_{|\beta|<M} W(\beta;\mu_{1:n}) - \frac{1}{n}\sum_{i=1}^n \mu_i \mathbb{E}_{Y_{1:n}\sim\mathcal{N}(\mu_{1:n},I_n)}[a_i(Y_1,\ldots,Y_n)]\right\} \ge (\text{A.3}).$$

Let $\varphi(Y_1,\ldots,Y_n) = \frac{1}{n}\sum_i a_i(Y_1,\ldots,Y_n) \in [0,1]$. When viewed as a test against $H_0 : \mu = \mu_-$ for $H_1 : \mu = \mu_+$, g_- is its type I error and $1-g_+$ is its type II error. Then

$$\max(1-g_+, g_-) \ge \frac{(1-g_+)+(g_-)}{2} \ge \frac{1}{2}\left(1 - \mathrm{TV}\left(\mathcal{N}(\mu_+, I_n), \mathcal{N}(\mu_-, I_n)\right)\right).$$

By Pinsker's inequality, $\mathrm{TV}\left(\mathcal{N}(\mu_+, I_n), \mathcal{N}(\mu_-, I_n)\right) \le h$. Thus,

$$(\text{A.3}) \ge \frac{h}{\sqrt{n}}\left(\Phi\left(\frac{h}{\sqrt{n}} + M\right) - 1 + \frac{1}{2} - \frac{h}{2}\right) \ge \frac{h}{\sqrt{n}}\left(c_M - \frac{h}{2}\right) \gtrsim_M \frac{1}{\sqrt{n}}$$

for $c_M = \Phi(M) - 1/2 > 0$ and the choice $h = c_M$. This completes the proof. □

Proof of Theorem 3. Decompose the regret as

$$\mathsf{Regret}_n(\hat{\beta}) = \mathbb{E}\left[W(\beta^*) - W(\hat{\beta})\right] + \mathbb{E}\left[W(\hat{\beta}) - u(\hat{\beta})\right].$$

Theorem C.2 shows that the first term is $O((\log n)^6 n^{-1})$. To handle the second term $\mathbb{E}\left[W(\hat{\beta}) - u(\hat{\beta})\right]$, we use a leave-one-out stability argument. Define Y_{-i} to be the vector of observations without Y_i and

$$\hat{\beta}^{(-i)} = \operatorname*{argmax}_{\beta} \sum_{j \ne i} g_n(Y_j, z_j, \beta), \tag{A.4}$$

where

$$g_n(y, z, \beta) = \frac{(y-k)}{2} + \frac{(y-k)}{\pi}\,\mathrm{Si}\left(\frac{\lambda_n(y - \delta(\beta, z))}{\sigma}\right) - \sigma\lambda_n\,\mathrm{sinc}\left(\frac{\lambda_n(y - \delta(\beta, z))}{\sigma}\right)$$

are the summands in the ASSURE* estimator. We will show that for each i, $\hat{\beta}^{(-i)}$ only differs from $\hat{\beta}$ by $\tilde{O}_P(1/n)$. To explain why this will suffice, first write

$$\begin{aligned} W(\hat{\beta}) - \mathbb{E}u(\hat{\beta}) = W(\hat{\beta}) &- \frac{1}{n}\sum_{i=1}^{n}(\mu_i - k_i)\mathbb{P}(Y_i \ge \delta(\hat{\beta}^{(-i)}, z_i)) \\ &+ \frac{1}{n}\sum_{i=1}^{n}(\mu_i - k_i)\left(\mathbb{P}(Y_i \ge \delta(\hat{\beta}^{(-i)}, z_i)) - \mathbb{P}(Y_i \ge \delta(\hat{\beta}, z_i))\right). \end{aligned}$$

Thus

$$\mathbb{E}\left[W(\hat{\beta}) - u(\hat{\beta})\right] = \frac{1}{n}\sum_{i=1}^{n}(\mu_i - k_i)\mathbb{E}\left[\Phi\left(\frac{\mu_i - \delta(\hat{\beta}, z_i)}{\sigma_i}\right) - \Phi\left(\frac{\mu_i - \delta(\hat{\beta}^{(-i)}, z_i)}{\sigma_i}\right)\right] \tag{A.5}$$

$$+ \frac{1}{n}\sum_{i=1}^{n}(\mu_i - k_i)\left(\mathbb{P}(Y_i \ge \delta(\hat{\beta}^{(-i)}, z_i)) - \mathbb{P}(Y_i \ge \delta(\hat{\beta}, z_i))\right). \tag{A.6}$$

Using Lipschitzness of Φ and $\delta(\cdot, z_i)$, we will upper bound (A.5) by

$$\frac{1}{n}\sum_{i=1}^{n}\frac{L_i|\mu_i - k_i|}{\sigma_i}\mathbb{E}|\hat{\beta} - \hat{\beta}^{(-i)}|$$

for some constants L_i. To handle (A.6), we truncate. Fix a truncation threshold $E_n := C(\log n)^\gamma/n$ for C, γ constants to be chosen later.

$$\begin{aligned}
&\mathbb{P}(Y_i \geq \delta(\hat{\beta}^{(-i)}, z_i)) - \mathbb{P}(Y_i \geq \delta(\hat{\beta}, z_i)) \\
&= \mathbb{E}\left[\mathbb{P}(Y_i \geq \delta(\hat{\beta}^{(-i)}, z_i) \mid Y_{-i}) - \mathbb{P}(Y_i \geq \delta(\hat{\beta}, z_i) \mid Y_{-i})\right] \\
&\leq \mathbb{E}\left[\mathbb{P}\left\{|Y_i - \delta(\hat{\beta}^{(-i)}, z_i)| \leq |\delta(\hat{\beta}^{(-i)}, z_i) - \delta(\hat{\beta}, z_i)| \mid Y_{-i}\right\}\right] \\
&\leq \mathbb{E}\left[\mathbb{P}(|Y_i - \delta(\hat{\beta}^{(-i)}, z_i)| \leq L_i|\hat{\beta}^{(-i)} - \hat{\beta}| \mid Y_{-i})\right] \\
&\leq \mathbb{E}\left[\mathbb{P}(|Y_i - \delta(\hat{\beta}^{(-i)}, z_i)| \leq L_i E_n \mid Y_{-i})\right] + \mathbb{P}(|\hat{\beta}^{(-i)} - \hat{\beta}| > E_n) \\
&\leq \mathbb{E}\left[\Phi\left(\frac{\delta(\hat{\beta}^{(-i)}, z_i) + L_i E_n - \mu_i}{\sigma_i}\right) - \Phi\left(\frac{\delta(\hat{\beta}^{(-i)}, z_i) - L_i E_n - \mu_i}{\sigma_i}\right)\right] \\
&\quad + \mathbb{P}(|\hat{\beta}^{(-i)} - \hat{\beta}| > E_n).
\end{aligned}$$

Using Lipschitzness again, we find that (A.6) is bounded above by

$$|(\text{A.6})| \leq \frac{1}{n}\sum_{i=1}^n \frac{L_i|\mu_i - k_i|}{\sigma_i} E_n + \frac{1}{n}\sum_{i=1}^n |\mu_i - k_i| \mathbb{P}(|\hat{\beta}^{(-i)} - \hat{\beta}| > E_n). \tag{A.7}$$

By the compactness assumption on the search space for β, we may bound

$$\mathbb{E}[|\hat{\beta}^{(-i)} - \hat{\beta}|] \leq M\mathbb{P}(|\hat{\beta}^{(-i)} - \hat{\beta}| > E_n) + E_n$$

for some constant M depending on the diameter of the search space. Therefore we may bound $\mathbb{E}\left[W(\hat{\beta}) - u(\hat{\beta})\right]$ by

$$\frac{1}{n}\sum_{i=1}^n \left(1 + \frac{L_i}{\sigma_i}\right)|\mu_i - k_i|\left(E_n + \mathbb{P}(|\hat{\beta}^{(-i)} - \hat{\beta}| > E_n)\right). \tag{A.8}$$

Theorem C.4 shows that

$$\mathbb{P}\left(|\hat{\beta}^{(-i)} - \hat{\beta}| > \frac{4MB\log n}{\kappa n}\right) \leq O\left(\frac{1}{n}\right),$$

so we may take $E_n = O(\frac{\log n}{n})$. Thus

$$\begin{aligned}
\mathbb{E}\left[W(\hat{\beta}) - u(\hat{\beta})\right] &\leq \frac{B^3}{n}\sum_{i=1}^n \left(E_n + \mathbb{P}(|\hat{\beta}^{(-i)} - \hat{\beta}| > E_n)\right) \\
&\leq B^3\left(O\left(\frac{\log n}{n}\right) + O\left(\frac{1}{n}\right)\right) = O\left(\frac{\log n}{n}\right).
\end{aligned}$$

Combining this with the conclusion of Theorem C.2 gives the result. □

Proof of Theorem 4. Throughout the proof let $a_{1:n}$ denote a vector of binary selections, where the ith index is $a_i(Y_1, \ldots, Y_n) \in \{0, 1\}$. We will first exhibit a class of priors that are supported on Θ. Proposition B.2 verifies that a set of assumptions Assumption B.1(1)–Assumption B.1(3) are sufficient for Assumption 1.

Consider the priors $G_p^{\otimes n}$, where $G_p := \mathsf{Rad}_p$ is the one-dimensional Rademacher prior with $\mu_i = +1$ with probability p and $\mu_i = -1$ with probability $1 - p$. Let $G_p^{\otimes n}|A$ be the conditional distribution of $\mu_{1:n}$ conditioned on $A := \{\min_i \mu_i < \max_i \mu_i\} \cap \{|\overline{\mu}| < C'\}$ for some absolute constant C' small enough. Then $G_p^{\otimes n}|A$ satisfies Assumption B.1(1) with $B = 1$ and, choosing κ small enough, it satisfies Assumption B.1(3) since $\frac{1}{n}\sum_{i=1}^n \mu_i^2 = 1$. Lastly, Assumption B.1(2) is implied if $|\overline{\mu}| < \frac{m_1}{\varphi(1)}\overline{\sigma}^2$. Since μ_i is either $+1$ or -1, this is in turn implied by $|\overline{\mu}| < C'$ for C' small enough.

Next, observe the following inequalities:

$$\sup_{\mu_{1:n}\in\Theta} \mathsf{Regret}_n(a_{1:n}) \geq \mathbb{E}_{G_p^{\otimes n}|A}\left[\mathsf{Regret}_n(a_{1:n})\right] \tag{A.9}$$

$$= \frac{\mathbb{E}_{G_p^{\otimes n}}\left[\mathsf{Regret}_n(a_{1:n})\mathbf{1}(A)\right]}{\mathbb{P}_{G_p^{\otimes n}}(A)} \tag{A.10}$$

$$= \frac{\mathbb{E}_{G_p^{\otimes n}}\left[\mathsf{Regret}_n(a_{1:n})\right] - \mathbb{E}_{G_p^{\otimes n}}\left[\mathsf{Regret}_n(a_{1:n})\mathbf{1}(A^c)\right]}{\mathbb{P}_{G_p^{\otimes n}}(A)} \tag{A.11}$$

$$\geq \frac{\mathbb{E}_{G_p^{\otimes n}}\left[\mathsf{Regret}_n(a_{1:n})\right] - \mathbb{P}_{G_p^{\otimes n}}(A^c)}{1 - \mathbb{P}_{G_p^{\otimes n}}(A^c)}. \tag{A.12}$$

Observe that $\mathbb{P}_{G_p^{\otimes n}}(A^c)$ is exponentially small in n for all p in a small neighborhood of $\frac{1}{2}$.

We will thus focus on analyzing the term $\mathbb{E}_{G_p^{\otimes n}}\left[\mathsf{Regret}_n(a_{1:n})\right]$. We have

$$\begin{aligned}\mathbb{E}_{G_p^{\otimes n}}\left[\mathsf{Regret}_n(a_{1:n})\right] &= \mathbb{E}_{G_p^{\otimes n}}\left\{\sup_\beta W(\beta) - \frac{1}{n}\sum_{i=1}^n \mu_i \mathbb{E}\left[a_i(Y_1, \ldots, Y_n)\right]\right\} \\ &\geq \sup_\beta \mathbb{E}_{G_p^{\otimes n}} W(\beta) - \mathbb{E}_{G_p^{\otimes n}}\left[\frac{1}{n}\sum_{i=1}^n \mu_i \mathbb{E}\left[a_i(Y_1, \ldots, Y_n)\right]\right].\end{aligned}$$

Since $W(\beta) = \frac{1}{n}\sum_{i=1}^n \mu_i \mathbb{P}(Y_i \geq \beta)$ and μ_i are independent, observe that

$$\sup_\beta \mathbb{E}_{G_p^{\otimes n}} W(\beta) = \sup_\beta \mathbb{E}_{G_p}\left[\mu \mathbb{P}(Y \geq \beta \mid \mu)\right] = \mathbb{E}_{G_p}\mathbb{E}_Y[\mu \mathbf{1}\left\{\mathbb{E}_{G_p}[\mu \mid Y] > 0\right\}],$$

which follows from monotonicity of the posterior expectation in Y. This is the welfare of the Bayes-optimal decision (see e.g. Azevedo et al. (2020)). Therefore, we have lower bounded $\sup_{\mu_{1:n}\in\Theta} \mathsf{Regret}_n(a_{1:n})$ by the Bayes regret of any product prior $G_p^{\otimes n}$ up to some additive errors and multiplicative factors that are exponentially close to 1 in n.

We will finish the proof using Le Cam's two point argument. By Lemma B.3 we may restrict to decision rules which are monotone: $a_i(Y) = \mathbf{1}\left\{Y_i > \delta_i(Y_{(-i)})\right\}$. For a prior G on $\{-1,+1\}$, let $m_G(y) = \mathbb{E}_G[\mu \mid y]$ be the posterior mean, let $a^*_G(y) = \mathbf{1}\{m_G(y) > 0\}$ be the Bayes optimal decision, and let $R(G, a_{1:n})$ denote the Bayes regret under the product prior $G^{\otimes n}$. We can equivalently write this Bayes regret as

$$\begin{aligned} R(G, a_{1:n}) &:= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{G^{\otimes n}}\left[\mu_i \left(a^*_G(Y_i) - a_i(Y_{1:n})\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{G^{\otimes n}}\left[m_G(Y_i)\left(a^*_G(Y_i) - a_i(Y_{1:n})\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{G^{\otimes n}}\left[|m_G(Y_i)|\,\mathbf{1}\left\{a^*_G(Y_i) \neq a_i(Y_{1:n})\right\}\right]. \end{aligned} \tag{A.13}$$

where (A.13) is verified by Lemma C.8.

Take

$$G_0 := \mathsf{Rad}_{1/2-1/(4\sqrt{n})}, \qquad G_1 := \mathsf{Rad}_{1/2+1/(4\sqrt{n})}.$$

By Lemma C.10, for every monotone rule $a_{1:n}$ there is an absolute constant $c > 0$ such that

$$R(G_0, a_{1:n}) + R(G_1, a_{1:n}) \geq \frac{c}{n}$$

for all sufficiently large n. Hence

$$\max\left(R(G_0, a_{1:n}), R(G_1, a_{1:n})\right) \geq \frac{c}{2n}.$$

Applying (A.12) separately under $G_0^{\otimes n}$ and $G_1^{\otimes n}$, and using $\mathbb{P}_{G_0^{\otimes n}}(A^c), \mathbb{P}_{G_1^{\otimes n}}(A^c) = O(e^{-n})$, we obtain

$$\begin{aligned} \sup_{\mu_{1:n}\in\Theta} \mathsf{Regret}_n(a_{1:n}) &\geq \max\left(\frac{R(G_0, a_{1:n}) - \mathbb{P}_{G_0^{\otimes n}}(A^c)}{1 - \mathbb{P}_{G_0^{\otimes n}}(A^c)}, \frac{R(G_1, a_{1:n}) - \mathbb{P}_{G_1^{\otimes n}}(A^c)}{1 - \mathbb{P}_{G_1^{\otimes n}}(A^c)}\right) \\ &\geq \frac{c}{2n} - O(e^{-n}) = \Omega(n^{-1}), \end{aligned}$$

as desired. □

Proof of Theorem 5. For each i, let $Z_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Recall the definition of w_h in (2.7) and define $w_{h,i}(y;\delta) = w_h(y;\sigma_i,\delta)$. By the triangle inequality,

$$\begin{aligned} &\left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\hat{w}_{h,i}(\sqrt{m}\bar{Y}_i;\delta_i)\right] - W(\delta)\right| \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \left|\mathbb{E}\left[\hat{w}_{h,i}(\sqrt{m}\bar{Y}_i;\delta_i) - w_{h,i}(\sqrt{m}\bar{Y}_i;\delta_i)\right]\right| \end{aligned}$$

$$
\begin{aligned}
&+ \frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}\left[w_{h,i}(\sqrt{m}\bar{Y}_i;\delta_i) - w_{h,i}(Z_i;\delta_i)\right]\right| \\
&+ \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[w_{h,i}(Z_i;\delta_i)\right] - W(\delta)\right| \\
&=: T_1 + T_2 + T_3.
\end{aligned}
$$

The terms have distinct roles: T_1 controls the effect of estimating σ_i^2, T_2 controls the error from replacing the non-Gaussian sample averages by Gaussian analogues, and T_3 is the residual bias from targeting the Gaussian welfare rather than $W(\delta)$.

By Lemma C.12, $T_1 \lesssim \sqrt{\log n}/\sqrt{m}$. By Lemma C.13, $T_2 \lesssim (\log n)^2/\sqrt{m}$. The idea for T_2 is to use the Lindeberg swapping argument, defined as follows. For each i, relabel the micro-data as $Y_1, \ldots, Y_m$, introduce Gaussian variables $Z_1, \ldots, Z_m$ with the same mean and variance, and define hybrid sums S_k in which the first k terms come from Z_t and the remaining $m-k$ terms are the original Y_t. Then

$$
\mathbb{E}\left[w_{h,i}(Z_i;\delta_i)\right] - \mathbb{E}\left[w_{h,i}(\sqrt{m}\bar{Y}_i;\delta_i)\right] = \sum_{k=0}^{m-1}\mathbb{E}\left[w_{h,i}(S_{k+1};\delta_i) - w_{h,i}(S_k;\delta_i)\right].
$$

Each increment is expanded around the common background sum S_k' obtained by removing the swapped variable. Because Y_{k+1} and Z_{k+1} have the same first two moments, the constant, linear, and quadratic Taylor terms cancel after taking expectations, so only a third-order remainder remains. Lemma C.13 shows that $|w_{h,i}^{(3)}(x;\delta_i)| \lesssim \sigma_i^{-2}h^{-4} + |x|\sigma_i^{-3}h^{-3}$, so one swap costs at most $P_i m^{-3/2}(h^{-4}+h^{-3})$ for a polynomial P_i in the moments. Summing over the m swaps gives $P_i(h^{-4}+h^{-3})/\sqrt{m}$, which is $P_i(\log n)^2/\sqrt{m}$ since $h = 1/\sqrt{2\log n}$.

Finally, Lemma C.14 gives $T_3 \lesssim (n\log n)^{-1} + m^{-1/2}$. Combining the three bounds proves the claim. □

Appendix B. Miscellany

We first state some straightforward computations.

Lemma B.1 (Properties of the sinc kernel)**.** *The map $x \mapsto \frac{1}{h}\operatorname{sinc}\left(\frac{x}{h}\right)$ is bounded above by Ch^{-1}, and further has a derivative bounded uniformly by Ch^{-2} for an absolute constant C.*

Lemma B.2 (Derivatives of Ψ)**.** *Let Ψ_C, Ψ_{CC} denote the first and second partial derivatives in C. They are given by the following formulas.*

$$
\Psi_C(Y_i, z_i, C) = -\frac{(Y_i - k_i)}{h\sigma_i}\operatorname{sinc}\left(\frac{Y_i - C}{h\sigma_i}\right) + \frac{1}{h^2}\operatorname{sinc}'\left(\frac{Y_i - C}{h\sigma_i}\right) \tag{B.1}
$$

$$
\Psi_{CC}(Y_i, z_i, C) = \frac{(Y_i - k_i)}{h^2\sigma_i^2}\operatorname{sinc}'\left(\frac{Y_i - C}{h\sigma_i}\right) - \frac{1}{\sigma_i h^3}\operatorname{sinc}''\left(\frac{Y_i - C}{h\sigma_i}\right) \tag{B.2}
$$

Lemma B.3 (Reduction to monotone decision rules). *Fix $\sigma_{1:n}, k_{1:n}$. Let $a_i(Y_i; Y_{-i}) \in [0,1]$ be a decision rule and assume $a_i(\cdot, Y_{-i})$ is not almost surely zero or almost surely one. There exists a threshold rule $a_i^*(\cdot; Y_{-i}) = \mathbf{1}(Y_i \geq \delta_i(Y_{-i}))$, where $\delta_i(Y_{-i}) \in \mathbb{R}$, such that*

$$\mathbb{E}_{\mu_{1:n}}[(\mu_i - k_i)a_i(Y_i; Y_{-i})] \leq \mathbb{E}_{\mu_{1:n}}[(\mu_i - k_i)a_i^*(Y_i; Y_{-i})]$$

for all $\mu_{1:n} \in \mathbb{R}^n$.

Proof. Let $\gamma_0(Y_{-i}) = \mathbb{E}_{\mu_i = k_i}[a_i(Y_i; Y_{-i}) \mid Y_{-i}] \in [0,1]$. Define $a_i^*(y; Y_{-i}) = \mathbf{1}(y \geq c(Y_{-i}))$ such that

$$\mathbb{E}_{\mu_i = k_i}\left[\mathbf{1}(Y_i \geq c(Y_{-i})) \mid Y_{-i}\right] = \gamma_0(Y_{-i}).$$

Note that $a_i^*(y; Y_{-i})$ is the uniformly most powerful test for $H_0 : \mu_i \leq k_i$ with *conditional size* $\gamma_0(Y_{-i})$, against $H_1 : \mu_i > k_i$. Likewise, $1 - a_i^*(y; Y_{-i})$ is the UMP test when we swap the null and the alternative. $a_i(y; Y_{-i})$ is a randomized test that has the same conditional size $\gamma_0(Y_{-i})$.

Let $\mu_i > k_i$, then $\mathbb{E}_{\mu_i}[a_i(Y_i; Y_{-i}) \mid Y_{-i}] \leq \mathbb{E}_{\mu_i}[a_i^*(Y_i; Y_{-i}) \mid Y_{-i}]$ since a_i^* is weakly more powerful than a_i. Conversely, if $\mu_i < k_i$, then $\mathbb{E}_{\mu_i}[a_i(Y_i; Y_{-i}) \mid Y_{-i}] \geq \mathbb{E}_{\mu_i}[a_i^*(Y_i; Y_{-i}) \mid Y_{-i}]$ since $1 - a^*$ is more powerful than $1 - a_i$. Thus

$$(\mu_i - k_i)\mathbb{E}_{\mu_i}[a_i(Y_i; Y_{-i}) \mid Y_{-i}] \leq (\mu_i - k_i)\mathbb{E}_{\mu_i}[a_i^*(Y_i; Y_{-i}) \mid Y_{-i}].$$

The conclusion follows by integrating out Y_{-i}. □

B.1. **Lack of an unbiased estimator.** We present an argument that is referenced in Stefanski (1989) without proof. By Lemma B.4, if an unbiased estimator for $W(\beta)$ were to exist, then one could unbiasedly estimate $\mu\Phi\left(\frac{\mu - c}{\sigma}\right)$. This means that one could estimate $\varphi(\mu/\sigma)/\sigma$ unbiasedly. Proposition B.1 then shows that $\varphi(\mu)$ cannot be unbiasedly estimated, at least not with estimators with reasonable tail behavior.

