Compound Selection Decisions: An Almost SURE Approach

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

Abstract. We propose methods for *compound selection decisions* in a Gaussian sequence model. Given unknown, fixed parameters $\mu_{1:n}$, known $\sigma_{1:n}$, observations $Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, and known costs k_i, the decision maker chooses thresholds δ_i to maximize $\frac{1}{n}\sum_{i=1}^n \mathbf{1}(Y_i > \delta_i)(\mu_i - k_i)$. Inspired by Stein's unbiased risk estimate (SURE), we introduce a family of estimators, which we term ASSURE, for the expected utility of a set of proposed thresholds. ASSURE enables practitioners to select thresholds within a pre-specified class by maximizing an estimated objective, borrowing strength across noisy estimates in compound settings. A leading member of this family, ASSURE*, is nearly unbiased for the true expected utility and converges at near-parametric rates. We show that ASSURE* has favorable regret properties. We apply ASSURE to the selection of Census tracts for economic opportunity, the identification of discriminating firms, and the analysis of p-value decision procedures in A/B testing.

Date: March 26, 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, Princeton, Stanford, UC Davis, and UC Santa Cruz for helpful comments. Jiafeng Chen is grateful for the support of 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 was used to check the paper for consistency and clarity.

1. Introduction

In settings with many parallel noisy estimates, researchers often aim to *select* units with large underlying parameters. Many such problems can be modeled as follows: For parallel settings $i = 1, \ldots, n$, a researcher observes a noisy estimate Y_i, a standard error σ_i, some covariate information 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 $Y_i > \delta_i$.

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 total net payoff of the selection decisions, normalized by n:

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

The statistical challenge is to choose $\delta_{1:n}$ from the data so as to maximize expected utility. The selection rule therefore depends not only on the observed contextual information (σ_i, x_i, k_i) but also on the noisy signal Y_i. In particular, μ_i may exhibit heterogeneity that is not fully explained by the observed covariates alone, making Y_i useful.[2]

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

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

[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] This type of problem appears 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.*, 2018; 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).

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*.

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, as a randomly selected unit from the sample is likely to have $\mu_i < k_i$—a fundamental insight of Robbins (1951/1985).

Existing approaches for solving compound selection problems are either indirect for (1.1)—choosing selection decisions based, e.g., on optimal mean-squared error predictions for μ_i or on hypothesis testing (Manski, 2019)—or rely on plausibly 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 are analogues of James–Stein estimators and likely improve mean-squared error for μ_i relative to the raw estimates Y_i, but it is not obvious that better prediction implies 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 generates optimal selection decisions. However, this model is very plausibly misspecified (Chen, forthcoming; Bonhomme and Denis, 2024; Müller and Watson, 2024).

This paper develops robust methods both for evaluating and for generating 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 decision rule:

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

We show that $\hat{W}$ converges to W for fixed $\mu_{1:n}$ and $z_{1:n}$, integrating only over the randomness in Y_i under (1.2). Thus, the validity of evaluations does not depend on correctly specifying any model for the unknown parameters.

Second, this welfare estimator enables researchers to generate selection decisions by optimizing within a user-chosen class of decision rules. So long as our estimate of welfare 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 parameters.

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 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). Since SURE targets mean-squared error prediction, it is insufficient 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. 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 $F_h(y-\delta)$ is a smooth approximation of $\mathbf{1}(y > \delta)$. We propose a class of estimators, which we term ASSURE, for **A**lmos**t** **SURE**:

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

Like SURE, ASSURE follows from Stein's identity (Stein, 1981). Unlike SURE, an exactly unbiased estimator of welfare does not exist due to the non-smoothness of $\mathbf{1}(Y > \delta)$. This necessitates smoothing through $F_h(y-\delta)$.

Different smoothing kernels produce different bias properties for the welfare estimator. We focus on two choices. First, our leading choice of $\mathsf{k}(\cdot)$ is the sinc kernel $\mathsf{k}(x) := \frac{\sin x}{\pi x}$. 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, forthcoming; Ignatiadis and Sun, 2025).[3] We thus call this estimator ASSURE-CB. Because the

[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 (forthcoming) uses this *coupled bootstrap* idea to estimate (1.1) for selection decisions that depend on Y_{i1}.

Gaussian kernel is only a second-order kernel, it does not fully exploit the smoothness of the estimand, so ASSURE-CB converges more slowly than ASSURE*.

Evaluation of decision rules naturally leads to choosing decisions that optimize estimated welfare, among a user-chosen 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, motivated by empirical Bayes (EB) posterior means under a correlated random effects model in which $\mu_i \mid \sigma_i, k_i, x_i \sim \mathcal{N}(m(\sigma_i, k_i, x_i; \beta), s^2(\sigma_i, k_i, x_i; \beta))$ is

$$\mathcal{D}_{\text{CLOSE-GAUSS}} = \{\delta_i : \mathbf{1}(y > \delta_i) = \mathbf{1}\left\{a(\sigma_i, x_i, k_i; \beta)y + b(\sigma_i, x_i, k_i; \beta) > k_i\right\}\},$$

for functions $a(\cdot;\beta), b(\cdot;\beta)$ parametrized by $\beta \in \mathbb{R}^d$. This class is proposed by Weinstein *et al.* (2018) and called CLOSE-GAUSS by Chen (forthcoming); it nests screening based on James–Stein estimators and various versions of 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 $\max_i \sigma_i^2 = \Omega(1)$) and ASSURE* converges to true welfare. 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. Both the upper and lower bounds for regret are novel analyses. 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 tuning β in $\mathcal{D}_{\text{CLOSE-GAUSS}}$.

These favorable properties rely on the Gaussian distributional assumption (1.2), which, though standard, ignores that the central limit theorem only provides an 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 means are within some $1/\sqrt{m}$ neighborhood. We find that the performance of ASSURE* degrades gracefully in the quality of the Gaussian approximation. 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 as $m \to \infty$. 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 (Jiang, 2020; Gu and Koenker, 2023; Walters, 2024; Kline *et al.*, 2024; Chen, forthcoming). 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 by simply 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 is predicted by the contextual information (Chen, forthcoming) or how μ_i correlates with μ_j (Bonhomme and Denis, 2024)—could harm the selection decision. The full joint distribution of $\mu_{1:n}$ conditional on all covariates is exceedingly difficult to model and estimate flexibly; feasible empirical Bayes models are thus usually at risk of misspecification. A similar robustness concern motivates the fixed effects panel literature (Dano *et al.*, 2025), relative to correlated random effects.

ASSURE offers a robust way to use empirical Bayes models as working approximations rather than as maintained assumptions. In practice, one could choose $\mathcal{D}$ as selection decision rules implied by a (parametric) empirical Bayes model.[4] Optimizing within this class using ASSURE weakly improves on the empirical Bayes decisions under an estimated prior $\hat{P}$. When the model for P is well-specified, ASSURE-based decisions do little harm. When such a model is badly misspecified, ASSURE-based decisions nevertheless choose optimal decision rules within the implied class. If the class includes simple decision rules like $\delta_i = k_i$, selection via ASSURE nevertheless assures practitioners that decisions are no worse than the simple benchmarks. We present Monte Carlo exercises in Sections 2.3 and 5 and three empirical applications in Section 6 that illustrate this point.

[4]Empirical Bayesians solve selection problems by thresholding on the posterior mean $\mathbb{E}_P[\mu_i \mid Y_i, x_i, \sigma_i] > k_i$. Provided the posterior mean is monotone in Y_i (which is true for Gaussian Y), this can be inverted to a threshold decision of the form $\mathbf{1}(Y_i > \delta(x_i, \sigma_i; P))$. When P is parametric, this defines a class of decision rules $\mathcal{D}$.

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 and an extension of our framework to Poisson observations.

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. 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 $[n] := \{1, \dots, n\}$.

2. Setup and Methodology

Let $Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2), i \in [n]$ with known σ_i, known costs k_i, with potentially related 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.[5] 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.

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

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

This section introduces estimators $\hat{W}_n(\beta)$ for $W(\beta)$ and studies their estimation error $\hat{W}_n - W$. Such an estimator immediately enables the evaluation of a decision parametrized by a fixed β. Naturally, an estimator $\hat{W}_n$ with small estimation error is valuable for evaluating decision rules.

Moreover, because $\hat{W}_n$ provides an empirical objective, practitioners can also choose β by optimizing $\hat{W}_n(\cdot)$:

$$\hat{\beta} \in \operatorname*{argmax}_{\beta \in \mathcal{B}} \hat{W}_n(\beta). \tag{2.3}$$

[5] If $(Y_1, \dots, Y_n) \sim \mathcal{N}((\mu_1, \dots, \mu_n)', \Sigma)$ 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; **?**)

Accurate estimation of W is also important for the estimated decision rule $\hat{\beta}$ to perform well. Using bounds on $\hat{W}_n - W$, our subsequent theoretical results control the difference, or *regret*, between the welfare of the best decision rule in $\mathcal{D}$ and the expected utility achieved by $\hat{\beta}$:

$$\mathsf{Regret}_n = \sup_{\beta \in \mathcal{B}} W(\beta) - \mathbb{E}_{\mu_{1:n}}[u(\hat{\beta})]. \tag{2.4}$$

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

2.1. **The ASSURE estimators for welfare.** We now motivate the functional form in (1.4) through Stein's identity. Let $F(\cdot)$ be differentiable with derivative $f(\cdot)$ and fix some $k, \delta, \mu \in \mathbb{R}$ and $\sigma^2 > 0$. Simple manipulations of Stein's identity reveal that $(Y-k)F(Y-\delta) - \sigma^2 f(Y-\delta)$ is an unbiased estimator for $(\mu - k)\mathbb{E}[F(Y-\delta)]$:

$$\mathbb{E}[(Y-k)F(Y-\delta) - \sigma^2 f(Y-\delta)] = (\mu-k)\mathbb{E}[F(Y-\delta)]. \tag{2.5}$$

On the other hand, with (2.2), estimating $W(\beta)$ reduces to estimation of the parameter

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

Accordingly, we seek a differentiable function 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 $\sigma^2 \frac{d}{dy} F_h\left(\frac{y-\delta}{\sigma}\right) = \sigma \mathsf{k}_h\left(\frac{y-\delta}{\sigma}\right)$

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). 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, for fixed β, ASSURE* achieves squared bias and variance of order $\tilde{O}(1/n)$.

Proposition 1 (ASSURE* has low bias)**.** *Fix $\mu, k \in \mathbb{R}$. Let $\delta(z_i;\beta) = \beta$,*[6] *then*

(1) Let $Y \sim \mathcal{N}(\mu, \sigma^2)$. The bias of ASSURE* *is exponentially decreasing in $1/h$: For all $h > 0$,*

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

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

(2) For $h = 1/\sqrt{2\log n}$, the pointwise squared error of ASSURE* *is $O(\log^2 n/n)$:*

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

Figure 1 illustrates ASSURE*'s rapidly vanishing bias as a function of $h := 1/\sqrt{2\log n}$.

Unlike SURE, ASSURE* and the broader ASSURE family are not exactly unbiased for welfare, and 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).

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

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

[6] For simplicity, we consider $\delta(z_i;\beta) = \beta$. Our regret results rely on analogous statements for general δ.

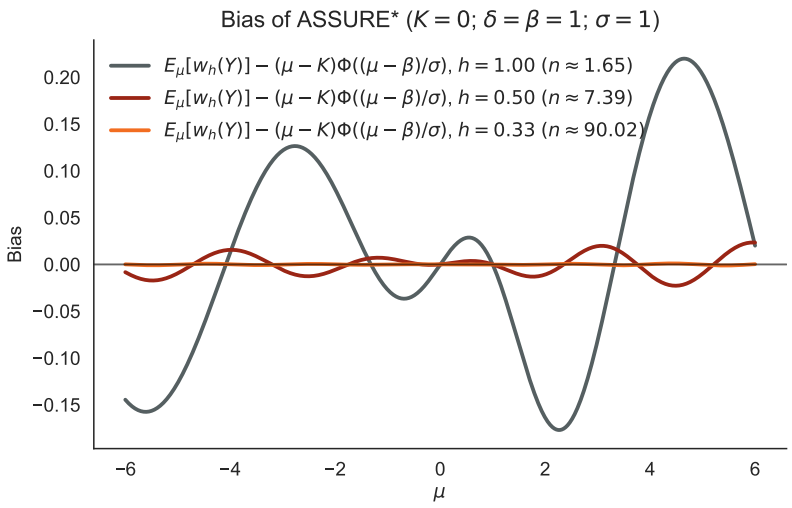


FIGURE 1. Illustration for the bias of ASSURE*.

whereas 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:[7] By taking the Fourier transform of both sides, we would like a kernel for which

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

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 $\hat{\varphi}(\cdot)$ has Gaussian tails, this truncation leaves it essentially unchanged—allowing the bias to decay exponentially in $1/h$.

2.2. **Comparison with 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, forthcoming). For $Y \sim \mathcal{N}(\mu, \sigma^2)$, we can construct two independent samples by adding and subtracting an independent Gaussian noise Q

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

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

This *coupled bootstrap* procedure is a synthetic version of sample-splitting.[8] 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 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*.[9]

2.3. **Discussion.** We now contextualize our approach in the literature and compare it against a few alternatives. We also disambiguate our regret notion (2.4) from the regret notions in these literatures. We illustrate these comparisons using a classic example from Robbins (1951/1985).

Example 1. 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 be the class of $\delta_i(\beta) = \beta$ in the decision problem (1.1). 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 1 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 Kurisu, 2022), where

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

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

[9] ASSURE* is $\tilde{O}(n^{-1/2})$-consistent for welfare, whereas ASSURE-CB is only $\tilde{O}(n^{-2/5})$-consistent. Appendix D provides a more formal comparison.

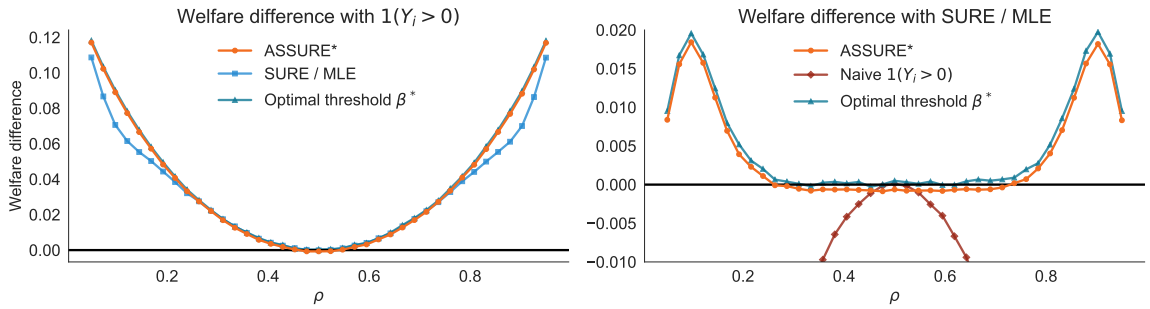


FIGURE 2. Welfare comparisons of different methods for fitting decision rules for Example 1, as a function of the proportion ρ of $\mu_i = +1$.

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. □

By contrast, our goal is asymptotically vanishing regret in (2.4), uniformly over suitable sequences $(\mu_i, z_i)_{i=1}^n$. Specializing to Example 1, (2.4) is instead

$$\mathsf{Regret}_n = \rho\Phi\left(\frac{1}{2}\log\frac{\rho}{1-\rho} + 1\right) - (1-\rho)\Phi\left(\frac{1}{2}\log\frac{\rho}{1-\rho} - 1\right) - \mathbb{E}[u(\hat{\beta})].$$

These desiderata are distinct. Our later results show that ASSURE* ensures vanishing Regret_n, but its worst-case $\mathsf{FreqRegret}$ for any finite n may not be optimal. Conversely, the empirical success rule $\mathbf{1}(Y_i > 0)$ achieves minimax-$\mathsf{FreqRegret}$ (Proposition 1, Stoye, 2012), but its regret (2.4) 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 1: 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: If its estimate is positive, this provides evidence that the μ_i's are positive on average. If we are confident that $\frac{1}{n}\sum_{i=1}^n (\mu_i - k_i) > 0$, less conservative decision rules such as $\mathbf{1}\{Y_i \geq -\epsilon\}$ can improve the compound welfare, since any mildly negative realizations of Y_i are then more likely due to chance.

This is especially stark in Example 1, 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* exploits this information through estimating the objective function.

Empirical Bayes. 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: $\mathbb{E}_P[\mu_i \mid Y_i, z_i] > k_i$. However, this additional structure hinges on modeling and estimating the distribution of μ_i well, which can make subsequent procedures and guarantees less robust.[10]

Our procedure is complementary to empirical Bayes. EB-style modeling for μ_i is powerful at generating classes of decision rules $\mathcal{D}$ that are reasonable—indeed optimal if the model on μ_i happens to be correct. Choosing a member of $\mathcal{D}$ with ASSURE provides guarantees for the compound selection problem directly—thus not requiring researchers 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.''

As an example decision class, following Chen (forthcoming), we use CLOSE-GAUSS to refer to a family of models modeling μ_i as correlated Gaussian random effects $\mu_i \mid z_i \sim \mathcal{N}(m_0(z_i;\beta), s_0^2(z_i;\beta))$, where m_0, s_0 are flexibly parametrized by β.[11] This motivates using ASSURE to choose among the following threshold decision rules, derived from inverting $\mathbb{E}[\mu_i \mid Y_i, z_i] > k_i$ in Y_i :

$$\mathcal{D}_{\text{CLOSE-GAUSS}} = \left\{ \delta(z_i;\beta) = k_i + \frac{\sigma_i^2}{s_0^2(z_i;\beta)}(k_i - m_0(z_i;\beta)) : \beta \in \mathcal{B} \right\}. \tag{2.11}$$

Formally, empirical Bayes studies a different regret criterion that averages over draws of $\mu_{1:n}$ and benchmarks against the oracle Bayes rule (Jiang, 2020; Soloff *et al.*, 2024; Chen, forthcoming):

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

[10] A similar concern—about the robustness of correlated random effect models—motivates the literature of fixed effect nonlinear panels (Dano *et al.*, 2025).

[11] This prior is first analyzed in the empirical Bayes literature by Weinstein *et al.* (2018). It is also closely related to correlated random-effects models in panel data (Chamberlain, 1984).

where $u(\hat{P}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(\mathbb{E}_{\hat{P}}[\mu_i \mid Y_i, z_i] > k_i)(\mu_i - k_i)$. Relative to the regret (2.4), $\mathsf{EBRegret}_n$ (i) additionally integrates over $\mu_{1:n} \mid z_{1:n} \sim P$, (ii) chooses the benchmark as $\mathbf{1}(\mathbb{E}_P[\mu_i \mid Y_i, z_i] > k_i)$, and (iii) typically considers the class of posterior decision rules indexed by an estimate of P. When the class of decision rules $\mathcal{D}$ is exactly the class of posterior thresholding decisions, controlling (2.4) automatically controls $\mathsf{EBRegret}_n$:

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

When the prior is misspecified, however, that benchmark can be misleading for the fixed-parameter compound problem, which is exactly where ASSURE* can improve on EB. To illustrate, in Example 1 and Figure 2, we consider using a misspecified normal prior where $\mu_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(m_0, s_0^2)$. The implied posterior mean, given by $\mathbb{E}[\mu_i \mid Y_i] = \frac{s_0^2}{s_0^2+1}Y_i + \frac{m_0}{s_0^2+1}$, is therefore also misspecified. An empirical Bayesian may estimate (m_0, s_0^2) by maximum likelihood, which, interestingly, in Example 1, yields the same selection decisions as choosing (m_0, s_0^2) by SURE. The misspecification of the prior *misaligns* optimal selection and estimation decisions (Manski, 2021), causing EB-MLE/SURE to underperform ASSURE*.

