Compound Selection Decisions: An Almost SURE Approach

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

Abstract. We propose methods to construct *compound selection decisions* in a Gaussian sequence model. Given unknown, fixed parameters $\mu_{1:n}$, known $\sigma_{1:n}$, and observations $Y_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$, the decision maker chooses a subset $S \subset [n]$ to maximize utility $\frac{1}{n}\sum_{i \in S}(\mu_i - K_i)$ for known costs K_i. Inspired by Stein's unbiased risk estimate (SURE), we introduce an *almost* unbiased estimator, ASSURE, for the expected utility of a proposed decision rule. ASSURE selects a welfare-maximizing rule within a pre-specified class by optimizing the estimated welfare, thereby borrowing strength across noisy estimates. We show that, within the pre-specified class, ASSURE's decisions are asymptotically no worse than the optimal (infeasible) rule. 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: November 17, 2025. 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, Nikos Ignatiadis, Guido Imbens, Patrick Kline, Charles Manski, and seminar and workshop participants at Berkeley, Georgetown, Princeton, Northwestern, Stanford, UC Davis, and UC Santa Cruz for helpful comments. 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).

1. Introduction

In settings with many parallel noisy estimates, researchers often aim to *select* units with large underlying parameters—often screening using shrinkage estimates. Yet shrinkage and selection objectives can be misaligned: Good estimates need not yield good selections, and good selections need not rely on good estimates (Manski, 2021). This paper directly studies these *compound selection* problems[1] and proposes methods that optimize selection decisions while still "borrowing strength" across different noisy estimates (Robbins, 1951/1985).

We first describe a standard empirical-Bayes and compound-decision setup (see, e.g., Jiang and Zhang, 2009; Efron, 2012; Jiang, 2020; Walters, 2024; Soloff *et al.*, 2024; Chen, 2025). Researchers observe noisy estimates $Y_{1:n} = (Y_1, \ldots, Y_n)'$ and standard errors $\sigma_{1:n}$ for unknown *parameters* $\mu_{1:n}$, in parallel settings indexed by $i = 1, \ldots, n$. Motivated by central limit theorems for the procedure generating Y_i, we model the signals as Gaussian with known variance: $Y_i \mid \mu_i, \sigma_i, X_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$. Here, X_i are covariates that do not predict the noise $Y_i - \mu_i$. We seek binary decisions $a_i \in \{0, 1\}$ to maximize the compound utility

$$\frac{1}{n}\sum_{i=1}^{n} a_i \left(\mu_i - K_i\right). \tag{1.1}$$

The objective (1.1) is the average payoff of selection decisions over units $1, \ldots, n$, where the payoff of selecting unit i is $\mu_i - K_i$, for some known cost K_i. (1.1) can be viewed as a compound version of the treatment choice problem (Manski, 2004), where normality is motivated by local asymptotic approximations (Hirano and Porter, 2009, 2016; Hirano, 2023).

For concreteness,[2] in Bergman *et al.* (2024), Y_i is the estimated economic mobility of a Census tract, estimated on Census microdata. μ_i is true economic mobility, defined as the population economic outcome of poor children in adulthood in the tract. We

[1]In this paper, we refer to *compound decision*—often used interchangeably with "empirical Bayes" and "shrinkage" in the literature—specifically as decision-theoretic frameworks in which (i) the true parameters are fixed or conditioned upon and (ii) the loss function averages over units. *Empirical Bayes* additionally posits a random effect model for the distribution of the true parameters, and evaluates decisions by integrating over the distribution of the parameters.

[2]This type of problem also appears in other 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 treatment choice (Kitagawa and Tetenov, 2018; Athey and Wager, 2021; Moon, 2025). This problem also appears in the statistical literature under the name *empirical Bayes testing with linear loss* (Liang, 1988, 2000, 2004; Karunamuni, 1996).

may 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. The objective (1.1) then rewards selecting an above-breakeven tract, punishes selecting a below-breakeven tract, and weighs the rewards and penalties according to the distance to K. Since Y_i are fixed effect estimates, their normality—similarly invoked in Armstrong *et al.* (2022); Mogstad *et al.* (2024); Andrews *et al.* (2024); Chen (2025)—is motivated by the central limit theorem applied to regression specifications that generate them (Chetty *et al.*, 2018).

Because the Gaussian family has monotone likelihood ratio (Karlin and Rubin, 1956), admissible selection decisions are increasing in Y_i.[3] We thus identify selection decisions with *threshold rules* $\delta_i(Y_{-i})$ such that $a_i = \mathbf{1}(Y_i > \delta_i)$.[4] For a given threshold rule $\delta_{1:n}(Y_{1:n})$, define the *compound welfare* (cf. frequentist risk Wald (1950))

$$W(\delta_{1:n}, \mu_{1:n}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Y_i \sim \mathsf{N}(\mu_i, \sigma_i^2)} \left[\mathbf{1}(Y_i > \delta_i(Y_{-i}))\,(\mu_i - K_i)\right] \tag{1.2}$$

as the expected utility, integrating solely over $Y_{1:n}$. Optimal decisions should maximize (1.2), but doing so is infeasible since welfare depends on the unknown parameters $\mu_{1:n}$. For estimation problems minimizing MSE, one feasible approach is *Stein's unbiased risk estimate* (SURE, Stein, 1981). SURE *estimates* $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Y_i\sim\mathsf{N}(\mu_i,\sigma_i^2)}[(\delta(Y_i) - \mu_i)^2]$ and produces decisions by optimizing the estimated objective over a parametrized class of decisions (Xie *et al.*, 2012; Kwon, 2023; Cheng *et al.*, 2025).

Inspired by SURE, we construct a novel estimator $\hat{W}(\delta_{1:n})$ of $W(\delta_{1:n})$. Unlike with MSE, it turns out that an exactly unbiased estimator for (1.2) does not exist (Stefanski, 1989), but an *almost* unbiased one—whose bias decays exponentially in a tuning parameter—does. We thus term our approach ASSURE, for **A**lmo**s**t **SURE**. This almost-unbiasedness is an appealing and nontrivial property and can be viewed as a debiased version of (synthetic) sample-splitting-based approaches (Oliveira *et al.*, 2024; Chen, 2025; Ignatiadis and Sun, 2025).[5]

As with SURE, we optimize the estimated welfare over a user-chosen class of decisions $\delta \in \mathcal{D}$. For instance, $\mathcal{D}$ may be the class of decisions that screens on linear

[3] See Lemma E.1 for a formal result.

[4] The threshold rules depend on i through the contextual information $Z_i = (\sigma_i, K_i, X_i)$, which we suppress in the notation.

[5] That is, $Y_{i+} = Y_i + \epsilon W_i$ and $Y_{i-} = Y_i - \frac{1}{\epsilon} W_i$ for independent $W_i \sim \mathsf{N}(0, \sigma_i^2)$ satisfy $Y_{i+} \perp\!\!\!\perp Y_{i-} \mid \mu_i, \sigma_i^2$. Proposition 1 in Chen (2025) uses this *coupled bootstrap* idea to estimate (1.1) for decisions that depend on Y_{i+}.

shrinkage rules

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

for functions $a(\cdot), b(\cdot)$, parametrized by a finite-dimensional $\beta \in B \subset \mathbb{R}^d$, that depends on contextual information (σ_i, X_i, K_i) for some covariates X_i (Weinstein *et al.*, 2018; Chen, 2025). We restrict $\mathcal{D}$ for familiar bias-variance reasons: Optimizing over too large a class of decisions risks overfitting to the noise $\hat{W} - W$. In practice, these restrictions can come from external preference for simple decisions, similar to Kitagawa and Tetenov (2018); Sudijono *et al.* (2024); Crippa (2025), or from a benchmark (empirical) Bayesian model on the parameters $\mu_{1:n}$, similar to Kwon (2023) and Cheng *et al.* (2025). $\mathcal{D}_{\text{CLOSE-GAUSS}}$, for instance, can be motivated either by simplicity preferences or by a correlated random effects model in which $\mu_i \mid \sigma_i, K_i, X_i \sim \mathsf{N}(m(\sigma_i, K_i, X_i; \beta), s^2(\sigma_i, K_i, X_i; \beta))$, under which posterior means $\mu_i \mid Y_i, \sigma_i, K_i, X_i$ are linear in Y_i.[6]

Decisions selected by ASSURE enjoy strong optimality guarantees. As the number of parameters diverges ($n \to \infty$), we show that the performance gap between the estimated optimal decision and the optimizer of (1.2) 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 $\mathcal{D}_{\text{CLOSE-GAUSS}}$, since many popular shrinkage methods are nested within $\mathcal{D}_{\text{CLOSE-GAUSS}}$, decisions tuned by ASSURE would asymptotically improve over screening on these status-quo shrinkage methods. The bulk of the paper concerns Gaussian estimates, but similar arguments pertain to Poisson Y_i as well; we present these related results as an extension.

Our analysis complements and robustifies *empirical Bayes* analyses (Jiang, 2020; Gu and Koenker, 2023; Walters, 2024; Kline *et al.*, 2024; Chen, 2025) by focusing on the compound objective (1.2) rather than its empirical Bayesian counterpart

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

Empirical Bayes analyses treat $\mu_{1:n}$ as random effects and integrates over their distribution; in contrast, our compound perspective is analogous to fixed effects in panel data (Dano *et al.*, 2025). The benefit of treating $\mu_{1:n} \sim P$ as random is that the optimal selection decision simply screens on the posterior mean $\mathbb{E}_P[\mu_i \mid Y_i, \sigma_i, K_i, X_i]$: For an estimate $\hat{P}$, an empirical Bayesian selects via $\mathbf{1}(\mathbb{E}_{\hat{P}}[\mu_i \mid Y_i, \sigma_i, K_i, X_i] > K_i)$.

[6] Chen (2025) refers to this random effects model as CLOSE-GAUSS.

That benefit comes with two costs that our framework avoids. First, empirical Bayesians have to carefully model P. Failing to model this distribution well—either misspecifying how μ_i is predicted by the contextual information (Chen, 2025) or how μ_i correlates with μ_j (Bonhomme and Denis, 2024)—could harm the selection decision. The compound framework avoids these problems entirely by conditioning on $\mu_{1:n}$. Second, optimality for W_{EB} is with respect to a new draw of units $\mu_i \sim P$. In Bergman *et al.* (2024), for instance, empirical Bayes evaluates performance over a randomly drawn new Census tract, whereas optimality with respect to (1.1) only imagines alternative draws of the estimation error in economic mobility.

To be sure, a SURE-based compound decision is limited to optimality within a class of decision rules $\mathcal{D}$. Empirical Bayes methods can match the performance of the optimal decision under P, without constraining to a class $\mathcal{D}$. This limitation of ASSURE can be mitigated by choosing $\mathcal{D}$ via some working empirical Bayes model. When a model of P implies a parsimonious class of decisions rules through $a(Y_i) = \mathbf{1}(\mathbb{E}_P[\mu_i \mid Y_i] > 0)$,[7] optimizing within this class using ASSURE weakly improves on the empirical Bayes decisions under an estimated prior $\hat{P}$. Thus, our method can be combined with (parametric) empirical Bayesian modeling: it *fine-tunes* such a model without harming performance.

We illustrate ASSURE in three empirical applications: selecting census tracts to maximize economic mobility (Chetty *et al.*, 2018; Bergman *et al.*, 2024), picking innovations in A/B testing (Azevedo *et al.*, 2020), and identifying discrimination in large firms (Kline *et al.*, 2022, 2024). These applications show that ASSURE improves performance and provides robustness relative to common plug-in empirical Bayes methods. The ASSURE-estimated welfare $\hat{W}$ also assesses whether a status-quo decision rule attains reasonable welfare under assumed costs K_i.

This paper is structured as follows. Section 2 defines the ASSURE estimator and discusses its bias and variance. Section 3 presents several theoretical guarantees on selecting a decision using ASSURE, showing upper and lower bounds on regret. We next highlight extensions to other observation distributions like the Poisson and to complicated decision procedures in Section 4. Simulation results and empirical applications are discussed in Section 5 and Section 6.

[7] Posterior means with Gaussian Y are monotone in Y, and thus this threshold rule can be equivalently written as $Y_i > \delta_i$.