Lemma B.4. *Let $Y \sim \mathcal{N}(\mu, \sigma^2)$. Then*

$$\mathbb{E}[Y\mathbf{1}\{Y \geq c\}] = \frac{\sigma^2}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mu - c)^2}{2\sigma^2}\right) + \mu\Phi\left(\frac{\mu - c}{\sigma}\right).$$

Proof of Lemma B.4. Rewrite the expression as

$$\begin{aligned}
\mathbb{E}[Y\mathbf{1}\{Y \geq c\}] &= \mathbb{E}[(Y - \mu)\mathbf{1}\{Y - \mu \geq c - \mu\}] + \mu\mathbb{P}(Y \geq c) \\
&= \int_{c-\mu}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx + \mu\left(1 - \Phi\left(\frac{c - \mu}{\sigma}\right)\right) \\
&= -\frac{\sigma}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right)\Big|_{c-\mu}^{\infty} + \mu\left(1 - \Phi\left(\frac{c - \mu}{\sigma}\right)\right)
\end{aligned}$$

$$= \frac{\sigma}{\sqrt{2\pi}} \exp\left(-\frac{(c-\mu)^2}{2\sigma^2}\right) + \mu\left(1 - \Phi\left(\frac{c-\mu}{\sigma}\right)\right). \qquad \square$$

Proposition B.1 (Stefanski (1989)). *Let $Y \sim \mathcal{N}(0,1)$. Let $f : \mathbb{R} \to \mathbb{R}$ be some real-valued function where $\mathbb{E}_{Y\sim\mathcal{N}(\mu,1)}[|f(Y)|] < \infty$ for every $\mu \in \mathbb{R}$. Define $m : \mathbb{C} \to \mathbb{C}$ as $m(z) = \int f(y)\varphi(y-z)\,dy$. If m is entire, then there is some $\mu \in \mathbb{R}$ where $\mathbb{E}_\mu[f(Y)] \neq \varphi(\mu)$.*

Proof. Towards contradiction, suppose $m(\mu) = \mathbb{E}_\mu[f(Y)] = \varphi(\mu)$ over $\mu \in \mathbb{R}$. Since $\varphi(\cdot)$ is analytic, we must have $m(z) = \varphi(z)$ over $\mathbb{C}$. Let $z = it$, we have that for all $t \in \mathbb{R}$,

$$\int_{\mathbb{R}} f(y)\varphi(y)e^{ity}dye^{t^2/2} = \frac{1}{\sqrt{2\pi}}e^{t^2/2} \implies \int_{\mathbb{R}} f(y)\varphi(y)e^{ity}dy = \frac{1}{\sqrt{2\pi}}$$

Note that $\int |f(y)\varphi(y)|dy = \int |f(y)|\varphi(y)\,dy = \mathbb{E}_{Y\sim\mathcal{N}(0,1)}[|f(Y)|] < \infty$. Thus $f\varphi$ is L^1. By the Riemann–Lebesgue lemma, the Fourier transform vanishes at infinity:

$$0 = \lim_{|t|\to\infty} \int_{\mathbb{R}} f(y)\varphi(y)e^{ity}dy = \frac{1}{\sqrt{2\pi}}.$$

This is a contradiction. $\square$

It remains to discuss when $m(z)$ is entire. A sufficient condition is that $f(Y)$ is integrable under a slightly noisier Gaussian: $\mathbb{E}_{Y\sim\mathcal{N}(0,1+\epsilon^2)}[|f(Y)|] < \infty$. When this is true, we can then differentiate the real and imaginary parts of $m(z)$ under the integral sign and verify that the partials are continuous and satisfy the Cauchy–Riemann equations.

B.2. Example decision rule class for Theorem 3.

Example 3. Consider the class of simple threshold decision rules $\delta(z;\beta) = \beta$ with $\beta \in \mathbb{R}$. It is unclear whether there are interesting examples which satisfy Assumptions 1(1) to 1(4). The following assumption gives such a class of $\mu_{1:n}$. Fix $\kappa_0 > 0, B > 0$ and let $\overline{\mu} := \frac{1}{n}\sum_{i=1}^n \mu_i, \overline{\sigma}^2 := \frac{1}{n}\sum_{i=1}^n (\mu_i - \overline{\mu})^2$.

Assumption B.1. *(1)* $-B < \min_i \mu_i < 0$ *and* $0 < \max_i \mu_i < B$.
(2) Letting $m_B := \inf_{[B,3B]} t\varphi(t)$, $|\overline{\mu}| \le \overline{\sigma}^2 \frac{m_B}{\varphi(B)}$.
(3) $\frac{1}{n}\sum_{i=1}^n \mu_i^2 > \kappa_0/\varphi(3B)$.

Proposition B.2. *Fix constants B, κ_0 as above. Assume $\sigma_i^2 \equiv 1$ and Assumptions B.1(1) to B.1(3). Then Assumptions 1(1) to 1(4) hold with $\kappa = \kappa_0$ and with some $\xi > 0$ depending only on B, κ_0.*

Proof. By Assumption B.1(1), it is easy to see that Assumption 1(1) holds. Now without loss, assume that $\mu_1 \le \mu_2 \le \cdots \le \mu_n$. In this setting, $W(\beta) = \frac{1}{n}\sum_{i=1}^n \mu_i\Phi(\mu_i - \beta), W'(\beta) = -\frac{1}{n}\sum_{i=1}^n \mu_i\varphi(\mu_i - \beta)$, and $W''(\beta) = \frac{1}{n}\sum_{i=1}^n \mu_i(\beta - \mu_i)\varphi(\mu_i - \beta)$. A result of Schönberg (1948) shows that φ is a totally positive kernel and, for any sequence of $c_1 \le c_2 \le \ldots \le c_n$,

the number of sign changes in the vector $(W'(c_1), \ldots, W'(c_n))$ is at most that in the vector $(\mu_1, \ldots, \mu_n)$. Note that since $(\mu_1, ..., \mu_n)$ changes sign only once, $(W'(c_1), \ldots, W'(c_n))$ changes sign at most once. This implies that there cannot be multiple local maximizers.

Next, we claim that Assumption B.1(2) implies that $W'(-2B) > 0$ and $W'(2B) < 0$. If so, this would show Assumption 1(2) holds and $\beta^* \in [-2B, 2B]$. First, let $f_+(x) = \varphi(x - 2B)$. Let $\overline{f}_+ := \frac{1}{n}\sum_{i=1}^n f_+(\mu_i)$. Observe that

$$-W'(2B) = \frac{1}{n}\sum_{i=1}^n \mu_i f_+(\mu_i) = \overline{\mu}\overline{f}_+ + \frac{1}{n}\sum_{i=1}^n (\mu_i - \overline{\mu})\left(f_+(\mu_i) - f_+(\overline{\mu})\right) \tag{B.3}$$

By the mean-value theorem, we have for some ξ_i between $\mu_i, \overline{\mu}$ that $f'_+(\xi_i)(\mu_i - \overline{\mu}) = f_+(\mu_i) - f_+(\overline{\mu})$. Plugging this into the equation above,

$$\frac{1}{n}\sum_{i=1}^n \mu_i f_+(\mu_i) = \overline{\mu}\overline{f}_+ + \frac{1}{n}\sum_{i=1}^n f'_+(\xi_i)(\mu_i - \overline{\mu})^2.$$

We have $f'_+(x) = (2B - x)\varphi(x - 2B)$ and $f'_+(\xi_i) \geq m_B = \inf_{t\in[B,3B]} t\varphi(t) > 0$. Moreover, $\overline{f}_+ \leq \varphi(B)$. Thus, $-W'(2B) \geq -|\overline{\mu}|\varphi(B) + \bar{\sigma}^2 m_B > 0$. We can repeat the argument in the other direction with $f_- := \varphi(x + 2B)$. Doing so shows that $-W'(-2B) \leq |\overline{\mu}|\varphi(B) - \bar{\sigma}^2 m_B < 0$. This completes the claim.

Next, notice that $W''(\beta^*) = \frac{1}{n}\sum_i \mu_i(\beta^* - \mu_i)\varphi(\mu_i - \beta^*) = -\frac{1}{n}\sum_i \mu_i^2\varphi(\mu_i - \beta^*)$. where the second equality follows from the first order condition. Then boundedness, $\beta^* \in [-2B, 2B]$, and Assumption B.1(3) imply $W''(\beta^*) < -\frac{\kappa_0}{\varphi(3B)}\varphi(3B) = -\kappa_0$.

Now, for β such that $|\beta - \beta^*| \leq R$, we have

$$\begin{aligned} W''(\beta) &= \beta\left(\frac{1}{n}\sum_{i=1}^n \mu_i\varphi(\mu_i - \beta)\right) - \frac{1}{n}\sum_{i=1}^n \mu_i^2\varphi(\mu_i - \beta) \\ &\leq (2B + R)\, R \sup_{|\beta' - \beta^*| \leq R} |W''(\beta')| - \kappa_0 \frac{\varphi(3B + R)}{\varphi(3B)}. \end{aligned}$$

Notice that for $\mu_i \in [-B, B]$,

$$\left|\frac{1}{n}\sum_{i=1}^n \mu_i(\beta - \mu_i)\varphi(\mu_i - \beta)\right| \leq B\left|\sup_x x\varphi(x)\right| \leq B/4.$$

Thus

$$W''(\beta) \leq (2B + R)\frac{BR}{4} - \kappa_0 \frac{\varphi(3B + R)}{\varphi(3B)}$$

Now by the intermediate value theorem, for some R small enough, less than κ_0/B^2, $W''(\beta) < -\kappa_0/2$ for all $|\beta - \beta^*| \leq R$. This choice of R only depends on B and κ_0. By the uniqueness of local maximizer and a Taylor expansion argument, we conclude that

there exists $\xi > 0$ such that $\sup_{|\beta-\beta^*|>\kappa_0/B} W(\beta) < \sup_{|\beta-\beta^*|>R} W(\beta) < W(\beta^*) - \xi$. This proves Assumption 1(4). □

Online Appendix
for "Compound Selection Decisions: An Almost SURE Approach"

Jiafeng Chen, Lihua Lei, Timothy Sudijono, Liyang Sun, and Tian Xie

July 22, 2026

Contents

Appendix C. Auxiliary results for proofs OA-1
C.1. Auxiliary results for Proposition 1 OA-1
C.2. Auxiliary results for Theorem 1 OA-3
C.3. Tail bound for regret OA-10
C.4. Auxiliary results for Theorem 3 OA-11
C.5. Auxiliary results for Theorem 4 OA-26
C.6. Auxiliary results for Theorem 5 OA-28
Appendix D. Empirical processes for i.n.i.d. data OA-31
D.1. Examples of VC classes OA-35
Appendix E. Additional details for Section 6.1 OA-36

Appendix C. Auxiliary results for proofs

This appendix collects auxiliary lemmas and intermediate arguments organized by the main-text proof they support.

C.1. Auxiliary results for Proposition 1.

Lemma C.1. *Let $Y \sim \mathcal{N}(\mu, \sigma^2)$ and ω, C be constants. Then:*

$$\begin{aligned}
\mathbb{E}\sin\omega(Y-C) &= e^{-\frac{1}{2}\sigma^2\omega^2}\sin(\omega(\mu-C)) \\
\mathbb{E}\cos\omega(Y-C) &= e^{-\frac{1}{2}\sigma^2\omega^2}\cos(\omega(\mu-C)) \\
\mathbb{E}\left[Y\sin(\omega(Y-C))\right] &= e^{-\frac{1}{2}\sigma^2\omega^2}\left[\mu\sin(\omega(\mu-C)) + \omega\sigma^2\cos(\omega(\mu-C))\right] \\
&= e^{-\frac{1}{2}\sigma^2\omega^2}\left[\mu\sin(\omega(\mu-C))\right] + \sigma^2\omega\mathbb{E}\left[\cos(\omega(Y-C))\right].
\end{aligned}$$

Proof. For the first claim, observe that

$$\begin{aligned}
\mathbb{E}\left[\sin\omega(Y-C)\right] &= \frac{1}{2i}\mathbb{E}\left[e^{i\omega(Y-C)} - e^{-i\omega(Y-C)}\right] = \frac{1}{2i}\left(e^{i\omega\mu-\frac{1}{2}\sigma^2\omega^2-i\omega C} - e^{-i\omega\mu-\frac{1}{2}\sigma^2\omega^2+i\omega C}\right) \\
&= \frac{1}{2i}e^{-\frac{1}{2}\sigma^2\omega^2}\left(e^{i\omega\mu-i\omega C} - e^{-i\omega\mu+i\omega C}\right) \\
&= e^{-\frac{1}{2}\sigma^2\omega^2}\sin(\omega(\mu-C)).
\end{aligned}$$

An analogous argument with cosine shows the second claim. For the third claim, use Stein's identity $\mathbb{E}[Yf(Y)] = \mu\mathbb{E}f(Y) + \sigma^2\mathbb{E}[f'(Y)]$, and then apply the first two equalities of the lemma. □

Lemma C.2. *In the proof of Proposition 1, we have, for $z = (k, \sigma)$,*

$$(\mu - k)\Phi\left(\frac{\mu - C}{\sigma}\right) = \frac{1}{2}(\mu - k) + \frac{1}{\pi}(\mu - k)\int_0^\infty \frac{1}{\omega}e^{-\frac{1}{2}\omega^2}\sin(\omega(\mu - C)/\sigma)d\omega$$
$$\mathbb{E}\Psi_h(Y, z, C) = \frac{1}{2}(\mu - k) + \frac{1}{\pi}(\mu - k)\int_0^{1/h} \frac{1}{\omega}e^{-\frac{1}{2}\omega^2}\sin(\omega(\mu - C)/\sigma)d\omega$$

interpreted as Lebesgue integrals.

Proof. We consider the fact that for all $a \in \mathbb{R}$,[24]

$$\int_0^\infty \frac{\sin ax}{x}dx = \frac{\pi}{2}\text{sign}(a). \tag{C.1}$$

Then

$$\begin{aligned}
(\mu - k)\Phi\left(\frac{\mu - C}{\sigma}\right) &= (\mu - k)\int_{-\infty}^\infty \mathbf{1}\{y - C \geq 0\}\frac{1}{\sigma}\varphi\left(\frac{y - \mu}{\sigma}\right)dy \\
&= (\mu - k)\int_{-\infty}^\infty \left(\frac{1}{2} + \frac{1}{2}\text{sign}\left(\frac{y - C}{\sigma}\right)\right)\frac{1}{\sigma}\varphi\left(\frac{y - \mu}{\sigma}\right)dy \\
&= \frac{1}{2}(\mu - k) + \frac{1}{\pi}(\mu - k)\int_{-\infty}^\infty \lim_{h\downarrow 0}\int_0^{1/h}\frac{\sin(\omega(y - C)/\sigma)}{\omega}d\omega\frac{1}{\sigma}\varphi\left(\frac{y - \mu}{\sigma}\right)dy \\
&= \frac{1}{2}(\mu - k) + \frac{1}{\pi}(\mu - k)\lim_{h\downarrow 0}\int_{-\infty}^\infty\int_0^{1/h}\frac{\sin(\omega(y - C)/\sigma)}{\omega}d\omega\frac{1}{\sigma}\varphi\left(\frac{y - \mu}{\sigma}\right)dy.
\end{aligned}$$

The last step follows from dominated convergence theorem, noting that

$$\sup_h\left|\int_0^{1/h}\frac{\sin(\omega(y - C)/\sigma)}{\omega}d\omega\right| = \sup_h|\text{Si}((y - C)/h\sigma)| = K < \infty$$

if $x > C$, for an absolute constant K. A similar calculation shows when $x - C < 0$ the integral is bounded above by the same uniform constant K. Thus an integrable dominating function is given by $K\varphi((y - \mu)/\sigma)/\sigma$. Now applying Fubini's theorem, we can switch the order of integration and apply Lemma C.1:

$$\frac{1}{\pi}(\mu - k)\lim_{h\downarrow 0}\int_0^{1/h}\frac{1}{\omega}\left(\int_{-\infty}^\infty \sin(\omega(y - C)/\sigma)\frac{1}{\sigma}\varphi\left(\frac{y - \mu}{\sigma}\right)dy\right)d\omega$$

[24] The integral is interpreted as an improper Riemann integral, so

$$\int_0^\infty \frac{\sin ax}{x}dx = \lim_{h\downarrow 0}\int_0^{1/h}\frac{\sin ax}{x}dx.$$

The truncated integrals $\int_0^{1/h}\frac{\sin ax}{x}dx$ can be treated as Lebesgue integrals.

$$= \frac{1}{\pi} \lim_{h \downarrow 0} \int_0^{1/h} \frac{1}{\omega} e^{-\frac{1}{2}\omega^2} (\mu - k) \sin(\omega(\mu - C)/\sigma) d\omega$$
$$= \frac{1}{\pi} \int_0^{\infty} \frac{1}{\omega} e^{-\frac{1}{2}\omega^2} (\mu - k) \sin(\omega(\mu - C)/\sigma) d\omega.$$

This proves the first claim. For the second claim, note that

$$\frac{1}{\pi} \int_0^{1/h} \frac{1}{\omega} e^{-\frac{1}{2}\omega^2} (\mu - k) \sin(\omega(\mu - C)/\sigma) d\omega$$
$$= \frac{1}{\pi} \int_0^{1/h} \frac{1}{\omega} \int_{-\infty}^{\infty} \left[(y - k) \sin\left(\frac{\omega(y - C)}{\sigma}\right) - \omega\sigma \cos\left(\frac{\omega(y - C)}{\sigma}\right) \right] \frac{1}{\sigma} \varphi\left(\frac{y - \mu}{\sigma}\right) dy d\omega$$
$$= \frac{1}{\pi} \int_0^{1/h} \int_{-\infty}^{\infty} \left[\frac{(y - k) \sin\left(\frac{\omega(y-C)}{\sigma}\right)}{\omega} - \sigma \cos\left(\frac{\omega(y - C)}{\sigma}\right) \right] \frac{1}{\sigma} \varphi\left(\frac{y - \mu}{\sigma}\right) dy d\omega.$$

Again we may apply Fubini noting that the integral over ω, being over a compact domain, is finite, yielding

$$\frac{1}{\pi} \int_{-\infty}^{\infty} \int_0^{1/h} \left[(y - k) \frac{\sin(\omega(y - C)/\sigma)}{\omega} - \sigma \cos(\omega(y - C)/\sigma) \right] d\omega \frac{1}{\sigma} \varphi\left(\frac{y - \mu}{\sigma}\right) dy$$
$$= \frac{1}{\pi} \int_{-\infty}^{\infty} \left[\int_0^{1/h} (y - k) \frac{\sin(\omega(y - C)/\sigma)}{\omega} d\omega - \sigma \frac{\sin((y - C)/h\sigma)}{(y - C)/\sigma} \right] \frac{1}{\sigma} \varphi\left(\frac{y - \mu}{\sigma}\right) dy$$
$$= \mathbb{E}\left[\frac{1}{\pi} (Y - k) \operatorname{Si}\left(\frac{Y - C}{h\sigma}\right) - \frac{\sigma}{h} \operatorname{sinc}\left(\frac{Y - C}{h\sigma}\right) \right]$$
$$= \mathbb{E}\left[\Psi_h(Y; z, C)\right] - \frac{1}{2}(\mu - k).$$

This proves the second claim. □

C.2. **Auxiliary results for Theorem 1.** To state the proofs, define the norms $\|D\|_{p,n}^p := \frac{1}{n} \sum_{i=1}^n |D(z_i)|^p$ and $\|D\|_n := \|D\|_{2,n}$. Let $s(z_i) = \sigma_i$ and for $q \geq 1$, $s_q := \|s\|_{q,n}$. Let $m(z_i) = \mu_i - k_i$ and $m_q := \|m\|_{q,n}$.

Proposition C.1. *In the proof of Theorem 1, we have*

$$\mathbb{E} \sup_{\beta \in \mathcal{B}} |A_1(\beta)| \lesssim m_2 \frac{\sqrt{V(\mathcal{D})}}{\sqrt{n}}.$$

Proof. We will handle this with i.n.i.d. empirical process theory applied to the function class $\mathcal{F} = \{f_\beta : \beta \in \mathcal{B}\}$ with $f_\beta(y, \mu, z) = (\mu - k)\mathbf{1}\{y - \delta(z; \beta) \geq 0\}$. The class $\mathcal{F}$ has a uniform envelope function given by $F = |\mu - k|$. Moreover, Lemma D.1 shows that $\mathcal{F}$ is a VC subgraph class with index $O(V(\mathcal{D}))$: in particular, $y - \delta(\beta, z)$ is a VC subgraph function class ranging over $\beta \in \mathcal{B}$, and so is $\mathbf{1}\{y - \delta(\beta, z) \geq 0\}$. Multiplying by the fixed function $(\mu - k)$ and applying Lemma D.1(2) gives the claim. Next, we apply Theorem D.1

with $V_i = (Y_i, \mu_i, z_i)$ and Lemma D.2,

$$\begin{aligned}
\mathbb{E}\sup_{\beta\in\mathcal{B}}|A_1(\beta)| &= \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\beta\in\mathcal{B}}\frac{n}{\sqrt{n}}|A_1(\beta)| = \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\beta\in\mathcal{B}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{f_\beta(Y_i,\mu_i,z_i) - \mathbb{E}[f_\beta(Y_i,\mu_i,z_i)]\} \\
&\le \frac{1}{\sqrt{n}}J(1,\mathcal{F}\mid F,\mathbb{L}^2)\left(\frac{1}{n}\sum_{i=1}^{n}|\mu_i - k_i|^2\right)^{1/2} && \text{(Theorem D.1)} \\
&\lesssim \frac{1}{\sqrt{n}}\left(\frac{1}{n}\sum_{i=1}^{n}|\mu_i - k_i|^2\right)^{1/2}\int_0^1\sqrt{1+V(\mathcal{D})\ln\frac{1}{\epsilon}}d\epsilon && \text{(Lemma D.2)} \\
&\lesssim \sqrt{V(\mathcal{D})}\left(\frac{1}{n}\sum_{i=1}^{n}(\mu_i - k_i)^2\right)^{1/2}\frac{1}{\sqrt{n}}. && \square
\end{aligned}$$

Lemma C.3. *Let $Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ and let $\lambda > 0$. Uniformly over $C \in \mathbb{R}$,*

$$\mathbb{E}\left[\left(\sigma_i\lambda\,\mathrm{sinc}\left(\lambda\frac{Y_i - C}{\sigma_i}\right)\right)^2\right] \lesssim \lambda\sigma_i^2.$$

Consequently, for $h_\beta(y,z) := \sigma\lambda_n\,\mathrm{sinc}(\lambda_n(y-\delta(z;\beta))/\sigma)$, $\sup_{\beta\in\mathcal{B}}\frac{1}{n}\sum_{i=1}^{n}P_ih_\beta^2 \lesssim \lambda_n s_2^2$.