3. Theoretical assurances

We now turn to statistical guarantees on ASSURE*. Our main results are upper and matching lower bounds for regret (2.4). 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, define the norms $\|D\|_{p,n}^p := \frac{1}{n}\sum_{i=1}^{n}|D(z_i)|^p$ and $\|D\|_n := \|D\|_{2,n}$ and let $s(z_i) = \sigma_i$, $m(z_i) = \mu_i - k_i$, and $\nu(z_i) = (\mu_i - k_i)/\sigma_i$. Further define, for $q \geq 1$, $s_q := \|s\|_{q,n}$, $m_q := \|m\|_{q,n}$, and $\nu_q := \|\nu\|_{q,n}$.

Theorem 1 ($\tilde{O}(1/\sqrt{n})$-regret for ASSURE*)**.** *Suppose that the class of decision rules $\delta(z;\beta)$ forms a VC subgraph class with index $V(\mathcal{D})$ and envelope function D. Let $\hat{\beta} = \operatorname{argmax}_\beta \hat{W}(\beta)$ be the optimal decision chosen by* ASSURE*. *Then, its corresponding regret satisfies*

$$\mathsf{Regret}_n \lesssim \frac{m_1}{n\log n} + M\frac{\log n}{\sqrt{n}}$$

for $M := \sqrt{V(\mathcal{D})}(s_2 + m_2 + (1+\nu_4)(\|D\|_n + \|D\|_{4,n}))$.

Thus, assuming that s_2, m_2, ν_4 are bounded and the envelope function D is controlled, the procedure achieves $\frac{\log n}{\sqrt{n}}$ regret in general. A leading case is if all μ_i, k_i, σ_i are bounded, collected in the following corollary.

Corollary 1. *Assume that $\mu_{1:n}, \sigma_{1:n}, \sigma_{1:n}^{-1}, k_{1:n}$ are bounded by an absolute constant B_1. Let $\mathcal{D}$ be VC subgraph with envelope function D uniformly bounded by a constant B_2. Then, for $\hat{\beta} = \operatorname{argmax}_\beta \hat{W}(\beta)$, we have $\sup_{\mu_{1:n}} \mathsf{Regret}_n \lesssim \frac{\log n}{\sqrt{n}}$.*

Broadly speaking, Theorem 1 follows a standard empirical-risk-minimization argument. Regret can be controlled by controlling the empirical process $\sup_\beta |\hat{W}(\beta) - W(\beta)|$. 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 (2.4) 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. **Examples.** We now enumerate several decision rule classes that are VC-subgraph, which we check in Appendix E.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, common in digital experimentation (Kohavi *et al.*, 2020) and medicine (Manski, 2019).

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

(4) Gaussian empirical Bayes models:

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

This is a family of decision rule classes by parametrizing s_0, m_0 as different functions of β.[12]

Gaussian empirical Bayes models put 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 decision rules. Particular parametrizations of m_0, s_0^2 yield important examples. We shall refer to them with different names listed in Table 1.

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

Theorem 2 (Matching Regret Lower Bounds)**.** *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*

[12]Whether this type of decision rules is VC-subgraph is less obvious: Proposition E.2, via model theory, shows that parametric CLOSE-GAUSS is VC-subgraph.

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, forthcoming)	$a_1 + a_2 \log \sigma_i$	$\exp(b_1 + b_2 \log \sigma_i)$

TABLE 1. Examples of Gaussian empirical Bayes models where $\mu_i \mid z_i \sim \mathcal{N}(m_0(z_i), s_0(z_i)^2)$.

rules $a_i(Y_1, \ldots, Y_n) \in [0, 1]$,

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

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, C is a constant that only depends on M.

Here, the regret notion in Theorem 2 compares the welfare of an arbitrary binary selection decision $a_i(Y_1, \ldots, Y_n)$ to the expected welfare of the best decision in the thresholding class restricted to $|\beta| \leq M$. Lower bounds for this quantity thus imply lower bounds when we allow the oracle to be more flexible than using $\delta = \beta$ and constrain the decision maker to threshold decision rules.[13] Theorem 2 is derived by an application of Le Cam's two-point method, considering cases in which, for some $h > 0$, either $\mu_i = h/\sqrt{n}$ for all i or $\mu_i = -h/\sqrt{n}$ for all i.[14] In this case, the optimal β^* is either $+M$ or $-M$, meaning that $W(\cdot)$ is maximized on the boundary instead of at a local maximum. 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.

3.4. **Fast rates for ASSURE*.** The performance for ASSURE* can be better than the $\tilde{O}(n^{-1/2})$ rate predicts. 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). 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

[13] 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.

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

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.

For simplicity, we will restrict to a case where the decision rules are parametrized by a scalar parameter ($d = 1$). 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^* 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. Furthermore, suppose that β is restricted to a compact set Θ which includes β^*. Lastly, suppose that the derivatives $\delta'(\beta, z), \delta''(\beta, z)$ form VC-subgraph function classes $\mathcal{D}', \mathcal{D}''$ with indices $V(\mathcal{D}')$ and $V(\mathcal{D}'')$. 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 $V(\mathcal{D}'), V(\mathcal{D}''), \Theta$.

Appendix C.4.2 illustrates a class of decision rules and lower-level assumptions on $\mu_{1:n}$ for which Theorem 3 applies. Theorem 3 exploits the fact that $\hat{\beta}$ is a local maximum, which enables the approximation $W(\hat{\beta}) \approx W(\beta^*) + W''(\beta^*)(\hat{\beta} - \beta^*)^2$. By analogously Taylor approximating $0 = \hat{W}'(\hat{\beta})$, we can show 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.

Finally, we present a matching lower bound for the fast rate.

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 intuition for Theorem 4 is to lower bound the compound regret by the Bayes regret for a well-chosen prior and then further lower bound the Bayes regret by Le Cam's two point argument, building on Polyanskiy and Wu (2021).[15]

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,[16] a reasonable worry is whether our construction is sensitive to the quality of the Gaussian approximation.

This section thus considers Y_i explicitly as averages of i.i.d. random variables, and analyze 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 to 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.

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 $G_1^{(m)}, \dots, G_n^{(m)}$ and $j = 1, \dots, m$.[17] The decision-maker observes the sample mean of cell i, $\bar{Y}_i := \frac{1}{m}\sum_{j=1}^m Y_{ij}$. Assume that costs are zero $k_i = 0$. We will consider a local asymptotic regime similar to Hirano and Porter (2009), where the micro-data distribution $G_i^{(m)}$ has mean $\mu_i^{(m)} := \mu_i/\sqrt{m}$ and standard deviation σ_i. Let $F_i^{(m)}(t)$ be the CDF of the t-statistic, let $\bar{F}_i^{(m)}(t) := 1 - F_i^{(m)}(t)$ denote the corresponding survival function, and let $\delta_i = \sqrt{m}\delta_i^{(m)}$ be a rescaled version

[15] 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 $Y_1, \dots, Y_n$ 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.

[16] 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 means consistent estimation of certain moments is sufficient.

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

of the threshold. This rescaling allows us to write:

$$\mathbb{E}[\mathbf{1}(\bar{Y}_i > \delta_i^{(m)})] = \bar{F}_i^{(m)}\left(\frac{\delta_i - \mu_i}{\sigma_i}\right) \quad \bar{F}_i^{(m)}(t) := \mathbb{P}_{G_i^{(m)}}\left(\sqrt{m}(\bar{Y}_i - \mu_i^{(m)})/\sigma_i > t\right).$$

Importantly, we rescale the welfare function accordingly: Let

$$W(\delta) := \frac{1}{n}\sum_{i=1}^{n} \underbrace{(\sqrt{m}\cdot\mu_i^{(m)})}_{\text{scaling by } \sqrt{m}} \underbrace{\bar{F}_i^{(m)}\left(\frac{\delta_i - \mu_i}{\sigma_i}\right)}_{\text{true selection prob.}} = \frac{1}{n}\sum_{i=1}^{n} \mu_i \bar{F}_i^{(m)}\left(\frac{\delta_i - \mu_i}{\sigma_i}\right). \tag{4.1}$$

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 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

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

Theorem 5 (ASSURE* bias with non-Gaussian micro-data)**.** *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 averages of polynomials in $|\mu_i|, \sigma_i, \mu_{3,i}, \mu_{4,i}$.

By independence and the bound $\mathrm{Var}(\hat{w}_{h,i}(Y_i;\delta)) = O((\log n)^2)$, we expect 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 $m \to \infty$. 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 approach that does not use this information: Consider an estimator that simply replaces μ_i with $\bar{Y}_i$:

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

Maximizing $\hat{W}_{\text{CES}}$ over fixed thresholds $\delta_i = \beta$ results in the empirical success rule $\mathbf{1}(\bar{Y}_i > 0)$ (Manski, 2004). Under mild assumptions, it is not difficult to show that the bias at this scale does not vanish,[18] leaving the mean-squared error no better than $1 + 1/n$:

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

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. **Complex decisions.** We discuss how ASSURE can be used to *ensemble* complex decision rules. To motivate, practitioners may consider multiple methods and want to combine them. For instance, one model may be 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(NPMLE). Another may be 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 $m^{(-i)}_{\hat{G}_1}(y, z)$ be a posterior mean function fitted on the data (Y_{-i}, z_{-i}). Let $m^{(-i)}_{\hat{G}_2}(y, z)$ 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 prediction is greater than the implementation cost:

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

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

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

(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[19]

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

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

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

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

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

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

for parameters $(\tilde{a}, \tilde{b}, s_0^2)$. In particular, under this empirical Bayes model, since costs are zero, the threshold 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.

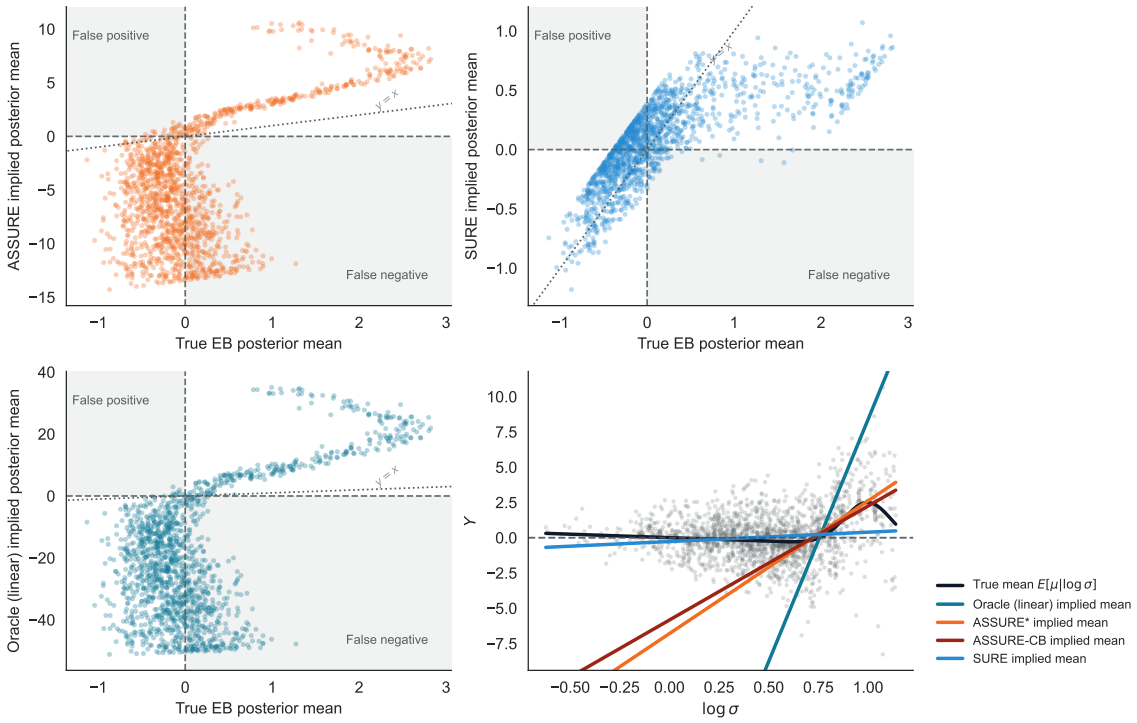


FIGURE 3. Results from the DGP (5.1) with a misspecified prior.

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 $m_0(\log\sigma_i)$, the band $m_0(\log\sigma_i) \pm \sqrt{0.2}$, 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.[20] Table 2 shows the average utility across 1000 draws. We see from 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

[20]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 substracting by the average of σ_i^2. □

over $\mathcal{D}$ to one, ASSURE* has relative performance 0.857 while SURE and plug-in has relative performance only about 0.35.

To unpack why ASSURE* has good performance, it is helpful to compare results on the same empirical-Bayes scale. 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 $m_0(\log \sigma_i) > 0$.

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 optimization methods à la Chernozhukov and Hong (2003).

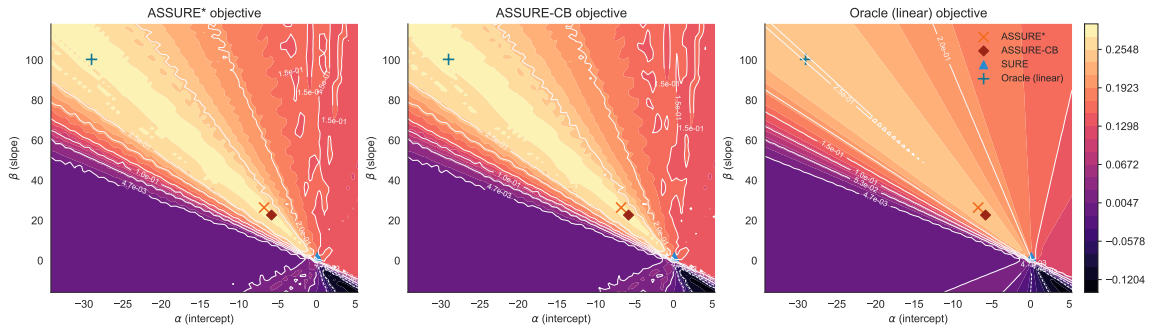


FIGURE 4. Objective function landscape of (5.1).

It does not appear that the higher-order sinc kernel *per se* results in worse optimization landscape than Gaussian kernel, however.

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 experimentation-program 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 to recommend to low-income households using noisy estimates from Chetty *et al.* (2018). Our implementation follows Chen (forthcoming), 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.* (2018) as σ_i.

Bergman *et al.* (2024) are interested in selecting the top third of census tracts. We interpret that targeting rule as implicitly defining a constant implementation cost $k_i = k$. To calibrate k, we compute posterior means under Fay–Herriot and let

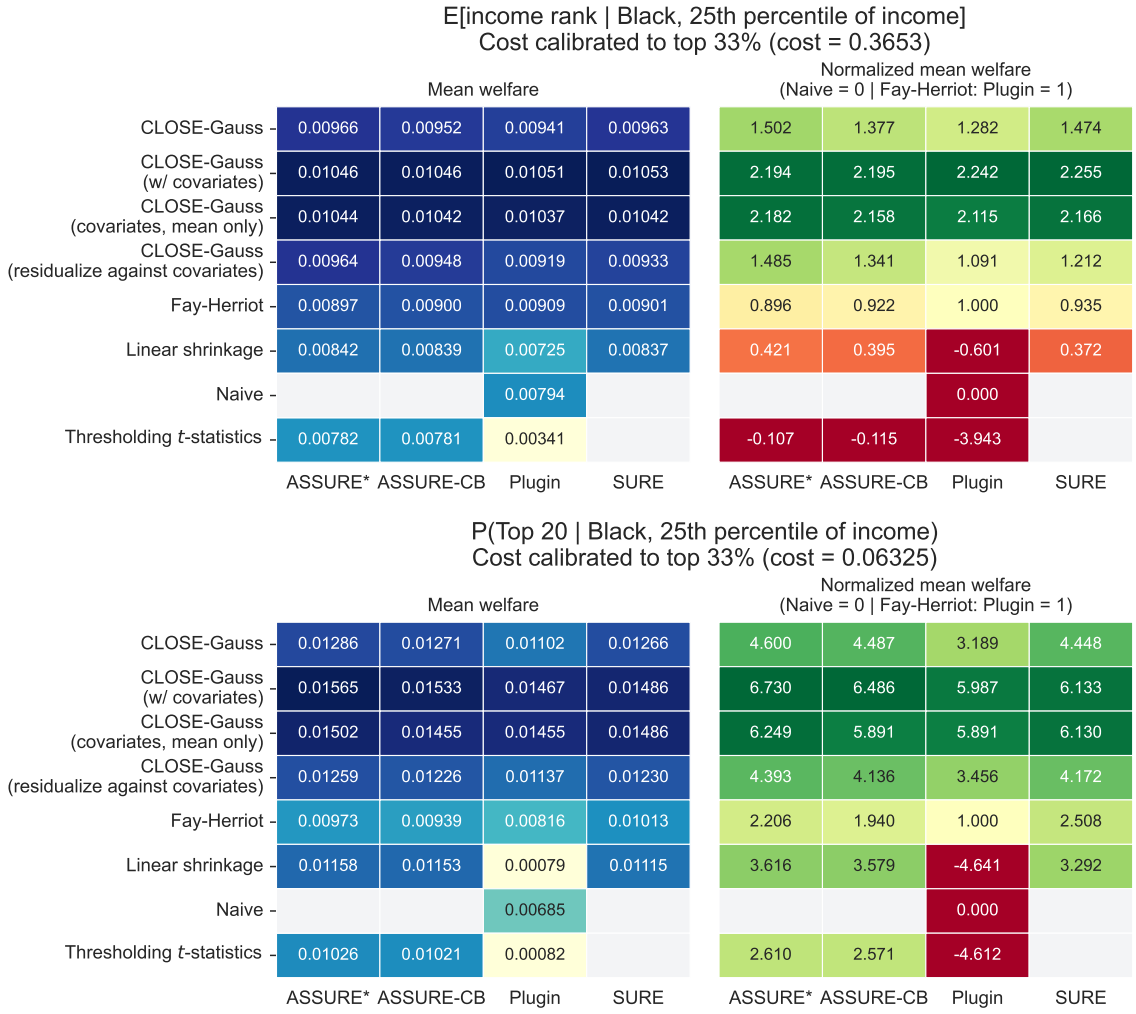


FIGURE 5. Absolute and relative welfare for two different economic mobility measures in Bergman *et al.* (2024) application.

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. □

$k_{\text{calibrated}}$ 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.

To evaluate the performance of different selection decisions, following Chen (forthcoming), 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; see Appendix G. We consider a few instances of decision rules discussed in Table 1 and Section 3.2: (i) Linear shrinkage, (ii) Fay–Herriot, (iii) Parametric CLOSE-GAUSS, (iv) Parametric CLOSE-GAUSS with covariates and its variants, and (v) Thresholding t-statistics.

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 $\beta = 1.96$.