2. Setup & Methodology

Let $Y_i \sim \mathsf{N}(\mu_i, \sigma_i^2), i \in [n] = \{1, \dots, n\}$ with known σ_i, known costs K_i, with potentially related known covariates X_i. We treat the contextual information $Z_i := (X_i, K_i, \sigma_i)$ and the parameters μ_i as fixed, and take all probability statements over $Y_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$. For simplicity, we assume that Y_i are mutually independent across i.[8] We also restrict to *separable* thresholds of the form $\delta_i(Y_{-i}) = \delta(Z_i; \beta)$ in all classes $\mathcal{D}$ that we consider. That is, the threshold for unit i does not depend on information for unit $j \neq i$ (except through estimation of β).

Having parametrized thresholds with β, we can rewrite (1.2) as

$$W(\beta) := \frac{1}{n}\sum_{i=1}^{n}(\mu_i - K_i)\,\mathbb{P}\left(Y_i \geq \delta(Z_i;\beta)\right) = \frac{1}{n}\sum_{i=1}^{n}(\mu_i - K_i)\Phi\left(\frac{\mu_i - \delta(Z_i;\beta)}{\sigma_i}\right) \tag{2.1}$$

$$u(\beta) := \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(Y_i > \delta(Z_i;\beta)) \cdot (\mu_i - K_i). \tag{2.2}$$

Directly optimizing (2.1) is infeasible since it depends on the unknown $\mu_{1:n}$. However, if we had access to an estimator $\hat{W}_n(\beta)$, we can optimize $\hat{W}_n$ over a class of threshold rules $\mathcal{D} = \left\{\delta(\cdot;\beta) : \beta \in \mathcal{B} \subset \mathbb{R}^d\right\}$ (e.g., $\mathcal{D} = \mathcal{D}_{\text{CLOSE-GAUSS}}$), yielding:

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

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 $\hat{\beta}$:

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

Intuitively, for Regret_n to be small, it is critical that our estimator $\hat{W}_n$ for welfare is accurate. A good welfare estimator for the compound selection problem is the focus of this paper.

2.1. The ASSURE estimator for welfare.

We propose the following estimator for (2.1) with good regret properties, building on the deconvolution literature (Kolmogorov, 1950; Tate, 1959; Liang, 2000; Pensky, 2017; Zhou and Li, 2019). Define the sinc, sine integral, and cumulative sinc functions as follows:

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

[8] If $(Y_1, \dots, Y_n) \sim \mathsf{N}((\theta_1, \dots, \theta_n)', \Sigma)$ for some known Σ, then our arguments are generalizable by studying the conditional distributions $Y_i \mid Y_{-i}$.

Definition 1 (ASSURE)**.** For a bandwidth parameter h, define

$$w_h(Y_i; Z_i, \beta) := (Y_i - K_i)\,\mathrm{Csinc}\left(\frac{Y_i - \delta(Z_i;\beta)}{\sigma_i h}\right) - \frac{\sigma_i}{h}\,\mathrm{sinc}\left(\frac{Y_i - \delta(Z_i;\beta)}{\sigma_i h}\right). \quad (2.5)$$

Then the ASSURE estimator is given by

$$\widehat{W}_n(\beta) := \frac{1}{n}\sum_{i=1}^{n} w_{1/\sqrt{2\log n}}(Y_i; Z_i, \beta). \quad (2.6)$$

We will define the scaling parameter $\lambda_n = \sqrt{2\log n}$ and set the bandwidth as $h = 1/\lambda_n$. With this choice of h, we will usually omit dependence of w on h.

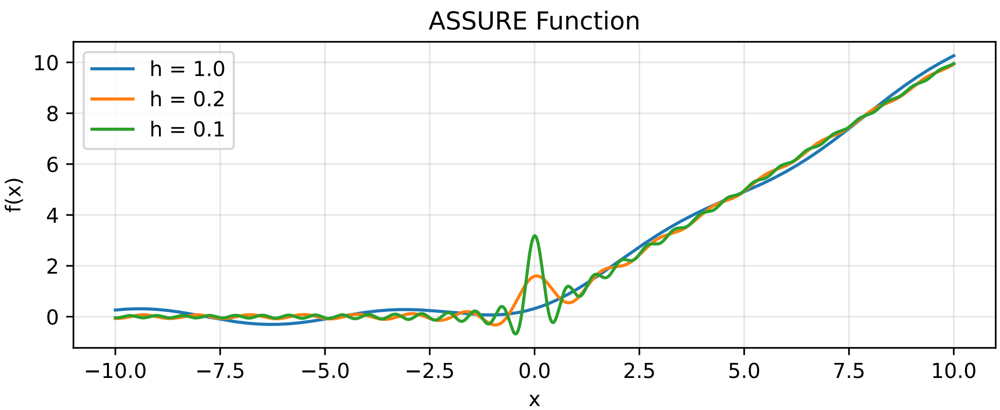


FIGURE 1. A plot of the function $\frac{y}{2}+\frac{y}{\pi}\,\mathrm{Si}((y-C)/h)-\frac{1}{h}\,\mathrm{sinc}((y-C)/h)$, for $C = 1$ and various values of the bandwidth h.

Figure 1 visualizes the function w_h in (2.5) for $K = 0$ and $\sigma = 1$. The estimator w_h is chosen due to its excellent bias properties: The following proposition shows that its bias decays exponentially in $1/h$.

Proposition 1 (ASSURE has low bias)**.** *Fix $\mu, K \in \mathbb{R}$. Let $Y \sim \mathsf{N}(\mu, \sigma^2)$ and $\delta(Z_i;\beta) = \beta$, then $\lim_{h\to 0}\mathbb{E}_\mu w_h(Y;(\sigma,K),\beta) = (\mu-K)\Phi\left(\frac{\mu-\beta}{\sigma}\right)$. The bias attains the following bound: For all $h > 0$,*

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

Thus, with $h := h_n := \lambda_n^{-1} := 1/\sqrt{2\log n}$,[9] *the bias is bounded by* $C|\mu - K|\frac{1}{n\log n}$. *Under this choice, pointwise in* β,

$$\mathbb{E}\left[\left(\hat{W}_n(\beta) - W(\beta)\right)^2\right] \lesssim \frac{(\log n)^2}{n}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i^2\right) + \frac{\log n}{n}\left(\frac{1}{n}\sum_{i=1}^n (\delta(Z_i;\beta) - K_i)^2\right). \tag{2.7}$$

Unlike SURE, ASSURE is biased. The bandwidth term h_n is chosen to ensure the bias contributes negligibly to MSE for estimating $W(\beta)$. For $W(\beta)$, it turns out that unbiased estimators (with reasonable growth behavior) do not exist: we formalize in Proposition E.1 an argument in Stefanski (1989).

The next subsection provides heuristic motivation for the functional form (2.5) in ASSURE and connects it to SURE, sample splitting, and to a literature on estimation using Fourier transforms. The subsection after compares our SURE-type analysis (2.3) to other approaches to compound selection problems, such as empirical Bayes.

2.2. **Intuition for ASSURE.** Inspecting (2.1), estimating $W(\beta)$ can be reduced to a Gaussian estimation problem: Given $Y \sim \mathsf{N}(\mu, \sigma^2)$ and $K, \delta \in \mathbb{R}$, we would like to estimate the parameter $(\mu - K)\Phi\left(\frac{\mu-\delta}{\sigma}\right)$. Inspired by SURE, a natural starting point is *Stein's identity*: For differentiable $F(Y)$ with derivative $f(Y)$,

$$\mathbb{E}_{Y\sim\mathsf{N}(\mu,\sigma^2)}[(Y-\mu)F(Y)] = \sigma^2\mathbb{E}_{Y\sim\mathsf{N}(\mu,\sigma^2)}[f(Y)].$$

We can rearrange to obtain

$$\mathbb{E}[(Y-K)F(Y) - \sigma^2 f(Y)] = (\mu - K)\mathbb{E}[F(Y)]. \tag{2.8}$$

Thus, $(Y-K)F(Y)-\sigma^2 f(Y)$ is an unbiased estimator for $(\mu-K)\mathbb{E}[F(Y)]$. To obtain a low-bias estimator for $(\mu - K)\Phi\left(\frac{\mu-\delta}{\sigma}\right)$, we could find some differentiable F whose expectation $\mathbb{E}[F(Y)]$ is approximately $\Phi\left(\frac{\mu-\delta}{\sigma}\right) = \mathbb{E}[\mathbf{1}(Y > \delta)]$.

Having reduced the problem in this way, by rescaling if necessary, we may assume $\sigma = 1, \delta = 0$ without loss of generality. A natural idea is to smooth the Heaviside function $H(y) := \mathbf{1}(y > 0)$ with some kernel $k_h(y) = \frac{1}{h}k\left(\frac{y}{h}\right)$—for some kernel $k(\cdot)$ and bandwidth h—so that the resulting function is differentiable. That is, we may

[9] We upper bound bias with $h^2e^{-Ch^2}$ and variance with $n^{-1}\left(h^{-4}C_\sigma + h^{-2}C_{\mu-K}\right)$ for constants C that depends on (σ_i, μ_i, K_i). Minimization of corresponding MSE bounds with respect to h shows that $h_n^{-1} \asymp \sqrt{C_1\log n - \log C_{2,\sigma} - C_3 \log\log n} \asymp \sqrt{C_1 \log n}$. The leading constant C_1 does not depend on σ_i^2.

consider convolving H with the kernel:

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

Which kernel should we choose? A natural but suboptimal idea is to use the Gaussian kernel $k(\cdot) = \varphi(\cdot)$. Doing so yields an estimator that has $O(h^2)$-bias

$$\mathbb{E}\left[(Y-K)\Phi\left(\frac{Y}{h}\right) - \varphi\left(\frac{Y}{h}\right)\right] = (\mu - K)\Phi(\mu) + O(h^2). \tag{2.9}$$

Interestingly, this choice of kernel has a natural interpretation as *coupled bootstrap* (Oliveira *et al.*, 2024; Leiner *et al.*, 2023; Ignatiadis and Sun, 2025; Chen, 2025). For $Y \sim \mathsf{N}(\mu, 1)$, we can construct two independent samples by adding and subtracting an independent Gaussian noise Q

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

This *coupled bootstrap* procedure is a synthetic version of sample-splitting.[10] With (2.10), the welfare of selection decisions based on $Y_1 > 0$ can be unbiasedly estimated by $(Y_2 - K)\mathbf{1}(Y_1 > 0)$, since Y_2 acts as fresh "testing data" that is independent of the "training data" Y_1. The estimator (2.9) is exactly the Rao–Blackwellization of coupled bootstrap:

$$\mathbb{E}[(Y_2 - K)\mathbf{1}(Y_1 > 0) \mid Y] = (Y-K)\Phi\left(\frac{Y}{h}\right) - \varphi\left(\frac{Y}{h}\right).$$

This $O(h^2)$ bias, however, means that the regret rate for coupled bootstrap is $\tilde{O}(n^{-4/5})$ (Theorem A.1).[11]

Instead of a Gaussian kernel, the ASSURE estimator uses the sinc kernel for improved bias (Davis, 1975; Tsybakov, 2009). Observe that the expectation $\mathbb{E}[F_h]$ convolves F_h with a Gaussian kernel:

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

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: $k_h \star \varphi \approx \varphi$. The sinc kernel is motivated by this approximation in frequency space:[12] By taking the

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

[11]Throughout, we use $\tilde{O}(\cdot)$ to denote the analogue of big-O notation but ignoring logarithmic terms in n.

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

Fourier transform of both sides, we would like a kernel for which

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

The sinc kernel has Fourier transform $\mathbf{1}(|\omega| < 1/h)$, 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 $1/h$ (Proposition 1).