Proof. By the change of variables $u = \lambda(Y_i - C)/\sigma_i$,

$$\mathbb{E}\left[\left(\sigma_i\lambda\,\mathrm{sinc}\left(\lambda\frac{Y_i - C}{\sigma_i}\right)\right)^2\right] = \sigma_i^2\lambda\int_{\mathbb{R}}\mathrm{sinc}^2(u)\varphi\left(\frac{C-\mu_i}{\sigma_i} + \frac{u}{\lambda}\right)du \lesssim \lambda\sigma_i^2,$$

since $\sup_x\varphi(x) < \infty$ and $\int_{\mathbb{R}}\mathrm{sinc}^2(u)du < \infty$. Averaging over i gives the second claim. $\square$

Lemma C.4. *Let $\rho_\lambda(t) := \frac{1}{\pi}\mathrm{Si}(\lambda t) - \frac{1}{2}\mathrm{sgn}(t)$ for $\lambda > 1$. Let $Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Then, uniformly over $C \in \mathbb{R}$, $\mathbb{E}\left[(Y_i - k_i)^2\rho_\lambda\left(\frac{Y_i - C}{\sigma_i}\right)^2\right] \lesssim \lambda^{-1}\left\{\sigma_i^2 + (\mu_i - k_i)^2\right\}$. Consequently, for $r_\beta(y,z) := (y-k)\left[\frac{1}{\pi}\mathrm{Si}\left(\lambda_n\frac{y-\delta(z;\beta)}{\sigma}\right) - \frac{1}{2}\mathrm{sgn}\left(\frac{y-\delta(z;\beta)}{\sigma}\right)\right]$, we have*

$$\sup_{\beta\in\mathcal{B}}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}r_\beta(Y_i,z_i)^2 \lesssim \lambda_n^{-1}(s_2^2 + m_2^2).$$

Proof. First note that $|\rho_\lambda(t)| \lesssim 1 \wedge \frac{1}{\lambda|t|}$. Indeed, for $t > 0$, by (C.1), $\rho_\lambda(t) = -\frac{1}{\pi}\int_{\lambda t}^{\infty}\frac{\sin u}{u}\,du$, and integration by parts gives

$$\left|\int_a^\infty\frac{\sin u}{u}\,du\right| \le \frac{|\cos a|}{a} + \int_a^\infty\frac{|\cos u|}{u^2}\,du \le \frac{1}{a} + \int_a^\infty\frac{1}{u^2}\,du = \frac{2}{a}, \quad a > 0.$$

The case $t < 0$ follows by oddness, while the uniform boundedness follows from the boundedness of Si.

Set

$$X_i := \frac{Y_i - \mu_i}{\sigma_i} \sim \mathcal{N}(0,1), \qquad a := \frac{\mu_i - C}{\sigma_i}.$$

Then $(Y_i - C)/\sigma_i = X_i + a$ and $Y_i - k_i = (\mu_i - k_i) + \sigma_i X_i$. Thus it is enough to show, uniformly over $a \in \mathbb{R}$,

$$\mathbb{E}\rho_\lambda(X_i + a)^2 \lesssim \lambda^{-1}, \qquad \mathbb{E}X_i^2\rho_\lambda(X_i + a)^2 \lesssim \lambda^{-1}.$$

For the first display, using the boundedness of the normal density,

$$\mathbb{E}\rho_\lambda(X_i + a)^2 \lesssim \int_{\mathbb{R}} \left(1 \wedge \frac{1}{\lambda^2 t^2}\right) \varphi(t - a)\, dt \lesssim \int_{\mathbb{R}} \left(1 \wedge \frac{1}{\lambda^2 t^2}\right) dt \lesssim \lambda^{-1}.$$

For the second display,

$$\begin{aligned}\mathbb{E}X_i^2\rho_\lambda(X_i + a)^2 &= \int_{\mathbb{R}} (t - a)^2 \rho_\lambda(t)^2 \varphi(t - a)\, dt \lesssim \int_{\mathbb{R}} (t^2 + a^2)\rho_\lambda(t)^2\varphi(t - a)\, dt \\ &\leq \underbrace{\int_{\mathbb{R}} t^2 \cdot \frac{1}{\lambda^2 t^2}\varphi(t - a)\, dt}_{\lesssim \lambda^{-1}} + a^2 \int_{\mathbb{R}} \rho_\lambda(t)^2\varphi(t - a)\, dt\end{aligned}$$

It remains to bound $a^2 \int_{\mathbb{R}} \rho_\lambda(t)^2\varphi(t - a)\, dt$, uniformly over a. If $|a| \leq 2$, this is $\lesssim \lambda^{-1}$ by the previous bound. If $|a| > 2$, split the integral over $|t| \leq |a|/2$ and $|t| > |a|/2$. On the first region, $|t - a| \geq |a|/2$, so

$$a^2 \int_{|t| \leq |a|/2} \rho_\lambda(t)^2\varphi(t - a)\, dt \lesssim a^2 e^{-a^2/8} \int_{\mathbb{R}} \left(1 \wedge \frac{1}{\lambda^2 t^2}\right) dt \lesssim \lambda^{-1},$$

where the last inequality is due to that $a^2 e^{-a^2/8}$ is bounded over $a \in \mathbb{R}$. On the second region, $\rho_\lambda(t)^2 \lesssim \frac{1}{\lambda^2 a^2}$, and hence $a^2 \int_{|t| > |a|/2} \rho_\lambda(t)^2\varphi(t - a)\, dt \lesssim \lambda^{-2} \lesssim \lambda^{-1}$. Therefore

$$\mathbb{E}X_i^2\rho_\lambda(X_i + a)^2 \lesssim \lambda^{-1}$$

uniformly in a.

Combining the preceding bounds,

$$\mathbb{E}\left[(Y_i - k_i)^2\rho_\lambda\left(\frac{Y_i - C}{\sigma_i}\right)^2\right] \lesssim \lambda^{-1}\left\{\sigma_i^2 + (\mu_i - k_i)^2\right\}.$$

Averaging over i and taking the supremum over β gives the stated bound. □

Lemma C.5. *Let*

$$\rho_\lambda(t) := \frac{1}{\pi}\,\mathrm{Si}(\lambda t) - \frac{1}{2}\mathrm{sgn}(t),$$

fix $M \geq 1$, and define

$$\mathcal{R}_M := \left\{(y - k)\rho_{\lambda_n}\left(\frac{y - \delta(z;\beta)}{\sigma}\right)\mathbf{1}\{|y - k| \leq M\} : \beta \in \mathcal{B}\right\}.$$

Suppose $L_\sigma \geq 1$ satisfies $D(z)/\sigma \leq L_\sigma$, where D is the envelope function of the decision class $\mathcal{D}$. Let $0 < b \leq 1$ and $F_M(y, z) \geq C|y - k|\mathbf{1}\{|y - k| \leq M\}$ for some constant $C > 0$

and is an envelope for $\mathcal{R}_M$. Then

$$J(b, \mathcal{R}_M \mid F_M, L_2) \lesssim b\sqrt{V(\mathcal{D}) \log \frac{e\lambda_n L_\sigma}{b}}.$$

Proof. Define

$$\mathcal{G}_M := \left\{ \frac{1}{\pi}(y-k) \operatorname{Si}\left(\lambda_n \frac{y-\delta(z;\beta)}{\sigma}\right) \mathbf{1}\{|y-k| \le M\} : \beta \in \mathcal{B} \right\},$$
$$\mathcal{S}_M := \left\{ \frac{1}{2}(y-k)\operatorname{sgn}\left(\frac{y-\delta(z;\beta)}{\sigma}\right) \mathbf{1}\{|y-k| \le M\} : \beta \in \mathcal{B} \right\}.$$

For an arbitrary discrete probability measure Q, define the reweighted measure

$$dQ_{F_M}(y,z) := \frac{F_M(y,z)^2}{\|F_M\|_{Q,2}^2} dQ(y,z).$$

Let Q be any such measure and $0 < \eta \le 1$. Since $\mathcal{R}_M \subset \mathcal{G}_M - \mathcal{S}_M$, the product cover gives

$$N(\eta\|F_M\|_{Q,2}, \mathcal{R}_M, L_2(Q)) \le N(\eta\|F_M\|_{Q,2}/2, \mathcal{G}_M, L_2(Q)) \cdot N(\eta\|F_M\|_{Q,2}/2, \mathcal{S}_M, L_2(Q)).$$

We next construct an entropy bound for $\mathcal{G}_M$. Write

$$g_{\beta,M}(y,z) := \frac{1}{\pi}(y-k) \operatorname{Si}\left(\lambda_n \frac{y-\delta(z;\beta)}{\sigma}\right) \mathbf{1}\{|y-k| \le M\}.$$

The $\mathcal{G}_M$ has the deterministic envelope $G_M(y,z) := c|y-k|\mathbf{1}\{|y-k| \le M\} \le cF_M$, for some absolute constant $c > 0$ because $\operatorname{Si}(\cdot)$ is bounded. Define

$$\mathcal{D}/\sigma := \{(y,z) \mapsto \delta(z;\beta)/\sigma : \beta \in \mathcal{B}\}.$$

We claim that, for some absolute constants $c > 0$ and every $0 < \eta < 1$,

$$N(\eta\|F_M\|_{Q,2}, \mathcal{G}_M, L_2(Q)) \le N\left(c\frac{\eta}{\lambda_n}, \mathcal{D}/\sigma, L_2(Q_{F_M})\right). \tag{C.2}$$

Indeed, take any $c\eta/\lambda_n$-cover of $\mathcal{D}/\sigma$ in the $L_2(Q_{F_M})$ metric. For any element $\delta(\cdot;\beta)/\sigma$ in this class, let $\tilde{d}/\sigma$ be a covering element satisfying $\left\|(\delta(\cdot;\beta) - \tilde{d})/\sigma\right\|_{Q_{F_M},2} \le c\eta/\lambda_n$. Since Si is Lipschitz and the truncation is independent of β, we have for some universal constant $C > 0$,

$$\left\| g_{\beta,M}(y,z) - \frac{1}{\pi}(y-k) \operatorname{Si}\left(\lambda_n \frac{y-\tilde{d}(z)}{\sigma}\right) \mathbf{1}\{|y-k| \le M\} \right\|_{Q,2}$$
$$\le C\lambda_n \left(\int (y-k)^2 \mathbf{1}\{|y-k| \le M\} \left(\frac{\delta(z;\beta) - \tilde{d}(z)}{\sigma}\right)^2 dQ(y,z) \right)^{1/2}$$

$$\leq C\lambda_n \left(\int F_M(y,z)^2 \left(\frac{\delta(z;\beta) - \tilde{d}(z)}{\sigma} \right)^2 dQ(y,z) \right)^{1/2}$$
$$= C\lambda_n \|F_M\|_{Q,2} \left\| (\delta(\cdot;\beta) - \tilde{d})/\sigma \right\|_{Q_{F_M},2} \leq \eta \|F_M\|_{Q,2}.$$

Thus the pushforward of the $L_2(Q_{F_M})$ cover of $\mathcal{D}/\sigma$ through the map

$$d/\sigma \mapsto \frac{1}{\pi}(y-k)\,\mathrm{Si}\left(\lambda_n \frac{y - d(z)}{\sigma}\right) \mathbf{1}\{|y-k| \leq M\}$$

is an $\eta\|F_M\|_{Q,2}$-cover of $\mathcal{G}_M$ in $L_2(Q)$.

The class $\mathcal{D}/\sigma$ is VC-subgraph with index of order $V(\mathcal{D})$ and envelope D/σ. Since $D/\sigma \leq L_\sigma$ on the fixed design support, $c\frac{\eta}{\lambda_n} \geq c\frac{\eta}{\lambda_n L_\sigma}\|D/\sigma\|_{Q_{F_M},2}$. Applying Lemma D.2 to $\mathcal{D}/\sigma$ gives

$$\log N(\eta\|F_M\|_{Q,2}, \mathcal{G}_M, L_2(Q)) \lesssim V(\mathcal{D}) \log \frac{e\lambda_n L_\sigma}{\eta}.$$

For $\mathcal{S}_M$, note that $\mathrm{sgn}((y - \delta(z;\beta))/\sigma) = 2\mathbf{1}\{y \geq \delta(z;\beta)\} - 1$. Therefore $\mathcal{S}_M$ is a VC-subgraph class with envelope $F_M/2$ and index $O(V(\mathcal{D}))$, so

$$\log N(\eta\|F_M\|_{Q,2}, \mathcal{S}_M, L_2(Q)) \lesssim V(\mathcal{D}) \log \frac{e}{\eta}.$$

The preceding product cover implies, after changing constants, that for every $0 < \eta \leq 1$,

$$\log N(\eta\|F_M\|_{Q,2}, \mathcal{R}_M, L_2(Q)) \lesssim V(\mathcal{D}) \log \frac{e\lambda_n L_\sigma}{\eta} + V(\mathcal{D}) \log \frac{e}{\eta}.$$

Since $\lambda_n \geq 1$ and $L_\sigma \geq 1$, the second term is absorbed into the first one, so

$$\sup_Q \log N(\eta\|F_M\|_{Q,2}, \mathcal{R}_M, L_2(Q)) \lesssim V(\mathcal{D}) \log \frac{e\lambda_n L_\sigma}{\eta},$$

where the supremum is over discrete probability measures Q. Therefore

$$J(b, \mathcal{R}_M \mid F_M, L_2) = \int_0^b \sup_Q \sqrt{1 + \log N(\eta\|F_M\|_{Q,2}, \mathcal{R}_M, L_2(Q))}\, d\eta$$
$$\lesssim \sqrt{V(\mathcal{D})} \int_0^b \sqrt{\log \frac{e\lambda_n L_\sigma}{\eta}}\, d\eta.$$

Finally,

$$\int_0^b \sqrt{\log \frac{e\lambda_n L_\sigma}{\eta}}\, d\eta = b \int_0^1 \sqrt{\log \frac{e\lambda_n L_\sigma}{b} + \log \frac{1}{u}}\, du \lesssim b\sqrt{\log \frac{e\lambda_n L_\sigma}{b}},$$

which gives

$$J(b, \mathcal{R}_M \mid F_M, L_2) \lesssim b\sqrt{V(\mathcal{D}) \log \frac{e\lambda_n L_\sigma}{b}},$$

after enlarging the absolute constant. □

Proposition C.2. *Let* $\lambda_n = \sqrt{2\log n}$

$$L_{\sigma,n} := 1 \vee \max_{1\le i\le n} \frac{D(z_i)}{\sigma_i}, \qquad \omega_{4,n}^4 := m_4^4 + s_4^4, \qquad U_{H,n} := \max_{1\le i\le n}\{\sigma_i + D(z_i)\}.$$

Further define $M := \log(eL_{\sigma,n})\left\{V(\mathcal{D})(1+s_2+m_2+\|D\|_n + U_{H,n}) + \omega_{4,n}^4\right\}$. *Then, in the proof of Theorem 1,*

$$\mathbb{E}\sup_{\beta\in\mathcal{B}}|A_2(\beta)| \lesssim M\frac{(\log n)^{1/4}\sqrt{\log\log n}}{\sqrt{n}}.$$

Proof. Write

$$S_n := s_2 + m_2, \quad \rho_{\lambda_n}(t) := \frac{1}{\pi}\mathrm{Si}(\lambda_n t) - \frac{1}{2}\mathrm{sgn}(t).$$

Let

$$P_i f := \mathbb{E}[f(Y_i,z_i)], \qquad \bar{P}f := \frac{1}{n}\sum_{i=1}^n P_i f, \qquad \mathbb{G}_n f_\beta := \frac{1}{\sqrt{n}}\sum_{i=1}^n\{f_\beta(Y_i,z_i) - P_i f_\beta\}.$$

Define

$$q_\beta(y,z) := \frac{1}{2}(y-k)\mathrm{sgn}\left(\frac{y-\delta(z;\beta)}{\sigma}\right), \quad \mathcal{Q} := \{q_\beta : \beta\in\mathcal{B}\},$$
$$r_\beta(y,z) := (y-k)\rho_{\lambda_n}\left(\frac{y-\delta(z;\beta)}{\sigma}\right), \quad \mathcal{R} := \{r_\beta : \beta\in\mathcal{B}\},$$
$$h_\beta(y,z) := \sigma\lambda_n\,\mathrm{sinc}\left(\lambda_n\frac{y-\delta(z;\beta)}{\sigma}\right), \quad \mathcal{H} := \{h_\beta : \beta\in\mathcal{B}\}.$$

We write $\|\mathbb{G}_n\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}}|\mathbb{G}_n f|$. We then have

$$\mathbb{E}\sup_{\beta\in\mathcal{B}}|A_2(\beta)| \le \underbrace{\frac{1}{2}\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n(Y_i-\mu_i)\right|}_{\lesssim s_2/\sqrt{n}} + \frac{1}{\sqrt{n}}\left\{\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{Q}} + \mathbb{E}\|\mathbb{G}_n\|_{\mathcal{R}} + \mathbb{E}\|\mathbb{G}_n\|_{\mathcal{H}}\right\}, \quad \text{(C.3)}$$

We begin with $\|\mathbb{G}_n\|_{\mathcal{Q}}$. Since $\mathrm{sgn}((y-\delta(z;\beta))/\sigma) = 2\mathbf{1}\{y\ge\delta(z;\beta)\} - 1$, $\mathcal{Q}$ is VC-subgraph with index $O(V(\mathcal{D}))$ and envelope $|y-k|$. Therefore Theorem D.1 and Lemma D.2 give

$$\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{Q}} \lesssim \sqrt{V(\mathcal{D})}(s_2+m_2). \quad \text{(C.4)}$$

Moving on to $\|\mathbb{G}_n\|_{\mathcal{R}}$, set $M_n := n^{1/4}$ and truncate r_β:

$$r_{\beta,M_n}(y,z) := r_\beta(y,z)\mathbf{1}\{|y-k|\le M_n\}, \qquad \mathcal{R}_{M_n} := \{r_{\beta,M_n} : \beta\in\mathcal{B}\}.$$

For a sufficiently large absolute constant C,

$$F_{R,M_n}(y,z) := C\left(|y-k|\mathbf{1}\{|y-k| \le M_n\} + S_n\right)$$

is an envelope for $\mathcal{R}_{M_n}$, satisfies $\|F_{R,M_n}\|_{\bar{P},2} \asymp S_n, F_{R,M_n} \le C(M_n + S_n)$, and, by Lemma C.4,

$$\sup_{\beta\in\mathcal{B}} \bar{P} r^2_{\beta,M_n} \le \sup_{\beta\in\mathcal{B}} \bar{P} r^2_{\beta} \lesssim \lambda_n^{-1} S_n^2 \lesssim \lambda_n^{-1} \|F_{R,M_n}\|^2_{\bar{P},2}. \tag{C.5}$$

Apply Lemma C.5 with F_M therein as F_{R,M_n}. We obtain

$$J\left(\lambda_n^{-1/2}, \mathcal{R}_{M_n} \mid F_{R,M_n}, L_2\right) \lesssim \lambda_n^{-1/2}\sqrt{V(\mathcal{D})\log\left(e\lambda_n^{3/2}L_{\sigma,n}\right)}. \tag{C.6}$$

Applying Corollary D.1 with the localization condition (C.5) and (C.6) yields

$$\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{R}_{M_n}} \lesssim \lambda_n^{-1/2} S_n \sqrt{V(\mathcal{D})\log\left(e\lambda_n^{3/2}L_{\sigma,n}\right)} + \frac{(M_n+S_n)V(\mathcal{D})\log\left(e\lambda_n^{3/2}L_{\sigma,n}\right)}{\sqrt{n}}. \tag{C.7}$$

Moreover, since ρ_{λ_n} is bounded,

$$\mathbb{E}\sup_{\beta\in\mathcal{B}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \{(r_\beta - r_{\beta,M_n})(Y_i,z_i) - P_i(r_\beta - r_{\beta,M_n})\}\right| \tag{C.8}$$

$$\le \mathbb{E}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \left\{\sup_{\beta\in\mathcal{B}}|r_\beta - r_{\beta,M_n}|(Y_i,z_i) + P_i \sup_{\beta\in\mathcal{B}}|r_\beta - r_{\beta,M_n}|\right\}\right| \tag{C.9}$$

$$\lesssim \sqrt{n}\,\frac{\omega^4_{4,n}}{M_n^3}. \tag{C.10}$$

Indeed, (C.10) follows from $|r_\beta - r_{\beta,M_n}| \lesssim |y-k|\mathbf{1}\{|y-k| > M_n\}$, $\mathbb{E}|X|\mathbf{1}\{|X| > M_n\} \le \mathbb{E}|X|^4/M_n^3$, and $\frac{1}{n}\sum_{i=1}^n P_i|Y_i - k_i|^4 \lesssim s_4^4 + m_4^4 = \omega^4_{4,n}$.

Combining (C.7) and (C.10), and substituting $M_n = n^{1/4}$, gives

$$\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{R}} \lesssim \lambda_n^{-1/2} S_n \sqrt{V(\mathcal{D})\log\left(e\lambda_n^{3/2}L_{\sigma,n}\right)} + \frac{(n^{1/4}+S_n)V(\mathcal{D})\log\left(e\lambda_n^{3/2}L_{\sigma,n}\right)}{\sqrt{n}} + \frac{\omega^4_{4,n}}{n^{1/4}}. \tag{C.11}$$

Finally, we bound $\|\mathbb{G}_n\|_{\mathcal{H}}$. Let $F_H(y,z) := C\lambda_n(\sigma + D(z))$. For C large enough this is an envelope for $\mathcal{H}$, and

$$\|F_H\|_{\bar{P},2} \lesssim \lambda_n(s_2 + \|D\|_n), \qquad F_H \le C\lambda_n U_{H,n}.$$

By Lemma C.3,

$$\sup_{\beta\in\mathcal{B}} \bar{P} h^2_\beta \lesssim \lambda_n s_2^2 \lesssim \lambda_n^{-1}\|F_H\|^2_{\bar{P},2}. \tag{C.12}$$

Since sinc' is bounded, $|h_\beta(y,z) - h_{\tilde\beta}(y,z)| \lesssim \lambda_n^2 |\delta(z;\beta) - \delta(z;\tilde\beta)|$. Thus, for every discrete probability measure Q and $0 < \eta < 1$,

$$N\left(\eta \|F_H\|_{Q,2}, \mathcal{H}, L_2(Q)\right) \le N\left(c\frac{\eta}{\lambda_n}\|D\|_{Q_z,2}, \mathcal{D}, L_2(Q_z)\right),$$

where Q_z is the marginal measure on z. By Lemma D.2,

$$J\left(\lambda_n^{-1/2}, \mathcal{H} \mid F_H, L_2\right) \lesssim \lambda_n^{-1/2}\sqrt{V(\mathcal{D})\log(e\lambda_n)}. \tag{C.13}$$

Applying Corollary D.1 with (C.12) and (C.13) gives

$$\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{H}} \lesssim \lambda_n^{1/2}\sqrt{V(\mathcal{D})\log(e\lambda_n)}(s_2 + \|D\|_n) + \frac{\lambda_n U_{H,n} V(\mathcal{D})\log(e\lambda_n)}{\sqrt{n}}. \tag{C.14}$$

Combining (C.3), (C.4), (C.11), and (C.14), and using $V(\mathcal{D}) \ge 1$, $\lambda_n = \sqrt{2\log n}$, $\log(eL_{\sigma,n}) \ge 1$, gives

$$\mathbb{E}\sup_{\beta\in\mathcal{B}}|A_2(\beta)| \lesssim M\frac{(\log n)^{1/4}\sqrt{\log\log n}}{\sqrt{n}}. \qquad \square$$

C.3. **Tail bound for regret.** In addition to controlling the expected welfare, empirical process techniques can also get high probability guarantees on the *realized* in-sample regret, defined as the quantity $u(\beta^*) - u(\hat\beta)$. We give a basic tail bound result for ASSURE* using the i.n.i.d. extension of van der Vaart and Wellner (2023), Theorem 2.14.23 from Appendix D.