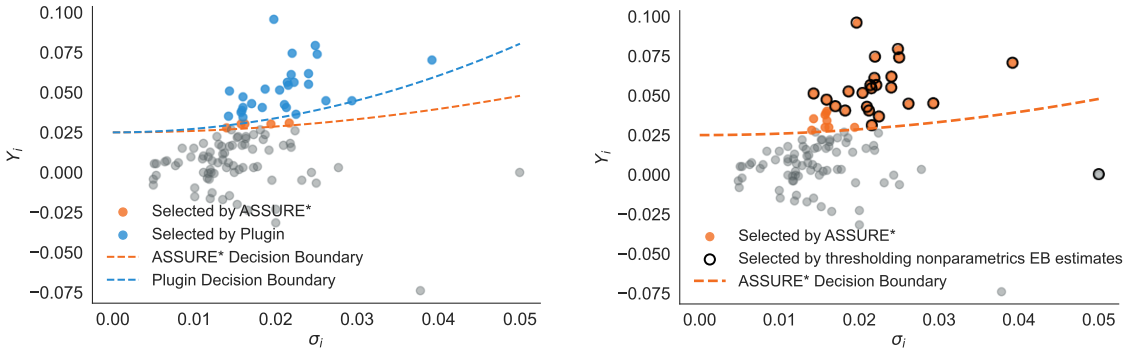
Figure 5 displays results for calibrating k to select the top third. The left panels report absolute welfare and the right panels report welfare normalized so that the performance of $\mathbf{1}(Y_i > k)$ achieves zero relative welfare and the plug-in Fay–Herriot decision achieves relative welfare equal 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. For less flexible decision classes, however, 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 through five waves of the experiment. For each job, fictitious applications were sent in pairs, one with a distinctly white-sounding name and the other with a distinctly Black-sounding name. Other attributes in the resume were randomized. The difference in callback rates between races was then averaged across jobs within a firm to produce the estimate Y_i.

Using a false-discovery-rate procedure at the 5% level, Kline *et al.* (2022) identify 23 firms that discriminate against Black applicants, and none that discriminate against White 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 for 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 large collections 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

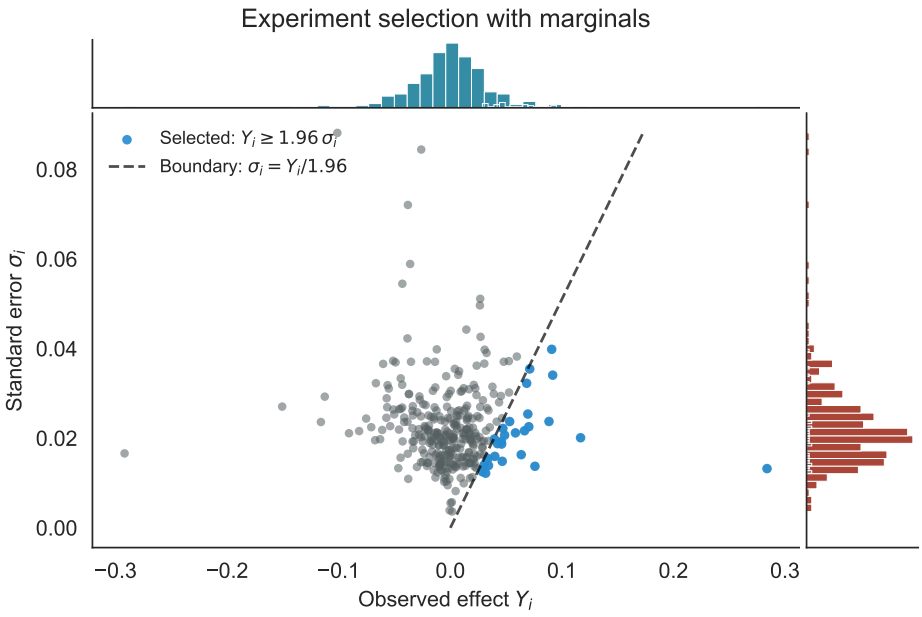


FIGURE 7. Experimentation-program data with the conventional significance threshold

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. □

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 Section 3.2, $\mathcal{D} = \{\delta(z;\beta) = k + \sigma\beta\}$. 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 $\hat{\beta} \approx 0$.

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 literally. Even so, the shape of the estimated welfare curve is informative: If implementation is nearly costless, the preferred threshold is

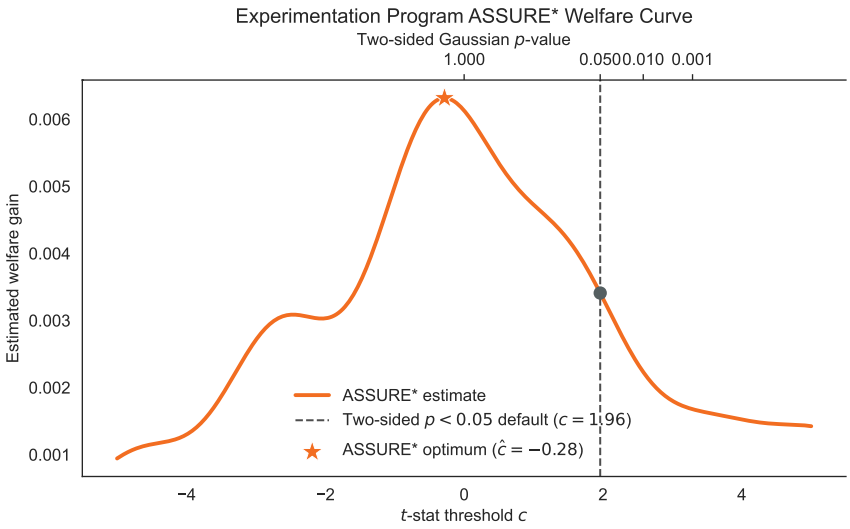


FIGURE 8. ASSURE* welfare curve for t-statistic thresholds.

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. □

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.

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 accurates estimates

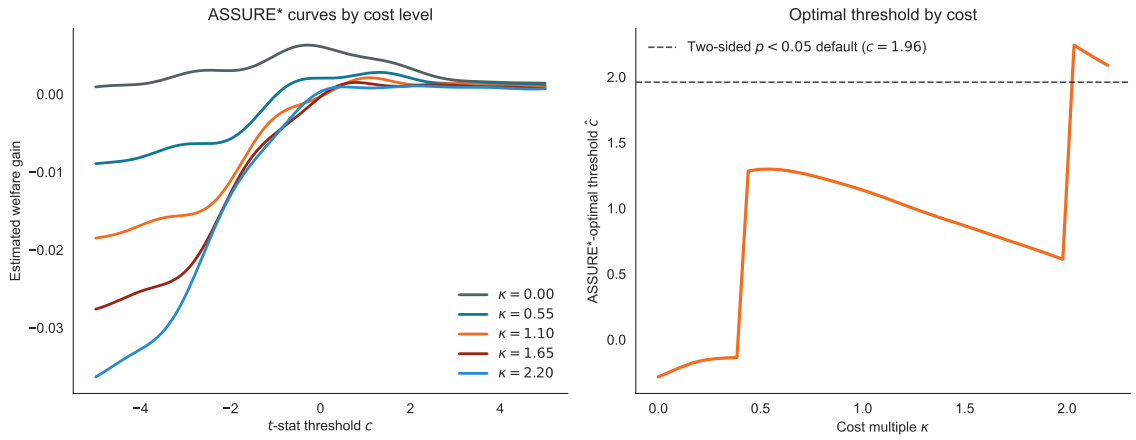


FIGURE 9. ASSURE* welfare and optimal thresholds by cost level.

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 $c = 1.96$.

welfare, optimizing ASSURE* also leads to attractive selection decisions. We present both theoretical guarantees and practical applications illustrating these advantages.

References

ASCHENBRENNER, M., DOLICH, A., HASKELL, D., MACPHERSON, D. and STARCHENKO, S. (2013). Vapnik–chervonenkis density in some theories without the independence property, ii. 90

AUDIBERT, J.-Y. and TSYBAKOV, A. B. (2005). Fast learning rates for plug-in classifiers under the margin condition. *arXiv preprint math/0507180*. 14, 16

— and — (2007). Fast learning rates for plug-in classifiers. 16

AZEVEDO, E. M., DENG, A., MONTIEL OLEA, J. L., RAO, J. and WEYL, E. G. (2020). A/b testing with fat tails. *Journal of Political Economy*, **128** (12), 4614–000. 2, 27, 29, 41

BERGMAN, P., CHETTY, R., DELUCA, S., HENDREN, N., KATZ, L. F. and PALMER, C. (2024). Creating moves to opportunity: Experimental evidence on barriers to neighborhood choice. *American Economic Review*, **114** (5), 1281–1337. 2, 3, 24, 25, 94

BERMAN, R. and VAN DEN BULTE, C. (2022). False discovery in a/b testing. *Management Science*, **68** (9), 6762–6782. 29

Bonhomme, S. and Denis, A. (2024). Estimating heterogeneous effects: applications to labor economics. *Labour Economics*, **91**, 102638. 3, 6

Brown, L. D., Greenshtein, E. and Ritov, Y. (2013). The poisson compound decision problem revisited. *Journal of the American Statistical Association*, **108** (502), 741–749. 92

Chamberlain, G. (1984). Panel data. In Z. Griliches and M. D. Intriligator (eds.), *Handbook of Econometrics*, vol. 2, *22*, 1st edn., Elsevier, pp. 1247–1318. 13

Chen, J. (forthcoming). Empirical bayes when estimation precision predicts parameters. *Econometrica*. 3, 4, 5, 6, 10, 13, 16, 24, 26, 94, 95

Cheng, X., Ho, S. C. and Schorfheide, F. (2025). Optimal estimation of two-way effects under limited mobility. *arXiv preprint arXiv:2506.21987*. 2, 4, 5, 13

Chernozhukov, V. and Hong, H. (2003). An mcmc approach to classical estimation. *Journal of econometrics*, **115** (2), 293–346. 23

Chetty, R., Friedman, J. N., Hendren, N., Jones, M. R. and Porter, S. R. (2018). *The opportunity atlas: Mapping the childhood roots of social mobility*. Tech. rep., National Bureau of Economic Research. 2, 24

—, — and Rockoff, J. E. (2014). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American economic review*, **104** (9), 2593–2632. 2

Chou, W., Gray, C., Kallus, N., Bibaut, A. and Ejdemyr, S. (2025). Evaluating decision rules across many weak experiments. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 4365–4374. 27

Crippa, F. (2025). Regret analysis in threshold policy design. *Journal of Econometrics*, **249**, 105998. 5, 16

Dano, K., Honoré, B. E. and Weidner, M. (2025). Binary choice logit models with general fixed effects for panel and network data. *arXiv preprint arXiv:2508.11556*. 6, 13

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

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, vol. 1. Cambridge University Press. 3, 13

— (2016). Empirical bayes deconvolution estimates. *Biometrika*, **103** (1), 1–20. 27

— and Hastie, T. (2021). *Computer age statistical inference, student edition: algorithms, evidence, and data science*, vol. 6. Cambridge University Press. 91

Fay III, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of james-stein procedures to census data. *Journal of the American Statistical Association*, **74** (366a), 269–277. 16

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

Ghosh, S., Ignatiadis, N., Koehler, F. and Lee, A. (2025). Stein's unbiased risk estimate and hyvarinen's score matching. *arXiv preprint arXiv:2502.20123*. 13

Gu, J. and Koenker, R. (2023). Invidious comparisons: Ranking and selection as compound decisions. *Econometrica*, **91** (1), 1–41. 6

Hirano, K. (2023). A comment on: "invidious comparisons: Ranking and selection as compound decisions" by jiaying gu and roger koenker. *Econometrica*, **91** (1), 43–46. 6

— and Porter, J. R. (2009). Asymptotics for statistical treatment rules. *Econometrica*, **77** (5), 1683–1701. 3, 5, 18

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

Hudson, H. M. (1978). A natural identity for exponential families with applications in multiparameter estimation. *The Annals of Statistics*, **6** (3), 473–484. 90

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

Ishihara, T. and Kurisu, D. (2022). Shrinkage methods for treatment choice. *arXiv preprint arXiv:2210.17063*. 11

Jana, S., Polyanskiy, Y., Teh, A. Z. and Wu, Y. (2023). Empirical bayes via erm and rademacher complexities: the poisson model. In *The Thirty Sixth Annual Conference on Learning Theory*, PMLR, pp. 5199–5235. 91

—, — and Wu, Y. (2022). Optimal empirical bayes estimation for the poisson model via minimum-distance methods. *arXiv preprint arXiv:2209.01328*. 91, 92

Jiang, W. (2020). On general maximum likelihood empirical bayes estimation of heteroscedastic iid normal means. 6, 13

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

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

Kline, P., Rose, E. K. and Walters, C. R. (2022). Systemic discrimination among large us employers. *The Quarterly Journal of Economics*, **137** (4), 1963–2036.

2, 26, 27

Kline, P. M., Rose, E. K. and Walters, C. R. (2024). Discrimination report cards: An empirical bayes ranking approach. 6

Kohavi, R., Tang, D. and Xu, Y. (2020). *Trustworthy online controlled experiments: A practical guide to a/b testing*. Cambridge University Press. 15

Kolmogorov, A. N. (1950). Unbiased estimates. *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya*, **14** (4), 303–326. 8, 10

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

Lei, L. (2025). Discussion of "data fission: Splitting a single data point"–some asymptotic results for data fission. *Journal of the American Statistical Association*, **120** (549), 147–150. 4

Leiner, J., Duan, B., Wasserman, L. and Ramdas, A. (2023). Data fission: splitting a single data point. *Journal of the American Statistical Association*, pp. 1–12. 4, 10

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

— (2000). On an empirical bayes test for a normal mean. *Annals of statistics*, pp. 648–655. 2, 8, 9

Liang, T. C. (2004). On optimal convergence rate of empirical bayes tests. *Statistics & probability letters*, **68** (2), 189–198. 2, 18

Luo, Y., Fithian, W. and Lei, L. (2025). Improving knockoffs with conditional calibration. *The Annals of Statistics*, **53** (5), 2283–2302. 7

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica*, **72** (4), 1221–1246. 3, 20

— (2009). *Identification for prediction and decision*. Harvard University Press. 11

— (2019). Treatment choice with trial data: Statistical decision theory should supplant hypothesis testing. *The American Statistician*, **73** (sup1), 296–304. 3, 15

— (2021). Econometrics for decision making: Building foundations sketched by haavelmo and wald. *Econometrica*, **89** (6), 2827–2853. 3, 14

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

Montiel Olea, J. L., O'Flaherty, B. and Sethi, R. (2021). Empirical bayes counterfactuals in poisson regression with an application to police use of deadly force. *Available at SSRN 3857213*. 91

Müller, U. K. and Watson, M. W. (2024). Spatial unit roots and spurious regression. *Econometrica*, **92** (5), 1661–1695. 3

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

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

Onshuus, A. and Quiroz, A. J. (2015). Metric entropy estimation using o-minimality theory. *arXiv preprint arXiv:1511.07098*. 90

Pensky, M. (2017). Minimax theory of estimation of linear functionals of the deconvolution density with or without sparsity. 8, 10

Polyanskiy, Y. and Wu, Y. (2021). Sharp regret bounds for empirical bayes and compound decision problems. *arXiv preprint arXiv:2109.03943*. 16, 18

Ponomarev, K. and Semenova, V. (2024). On the lower confidence band for the optimal welfare. *arXiv preprint arXiv:2410.07443*. 16

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

Schönberg, I. J. (1948). On variation-diminishing integral operators of the convolution type. *Proceedings of the National Academy of Sciences*, **34** (4), 164–169. 72

Soloff, J. A., Guntuboyina, A. and Sen, B. (2024). Multivariate, heteroscedastic empirical bayes via nonparametric maximum likelihood. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, p. qkae040. 13

Stefanski, L. A. (1989). Unbiased estimation of a nonlinear function a normal mean with application to measurement err oorf models. *Communications in Statistics-Theory and Methods*, **18** (12), 4335–4358. 9, 10, 44, 45

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

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

Sudijono, T., Ejdemyr, S., Lal, A. and Tingley, M. (2024). Optimizing returns from experimentation programs. *arXiv preprint arXiv:2412.05508*. 5, 27, 28, 29

Tate, R. F. (1959). Unbiased estimation: functions of location and scale parameters. *The Annals of Mathematical Statistics*, pp. 341–366. 8, 10

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer Series in Statistics, Springer. 9

Van den Dries, L. (1998). *Tame topology and o-minimal structures*, vol. 248. Cambridge university press. 90

van der Vaart, A. and Wellner, J. A. (2023). Empirical processes. In *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer, pp. 127–384. 53, 65, 83, 84, 85, 86, 87, 88, 89, 90

van der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes*, Springer, pp. 16–28. 83, 84

Vershynin, R. (2009). High-dimensional probability. 15

Walters, C. (2024). Empirical bayes methods in labor economics. In *Handbook of Labor Economics*, vol. 5, Elsevier, pp. 183–260. 3, 6

Weinstein, A., Ma, Z., Brown, L. D. and Zhang, C.-H. (2018). Group-linear empirical bayes estimates for a heteroscedastic normal mean. *Journal of the American Statistical Association*, **113** (522), 698–710. 5, 13

Wilkie, A. J. (1996). 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*, **9** (4), 1051–1094. 90

Xie, X., Kou, S. and Brown, L. D. (2012). Sure estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association*, **107** (500), 1465–1479. 4

Zhou, F. and Li, P. (2019). A fourier analytical approach to estimation of smooth functions in gaussian shift model. *arXiv preprint arXiv:1911.02010*. 8, 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

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

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

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

$$\begin{aligned}\frac{|\mu - k|}{\pi}\int_{1/h}^\infty \frac{1}{\omega} e^{-\frac{1}{2}\omega^2} d\omega &\leq \frac{h|\mu - k|}{\pi}\int_{1/h}^\infty e^{-\frac{1}{2}\omega^2} d\omega \\ &\lesssim h^2|\mu - k|e^{-1/2h^2}.\end{aligned}$$

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 $\left(\frac{1}{n}\sum_{i=1}^n |\mu_i - k_i|\right) h^2 e^{-1/2h^2}$.