Remark 1 (Intuition for behavior under approximate normality of Y)**.** The identity (2.8) is also suggestive of ASSURE's robustness to viewing Y as approximately Gaussian—though we leave a formal analysis to future work. Let $Z \sim \mathsf{N}(\mu, \sigma^2)$ and suppose Y has mean μ and variance σ^2, but is not necessarily Gaussian. Then

$$\begin{aligned}
&\mathbb{E}[(Y-K)F(Y) - \sigma^2 f(Y)] \\
&= (\mu - K)\mathbb{E}[F(Y)] + \mathbb{E}[(Y-\mu)F(Y) - \sigma^2 f(Y)] \\
&= (\mu - K)\mathbb{E}[F(Z)] + (\mu - K)\underbrace{\{\mathbb{E}[F(Y)] - \mathbb{E}[F(Z)]\}}_{A} + \underbrace{\mathbb{E}[(Y-\mu)F(Y) - \sigma^2 f(Y)]}_{B}
\end{aligned}$$

The term A measures the discrepancy between the distributions of Y and Z in terms of $\mathbb{E}[F(\cdot)]$. This term is controlled with various weak convergence arguments, such as controlling the Wasserstein-1 distance between Y and Z. The term B measures the extent to which Y does not obey Stein's identity. It is often controlled directly by central limit theorems using Stein's method. As long as F_h is sufficiently well-behaved so that A and B are small by the approximate Gaussianity of Y, we would like $\mathbb{E}[F_h(Z)] \approx \Phi(\mu)$, for which the sinc kernel is optimizing. Simulation studies in Sections 5 and 6 show robustness to the Gaussian assumption.

2.3. **Discussion.** We now contextualize our approach in the SURE literature and compare it against a few alternatives. We also disambiguate our regret notion (2.4) from the regret notions in these literatures.

Stein's unbiased risk estimate. Our approach exactly mimics SURE: For $Y \sim \mathsf{N}(\mu, \sigma^2)$ and a differentiable estimator $\delta(Y)$, the squared error risk of δ is unbiasedly estimable by $T(Y; \delta, \sigma) = (\delta(Y) - Y)^2 + 2\sigma^2\delta'(Y) - \sigma^2$ (Stein, 1981), so that

$$\mathbb{E}_\mu\left[T(Y; \delta, \sigma)\right] = \mathbb{E}_\mu[(\delta(Y) - \mu)^2]. \tag{SURE}$$

Thus, for compound *estimation* problems—*predicting* μ_i using $\delta(Y_i; Z_i, \beta)$—one could form a risk estimator $\hat{R}_{\text{SURE}}(\beta) := \frac{1}{n}\sum_{i=1}^n T(Y; \delta(\cdot, \beta), \sigma_i)$. This is the strategy pursued in Xie *et al.* (2012); Kwon (2023); Cheng *et al.* (2025). This research often invokes selection as motivation, but the shrinkage approaches are designed for estimation; they are not guaranteed to be good for selection (Manski, 2021). We target the selection problem directly.

Empirical Bayes. A popular framework for compound decision problems is *empirical Bayes* (Efron, 2012), which models μ_i as random effects. This additional distributional structure characterizes optimal decision rules as posterior quantities involving the estimable distribution of μ_i (Liang, 1988, 2000, 2004; Karunamuni, 1996; Gupta and Li, 2005; Gu and Koenker, 2023; Kline *et al.*, 2024). However, the additional structure hinges on modeling and estimating the distribution of μ_i well, which can make subsequent procedures and guarantees less robust.[13] In particular, our procedure can be viewed as a compound analogue of Liang (2000).

Our procedure is also complementary to empirical Bayes. Empirical Bayes-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. For concreteness, a model for μ_i that performs well in the empirical exercise in Chen (2025) is $\mu_i \mid Z_i \sim \mathsf{N}(m_0(Z_i;\beta), s_0^2(Z_i;\beta))$, where m_0, s_0 are flexibly parametrized by β. Under this prior,

$$\mathbb{E}[\mu_i \mid Y_i, Z_i] = m_0(Z_i;\beta) + \frac{s_0^2(Z_i;\beta)}{\sigma_i^2 + s_0^2(Z_i;\beta)}(Y_i - m_0(Z_i;\beta))$$

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 :

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

From a formal perspective, the empirical Bayes literature (Jiang, 2020; Soloff *et al.*, 2024; Chen, 2025) often considers the difference in W_{EB} between the oracle Bayes decision rule and the empirical Bayes decision rule:

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

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

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 criterion (2.4) that we subsequently study, $\mathsf{EBRegret}_n$ (i) takes another expectation under $\mu_{1:n} \mid Z_{1:n} \sim P$, (ii) chooses the benchmark as the oracle decision rule $\mathbf{1}(\mathbb{E}_P[\mu_i \mid Y_i, Z_i] > K_i)$, 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 $\mathsf{EBRegret}_n$:

$$\mathsf{EBRegret}_n = \mathbb{E}_{\substack{\mu_{1:n}\sim P \\ Y_i \mid \mu_i, Z_i \sim \mathsf{N}(\mu_i, \sigma_i^2)}} \left[W(P) - u(\hat{P})\right] \le \mathbb{E}_{\mu_{1:n}\sim P}\left[\mathsf{Regret}_n\right] = \mathbb{E}_{\mu_{1:n}\sim P}\left[(2.4)\right].$$

Treatment choice and policy learning. The decision problem (1.1) is a compound version of the "testing an innovation" problem in Manski (2009), where μ_i represents the average treatment effect of a new treatment relative to a known status quo. To connect to the treatment choice literature (Manski, 2004; Stoye, 2012; Kitagawa and Tetenov, 2018; Athey and Wager, 2021), the compound aspect may represent n parallel innovations that the decision-maker is simultaneously entertaining, or it may represent discrete covariates taking on n values, wherein μ_i is the conditional average treatment effect on covariate value i.[14] The normality assumption on Y_i corresponds to the limit experiment in Hirano and Porter (2009), though rigorously justifying this approximation through Le Cam–Hajek-style arguments remains important future work (Hirano, 2023).

Our statistical perspective is distinct from the treatment-choice literature. First, we aim for decision rules whose regret (2.4) converges to zero as $n \to \infty$, uniformly over sequences $(\mu_i, Z_i)_{i=1}^n$. These decision rules are not necessarily minimax-$\mathsf{FreqRegret}$ for any finite n, for $\mathsf{FreqRegret}$ defined as follows (e.g., in Ishihara and Kurisu, 2022):

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

This means that ASSURE may suffer from higher worst-case $\mathsf{FreqRegret}$ for any finite n. Conversely, minimax-$\mathsf{FreqRegret}$ decision rules, like the empirical success rule $a(Y_i) = \mathbf{1}(Y_i > 0)$ for $K_i = 0$ (Proposition 1, Stoye, 2012), have non-vanishing worst-case regret $\sup_{\mu_{1:n}} \mathsf{Regret}_n = \Omega(1)$. That is, these decisions can be substantially improved from the perspective of (2.4). See Remark 2 for further comparisons and intuition.

[14]Though, from this perspective, the objective function (1.1) assumes that each covariate cell is equally probable.

Second, our asymptotic regime is different from the policy learning literature (Kitagawa and Tetenov, 2018; Athey and Wager, 2021; Mbakop and Tabord-Meehan, 2021). If a unit i is taken to be a covariate cell, we consider a regime where the number of covariate cells is large, perhaps due to finer discretization of continuous covariates. In contrast, the standard asymptotic regime in the policy learning literature considers increasing sample size within each covariate cell, which can be approximately written as $\sigma_i^2 \asymp c_i/m$ for $m \to \infty$. Thus the regret rates we obtain are not directly comparable to those in Kitagawa and Tetenov (2018).

Remark 2 (Comparisons to minimax-FreqRegret)**.** The comparison between ASSURE-based decision rules and minimax-FreqRegret decision rules is analogous to the fact that, in estimation settings, the maximum likelihood estimator $\delta(Y_i) = Y_i$ is a minimax estimator for $\mu_{1:n}$ in any dimension, but its squared error risk is dominated by the James–Stein estimator for $n \geq 3$ (Lehmann and Casella, 1998). Indeed, the James–Stein estimator is a minimizer of SURE under homoskedasticity, up to a degree-of-freedom adjustment. Related examples, for the advantages of compound decisions, are found in Robbins (1951/1985); Sun and Cai (2007); Koenker and Gu (2024).

To intuitively understand how ASSURE improves over the empirical success rule $\mathbf{1}(Y_i \geq K_i)$, note that certain aggregate statistics like $m := \frac{1}{n}\sum_{i=1}^n(\mu_i - K_i)$ are precisely estimable from the data. Estimates of these statistics are informative. For example, supposing that $K_i = 0$, if our estimate for m is greater than zero, then we have some evidence that the configuration of μ_i is on average positive. If we are confident that $m > 0$, we should perform better by considering less conservative decision rules such as $\delta_i = \mathbf{1}\{Y_i \geq -\epsilon\}$, since any mildly negative estimate Y_i are now more likely due to chance rather than a negative μ_i. ASSURE thus adapts the selection decision to features of the data, implicitly betting on the estimates for m to be accurate. Fixed decision rules—often motivated by concerns about the estimation error for m—lack this adaptivity.

3. Theoretical assurances

This section presents statistical guarantees on ASSURE. The main results are upper and matching lower bounds for expected regret (2.4). We first present general upper and lower bounds for $\tilde{O}(1/\sqrt{n})$ regret. 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.

Theorem 1 ($\tilde{O}(1/\sqrt{n})$-regret for ASSURE)**.** *Define the quantities*

$$s_k := \left(\frac{1}{n}\sum_{i=1}^n \sigma_i^k\right)^{1/k} \qquad m_k := \left(\frac{1}{n}\sum_{i=1}^n |\mu_i - K_i|^k\right)^{1/k} \qquad \nu_k := \left(\frac{1}{n}\sum_{i=1}^n \frac{|\mu_i - K_i|^k}{\sigma_i^k}\right)^{1/k}.$$

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, for $\hat{\beta}$, the regret satisfies*

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

with $M := \sqrt{V(\mathcal{D})}(s_2 + m_2 + (1+\nu_4)(\mathbb{E}\|D\|_n + (\mathbb{E}\|D\|_{4,n}^2)^{1/2}))$.

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. Constants in the regret rate depend on various norms of the parameters. If all parameters are bounded, we have the following corollary, showing a simple $\log n/\sqrt{n}$ rate.

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, not depending on n. 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 $\sup_{\mu_{1:n}} \mathsf{Regret}_n = O\left(\frac{\log n}{\sqrt{n}}\right)$.*

Theorem 1 builds on a standard empirical-risk-minimization argument:[15] from the optimality of β^* and $\hat{\beta}$, recalling $u(\beta)$ from (2.2),

$$\mathsf{Regret}_n \le 2\mathbb{E}\left[\sup_\beta |\hat{W}(\beta) - W(\beta)|\right] + 2\mathbb{E}\left[\sup_\beta |u(\beta) - W(\beta)|\right]. \tag{3.1}$$

The $1/\sqrt{n}$-rate follows from analyzing the empirical process $\beta \mapsto \hat{W}(\beta) - W(\beta)$. Proposition 1 shows that $|\hat{W}(\beta) - W(\beta)| = O_P(\log n/n)$ pointwise. Empirical process arguments control this uniformly.[16] We modify standard empirical process arguments to accommodate independent but not identically distributed data in our setting.

[15]See, e.g., 8.4.3. in Vershynin (2009). Regret guarantees (2.4) are related to, but are distinct from, excess risk control in empirical risk minimization. Standard empirical risk minimization (e.g., Theorem 8.4.4 in Vershynin (2009)) controls—in our notation—the welfare gap $W(\beta^*) - \mathbb{E}[W(\hat{\beta})]$. This is a different quantity than (2.4) since $\mathbb{E}[W(\hat{\beta})] \neq \mathbb{E}[u(\hat{\beta})]$. $\mathbb{E}[W(\hat{\beta})]$ imagines evaluating $\hat{\beta}$ on a new draw of Y_i, whereas $\mathbb{E}[u(\hat{\beta})]$ evaluates $\hat{\beta}$ on the same sample of draws $Y_{1:n}$.

[16]The VC subgraph assumption is standard but not crucial, as long as an appropriate covering number of the decision class is controlled. We state our results in terms of the VC subgraph dimension for simplicity.

3.2. **Examples.** We now give several examples of function classes with finite VC dimension.