Theorem C.1. *Fix $\delta > 0$. Under the assumptions of Theorem 1, with probability at least $1 - \delta$*

$$u(\beta^*) - u(\hat\beta) \lesssim \frac{m_1}{n\log n} + \frac{m_2 + (M + s_2)\sqrt{\log n}}{\sqrt{n}}\sqrt{\log\frac{2}{\delta}}$$

where M is the constant in the statement of Proposition C.2. The same bound is true for $W(\beta^*) - W(\hat\beta)$.

Proof. The optimality of $\hat\beta$ for $\hat W_n(\cdot)$ shows that

$$u(\beta^*) - u(\hat\beta) \le u(\beta^*) - \hat W_n(\beta^*) + \hat W_n(\hat\beta) - u(\hat\beta) \le 2\sup_{\beta\in\mathcal{B}}|u(\beta) - \hat W_n(\beta)|$$

Combined with the uniform bias bound on $\hat W_n$ from Proposition 1, we have

$$u(\beta^*) - u(\hat\beta) \le 2\sup_{\beta\in\mathcal{B}}\left|u(\beta) - \hat W_n(\beta) - \mathbb{E}\left[u(\beta) - \hat W_n(\beta)\right]\right| + \frac{2m_1}{n\log n}.$$

Let $\|\mathbb{G}_n\|_{\mathcal{F}} := \sqrt{n}\sup_{\beta\in\mathcal{B}}\left|u(\beta) - \hat W_n(\beta) - \mathbb{E}\left[u(\beta) - \hat W_n(\beta)\right]\right|$.

Here, $\mathcal{F}$ is the function class with members, indexed by $\beta \in \mathcal{B}$,

$$f_\beta(y,\mu,z) = (\mu - k)\mathbf{1}\{y > \delta(z;\beta)\} - w_h(y;z,\beta)$$

with $z = (k, \sigma, x)$ and w_h defined in (2.7). We will control the sub-Gaussian norm of $\|\mathbb{G}_n\|_{\mathcal{F}}$. Notice that an envelope function for this class is given by $|\mu - k| + C|y - k| + \lambda_n \sigma$, which is in turn bounded above by

$$F(y, \mu, k, \sigma) := C(|\mu - k| + |y - \mu| + \lambda_n \sigma)$$

for some constant C. Applying Proposition D.1 with $V_i = (Y_i, \mu_i, z_i)$, the sub-Gaussian norm is bounded above by a constant multiple times

$$\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}} + \left(\frac{1}{n}\sum_{i=1}^{n}(\mu_i - k_i)^2 + \lambda_n^2 \sigma_i^2\right)^{1/2} \lesssim \mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}} + m_2 + s_2 \lambda_n$$

By definition of the Orlicz norm ψ_2, any random variable satisfies $\mathbb{P}(|Y| > t) \leq 2\exp\left(-t^2/\|Y\|_{\psi_2}^2\right)$. Equivalently, with probability $\geq 1 - \delta$, $|Y| < \|Y\|_{\psi_2}\sqrt{\log \frac{2}{\delta}}$. Combining this with the bound on Equation (A.2) in Theorem 1 yields

$$u(\beta^*) - u(\hat{\beta}) \lesssim \frac{m_1}{n \log n} + \frac{M(\log n)^{1/4}\sqrt{\log\log n} + m_2 + s_2\sqrt{\log n}}{\sqrt{n}}\sqrt{\log \frac{2}{\delta}},$$

from which the statement follows.

The analogous result follows for $W(\beta^*) - W(\hat{\beta})$ by first noting that

$$W(\beta^*) - W(\hat{\beta}) \leq W(\beta^*) - \hat{W}_n(\beta^*) + \hat{W}_n(\hat{\beta}) - W(\hat{\beta}) \leq 2\sup_{\beta \in \mathcal{B}} |\hat{W}_n(\beta) - W(\beta)|,$$

and then repeating the same steps of the proof. □

C.4. **Auxiliary results for Theorem 3.** We start with a lemma calculating the derivatives of the average welfare. Since the computation is straightforward, we omit the proof.

Lemma C.6 (Derivatives of the welfare)**.** *Assume the decision thresholds $\delta(\cdot, Z)$ are smooth for each Z, and set $T_i(\beta) := (\mu_i - \delta(z_i; \beta))/\sigma_i$. Then the derivatives of the averaged welfare are given by*

$$W'(\beta) := -\frac{1}{n}\sum_{i=1}^{n}(\mu_i - k_i)\varphi\left(T_i(\beta)\right)\frac{\delta'(\beta, z_i)}{\sigma_i}$$

$$W''(\beta) := -\frac{1}{n}\sum_{i=1}^{n}(\mu_i - k_i)\left[\varphi\left(T_i(\beta)\right)\frac{\delta''(\beta, z_i)}{\sigma_i} - \varphi'\left(T_i(\beta)\right)\frac{\delta'(\beta, z_i)^2}{\sigma_i^2}\right]$$

and

$$W'''(\beta) := -\frac{1}{n}\sum_{i=1}^{n}(\mu_i - k_i)A_i$$

$$A_i := \varphi\left(T_i(\beta)\right)\frac{\delta'''(\beta, z_i)}{\sigma_i} - 3\varphi'\left(T_i(\beta)\right)\frac{\delta''(\beta, z_i)\delta'(\beta, z_i)}{\sigma_i^2} + \varphi''\left(T_i(\beta)\right)\frac{\delta'(\beta, z_i)^3}{\sigma_i^3}.$$

Under Assumption 1(1), there exists a constant C_B that only depends on B such that

$$\sup_{\beta} |W'(\beta)| + |W''(\beta)| + |W'''(\beta)| \leq C_B.$$

Theorem C.2. *Under the assumptions of Theorem 3, for n large enough, we have*

$$\left|\hat{\beta} - \beta^*\right| \leq \frac{2C}{\kappa} \frac{(\log n)^3}{\sqrt{n}}$$

with probability at least $1 - \frac{2}{n}$, for a constant C depending only on B and the VC indices of $\mathcal{D}, \mathcal{D}'$. As a result,

$$\mathbb{E}\left[W(\beta^*) - W(\hat{\beta})\right] \leq C \frac{(\log n)^6}{n}$$

for an absolute constant C depending only on B, κ, and the VC indices of $\mathcal{D}, \mathcal{D}'$.

Proof. Firstly, inspect the derivatives of the welfare from Lemma C.6. By the results of Theorem C.1 we have with probability at least $1 - \delta$

$$W(\beta^*) - W(\hat{\beta}) \leq c \left\{ \frac{m_1}{n \log n} + \frac{m_2 + (M + s_2)\sqrt{\log n}}{\sqrt{n}} \sqrt{\log \frac{2}{\delta}} \right\},$$

where c is a universal constant in Theorem C.1 and M is the constant in the statement of Proposition C.2. Substituting $\delta = 1/n$ and using Assumption 1(1), we have

$$W(\beta^*) - W(\hat{\beta}) \leq CB \frac{\log n}{\sqrt{n}} \tag{C.15}$$

with probability at least $1 - 1/n$ and some absolute constant C. Let A_n be this event.

Take n large enough such that the right hand side of Eq. (C.15) is less than ξ. Using Assumption 1(4), we conclude that on A_n, $|\beta^* - \hat{\beta}| \leq \kappa/B$. Now, Taylor expand W' around β^* to obtain

$$W'(\beta^*) - W'(\hat{\beta}) = -W''(\beta^*)(\beta^* - \hat{\beta}) - \frac{W'''(\tilde{\beta})}{2}(\beta^* - \hat{\beta})^2 \tag{C.16}$$

for some $\tilde{\beta}$ between $\hat{\beta}$ and β^*. By Lemma C.6 and the reverse triangle inequality $|a - b| \geq |a| - |b|$ we have

$$\left|W'(\beta^*) - W'(\hat{\beta})\right| \geq |W''(\beta^*)| \left|\beta^* - \hat{\beta}\right| - \frac{|W'''(\tilde{\beta})|}{2}(\beta^* - \hat{\beta})^2 \geq \kappa |\beta^* - \hat{\beta}| - \frac{C_B}{2}(\beta^* - \hat{\beta})^2,$$

where the last inequality follows from Assumption 1(3). Combining this with the Taylor expansion in Eq. (C.16), for n large enough, on the event A_n,

$$\frac{\kappa}{2} \left|\beta^* - \hat{\beta}\right| \leq \left|W'(\beta^*) - W'(\hat{\beta})\right|. \tag{C.17}$$

Next, by Theorem C.5, with probability at least $1 - 1/n$

$$\left|W'(\hat{\beta}) - W'(\beta^*)\right| \leq \frac{C}{n\sqrt{\log n}} + \frac{C(\log n)^{5/2}}{\sqrt{n}}\sqrt{\log(2n)} \tag{C.18}$$

for a constant C depending only on B and the VC indices of $\mathcal{D}, \mathcal{D}'$. Let $\tilde{A}_n$ denote this event. Combining equations (C.17) and (C.18), we conclude that on $A_n \cap \tilde{A}_n$, which has probability at least $1 - \frac{2}{n}$, that

$$\left|\beta^* - \hat{\beta}\right| \leq \frac{2C}{\kappa}\frac{(\log n)^3}{\sqrt{n}}$$

for a C depending only on B and $V(\mathcal{D}), V(\mathcal{D}')$. This proves the first claim.

Next, notice that $W'(\beta^*) = 0$. By Taylor expansion of W around β^*, we have

$$W(\beta^*) - W(\hat{\beta}) \lesssim C_B(\hat{\beta} - \beta^*)^2 + C_B(\hat{\beta} - \beta^*)^3.$$

Then for n large enough,

$$\begin{aligned}
\mathbb{E}\left[W(\beta^*) - W(\hat{\beta})\right] &\leq \mathbb{E}\left[W(\beta^*) - W(\hat{\beta}); A_n \cap \tilde{A}_n\right] + \mathbb{E}\left[W(\beta^*) - W(\hat{\beta}); (A_n \cap \tilde{A}_n)^c\right] \\
&\lesssim C_B\mathbb{E}\left[(\beta^* - \hat{\beta})^2; A_n \cap \tilde{A}_n\right] + B\mathbb{P}\left((A_n \cap \tilde{A}_n)^c\right) \\
&\lesssim \frac{C_B C^2}{\kappa^2}\frac{(\log n)^6}{n} + \frac{B}{n}.
\end{aligned}$$

This concludes the proof. □

Theorem C.3 (Uniform bounds on the derivative bias)**.** *We have*

$$\begin{aligned}
|W'(\beta) - \mathbb{E}\hat{W}_n'(\beta)| &\lesssim \left(\frac{1}{n}\sum_{i=1}^n \frac{|\mu_i - k_i|}{\sigma_i}|\delta'(\beta, z_i)|\right) he^{-1/(2h^2)} \\
|W''(\beta) - \mathbb{E}\hat{W}_n''(\beta)| &\lesssim \\
&\left[h\left(\frac{1}{n}\sum_{i=1}^n \frac{|\mu_i - k_i|}{\sigma_i}|\delta''(\beta, z_i)|\right) + \left(\frac{1}{n}\sum_{i=1}^n \frac{|\mu_i - k_i|}{\sigma_i^2}\delta'(\beta, z_i)^2\right)\right] e^{-1/(2h^2)}
\end{aligned}$$

With the choice of $h = 1/\sqrt{2\log n}$ and uniform upper bounds on δ', the bias for the first derivative is on the order of $\frac{1}{n\sqrt{\log n}}\left(\frac{1}{n}\sum_{i=1}^n \frac{|\mu_i - k_i|}{\sigma_i}\right)$.

Proof of Theorem C.3. Lemma C.2 shows

$$\begin{aligned}
(\mu - k)\Phi\left(\frac{\mu - C}{\sigma}\right) &= \frac{1}{2}(\mu - k) + \frac{1}{\pi}(\mu - k)\int_0^\infty \frac{1}{\omega}e^{-\frac{1}{2}\omega^2}\sin(\omega(\mu - C)/\sigma)d\omega \\
\mathbb{E}\Psi_h(Y, z, C) &= \frac{1}{2}(\mu - k) + \frac{1}{\pi}(\mu - k)\int_0^{1/h} \frac{1}{\omega}e^{-\frac{1}{2}\omega^2}\sin(\omega(\mu - C)/\sigma)d\omega
\end{aligned}$$

We may differentiate under the Lebesgue integral in C, which is justified by dominated convergence theorem. This yields

$$-\frac{\mu-k}{\sigma}\varphi\left(\frac{\mu-C}{\sigma}\right) = -\frac{1}{\pi}\frac{\mu-k}{\sigma}\int_0^\infty e^{-\frac{1}{2}\omega^2}\cos(\omega(\mu-C)/\sigma)d\omega$$
$$\frac{d}{dC}\mathbb{E}\Psi_h(Y,z,C) = -\frac{1}{\pi}\frac{\mu-k}{\sigma}\int_0^{1/h} e^{-\frac{1}{2}\omega^2}\cos(\omega(\mu-C)/\sigma)d\omega.$$

By the chain rule, we have

$$\hat{W}_n'(\beta) = \frac{1}{n}\sum_{i=1}^n \frac{d\Psi_h}{dC}(Y_i, z_i, \delta(z_i;\beta))\delta'(\beta, z_i)$$

By the formulas for derivatives of Ψ in Lemma B.2 and Lemma B.1 shows, $|\frac{d\Psi_h}{dC}| \le K_h|Y-k|/\sigma$ for some constant K_h depending on h. This provides a dominating integrable function which justifies DCT, yielding $\frac{d}{dC}\mathbb{E}\Psi_h = \mathbb{E}\frac{d\Psi_h}{dC}$. Therefore,

$$\begin{aligned}|W'(\beta) - \mathbb{E}\hat{W}_n'(\beta)| &\le \frac{1}{\pi}\left(\frac{1}{n}\sum_{i=1}^n \frac{|\mu_i-k_i|}{\sigma_i}\delta'(\beta,z_i)\right)\int_{1/h}^\infty e^{-\frac{1}{2}\omega^2}\cos(\omega(\mu-C)/\sigma)d\omega \\ &\lesssim \left(\frac{1}{n}\sum_{i=1}^n \frac{|\mu_i-k_i|}{\sigma_i}\delta'(\beta,z_i)\right) h e^{-1/2h^2}\end{aligned}$$

using the constant upper bound for $\cos$ and the Mills ratio bound.

For the second derivative, the argument is similar. For notational clarity we will write Ψ_C, Ψ_{CC} to be the first and second derivatives in C of Ψ_h. Applying DCT again to justify the interchange of derivative and integral, we obtain

$$\frac{\mu-k}{\sigma^2}\varphi'\left(\frac{\mu-C}{\sigma}\right) = -\frac{1}{\pi}\frac{\mu-k}{\sigma^2}\int_0^\infty \omega e^{-\frac{1}{2}\omega^2}\sin(\omega(\mu-C)/\sigma)d\omega$$
$$\mathbb{E}\Psi_{CC}(Y,z,C) = -\frac{1}{\pi}\frac{\mu-k}{\sigma^2}\int_0^{1/h} \omega e^{-\frac{1}{2}\omega^2}\sin(\omega(\mu-C)/\sigma)d\omega.$$

Next, notice that the second derivative of ASSURE* is given by

$$\hat{W}_n''(\beta) = \frac{1}{n}\sum_{i=1}^n \Psi_{CC}(Y_i,z_i,\delta(z_i;\beta))\delta'(\beta,z_i)^2 + \Psi_C(Y_i,z_i,\delta(z_i;\beta))\delta''(\beta,z_i). \quad \text{(C.19)}$$

Thus,

$$\begin{aligned}|\mathbb{E}\hat{W}_n''(\beta) - W''(\beta)| \lesssim &\left(\frac{1}{n}\sum_{i=1}^n \frac{|\mu_i-k_i|}{\sigma_i^2}\delta'(\beta,z_i)^2\right)\sup_\mu \int_{1/h}^\infty \omega e^{-\frac{1}{2}\omega^2}|\sin(\omega(\mu-C)/\sigma)|d\omega \\ &+ \left(\frac{1}{n}\sum_{i=1}^n \frac{|\mu_i-k_i|}{\sigma_i}|\delta''(\beta,z_i)|\right)\sup_\mu \int_{1/h}^\infty e^{-\frac{1}{2}\omega^2}|\cos(\omega(\mu-C)/\sigma)|d\omega.\end{aligned}$$

The conclusion follows from bounding $\cos, \sin$ above by 1 and bounding the resulting integrals. □

Theorem C.4. *In the setting and notation of Theorem 3, we have*

$$\mathbb{P}\left(\exists i : |\hat{\beta}^{(-i)} - \hat{\beta}| > \frac{4MB\log n}{\kappa n}\right) = O\left(\frac{1}{n}\right),$$

for some universal constant $M > 0$.

Proof of Theorem C.4. Recall the function Ψ in the ASSURE* estimator of (A.1). Let Ψ_C, Ψ_{CC} denote the first and second partial derivatives in C, given in Lemma B.2. To control $\hat{\beta} - \hat{\beta}^{(-i)}$, we will apply a standard Taylor expansion using the first-order conditions

$$\sum_{i=1}^{n} \Psi_C(Y_i, z_i, \delta(\hat{\beta}, z_i))\delta'(\hat{\beta}, z_i) = 0 \tag{C.20}$$

$$\sum_{j\neq i} \Psi_C(Y_j, z_j, \delta(\hat{\beta}^{(-i)}, z_j))\delta'(\hat{\beta}^{(-i)}, z_j) = 0, \quad \forall i = 1, \ldots, n. \tag{C.21}$$

By Taylor's theorem applied to $\sum_{i=1}^{n} \Psi_C(Y_i, z_i, \delta(\bullet, z_i))\delta'(\bullet, z_i)$, we have

$$\begin{aligned}
0 &= \frac{1}{n}\sum_{i=1}^{n} \Psi_C(Y_i, z_i, \delta(\hat{\beta}, z_i))\delta'(\hat{\beta}, z_i) \\
&= \frac{1}{n}\sum_{j=1}^{n} \Psi_C(Y_j, z_j, \delta(\hat{\beta}^{(-i)}, z_j))\delta'(\hat{\beta}^{(-i)}, z_j) + \Delta_{n,i}(\hat{\beta} - \hat{\beta}^{(-i)}) \\
&= \frac{1}{n}\Psi_C(Y_i, z_i, \delta(\hat{\beta}^{(-i)}, z_i))\delta'(\hat{\beta}^{(-i)}, z_i) + \Delta_{n,i}(\hat{\beta} - \hat{\beta}^{(-i)}).
\end{aligned}$$

where

$$\Delta_{n,i} := \left(\frac{1}{n}\sum_{j=1}^{n} \Psi_{CC}(Y_j, z_j, \delta(\tilde{\beta}_i, z_j))\delta'(\tilde{\beta}_i, z_j)^2 + \Psi_C(Y_j, \tilde{\beta}_i)\delta''(\tilde{\beta}_i, z_j)\right) \tag{C.22}$$

$$= \hat{W}_n''(\tilde{\beta}_i). \tag{C.23}$$

for some $\tilde{\beta}_i$ between $\hat{\beta}$ and $\hat{\beta}^{(-i)}$. By Assumption 1(1), $|\delta'(\hat{\beta}^{(-i)}, z_i)| \leq B$. By Lemma B.2, it is not difficult to show that

$$|\Psi_C(Y_i, z_i, \delta(\hat{\beta}^{(-i)}, z_i))| \lesssim \lambda_n^2 + \lambda_n|\delta(\hat{\beta}^{(-i)}, z_i)| \leq MB\lambda_n^2,$$

for some universal constant $M > 0$. Thus,

$$\left|\hat{\beta} - \hat{\beta}^{(-i)}\right| \leq \frac{MB\lambda_n^2}{n|\Delta_{n,i}|}. \tag{C.24}$$

By Lemma C.7 and a union bound, with probability at least $1 - O(n^{-1})$ we have $|\Delta_{n,i}| \geq \kappa/2$ for all i. On this event, (C.24) yields $\left|\hat{\beta} - \hat{\beta}^{(-i)}\right| \leq \frac{2MB\lambda_n^2}{\kappa n}$. Since $\lambda_n^2 = 2\log n$, enlarging

the absolute constant M if necessary gives

$$\left|\hat{\beta} - \hat{\beta}^{(-i)}\right| \leq \frac{4MB\log n}{\kappa n}.$$

This proves the theorem. □

Lemma C.7. *In the setting and notation of Theorem C.4, let $\tilde{\beta}_i$ be any point between $\hat{\beta}$ and $\hat{\beta}^{(-i)}$, and define $\Delta_{n,i} := \hat{W}_n''(\tilde{\beta}_i)$. Then, for n large enough,*

$$\mathbb{P}\left(|\Delta_{n,i}| \geq \kappa/2\right) \geq 1 - O\left(\frac{1}{n^2}\right).$$

Proof. Write

$$I_n := \sup_\beta \left|\hat{W}_n''(\beta) - \mathbb{E}\left[\hat{W}_n''(\beta)\right]\right|, \quad II_n := \sup_\beta \left|\mathbb{E}\left[\hat{W}_n''(\beta)\right] - W''(\beta)\right|,$$

$$III_n := C_B \max_i \max\left(|\beta^* - \hat{\beta}^{(-i)}|, |\beta^* - \hat{\beta}|\right),$$

where C_B is a constant depending only on B. Since W''' is uniformly bounded by C_B and $\tilde{\beta}_i$ lies between $\hat{\beta}$ and $\hat{\beta}^{(-i)}$, we have $\left|W''(\tilde{\beta}_i) - W''(\beta^*)\right| \leq III_n$. Hence, by the reverse triangle inequality and Assumption 1(3), $|\Delta_{n,i}| \geq \kappa - I_n - II_n - III_n$.