Next, we establish the second claim

$$\mathbb{E}\left[\left(\hat{W}_n(\beta) - W(\beta)\right)^2\right] = O\left(\frac{(\log n)^2}{n}\right),$$

pointwise in β. Recall the Gaussian Poincaré inequality, which states that for $Y \sim \mathcal{N}(0, 1), \operatorname{Var} f(Y) \leq \mathbb{E}\left[f'(Y)^2\right]$ where f is a differentiable function. Let $\tilde{Y} \sim \mathcal{N}(\mu, \sigma^2)$. Applying the Gaussian Poincaré inequality to the function $f(\mu + \sigma x)$, we find

$$\operatorname{Var} f(\tilde{Y}) \leq \sigma^2 \mathbb{E}\left[f'(\tilde{Y})^2\right].$$

It is easy to verify that

$$|\partial_y w_h(y; z_i, \beta)| \lesssim \frac{1}{h^2} + \frac{|\delta(z_i; \beta) - k_i|}{\sigma_i h}. \tag{A.2}$$

Indeed, $\partial_y w_h(y; z_i, \beta) = \partial_y \Psi_h(y; z_i, \delta(z_i; \beta))$ and we can compute

$$\begin{aligned}\partial_y \Psi_h(y; z_i, C) &= \frac{y - k_i}{\sigma_i h} \operatorname{sinc}\left(\frac{y - C}{\sigma_i h}\right) + \operatorname{Csinc}\left(\frac{y - C}{\sigma_i h}\right) - \frac{1}{h^2} \operatorname{sinc}'\left(\frac{y - C}{\sigma_i h}\right) \\ &= \frac{y - C}{\sigma_i h} \operatorname{sinc}\left(\frac{y - C}{\sigma_i h}\right) + \frac{C - k_i}{\sigma_i h} \operatorname{sinc}\left(\frac{y - C}{\sigma_i h}\right) \\ &\quad + \operatorname{Csinc}\left(\frac{y - C}{\sigma_i h}\right) - \frac{1}{h^2} \operatorname{sinc}'\left(\frac{y - C}{\sigma_i h}\right)\end{aligned}$$

Recall from the definition of Csinc that it is uniformly bounded since the sine integral function is. From the calculation in Lemma H.1, it is easily seen that $\operatorname{sinc}, \operatorname{sinc}'$ are uniformly bounded as well. Finally, $y \operatorname{sinc} y = \frac{1}{\pi} \sin y$ which is also uniformly bounded. (A.2) follows. Now, applying the Gaussian Poincaré inequality with w_h, defined in (2.7), and using (A.2), we obtain

$$\operatorname{Var}(w_{1/\sqrt{2\log n}}(Y_i, z_i, \beta)) \lesssim h^{-4}\sigma_i^2 + h^{-2}\left|\delta(z_i;\beta) - k_i\right|^2.$$

Thus by independence of the data Y_i, the variance of $\hat{W}_n(\beta)$ can be upper bounded by a constant multiple times

$$\frac{(\log n)^2}{n}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i^2\right) + \frac{\log n}{n}\left(\frac{1}{n}\sum_{i=1}^n (\delta(z_i;\beta) - k_i)^2\right).$$

The bias-variance decomposition for mean-squared error gives the result. □

Proof of Theorem 1. The optimality of $\hat{\beta}$ for $\hat{W}_n$ implies $\hat{W}_n(\hat{\beta}) \geq \hat{W}_n(\beta^*)$ almost surely. Recalling $u(\beta)$ from (2.1), we have $\mathbb{E}[u(\beta^*)] = W(\beta^*)$ and

$$\begin{aligned}
\mathsf{Regret}_n &= W(\beta^*) - \mathbb{E}[u(\hat{\beta})] = \mathbb{E}[u(\beta^*) - u(\hat{\beta})] \\
&\leq \mathbb{E}\left[u(\beta^*) - \hat{W}_n(\beta^*) + \hat{W}_n(\hat{\beta}) - u(\hat{\beta})\right] \leq 2\mathbb{E}\left[\sup_\beta |u(\beta) - \hat{W}_n(\beta)|\right] \\
&\leq 2\mathbb{E}\left[\sup_\beta |\hat{W}(\beta) - W(\beta)|\right] + 2\mathbb{E}\left[\sup_\beta |u(\beta) - W(\beta)|\right]. \qquad \text{(A.3)}
\end{aligned}$$

Starting from (A.3), we use the bias bound in Proposition 1 to bound:

$$\text{(A.3)} \lesssim \mathbb{E}\left[\sup_\beta |u(\beta) - W(\beta)|\right] + \mathbb{E}\left[\sup_\beta |\hat{W}(\beta) - \mathbb{E}\hat{W}(\beta)|\right] + \underbrace{\frac{m_1}{n\log n}}_{\text{by Proposition 1}}.$$

Letting $A_1(\beta) := u(\beta) - W(\beta)$, $A_2(\beta) := \hat{W}(\beta) - \mathbb{E}\hat{W}(\beta)$, we will next individually bound $\mathbb{E}\sup_\beta |A_1(\beta)|, \mathbb{E}\sup_\beta |A_2(\beta)|$. We defer the bounds for these two terms to Propositions C.1 and C.2, which prove that

$$\begin{aligned}
\mathbb{E}\sup_\beta |A_1(\beta)| &\lesssim m_2 \frac{\sqrt{V(\mathcal{D})}}{\sqrt{n}} \\
\mathbb{E}\sup_\beta |A_2(\beta)| &\lesssim \frac{\log n}{\sqrt{n}}(s_2 + m_2 + (\nu_4 + 1)\sqrt{V(\mathcal{D})}(\|D\|_n + \|D\|_{4,n})).
\end{aligned}$$

Since $V(\mathcal{D}) \geq 1$,

$$\mathsf{Regret}_n \leq (\text{A.3}) \lesssim M \frac{\log n}{\sqrt{n}},$$

with $M := \sqrt{V(\mathcal{D})}(s_2 + m_2 + (1+\nu_4)(\|D\|_n + \|D\|_{4,n}))$ as desired. □

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

$$\frac{h}{\sqrt{n}}\left(\Phi\left(\frac{h}{\sqrt{n}} + M\right) - 1 + \max(1 - g_+, g_-)\right) \tag{A.4}$$

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

$$\sup_{\mu_{1:n}\in[-1,1]^n}\left\{\sup_{|\beta|<M} W(\beta; \mu_{1:n}) - \frac{1}{n}\sum_{i=1}^n \mu_i \mathbb{E}_{Y_{1:n}\sim\mathcal{N}(\mu_{1:n}, I_n)}[a_i(Y_1, \ldots, Y_n)]\right\} \geq (\text{A.4}).$$

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

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

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

$$(\text{A.4}) \geq \frac{h}{\sqrt{n}}\left(\Phi\left(\frac{h}{\sqrt{n}} + M\right) - 1 + \frac{1}{2} - \frac{h}{2}\right) \geq \frac{h}{\sqrt{n}}\left(c_M - \frac{h}{2}\right) \gtrsim_M \frac{1}{\sqrt{n}}$$

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

$$\hat{\beta}^{(-i)} = \operatorname*{argmax}_{\beta} \sum_{j \neq i} g_n(Y_j, z_j, \beta), \tag{A.5}$$

where

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

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

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

Thus

$$\mathbb{E}\left[W(\hat{\beta}) - u(\hat{\beta})\right] = \frac{1}{n}\sum_{i=1}^{n}(\mu_i - k_i)\mathbb{E}\left[\Phi\left(\frac{\mu_i - \delta(\hat{\beta}, z_i)}{\sigma_i}\right) - \Phi\left(\frac{\mu_i - \delta(\hat{\beta}^{(-i)}, z_i)}{\sigma_i}\right)\right] \tag{A.6}$$

$$+ \frac{1}{n}\sum_{i=1}^{n}(\mu_i - k_i)\left(\mathbb{P}(Y_i \geq \delta(\hat{\beta}^{(-i)}, z_i)) - \mathbb{P}(Y_i \geq \delta(\hat{\beta}, z_i))\right). \tag{A.7}$$

Using Lipschitzness of Φ and $\delta(\cdot, z_i)$, we will upper bound (A.6) 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.7), 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.7) is bounded above by

$$|(\text{A.7})| \leq \frac{1}{n}\sum_{i=1}^n \frac{L_i|\mu_i - k_i|}{\sigma_i} E_n + \frac{1}{n}\sum_{i=1}^n |\mu_i - k_i| \mathbb{P}(|\hat{\beta}^{(-i)} - \hat{\beta}| > E_n). \tag{A.8}$$

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

$$\frac{1}{n}\sum_{i=1}^n \left(1 + \frac{L_i}{\sigma_i}\right)|\mu_i - k_i| \left(E_n + \mathbb{P}(|\hat{\beta}^{(-i)} - \hat{\beta}| > E_n)\right). \tag{A.9}$$

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 $a_i(Y_1, \dots, Y_n) \in \{0, 1\}$. We will first exhibit a class of priors that are supported on Θ. Proposition C.7 verifies that a set of assumptions Assumption C.1(1)–Assumption C.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 C.1(1) with $B = 1$ and, choosing κ small enough, it satisfies Assumption C.1(3) since $\frac{1}{n}\sum_{i=1}^n \mu_i^2 = 1$. Lastly, Assumption C.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:

$$\sup_{\mu_{1:n}\in\Theta} \mathsf{Regret}_n(a_{1:n}) \geq \mathbb{E}_{G_p^{\otimes n}|A}\left[\mathsf{Regret}_n(a_{1:n})\right] \tag{A.10}$$

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

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

$$\geq \frac{\mathbb{E}_{G_p^{\otimes n}}\left[\mathsf{Regret}_n(a_{1:n})\right] - \mathbb{P}_{G_p^{\otimes n}}(A^c)}{1 - \mathbb{P}_{G_p^{\otimes n}}(A^c)}. \tag{A.13}$$

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

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

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.1 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

$$\begin{aligned} R(G, a_{1:n}) &:= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{G^{\otimes n}}\left[\mu_i \left(a^*_G(Y_i) - a_i(Y_{1:n})\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{G^{\otimes n}}\left[m_G(Y_i) \left(a^*_G(Y_i) - a_i(Y_{1:n})\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{G^{\otimes n}}\left[|m_G(Y_i)|\, \mathbf{1}\left\{a^*_G(Y_i) \neq a_i(Y_{1:n})\right\}\right]. \end{aligned} \tag{A.14}$$

where (A.14) is verified by Lemma C.5.

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.7, 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.13) 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,

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

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.9, $T_1 \lesssim \sqrt{\log n}/\sqrt{m}$. By Lemma C.10, $T_2 \lesssim (\log n)^2/\sqrt{m}$. The broad idea for T_2 is 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.10 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.11 gives $T_3 \lesssim (n\log n)^{-1} + m^{-1/2}$. Combining the three bounds proves the claim. □

Appendix B. Miscellany

Lemma B.1 (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.2, an unbiased estimator for $W(\beta)$ exists if and only if one for $\varphi(\mu)$ exists. The following result shows that $\varphi(\mu)$ cannot be unbiasedly estimated, at least not with estimators with reasonable tail behavior.

Lemma B.2. *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.2. Rewrite the expression as

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

□

Proposition B.1 (Stefanski (1989))**.** *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. □

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.

Online Appendix

March 26, 2026

Contents

Appendix C. Auxiliary results for proofs 47
C.1. Auxiliary results for Proposition 1 47
C.2. Auxiliary results for Theorem 1 49
C.3. Tail bound for regret 53
C.4. Auxiliary results for Theorem 3 54
C.5. Auxiliary results for Theorem 4 74
C.6. Auxiliary results for Theorem 5 76
Appendix D. Results for ASSURE-CB 80
Appendix E. Empirical processes for i.n.i.d. data 86
E.1. Examples of VC classes 89
Appendix F. Other observation distributions: the Poisson case 90
F.1. Proofs for Poisson extension 92
Appendix G. Additional details for Section 6.1 94
Appendix H. Miscellany 102

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 $z = (k, \sigma)$,*

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

interpreted as Lebesgue integrals.

Proof. We consider the fact that for all $a \in \mathbb{R}$,[21]

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

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

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

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:

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

[21] 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.

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

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

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

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

This proves the second claim. □

C.2. **Auxiliary results for Theorem 1.**

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

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

Proof. We will handle this with i.n.i.d. empirical process theory applied to the function class $\mathcal{F}=\left\{f_\beta:\beta\in\mathbb{R}^d\right\}$ with $f_\beta(y,\mu,z)=(\mu-k)\mathbf{1}\left\{y-\delta(z;\beta)\geq 0\right\}$. The class $\mathcal{F}$ has a uniform envelope function given by $F=|\mu-k|$. Moreover, Lemma E.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 β, and so is $\mathbf{1}\left\{y-\delta(\beta,z)\geq 0\right\}$. Multiplying by the fixed function $(\mu-k)$ and applying Lemma E.1(2) gives the claim. Next, we

apply Theorem E.1 with $V_i = (Y_i, \mu_i, z_i)$ and Lemma E.2,

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

Proposition C.2. *In the proof of Theorem 1, we have*

$$\mathbb{E}\left|\sup_\beta A_2(\beta)\right| \lesssim \frac{\log n}{\sqrt{n}}(s_2 + m_2 + (\nu_4+1)\sqrt{V(\mathcal{D})}(\|D\|_n + \|D\|_{4,n})).$$

Proof. Define the function classes $\mathcal{G} = \left\{g_\beta : \beta \in \mathbb{R}^d\right\}, \mathcal{H} = \left\{h_\beta : \beta \in \mathbb{R}^d\right\}$ where

$$\begin{aligned}
g_\beta(y,z) &:= \frac{1}{\pi}(y-k)\,\mathrm{Si}\left(\lambda_n(y - \delta(z;\beta))/\sigma\right) \\
h_\beta(y,z) &:= \sigma\lambda_n\,\mathrm{sinc}\left(\lambda_n(y - \delta(z;\beta))/\sigma\right)
\end{aligned}$$

for $\lambda_n = 1/h_n = 1/\sqrt{2\log n}$. Then:

$$\begin{aligned}
\mathbb{E}\sup_\beta |A_2(\beta)| \le\ & \frac{1}{\sqrt{n}}\mathbb{E}\sup_\beta \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n g_\beta(Y_i,\mu_i,z_i) - \mathbb{E}g_\beta(Y_i,\mu_i,z_i)\right| \\
&+ \frac{1}{\sqrt{n}}\mathbb{E}\sup_\beta \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n h_\beta(Y_i,\mu_i,z_i) - \mathbb{E}h_\beta(Y_i,\mu_i,z_i)\right| \\
&+ \underbrace{\frac{1}{2}\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n (Y_i - \mu_i)\right|}_{\lesssim \frac{1}{\sqrt{n}}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i^2\right)^{1/2}}.
\end{aligned}$$

We can therefore focus on bounding

$$\mathbb{E}\left|\sup_\beta \frac{1}{\sqrt{n}}\sum_{i=1}^n g_\beta(Y_i,\mu_i,z_i) - \mathbb{E}g_\beta(Y_i,\mu_i,z_i)\right|$$

and the analogous empirical process h_β using variants of i.n.i.d. empirical process theory. The function class $\mathcal{G}$ has a uniform envelope function $G(y) = c|y-k|$ for some absolute constant c, since the $\mathrm{Si}(\cdot)$ function is bounded. Following the proof of Theorem E.1 with $V_i = (Y_i, z_i)$, we may apply symmetrization and a maximal inequality to obtain

$$\mathbb{E}\left|\sup_\beta \frac{1}{\sqrt{n}}\sum_{i=1}^n g_\beta(Y_i,\mu_i,z_i) - \mathbb{E}g_\beta(Y_i,\mu_i,z_i)\right| \le \mathbb{E}\left[\int_0^{\eta_n}\sqrt{1+\log N(\epsilon,\mathcal{G},\mathbb{L}_2(\mathbb{P}_n))}d\epsilon\right], \tag{C.1}$$

with $\eta_n := \sup_\beta \|g_\beta\|_n$.

Now, we claim that for some constant C,

$$N(C\lambda_n\hat{\nu}_4\epsilon, \mathcal{G}, \mathbb{L}_2(\mathbb{P}_n)) \le N(\epsilon, \mathcal{D}, \mathbb{L}_4(\mathbb{P}_n)). \tag{C.2}$$

where $\hat{\nu}_4 := \left(\frac{1}{n}\sum_{i=1}^n \frac{(Y_i-k_i)^4}{\sigma_i^4}\right)^{1/4}$. Indeed, 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 Si is Lipschitz for some constant C,

$$\begin{aligned}
&\left\|g_\beta(y,\mu,z) - \frac{1}{\pi}(y-k)\,\mathrm{Si}(\lambda_n(y-\tilde{d})/\sigma)\right\|_{\mathbb{L}^2(\mathbb{P}_n)} \\
&\le C\lambda_n\left(\frac{1}{n}\sum_{i=1}^n \frac{(Y_i-k_i)^2}{\sigma_i^2}\left(\delta(z_i;\beta)-\tilde{d}\right)^2\right)^{1/2} \\
&\le C\lambda_n\left(\frac{1}{n}\sum_{i=1}^n \frac{(Y_i-k_i)^4}{\sigma_i^4}\right)^{1/4}\left(\frac{1}{n}\sum_{i=1}^n\left(\delta(z_i;\beta)-\tilde{d}\right)^4\right)^{1/4} \qquad \text{(Cauchy–Schwarz)} \\
&\le C\lambda_n\hat{\nu}_4\epsilon.
\end{aligned}$$

This shows the pushforward of the $\mathbb{L}^4(\mathbb{P}_n)$ ϵ-cover through the function $\frac{1}{\pi}(y-k)\,\mathrm{Si}(\lambda_n(y-\cdot)/\sigma)$ is a $C\lambda_n\hat{\nu}_4\epsilon$-cover in the $\mathbb{L}^2(\mathbb{P}_n)$ seminorm.

Returning to (C.1), using (C.2), we can bound the entropy integral in (C.1) as

$$\begin{aligned}
&\int_0^{\eta_n} \sqrt{1+\log N(\epsilon, \mathcal{G}, \mathbb{L}_2(\mathbb{P}_n))}d\epsilon \\
&\leq \int_0^{\eta_n} \sqrt{1+\log N\left(\frac{\epsilon}{C\lambda_n\hat{\nu}_4}, \mathcal{D}, \mathbb{L}_4(\mathbb{P}_n)\right)}d\epsilon \\
&= C\lambda_n\hat{\nu}_4\|D\|_{4,n}\int_0^{\eta_n/C\lambda_n\hat{\nu}_4\|D\|_{4,n}} \sqrt{1+\log N\left(\epsilon\|D\|_{n,4}, \mathcal{D}, \mathbb{L}^4(\mathbb{P}_n)\right)}d\epsilon \\
&\leq \left(\eta_n + C\lambda_n\hat{\nu}_4\|D\|_{4,n}\int_0^1 \sqrt{1+\log N\left(\epsilon\|D\|_{n,4}, \mathcal{D}, \mathbb{L}^4(\mathbb{P}_n)\right)}d\epsilon\right). \qquad \text{(C.3)}
\end{aligned}$$

The last line is obtained by splitting the integral into two integrals, one over $[0,1]$ and the other $[1,\infty)$, if needed. The covering number for $\epsilon \geq 1$ is equal to 1, so we may bound the integral for $\epsilon \in [1, \eta_n/C\lambda_n\hat{\nu}_4\|D\|_{4,n})$ by $(\eta_n/C\lambda_n\hat{\nu}_4\|D\|_{4,n} - 1)_+$. Applying the envelope function bound for $\mathcal{G}$ and Lemma E.2,

$$\begin{aligned}
\text{(C.3)} &\lesssim \|G\|_n + \lambda_n\hat{\nu}_4\|D\|_{4,n}\int_0^1 \sqrt{1+\log V(\mathcal{D}) + 4V(\mathcal{D})\log\frac{16e}{\epsilon}}d\epsilon \qquad \text{(C.4)} \\
&\leq \|G\|_n + \lambda_n\hat{\nu}_4\|D\|_{4,n}\sqrt{V(\mathcal{D})}. \qquad \text{(C.5)}
\end{aligned}$$

Taking expectations over the data and applying Cauchy–Schwarz, while using that $\|D\|_{4,n}$ is deterministic, we obtain

$$\mathbb{E}\|G\|_n + \lambda_n\sqrt{V(\mathcal{D})}(\mathbb{E}\hat{\nu}_4^2)^{1/2}\|D\|_{4,n}$$

It remains to bound the last few terms. Applying Jensen's inequality, we have

$$\begin{aligned}
\mathbb{E}\|G\|_n &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n (Y_i - k_i)^2\right)^{1/2} \lesssim \left(\frac{1}{n}\sum_{i=1}^n \sigma_i^2 + (\mu_i - k_i)^2\right)^{1/2} \lesssim s_2 + m_2 \\
\mathbb{E}\hat{\nu}_4^2 &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \frac{(Y_i - k_i)^4}{\sigma_i^4}\right)^{1/2} \lesssim \left(1 + \frac{1}{n}\sum_{i=1}^n (\mu_i - k_i)^4/\sigma_i^4\right)^{1/2} \lesssim 1 + \nu_4^2.
\end{aligned}$$