Example 1 (Simple truncation rules)**.** Consider decision rules

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

Let β be restricted to a compact set $[-M, M]$. $\mathcal{D}_{\text{threshold}}$ is VC subgraph with a uniform envelope function $k + M$, satisfying the requirements of Theorem 1 and Corollary 1. $\mathcal{D}_{\text{threshold}}$ is common; see Crippa (2025) for several examples.

Example 2 (Thresholding t-statistics)**.** Consider decision rules

$$\mathcal{D}_{t\text{-stat}} = \{\delta(z;\beta) = k + \beta\sigma : \beta \in [-M, M]\}$$

Such rules nest selection via one-sided p-values, common in digital experimentation (Sudijono *et al.*, 2024) and medicine (Manski, 2019).[17] For σ_i bounded, $\mathcal{D}$ is VC subgraph with a uniformly bounded envelope function.

Example 3 (Finite Classes of Decision Rules)**.** Suppose now a discrete decision class $\mathcal{D}_{\text{finite}}$ consisting of p different decision thresholds

$$\left\{\delta^{(1)}(z), \ldots, \delta^{(p)}(z)\right\}.$$

A finite class of functions is VC subgraph. If $\sup_{i=1,\ldots,p} \delta^{(i)}(z)$ is uniformly bounded, then Theorem 1 and Corollary 1 apply. By standard arguments, convex combinations of these decision rules are also VC subgraph.

Example 4 (Gaussian Empirical Bayes Models)**.** Recall in Eq. (2.11) the decision rules obtained by imposing the prior

$$\mu_i \mid Z_i \sim \mathsf{N}(m_0(Z_i;\beta), s_0^2(Z_i;\beta)),$$

and selecting a unit if its posterior mean is greater than K_i. Rearranging this yields

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

[17]These decision rules can be shown to be Bayes optimal for a class of exponentially tilted improper priors.

Particular choices of m_0, s_0^2 yield important examples. For example, taking both $m_0 = \mu_0, s_0^2 = \tau^2$ to be constants independent on Z_i generates *linear shrinkage* decision rules where the parameters are $\beta = (\mu_0, \tau)$:

$$\mathcal{D}_{\text{linear-shrink}} := \left\{ \delta(Z_i; \beta) = K_i + \frac{\sigma_i^2}{\tau^2}(K_i - \mu_0) : \beta \in \mathbb{R}^2 \right\}. \tag{3.2}$$

Next, taking s_0^2 to be a constant and m_0 to be the linear model $X_i^\top \beta$ where X_i are auxiliary covariates yields the well-known *Fay-Herriot model* (Fay III and Herriot, 1979):

$$\mathcal{D}_{\text{Fay-Herriot}} := \left\{ K_i + \frac{\sigma_i^2}{A}(K_i - X_i^\top \beta) : (A, \beta) \in \mathbb{R}^{1+p} \right\}. \tag{3.3}$$

For either model, empirical Bayes procedures typically plug in estimates of μ_0, τ^2, A using the method of moments or maximum likelihood. See Ignatiadis and Wager (2019) and Luo *et al.* (2023) for other empirical Bayes methods which utilize covariates or side information.

Finally, Chen (2025) discusses fitting models $m_0(\sigma_i), s_0^2(\sigma_i)$ only depending on σ_i. Chen (2025) considers fitting m_0, s_0^2 nonparametrically and also parametrically using the model

$$\mathcal{D}_{\text{CLOSE-GAUSS}} = \left\{ K_i + \frac{\sigma_i^2}{s_0^2(Z_i; \beta)}(K_i - m_0(Z_i; \beta)) : \beta \in \mathcal{B} \right\} \tag{3.4}$$

$$m_0(Z_i; \beta) = a_1 + a_2 \sigma_i \tag{3.5}$$

$$s_0^2(Z_i; \beta) = \exp(b_1 + b_2 \log \sigma_i), \tag{3.6}$$

with $\beta = (a_1, a_2, b_1, b_2) \in \mathbb{R}^4$. ASSURE can be used to tune the parameters β for the parametric specification. We also discuss how to use ASSURE to improve on the nonparametric method in Section 4.

These decision threshold classes are VC subgraph: the thresholds for linear shrinkage (3.2) and the Fay-Herriot model (3.3) can be reparametrized to be linear. Proposition B.4, via model theory, shows that parametric CLOSE-GAUSS is VC-subgraph.

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 $W(\beta; \mu_{1:n})$ is defined with respect to the thresholds $\delta(Z_i; \beta) = \beta \in \mathbb{R}$. Let $M > 0$, then over* all *decision rules $a_i(Y_1, \ldots, Y_n) \in [0, 1]$,*

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

where

$$\mathsf{Regret}_n(a_{1:n}) := \sup_{|\beta|\le M} W(\beta;\mu_{1:n}) - \frac{1}{n}\sum_{i=1}^n \mu_i \mathbb{E}_{\mu_{1:n}}[a_i(Y_1,\ldots,Y_n)]. \tag{3.7}$$

Here, C is a constant that only depends on M.

Here, the regret notion in Theorem 2 compares the welfare of an arbitrary decision $a_i(Y_1,\ldots,Y_n)$ to the expected welfare of the best decision in the thresholding class restricted to $|\beta| \le M$. Lower bounds for this quantity thus imply lower bounds when we allow the oracle to be more flexible than using $\delta = \beta$.[18]

Theorem 2 is derived by 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. 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 $\tilde{O}(1/n)$.

3.4. **Fast Rates for ASSURE.** In particular settings, the performance for ASSURE can be better than the $\tilde{O}(n^{-1/2})$ rate predicts. Technically, this is due to the fact that the derivatives of ASSURE also estimate the derivatives of welfare (1.2) as shown in Theorem B.3. This analysis shares similarities with the kernel estimator introduced by Liang (2000) for the random effect case.

The assumptions needed for the improved rate are reminiscent of the *margin condition* used to achieve 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 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 in the proof, we will restrict to the $d = 1$ case which covers a range of examples. We leave the multidimensional extension to future work. Suppose the following conditions hold.

Assumption 1. *(A1) (Boundedness). Suppose that $\mu_i, K_i, \sigma_i, \sigma_i^{-1}$ are all bounded in absolute value by a constant B. Furthermore, suppose that the decision*

[18] 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 Eq. (12) for the mean square case.

rules $\delta(\cdot, Z_i)$ are smooth and L_i-Lipschitz for each i. Further, assume that the decision rules and all derivatives are uniformly bounded by a uniform constant B.

(A2) *(Unimodality). There exists a global maximizer β^* at which $W(\beta^*) > 0$; moreover β^* is the unique local and global maximizer.*

(A3) *(Curvature). $W''(\beta^*) < -\kappa$ for some positive $\kappa > 0$, not depending on $\mu_1, \ldots, \mu_n$.*

(A4) *(Well-separated Maximum). There exists $\xi > 0$ not depending on $\mu_1, \ldots, \mu_n$ such that*

$$\sup_{|\beta-\beta^*|>\kappa/B} W(\beta) < W(\beta^*) - \xi.$$

For convenience in the proof, we additional conditions on the decision rules and search space.

Theorem 3 (Fast Rates for ASSURE). *Suppose Assumption 1 holds. Furthermore, suppose that β is restricted to a compact space Θ of finite volume, which includes β^*. Lastly, suppose that the derivatives $\delta'(\beta, z), \delta''(\beta, z)$ form VC subgraph function classes with index $V(\mathcal{D}'), V(\mathcal{D}'')$. Then the regret of* ASSURE *satisfies*

$$\mathbb{E}\left[W(\beta^*) - u(\widehat{\beta})\right] = O\left(\frac{(\log n)^5}{n}\right).$$

Example 5 in Appendix B.5.3 illustrates a class of decision rules and lower-level assumptions on $\mu_{1:n}$ for which Theorem 3 applies. A heuristic argument for Theorem 3 is as follows. If $\hat{\beta}$ is a local maximum of $\hat{W}$, then from a first-order Taylor expansion,

$$\begin{aligned} 0 = \hat{W}'(\hat{\beta}) &\approx \hat{W}'(\beta^*) + \hat{W}''(\beta^*)(\hat{\beta} - \beta^*). \\ &= \hat{W}'(\beta^*) - W'(\beta^*) + \hat{W}''(\beta^*)(\hat{\beta} - \beta^*). \end{aligned}$$

It can be shown that $\hat{W}'$ estimates W' uniformly at rate $\tilde{O}(1/\sqrt{n})$, from which it follows that $\hat{\beta} - \beta^* = \tilde{O}_P(1/\sqrt{n})$. Another Taylor expansion $W(\hat{\beta}) \approx W(\beta^*) + W''(\beta^*)(\hat{\beta} - \beta^*)^2$ shows that the welfare gap is $\tilde{O}(1/n)$: $W(\hat{\beta}) - W(\beta^*) = \tilde{O}_P(1/n)$. Finally, we consider a leave-one-out stability argument to control $\mathbb{E}[W(\hat{\beta}) - u(\hat{\beta})]$.[19]

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

Theorem 4 (Fast Regret Lower Bounds). *Consider the homoskedastic $\sigma_i = 1$ costless case with the decision rules $\{Y_i \geq \beta\}$ ranging over $\beta \in \mathbb{R}$. Let $\Theta = \Theta_{n,\kappa,\xi} \subseteq \mathbb{R}^n$ be the space of $(\mu_1, \ldots, \mu_n)$ satisfying the assumptions (A1)-(A4) for a fixed ξ, κ. Then*

[19] The log factors are likely not tight and removable with a more refined application of empirical process theory.

there exists choices of ξ *and* κ *such that*

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

with $\mathsf{Regret}_n(a_{1:n})$ *defined as in* (3.7).

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.[20] We will show that the Bayes regret for this problem can be interpreted as a type of weighted classification loss, building on an insight by Polyanskiy and Wu (2021). As a result, Theorem 4 may be of independent interest for empirical Bayes testing with linear loss.

4. **Extensions**

4.1. **Other observation distributions.** The approach—finding an estimator for W and optimizing this estimator over decisions—can be extended to settings where the observation distribution is non-Gaussian. For certain exponential families like the Poisson distribution and exponential distribution, one can find an exactly unbiased estimate for the welfare W. The intuition for these cases is to exploit a formula akin to Tweedie's formula. We focus on the Poisson setting as it frequently appears in empirical Bayes and compound decisions.[21]

In this setting, let $\mu_i \geq 0$ and $Y_i \sim \mathsf{Poi}(\mu_i)$. Again, associated with each decision problem is a cost K_i and auxiliary information Z_i.[22] Consider any integer-valued decision rule $\delta(Z_i;\beta)$ parametrized by β. Then the estimator

$$\widehat{W}_{\mathsf{Poi}}(\beta) := \frac{1}{n}\sum_{i=1}^{n}\left(Y_i \mathbf{1}\left\{Y_i \geq \delta(Z_i;\beta)+1\right\} - K_i \mathbf{1}\left\{Y_i \geq \delta(Z_i;\beta)\right\}\right) \tag{4.1}$$

is the analog for ASSURE.

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

[21]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.

[22]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.

Proposition 2. *The Poisson* ASSURE *estimator* $\widehat{W}_{\mathsf{Poi}}(\beta)$ *is unbiased for the welfare*

$$W(\beta;\mu_{1:n}) := W(\beta) := \frac{1}{n}\sum_{i=1}^{n}(\mu_i - K_i)\,\mathbb{P}(Y_i \geq \delta(Z_i;\beta)). \tag{4.2}$$

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 5 (Regret upper bounds for the Poisson Case)**.** *Let* $m_k := \left(\frac{1}{n}\sum_{i=1}^{n}\mu_i^k\right)^{1/k}$. *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 5 may be further relaxed. Mirroring the argument of the Gaussian case, we have a matching lower bound. Proofs of these results are in Appendix C.