The three error terms are all $o(1)$ with probability $1 - O(n^{-1})$. First, by Assumption 1(1) and Theorem C.3, $II_n \leq \left(\frac{B^3}{\sqrt{2\log n}} + B^5\right)\frac{1}{n} = O(n^{-1})$. Second, Theorem C.6 with $\delta = n^{-1}$ gives $I_n \leq C\frac{(\log n)^4}{\sqrt{n}}$ with probability at least $1 - 1/n$. Third, Theorem C.2 gives $|\hat{\beta} - \beta^*| \leq C\frac{(\log n)^3}{\sqrt{n}}$ with probability at least $1 - 2/n$. Let $\beta^{(-i)*}$ maximize the leave-one-out welfare $W^{(-i)}$. The proof of Theorem C.2 applies verbatim to $W^{(-i)}$, except that δ is set to be n^{-2}, so

$$|\hat{\beta}^{(-i)} - \beta^{(-i)*}| \leq C\frac{(\log n)^3}{\sqrt{n}}$$

with probability at least $1 - O(n^{-2})$. Applying the union bound, we obtain that, with probability $1 - O(n^{-1})$

$$|\hat{\beta}^{(-i)} - \beta^{(-i)*}| \leq C\frac{(\log n)^3}{\sqrt{n}}, \quad \forall i = 1, \ldots, n.$$

Moreover, W and $W^{(-i)}$ differ uniformly by $O(n^{-1})$ under Assumption 1(1), hence

$$W(\beta^*) - W(\beta^{(-i)*}) = O(n^{-1}).$$

Since $W''(\beta^*) < -\kappa$ and W''' is uniformly bounded, there exists $r > 0$ such that $W''(\beta) \leq -\kappa/2$ for all $|\beta - \beta^*| \leq r$. Because Θ is compact and β^* is the unique global maximizer of W,

$$\xi_r := W(\beta^*) - \sup_{|\beta - \beta^*| \geq r} W(\beta) > 0.$$

For n large enough, the bound above is smaller than ξ_r, so $|\beta^{(-i)*} - \beta^*| < r$. Therefore Taylor'xs theorem gives

$$W(\beta^*) - W(\beta^{(-i)*}) = -\frac{1}{2}W''(\bar{\beta}_i)\left(\beta^{(-i)*} - \beta^*\right)^2 \geq \frac{\kappa}{4}\left|\beta^{(-i)*} - \beta^*\right|^2$$

for some $\bar{\beta}_i$ between β^* and $\beta^{(-i)*}$. Consequently, $|\beta^{(-i)*} - \beta^*| = O(n^{-1/2})$, and, by a union bound, $III_n = O\left(\frac{(\log n)^3}{\sqrt{n}}\right)$ with probability $1 - O(n^{-1})$. For n large enough, the sum of the bounds for I_n, II_n, and III_n is at most $\kappa/2$. The conclusion follows by a union bound. □

Theorem C.5. *Under the same assumptions of Theorem 3,*

$$\mathbb{E}|W'(\beta^*) - W'(\hat{\beta})| \leq \frac{C}{n\sqrt{\log n}} + \frac{C(\log n)^{5/2}}{\sqrt{n}}\left(\sqrt{\mathcal{V}(\mathcal{D}')} + \sqrt{\mathcal{V}(\mathcal{D})}\right).$$

for an absolute constant C depending on B. Moreover, fix any $\delta \in (0,1)$. Then with probability at least $1-\delta$, we have

$$|W'(\beta^*) - W'(\hat{\beta})| \leq \frac{C}{n\sqrt{\log n}} + \frac{C(\log n)^{5/2}}{\sqrt{n}}\left(\sqrt{\mathcal{V}(\mathcal{D}')} + \sqrt{\mathcal{V}(\mathcal{D})}\right)\sqrt{\ln\frac{2}{\delta}}.$$

Proof of Theorem C.5. By Assumption 1(2), $W'(\beta^*) = 0$ and by definition we have $\hat{W}'(\hat{\beta}) = 0$. Therefore,

$$|W'(\beta^*) - W'(\hat{\beta})| = |W'(\hat{\beta})| = |W'(\hat{\beta}) - \hat{W}_n'(\hat{\beta})| \leq \sup_\beta |W'(\beta) - \hat{W}_n'(\beta)|,$$

which can in turn be bounded above by

$$\sup_\beta |W'(\beta) - \mathbb{E}\hat{W}_n'(\beta)| + \sup_\beta |\hat{W}_n'(\beta) - \mathbb{E}\hat{W}_n'(\beta)|. \tag{C.25}$$

The first term represents the bias of the derivative of the ASSURE* estimator, which we bounded in Theorem C.3 by, up to universal constants,

$$\sup_\beta \left(\frac{1}{n}\sum_{i=1}^n \frac{|\mu_i - k_i|}{\sigma_i}|\delta'(\beta, z_i)|\right)\frac{1}{n\sqrt{2\log n}}.$$

We will focus on bounding the second term $\sup_\beta |\hat{W}_n'(\beta) - \mathbb{E}\hat{W}_n'(\beta)|$ using i.n.i.d. empirical process theory. The proof structure follows exactly the proof of Theorem 1. Recalling the notation in Eq. (B.1), we need to bound the quantity

$$n^{-1/2}\sup_\beta \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\Psi_C(Y_i, z_i, \delta(z_i;\beta))\delta'(\beta, z_i) - \mathbb{E}\left[\Psi_C(Y_i, z_i, \delta(z_i;\beta))\delta'(\beta, z_i)\right]\right)\right|$$

where recall that $\Psi_C(Y_i, z_i, C)$ is given by

$$-\lambda_n\frac{(Y_i - k_i)}{\sigma_i}\operatorname{sinc}\left(\lambda_n\left(\frac{Y_i - C}{\sigma_i}\right)\right) + \lambda_n^2\operatorname{sinc}'\left(\lambda_n\left(\frac{Y_i - C}{\sigma_i}\right)\right).$$

Throughout this proof, let $\mathcal{F}^{(1)}$ denote the function class consisting of the functions

$$f_\beta^{(1)}(y,z) = -\lambda_n \left(\frac{y-k}{\sigma}\right) \operatorname{sinc}\left(\lambda_n \left(y - \delta(z;\beta)\right)/\sigma\right) \delta'(\beta, z) \tag{C.26}$$

indexed by β, and let $\mathcal{F}^{(2)}$ denote the function class consisting of the functions

$$f_\beta^{(2)}(y,z) = \lambda_n^2 \operatorname{sinc}'\left(\lambda_n \left(y - \delta(z;\beta)\right)/\sigma\right) \delta'(\beta, z). \tag{C.27}$$

Let $\|\mathbb{G}\|_{\mathcal{F}^{(1)}} = \sup_\beta \left|\frac{1}{\sqrt{n}} \sum_{i=1}^n f_\beta^{(1)}(Y_i, z_i) - \mathbb{E} f_\beta^{(1)}(Y_i, z_i)\right|$ and define the analogous quantity $\|\mathbb{G}\|_{\mathcal{F}^{(2)}}$. Finally, let $\mathcal{F}$ denote the sum function class $\mathcal{F}^{(1)} + \mathcal{F}^{(2)}$, comprised of the functions $\Psi_C(y, z, \delta(z;\beta))\delta'(\beta, z)$.

Using a uniform bound for sinc and sinc' alongside Assumption 1(1), envelope functions for $\mathcal{F}^{(1)}, \mathcal{F}^{(2)}$ are given by

$$F^{(1)}(y) := BL\lambda_n \cdot \frac{|y-k|}{\sigma} \quad F^{(2)}(y) := BL\lambda_n^2.$$

for some absolute constant L. Propositions C.3 and C.4 bound these terms, yielding

$$\mathbb{E}\|\mathbb{G}\|_{\mathcal{F}} \le C(\log n)^{5/2} \left(\sqrt{\mathcal{V}(\mathcal{D}')} + \sqrt{\mathcal{V}(\mathcal{D})}\right).$$

Combining these propositions and the bias bound, we obtain

$$\mathbb{E} \sup_\beta |W'(\beta) - \hat{W}_n'(\beta)| \le \frac{B}{n\sqrt{\log n}} + \frac{C(\log n)^{5/2}}{\sqrt{n}} \left(\sqrt{\mathcal{V}(\mathcal{D}')} + \sqrt{\mathcal{V}(\mathcal{D})}\right)$$

for an absolute constant C. This is the first claim.

To get the second claim, we will apply Proposition D.1 with $V_i = (Y_i, z_i)$ using the envelope function

$$C\left(\lambda_n \frac{|y-k|}{\sigma} + \lambda_n^2\right).$$

The random variable $\lambda_n |Y_i - k_i|/\sigma_i + \lambda_n^2$ is easily seen to be sub-Gaussian with parameter $\lesssim \lambda_n^2 + \lambda_n |\mu_i - k_i|/\sigma_i$. This shows

$$\begin{aligned}
\left\|\|\mathbb{G}\|_{\mathcal{F}}\right\|_{\psi_2} &\lesssim \mathbb{E}\|\mathbb{G}\|_{\mathcal{F}} + \left(\frac{1}{n}\sum_{i=1}^n \left(\lambda_n^4 + \lambda_n^2 (\mu_i - k_i)^2/\sigma_i^2\right)\right)^{1/2} \\
&\lesssim \mathbb{E}\|\mathbb{G}\|_{\mathcal{F}} + \lambda_n^2 + \lambda_n \nu_2 \lesssim (\log n)^{5/2} \left(\sqrt{\mathcal{V}(\mathcal{D}')} + \sqrt{\mathcal{V}(\mathcal{D})}\right).
\end{aligned}$$

We conclude by the definition of the Orlicz norm. □

Theorem C.6. *Under the same assumptions of Theorem 3,*

$$\mathbb{E} \sup_\beta |\hat{W}_n''(\beta) - \mathbb{E}\hat{W}_n''(\beta)| \le C \frac{(\log n)^{7/2}}{\sqrt{n}}$$

for a constant C depending only on B and the VC indices of $\mathcal{D}, \mathcal{D}', \mathcal{D}''$. Moreover, fix any $\delta \in (0,1)$. Then with probability greater than $1-\delta$, we have

$$\sup_\beta |\hat{W}_n''(\beta) - \mathbb{E}\hat{W}_n''(\beta)| \leq C\frac{(\log n)^{7/2}}{\sqrt{n}}\sqrt{\log\frac{2}{\delta}}.$$

Proof of Theorem C.6. Recall that

$$\hat{W}_n''(\beta) := \frac{1}{n}\sum_{j=1}^n \Psi_{CC}(Y_j, z_j, \delta(\beta, z_j))\delta'(\beta, z_j)^2 + \Psi_C(Y_j, \beta)\delta''(\beta, z_j).$$

Let $A(\beta) := \frac{1}{\sqrt{n}}\sum_{i=1}^n \Psi_{CC}(Y_i, z_i, \delta(z_i;\beta))\delta'(\beta, z_i)^2$ and $B(\beta) := \frac{1}{\sqrt{n}}\sum_{i=1}^n \Psi_C(Y_i, z_i, \delta(z_i;\beta))\delta''(\beta, z_i)$. By the triangle inequality, it suffices to bound

$$n^{-1/2}\left(\sup_\beta |A(\beta) - \mathbb{E}A(\beta)| + \sup_\beta |B(\beta) - \mathbb{E}B(\beta)|\right)$$

in expectation and with high probability. The argument of Theorem C.5, replacing δ' with δ'', establishes the bound

$$\mathbb{E}\left[\sup_\beta |B(\beta) - \mathbb{E}B(\beta)|\right] \leq C(\log n)^{5/2} \tag{C.28}$$

for a constant C depending on the VC indices of $\mathcal{D}, \mathcal{D}''$. Therefore, we will focus on bounding $\sup_\beta |A(\beta) - \mathbb{E}A(\beta)|$.

From Lemma B.2, Ψ_{CC} is given by

$$\Psi_{CC}(Y_i, z_i, C) = \lambda_n^2\frac{(Y_i - k_i)}{\sigma_i^2}\operatorname{sinc}'\left(\lambda_n\left(\frac{Y_i - C}{\sigma_i}\right)\right) - \frac{\lambda_n^3}{\sigma_i}\operatorname{sinc}''\left(\lambda_n\left(\frac{Y_i - C}{\sigma_i}\right)\right)$$

We emulate the proofs of Theorem 1 and Theorem C.5. Throughout this proof, define the following functions classes indexed by β:

$$\mathcal{F}^{(0)} := \left\{f_\beta^{(0)}(y,z) = \lambda_n^2\frac{y-k}{\sigma^2}\operatorname{sinc}'\left(\lambda_n\left(\frac{y-\delta(z;\beta)}{\sigma}\right)\right) : \beta \in \mathbb{R}\right\} \tag{C.29}$$

$$\mathcal{F}^{(1)} := \left\{f_\beta^{(1)}(y,z) = \lambda_n^2\frac{y-k}{\sigma^2}\operatorname{sinc}'\left(\lambda_n\left(\frac{y-\delta(z;\beta)}{\sigma}\right)\right)\delta'(\beta,z)^2 : \beta \in \mathbb{R}\right\} \tag{C.30}$$

$$\mathcal{F}^{(2)} := \left\{f_\beta^{(2)}(y,z) = \frac{\lambda_n^3}{\sigma}\operatorname{sinc}''\left(\lambda_n\left(\frac{y-\delta(z;\beta)}{\sigma}\right)\right)\delta'(\beta,z)^2 : \beta \in \mathbb{R}\right\} \tag{C.31}$$

$$\mathcal{D}'^2 := \left\{\delta'(\beta,z)^2 : \beta \in \mathbb{R}\right\}, \text{ which is also VC-subgraph} \tag{C.32}$$

Let $\|\mathbb{G}\|_{\mathcal{F}^{(1)}} = \sup_\beta |\frac{1}{\sqrt{n}}\sum_{i=1}^n f_\beta^{(1)}(Y_i, z_i) - \mathbb{E}f_\beta^{(1)}(Y_i, z_i)|$ and define the analogous quantity $\|\mathbb{G}\|_{\mathcal{F}^{(2)}}$. Finally, let $\mathcal{F}^{(1)} - \mathcal{F}^{(2)}$ denote the difference function class, comprised of the

functions $\Psi_{CC}(y, z, \delta(z;\beta))\delta'(\beta, z)^2$. Using a uniform bound for sinc' and sinc'' alongside Assumption 1(1), envelope functions for $\mathcal{F}^{(1)}, \mathcal{F}^{(2)}$ are given by

$$F^{(1)} := CB^3\lambda_n^2 \cdot \frac{|y-k|}{\sigma} \tag{C.33}$$

$$F^{(2)} := CB^3\lambda_n^3. \tag{C.34}$$

for some absolute constant C. Furthermore notice that $\mathcal{D}'^2$ is a VC subgraph class, by composing Lemma D.1(5) and Lemma 2.6.20.viii of van der Vaart and Wellner (2023). To proceed, write $\sup_\beta |A(\beta) - \mathbb{E}A(\beta)| \le \|\mathbb{G}\|_{\mathcal{F}^{(1)}} + \|\mathbb{G}\|_{\mathcal{F}^{(2)}}$. Propositions C.5 and C.6 show that

$$\mathbb{E}\|\mathbb{G}\|_{\mathcal{F}^{(1)}} \le C(\log n)^{5/2} \quad \mathbb{E}\|\mathbb{G}\|_{\mathcal{F}^{(2)}} \le C(\log n)^{7/2}$$

for a constant C that depend on B and the VC indexes of $\mathcal{D}, \mathcal{D}'^2$. Combining this with Eq. (C.28) we obtain the first claim: $\mathbb{E}\sup_\beta |\hat{W}_n''(\beta) - \mathbb{E}\hat{W}_n''(\beta)| \le C\frac{(\log n)^{7/2}}{\sqrt{n}}$.

To obtain the second claim, apply Proposition D.1 with $V_i = (Y_i, z_i)$ and envelope function $F^{(1)} + F^{(2)} = CB^3\left(\lambda_n^3 + \lambda_n^2\frac{|y-k|}{\sigma}\right)$. It is not hard to see that $F^{(1)}(Y_i, z_i) + F^{(2)}(Y_i, z_i)$ is sub-Gaussian with parameter at most a constant multiple times $B^3\left(\lambda_n^3 + \lambda_n^2\frac{|\mu_i - k_i|}{\sigma_i}\right)$. Plugging this into Proposition D.1 plus the definition of the Orlicz norm yields

$$\begin{aligned}\sup_\beta |\hat{W}_n''(\beta) - \mathbb{E}\hat{W}_n''(\beta)| &\le \left(\mathbb{E}\sup_\beta |\hat{W}_n''(\beta) - \mathbb{E}\hat{W}_n''(\beta)| + C\frac{(\lambda_n^3 + \lambda_n^2\nu_2)}{\sqrt{n}}\right)\sqrt{\log\frac{2}{\delta}} \\ &\le C\frac{(\log n)^{7/2}}{\sqrt{n}}\sqrt{\log\frac{2}{\delta}}\end{aligned}$$

with probability at least $1 - \delta$. □

C.4.1. *Proofs for uniform bounds for* ASSURE* *derivatives.* The structure of the following results are all analogous to Theorem 1.

Proposition C.3. *Under the assumptions of Theorem C.5, we have*

$$\mathbb{E}\|\mathbb{G}\|_{\mathcal{F}^{(1)}} \le C(\log n)^{3/2}\left(\sqrt{\mathcal{V}(\mathcal{D}')} + \sqrt{\mathcal{V}(\mathcal{D})}\right),$$

for a constant C depending on B, where $\mathcal{F}^{(1)}$ is the class of functions defined in (C.26).

Proof. Following the proof of Theorem D.1 as before with $V_i = (Y_i, z_i)$, we may apply symmetrization and a maximal inequality to obtain

$$\mathbb{E}\left[\|\mathbb{G}\|_{\mathcal{F}^{(1)}}\right] \le \mathbb{E}\left[\int_0^{\eta_n}\sqrt{1 + \log N(\epsilon, \mathcal{F}^{(1)}, \mathbb{L}_2(\mathbb{P}_n))}d\epsilon\right], \tag{C.35}$$

with $\eta_n := \sup_\beta \left\| f_\beta^{(1)} \right\|_n$. We will bound the covering number in two steps. First, consider an ϵ-cover M_ϵ in the $\mathbb{L}^2(\mathbb{P}_n)$ metric for the function class

$$\mathcal{F}^{(0)} := \left\{ f_\beta^{(0)}(y,z) := -\lambda_n \left(\frac{y-k}{\sigma} \right) \operatorname{sinc}\left(\lambda_n \left(y - \delta(z;\beta) \right) / \sigma \right) : \beta \in \mathbb{R} \right\} \tag{C.36}$$

and a ϵ-cover in the $\mathbb{L}^4(\mathbb{P}_n)$ metric N_ϵ for the function class $\mathcal{D}' := \{\delta'(\beta, z) : \beta \in \mathbb{R}\}$. Moreover, take the cover such that all elements of M_ϵ are bounded above by the envelope $C\lambda_n \left| \frac{y-k}{\sigma} \right|$ for an absolute constant C. Take $d(z) \in N_\epsilon$ which is ϵ-close to $\delta'(\beta, z)$ and similarly, take $g(y,z) \in M_\epsilon$ which is ϵ-close to $f_\beta^{(0)}(y,z)$. By assumption, we can take the net N_ϵ such that $d(z_i) \leq B$ for all i. Then

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \left(f_\beta^{(0)}(Y_i, z_i)\delta'(\beta, z_i) - g(Y_i, z_i)d(z_i) \right)^2 \\
&\lesssim \frac{1}{n}\sum_{i=1}^n f_\beta^{(0)}(Y_i, z_i)^2 \left(\delta'(\beta, z_i) - d(z_i) \right)^2 + \frac{1}{n}\sum_{i=1}^n d(z_i)^2 \left(f_\beta^{(0)}(Y_i, z_i) - g(Y_i, z_i) \right)^2 \\
&\leq \lambda_n^2 \left(\frac{1}{n}\sum_{i=1}^n \frac{(Y_i - k_i)^4}{\sigma_i^4} \right)^{1/2} \epsilon^2 + B^2\epsilon^2
\end{aligned}$$

where we used the envelope bound and Cauchy Schwarz on the last line. Thus, we conclude the following relation between covering numbers:

$$N(C\left(\lambda_n \hat{\nu}_4 + B\right)\epsilon, \mathcal{F}^{(1)}, \mathbb{L}_2(\mathbb{P}_n)) \leq N(\epsilon, \mathcal{D}', \mathbb{L}_4(\mathbb{P}_n)) \cdot N(\epsilon, \mathcal{F}^{(0)}, \mathbb{L}_2(\mathbb{P}_n)), \tag{C.37}$$

where $\hat{\nu}_4 := \left\| \frac{y-k}{\sigma} \right\|_{\mathbb{L}^4(\mathbb{P}_n)}$. We will use this notation in other proofs. Let $\hat{\zeta}_4 := \left\| \frac{y-k}{\sigma^2} \right\|_{\mathbb{L}^4(\mathbb{P}_n)}$. By a Lipschitzness argument analogous to the one used to obtain Eq. (C.2), we also have

$$N(C\lambda_n^2 \hat{\zeta}_4 \epsilon, \mathcal{F}^{(0)}, \mathbb{L}_2(\mathbb{P}_n)) \leq N(\epsilon, \mathcal{D}, \mathbb{L}_4(\mathbb{P}_n)). \tag{C.38}$$

Putting these bounds into Eq. (C.35), we have

$$\begin{aligned}
\mathbb{E}\|\mathbb{G}\|_{\mathcal{F}^{(1)}} \lesssim\ & \mathbb{E}\left[\int_0^{\eta_n} \sqrt{1 + \log N\left(\frac{\epsilon}{C\lambda_n^2 \hat{\zeta}_4 (\lambda_n \hat{\nu}_4 + B)}, \mathcal{D}, \mathbb{L}^4(\mathbb{P}_n) \right)} d\epsilon \right] \\
&+ \mathbb{E}\left[\int_0^{\eta_n} \sqrt{1 + \log N\left(\frac{\epsilon}{C(\lambda_n \hat{\nu}_4 + B)}, \mathcal{D}', \mathbb{L}^4(\mathbb{P}_n) \right)} d\epsilon \right]
\end{aligned}$$

By change of variables and the uniform envelope function B for $\mathcal{D}, \mathcal{D}'$ we then have

$$\mathbb{E}\|\mathbb{G}\|_{\mathcal{F}^{(1)}} \lesssim \mathbb{E}\left\| F^{(1)} \right\|_{\mathbb{L}^2(\mathbb{P}_n)} + \lambda_n^3 \sqrt{\mathcal{V}(\mathcal{D})}\mathbb{E}\left[\hat{\zeta}_4 (\hat{\nu}_4 + B) \right] + \lambda_n \sqrt{\mathcal{V}(\mathcal{D}')}\mathbb{E}[\hat{\nu}_4 + B].$$

Using Cauchy Schwarz repeatedly and Assumption 1(1), it is not difficult to observe the bound $\mathbb{E}\|\mathbb{G}\|_{\mathcal{F}^{(1)}} \leq C\lambda_n^3 \left(\sqrt{\mathcal{V}(\mathcal{D}')} + \sqrt{\mathcal{V}(\mathcal{D})} \right)$ for a constant C depending on B. □

Proposition C.4. *Under the assumptions of Theorem C.5, we have*

$$\mathbb{E}\|\mathbb{G}\|_{\mathcal{F}^{(2)}} \leq C(\log n)^{5/2}\left(\sqrt{\mathcal{V}(\mathcal{D}')} + \sqrt{\mathcal{V}(\mathcal{D})}\right).$$

for a constant C depending on B, where $\mathcal{F}^{(2)}$ is the class of functions defined in (C.27).