Thus

$$\begin{aligned}
&\mathbb{E}\sup_\beta \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n g_\beta(Y_i, \mu_i, z_i) - \mathbb{E}g_\beta(Y_i, \mu_i, z_i)\right| \\
&\lesssim \left(s_2 + m_2 + \log n\sqrt{V(\mathcal{D})}(1+\nu_4)\|D\|_{4,n}\right).
\end{aligned}$$

The argument for the function class $\mathcal{H}$ is similar but more straightforward. We take the uniform envelope function $H = C\sigma\lambda_n$ for some constant C, since the sinc

function has a uniform upper bound. Following the proof of Theorem E.1 again with $V_i = (Y_i, z_i)$, we obtain

$$\mathbb{E}\left|\sup_\beta \frac{1}{\sqrt{n}}\sum_{i=1}^n h_\beta(Y_i,\mu_i,z_i) - \mathbb{E}h_\beta(Y_i,\mu_i,z_i)\right| \le \mathbb{E}\left[\int_0^{\eta_n}\sqrt{1+\log N(\epsilon,\mathcal{H},\mathbb{L}_2(\mathbb{P}_n))}d\epsilon\right], \tag{C.6}$$

with $\eta_n := \sup_\beta \|h_\beta\|_n$ this time. Now because the sinc$(\cdot)$ function is Lipschitz with some constant C,

$$N(C\lambda_n^2\epsilon,\mathcal{H},\mathbb{L}_2(\mathbb{P}_n)) \le N(\epsilon,\mathcal{D},\mathbb{L}_2(\mathbb{P}_n)). \tag{C.7}$$

Applying the same logic as before, we obtain an upper bound of

$$\begin{aligned}
&\mathbb{E}\eta_n + \lambda_n^2\|D\|_n\int_0^1\sqrt{1+\log N(\epsilon\|D\|_n,\mathcal{H},\mathbb{L}_2(\mathbb{P}_n))}d\epsilon \\
&\lesssim \|H\|_n + \lambda_n^2\|D\|_n\sqrt{V(\mathcal{D})} \\
&\lesssim (\log n)\left(s_2 + \sqrt{V(\mathcal{D})}\|D\|_n\right).
\end{aligned}$$

Combining the pieces, we have

$$\begin{aligned}
\mathbb{E}\left|\sup_\beta A_2(\beta)\right| &\lesssim \frac{1}{\sqrt{n}}\left(s_2 + m_2 + \log n\sqrt{V(\mathcal{D})}(1+\nu_4)\|D\|_{4,n}\right) \\
&\quad + \frac{\log n}{\sqrt{n}}\left(s_2 + \sqrt{V(\mathcal{D})}\|D\|_n\right) + s_2 \\
&\lesssim \frac{\log n}{\sqrt{n}}(s_2 + m_2 + (\nu_4+1)\sqrt{V(\mathcal{D})}(\|D\|_n + \|D\|_{4,n})). \qquad \square
\end{aligned}$$

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 E.

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

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

where M is the constant in the statement of Theorem 1. The same bound is true for $W(\beta^*) - W(\hat{\beta})$.

Proof. The optimality of $\hat{\beta}$ for $\hat{W}_n(\cdot)$ shows that

$$u(\beta^*) - u(\hat{\beta}) \leq u(\beta^*) - \hat{W}_n(\beta^*) + \hat{W}_n(\hat{\beta}) - u(\hat{\beta}) \leq 2 \sup_\beta |u(\beta) - \hat{W}_n(\beta)|$$

Combined with the uniform bias bound on $\hat{W}_n$ from Proposition 1, we have

$$u(\beta^*) - u(\hat{\beta}) \leq 2 \sup_\beta \left| u(\beta) - \hat{W}_n(\beta) - \mathbb{E}\left[u(\beta) - \hat{W}_n(\beta)\right] \right| + \frac{2m_1}{n \log n}.$$

Let

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

Here, $\mathcal{F}$ is the function class with members

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

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 E.1 with $V_i = (Y_i, \mu_i, z_i)$, the sub-Gaussian norm is bounded above by a constant multiple times

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

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 bound with Propositions C.1 and C.2 completes the proof.

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

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

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.3 (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

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

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

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

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,

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

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.3. By the results of Theorem C.1 we have with probability at least $1 - \delta$

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

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

$$W(\beta^*) - W(\hat{\beta}) \leq CB\frac{(\log n)^{3/2}}{\sqrt{n}} \tag{C.8}$$

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.8) 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

$$W'(\beta^*) - W'(\hat{\beta}) = -W''(\beta^*)(\beta^* - \hat{\beta}) - \frac{W'''(\tilde{\beta})}{2}(\beta^* - \hat{\beta})^2 \tag{C.9}$$

for some $\tilde{\beta}$ between $\hat{\beta}$ and β^*. By Lemma C.3 and the reverse triangle inequality $|a - b| \geq |a| - |b|$ we have

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

where the last line follows from Assumption 1(3). Combining this with the Taylor expansion in Eq. (C.9), for n large enough, on the event A_n,

$$\frac{\kappa}{2}\left|\beta^* - \hat{\beta}\right| \leq \left|W'(\beta^*) - W'(\hat{\beta})\right|. \tag{C.10}$$

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

$$\left|W'(\hat{\beta}) - W'(\beta^*)\right| \leq \frac{C}{n\sqrt{\log n}} + \frac{C(\log n)^{5/2}}{\sqrt{n}}\sqrt{\log(2n)} \tag{C.11}$$

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.10) and (C.11), 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 constant C depending only on B and the VC indices of $\mathcal{D}, \mathcal{D}'$. 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*

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

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

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

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 H.2 and Lemma H.1 shows, $|\frac{d\Psi_h}{dC}| \leq K_h|Y - k|/\sigma$ 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,

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

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

$$\hat{W}_n''(\beta) = \frac{1}{n}\sum_{i=1}^n \Psi_{CC}(Y_i, z_i, \delta(z_i; \beta))\delta'(\beta, z_i)^2 + \Psi_C(Y_i, z_i, \delta(z_i; \beta))\delta''(\beta, z_i). \quad \text{(C.12)}$$

Thus,

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

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*

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

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 H.2. To control $\hat{\beta} - \hat{\beta}^{(-i)}$, we will apply a standard Taylor expansion using the first-order

conditions

$$\sum_{i=1}^{n} \Psi_C(Y_i, z_i, \delta(\hat{\beta}, z_i))\delta'(\hat{\beta}, z_i) = 0 \tag{C.13}$$

$$\sum_{j\neq i} \Psi_C(Y_j, z_j, \delta(\hat{\beta}^{(-i)}, z_j))\delta'(\hat{\beta}^{(-i)}, z_j) = 0, \quad \forall i = 1, \ldots, n. \tag{C.14}$$

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

$$\Delta_{n,i} := \left(\frac{1}{n}\sum_{j=1}^{n} \Psi_{CC}(Y_j, z_j, \delta(\tilde{\beta}_i, z_j))\delta'(\tilde{\beta}_i, z_j)^2 + \Psi_C(Y_j, \tilde{\beta}_i)\delta''(\tilde{\beta}_i, z_j)\right) \tag{C.15}$$

$$= \hat{W}_n''(\tilde{\beta}_i). \tag{C.16}$$

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 H.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,

$$\left|\hat{\beta} - \hat{\beta}^{(-i)}\right| \leq \frac{MB\lambda_n^2}{n|\Delta_{n,i}|}. \tag{C.17}$$

By Lemma C.4 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.17) 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.4. *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*

$$\Delta_{n,i} := \hat{W}''_{n,i}(\tilde{\beta}_i).$$

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's 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,

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

which can in turn be bounded above by

$$\sup_\beta |W'(\beta) - \mathbb{E}\hat{W}_n'(\beta)| + \sup_\beta |\hat{W}_n'(\beta) - \mathbb{E}\hat{W}_n'(\beta)|. \tag{C.18}$$

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. (H.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

$$f_\beta^{(1)}(y, z) = -\lambda_n \left(\frac{y - k}{\sigma} \right) \operatorname{sinc}\left(\lambda_n \left(y - \delta(z; \beta) \right) / \sigma \right) \delta'(\beta, z) \tag{C.19}$$

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

$$f_\beta^{(2)}(y, z) = \lambda_n^2 \operatorname{sinc}'\left(\lambda_n \left(y - \delta(z; \beta) \right) / \sigma \right) \delta'(\beta, z). \tag{C.20}$$

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

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

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

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

Combining these propositions and the bias bound, we obtain

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

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

To get the second claim, we will apply Proposition E.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,*

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

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

$$A(\beta) := \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \Psi_{CC}(Y_i, z_i, \delta(z_i; \beta))\delta'(\beta, z_j)^2$$

$$B(\beta) := \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \Psi_{C}(Y_i, z_i, \delta(z_i; \beta))\delta''(\beta, z_j).$$

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

$$\mathbb{E}\left[\sup_\beta |B(\beta) - \mathbb{E}B(\beta)|\right] \le C(\log n)^{5/2} \tag{C.21}$$

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 H.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 Theorems Theorem 1 and Theorem C.5. Throughout this proof, define the following functions classes indexed by β:

$$\mathcal{F}^{(0)} := \left\{ f_\beta^{(0)}(y, z) = \lambda_n^2 \frac{y - k}{\sigma^2}\operatorname{sinc}'\left(\lambda_n\left(\frac{y - \delta(z;\beta)}{\sigma}\right)\right) : \beta \in \mathbb{R}\right\} \tag{C.22}$$

$$\mathcal{F}^{(1)} := \left\{ f_\beta^{(1)}(y, z) = \lambda_n^2 \frac{y - k}{\sigma^2}\operatorname{sinc}'\left(\lambda_n\left(\frac{y - \delta(z;\beta)}{\sigma}\right)\right)\delta'(\beta, z)^2 : \beta \in \mathbb{R}\right\} \tag{C.23}$$

$$\mathcal{F}^{(2)} := \left\{ f_\beta^{(2)}(y, z) = \frac{\lambda_n^3}{\sigma}\operatorname{sinc}''\left(\lambda_n\left(\frac{y - \delta(z;\beta)}{\sigma}\right)\right)\delta'(\beta, z)^2 : \beta \in \mathbb{R}\right\} \tag{C.24}$$

$$\mathcal{D}'^2 := \left\{\delta'(\beta, z)^2 : \beta \in \mathbb{R}\right\}. \tag{C.25}$$

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

$$F^{(1)} := CB^3\lambda_n^2 \cdot \frac{|y-k|}{\sigma} \tag{C.26}$$

$$F^{(2)} := CB^3\lambda_n^3. \tag{C.27}$$

for some absolute constant C. Furthermore notice that $\mathcal{D}'^2$ is a VC subgraph class, by composing Lemma E.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 Proposition C.5 and Proposition C.6 show that

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

for a constant C that depend on B and the VC indexes of $\mathcal{D}, \mathcal{D}'^2$. Combining this with Eq. (C.21) 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 E.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 E.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*

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

for a constant C depending on B, where $\mathcal{F}^{(1)}$ is the class of functions defined in (C.19).

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

$$\mathbb{E}\left[\|\mathbb{G}\|_{\mathcal{F}^{(1)}}\right] \leq \mathbb{E}\left[\int_0^{\eta_n} \sqrt{1 + \log N(\epsilon, \mathcal{F}^{(1)}, \mathbb{L}_2(\mathbb{P}_n))}d\epsilon\right], \tag{C.28}$$

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

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

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:

$$N(C\left(\lambda_n\hat{\nu}_4 + B\right)\epsilon, \mathcal{F}^{(1)}, \mathbb{L}_2(\mathbb{P}_n)) \leq N(\epsilon, \mathcal{D}', \mathbb{L}_4(\mathbb{P}_n)) \cdot N(\epsilon, \mathcal{F}^{(0)}, \mathbb{L}_2(\mathbb{P}_n)). \tag{C.30}$$

Let $\hat{\zeta}_4 := \left\|\frac{y-k}{\sigma^2}\right\|_{\mathbb{L}^4(\mathbb{P}_n)}$. By an argument analogous to the one used to obtain Eq. (C.2), we also have

$$N(C\lambda_n^2\hat{\zeta}_4\epsilon, \mathcal{F}^{(0)}, \mathbb{L}_2(\mathbb{P}_n)) \leq N(\epsilon, \mathcal{D}, \mathbb{L}_4(\mathbb{P}_n)). \tag{C.31}$$

Putting these bounds into Eq. (C.28), we have

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

Following the change of variables argument used to obtain Equation (C.5) 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, as desired. □

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.20).

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

$$\mathbb{E}\left[\|\mathbb{G}\|_{\mathcal{F}^{(2)}}\right] \leq \mathbb{E}\left[\int_0^{\eta_n} \sqrt{1+\log N(\epsilon, \mathcal{F}^{(2)}, \mathbb{L}_2(\mathbb{P}_n))} d\epsilon\right], \tag{C.32}$$

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

$$\mathcal{S} := \left\{s_\beta(y,z) := \lambda_n^2 \operatorname{sinc}'\left(\lambda_n\left(y-\delta(z;\beta)\right)/\sigma\right) : \beta \in \mathbb{R}\right\} \tag{C.33}$$

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 $d(z_i) \le B$ for all i. Then

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

for a constant C that only depends on B, where we used the envelope bound on the last line. Thus,

$$N(C\left(\lambda_n^2+B\right)\epsilon, \mathcal{F}^{(2)}, \mathbb{L}_2(\mathbb{P}_n)) \le N(\epsilon, \mathcal{D}', \mathbb{L}_2(\mathbb{P}_n)) \cdot N(\epsilon, \mathcal{S}, \mathbb{L}_2(\mathbb{P}_n)) \tag{C.34}$$

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

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

Putting these bounds into Eq. (C.32), we have

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

Following the argument of Proposition C.2 again we then have

$$\mathbb{E}\|\mathbb{G}\|_{\mathcal{F}^{(2)}} \lesssim \mathbb{E}\left\|F^{(2)}\right\|_n + \lambda_n^3(\lambda_n^2+B)\sqrt{V(\mathcal{D})} + (\lambda_n^2+B)\sqrt{V(\mathcal{D}')}, \tag{C.35}$$

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

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

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*

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

for a constant C that depends on B, where $\mathcal{F}^{(1)}$ is defined in (C.23).

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 E.1 as before

with $V_i = (Y_i, z_i)$, we may apply symmetrization and a maximal inequality to obtain

$$\mathbb{E}\left[\|\mathbb{G}\|_{\mathcal{F}^{(1)}}\right] \leq \mathbb{E}\left[\int_0^{\eta_n} \sqrt{1 + \log N(\epsilon, \mathcal{F}^{(1)}, \mathbb{L}_2(\mathbb{P}_n))} d\epsilon\right], \tag{C.36}$$

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.22) 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

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

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

$$N(CB\left(\lambda_n^2\hat{\nu}_4 + B\right)\epsilon, \mathcal{F}^{(1)}, \mathbb{L}_2(\mathbb{P}_n)) \leq N(\epsilon, \mathcal{D}'^2, \mathbb{L}_4(\mathbb{P}_n)) \cdot N(\epsilon, \mathcal{F}^{(0)}, \mathbb{L}_2(\mathbb{P}_n)). \tag{C.37}$$

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

$$N(CB^2\lambda_n^3\hat{\nu}_4\epsilon, \mathcal{F}^{(0)}, \mathbb{L}_2(\mathbb{P}_n)) \leq N(\epsilon, \mathcal{D}, \mathbb{L}_4(\mathbb{P}_n)). \tag{C.38}$$

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,

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

Putting these bounds into Eq. (C.36), we have

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

Following the change of variables argument leading up to Equation (C.5) 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.26). 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.24).

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,

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

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

$$\mathcal{S} := \left\{ s_\beta(y,z) := \frac{\lambda_n^3}{\sigma} \operatorname{sinc}''\left(\lambda_n\left(y-\delta(z;\beta)\right)/\sigma\right) : \beta \in \mathbb{R} \right\} \tag{C.40}$$

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,

$$N(CB\left(\lambda_n^3+B^2\right)\epsilon, \mathcal{F}^{(2)}, \mathbb{L}_2(\mathbb{P}_n)) \leq N(\epsilon, \mathcal{D}'^2, \mathbb{L}_2(\mathbb{P}_n)) \cdot N(\epsilon, \mathcal{S}, \mathbb{L}_2(\mathbb{P}_n)) \tag{C.41}$$

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.39), we have

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

Following the argument of Proposition C.2 again we then have

$$\mathbb{E}\|\mathbb{G}\|_{\mathcal{F}^{(2)}} \lesssim \mathbb{E}\left\|F^{(2)}\right\|_n + \lambda_n^4(\lambda_n^3+B^2)\sqrt{V(\mathcal{D})} + (\lambda_n^3+B^2)\sqrt{V(\mathcal{D}'^2)}. \tag{C.42}$$

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

C.4.2. *Example decision rule class for Theorem 3.*

Example 2. 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 C.1. *(1)* $-B < \min_i \mu_i < 0$ *and* $0 < \max_i \mu_i < B$.

(2) Letting $m_B := \inf_{[B,3B]} t\varphi(t)$,

$$|\overline{\mu}| \leq \overline{\sigma}^2 \frac{m_B}{\varphi(B)}.$$

(3) $\frac{1}{n}\sum_{i=1}^n \mu_i^2 > \kappa_0/\varphi(3B)$.

Proposition C.7. *Fix constants* B, κ_0 *as above. Assume* $\sigma_i^2 \equiv 1$ *and Assumptions C.1(1) to C.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 C.1(1), it is easy to see that Assumption 1(1) holds. Now without loss, assume that $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_n$. 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 $c_1 \leq c_2 \leq ... \leq c_n$, 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 C.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

$$-W'(2B) = \frac{1}{n}\sum_{i=1}^n \mu_i f_+(\mu_i) = \overline{\mu}\overline{f}_+ + \frac{1}{n}\sum_{i=1}^n (\mu_i - \overline{\mu})\left(f_+(\mu_i) - f_+(\overline{\mu})\right) \tag{C.43}$$

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 $f_- := \varphi(x+2B)$. Observe the analogous decomposition

$$\frac{1}{n}\sum_{i=1}^n \mu_i f_-(\mu_i) = \overline{\mu}\overline{f}_- + \frac{1}{n}\sum_{i=1}^n (\mu_i - \overline{\mu})\left(f_-(\mu_i) - f_-(\overline{\mu})\right) \tag{C.44}$$

By the mean-value theorem again, 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 $f'_-(x) = -(x+2B)\varphi(x+2B) \leq -m_B$. Moreover, $\overline{f}_- \leq \varphi(B)$. Thus,

$$-W'(-2B) \leq |\overline{\mu}|\varphi(B) - \sigma^2 m_B < 0.$$

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 C.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

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

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). □

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

Lemma C.5. *In the proof of Theorem 4, we have that Bayes regret is equal to (A.14).*

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

$$\begin{aligned} m_G(Y_i)\,(a^*_G(Y_i) - a_i(Y_{1:n})) &= |m_G(Y_i)|\,(1 - a_i(Y_{1:n})) \\ &= |m_G(Y_i)|\,\mathbf{1}\left\{a^*_G(Y_i) \neq a_i(Y_{1:n})\right\}. \end{aligned}$$

In the other case where $m_G(Y_i) < 0$ so that $a^*_G(Y_i) = 0$, then

$$\begin{aligned} m_G(Y_i)\,(a^*_G(Y_i) - a_i(Y_{1:n})) &= -m_G(Y_i)a_i(Y_{1:n}) \\ &= |m_G(Y_i)|\,a_i(Y_{1:n}) \\ &= |m_G(Y_i)|\,\mathbf{1}\left\{a^*_G(Y_i) \neq a_i(Y_{1:n})\right\}. \end{aligned}$$

□

Lemma C.6. *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,

$$\mathrm{TV}\left(p_{G_0}^{(-j)}, p_{G_1}^{(-j)}\right) \le \sqrt{\frac{n-1}{2}\mathrm{KL}\left(p_{G_0}, p_{G_1}\right)}.$$

By the data processing inequality,

$$\mathrm{KL}\left(p_{G_0}, p_{G_1}\right) \le \mathrm{KL}\left(\mathrm{Bern}\left(\frac{1}{2}-\frac{1}{4\sqrt{n}}\right), \mathrm{Bern}\left(\frac{1}{2}+\frac{1}{4\sqrt{n}}\right)\right).$$

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

$$\alpha_{n-1} \ge 1-\sqrt{\frac{n-1}{2n}} \ge 1-\frac{\sqrt{2}}{2} = \alpha.$$

□

Lemma C.7. *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 $a_i(Y_{1:n}) = \mathbf{1}\{Y_i > \delta_i(Y_{-i})\}$,

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

for all sufficiently large n.