Theorem 6 (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* (3.7) *except with* μ_i *replaced with* $\mu_i - K$*.*

Remark 3. 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

$$\widehat{\mu}(y) := (y+1)\frac{N_{y+1}}{N_y} \tag{4.3}$$

where $N_y := \#\{i : Y_i = y\}$. The estimator can also be used for selection in the strict compound case. That is, select all indices i for which the Robbins estimator is greater than K. By rearranging $\widehat{\mu}(Y_i) \geq K$, one can see that truncating the Robbins estimator bears a passing resemblance to the ASSURE decision. However, truncating the Robbins estimator does not respect monotonicity. The estimator (4.1) regularizes and enforces this monotonicity, giving a more stable and intuitive alternative to decisions based on the Robbins estimator (Brown *et al.*, 2013; Jana *et al.*, 2022).

4.2. **Complex Decisions.** We discuss how to extend the ASSURE framework to accommodate complex, non-separable decision rules $\delta(\beta, Z_{1:n}, Y_{-i})$ which may depend on the outcomes and auxiliary information of other units. For example, an analyst might consider making decisions using empirical Bayes posterior mean estimates $\mathbb{E}_{\hat{G}}[\mu_i \mid \sigma_i, Y_i]$ and flexible machine-learning models $\hat{f}(X_i)$ that estimate the regression $\mathbb{E}[Y_i \mid X_i]$. In both of these cases, it would be desirable to estimate $\hat{G}$ and model $\hat{f}$ using the data itself.

By a similar argument to Proposition 1, one can create a near unbiased estimator of the welfare of this decision rule using a leave-one-out cross fitting construction. Redefine the ASSURE summand (2.5) with the decision rule $\delta(\beta, Z_{1:n}, Y_{-i})$, so that

$$w_h(Y_i; \beta, Z_i, Y_{-i}) := (Y_i - K_i)\,\mathrm{Csinc}\left(\frac{Y_i - \delta(\beta, Z_{1:n}, Y_{-i})}{\sigma_i h}\right) - \frac{\sigma_i}{h}\,\mathrm{sinc}\left(\frac{Y_i - \delta(\beta, Z_{1:n}, Y_{-i})}{\sigma_i h}\right).$$

Then,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \mathbb{E} w_h(Y_i; \beta, Z_i, Y_{-i}) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}\left[w_h(Y_i; \beta, Z_i, Y_{-i}) \mid Y_{-i}\right]\right] \\ &\approx \frac{1}{n}\sum_{i=1}^{n} \mu_i \mathbb{E}\left[\Phi\left(\mu_i - \delta(\beta, Z_{1:n}, Y_{-i})\right)\right],\end{aligned}$$

which can be interpreted as the average welfare of using the non-separable decision rule. Proposition 1 shows that the bias in the above approximation is small.

Given this observation, a simple construction is the following. Let $m_{\hat{G}_1}^{(-i)}(y, \sigma)$ be a posterior mean function obtained by applying an empirical Bayes method on the data (Y_{-i}, σ_{-i}). Let $m_{\hat{G}_2}^{(-i)}(y, \sigma)$ denote the same for a different empirical Bayes method. Finally, let $\widehat{f}^{(-i)}$ denote a machine learning model trained using the data (Y_{-i}, X_{-i}). Consider the class of decision rules which selects a unit whenever the ensemble prediction of these models is greater than the implementation cost:

$$b_1 m_{\hat{G}_1}^{(-i)}(Y_i, \sigma_i) + b_2 m_{\hat{G}_2}^{(-i)}(Y_i, \sigma_i) + b_3 \widehat{f}^{(-i)}(X_i) \geq K_i \tag{4.4}$$

where the parameters $\beta = (b_1, b_2, b_3)$ are restricted to be in the unit simplex. Extensions to multiple empirical Bayes posterior mean models or machine learning models is straightforward. By the monotonicity of the posterior mean function, the left hand side of (4.4) is increasing in Y_i: the decision is equivalent to $Y_i \geq \delta(\beta, Z_{1:n}, Y_{-i})$ for some function δ. As a result, ASSURE can be used to tune the parameter β. Under mild stability assumptions and for large enough samples, ASSURE is expected to perform no worse than the constituent models since the decision class nests each model individually. This ensemble class thus immediately enables ASSURE to improve on any fixed empirical Bayes decision method and illustrates the utility of the framework. We defer a detailed theoretical analysis of the ensembling method for future work.

5. Simulation Results

This section conducts two calibrated simulations where the μ_i's are constructed by sampling from some estimated empirical Bayes model. In each draw of the simulation, we then sample $Y_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$. Code to replicate these simulations and the empirical applications in the next section may be found on Github.

5.1. **Calibrated Opportunity Atlas Simulation.** We first consider a calibrated simulation based on the Opportunity Atlas (OA) dataset from (Chetty *et al.*, 2018; Bergman *et al.*, 2024). See Section 6.1 for further background on the underlying dataset. The data for this simulation comes from a simulation exercise of Chen (2025) and is generated by sampling $\mu_{1:n}, n \approx 10^4$, from a Monte Carlo sample of an empirical Bayes prior fitted on the OA dataset. We compare several classes of methods by evaluating the in-sample welfare obtained by the chosen decision using the known μ_i. In particular, we compare (a) the linear shrinkage class in Eq. (3.2), (b) the Fay–Herriot class of Eq. (3.3), and (c) the CLOSE-GAUSS decision class of Eq. (2.11). In each of these three classes, we compare (i) empirical Bayes with plug-in estimates of prior parameters against the decisions selected by (ii) ASSURE and (iii) coupled bootstrap (2.9). In addition we include two NPMLE methods, and finally an ensemble tuned using ASSURE as in Section 4. Complete details on the simulation setting are given in Appendix D, alongside further numerical results.

Figure 2 presents our results. For the linear shrinkage and Fay-Herriot classes, ASSURE consistently produces decisions that improve over the respective empirical Bayes baselines. These improvements highlight that ASSURE robustifies empirical Bayes procedures, improving performance when empirical Bayes models are misspecified. Indeed, these respective empirical Bayes methods assume that the μ_i are drawn from Gaussian priors with a fixed constant variance. This assumption is questionable in the present dataset, as can be seen in Figure 4. When the model is misspecified, ASSURE selects a better-performing set of parameters within the class. This gain can be significant, as seen in the linear shrinkage class.

For fairly well-specified empirical Bayes models, ASSURE does little harm and performs comparably to the respective empirical Bayes method, as shown by the comparison between methods CLOSE-GAUSS and the ASSURE-chosen decision in

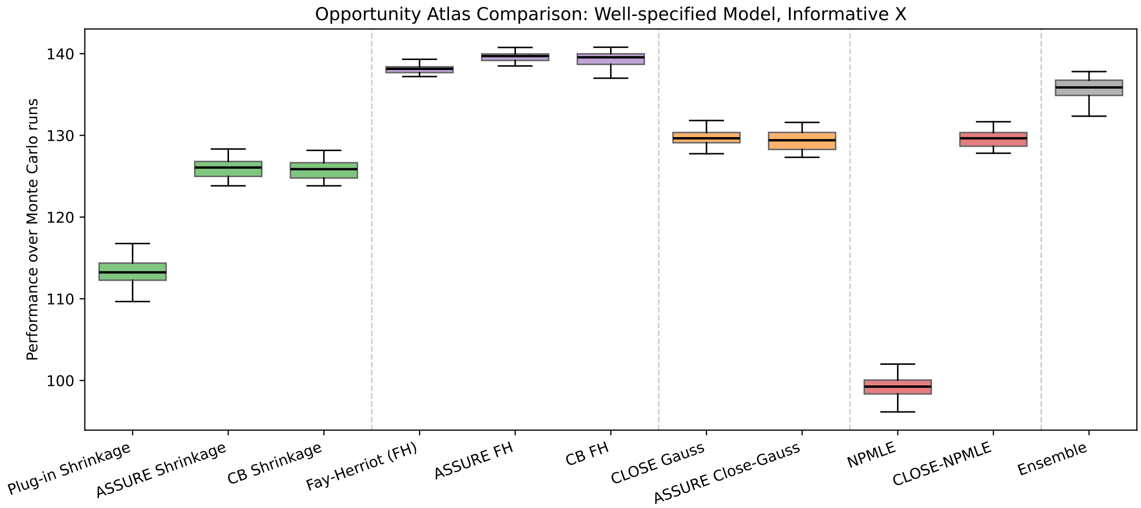


FIGURE 2. Semisynthetic simulation comparison on Opportunity Atlas dataset. Box and whisker plots summarize 40 Monte Carlo runs where only Y_i are redrawn. Constant costs are taken with $K = 0.361$, which is motivated in Section 6.1.

the class. As expected, ensembling produces decisions which outperform CLOSE-GAUSS and CLOSE-NPMLE. Because covariates are predictive of the true effects, we expect the ensemble method to strictly outperform the empirical Bayes counterpart. Across these simulations, ASSURE and coupled bootstrap have similar performance. Although theory suggests a better rate ($\tilde{O}(n^{-0.5})$ vs. $\tilde{O}(n^{-0.4})$), the difference appears negligible in our samples; it may nevertheless be economically meaningful at large n. The performance of our methods appears robust to the normality assumption, as shown in Figure 11. Appendix D details additional specifications where both normality is misspecified and the covariates are uninformative.

5.2. **Calibrated Experimentation Program Dataset.** Next, we present results on a semisynthetic dataset derived from the experimentation program dataset of Section 6.2. The specification is somewhat challenging due to smaller size of the dataset ($n \approx 330$) and the large heterogeneity in the σ_i, as seen in Figure 7. On this dataset, we evaluate three classes of decision rules: (a) the t-statistic/p-value threshold rules given in Example 2, (b) linear shrinkage rules, and (c) the Fay-Herriot decision rules. We also consider the performance of standard nonparametric maximum likelihood (Jiang, 2020).

Figure 3 shows the results. Importantly, both ASSURE and the coupled bootstrap method consistently identify a better choice of t-statistic threshold than the fixed

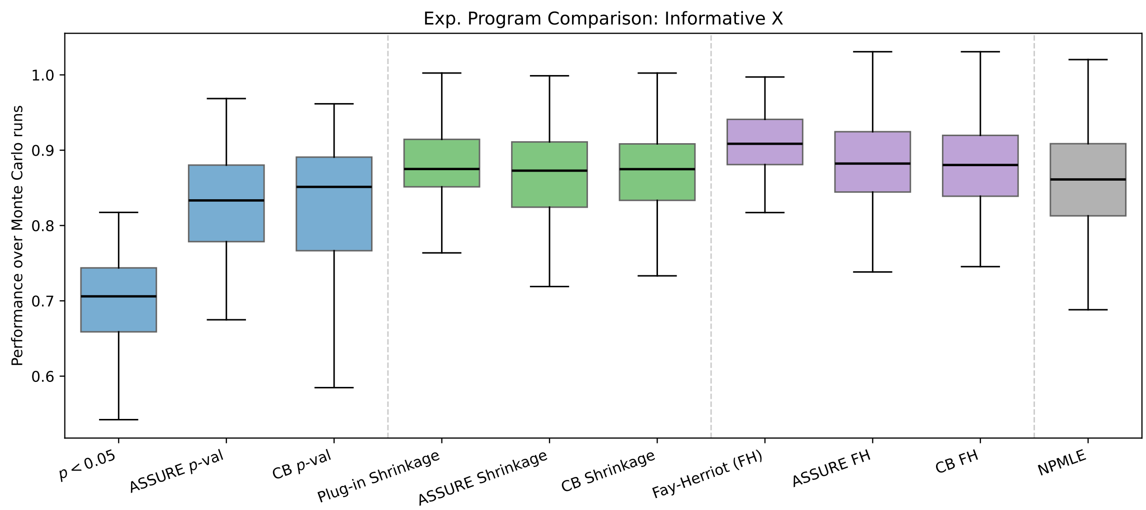


FIGURE 3. Experimentation program semisynthetic simulation, with $X_i = \mu_i + \sigma_i t_{10}$, $(\rho = 0.7)$. Boxplots show 100 Monte Carlo comparisons where Y_i is rerandomized.

decision rule corresponding to $p < 0.05$. The performance of both ASSURE and coupled bootstrap are comparable to the corresponding empirical Bayes methods in the linear shrinkage class and Fay–Herriot class, but somewhat more variable due to the smaller size of this dataset. In contrast to the data from the Opportunity Atlas in Section 6.1, the empirical Bayes assumptions underlying the plug-in methods are not unreasonable for this dataset, and thus ASSURE does not provide improvements. Nevertheless, even with a good empirical Bayes model, applying ASSURE does not significantly harm performance.