Proof. Following the proof of Theorem D.1 as before with $V_i = (Y_i, z_i)$, we may apply symmetrization and a maximal inequality to obtain

$$\mathbb{E}\left[\|\mathbb{G}\|_{\mathcal{F}^{(2)}}\right] \leq \mathbb{E}\left[\int_0^{\eta_n} \sqrt{1 + \log N(\epsilon, \mathcal{F}^{(2)}, \mathbb{L}_2(\mathbb{P}_n))}d\epsilon\right], \tag{C.39}$$

with $\eta_n := \sup_\beta \left\|f_\beta^{(2)}\right\|_n$. We will bound the covering number in two steps. First, consider an ϵ-cover M_ϵ in the $\mathbb{L}^2(\mathbb{P}_n)$ metric for the function class

$$\mathcal{S} := \left\{s_\beta(y,z) := \lambda_n^2 \operatorname{sinc}'\left(\lambda_n\left(y - \delta(z;\beta)\right)/\sigma\right) : \beta \in \mathbb{R}\right\} \tag{C.40}$$

and a ϵ-cover in the $\mathbb{L}^2(\mathbb{P}_n)$ metric N_ϵ for the function class $\mathcal{D}' := \{\delta'(\beta, z) : \beta \in \mathbb{R}\}$. Moreover, take the cover such that all elements of M_ϵ are bounded above by the envelope $C\lambda_n^2$ for an absolute constant C. Take $d(z) \in N_\epsilon$ which is ϵ-close to $\delta'(\beta, z)$ and similarly, take $g(y, z) \in M_\epsilon$ which is ϵ-close to $s_\beta(y, z)$. By assumption, we can take the net N_ϵ such that $d(z_i) \leq B$ for all i. Then

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \left(s_\beta(Y_i, z_i)\delta'(\beta, z_i) - g(Y_i, z_i)d(z_i)\right)^2 \\
\lesssim &\frac{1}{n}\sum_{i=1}^n s_\beta(Y_i, z_i)^2\left(\delta'(\beta, z_i) - d(z_i)\right)^2 + \frac{1}{n}\sum_{i=1}^n d(z_i)^2\left(s_\beta(Y_i, z_i) - g(Y_i, z_i)\right)^2 \\
\leq &C\lambda_n^4\epsilon^2 + B^2\epsilon^2, && \text{(Envelope bound)}
\end{aligned}$$

for a constant C that only depends on B. Thus,

$$N(C\left(\lambda_n^2 + B\right)\epsilon, \mathcal{F}^{(2)}, \mathbb{L}_2(\mathbb{P}_n)) \leq N(\epsilon, \mathcal{D}', \mathbb{L}_2(\mathbb{P}_n)) \cdot N(\epsilon, \mathcal{S}, \mathbb{L}_2(\mathbb{P}_n)) \tag{C.41}$$

for some absolute constant C. Since sinc' is Lipschitz and $\sigma_i^{-1} \leq B$, we also have

$$N(B\lambda_n^3\epsilon, \mathcal{S}, \mathbb{L}_2(\mathbb{P}_n)) \leq N(\epsilon, \mathcal{D}, \mathbb{L}_2(\mathbb{P}_n)).$$

Putting these bounds into Eq. (C.39), we have

$$\begin{aligned}
\mathbb{E}\|\mathbb{G}\|_{\mathcal{F}^{(2)}} \lesssim &\mathbb{E}\left[\int_0^{\eta_n} \sqrt{1 + \log N\left(\frac{\epsilon}{C\lambda_n^3(\lambda_n^2 + B)}, \mathcal{D}, \mathbb{L}^2(\mathbb{P}_n)\right)}d\epsilon\right] \\
&+ \mathbb{E}\left[\int_0^{\eta_n} \sqrt{1 + \log N\left(\frac{\epsilon}{C(\lambda_n^2 + B)}, \mathcal{D}', \mathbb{L}^2(\mathbb{P}_n)\right)}d\epsilon\right]
\end{aligned}$$

By change of variables on the entropy integral, we have

$$\mathbb{E}\|\mathbb{G}\|_{\mathcal{F}^{(2)}} \lesssim \mathbb{E}\left\|F^{(2)}\right\|_n + \lambda_n^3(\lambda_n^2+B)\sqrt{V(\mathcal{D})} + (\lambda_n^2+B)\sqrt{V(\mathcal{D}')}, \tag{C.42}$$

Using the definition of the envelope $F^{(2)}$ and the fact that $\lambda_n \geq 1$ for n large enough,

$$\mathbb{E}\|\mathbb{G}\|_{\mathcal{F}^{(2)}} \leq C\lambda_n^5\left(\sqrt{\mathcal{V}(\mathcal{D}')} + \sqrt{\mathcal{V}(\mathcal{D})}\right),$$

which is the stated bound since $\lambda_n^5 \asymp (\log n)^{5/2}$. □

Proposition C.5. *Under the setting and assumptions of Theorem C.6, we have*

$$\mathbb{E}\|\mathbb{G}\|_{\mathcal{F}^{(1)}} \leq C(\log n)^{5/2}\left(\sqrt{\mathcal{V}(\mathcal{D}'^2)} + \sqrt{\mathcal{V}(\mathcal{D})}\right),$$

for a constant C that depends on B, where $\mathcal{F}^{(1)}$ is defined in (C.30).

Proof. Throughout this proof, C will refer to an absolute constant which may depend on B and will change from line to line. Following the proof of Theorem D.1 as before with $V_i = (Y_i, z_i)$, we may apply symmetrization and a maximal inequality to obtain

$$\mathbb{E}\left[\|\mathbb{G}\|_{\mathcal{F}^{(1)}}\right] \leq \mathbb{E}\left[\int_0^{\eta_n} \sqrt{1+\log N(\epsilon, \mathcal{F}^{(1)}, \mathbb{L}_2(\mathbb{P}_n))}d\epsilon\right], \tag{C.43}$$

with $\eta_n := \sup_\beta \left\|f_\beta^{(1)}\right\|_n$. We will bound the covering number in two steps. First, consider an ϵ-cover M_ϵ in the $\mathbb{L}^2(\mathbb{P}_n)$ metric for the function class $\mathcal{F}^{(0)}$ in (C.29) and a ϵ-cover in the $\mathbb{L}^4(\mathbb{P}_n)$ metric N_ϵ for the function class $\mathcal{D}'^2$. Moreover, take the cover such that all elements of M_ϵ are bounded above by the envelope $C\lambda_n^2\left|\frac{y-k}{\sigma^2}\right|$ for an absolute constant C. Take $d(z) \in N_\epsilon$ which is ϵ-close to $\delta'(\beta,z)^2$ and similarly, take $g(y,z) \in M_\epsilon$ which is ϵ-close to $f_\beta^{(0)}(y,z)$. By assumption, we can take the net N_ϵ such that $d(z_i) \leq B^2$ for all i. Then

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \left(f_\beta^{(0)}(Y_i,z_i)\delta'(\beta,z_i)^2 - g(Y_i,z_i)d(z_i)\right)^2 \\
&\lesssim \frac{1}{n}\sum_{i=1}^n f_\beta^{(0)}(Y_i,z_i)^2\left(\delta'(\beta,z_i)^2 - d(z_i)\right)^2 + \frac{1}{n}\sum_{i=1}^n d(z_i)^2\left(f_\beta^{(0)}(Y_i,z_i) - g(Y_i,z_i)\right)^2 \\
&\leq \lambda_n^4\left(\frac{1}{n}\sum_{i=1}^n \frac{(Y_i-k_i)^4}{\sigma_i^8}\right)^{1/2}\epsilon^2 + B^4\epsilon^2 \qquad \text{(Envelope bound and Cauchy–Schwarz)} \\
&\leq B^2\lambda_n^4\left(\frac{1}{n}\sum_{i=1}^n \frac{(Y_i-k_i)^4}{\sigma_i^4}\right)^{1/2}\epsilon^2 + B^4\epsilon^2
\end{aligned}$$

Thus, we conclude the following relation between covering numbers:

$$N(CB\left(\lambda_n^2\hat{\nu}_4 + B\right)\epsilon, \mathcal{F}^{(1)}, \mathbb{L}_2(\mathbb{P}_n)) \leq N(\epsilon, \mathcal{D}'^2, \mathbb{L}_4(\mathbb{P}_n)) \cdot N(\epsilon, \mathcal{F}^{(0)}, \mathbb{L}_2(\mathbb{P}_n)). \tag{C.44}$$

By an argument analogous to the one used to obtain Eq. (C.2), we also have

$$N(CB^2\lambda_n^3\hat{\nu}_4\epsilon, \mathcal{F}^{(0)}, \mathbb{L}_2(\mathbb{P}_n)) \leq N(\epsilon, \mathcal{D}, \mathbb{L}_4(\mathbb{P}_n)). \tag{C.45}$$

To show this, take any ϵ-cover in the $\mathbb{L}^4(\mathbb{P}_n)$ metric of the decision function space $\mathcal{D}$. Take any $\delta(\cdot;\beta) \in \mathcal{D}$, and suppose the function $\tilde{d}$ is ϵ-close to this function. Then because sinc' is Lipschitz for some constant C,

$$\begin{aligned}
&\left\| f_\beta^{(0)}(y,z) - \lambda_n^2 \frac{(y-k)}{\sigma_i^2} \mathrm{sinc}'(\lambda_n(y-\tilde{d})/\sigma) \right\|_{\mathbb{L}^2(\mathbb{P}_n)} \\
&\leq C\lambda_n^3 \left(\frac{1}{n}\sum_{i=1}^n \frac{(Y_i-k_i)^2}{\sigma_i^6} \left(\delta(z_i;\beta) - \tilde{d}\right)^2 \right)^{1/2} \\
&\leq CB^2\lambda_n^3 \left(\frac{1}{n}\sum_{i=1}^n \frac{(Y_i-k_i)^4}{\sigma_i^4} \right)^{1/4} \left(\frac{1}{n}\sum_{i=1}^n \left(\delta(z_i;\beta) - \tilde{d}\right)^4 \right)^{1/4} \leq CB^2\lambda_n^3\hat{\nu}_4\epsilon.
\end{aligned}$$

(Cauchy–Schwarz, Assumption 1(1))

Putting these bounds into (C.43), we have

$$\begin{aligned}
\mathbb{E}\|\mathbb{G}\|_{\mathcal{F}^{(1)}} \lesssim\ & \mathbb{E}\left[\int_0^{\eta_n} \sqrt{1 + \log N\left(\frac{\epsilon}{CB^3\lambda_n^3\hat{\nu}_4(\lambda_n^2\hat{\nu}_4 + B)}, \mathcal{D}, \mathbb{L}^4(\mathbb{P}_n) \right)} d\epsilon \right] \\
&+ \mathbb{E}\left[\int_0^{\eta_n} \sqrt{1 + \log N\left(\frac{\epsilon}{CB(\lambda_n^2\hat{\nu}_4 + B)}, \mathcal{D}'^2, \mathbb{L}^4(\mathbb{P}_n) \right)} d\epsilon \right].
\end{aligned}$$

By a change of variables argument and using the uniform envelope functions B, B^2 for $\mathcal{D}, \mathcal{D}'^2$ we then have

$$\mathbb{E}\|\mathbb{G}\|_{\mathcal{F}^{(1)}} \lesssim \mathbb{E}\left\|F^{(1)}\right\|_{\mathbb{L}^2(\mathbb{P}_n)} + \lambda_n^5\sqrt{\mathcal{V}(\mathcal{D})}\mathbb{E}\left[\hat{\nu}_4(\hat{\nu}_4 + B)\right] + \lambda_n^2\sqrt{\mathcal{V}(\mathcal{D}'^2)}\mathbb{E}[\hat{\nu}_4 + B],$$

where $F^{(1)}$ is the envelope function defined in (C.33). Using Cauchy Schwarz repeatedly and Assumption 1(1), it is not difficult to observe the bound

$$\mathbb{E}\|\mathbb{G}\|_{\mathcal{F}^{(1)}} \leq C\lambda_n^5\left(\sqrt{\mathcal{V}(\mathcal{D}'^2)} + \sqrt{\mathcal{V}(\mathcal{D})}\right)$$

for a constant C depending on B, as desired. □

Proposition C.6. *Under the setting and assumptions of Theorem C.6, we have*

$$\mathbb{E}\|\mathbb{G}\|_{\mathcal{F}^{(2)}} \leq C(\log n)^{7/2}\left(\sqrt{\mathcal{V}(\mathcal{D}'^2)} + \sqrt{\mathcal{V}(\mathcal{D})}\right).$$

for a constant C depending on the bound B, where $\mathcal{F}^{(2)}$ is defined in (C.31).

Proof. Throughout this proof, C will refer to an absolute constant which may depend on B and will change from line to line. As before,

$$\mathbb{E}\left[\|\mathbb{G}\|_{\mathcal{F}^{(2)}}\right] \leq \mathbb{E}\left[\int_0^{\eta_n} \sqrt{1+\log N(\epsilon, \mathcal{F}^{(2)}, \mathbb{L}_2(\mathbb{P}_n))}d\epsilon\right], \tag{C.46}$$

with $\eta_n := \sup_\beta \left\|f_\beta^{(2)}\right\|_n$. We will bound the covering number in two steps. First, consider an ϵ-cover M_ϵ in the $\mathbb{L}^2(\mathbb{P}_n)$ metric for the function class

$$\mathcal{S} := \left\{ s_\beta(y,z) := \frac{\lambda_n^3}{\sigma} \operatorname{sinc}''\left(\lambda_n\left(y-\delta(z;\beta)\right)/\sigma\right) : \beta \in \mathbb{R} \right\} \tag{C.47}$$

and a ϵ-cover in the $\mathbb{L}^2(\mathbb{P}_n)$ metric N_ϵ for the function class $\mathcal{D}'^2$. Moreover, take the cover such that all elements of M_ϵ are bounded above by the envelope $CB\lambda_n^3$ for an absolute constant C. Take $d(z) \in N_\epsilon$ which is ϵ-close to $\delta'(\beta,z)^2$ and similarly, take $g(y,z) \in M_\epsilon$ which is ϵ-close to $s_\beta(y,z)$. By Assumption 1(1), we can take the net N_ϵ such that $d(z_i) \leq B^2$ for all i. Then

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \left(s_\beta(Y_i,z_i)\delta'(\beta,z_i)^2 - g(Y_i,z_i)d(z_i)\right)^2 \\
&\lesssim \frac{1}{n}\sum_{i=1}^n s_\beta(Y_i,z_i)^2\left(\delta'(\beta,z_i)^2 - d(z_i)\right)^2 + \frac{1}{n}\sum_{i=1}^n d(z_i)^2\left(s_\beta(Y_i,z_i)-g(Y_i,z_i)\right)^2 \\
&\leq CB^2\lambda_n^6\epsilon^2 + B^4\epsilon^2,
\end{aligned}$$

where we used the envelope bound on the last line. Thus,

$$N(CB\left(\lambda_n^3+B^2\right)\epsilon, \mathcal{F}^{(2)}, \mathbb{L}_2(\mathbb{P}_n)) \leq N(\epsilon, \mathcal{D}'^2, \mathbb{L}_2(\mathbb{P}_n)) \cdot N(\epsilon, \mathcal{S}, \mathbb{L}_2(\mathbb{P}_n)) \tag{C.48}$$

for some absolute constant C. Since sinc'' is Lipschitz and $\sigma_i^{-1} \leq B$, we also have

$$N(CB^2\lambda_n^4\epsilon, \mathcal{S}, \mathbb{L}_2(\mathbb{P}_n)) \leq N(\epsilon, \mathcal{D}, \mathbb{L}_2(\mathbb{P}_n)).$$

Putting these bounds into Eq. (C.46), we have

$$\begin{aligned}
\mathbb{E}\|\mathbb{G}\|_{\mathcal{F}^{(2)}} \lesssim\ & \mathbb{E}\left[\int_0^{\eta_n} \sqrt{1+\log N\left(\frac{\epsilon}{CB^4\lambda_n^4(\lambda_n^3+B^2)}, \mathcal{D}, \mathbb{L}^2(\mathbb{P}_n)\right)}d\epsilon\right] \\
&+ \mathbb{E}\left[\int_0^{\eta_n} \sqrt{1+\log N\left(\frac{\epsilon}{CB^2(\lambda_n^3+B^2)}, \mathcal{D}'^2, \mathbb{L}^2(\mathbb{P}_n)\right)}d\epsilon\right]
\end{aligned}$$

Change of variables for the entropy integral gives as before

$$\mathbb{E}\|\mathbb{G}\|_{\mathcal{F}^{(2)}} \lesssim \mathbb{E}\left\|F^{(2)}\right\|_n + \lambda_n^4(\lambda_n^3+B^2)\sqrt{V(\mathcal{D})} + (\lambda_n^3+B^2)\sqrt{V(\mathcal{D}'^2)}. \tag{C.49}$$

Using the definition of the envelope $F^{(2)}$, we conclude the proof. □

C.5. **Auxiliary results for Theorem 4.**

Lemma C.8. *In the proof of Theorem 4, we have that Bayes regret is equal to (A.13).*

Proof. To see this, consider the case where $m_G(Y_i) > 0$ so that $a^*_G(Y_i) = 1$. Then

$$m_G(Y_i)\left(a^*_G(Y_i) - a_i(Y_{1:n})\right) = |m_G(Y_i)|\left(1 - a_i(Y_{1:n})\right) = |m_G(Y_i)|\,\mathbf{1}\left\{a^*_G(Y_i) \neq a_i(Y_{1:n})\right\}.$$

In the other case where $m_G(Y_i) < 0$ so that $a^*_G(Y_i) = 0$, then

$$\begin{aligned} m_G(Y_i)\left(a^*_G(Y_i) - a_i(Y_{1:n})\right) &= -m_G(Y_i)a_i(Y_{1:n}) = |m_G(Y_i)|\,a_i(Y_{1:n}) \\ &= |m_G(Y_i)|\,\mathbf{1}\left\{a^*_G(Y_i) \neq a_i(Y_{1:n})\right\}. \end{aligned}$$

□

Lemma C.9. *In the proof of Theorem 4, let*

$$G_0 := \mathsf{Rad}_{1/2-1/(4\sqrt{n})}, \qquad G_1 := \mathsf{Rad}_{1/2+1/(4\sqrt{n})}.$$

Let p_{G_i} denote the marginal density of a single observation under G_i, and let $p^{(-j)}_{G_i}$ denote the joint density of $(Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_n)$ under $G_i^{\otimes n}$. Then, for all sufficiently large n,

$$\alpha_{n-1} := \int_{\mathbb{R}^{n-1}} \min\left(p^{(-j)}_{G_0}(\mathbf{y}_{-j}), p^{(-j)}_{G_1}(\mathbf{y}_{-j})\right) d\mathbf{y}_{-j} \geq \alpha,$$

where $\alpha := 1 - \sqrt{2}/2 > 0$.

Proof. Since $p^{(-j)}_{G_i}$ is the $(n-1)$-fold product of p_{G_i}, we have $\alpha_{n-1} = 1 - \mathrm{TV}\left(p^{(-j)}_{G_0}, p^{(-j)}_{G_1}\right)$. By Pinsker's inequality and tensorization of the Kullback-Leibler divergence,

$$\mathrm{TV}\left(p^{(-j)}_{G_0}, p^{(-j)}_{G_1}\right) \leq \sqrt{\frac{n-1}{2}\mathrm{KL}\left(p_{G_0}, p_{G_1}\right)}.$$

By the data processing inequality,

$$\mathrm{KL}\left(p_{G_0}, p_{G_1}\right) \leq \mathrm{KL}\left(\mathrm{Bern}\left(\frac{1}{2} - \frac{1}{4\sqrt{n}}\right), \mathrm{Bern}\left(\frac{1}{2} + \frac{1}{4\sqrt{n}}\right)\right).$$

The Bernoulli KL-divergence equals $\frac{1}{2\sqrt{n}} \log \frac{\frac{1}{2}+\frac{1}{4\sqrt{n}}}{\frac{1}{2}-\frac{1}{4\sqrt{n}}}$, which is at most $1/n$ for all sufficiently large n. Therefore $\alpha_{n-1} \geq 1 - \sqrt{\frac{n-1}{2n}} \geq 1 - \frac{\sqrt{2}}{2} = \alpha$. □

Lemma C.10. *In the proof of Theorem 4, let*

$$G_0 := \mathsf{Rad}_{1/2-1/(4\sqrt{n})}, \qquad G_1 := \mathsf{Rad}_{1/2+1/(4\sqrt{n})}.$$

Then there exists an absolute constant $c > 0$ such that, for every monotone decision rule $a_i(Y_{1:n}) = \mathbf{1}\{Y_i > \delta_i(Y_{-i})\}$,

$$R(G_0, a_{1:n}) + R(G_1, a_{1:n}) \geq \frac{c}{n}$$

for all sufficiently large n.

Proof. Let c_i^* denote the zero of the posterior mean $m_{G_i}(y)$, and define

$$w(y) := \min\left(|m_{G_0}(y)|, |m_{G_1}(y)|\right), \qquad q(\mathbf{y}) := \min\left(p_{G_0}^{\otimes n}(\mathbf{y}), p_{G_1}^{\otimes n}(\mathbf{y})\right).$$

Since $c_1^* < c_0^*$ and $a_i(Y_{1:n})$ is monotone in Y_i, for each fixed Y_{-i} we have

$$\mathbf{1}\left\{a_{G_0}^*(Y_i) \neq a_i(Y_{1:n})\right\} + \mathbf{1}\left\{a_{G_1}^*(Y_i) \neq a_i(Y_{1:n})\right\} \geq \mathbf{1}\left\{c_1^* < Y_i < c_0^*\right\}.$$

Hence, by (A.13),

$$\begin{aligned} R(G_0, a_{1:n}) + R(G_1, a_{1:n}) &\geq \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}^n} w(y_i)\mathbf{1}\left\{c_1^* < y_i < c_0^*\right\} q(\mathbf{y})\, d\mathbf{y} \\ &\geq \alpha_{n-1}\int_{c_1^*}^{c_0^*} w(y)\min\left(p_{G_0}(y), p_{G_1}(y)\right)\, dy, \end{aligned}$$

where α_{n-1} is defined in Lemma C.9. By Lemma C.9, $\alpha_{n-1} \geq \alpha := 1 - \sqrt{2}/2$ for all sufficiently large n.

Next, write $p_- := \frac{1}{2} - \frac{1}{4\sqrt{n}}$ and $p_+ := \frac{1}{2} + \frac{1}{4\sqrt{n}}$. The posterior mean under Rad_p is

$$m_{\mathsf{Rad}_p}(y) = \frac{pe^{2y} - (1-p)}{pe^{2y} + (1-p)} = \tanh\left(y - c_p^*\right), \qquad c_p^* := \frac{1}{2}\log\frac{1-p}{p}.$$

Therefore, with $c_n := c_{p_-}^* = -c_{p_+}^*$, we have

$$c_0^* = c_n, \qquad c_1^* = -c_n, \qquad c_n = \frac{1}{2}\log\frac{\frac{1}{2} + \frac{1}{4\sqrt{n}}}{\frac{1}{2} - \frac{1}{4\sqrt{n}}}.$$

Since $c_n = \operatorname{arctanh}\left(\frac{1}{2\sqrt{n}}\right)$, we have $c_n \geq \frac{1}{2\sqrt{n}}$.

For $y \in [-c_n, c_n]$, $|m_{G_0}(y)| = \tanh(c_n - y)$, $|m_{G_1}(y)| = \tanh(y + c_n)$. For all sufficiently large n we have $2c_n \leq 1$, and $\tanh(t) \geq t/2$ for $t \in [0, 1]$. Hence, on $[-c_n, c_n]$, $w(y) \geq \frac{1}{2}\min(c_n - y, y + c_n) = \frac{1}{2}(c_n - |y|)$. Moreover, for all sufficiently large n the interval $[-c_n, c_n]$ is contained in $[-1, 1]$. Since both p_- and p_+ lie in $[1/4, 3/4]$,

$$\min\left(p_{G_0}(y), p_{G_1}(y)\right) \geq \frac{1}{4}\varphi(2) \qquad \text{for all } y \in [-c_n, c_n].$$

Combining the preceding displays, we obtain

$$\begin{aligned} R(G_0, a_{1:n}) + R(G_1, a_{1:n}) &\geq \alpha\frac{\varphi(2)}{4}\int_{-c_n}^{c_n} w(y)\, dy \geq \alpha\frac{\varphi(2)}{8}\int_{-c_n}^{c_n}(c_n - |y|)\, dy = \alpha\frac{\varphi(2)}{8}c_n^2 \\ &\geq \alpha\frac{\varphi(2)}{32}\frac{1}{n}. \end{aligned}$$

This proves the claim. □

C.6. **Auxiliary results for Theorem 5.** Throughout this section, we fix $\delta_1, \dots, \delta_n$, set $h = 1/\sqrt{2\log n}$, and let $Z_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ independently. We record the three ingredients used in the proof of Theorem 5.