Proof. Let c_i^* denote the zero of the posterior mean $m_{G_i}(y)$, and define

$$w(y) := \min\left(\left|m_{G_0}(y)\right|, \left|m_{G_1}(y)\right|\right), \qquad q(\mathbf{y}) := \min\left(p_{G_0}^{\otimes n}(\mathbf{y}), p_{G_1}^{\otimes n}(\mathbf{y})\right).$$

Since $c_1^* < c_0^*$ and $a_i(Y_{1:n})$ is monotone in Y_i, for each fixed Y_{-i} we have

$$\mathbf{1}\left\{a_{G_0}^*(Y_i) \ne a_i(Y_{1:n})\right\} + \mathbf{1}\left\{a_{G_1}^*(Y_i) \ne a_i(Y_{1:n})\right\} \ge \mathbf{1}\left\{c_1^* < Y_i < c_0^*\right\}.$$

Hence, by (A.14),

$$\begin{aligned} R(G_0, a_{1:n}) + R(G_1, a_{1:n}) &\ge \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}^n} w(y_i)\mathbf{1}\left\{c_1^* < y_i < c_0^*\right\} q(\mathbf{y})\, d\mathbf{y} \\ &\ge \alpha_{n-1}\int_{c_1^*}^{c_0^*} w(y)\min\left(p_{G_0}(y), p_{G_1}(y)\right)\, dy, \end{aligned}$$

where α_{n-1} is defined in Lemma C.6. By Lemma C.6, $\alpha_{n-1} \geq \alpha := 1 - \sqrt{2}/2$ 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 $y \in [-c_n, c_n]$,

$$|m_{G_0}(y)| = \tanh(c_n - y), \qquad |m_{G_1}(y)| = \tanh(y + c_n).$$

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 subsection, we fix $\delta_1, \ldots, \delta_n$, 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.8. *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.9. *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.8 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,

$$\begin{aligned}
&\mathbb{E}\left[\left|\hat{w}_{h,i}(\sqrt{m}\bar{Y}_i;\delta) - w_{h,i}(\sqrt{m}\bar{Y}_i;\delta)\right|\right] \\
&\leq C\mathbb{E}\left[\left(\frac{1}{h} + \frac{|\sqrt{m}\bar{Y}_i|}{\sigma_i}\right)|\hat{\sigma}_i - \sigma_i|\right] + C\mathbb{E}\left[\left(\frac{1}{h} + |\sqrt{m}\bar{Y}_i|\right)\mathbf{1}(A_i^c)\right] \\
&\leq C\left(\frac{1}{h}\sqrt{\mathbb{E}(\hat{\sigma}_i - \sigma_i)^2} + \frac{\sqrt{m}}{\sigma_i}\sqrt{\mathbb{E}(\bar{Y}_i^2)\mathbb{E}(\hat{\sigma}_i - \sigma_i)^2}\right) \\
&\quad + C\left(\frac{1}{h}\mathbb{P}(A_i^c) + \sqrt{m}\sqrt{\mathbb{E}(\bar{Y}_i^2)\mathbb{P}(A_i^c)}\right).
\end{aligned}$$

Now

$$\mathbb{E}(\hat{\sigma}_i - \sigma_i)^2 \leq \frac{\mathrm{Var}(\hat{\sigma}_i^2)}{\sigma_i^2}, \qquad \mathbb{E}(\bar{Y}_i^2) = \frac{\mu_i^2 + \sigma_i^2}{m},$$

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.10. *Under the assumptions of Theorem 5,*

$$\frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}\left[w_{h,i}(\sqrt{m}\bar{Y}_i;\delta_i) - w_{h,i}(Z_i;\delta_i)\right]\right| \leq C\frac{(\log n)^2}{\sqrt{m}}.$$

Proof. Fix i and write $\delta = \delta_i$. Relabel the micro-data Y_{ij} as $Y_1, \ldots, Y_m$, and let $Z_1, \ldots, Z_m$ 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 $k = 0, \ldots, m$, and

$$S_k' := \frac{1}{\sqrt{m}}\left(Z_1 + \cdots + Z_k + Y_{k+2} + \cdots + Y_m\right),$$

for $k = 0, \ldots, m-1$. 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

$$|\mathbb{E}w_{h,i}(S_{k+1};\delta) - \mathbb{E}w_{h,i}(S_k;\delta)| \lesssim \frac{1}{m^{3/2}}\left(\mathbb{E}\left[|Y_{k+1}|^3 |w^{(3)}_{h,i}(\tilde{x})|\right] + \mathbb{E}\left[|Z_{k+1}|^3 |w^{(3)}_{h,i}(\tilde{x}')|\right]\right),$$

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

$$w^{(3)}_{h,i}(x;\delta) = \frac{3}{(\sigma_i h)^2}\operatorname{sinc}'\left(\frac{x-\delta}{\sigma_i h}\right) + \frac{x}{(\sigma_i h)^3}\operatorname{sinc}''\left(\frac{x-\delta}{\sigma_i h}\right) - \frac{1}{\sigma_i^2 h^4}\operatorname{sinc}'''\left(\frac{x-\delta}{\sigma_i h}\right),$$

so the uniform boundedness of the derivatives of sinc yields

$$|w^{(3)}_{h,i}(x;\delta)| \lesssim \frac{1}{\sigma_i^2 h^4} + \frac{|x|}{\sigma_i^3 h^3}.$$

Also,

$$|\tilde{x}| \le \frac{|Y_{k+1}|}{\sqrt{m}} + |S'_k|, \qquad |\tilde{x}'| \le \frac{|Z_{k+1}|}{\sqrt{m}} + |S'_k|,$$

and

$$\mathbb{E}|S'_k| \le \sqrt{\mathbb{E}[(S'_k)^2]} \lesssim \sqrt{\mu_i^2 + \sigma_i^2}.$$

Finally,

$$\mathbb{E}|Y_{k+1}|^3 \lesssim \frac{|\mu_i|^3}{m^{3/2}} + \mu_{3,i}, \qquad \mathbb{E}[Y^4_{k+1}] \lesssim \frac{\mu_i^4}{m^2} + \mu_{4,i},$$

and the Gaussian analogues satisfy

$$\mathbb{E}|Z_{k+1}|^3 \lesssim \frac{|\mu_i|^3}{m^{3/2}} + \sigma_i^3, \qquad \mathbb{E}[Z^4_{k+1}] \lesssim \frac{\mu_i^4}{m^2} + \sigma_i^4.$$

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.11. *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

$$\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[w_{h,i}(Z_i;\delta_i)\right] - W(\delta)\right| \leq \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[w_{h,i}(Z_i;\delta_i)\right] - W_G(\delta)\right| + |W_G(\delta) - W(\delta)|.$$

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. Results for ASSURE-CB

In this section, we derive the counterpart of Corollary 1 for ASSURE-CB to corroborate the rate comparison in Section 2.2. Define the ASSURE-CB estimator as

$$\hat{W}_{\mathrm{CB}}(\beta;\epsilon) = \frac{1}{n}\sum_{i=1}^{n}\left[(Y_i - k_i)\cdot\Phi\left(\frac{Y_i - \delta(z_i;\beta)}{\epsilon\sigma_i}\right) - \frac{\sigma_i}{\epsilon}\cdot\varphi\left(\frac{Y_i - \delta(z_i;\beta)}{\epsilon\sigma_i}\right)\right].$$

By arguments in Section 2.2, $\hat{W}_{\mathrm{CB}}(\beta;\epsilon)$ is an unbiased estimator for a welfare function

$$W(\beta;\epsilon) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\mu_i - k_i)\,\mathbf{1}\left\{Y_i - \epsilon\sigma_i Q_i \geq \delta(z_i;\beta)\right\}\right] = \mathbb{E}\left[V_n(\beta;\epsilon)\right],$$

where $Q_i \sim \mathcal{N}(0,1)$ and $Q_i \perp\!\!\!\perp Y_i, z_i$. Define the welfare function of our interest.

$$W(\beta;0) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\mu_i - k_i)\,\mathbf{1}\left\{Y_i \geq \delta(z_i;\beta)\right\}\right] = \mathbb{E}\left[V_n(\beta;0)\right].$$

Theorem D.1 (Main regret bound for ASSURE-CB)**.** *Let $\hat{\beta}(\epsilon)$ be the maximizer of $\hat{W}_{CB}(\beta;\epsilon)$, $\beta^*(\epsilon)$ be the maximizer of $W(\beta;\epsilon)$ and β_0 be the maximizer of $W(\beta;0)$. Under same conditions of Corollary 1, with $\epsilon_n \to 0$ the regret satisfies*

$$W(\beta_0;0) - \mathbb{E}\left[V_n(\hat{\beta}(\epsilon_n);0)\right] = O(\gamma(n)),$$

where

$$\gamma(n) = \max\left\{\epsilon_n^2, \frac{1}{\sqrt{n}}, \frac{\epsilon_n^{-\frac{1}{2}}(-\log\epsilon_n)^{\frac{1}{2}}}{\sqrt{n}}, \frac{\epsilon_n^{-1}(-\log\epsilon_n)}{n}\right\}.$$

Therefore the optimal choice of ϵ_n is $\epsilon_n = n^{-\frac{1}{5}}$, and then $\gamma(n) = n^{-\frac{2}{5}}(\log n)^{\frac{1}{2}}$.

Proof of Theorem D.1. The regret can be bounded by

$$\begin{aligned}
W(\beta_0;0) - \mathbb{E}\left[V_n(\hat{\beta}(\epsilon_n);0)\right] &= \mathbb{E}\left[V_n(\beta_0;0) - V_n(\hat{\beta}(\epsilon_n);0)\right] \\
&\le \mathbb{E}\left[V_n(\beta_0;0) - \hat{W}_{\mathrm{CB}}(\beta_0;\epsilon_n)\right] + \mathbb{E}\left[\hat{W}_{\mathrm{CB}}(\hat{\beta}(\epsilon_n);\epsilon_n) - V_n(\hat{\beta}(\epsilon_n);0)\right] \\
&\le 2\mathbb{E}\left[\sup_\beta\left|\hat{W}_{\mathrm{CB}}(\beta;\epsilon_n) - V_n(\beta;0)\right|\right] \\
&\le 2\mathbb{E}\left[\sup_\beta\left|\hat{W}_{\mathrm{CB}}(\beta;\epsilon_n) - W(\beta;\epsilon_n)\right|\right] \\
&\quad + 2\mathbb{E}\left[\sup_\beta\left|V_n(\beta;0) - \mathbb{E}\left[V_n(\beta;0)\right]\right|\right] \\
&\quad + 2\sup_\beta\left|W(\beta;\epsilon_n) - W(\beta;0)\right|.
\end{aligned}$$

With $\epsilon_n \to 0$, from Proposition D.1, we have

$$\sup_\beta\left|W(\beta;\epsilon_n) - W(\beta;0)\right| = O(\epsilon_n^2).$$

From Proposition D.2, we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_\beta\left|\hat{W}_{\mathrm{CB}}(\beta;\epsilon_n) - W(\beta;\epsilon_n)\right|\right] + \mathbb{E}\left[\sup_\beta\left|V_n(\beta;0) - \mathbb{E}\left[V_n(\beta;0)\right]\right|\right] \\
&= \max\left\{O\left(\frac{1}{\sqrt{n}}\right), O\left(\frac{\epsilon_n^{-\frac{1}{2}}(-\log\epsilon_n)^{\frac{1}{2}}}{\sqrt{n}}\right), O\left(\frac{\epsilon_n^{-1}(-\log\epsilon_n)}{n}\right)\right\}.
\end{aligned}$$

Combine the two results and we conclude the proof. □

Proposition D.1. *Under the same assumption of Corollary 1, if we let $\epsilon_n \to 0$, then we have*

$$\sup_{\beta} |W(\beta;\epsilon_n) - W(\beta;0)| = \|D\|_n O(\epsilon_n^2).$$

Proof. For any ϵ, since $Y_i - \epsilon\sigma_i\omega_i \mid z_i \sim \mathcal{N}(\mu_i, (1+\epsilon^2)\,\sigma_i^2)$,

$$W(\beta;\epsilon) = \frac{1}{n}\sum_{i=1}^{n} (\mu_i - k_i)\,\Phi\left(\frac{\mu_i - \delta(z_i;\beta)}{\sqrt{1+\epsilon^2}\sigma_i}\right).$$

By the mean value theorem,

$$\begin{aligned}
&|W(\beta;\epsilon) - W(\beta;0)| = \left|\frac{\partial}{\partial\epsilon} W(\beta;\epsilon)\big|_{\epsilon=\epsilon_1}\right| |\epsilon - 0| \\
&= \left|\frac{1}{n}\sum_{i=1}^{n} (\mu_i - k_i)\,\varphi\left(\frac{\mu_i - \delta(z_i;\beta)}{\sqrt{1+\epsilon_1^2}\sigma_i}\right)\frac{\mu_i - \delta(z_i;\beta)}{\sigma_i}\right| \frac{1}{(1+\epsilon_1^2)^{3/2}}\epsilon_1\epsilon \\
&\lesssim \|D\|_n \frac{1}{(1+\epsilon_1^2)^{3/2}}\epsilon_1\epsilon \le \epsilon^2\|D\|_n.
\end{aligned}$$

□

Proposition D.2. *Under the same assumption as Corollary 1, if we let $\epsilon_n \to 0$, then we have*

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\beta}\left|\hat{W}_{CB}(\beta;\epsilon_n) - W(\beta;\epsilon_n)\right|\right] + \mathbb{E}\left[\sup_{\beta}|V_n(\beta;0) - \mathbb{E}\left[V_n(\beta;0)\right]|\right] \\
&= \max\left\{O\left(\frac{1}{\sqrt{n}}\right), O\left(\frac{\epsilon_n^{-\frac{1}{2}}(-\log\epsilon_n)^{\frac{1}{2}}}{\sqrt{n}}\right), O\left(\frac{\epsilon_n^{-1}(-\log\epsilon_n)}{n}\right)\right\}.
\end{aligned}$$

Proof. Define $\mathcal{D} = \{\delta(z;\beta)\}$ and $V(\mathcal{D})$ as its VC-dimension. With ϵ_n, first we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\beta}\left|\hat{W}_{\mathrm{CB}}(\beta;\epsilon_n) - W(\beta;\epsilon_n)\right|\right] \\
&= \mathbb{E}\left[\sup_{\beta}\left|\hat{W}_{\mathrm{CB}}(\beta;\epsilon_n) - \mathbb{E}\left[\hat{W}_{\mathrm{CB}}(\beta;\epsilon_n)\right]\right|\right] \\
&\le \mathbb{E}\left[\sup_{\beta}\left|\frac{1}{n}\sum_{i=1}^{n}(Y_i - k_i)\,\Phi\left(\frac{Y_i - \delta(z_i;\beta)}{\epsilon_n\sigma_i}\right) - \mathbb{E}\left[(Y_i - k_i)\,\Phi\left(\frac{Y_i - \delta(z_i;\beta)}{\epsilon_n\sigma_i}\right)\right]\right|\right] \\
&\quad + \frac{1}{\epsilon_n}\mathbb{E}\left[\sup_{\beta}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\varphi\left(\frac{Y_i - \delta(z_i;\beta)}{\epsilon_n\sigma_i}\right) - \mathbb{E}\left[\sigma_i\varphi\left(\frac{Y_i - \delta(z_i;\beta)}{\epsilon_n\sigma_i}\right)\right]\right|\right] \\
&= T_{1,n} + \frac{1}{\epsilon_n}T_{2,n},
\end{aligned}$$

and

$$\mathbb{E}\left[\sup_{\beta}\left|V_n(\beta;0)-W(\beta;0)\right|\right]=\mathbb{E}\left[\sup_{\beta}\left|V_n(\beta;0)-\mathbb{E}\left[V_n(\beta;0)\right]\right|\right]=T_{3,n}.$$

It suffices to bound $T_{1,n}$, $\frac{1}{\epsilon_n}T_{2,n}$ and $T_{3,n}$. By the symmetrization lemma for the i.n.i.d case, we derive the bound as follows. Conditional on Y,Z, denote the empirical measure of as $\mathbb{P}_n$. First notice that the function class below is a VC-subgraph class:

$$\mathcal{F}_1=\left\{f_\beta(y,z)=(y-k)\,\Phi\left(\frac{y-\delta(z;\beta)}{\epsilon_n\sigma}\right)\right\}.$$

The result is from the following. First, $\left\{\frac{y-\delta(z;\beta)}{\epsilon_n\sigma}\right\}$ is a VC-subgraph class the VC dimension $V(\mathcal{D})$ by Lemma 2.6.18.vi of van der Vaart and Wellner (1996) from multiplicity. Since $t\mapsto\Phi(t)$ is a monotone function, by Lemma 2.6.18.viii of van der Vaart and Wellner (1996) and by Lemma 2.6.18.vi of van der Vaart and Wellner (1996) with multiplicity, $\mathcal{F}_1$ is a VC-subgraph class with the VC dimension $V(\mathcal{D})$.

Since $\mathcal{F}_1$ has an envelope function: $|f_\beta(y,z)|\leq|y-k|$, from Theorem 2.6.7 of van der Vaart and Wellner (1996), we have for the covering number

$$N\left(\eta\|y-k\|_{L_2,\mathbb{P}_n},\mathcal{F}_1,L_2(\mathbb{P}_n)\right)\leq KV(\mathcal{D})\left(16e\right)^{V(\mathcal{D})}\left(\frac{1}{\eta}\right)^{2(V(\mathcal{D})-1)},$$

where K is a universal constant. Applying the maximal inequality similar to Theorem 2.14.1 of van der Vaart and Wellner (2023), we obtain the following:

$$\begin{aligned}
T_{1,n}&\lesssim\mathbb{E}\left[\frac{1}{\sqrt{n}}\|y-k\|_{L_2,\mathbb{P}_n}\int_0^1\sqrt{1+\log N\left(\eta\|y-k\|_{L_2,\mathbb{P}_n},\mathcal{F}_1,L_2\left(\mathbb{P}_n\right)\right)}d\eta\right]\\
&\leq\mathbb{E}\left[\frac{1}{\sqrt{n}}\|y-k\|_{L_2,\mathbb{P}_n}\int_0^1\left(1+\log\left(A_1\left(\frac{1}{\eta}\right)^{A_2}\right)\right)^{\frac{1}{2}}d\eta\right]\\
&\leq A_3\frac{1}{\sqrt{n}}\mathbb{E}\left[\|y-k\|_{L_2,\mathbb{P}_n}\right]\\
&\leq A_3\frac{1}{\sqrt{n}}\left(\mathbb{E}\left[\|y-\mu\|_{L_2,\mathbb{P}_n}\right]+\|\mu-k\|_{L_2,\mathbb{P}_n}\right)\\
&\leq A_3\frac{1}{\sqrt{n}}\left(\sqrt{\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left(Y_i-\mu_i\right)^2\right]}+m_2\right)=A_3\frac{1}{\sqrt{n}}\left(s_2+m_2\right)
\end{aligned}$$

where s_q,m_q are defined before Theorem 1, and A_1,A_2,A_3 are universal constants.