6. Empirical Applications

6.1. The Opportunity Atlas. The Opportunity Atlas (Chetty *et al.*, 2018) provides census-tract level estimates of economic mobility as measured by a suite of children's outcomes in adulthood. For each tract, multiple economic indicators such as earnings and incarceration rates are estimated by strata such as parental income, race, and sex. Building on the Opportunity Atlas dataset, Bergman *et al.* (2024) conducted the randomized trial *Creating Moves to Opportunity* (CMTO), offering low-income families housing vouchers and assistance to move to neighborhoods with historically high upward mobility. Implicit in defining highly upward-mobile tracts is a compound selection problem where the goal is to select the top 1/3 upwardly-mobile tracts. Bergman *et al.* (2024) and Chen (2025) approach this selection problem using empirical Bayes.

We consider the related selection task as defined in (1.1) on a particular economic mobility outcome pictured in Figure 4, where the mobility measure is the average adult income rank of children whose parents are at the 25th percentile of income. For each Census tract i, we decide whether to select it as high-mobility.

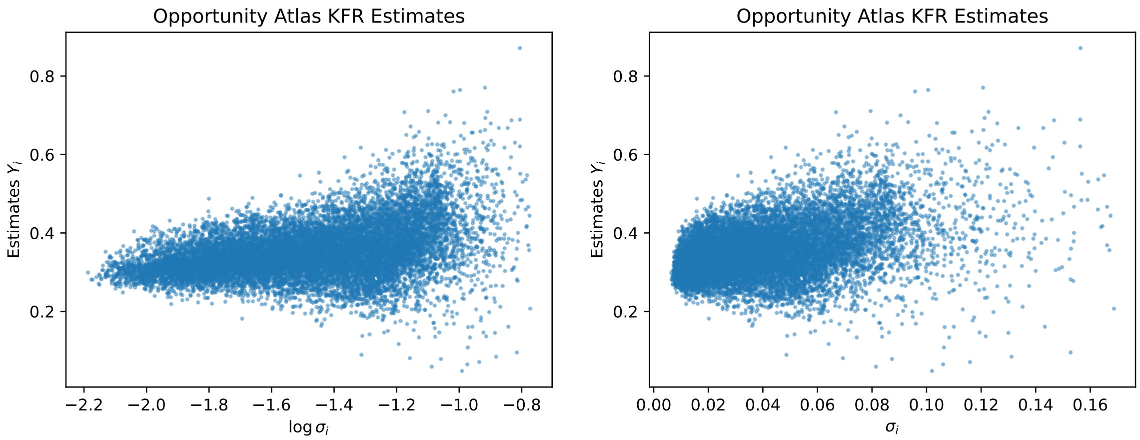


FIGURE 4. Visualization of Opportunity Atlas estimates from (Chetty *et al.*, 2018; Bergman *et al.*, 2024). Each point represents a census tract within the largest 20 Commuting Zones, measuring the household income rank in adulthood for Black children, with genders pooled, whose parents were at the 25th percentile of income.

We consider the class of linear shrinkage decision rules given in (3.2) with specifications of K_i to be discussed below. Recall that this class of decisions can be motivated by assuming that $\mu_i \sim \mathcal{N}(m_0, s_0^2)$, popular in many empirical economic contexts (Walters, 2024), which motivates the decision class:[23]

$$\mathcal{D}'_{\text{linear-shrink}} = \left\{ \delta(Z_i; \beta) = K + \beta \sigma_i^2 : \beta \in \mathbb{R} \right\}. \tag{6.1}$$

We compare fitting the plug-in empirical Bayes estimator discussed in Example 4 to the ASSURE chosen decision in the class (6.1). We consider three settings for costs K_i, which will be assumed constant for simplicity, given by $\{0.2, 0.361, 0.369\}$. The first specification corresponds roughly to the treatment cost in (Bergman *et al.*, 2024). The latter two choices are different specifications which emulate the top 1/3 selection exercise in Bergman. Details are given in Appendix D.

[23]With constant costs, the linear shrinkage decision class simplifies to this form.

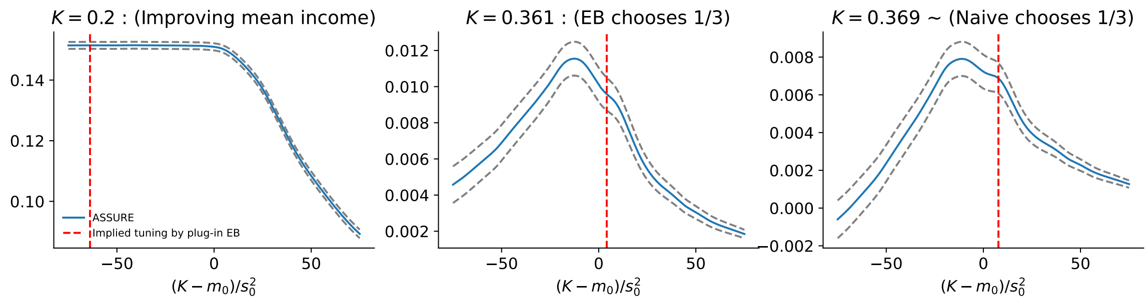


FIGURE 5. ASSURE estimate curves for linear shrinkage class with constant costs on the OA dataset of Figure 4. See text for an explanation of the different cost regimes. Grey lines indicate 1.96 times the standard errors of the ASSURE estimate, calculated pointwise.

Figure 5 shows the results. For $K = 0.2$, the majority of census tracts likely have a parameter μ_i which is greater than the cost $K = 0.2$, so the optimal decision is to select more tracts. In this case, the empirical Bayes plug-in and ASSURE chosen decision are comparable in performance. In the two other specifications $K \in \{0.361, 0.369\}$, the cost tradeoffs are more meaningful. The ASSURE welfare estimate suggests that the empirical Bayes plug-in is meaningfully suboptimal compared to the optimal decision, which takes β slightly less than zero. Figure 6 visualizes the difference, where ASSURE selects more aggressively.

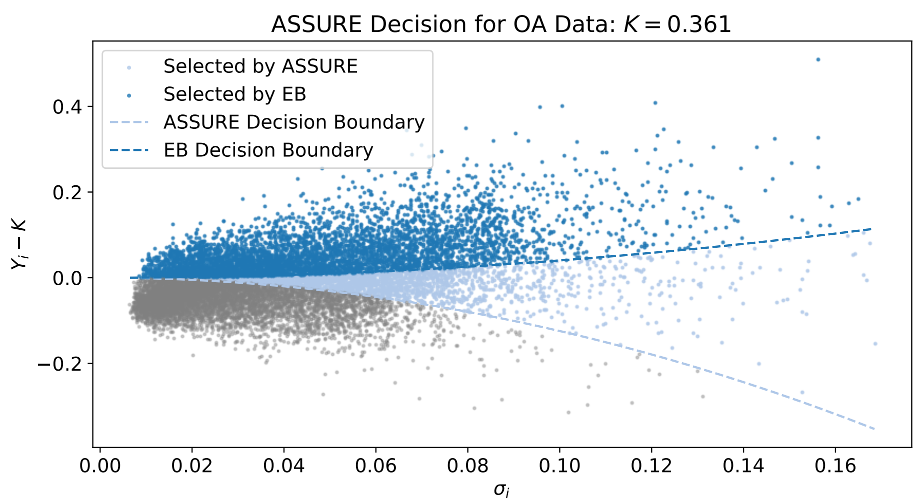


FIGURE 6. ASSURE decisions and empirical Bayes decisions overlayed on Opportunity Atlas data. $K = 0.361$, matching the middle panel of Figure 5.

This example shows that the ASSURE estimator has utility beyond just selecting a near-optimal decision through a black box. Consider again the left-hand side of Figure 5. The estimated curve suggests that decisions where $\beta \leq 0$ are all roughly equivalent in terms of expected welfare. Policy-makers can then choose a decision amongst this near-optimal subset on the basis of other criteria, such as to maximize another metric, or for interpretability. ASSURE therefore enables decision-makers to certify a status-quo decision if the estimated welfare is near-optimal. In other settings, it may also reveal that a status quo decision might be far from optimal, as discussed in the next empirical application.

6.2. **Experimentation Programs.** In the technology industry, experimentation programs are large collections of related A/B tests where the goal is to move a metric of interest (Azevedo *et al.*, 2020; Sudijono *et al.*, 2024; Chou *et al.*, 2025). By explicit randomization, the treatment effect estimates Y_i are plausibly normally distributed around their true unknown average treatment effects μ_i (Goldstein, 2007; Li and Ding, 2017). μ_i are measured in some metric of practical business interest. We take a dataset of anonymized treatment effects from Netflix (Sudijono *et al.*, 2024), consisting of treatment effects and standard deviations for 331 feature experiments. Figure 7 shows some descriptive statistics for the data.

Our goal is to select a subset S of innovations i to implement in the platform. Azevedo *et al.* (2020) explicitly introduce the problem of choosing a subset of indices i in order to maximize the welfare (1.1) in the Bayes setting where $\mu_{1:n} \overset{\text{i.i.d.}}{\sim} G$. Past work typically takes the empirical Bayes approach where G is estimated from large repositories of past A/B tests. In some cases, the Bayes assumption may be too strong. For example, innovations are often created as variants of the same idea, suggesting correlation between the returns to innovations. In new experimentation programs, no past data may be available to estimate G. In the absence of a trustworthy prior, the compound decision formulation is a compelling alternative.

The status quo decision procedures in the industry is the t-statistic decision class of Example 2. Traditionally, the chosen threshold corresponds to a one-sided p-value of 0.05. We use ASSURE to select the threshold β that maximizes the expected welfare of the decision $W(\beta) := \sum_{i=1}^{n}(\mu_i - K_i)\mathbb{P}(Y_i \geq \beta\sigma_i)$ with $K_i = 0$. Figure 8 shows the ASSURE estimate $\widehat{W}(\beta)$ as a function of the curve C. By maximizing the curve $\widehat{W}(\beta)$, ASSURE suggests to use a decision corresponding to $C \approx 0$.

The data in this application are anonymized: both the treatment effect estimates and standard deviations are obfuscated by random multiplicative factors. Thus, the

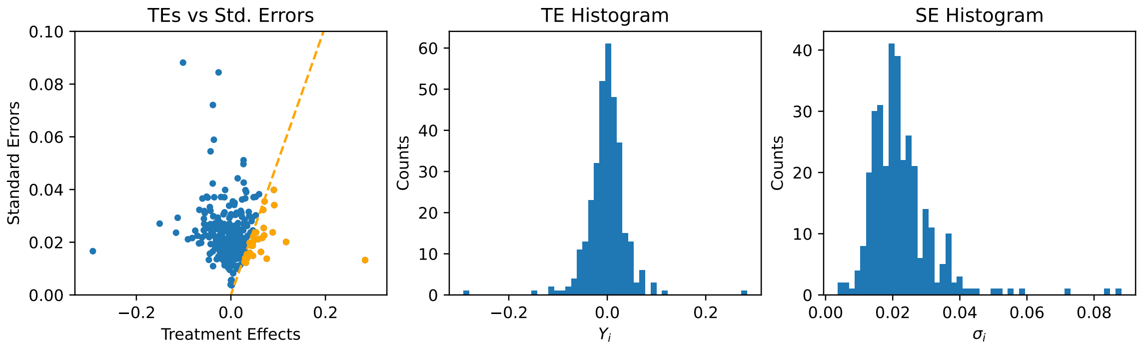


FIGURE 7. Left: Treatment effect estimate vs. standard deviation estimate for a dataset of 330 feature experiments from a large online tech company. The orange line and orange dots highlight tests which pass the traditional $p < 0.05$ criteria for statistical significance. Right: histogram of the treatment effects estimates. Both treatment effects and standard errors were multiplied with different random factors to preserve confidentiality.