Lemma C.11. *For*

$$F_x(s;\delta) := \frac{x}{2} + \frac{x}{\pi}\,\mathrm{Si}\left(\frac{x-\delta}{sh}\right) - \frac{s}{h}\,\mathrm{sinc}\left(\frac{x-\delta}{sh}\right),$$

the map $s \mapsto F_x(s;\delta)$ is $C(1/h + |x|/\epsilon)$-Lipschitz on $[\epsilon, \infty)$, uniformly in δ.

Proof. Write $y := (x-\delta)/h$. Then $F_x(s;\delta) = \frac{x}{2} + \frac{x}{\pi}\,\mathrm{Si}\left(\frac{y}{s}\right) - \frac{s^2}{h}\frac{\sin(y/s)}{\pi y}$. Differentiating in s gives

$$\frac{d}{ds}F_x(s;\delta) = -\frac{x}{\pi s}\sin\left(\frac{x-\delta}{sh}\right) - \frac{2}{h}\,\mathrm{sinc}\left(\frac{y}{s}\right) + \frac{1}{\pi h}\cos\left(\frac{y}{s}\right).$$

Using the uniform boundedness of $\sin$, $\cos$, and sinc, we obtain

$$\left|\frac{d}{ds}F_x(s;\delta)\right| \le C\left(\frac{1}{h} + \frac{|x|}{s}\right) \le C\left(\frac{1}{h} + \frac{|x|}{\epsilon}\right)$$

on $[\epsilon, \infty)$. The mean value theorem completes the proof. □

Lemma C.12. *Under the assumptions of Theorem 5,*

$$\frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}\left[\hat{w}_{h,i}(\sqrt{m}\bar{Y}_i;\delta_i) - w_{h,i}(\sqrt{m}\bar{Y}_i;\delta_i)\right]\right| \le C\frac{\sqrt{\log n}}{\sqrt{m}}.$$

Proof. Fix i and write $\delta = \delta_i$. Since

$$\hat{w}_{h,i}(x;\delta) = F_x(\hat{\sigma}_i;\delta), \qquad w_{h,i}(x;\delta) = F_x(\sigma_i;\delta),$$

Lemma C.11 controls the sensitivity to variance estimation. Let $A_i := \left\{\hat{\sigma}_i > \sigma_i/\sqrt{2}\right\}$. By Chebyshev's inequality,

$$\mathbb{P}(A_i^c) \le \mathbb{P}\left(\left|\hat{\sigma}_i^2 - \sigma_i^2\right| > \frac{\sigma_i^2}{2}\right) \le \frac{4\,\mathrm{Var}(\hat{\sigma}_i^2)}{\sigma_i^4}.$$

On A_i, Lipschitzness gives

$$\left|\hat{w}_{h,i}(\sqrt{m}\bar{Y}_i;\delta) - w_{h,i}(\sqrt{m}\bar{Y}_i;\delta)\right| \le C\left(\frac{1}{h} + \frac{|\sqrt{m}\bar{Y}_i|}{\sigma_i}\right)|\hat{\sigma}_i - \sigma_i|.$$

Also, using the boundedness of Si and sinc,

$$|\hat{w}_{h,i}(x;\delta) - w_{h,i}(x;\delta)|\,\mathbf{1}(A_i^c) \le C\left(\frac{1}{h} + |x|\right)\mathbf{1}(A_i^c).$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\left|\hat{w}_{h,i}(\sqrt{m}\bar{Y}_i;\delta) - w_{h,i}(\sqrt{m}\bar{Y}_i;\delta)\right|\right] \\
&\le C\mathbb{E}\left[\left(\frac{1}{h} + \frac{|\sqrt{m}\bar{Y}_i|}{\sigma_i}\right)|\hat{\sigma}_i - \sigma_i|\right] + C\mathbb{E}\left[\left(\frac{1}{h} + |\sqrt{m}\bar{Y}_i|\right)\mathbf{1}(A_i^c)\right] \\
&\le C\left(\frac{1}{h}\sqrt{\mathbb{E}(\hat{\sigma}_i - \sigma_i)^2} + \frac{\sqrt{m}}{\sigma_i}\sqrt{\mathbb{E}(\bar{Y}_i^2)\mathbb{E}(\hat{\sigma}_i - \sigma_i)^2}\right) \\
&\quad + C\left(\frac{1}{h}\mathbb{P}(A_i^c) + \sqrt{m}\sqrt{\mathbb{E}(\bar{Y}_i^2)\mathbb{P}(A_i^c)}\right).
\end{aligned}$$

Now $\mathbb{E}(\hat{\sigma}_i - \sigma_i)^2 \le \frac{\mathrm{Var}(\hat{\sigma}_i^2)}{\sigma_i^2}$ and $\mathbb{E}(\bar{Y}_i^2) = \frac{\mu_i^2+\sigma_i^2}{m}$, and $\mathrm{Var}(\hat{\sigma}_i^2) \lesssim \mu_{4,i}/m$. Substituting these bounds yields

$$\mathbb{E}\left[\left|\hat{w}_{h,i}(\sqrt{m}\bar{Y}_i;\delta) - w_{h,i}(\sqrt{m}\bar{Y}_i;\delta)\right|\right] \lesssim \frac{\mu_{4,i}^{1/2}}{\sqrt{m}}\left(\frac{1}{\sigma_i h} + \frac{\sqrt{\mu_i^2+\sigma_i^2}}{\sigma_i^2}\right) + \frac{\mu_{4,i}}{m\sigma_i^4 h}.$$

Since $h = 1/\sqrt{2\log n}$, the right-hand side is bounded by $C_i\sqrt{\log n}/\sqrt{m}$ for a polynomial C_i in $|\mu_i|, \sigma_i, \mu_{4,i}$. Averaging over i and absorbing the average polynomial into the constant C proves the claim. □

Lemma C.13. *Under the assumptions of Theorem 5,*

$$\frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}\left[w_{h,i}(\sqrt{m}\bar{Y}_i;\delta_i) - w_{h,i}(Z_i;\delta_i)\right]\right| \le C\frac{(\log n)^2}{\sqrt{m}}.$$

Proof. Fix i and write $\delta = \delta_i$. Relabel the micro-data Y_{ij} as $Y_1, \dots, Y_m$, and let $Z_1, \dots, Z_m$ be i.i.d. $\mathcal{N}(\mu_i/\sqrt{m}, \sigma_i^2)$. Define

$$S_k := \frac{1}{\sqrt{m}}\left(Z_1 + \cdots + Z_k + Y_{k+1} + \cdots + Y_m\right),$$

for $k = 0, \dots, m$, and

$$S_k' := \frac{1}{\sqrt{m}}\left(Z_1 + \cdots + Z_k + Y_{k+2} + \cdots + Y_m\right),$$

for $k = 0, \dots, m-1$. Then $S_0 = \sqrt{m}\bar{Y}_i$, S_m has the same distribution as Z_i, and

$$S_k = S_k' + \frac{Y_{k+1}}{\sqrt{m}}, \qquad S_{k+1} = S_k' + \frac{Z_{k+1}}{\sqrt{m}}.$$

Taylor-expanding $x \mapsto w_{h,i}(x;\delta)$ around S_k' to third order, and using that Y_{k+1} and Z_{k+1} are independent of S_k' and have matching first two moments, we obtain

$$|\mathbb{E}w_{h,i}(S_{k+1};\delta) - \mathbb{E}w_{h,i}(S_k;\delta)| \lesssim \frac{1}{m^{3/2}}\left(\mathbb{E}\left[|Y_{k+1}|^3|w_{h,i}^{(3)}(\tilde{x})|\right] + \mathbb{E}\left[|Z_{k+1}|^3|w_{h,i}^{(3)}(\tilde{x}')|\right]\right),$$

where $\tilde{x}$ lies between S_k and S_k' and $\tilde{x}'$ lies between S_{k+1} and S_k'.

Since $k_i = 0$ under the assumptions of Theorem 5, direct differentiation gives

$$w_{h,i}^{(3)}(x;\delta) = \frac{3}{(\sigma_i h)^2}\operatorname{sinc}'\left(\frac{x-\delta}{\sigma_i h}\right) + \frac{x}{(\sigma_i h)^3}\operatorname{sinc}''\left(\frac{x-\delta}{\sigma_i h}\right) - \frac{1}{\sigma_i^2 h^4}\operatorname{sinc}'''\left(\frac{x-\delta}{\sigma_i h}\right),$$

so the uniform boundedness of the derivatives of sinc yields

$$|w_{h,i}^{(3)}(x;\delta)| \lesssim \frac{1}{\sigma_i^2 h^4} + \frac{|x|}{\sigma_i^3 h^3}.$$

Also,

$$|\tilde{x}| \le \frac{|Y_{k+1}|}{\sqrt{m}} + |S_k'|, \qquad |\tilde{x}'| \le \frac{|Z_{k+1}|}{\sqrt{m}} + |S_k'|, \qquad \mathbb{E}|S_k'| \le \sqrt{\mathbb{E}[(S_k')^2]} \lesssim \sqrt{\mu_i^2 + \sigma_i^2}.$$

Finally,

$$\mathbb{E}|Y_{k+1}|^3 \lesssim \frac{|\mu_i|^3}{m^{3/2}} + \mu_{3,i}, \qquad \mathbb{E}[Y_{k+1}^4] \lesssim \frac{\mu_i^4}{m^2} + \mu_{4,i},$$

and the Gaussian analogues satisfy

$$\mathbb{E}|Z_{k+1}|^3 \lesssim \frac{|\mu_i|^3}{m^{3/2}} + \sigma_i^3, \qquad \mathbb{E}[Z_{k+1}^4] \lesssim \frac{\mu_i^4}{m^2} + \sigma_i^4.$$

Substituting these moment bounds into the previous display gives

$$|\mathbb{E}w_{h,i}(S_{k+1};\delta) - \mathbb{E}w_{h,i}(S_k;\delta)| \lesssim \frac{P_i}{m^{3/2}}\left(h^{-4} + h^{-3}\right),$$

where P_i is a polynomial in $|\mu_i|, \sigma_i, \mu_{3,i}, \mu_{4,i}, \sigma_i^{-1}$. Summing over $k = 0, \dots, m-1$ yields

$$\left|\mathbb{E}w_{h,i}(Z_i;\delta_i) - \mathbb{E}w_{h,i}(\sqrt{m}\bar{Y}_i;\delta_i)\right| \lesssim \frac{P_i}{\sqrt{m}}\left(h^{-4} + h^{-3}\right).$$

Since $h = 1/\sqrt{2\log n}$, this is bounded by $C_i(\log n)^2/\sqrt{m}$ for another polynomial C_i in $|\mu_i|, \sigma_i, \mu_{3,i}, \mu_{4,i}$. Averaging over i proves the claim. □

Lemma C.14. *Under the assumptions of Theorem 5,*

$$\left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[w_{h,i}(Z_i;\delta_i)\right] - W(\delta)\right| \le C\left(\frac{1}{n\log n} + \frac{1}{\sqrt{m}}\right).$$

Proof. Write

$$W_G(\delta) := \frac{1}{n}\sum_{i=1}^n \mu_i \bar{\Phi}\left(\frac{\delta_i - \mu_i}{\sigma_i}\right).$$

Then

$$\left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[w_{h,i}(Z_i;\delta_i)\right] - W(\delta)\right| \le \left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[w_{h,i}(Z_i;\delta_i)\right] - W_G(\delta)\right| + |W_G(\delta) - W(\delta)|.$$

For the first term, Proposition 1(1) with $k = 0$ gives

$$\left| \mathbb{E}\left[w_{h,i}(Z_i;\delta_i)\right] - \mu_i \bar{\Phi}\left(\frac{\delta_i - \mu_i}{\sigma_i}\right) \right| \leq |\mu_i| h^2 e^{-1/(2h^2)}.$$

Averaging over i and using $h = 1/\sqrt{2\log n}$ yields

$$\left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[w_{h,i}(Z_i;\delta_i)\right] - W_G(\delta) \right| \lesssim \frac{1}{n\log n}.$$

For the second term, Berry–Esseen gives $\sup_x \left|F_i^{(m)}(x) - \Phi(x)\right| \leq \frac{\mu_{3,i}}{\sigma_i^3\sqrt{m}}$, and therefore

$$|W_G(\delta) - W(\delta)| \leq \left(\frac{1}{n}\sum_{i=1}^{n} |\mu_i| \sup_x \left|\bar{F}_i^{(m)}(x) - \bar{\Phi}(x)\right|\right) \leq \left(\frac{1}{n}\sum_{i=1}^{n} \frac{|\mu_i|\mu_{3,i}}{\sigma_i^3}\right) \frac{1}{\sqrt{m}}.$$

Absorbing the average polynomial into the constant C proves the claim. □

Appendix D. Empirical processes for i.n.i.d. data

The next lemma records how the VC index changes under composition for the results in Lemma 2.6.20 of van der Vaart and Wellner (2023), alongside some additional useful results. Throughout this section, let $\psi_2 = e^{x^2} - 1$.

Lemma D.1. *Let $\mathcal{F}$ be a VC subgraph class with index $V(\mathcal{F})$ and let $g : \mathcal{X} \to \mathbb{R}$, $\phi : \mathcal{X} \to \mathcal{Y}$. Then*

(1) $g + \mathcal{F} = \{g + f : f \in \mathcal{F}\}$ is a VC subgraph class with VC index equal to $V(\mathcal{F})$.

(2) $g\mathcal{F} = \{gf : f \in \mathcal{F}\}$ is a VC subgraph class with VC index $\leq 2V(\mathcal{F}) + 1$.

(3) Let $f(x)$ be a VC subgraph class with index $V(\mathcal{F})$. Then the class of functions $\mathbf{1}\{f(x) > 0\}$ is also VC subgraph with index at most $V(\mathcal{F})$.

(4) Let $\mathcal{G}$ be the class of functions $g(x, y) : X \times Y \to \mathbb{R}$ such that $g(x, y) = f(x)$ for some $f \in \mathcal{F}$. Then $\mathcal{G}$ is VC subgraph with index equal to $V(\mathcal{F})$.

(5) $|\mathcal{F}| = \{|f| : f \in \mathcal{F}\}$ is a VC subgraph class.

Proof. Items (1) through (3) follow from inspecting the proofs of Lemma 2.6.20 of van der Vaart and Wellner (2023) parts i, vi, viii in conjunction with Problem 2.6.10. For item (4), this follows immediately by noting that the subgraphs of $\mathcal{G}$ shatter a set $\{((Y_i, y_i), t_i)\}_{i=1}^n$ if and only if the subgraphs of $\mathcal{F}$ shatter $\{(Y_i, t_i)\}_{i=1}^n$. For item (5), notice that the subgraph of $|f|$ is the union of the subgraphs of $f, -f$. Combining Lemma 2.6.20 parts ii and iv completes the proof. □

Lemma D.2 (VC function covering, Theorem 2.6.7 of van der Vaart and Wellner (2023))**.** *For a VC-subgraph class of functions with measurable envelope function F and $r \geq 1$ we*

have for any probability measure Q that

$$N(\epsilon\|F\|_{Q,r}, \mathcal{F}, L_r(Q)) \leq KV(\mathcal{F})(16e)^{V(\mathcal{F})}\left(\frac{1}{\epsilon}\right)^{rV(\mathcal{F})}$$

for a universal constant K and $0 < \epsilon < 1$.

Following van der Vaart and Wellner (2023) (p.330), for $0 < \delta \leq 1$, $r \geq 1$, a function class $\mathcal{F}$ with envelope F, and a probability measure Q, write

$$\|F\|_{Q,r}^r := \int |F|^r \, dQ, \qquad J(\delta, \mathcal{F} \mid F, L_r) := \sup_Q \int_0^\delta \sqrt{1 + \log N\left(\epsilon\|F\|_{Q,r}, \mathcal{F}, L_r(Q)\right)} \, d\epsilon,$$

where the supremum is over all discrete probability measures Q with $\|F\|_{Q,r} > 0$.

Write $\mathbb{G}_n$ to be the i.n.i.d. empirical process

$$\mathbb{G}_n f = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(f(V_i) - P_i f\right),$$

where $V_1, \ldots, V_n$ are independent random elements taking values in a common measurable space and $P_i f := \mathbb{E} f(V_i)$. The next theorem records a uniform entropy bound from van der Vaart and Wellner (2023) adapted to this setting. In the applications below, we will take $V_i = (Y_i, z_i)$ or $V_i = (Y_i, \mu_i, z_i)$, depending on the function class.

Theorem D.1 (Theorem 2.14.1 of van der Vaart and Wellner (2023))**.** *Let $V_1, \ldots, V_n$ be independent random elements taking values in a common measurable space, and let $\mathcal{F}$ be a measurable class of functions on this space with measurable envelope function F. Then for $p \geq 1$,*

$$\left\|\|\mathbb{G}_n\|_{\mathcal{F}}^*\right\|_p \lesssim \left\|J(\eta_n, \mathcal{F} \mid F, L_2)\|F\|_n\right\|_p \lesssim J(1, \mathcal{F} \mid F, L_2)\left(\frac{1}{n}\sum_{i=1}^n \|F(V_i)\|_{2\vee p}^2\right)^{1/2}$$

where $\|g\|_n^2 = \frac{1}{n}\sum_{i=1}^n g(V_i)^2$ is the $\mathbb{L}^2(\mathbb{P}_n)$-seminorm and $\eta_n = \|\|f\|_n\|_{\mathcal{F}}/\|F\|_n$.

Proof. The proof is verbatim from van der Vaart and Wellner (2023). Denote by $\mathbb{G}_n^o$ the symmetrized process given by $\mathbb{G}_n^o f := \frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(V_i)$. The same argument as Theorem 2.14.1 of van der Vaart and Wellner (2023) shows that conditional on $V_1, \ldots, V_n$, the sub-Gaussian norm of the symmetrized process satisfies

$$\|\|\mathbb{G}_n^o\|_{\mathcal{F}}\|_{\psi_2|V} \lesssim \int_0^{\eta_n} \sqrt{1 + \log N(\epsilon, \mathcal{F}, L_2(\mathbb{P}_n))} d\epsilon$$

where $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(V_i)$ and

$$\eta_n := \|\|f\|_n\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \sqrt{\frac{1}{n}\sum_{i=1}^n f(V_i)^2}.$$

A change of variable in the integral by $\epsilon \leftarrow \epsilon \|F\|_n$ gives

$$\begin{aligned}\left\| \|\mathbb{G}_n^o\|_{\mathcal{F}} \right\|_{\psi_2|V} &\lesssim \|F\|_n \int_0^{\eta_n/\|F\|_n} \sqrt{1 + \log N(\epsilon \|F\|_n, \mathcal{F}, L_2(\mathbb{P}_n))} d\epsilon \\ &\leq \|F\|_n J(\eta_n, \mathcal{F} \mid F, L_2)\end{aligned}$$

The rest of the proof follows as is. □

Theorem D.2 (Localized i.n.i.d. maximal inequality, Theorem 2.14.2 of van der Vaart and Wellner (2023))**.** *Let $V_1, \ldots, V_n$ be independent random elements taking values in a common measurable space, and let $\mathcal{F}$ be a measurable class of functions with measurable envelope F. Write*

$$\bar{P} f := \frac{1}{n} \sum_{i=1}^n P_i f, \qquad \|F\|_{\bar{P},2} := (\bar{P} F^2)^{1/2}.$$

Suppose that $F \leq 1$ and that, for some $\delta \in (0,1)$, $\sup_{f \in \mathcal{F}} \bar{P} f^2 \leq \delta^2 \bar{P} F^2$. If $\|F\|_{\bar{P},2} > 0$, then

$$\mathbb{E} \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J(\delta, \mathcal{F} \mid F, L_2) \left(1 + \frac{J(\delta, \mathcal{F} \mid F, L_2)}{\delta^2 \sqrt{n} \, \|F\|_{\bar{P},2}}\right) \|F\|_{\bar{P},2}.$$

If $\|F\|_{\bar{P},2} = 0$, then $\|\mathbb{G}_n\|_{\mathcal{F}} = 0$ almost surely.

Proof. Let $S := \mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}}$, $\sigma_n^2 := \sup_{f \in \mathcal{F}} \mathbb{P}_n f^2$, $\tau_n^2 := \mathbb{P}_n F^2$. The same symmetrization and conditional maximal inequality used in Theorem D.1 gives

$$S \lesssim \mathbb{E}\left[J\left(\frac{\sigma_n}{\tau_n}, \mathcal{F} \mid F, L_2\right) \tau_n \right],$$

with the convention that the integrand is zero when $\tau_n = 0$. Since $t \mapsto J(t, \mathcal{F} \mid F, L_2)$ is concave, Jensen's inequality gives

$$S \lesssim J\left(\frac{\sqrt{\mathbb{E}\sigma_n^2}}{\|F\|_{\bar{P},2}}, \mathcal{F} \mid F, L_2\right) \|F\|_{\bar{P},2}.$$

Also,

$$\mathbb{E}\sigma_n^2 \leq \sup_{f \in \mathcal{F}} \bar{P} f^2 + n^{-1/2} \mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}^2} \leq \delta^2 \|F\|_{\bar{P},2}^2 + C n^{-1/2} S,$$

where $\mathcal{F}^2 := \{f^2 : f \in \mathcal{F}\}$ and the last inequality follows from the contraction inequality, because $x \mapsto x^2$ is 2-Lipschitz on $[-1,1]$. Let $z := \sqrt{\mathbb{E}\sigma_n^2}/\|F\|_{\bar{P},2}$. If $z \leq \delta$, the claimed bound is immediate. If $z > \delta$, concavity implies

$$J(z, \mathcal{F} \mid F, L_2) \leq \frac{z}{\delta} J(\delta, \mathcal{F} \mid F, L_2).$$

Combining the preceding displays yields

$$z^2 \le \delta^2 + C\frac{J(\delta, \mathcal{F} \mid F, L_2)}{\delta\sqrt{n}\,\|F\|_{\bar{P},2}} z \implies z \le \delta + C\frac{J(\delta, \mathcal{F} \mid F, L_2)}{\delta\sqrt{n}\,\|F\|_{\bar{P},2}}. \qquad (x^2 \le a^2 + bx \implies x \le a + b)$$

Substituting this into the bound for S proves the theorem. □

The previous theorem is most useful when the envelope F is bounded by some other constant M, which may depend on n. The next corollary records this.

Corollary D.1. *Under the same setting as Theorem D.2, consider a function class $\mathcal{F}$ with envelope F uniformly bounded by a constant M. Then,*

$$\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J(\delta, \mathcal{F} \mid F, L_2)\left(1 + \frac{MJ(\delta, \mathcal{F} \mid F, L_2)}{\delta^2\sqrt{n}\,\|F\|_{\bar{P},2}}\right)\|F\|_{\bar{P},2}.$$

where the implicit constant in $\lesssim$ does not depend on M.