Similarly, first notice that the function class below is a VC-subgraph class:

$$\mathcal{F}_2 = \left\{ f_\beta(y,z) = \sigma\varphi\left(\frac{y-\delta(z;\beta)}{\epsilon_n\sigma}\right)\right\}.$$

The result is from the following. First, $\left\{\left|\frac{y-\delta(z;\beta)}{\epsilon_n\sigma}\right|\right\}$ is a VC subgraph class with dimension $V(\mathcal{D})$ by multiplicity (Lemma 2.6.18.vi of van der Vaart and Wellner (1996)) and part 5 of Lemma E.1. Since $t \mapsto t^2$ is a monotone mapping for non-negative t, and $t \mapsto e^{-t}$ is a monotone mapping, by Lemma 2.6.18.viii of van der Vaart and Wellner (1996) and by Lemma 2.6.18.vi of van der Vaart and Wellner (1996) with multiplicity, $\mathcal{F}_2$ is a VC-subgraph class with dimension $V(\mathcal{D})$.

Since $\mathcal{F}_2$ is bounded $|f_\beta(y,z)| \leq M$, from Theorem 2.6.7 of van der Vaart and Wellner (1996), we have for the covering number

$$N(\eta M, \mathcal{F}_2, L_2(\mathbb{P}_n)) \leq KV(\mathcal{D})(16e)^{V(\mathcal{D})}\left(\frac{1}{\eta}\right)^{2(V(\mathcal{D})-1)},$$

where K is a universal constant. For the entropy defined in Appendix E,

$$J(b, \mathcal{F}_2 \mid M, L_2) = \sup_Q \int_0^b \sqrt{1+\log N(\eta M, \mathcal{F}_2, L_2(Q))}d\eta,$$

we have $J(b, \mathcal{F}_2 \mid M, L_2) = O(b\sqrt{\log(1/b)})$ as $b \downarrow 0$ (van der Vaart and Wellner, 2023, before Theorem 2.14.1, p. 330)

Since Lemma D.1 shows that $\mathbb{E}f^2 \leq b_n^2 M^2$, with $b_n = \epsilon_n^{1/2}$ for $\forall f \in \mathcal{F}_2$, applying Theorem 2.14.2 in van der Vaart and Wellner (2023), we have

$$\begin{aligned} T_{2,n} &\lesssim \frac{1}{\sqrt{n}} J(b_n, \mathcal{F}_2 \mid M, L_2)\left(1 + \frac{J(b_n, \mathcal{F}_2 \mid M, L_2)}{b_n^2\sqrt{n}M}\right)M \\ &\lesssim \frac{1}{\sqrt{n}} b_n \sqrt{\log\frac{1}{b_n}} + \frac{1}{n}\log\left(\frac{1}{b_n}\right) \lesssim \max\left\{\frac{\epsilon_n^{\frac{1}{2}}\sqrt{(-\log\epsilon_n)}}{\sqrt{n}}, \frac{(-\log\epsilon_n)}{n}\right\}. \end{aligned}$$

For $T_{3,n}$, define the function class

$$\mathcal{F}_3 = \{f_\beta(y,z,\mu) = (\mu - k)\mathbf{1}\{y \geq \delta(z;\beta)\}\}.$$

By the additive, multiplicative and monotone composition properties of VC-subgraph classes (van der Vaart and Wellner, 1996, Lemma 2.6.18), $\mathcal{F}_3$ is a VC-subgraph class with dimension $V(\mathcal{D})$. Since $\mathcal{F}_3$ has an envelope function: $|f_\beta(y,z,\mu)| \leq |\mu - k|$, from Theorem 2.6.7 of van der Vaart and Wellner (1996) and the maximal inequality similar

to Theorem 2.14.1 of van der Vaart and Wellner (2023), we have

$$T_{3,n} \lesssim \mathbb{E}\left[\frac{1}{\sqrt{n}}\|\mu-k\|_{L_2,\mathbb{P}_n}\int_0^1 \sqrt{1+\log N\left(\eta\|\mu-k\|_{L_2,\mathbb{P}_n},\mathcal{F}_3,L_2\left(\mathbb{P}_n\right)\right)}d\eta\right]$$
$$\leq \mathbb{E}\left[\frac{1}{\sqrt{n}}m_2\int_0^1\left(1+\log\left(A_1\left(\frac{1}{\eta}\right)^{A_2}\right)\right)^{\frac{1}{2}}d\eta\right] \leq A_3\frac{1}{\sqrt{n}}m_2,$$

where s_q, m_q are defined before Theorem 1, and A_1, A_2, A_3 are universal constants.

As a conclusion, we have the bound:

$$T_{1,n}+\frac{1}{\epsilon_n}T_{2,n}+T_{3,n} \lesssim \frac{1}{\sqrt{n}}+\frac{1}{\epsilon_n}\max\left\{\frac{\epsilon_n^{\frac{1}{2}}\sqrt{(-\log\epsilon_n)}}{\sqrt{n}},\frac{(-\log\epsilon_n)}{n}\right\}+\frac{1}{\sqrt{n}}$$
$$\lesssim \max\left\{O\left(\frac{1}{\sqrt{n}}\right),O\left(\frac{\epsilon_n^{-\frac{1}{2}}(-\log\epsilon_n)^{\frac{1}{2}}}{\sqrt{n}}\right),O\left(\frac{\epsilon_n^{-1}(-\log\epsilon_n)}{n}\right)\right\}.$$

□

Lemma D.1. *Under the same assumption as Corollary 1, for $\epsilon_n \to 0$, we have*

$$\sup_\beta \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \varphi^2\left(\frac{Y_i-\delta(z;\beta)}{\epsilon_n\sigma_i}\right)\right] \leq \frac{1}{2\pi\sqrt{2}}\epsilon_n.$$

Proof. We have

$$\mathbb{E}\left[\varphi^2\left(\frac{Y_i-\delta(z_i;\beta)}{\epsilon_n\sigma_i}\right)\right] = \frac{1}{\sqrt{2\pi}}\mathbb{E}\left[\varphi\left(\sqrt{2}\frac{Y_i-\delta\left(z_i;\beta\right)}{\epsilon_n\sigma_i}\right)\right]$$
$$= \frac{1}{\sqrt{2\pi}}\int\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma_i}\exp\left\{-\frac{1}{2\sigma_i^2}\left[2\frac{\left(y-\delta(z_i;\beta)\right)^2}{\epsilon_n^2}+\left(y-\mu_i\right)^2\right]\right\}dy$$
$$= \frac{1}{\sqrt{2\pi}}\int\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma_i}\exp\left\{-\frac{1}{2\sigma_i^2}\left[\left(1+\frac{2}{\epsilon_n^2}\right)\left(y-\frac{2}{2+\epsilon_n^2}\delta(z_i;\beta)-\frac{\epsilon_n^2}{2+\epsilon_n^2}\mu_i\right)^2\right.\right.$$
$$\left.\left.+\frac{2}{2+\epsilon_n^2}\left(\delta(z_i;\beta)-\mu_i\right)^2\right]\right\}dy$$
$$= \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{1+\frac{2}{\epsilon_n^2}}}\varphi\left(\frac{\sqrt{2}}{\sigma_i\sqrt{2+\epsilon_n^2}}\left(\delta(z_i;\beta)-\mu_i\right)\right) \leq \frac{1}{2\pi\sqrt{2}}\epsilon_n, \forall\beta.$$

□

Appendix E. 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 E.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 E.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 E.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

$$\left\| \|\mathbb{G}_n^o\|_{\mathcal{F}} \right\|_{\psi_2 | V} \lesssim \int_0^{\eta_n} \sqrt{1 + \log N(\epsilon, \mathcal{F}, L_2(\mathbb{P}_n))} d\epsilon$$

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. □

Proposition E.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*

$$\left\| \|\mathbb{G}_n\|_{\mathcal{F}}^* \right\|_{\psi_2} \lesssim \mathbb{E}^* \|\mathbb{G}_n\|_{\mathcal{F}} + \left(\frac{1}{n} \sum_{i=1}^n \|F(V_i)\|_{\psi_2}^2 \right)^{1/2}$$

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 $U_i(f) = \frac{1}{\sqrt{n}} \left(f(V_i) - \mathbb{E} f(V_i) \right)$, indexed by $f \in \mathcal{F}$ so that $\|S_n\|^* = \|\mathbb{G}_n\|_{\mathcal{F}}$. Plugging this into the proposition, we obtain

$$\left\| \|\mathbb{G}_n\|_{\mathcal{F}}^* \right\|_{\psi_2} \lesssim E^* \|\mathbb{G}_n\|_{\mathcal{F}} + n^{-1/2} \left(\sum_{i=1}^n \left\| \sup_{f \in \mathcal{F}} |f(V_i) - \mathbb{E}[f(V_i)]| \right\|_{\psi_2}^2 \right)^{1/2} \tag{E.1}$$

$$\leq E^* \|\mathbb{G}_n\|_{\mathcal{F}} + n^{-1/2} \left(\sum_{i=1}^n \|F(V_i) + \mathbb{E}[F(V_i)]\|_{\psi_2}^2 \right)^{1/2} \tag{E.2}$$

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

$$E^* \|\mathbb{G}_n\|_{\mathcal{F}} + n^{-1/2} \left(\sum_{i=1}^n \|F(V_i)\|_{\psi_2}^2 \right)^{1/2},$$

from which the desired claim holds. □

Lemma E.3. *For any constant random variable C, $\|C\|_{\psi_2} \leq |C|/\sqrt{\ln 2}$. For any random variable X, $\|X\|_{\psi_2} \geq \mathbb{E}|X|$.*

Proof. Recall that for any random variable X,

$$\|X\|_{\psi_2} = \inf\left\{c : \mathbb{E}\psi_2\left(\frac{|X|}{c}\right) \leq 1\right\}.$$

For the first claim, if $C = 0$ there is nothing to prove. Otherwise, taking $c = |C|/\sqrt{\ln 2}$, we see that $\psi_2(|C|/c) \leq \psi_2(\sqrt{\ln 2}) = 1$. For the second claim, notice that for any c such that $\mathbb{E}\psi_2\left(\frac{|X|}{c}\right) \leq 1$, convexity of ψ_2 and Jensen's inequality implies that $\psi_2(\mathbb{E}|X|/c) \leq 1$. The explicit form of ψ_2 implies that $\mathbb{E}|X|/c \leq 1$. Thus $c > \mathbb{E}|X|$ as desired. □

E.1. **Examples of VC classes.**

Lemma E.4. *For fixed finite p, the decision rule classes in (1)–(3) in Section 3.2 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 E.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 E.1(1) shows that $\mathcal{D}_{t\text{-stat}}$ is VC-subgraph.

For the third class, let

$$\mathcal{V} := \text{span}\left\{\delta^{(1)}, \ldots, \delta^{(p)}\right\}.$$

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,

$$\mathcal{D}_{\text{finite}} = \text{conv}\left(\left\{\delta^{(1)}, \ldots, \delta^{(p)}\right\}\right) \subset \mathcal{V}.$$

Any subclass of a VC-subgraph class is again VC-subgraph, so $\mathcal{D}_{\text{finite}}$ is VC-subgraph as well. □

Proposition E.2 (VC-subgraph properties in Section 3.2)**.** *The decision threshold classes in Table 1 are all VC-subgraph.*

Proof of Proposition E.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 E.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:

$$f_\beta(k, \ell) = k + e^{2\ell} e^{-b_1 - b_2\ell}(k - a_1 - a_2 e^\ell)$$

indexed by $\beta = (a_1, a_2, b_1, b_2)$. The subgraph sets

$$S_\beta := \left\{(k, \ell, t) : t < k + e^{2\ell} e^{-b_1 - b_2\ell}(k - a_1 - a_2 e^\ell)\right\}$$

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 F. Other observation distributions: the Poisson case

ASSURE can be extended to settings where the observation distribution is non-Gaussian. For certain exponential families, one can find an exactly unbiased estimate for the welfare W. The idea is to exploit a generalization of Stein's identity, called Hudson's Lemma (Hudson, 1978), which holds for both continuous and discrete exponential families.

We focus on the Poisson setting as it frequently appears in empirical Bayes and compound decisions.[22] In this setting, let $\mu_i \geq 0$ and $Y_i \sim \mathsf{Poi}(\mu_i)$. Associated with each decision problem is a cost k_i and auxiliary information z_i.[23] Consider any integer-valued threshold $\delta(z_i; \beta)$ parametrized by β. Then the estimator

$$\hat{W}_{\mathsf{Poi}}(\beta) := \frac{1}{n} \sum_{i=1}^{n} \left(Y_i \mathbf{1}\left\{ Y_i \geq \delta(z_i;\beta) + 1 \right\} - k_i \mathbf{1}\left\{ Y_i \geq \delta(z_i;\beta) \right\} \right) \tag{F.1}$$

is the analog for the ASSURE estimator:

Proposition F.1. *The Poisson* ASSURE *estimator* $\hat{W}_{\mathsf{Poi}}(\beta)$ *is unbiased for the welfare*

$$W(\beta; \mu_{1:n}) := W(\beta) := \frac{1}{n} \sum_{i=1}^{n} (\mu_i - k_i)\, \mathbb{P}(Y_i \geq \delta(z_i;\beta)). \tag{F.2}$$

This follows as a direct consequence of Hudson's Lemma, so the proof is omitted. Under mild conditions, ASSURE achieves $1/\sqrt{n}$ regret, which is optimal. In contrast to the Gaussian case, the discreteness of the Poisson distribution suggests that no fast-rate regime is available.

Theorem F.1 (Regret upper bounds for the Poisson Case)**.** *Let* $m_q := \left(\frac{1}{n}\sum_{i=1}^{n} \mu_i^q\right)^{1/q}$ *for* $q \geq 1$*. Suppose the class of decision rules* $\mathcal{D}$ *is a VC subgraph class with index* $V(\mathcal{D})$*. Then* ASSURE *for Poisson observations has regret at most an absolute constant times*

$$\frac{1}{\sqrt{n}} \sqrt{V(\mathcal{D})}(m_2 + \sqrt{m_1}).$$

We prove the upper bound by appealing to the similar empirical process theory arguments as in the Gaussian case. We suspect some of the assumptions of Theorem F.1 may be further relaxed. This rate is the best possible, as the following lower bound shows.

Theorem F.2 (Matching Regret Lower Bounds)**.** *Fix costs* $k_i = k > 0$*, and let* Θ *be a compact interval containing a neighborhood of* k*. We have*

$$\inf_{a(\cdot):a_i \in \{0,1\}} \sup_{\mu_{1:n} \in \Theta^n} \mathsf{Regret}_n(a_{1:n}) = \Omega(n^{-1/2}),$$

where $\mathsf{Regret}_n(a_{1:n})$ *is defined analogously to* (3.1) *except with* μ_i *replaced with* $\mu_i - k$*.*

[22] See Chapter 6 of Efron and Hastie (2021) for an overview of classic applications of the Poisson model to insurance claims, the missing species problems, and medical applications. See Jana *et al.* (2022, 2023) for modern methods to this problem and also Montiel Olea *et al.* (2021) for econometric applications.

[23] For example, analogous to σ_i in the Gaussian case, z_i might include the known length t_i of the observation window for the Poisson outcome Y_i. If the unknown parameter μ_i denotes the Poisson rate per unit time, then $Y_i \sim \mathsf{Poi}(\mu_i t_i)$. The decision rule might incorporate the observation window t_i.

Remark 1. Consider the case $k_i = k$ and the threshold class of decisions $Y_i \geq \delta, \delta \in \mathbb{R}$. An alternative empirical Bayes approach to the Poisson selection problem is to threshold the *Robbins estimator* given by

$$\hat{\mu}(y) := (y+1)\frac{N_{y+1}}{N_y} \tag{F.3}$$

where $N_y := \#\{i : Y_i = y\}$. The estimator can also be used for selection by selecting all indices i for which the Robbins estimator is greater than k. By rearranging $\hat{\mu}(Y_i) \geq k$, one can see that truncating the Robbins estimator bears a resemblance to the ASSURE decision. However, truncating the Robbins estimator does not respect monotonicity. The estimator (F.1) regularizes and enforces this monotonicity, giving a more stable and intuitive alternative to decisions based on the Robbins estimator. Related issues of the Robbins estimator have been discussed in the literature (Brown *et al.*, 2013; Jana *et al.*, 2022).