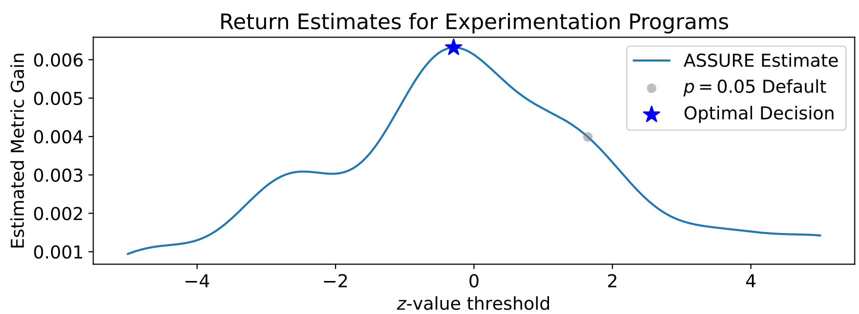


FIGURE 8. Blue line shows welfare estimate $\widehat{W}(\beta)$ for decisions of the form $Y_i \geq \beta\sigma_i$. Anonymized data on treatment effects from a single experimentation program. As such, the y-axis scale is obfuscated.

optimal decision cannot directly be interpreted, though we hope this section serves as a template for the reader to conduct similar exercises on their own datasets. Despite this, the shape of the estimated welfare curve suggests that a lenient threshold is more warranted than a conservative one, questioning the optimality of the industry-standard $p < 0.05$ decision rule. Other papers in the experimentation literature find similar conclusions under different frameworks Azevedo *et al.* (2020); Berman and Van den Bulte (2022); Sudijono *et al.* (2024).

As reflected in the literature, the suboptimality of $p < 0.05$ in the compound setting is not surprising when $K = 0$. In general, measuring costs for A/B tests is a difficult

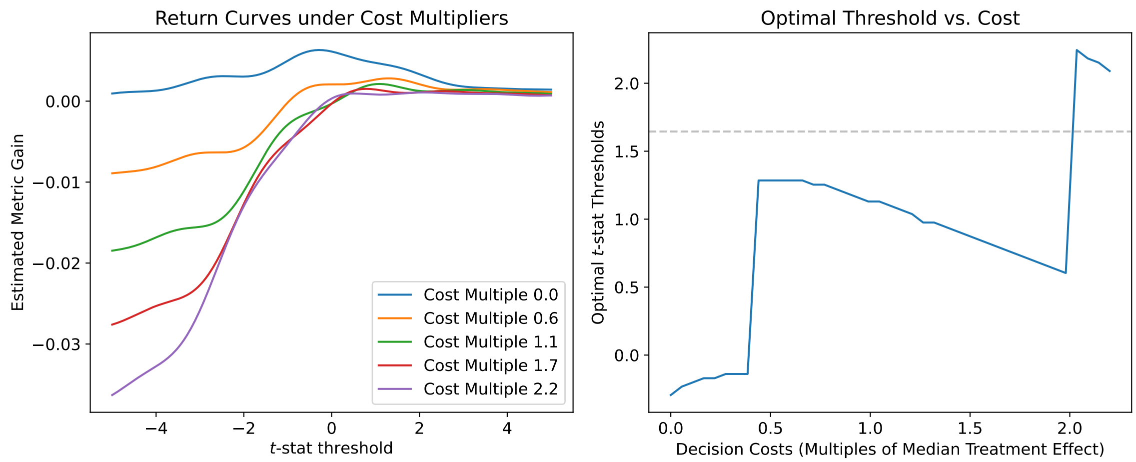


FIGURE 9. Left: ASSURE estimated welfare curves for t-statistic decision class $Y_i \geq K + \beta\sigma_i$ with constant costs K corresponding to multiples of the median absolute value of treatment effect. Right: optimal t-statistic decision threshold chosen by ASSURE. Grey dashed line corresponds to the $p < 0.05$ decision.

firm-specific exercise. Often costs are implicit, reflecting a host of factors including upfront costs of implementing and increasing future complexity of the platform. A useful exercise is to back out the implicit costs that justify the $p < 0.05$ decision rule and to use this as a heuristic benchmark when choosing a p-value threshold in practice. As a starting point, we set the cost for each decision equal to κ times the median treatment-effect size across the dataset, for various values of κ. The left panel shows the estimates $\widehat{W}(\beta)$ for various levels of κ. As κ grows, the optimal decision generally becomes more conservative. The right panel of Figure 9 shows the ASSURE-optimal t-statistic threshold as a function of the cost multiplier. The true optimal threshold is generally non-decreasing.[24] Crucially, Figure 9 suggests that the $p < 0.05$ rule can be rationalized by costs that are roughly twice the median treatment effect.

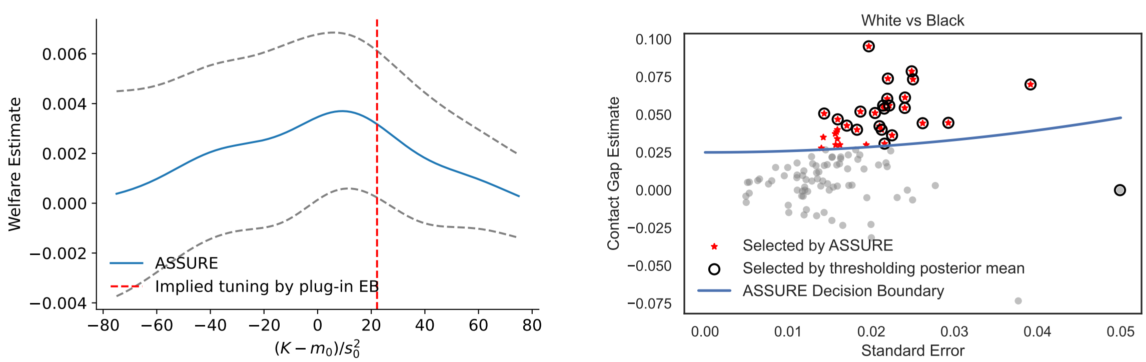
6.3. **Discrimination in Large Firms.** In this empirical illustration, we demonstrate how selection decisions at the individual level can differ when using ASSURE versus thresholding on nonparametric EB estimates. Kline *et al.* (2022) conducted a large correspondence experiment, sending up to 1,000 job applications to each of 108 firms. Here μ_i is the firm contact gap and Y_i is the average difference in callback

[24]The estimated threshold in Figure 9 is decreasing due to finite-sample effects and the form of the ASSURE estimator. The discontinuous jumps are however also present for the true optimal thresholds.

across pairs of resumes within a firm. Specifically, each firm in the sample had approximately 125 job listings throughout the experiment. For each job, 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 is then averaged over jobs to create Y_i. The normality assumption—also used in Kline et al. (2022)—is well justified by the paired randomization.

Based on an auditor with an "extensive margin" preference, Kline *et al.* (2022) found 23 firms to discriminate against Black applicants, controlling false discovery rates to the 5% level. To adapt their setting for our illustration, we instead consider an auditor with an "intensive margin" preference, whose objective is to maximize the expected welfare gain from an investigation selection rule S given by $\sum_{i=1}^{n}(\mu_i - K)\mathbb{P}\{i \in S\}$ where K is the cost of an investigation. Using the nonparametric EB estimates of Kline *et al.* (2022), adapted from Efron (2016), a Bayesian interpretation implies setting $K = 0.025$ to match with the selection of 23 firms,

To demonstrate how ASSURE can make different selections in practice, we consider the linear shrinkage class with constant cost $K = 0.025$ as specified in Eq. (6.1).



(A) Estimated Average Welfare (B) Selection

FIGURE 10. ASSURE and linear shrinkage rule on data from Kline *et al.* (2022)

As shown in Figure 10(b), comparing with the nonparametric EB estimates of Kline *et al.* (2022), the selections by ASSURE are largely concordant, including a few more but excluding a firm with noisy estimate. As argued before, the linear shrinkage class can also be motivated with a parametric normal prior. Figure 10(a) shows that directly plugging in estimates for (m_0, s_0) would give a higher threshold and select fewer firms than ASSURE.

7. Conclusion

In this work, we introduce ASSURE, a data-driven approach for compound selection decisions. By using a near-unbiased estimate of the welfare, ASSURE targets an instance-specific optimal decision while making only mild assumptions on the unknown parameters $\mu_{1:n}$. This robustifies empirical Bayes procedures where it may be difficult to justify distributional assumptions on μ_i. Section 6 discusses the pertinence of the Bayes assumption in our empirical applications. Another strength of ASSURE is its flexibility to validate a wide range of decision procedures, including those that use covariates and other auxiliary information. The construction of Section 4 allows decision-makers to improve on any black-box empirical Bayes decision by ensembling. In applications with large n, the ASSURE-optimal decision procedure consistently improves over plug-in empirical Bayes estimates. At the same time, the ASSURE estimator is also a useful tool for heuristic decision making: the nearly-unbiased welfare estimates produced by ASSURE allow a policy-maker to screen out clearly poorly-performing decision procedures or validate a fixed decision rule. For these theoretical and empirical reasons, we envision ASSURE as a useful tool for selection decisions in practice.

References

Andrews, I., Kitagawa, T. and McCloskey, A. (2024). Inference on winners. *The Quarterly Journal of Economics*, **139** (1), 305–358. 3

Armstrong, T. B., Kolesár, M. and Plagborg-Møller, M. (2022). Robust empirical bayes confidence intervals. *Econometrica*, **90** (6), 2567–2602. 3

Aschenbrenner, M., Dolich, A., Haskell, D., Macpherson, D. and Starchenko, S. (2013). Vapnik–chervonenkis density in some theories without the independence property, ii. 57

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, **89** (1), 133–161. 2, 12, 13

Audibert, J.-Y. and Tsybakov, A. B. (2005). Fast learning rates for plug-in classifiers under the margin condition. *arXiv preprint math/0507180*. 13, 17

— and — (2007). Fast learning rates for plug-in classifiers. 17

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, 5, 27, 28, 78

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, 5, 22, 24, 25, 88

Berman, R. and Van den Bulte, C. (2022). False discovery in a/b testing. *Management Science*, **68** (9), 6762–6782. 28

Bonhomme, S. and Denis, A. (2024). Estimating heterogeneous effects: applications to labor economics. *Labour Economics*, **91**, 102638. 5

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. 20

Chen, J. (2025). Empirical bayes when estimation precision predicts parameters. *arXiv preprint arXiv:2212.14444*. 2, 3, 4, 5, 9, 11, 16, 22, 25, 85, 86

Cheng, X., Ho, S. C. and Schorfheide, F. (2025). Optimal estimation of two-way effects under limited mobility. *arXiv preprint arXiv:2506.21987*. 2, 3, 4, 10

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. 3, 5, 22, 24, 25, 88

—, — 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. 4, 15, 17

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*. 4, 11

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. 2, 11

— (2016). Empirical bayes deconvolution estimates. *Biometrika*, **103** (1), 1–20. 30

— and Hastie, T. (2021). *Computer age statistical inference, student edition: algorithms, evidence, and data science*, vol. 6. Cambridge University Press. 19

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, 85

Goldstein, L. (2007). l^1 bounds in normal approximation. *The Annals of Probability*. 27

Gu, J. and Koenker, R. (2023). Invidious comparisons: Ranking and selection as compound decisions. *Econometrica*, **91** (1), 1–41. 4, 11

Gupta, S. S. and Li, J. (2005). On empirical bayes procedures for selecting good populations in a positive exponential family. *Journal of Statistical planning and Inference*, **129** (1-2), 3–18. 11

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. 2, 12

— and Porter, J. R. (2009). Asymptotics for statistical treatment rules. *Econometrica*, **77** (5), 1683–1701. 2, 12

— and — (2016). Panel asymptotics and statistical decision theory. *The Japanese Economic Review*, **67** (1), 33–49. 2

Ignatiadis, N. and Sun, D. L. (2025). Empirical bayes estimation via data fission. *Journal of the American Statistical Association*, **120** (549), 165–166. 3, 9

— and Wager, S. (2019). Covariate-powered empirical bayes estimation. *Advances in Neural Information Processing Systems*, **32**. 16

Ishihara, T. and Kurisu, D. (2022). Shrinkage methods for treatment choice. *arXiv preprint arXiv:2210.17063*. 12