Proof. Apply Theorem D.2 to the class of functions $\tilde{\mathcal{F}} := \{f/M : f \in \mathcal{F}\}$. Define $\tilde{F} := F/M$ as the envelope function for $\tilde{\mathcal{F}}$. We obtain

$$\mathbb{E}\|\mathbb{G}_n\|_{\tilde{\mathcal{F}}} \lesssim J(\delta, \tilde{\mathcal{F}} \mid \tilde{F}, L_2)\left(1 + \frac{J(\delta, \tilde{\mathcal{F}} \mid \tilde{F}, L_2)}{\delta^2\sqrt{n}\left\|\tilde{F}\right\|_{\bar{P},2}}\right)\left\|\tilde{F}\right\|_{\bar{P},2}.$$

First, note covering numbers remain unchanged after scaling. For any probability measure Q,

$$N\left(\epsilon\left\|\tilde{F}\right\|_{Q,r}, \tilde{\mathcal{F}}, L_r(Q)\right) = N\left(\epsilon\|F\|_{Q,r}, \mathcal{F}, L_r(Q)\right).$$

Thus, $\mathbb{E}\|\mathbb{G}_n\|_{\tilde{\mathcal{F}}} \lesssim J(\delta, \mathcal{F} \mid F, L_2)\left(1 + \frac{MJ(\delta, \mathcal{F}|F, L_2)}{\delta^2\sqrt{n}\,\|F\|_{\bar{P},2}}\right)\frac{\|F\|_{\bar{P},2}}{M}$. Multiplying by M gives the result. □

Proposition D.1 (Orlicz norm bounds, i.n.i.d. analogue of Theorem 2.14.23 van der Vaart and Wellner (2023))**.** *Let $V_1, \ldots, V_n$ be independent random elements taking values in a common measurable space, and let $\mathcal{F}$ be a class of measurable functions on this space with measurable envelope function F. Then*

$$\left\|\|\mathbb{G}_n\|^*_{\mathcal{F}}\right\|_{\psi_2} \lesssim \mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}} + \left(\frac{1}{n}\sum_{i=1}^{n}\|F(V_i)\|^2_{\psi_2}\right)^{1/2}$$

Proof. We recall Proposition A.1.6 of van der Vaart and Wellner (2023). Let $U_1, \ldots, U_n$ be independent mean zero stochastic processes indexed by an arbitrary set T. Set $S_n = \sum_{i=1}^n U_i$.

Then we have

$$\left\| \|S_n\|^* \right\|_{\psi_2} \leq K \left[\left\| \|S_n\|^* \right\|_1 + \left(\sum_{i=1}^n \left\| \|U_i\|^* \right\|_{\psi_2}^2 \right)^{1/2} \right],$$

where $\|X\|^*$ is shorthand for $\sup_t |X_t|$. Take $U_i(f) = \frac{1}{\sqrt{n}} (f(V_i) - \mathbb{E}f(V_i))$, indexed by $f \in \mathcal{F}$ so that $\|S_n\|^* = \|\mathbb{G}_n\|_{\mathcal{F}}$. Plugging this into the proposition, we obtain

$$\left\| \|\mathbb{G}_n\|_{\mathcal{F}}^* \right\|_{\psi_2} \lesssim E^* \|\mathbb{G}_n\|_{\mathcal{F}} + n^{-1/2} \left(\sum_{i=1}^n \left\| \sup_{f \in \mathcal{F}} |f(V_i) - \mathbb{E}[f(V_i)]| \right\|_{\psi_2}^2 \right)^{1/2} \tag{D.1}$$

$$\leq E^* \|\mathbb{G}_n\|_{\mathcal{F}} + n^{-1/2} \left(\sum_{i=1}^n \|F(V_i) + \mathbb{E}[F(V_i)]\|_{\psi_2}^2 \right)^{1/2} \tag{D.2}$$

Where we used the fact that if $|f| \leq |g|$ then $\||f(X)|\|_{\psi_2} \leq \||g(X)|\|_{\psi_2}$ for any random variable. Using the triangle inequality for the Orlicz norm and the fact that $\|X\|_{\psi_2} \geq \mathbb{E}|X|$ and $\|C\|_{\psi_2} \leq C/\sqrt{\ln 2}$ for any constant C, we can bound

$$\begin{aligned} \|F(V_i) + \mathbb{E}[F(V_i)]\|_{\psi_2} &\leq \|F(V_i)\|_{\psi_2} + \|\mathbb{E}[F(V_i)]\|_{\psi_2} \\ &\leq \|F(V_i)\|_{\psi_2} + \mathbb{E}[F(V_i)]/\sqrt{\ln 2} \lesssim \|F(V_i)\|_{\psi_2}. \end{aligned}$$

This gives a final bound by a constant multiple of $\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}} + n^{-1/2} \left(\sum_{i=1}^n \|F(V_i)\|_{\psi_2}^2 \right)^{1/2}$, from which the desired claim holds. □

D.1. **Examples of VC classes.**

Lemma D.3. *For fixed finite p, the decision rule classes in (1)–(3) in Example 1 are VC-subgraph classes.*

Proof. For the first class, $\mathcal{D}_{\text{threshold}} = \{z \mapsto k(z) + \beta : \beta \in [-M, M]\} \subset k + \{\beta \cdot 1 : \beta \in \mathbb{R}\}$, where $1(z) \equiv 1$. The class $\{\beta \cdot 1 : \beta \in \mathbb{R}\}$ is a one-dimensional vector space of measurable functions, hence VC-subgraph by Lemma 2.6.16 of van der Vaart and Wellner (2023). By Lemma D.1(1), adding the fixed function k preserves the VC-subgraph property, so $\mathcal{D}_{\text{threshold}}$ is VC-subgraph. For the second class, $\mathcal{D}_{t\text{-stat}} = \{z \mapsto k(z) + \beta\sigma(z) : \beta \in [-M, M]\} \subset k + \{\beta\sigma : \beta \in \mathbb{R}\}$. The class $\{\beta\sigma : \beta \in \mathbb{R}\}$ is again a one-dimensional vector space of measurable functions, so it is VC-subgraph by the same lemma. Another application of Lemma D.1(1) shows that $\mathcal{D}_{t\text{-stat}}$ is VC-subgraph. For the third class, let $\mathcal{V} := \operatorname{span}\left\{\delta^{(1)}, \ldots, \delta^{(p)}\right\}$. Because $p < \infty$, $\mathcal{V}$ is a finite-dimensional vector space of measurable functions, hence VC-subgraph by Lemma 2.6.16 of van der Vaart and Wellner (2023). Moreover, $\mathcal{D}_{\text{finite}} = \operatorname{conv}\left(\left\{\delta^{(1)}, \ldots, \delta^{(p)}\right\}\right) \subset \mathcal{V}$. Any subclass of a VC-subgraph class is again VC-subgraph, so $\mathcal{D}_{\text{finite}}$ is VC-subgraph as well. □

Proposition D.2 (VC-subgraph properties in Example 1)**.** *The decision threshold classes in Table 1 are all VC-subgraph.*

Proof of Proposition D.2. Let us consider the first claim. Notice that the linear shrinkage threshold rules are contained in the function class

$$d_{a,b}(k, \sigma^2) := k + a\sigma^2 k + b\sigma^2$$

for $a, b \in \mathbb{R}$. This is the fixed function $(k, \sigma^2) \mapsto k$ plus a two-dimensional vector space of functions. Combining Lemma 2.6.16 of van der Vaart and Wellner (2023) and Lemma D.1(1) establishes that this is VC subgraph. For the second claim, notice that Fay–Herriot decisions are contained in the function class

$$d_{b_0,\beta}(k, \sigma^2, x) := k + b_0\sigma^2 k - \sigma^2 x'\beta$$

indexed by $(b_0, \beta) \in \mathbb{R}^{1+p}$. The same argument as before establishes the claim.

Finally for the third claim, we will use the results of Onshuus and Quiroz (2015), which discuss a connection between VC dimension and the theory of o-minimality from model theory, a subfield of mathematical logic. See also Aschenbrenner et al. (2013) and Van den Dries (1998) for textbook references on this topic. Consider the decision class in parametric CLOSE-GAUSS as functions of the cost k and log standard deviation ℓ:

$$f_\beta(k, \ell) = k + e^{2\ell} e^{-b_1 - b_2\ell}(k - a_1 - a_2\ell)$$

indexed by $\beta = (a_1, a_2, b_1, b_2)$. The subgraph sets

$$S_\beta := \left\{(k, \ell, t) : t < k + e^{2\ell} e^{-b_1 - b_2\ell}(k - a_1 - a_2\ell)\right\}$$

are defined by a first-order formula in the structure $(\mathbb{R}, +, \cdot, 0, 1, <, e^x)$, which is o-minimal: see Fact 2 of Onshuus and Quiroz (2015) or Wilkie (1996). Theorem 2 of Onshuus and Quiroz (2015) implies that for any set of points $\{x_1, \ldots, x_n\}$ in $\mathbb{R}^3$, the collection of sets S_β pick out at most $O(n^4)$ subsets of $\{x_1, \ldots, x_n\}$. Thus the function class $\{f_\beta(k, \ell)\}_\beta$ is VC subgraph. □

Appendix E. Additional details for Section 6.1

We start from the replication code in Chen (2026) and follow Chen (2026) to generate the raw data and covariates for each target variable (mean rank and top-20 probability), for the largest 20 Commuting Zones in the US. Both Opportunity Atlas (OA, Chetty et al. (2026)) outcomes used in the paper retain $n = 10{,}056$ tracts.

Cost calibration and target shares. Let $\tau \in \{10, 33\}$ denote the target top share. For each outcome and each $\tau \in \{10, 33\}$, we first fit the plug-in Fay–Herriot rule to the OA data and denote the resulting posterior means by $\hat{m}_i^{\mathrm{FH}}$. The constant cost used in that scenario is then

$$k^{(\tau)} := \text{the } \left\lceil \frac{\tau}{100} n \right\rceil \text{th largest element of } \left\{ \hat{m}_i^{\mathrm{FH}} \right\}_{i=1}^{n}.$$

Equivalently, $k^{(\tau)}$ is the weakest Fay–Herriot posterior mean among the tracts in the calibrated top $\tau\%$. We then set $k_i = k^{(\tau)}$ for every tract in that scenario.

For completeness, Figure E.1 reports the coupled-bootstrap comparisons for the stricter top-10% calibration. The broad ranking of methods is similar to the main-text top-third exercise.

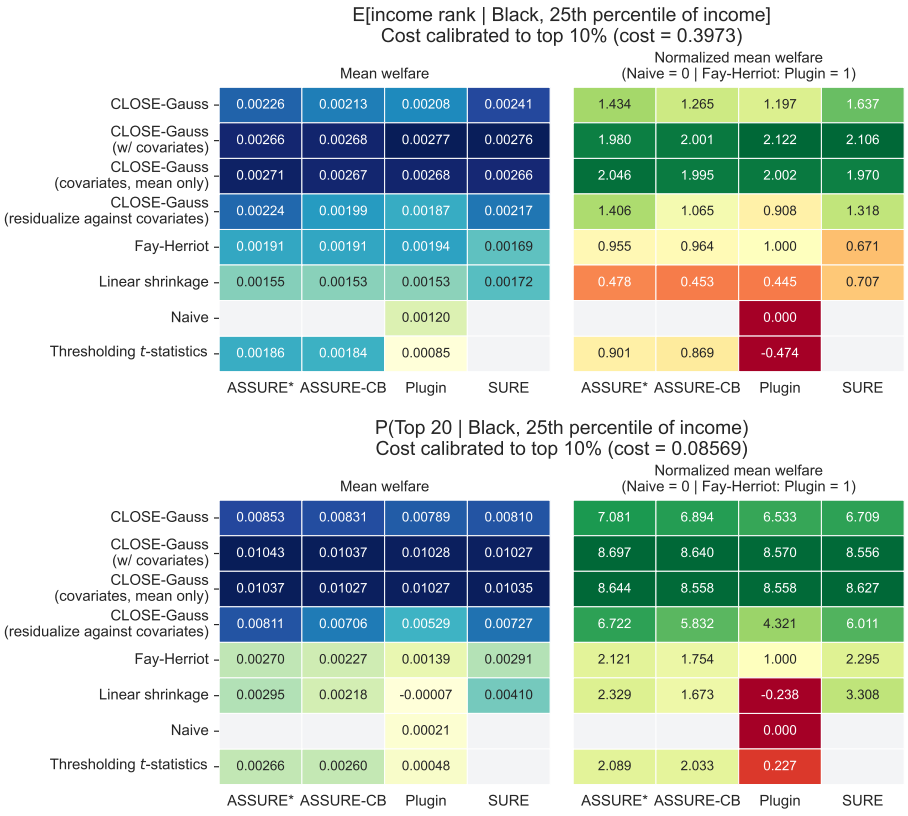


FIGURE E.1. Analogue of Figure 5 calibrated to selecting top 10%.

Notes. This is the analogue of Figure 5, but with costs calibrated to the weakest Fay–Herriot posterior mean among the top 10% of tracts rather than the top third. Results are averaged over 1000 coupled-bootstrap draws. □

Coupled bootstrap. Fix one outcome and one target share τ. Starting from the observed OA estimates $\{(Y_i, \sigma_i, x_i)\}_{i=1}^n$, one Monte Carlo draw generates i.i.d. perturbations $Q_i \sim \mathcal{N}(0,1)$ and forms $Y_i^{\text{tr}} = Y_i + \sqrt{w}\,\sigma_i Q_i, Y_i^{\text{val}} = Y_i - \sqrt{\frac{1}{w}}\,\sigma_i Q_i$, with $\sigma_i^{\text{tr}} = \sigma_i\sqrt{1+w}$ and $\sigma_i^{\text{val}} = \sigma_i\sqrt{1+\frac{1}{w}}$. The default implementation uses $w = 1/9$, which mimics a $90/10$ train/validation split. Any decision rule fitting only sees the training sample $\left\{(Y_i^{\text{tr}}, \sigma_i^{\text{tr}}, x_i, k^{(\tau)})\right\}$; the validation sample is used only for evaluation.

On draw b of the coupled bootstrap, if a fitted rule returns thresholds $\delta_i^{(b)}$ on draw b, the reported validation welfare is $\hat{W}_b^{\text{val}} := \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left(Y_i^{\text{tr}} > \delta_i^{(b)}\right)\left(Y_i^{\text{val}} - k^{(\tau)}\right)$.

Decision rule classes. The coupled-bootstrap OA exercise compares rules in Example 1. The naive rule is simply $\mathbf{1}(Y_i > k^{(\tau)})$. The t-statistic class uses thresholds of the form $\delta_i = k^{(\tau)} + \beta\sigma_i$. The linear-shrinkage class is the constant-mean, constant-variance Gaussian-prior class from Table 1, $\delta_i = k^{(\tau)} + \frac{\sigma_i^2}{\tau^2}\left(k^{(\tau)} - \mu_0\right)$. Under constant costs, threshold-based criteria depend on (μ_0, τ^2) only through the normalized slope $\beta = \frac{k^{(\tau)} - \mu_0}{\tau^2}$, and $\delta_i = k^{(\tau)} + \beta\sigma_i^2$. Fay–Herriot uses the conjugate-normal threshold implied by a prior mean $m_{0i} = x_i'\beta$ and a constant signal variance A.

Finally, CLOSE-GAUSS class uses $\delta_i = k_i + \frac{\sigma_i^2}{s_{0i}^2}(k_i - m_{0i})$, where the prior mean m_{0i} and prior variance s_{0i}^2 are modeled as functions of tract characteristics. In the plain CLOSE-GAUSS specification, m_{0i} and $\log s_{0i}^2$ are both linear in $(1, \log\sigma_i)$. In the covariate version they are linear in $(1, \log\sigma_i, x_i)$. In the "mean-only" version the prior mean is linear in $(1, \log\sigma_i, x_i)$ but the signal variance is constrained to be constant across tracts. In the "prefit" version we first regress Y_i on x_i, replace Y_i and k_i by residualized counterparts, fit the CLOSE-GAUSS rule using only $(1, \log\sigma_i)$ on those residuals, and then add the fitted covariate component back when interpreting thresholds in the original outcome scale.

Plug-in fitting. For each decision class, we may fit via plug-in:

- *Naive rule.* There is no statistical fit: the plug-in rule is simply $\delta_i = k^{(\tau)}$.
- *Thresholding t-statistics.* There is again no estimated nuisance parameter. The plug-in benchmark fixes $\delta_i = k^{(\tau)} + 1.96\,\sigma_i$.
- *Linear shrinkage.* Plug-in estimates the prior mean by the sample average

$$\hat{\mu}_{0,\text{plugin}} = \bar{Y}, \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i,$$

and the signal variance by the average residual variance proxy

$$\hat{\tau}^2_{\text{plugin}} = \max\left(\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2 - \frac{1}{n}\sum_{i=1}^n \sigma_i^2,\ \epsilon\right),$$

truncating at a small ϵ.

• *Fay–Herriot.* Plug-in first estimates β by ordinary least squares in the regression of Y_i on x_i. It then forms the residual variance proxy $\hat{V}_i = (Y_i - x_i'\hat{\beta})^2 - \sigma_i^2$ and sets the signal variance equal to the cross-tract average of $\hat{V}_i$, truncated below at a numerically negligible positive floor so that the fitted variance remains strictly positive: $\hat{A}_{\text{plugin}} = \max\left(\frac{1}{n}\sum_{i=1}^n \hat{V}_i,\ \epsilon\right)$.

• CLOSE-GAUSS *with covariates, mean only.* This is the same plug-in construction as Fay–Herriot, but with the enlarged design $z_i = (1, \log\sigma_i, x_i')'$ in place of x_i. Thus $\hat{m}_{0i} = z_i'\hat{\beta}$ and $\hat{s}_{0i}^2 = \hat{A}_{\text{plugin}}$, where $\hat{\beta}$ is the least-squares coefficient on z_i.

• CLOSE-GAUSS *and* CLOSE-GAUSS *with covariates.* These classes use $\delta_i = k_i + \frac{\sigma_i^2}{s_{0i}^2}(k_i - m_{0i})$, with m_{0i} linear in z_i and variance that is exponential-linear, up to a small variance floor $\underline{s}^2$: $s_{0i}^2 = \underline{s}^2 + \exp(z_i'\gamma)$. Plug-in first estimates β, $\hat{\beta}_{\text{plugin}} \in \arg\min_\beta \frac{1}{n}\sum_{i=1}^n (Y_i - z_i'\beta)^2$. Writing $\hat{m}_{0i} = z_i'\hat{\beta}_{\text{plugin}}$, it then forms

$$\hat{V}_i = (Y_i - \hat{m}_{0i})^2 - \sigma_i^2, \qquad \underline{s}^2 = \frac{1}{10}\max\left(\frac{1}{n}\sum_{i=1}^n \hat{V}_i,\ 10^{-10}\right).$$

The variance parameter γ is then chosen by nonlinear least squares:

$$\hat{\gamma}_{\text{plugin}} \in \arg\min_\gamma \frac{1}{n}\sum_{i=1}^n \left(\hat{V}_i - \underline{s}^2 - \exp(z_i'\gamma)\right)^2.$$

There are two additional truncation choices to prevent dividing by zero or over/underflow. First, the exponential index $z_i'\gamma$ is clipped to the interval $[-10, 10]$ before exponentiation, which rules out extremely small or large fitted signal variances. Second, the fitted prior mean can be clipped to the observed range of the estimation sample, $[\min_i Y_i, \max_i Y_i]$; the coupled-bootstrap OA runs leave this clipping turned off by default.

• *Prefit* CLOSE-GAUSS. In this class the data are first residualized on $(1, x_i')$:

$$R_i = Y_i - \hat{g}(x_i), \qquad k_i^R = k^{(\tau)} - \hat{g}(x_i),$$

where $\hat{g}$ is the least-squares projection of Y_i on $(1, x_i')$. Plug-in then applies the same procedure as in the plain CLOSE-GAUSS class, but to (R_i, σ_i, k_i^R) with design $(1, \log\sigma_i)$.

Methods for fitting decision classes. We state the fitting problem separately for each rule class. For a generic threshold family $\delta_i(\theta)$, ASSURE* estimates (2.4). The ASSURE-CB fitter chooses $\hat{\theta}_{\text{cb}} \in \text{argmax}_\theta \frac{1}{n}\sum_{i=1}^n (Y_i - k_i)\Phi\left(\frac{Y_i - \delta_i(\theta)}{\epsilon_n\sigma_i}\right) - \frac{\sigma_i}{\epsilon_n}\phi\left(\frac{Y_i - \delta_i(\theta)}{\epsilon_n\sigma_i}\right)$, with $\epsilon_n = n^{-0.2}$. For the Gaussian-prior classes, SURE minimizes

$$\text{SURE}(\theta) := \frac{1}{n}\sum_{i=1}^n \left(\hat{\mu}_i(\theta) - Y_i\right)^2 + \frac{2}{n}\sum_{i=1}^n \sigma_i^2 w_i(\theta),$$

where $\hat{\mu}_i(\theta) = w_i(\theta)Y_i + (1 - w_i(\theta))\, m_{0i}(\theta)$ and $w_i(\theta) = \frac{s_{0i}^2(\theta)}{s_{0i}^2(\theta) + \sigma_i^2}$. Some notable implementation choices include:

- CLOSE-GAUSS. Let $z_i = (1, \log \sigma_i)'$. For some small floor $\underline{s}^2$, this class uses

$$m_{0i}(\beta) = z_i'\beta, \qquad s_{0i}^2(\gamma) = \underline{s}^2 + \exp\left(z_i'\gamma\right),$$

so the threshold rule is $\delta_i(\beta, \gamma) = k^{(\tau)} + \frac{\sigma_i^2}{\underline{s}^2 + \exp\left(z_i'\gamma\right)} \left(k^{(\tau)} - z_i'\beta\right)$. Plug-in estimates (β, γ) by least squares on the mean side and nonlinear least squares on the residual variance proxy; ASSURE*, ASSURE-CB, and SURE then tune the same parameter vector.

- CLOSE-GAUSS *with covariates.* This class enlarges the previous design to $z_i = (1, \log \sigma_i, x_i')'$,
- *Prefit* CLOSE-GAUSS. In this variant we first regress Y_i on $(1, x_i')$ to obtain fitted values $\hat{g}(x_i)$ and residuals $R_i = Y_i - \hat{g}(x_i), k_i^R = k^{(\tau)} - \hat{g}(x_i)$. We then fit the plain CLOSE-GAUSS class to $R_i \sim \mathcal{N}\left(\mu_i - \hat{g}(x_i), \sigma_i^2\right)$ using only $(1, \log \sigma_i)$ in the Gaussian-prior model:

$$\delta_i^R(\beta, \gamma) = k_i^R + \frac{\sigma_i^2}{\underline{s}^2 + \exp\left(\gamma_0 + \gamma_1 \log \sigma_i\right)} \left(k_i^R - (\beta_0 + \beta_1 \log \sigma_i)\right).$$

The threshold in the original outcome scale is then $\delta_i(\beta, \gamma) = \hat{g}(x_i) + \delta_i^R(\beta, \gamma)$. Plug-in, ASSURE*, ASSURE-CB, and SURE are all applied to this residualized class.

Redundant variance intercept under constant costs when fitting ASSURE. Under constant costs, the thresholds on the ratio $(k_i - m_{0i})/s_{0i}^2$, so the intercept in the variance parametrization for s_{0i}^2 is redundant.[25] In the reported implementation, the optimizer fixes that intercept at its plug-in value and optimizes the remaining coordinates only.

Warm starts and optimizer fallback. The reported Gaussian-prior fits also use the following convention throughout. By default, ASSURE* is initialized at the ASSURE-CB optimum rather than at the raw plug-in fit. The ASSURE-CB fitter falls back to the plug-in fit whenever the optimized coupled-bootstrap objective is worse than the plug-in objective, so the fitter never accepts a numerically inferior point purely because the optimizer moved there.

Optimization heuristics. For the scalar t-statistic rule, both ASSURE* and ASSURE-CB are tuned by deterministic one-dimensional grid search rather than by generic nonlinear optimization.

For the Gaussian-prior classes, ASSURE*, ASSURE-CB, and SURE are all optimized multiple times, with added Gaussian noise, around the plug-in fitted parameters. In the OA analysis the default is three restarts for both ASSURE* and ASSURE-CB.

[25] This is exactly true if the small floor $\underline{s}^2$ were zero.