F.1. **Proofs for Poisson extension.**

Proof of Theorem F.1. Define $V(\beta) = \frac{1}{n}\sum_{i=1}^n (\mu_i - k_i)\,\mathbf{1}\{Y_i \geq \delta(z_i;\beta)\}$ be the in-sample welfare. The same decomposition as in Theorem 1 shows that the regret is bounded as

$$W(\beta^*) - \mathbb{E}[V(\hat{\beta})] \leq 2\mathbb{E}[\sup_\beta |V(\beta) - \hat{W}(\beta)|].$$

Thus it suffices to control $\frac{1}{\sqrt{n}}\mathbb{E}\sup_\beta \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n f_\beta(Y_i,\mu_i,z_i)\right|$ where

$$f_\beta(y,\mu,z) = y\mathbf{1}\{y \geq \delta(z;\beta)+1\} - \mu\mathbf{1}\{y \geq \delta(z;\beta)\}. \tag{F.4}$$

Equivalently, $f_\beta(y,\mu,z) = (y-\mu)\mathbf{1}\{y \geq \delta(z;\beta)+1\} - \mu\mathbf{1}\{y = \delta(z;\beta)\}$. Note that $\mathbb{E}f_\beta(Y_i,\mu_i,z_i) = 0$. Let $\mathcal{F}$ denote this function class, indexed by β, with envelope function $F := y + \mu$. By the argument of Theorem E.1 with $V_i = (Y_i,\mu_i,z_i)$, we may apply symmetrization and a maximal inequality to obtain

$$\mathbb{E}\sup_\beta \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n f_\beta(Y_i,\mu_i,z_i)\right| \leq \mathbb{E}\left[\int_0^{\eta_n} \sqrt{1+\log N(\epsilon,\mathcal{F},\mathbb{L}_2(\mathbb{P}_n))}d\epsilon\right], \tag{F.5}$$

with $\eta_n := \sup_\beta \|f_\beta\|_n$. Next, let $\mathcal{F}_1$ be the function class indexed by β consisting of the functions $y\mathbf{1}\{y \geq \delta(z;\beta)+1\}$ and $\mathcal{F}_2$ be the function class consisting of $\mu\mathbf{1}\{y \geq \delta(z;\beta)\}$. By Lemma E.1 these are VC subgraph classes with indices $O(V(\mathcal{D}))$. These have envelope functions $F_1 := y$ and $F_2 := \mu$, respectively. Next, note that

$$\log N(\epsilon,\mathcal{F},\mathbb{L}_2(\mathbb{P}_n)) \leq \log N(\epsilon,\mathcal{F}_1,\mathbb{L}_2(\mathbb{P}_n)) + \log N(\epsilon,\mathcal{F}_2,\mathbb{L}_2(\mathbb{P}_n))$$

Using this and a change of variables, we arrive at

$$\begin{aligned}\int_0^{\eta_n} \sqrt{1+\log N(\epsilon,\mathcal{F},\mathbb{L}_2(\mathbb{P}_n))}d\epsilon &\lesssim \|F_1\|_n \int_0^{\frac{\|F\|_n}{\|F_1\|_n}} \sqrt{1+\log N(\epsilon\|F_1\|_n,\mathcal{F}_1,\mathbb{L}_2(\mathbb{P}_n))}d\epsilon \\ &\quad + \|F_2\|_n \int_0^{\frac{\|F\|_n}{\|F_2\|_n}} \sqrt{1+\log N(\epsilon\|F_2\|_n,\mathcal{F}_2,\mathbb{L}_2(\mathbb{P}_n))}d\epsilon \\ &\lesssim \|F\|_n + \|F_1\|_n J(1,\mathcal{F}_1 \mid F_1,\mathbb{L}_2) + \|F_2\|_n J(1,\mathcal{F}_2 \mid F_2,\mathbb{L}_2) \\ &\lesssim \|F\|_n + \sqrt{V(\mathcal{D})}\left(\|F_1\|_n + \|F_2\|_n\right).\end{aligned}$$

Next, by Jensen's inequality observe the bounds

$$\mathbb{E}\|F_1\|_n \leq \left(\frac{1}{n}\sum_{i=1}^n \mu_i^2 + \mu_i\right)^{1/2}$$

and $\mathbb{E}\|F\|_n \leq \mathbb{E}\|F_1\|_n + \|F_2\|_n$ with $\|F_2\|_n = \left(\frac{1}{n}\sum_{i=1}^n \mu_i^2\right)^{1/2}$. Combining the pieces, we have

$$\mathbb{E}\sup_\beta \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n f_\beta(Y_i,\mu_i,z_i)\right| \lesssim \sqrt{V(\mathcal{D})}\left(\frac{1}{n}\sum_{i=1}^n \mu_i^2 + \mu_i\right)^{1/2}.$$

This gives the result. □

Proof of Theorem F.2. We can lower bound the compound total by restricting to two sets of parameters $\mu_1,\ldots,\mu_n$ given by $\mu_i = z_0 := k + hn^{-1/2}$ for all i, and $\mu_i = z_1 := k - hn^{-1/2}$ for all $i = 1,\ldots,n$. The parameter h can be chosen later. In the first case, notice that the optimal decision is to select all indices i (take $C_{\text{opt}} = 0$). In the second case, the optimal decision is to select no indices (take $C_{\text{opt}} = \infty$). Then:

$$\begin{aligned}\mathsf{RegretComp}_n(\Theta) \geq \inf_{a_i} \max\Bigg(&hn^{-1/2} - hn^{-1/2}\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{z_0}(a_i(Y_1,\ldots,Y_n)), \\ &hn^{-1/2}\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{z_1}(a_i(Y_1,\ldots,Y_n))\Bigg)\end{aligned}$$

This can be written as

$$\frac{h}{\sqrt{n}}\max\left(\mathbb{E}_{z_0}\left[1 - \frac{1}{n}\sum_{i=1}^n a_i(Y_1,\ldots,Y_n)\right], \mathbb{E}_{z_1}\left[\frac{1}{n}\sum_{i=1}^n a_i(Y_1,\ldots,Y_n)\right]\right)$$

Notice that this quantity is $h/\sqrt{n}$ times the maximum of the Type I and Type II errors of the randomized test which outputs $a_i \in \{0,1\}$ with probability $1/n$. But for any test, the minimum total error for a binary hypothesis test by Le Cam's lemma

is lower bounded by

$$\frac{1}{2}\left\{1 - d_{TV}\left(\mathsf{Poi}(z_1)^{\otimes n}, \mathsf{Poi}(z_0)^{\otimes n}\right)\right\}.$$

Now, applying Pinsker's inequality and tensorization yields

$$\begin{aligned} d_{TV}\left(\mathsf{Poi}(z_1)^{\otimes n}, \mathsf{Poi}(z_0)^{\otimes n}\right) &\leq \frac{1}{\sqrt{2}}\sqrt{d_{KL}\left(\mathsf{Poi}(z_1)^{\otimes n} \middle| \mathsf{Poi}(z_0)^{\otimes n}\right)} \\ &\leq \sqrt{\frac{n}{2}}\sqrt{d_{KL}\left(\mathsf{Poi}(z_1) \middle| \mathsf{Poi}(z_0)\right)} \\ &= \sqrt{n/2}\left[z_1 \ln \frac{z_1}{z_0} - (z_1 - z_0)\right]^{1/2} \end{aligned}$$

But

$$\begin{aligned} z_1 \ln \frac{z_1}{z_0} - (z_1 - z_0) &= z_1 \ln\left(1 + \frac{z_1 - z_0}{z_0}\right) - (z_1 - z_0) \\ &= z_1\left[\left(\frac{z_1 - z_0}{z_0}\right) - \frac{(z_1 - z_0)^2}{2z_0^2} + O\left((z_1 - z_0)^3\right)\right] - (z_1 - z_0) \\ &= \frac{(z_1 - z_0)^2}{z_0} - \frac{z_1(z_1 - z_0)^2}{2z_0^2} + O\left((z_1 - z_0)^3\right) \\ &= \frac{2z_0 - z_1}{2}\frac{(z_1 - z_0)^2}{z_0^2} + O\left((z_1 - z_0)^3\right) \\ &= O\left(\frac{h^2}{n}\right). \end{aligned}$$

Thus

$$d_{TV}\left(\mathsf{Poi}(z_1)^{\otimes n}, \mathsf{Poi}(z_0)^{\otimes n}\right) \leq O(h),$$

where h can be taken small enough so that the TV distance is less than $1/2$. Then the total risk is bounded below by $\Omega(n^{-1/2})$ as desired. □

Appendix G. Additional details for Section 6.1

We start from the replication code in Chen (forthcoming) and follow Chen (forthcoming) 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 OA 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. The main-text discussion emphasizes $\tau = 33$ to mirror the top-third targeting exercise in Bergman *et al.* (2024), but the same implementation also generates a stricter top-10% calibration. For each outcome and each $\tau \in \{10, 33\}$, we first fit the plug-in Fay–Herriot

rule to the observed OA data and denote the resulting posterior means by $\hat{m}_i^{\mathrm{FH}}$. Let

$$m_\tau := \left\lceil \frac{\tau}{100} n \right\rceil .$$

The constant cost used in that scenario is then

$$k^{(\tau)} := \text{the } m_\tau\text{th largest element of } \left\{\hat{m}_i^{\mathrm{FH}}\right\}_{i=1}^n .$$

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 G.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: flexible CLOSE-GAUSS specifications continue to dominate the naive, linear-shrinkage, and simple threshold rules. The gains are especially pronounced for the top-20 probability outcome, where the richer fitted rules substantially improve both mean welfare and normalized welfare relative to plug-in benchmarking.

One coupled-bootstrap draw. 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

$$Y_i^{\mathrm{tr}} = Y_i + \sqrt{w}\,\sigma_i Q_i, \qquad Y_i^{\mathrm{val}} = Y_i - \sqrt{\frac{1}{w}}\,\sigma_i Q_i,$$

with

$$\sigma_i^{\mathrm{tr}} = \sigma_i\sqrt{1+w}, \qquad \sigma_i^{\mathrm{val}} = \sigma_i\sqrt{1+\frac{1}{w}}.$$

The default implementation uses $w = 1/9$, which mimics a 90/10 train/validation split. The fitted rule sees only the training sample $\left\{(Y_i^{\mathrm{tr}}, \sigma_i^{\mathrm{tr}}, x_i, k^{(\tau)})\right\}$; the validation sample is used only for evaluation. The construction is such that Y_i^{tr} and Y_i^{val} are conditionally independent given (μ_i, σ_i, x_i). Thus, conditional on fixed $\mu_{1:n}$ and $x_{1:n}$, Y_i^{val} is an independent noisy proxy for μ_i against which a training-based decision rule can be assessed.

Reported welfare in one draw. If a fitted rule returns thresholds $\delta_i^{(b)}$ on draw b, the reported validation welfare is

$$\hat{W}_b^{\mathrm{val}} := \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left(Y_i^{\mathrm{tr}} > \delta_i^{(b)}\right)\left(Y_i^{\mathrm{val}} - k^{(\tau)}\right).$$

Proposition 1 in Chen (forthcoming) shows that this is unbiased for the utility of the chosen rule.

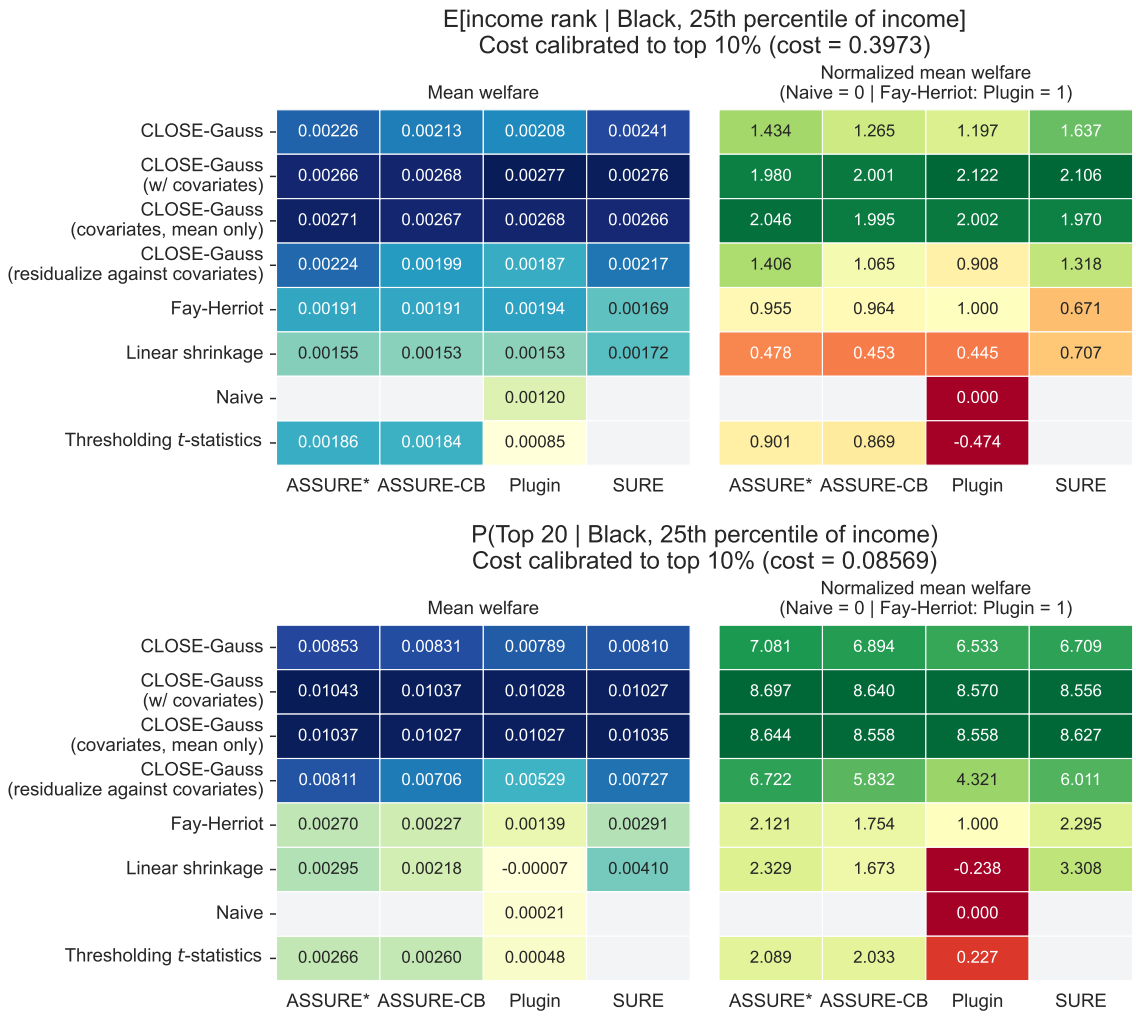


FIGURE G.1. Coupled-bootstrap comparisons for the top-10% OA calibration.

Notes. Same layout and normalization as in 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. □

Rule classes. The coupled-bootstrap OA exercise compares rules in Section 3.2. The naive rule is simply $\mathbf{1}(Y_i > k^{(\tau)})$. The t-statistic class uses thresholds of the form

$$\delta_i = k^{(\tau)} + c\sigma_i,$$

so it reduces to choosing a scalar cutoff c. 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

$$a = \frac{k^{(\tau)} - \mu_0}{\tau^2}, \qquad \delta_i = k^{(\tau)} + a\sigma_i^2.$$

Fay–Herriot uses the conjugate-normal threshold implied by a prior mean $m_{0i} = x_i'\beta$ and a constant signal variance A. The broader CLOSE-GAUSS class uses

$$\delta_i = k_i + \frac{\sigma_i^2}{s_{0i}^2}\left(k_i - m_{0i}\right),$$

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 and regularization choices. 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

$$\delta_i = k^{(\tau)} + 1.96\,\sigma_i,$$

so the conventional two-sided 5% rule is treated as the default member of this class.

- *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. Equivalently, the implied threshold can be written as

$$\delta_i = k^{(\tau)} + \frac{\sigma_i^2}{\hat{\tau}^2_{\text{plugin}}}\left(k^{(\tau)} - \hat{\mu}_{0,\text{plugin}}\right) = k^{(\tau)} + \hat{a}_{\text{plugin}}\sigma_i^2,$$

where

$$\hat{a}_{\text{plugin}} = \frac{k^{(\tau)} - \hat{\mu}_{0,\text{plugin}}}{\hat{\tau}^2_{\text{plugin}}}.$$

This is the empirical-Bayes plug-in associated with the constant-mean, constant-variance Gaussian prior. The only regularization is the negligible positive floor on $\hat{\tau}^2_{\text{plugin}}$, which guarantees a strictly positive fitted signal variance.

• *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

$$\hat{m}_{0i} = z_i'\hat{\beta}, \qquad \hat{s}^2_{0i} = \hat{A}_{\text{plugin}},$$

where $\hat{\beta}$ is the least-squares coefficient on z_i and $\hat{A}_{\text{plugin}}$ is the truncated average residual variance proxy.

• CLOSE-GAUSS *and* CLOSE-GAUSS *with covariates.* These classes use the generic threshold form

$$\delta_i = k_i + \frac{\sigma_i^2}{s^2_{0i}}\left(k_i - m_{0i}\right),$$

with m_{0i} linear in the chosen design z_i and

$$s^2_{0i} = \underline{s}^2 + \exp(z_i'\gamma).$$

Plug-in first estimates the mean side by ordinary least squares,

$$\hat{\beta}_{\text{plugin}} \in \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (Y_i - z_i'\beta)^2.$$

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:

$$\hat{\gamma}_{\text{plugin}} \in \arg\min_{\gamma} \frac{1}{n} \sum_{i=1}^{n} \left(\hat{V}_i - \underline{s}^2 - \exp(z_i'\gamma) \right)^2.$$

Two regularization choices matter here. 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 $(1, \log \sigma_i)$. Thus the mean side is fitted by least squares on the residualized outcome, and the variance side is fitted by the same nonlinear least-squares regression of the residual variance proxy on the log-variance index.

Redundant variance intercept under constant costs. This convention matters already at the level of defining the Gaussian-prior fitters. Under constant costs, the threshold formula depends on the ratio $(k_i - m_{0i})/s_{0i}^2$, so one variance intercept is redundant in the CLOSE-GAUSS parameterization. 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. After obtaining the ASSURE-CB optimum, the implementation also searches along the line segment connecting the plug-in and ASSURE-CB solutions and keeps the point on that segment with the

highest ASSURE* objective if it beats the direct optimizer output. Conversely, the ASSURE-CB fitter itself 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.

Fitters and objective functions. It is clearer to state the fitting problem separately for each rule class. For a generic threshold family $\delta_i(\theta)$, ASSURE* estimates (2.3). The ASSURE-CB fitter chooses

$$\hat{\theta}_{\mathrm{cb}} \in \operatorname*{argmax}_{\theta} \frac{1}{n}\sum_{i=1}^{n}(Y_i - k_i)\Phi\left(\frac{Y_i - \delta_i(\theta)}{\epsilon_n \sigma_i}\right) - \frac{\sigma_i}{\epsilon_n}\phi\left(\frac{Y_i - \delta_i(\theta)}{\epsilon_n \sigma_i}\right),$$

with $\epsilon_n = n^{-0.2}$. For the Gaussian-prior classes, SURE minimizes

$$\mathrm{SURE}(\theta) := \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_i(\theta) - Y_i\right)^2 + \frac{2}{n}\sum_{i=1}^{n}\sigma_i^2 w_i(\theta),$$

where

$$\hat{\mu}_i(\theta) = w_i(\theta)Y_i + (1 - w_i(\theta))\, m_{0i}(\theta), \qquad w_i(\theta) = \frac{s_{0i}^2(\theta)}{s_{0i}^2(\theta) + \sigma_i^2}.$$

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

$$z_i = (1, \log \sigma_i, x_i')',$$

and uses the same formulas

$$m_{0i}(\beta) = z_i'\beta, \qquad s_{0i}^2(\gamma) = \underline{s}^2 + \exp\left(z_i'\gamma\right).$$

Thus both the prior mean and the prior variance respond to $\log \sigma_i$ and the tract covariates. Plug-in, ASSURE*, ASSURE-CB, and SURE all optimize over (β, γ) in this richer family.

• *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

$$R_i = Y_i - \hat{g}(x_i), \qquad k_i^R = k^{(\tau)} - \hat{g}(x_i).$$

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.

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. The search interval is chosen from the empirical range of $(Y_i - k_i)/\sigma_i$, enlarged by a margin of 1, and the candidate grid always includes both 0 and 1.96. A coarse grid with 1201 points is searched first, after which the algorithm refines on a narrower neighborhood of the best coarse point using an 801-point grid. The value 0 is included because it nests the naive rule, while 1.96 is included because it is the conventional two-sided benchmark.

For the Gaussian-prior classes, ASSURE*, ASSURE-CB, and SURE all use multistart optimization around the plug-in fit. Additional starts are generated by Gaussian perturbations around the initial point, with coordinate-specific scales chosen from the inverse dispersion of the rule's design matrix so that poorly scaled covariates do not dominate the restart geometry. In the OA analysis the default is three restarts for both ASSURE* and ASSURE-CB.

Appendix H. Miscellany

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

Proof. The first claim follows since $|\operatorname{sinc}(y)| \leq \pi^{-1}$ per our definition. For the second, differentiating gives $h^{-2}\operatorname{sinc}'\left(\frac{x}{h}\right)$ where

$$\operatorname{sinc}'(y) = \frac{y\cos y - \sin y}{\pi y^2}.$$

An application of l'Hopital's rule shows that this function is zero at zero. From this and continuity, it is not difficult to see that $\operatorname{sinc}'(y)$ is uniformly bounded by some absolute constant. This gives the second claim. □

Lemma H.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.*

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

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