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. 19

—, — and Wu, Y. (2022). Optimal empirical bayes estimation for the poisson model via minimum-distance methods. *arXiv preprint arXiv:2209.01328*. 19, 20

Jiang, W. (2020). On general maximum likelihood empirical bayes estimation of heteroscedastic iid normal means. 2, 4, 11, 23

— and Zhang, C.-H. (2009). General maximum likelihood empirical bayes estimation of normal means. 2

Karlin, S. and Rubin, H. (1956). The theory of decision procedures for distributions with monotone likelihood ratio. *The Annals of Mathematical Statistics*, pp. 272–299. 3

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, 11

Kitagawa, T. and Tetenov, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, **86** (2), 591–616. 2, 4, 12, 13, 17

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, 5, 29, 30

Kline, P. M., Rose, E. K. and Walters, C. R. (2024). Discrimination report cards: An empirical bayes ranking approach. 4, 5, 11

Koenker, R. and Gu, J. (2017). Rebayes: an r package for empirical bayes mixture methods. *Journal of Statistical Software*, **82**, 1–26. 85

— and — (2024). Empirical bayes for the reluctant frequentist. *arXiv preprint arXiv:2404.03422*. 13

Kolmogorov, A. N. (1950). Unbiased estimates. *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya*, **14** (4), 303–326. 6, 9

Kwon, S. (2023). Optimal shrinkage estimation of fixed effects in linear panel data models. *arXiv preprint arXiv:2308.12485*. 2, 3, 4, 10

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*. Springer. 13

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. 9

Li, X. and Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, **112** (520), 1759–1769. 27

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, 11

— (2000). On an empirical bayes test for a normal mean. *Annals of statistics*, pp. 648–655. 2, 6, 11, 17

Liang, T. C. (2004). On optimal convergence rate of empirical bayes tests. *Statistics & probability letters*, **68** (2), 189–198. 2, 11, 19

Luo, J., Banerjee, T., Mukherjee, G. and Sun, W. (2023). Empirical bayes estimation with side information: A nonparametric integrative tweedie approach. *arXiv preprint arXiv:2308.05883*. 16

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica*, **72** (4), 1221–1246. 2, 12

— (2009). *Identification for prediction and decision*. Harvard University Press. 12

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

— (2021). Econometrics for decision making: Building foundations sketched by haavelmo and wald. *Econometrica*, **89** (6), 2827–2853. 2, 11

Mbakop, E. and Tabord-Meehan, M. (2021). Model selection for treatment choice: Penalized welfare maximization. *Econometrica*, **89** (2), 825–848. 13

Mogstad, M., Romano, J. P., Shaikh, A. M. and Wilhelm, D. (2024). Inference for ranks with applications to mobility across neighbourhoods and academic achievement across countries. *Review of Economic Studies*, **91** (1), 476–518. 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*. 19

Moon, S. (2025). Optimal policy choices under uncertainty. 2

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. 3, 9

Onshuus, A. and Quiroz, A. J. (2015). Metric entropy estimation using o-minimality theory. *arXiv preprint arXiv:1511.07098*. 57

Pensky, M. (2017). Minimax theory of estimation of linear functionals of the deconvolution density with or without sparsity. 6, 9

Polyanskiy, Y. and Wu, Y. (2021). Sharp regret bounds for empirical bayes and compound decision problems. *arXiv preprint arXiv:2109.03943*. 17, 19

Ponomarev, K. and Semenova, V. (2024). On the lower confidence band for the optimal welfare. *arXiv preprint arXiv:2410.07443*. 17

Robbins, H. (1951/1985). Asymptotically subminimax solutions of compound statistical decision problems. *Herbert Robbins Selected Papers*, pp. 7–24. 2, 13

Schönberg, I. J. (1948). On variation-diminishing integral operators of the convolution type. *Proceedings of the National Academy of Sciences*, **34** (4), 164–169. 76

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. 2, 11

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. 3, 8, 9, 89, 90

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pp. 1135–1151. 3, 10

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

Sudijono, T., Ejdemyr, S., Lal, A. and Tingley, M. (2024). Optimizing returns from experimentation programs. *arXiv preprint arXiv:2412.05508*. 4, 15, 27, 28

Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, **102** (479), 901–912. 13

Tate, R. F. (1959). Unbiased estimation: functions of location and scale parameters. *The Annals of Mathematical Statistics*, pp. 341–366. 6, 9

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

VAART, A. W. and WELLNER, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes*, Springer, pp. 16–28. 45, 46

VAN DEN DRIES, L. (1998). *Tame topology and o-minimal structures*, vol. 248. Cambridge university press. 57

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. 45, 46, 52, 53, 55, 57, 68

VERSHYNIN, R. (2009). High-dimensional probability. 14

WALD, A. (1950). Statistical decision functions. In *Breakthroughs in Statistics: Foundations and Basic Theory*, Springer, pp. 342–357. 3

WALTERS, C. (2024). Empirical bayes methods in labor economics. In *Handbook of Labor Economics*, vol. 5, Elsevier, pp. 183–260. 2, 4, 25

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. 4

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. 57

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. 3, 10

ZHOU, F. and LI, P. (2019). A fourier analytical approach to estimation of smooth functions in gaussian shift model. *arXiv preprint arXiv:1911.02010*. 6, 9

Online Appendix

November 17, 2025

Contents

Appendix A. Proofs for Sec. 2

Throughout, define

$$\Psi_h(Y_i; Z_i, C) := (Y_i - K_i)\,\mathrm{Csinc}\left(\frac{Y_i - C}{\sigma_i h}\right) - \frac{\sigma_i}{h}\,\mathrm{sinc}\left(\frac{Y_i - C}{\sigma_i h}\right) \tag{A.1}$$

Proof of Proposition 1. We leverage the fact that for all $a \in \mathbb{R}$,

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

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. Then

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

The last step follows from DCT, noting that

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

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 A.1:

$$\frac{1}{\pi}(\mu-K)\lim_{h\downarrow 0}\int_0^{1/h}\frac{1}{\omega}\left(\int_{-\infty}^{\infty}\sin(\omega(y-C)/\sigma)\frac{1}{\sigma}\varphi\left(\frac{y-\mu}{\sigma}\right)dy\right)d\omega$$
$$=\frac{1}{\pi}\lim_{h\downarrow 0}\int_0^{1/h}\frac{1}{\omega}e^{-\frac{1}{2}\omega^2}(\mu-K)\sin(\omega(\mu-C)/\sigma)d\omega$$
$$=\frac{1}{\pi}\int_0^{\infty}\frac{1}{\omega}e^{-\frac{1}{2}\omega^2}(\mu-K)\sin(\omega(\mu-C)/\sigma)d\omega$$
$$=\frac{1}{\pi}\lim_{h\downarrow 0}\int_0^{1/h}\frac{1}{\omega}\int_{-\infty}^{\infty}\left[(y-k)\sin\left(\frac{\omega(y-C)}{\sigma}\right)-\omega\sigma\cos\left(\frac{\omega(y-C)}{\sigma}\right)\right]\frac{1}{\sigma}\varphi\left(\frac{y-\mu}{\sigma}\right)dyd\omega$$
$$=\frac{1}{\pi}\lim_{h\downarrow 0}\int_0^{1/h}\int_{-\infty}^{\infty}\left[\frac{(y-k)\sin\left(\frac{\omega(y-C)}{\sigma}\right)}{\omega}-\sigma\cos\left(\frac{\omega(y-C)}{\sigma}\right)\right]\frac{1}{\sigma}\varphi\left(\frac{y-\mu}{\sigma}\right)dyd\omega.$$

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

$$\lim_{h\downarrow 0}\frac{1}{\pi}\int_{-\infty}^{\infty}\int_0^{1/h}\left[(y-k)\frac{\sin(\omega(y-C)/\sigma)}{\omega}-\sigma\cos(\omega(y-C)/\sigma)\right]d\omega\frac{1}{\sigma}\varphi\left(\frac{y-\mu}{\sigma}\right)dy$$
$$=\lim_{h\downarrow 0}\frac{1}{\pi}\int_{-\infty}^{\infty}\left[\int_0^{1/h}(y-k)\frac{\sin(\omega(y-C)/\sigma)}{\omega}d\omega-\sigma\frac{\sin((y-C)/h\sigma)}{(y-C)/\sigma}\right]\frac{1}{\sigma}\varphi\left(\frac{y-\mu}{\sigma}\right)dy$$
$$=\lim_{h\downarrow 0}\mathbb{E}\left[\frac{1}{\pi}(y-k)\,\mathrm{Si}\left(\frac{Y-C}{h\sigma}\right)-\frac{\sigma}{h}\,\mathrm{sinc}\left(\frac{Y-C}{h\sigma}\right)\right]$$

To finish the proof, note in the previous line of equalities that we showed

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

interpreted as Lebesgue integrals and $Z=(K,\sigma)$. Thus

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

which has an upper bound given by

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

The last line follows from the Mills ratio bound. This bound holds uniformly over C, σ. As a result, the bias for $\widehat{W}_h$ 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 Poincare inequality, which states that for $Y \sim \mathsf{N}(0,1)$, $\operatorname{Var} f(Y) \le \mathbb{E}\left[f'(Y)^2\right]$ where f is a differentiable function. Let $\tilde{Y} \sim \mathsf{N}(\mu, \sigma^2)$. Applying the Gaussian Poincare inequality to the function $f(\mu + \sigma x)$, we find

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

It is not difficult to see 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

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

Recall from the definition of Csinc that it is uniformly bounded since the Sine integral function is. From the calculation in Lemma E.2, it is easily seen that sinc, 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 Poincare inequality with w_h and using Eq. (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 $\widehat{W}_n(\beta)$ 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-square error gives the result.

□

Lemma A.1. *Let $Y \sim \mathsf{N}(\mu, \sigma^2)$ 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 Lemma A.2 and then apply the first two equalities of the lemma. □

Lemma A.2 (Gaussian IBP)**.** *Let $Y \sim N(\mu, \sigma^2)$. Then Gaussian integration by parts states that for any function f differentiable we have*

$$\mathbb{E}[Yf(Y)] = \mu\mathbb{E}f(Y) + \sigma^2\mathbb{E}[f'(Y)]$$

where the expectation is with respect to Y.

A.1. **Additional details for Section 2.2.** In this subsection, we derive the counterpart of Theorem 1 for coupled bootstrap to corroborate the rate comparison in Section 2.2. Define the coupled-bootstrap estimator as

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

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

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

where $\omega_i \sim N(0,1)$ and $\omega_i \perp\!\!\!\perp Y_i, Z_i$. Define the welfare function of our interest.

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

Theorem A.1 (Main Regret Bound for Coupled Bootstrap)**.** *Let $\hat{\beta}(\epsilon)$ be the maximizer of $\widehat{W}_{CB}(\beta;\epsilon)$, $\beta^*(\epsilon)$ be the maximizer of $W(\beta;\epsilon)$ and β_0 be the maximizer of $W(\beta;0)$. Under same conditions of Theorem 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

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

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

Proof of Theorem A.1. With $\epsilon_n \to 0$, from Proposition A.1, we have

$$W(\beta^*(\epsilon_n);\epsilon_n) = W(\beta_0;0) - O(\epsilon_n^2).$$

From Proposition A.2, we have

$$\mathbb{E}\left[V_n(\hat{\beta};\epsilon_n)\right] = W(\beta^*(\epsilon_n);\epsilon_n) - \max\left\{O\left(\frac{1}{\sqrt{n}}\right), O\left(\frac{\epsilon_n^{-\frac{1}{2}}\left(-\log\epsilon_n\right)^{\frac{1}{2}}}{\sqrt{n}}\right), O\left(\frac{\epsilon_n^{-1}\left(-\log\epsilon_n\right)}{n}\right)\right\}$$

Combine the two results and we conclude the proof. □

Proposition A.1. *Assume $Z_i = \sigma_i$. Under the same assumption of Theorem 1, if we let $\epsilon_n \to 0$, then we have*

$$W(\beta^*(\epsilon_n);\epsilon_n) = W(\beta_0;0) - \|D\|_{n,2}O(\epsilon_n^2).$$

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

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

By the mean value theorem and the envelope theorem,

$$
\begin{aligned}
&|W(\beta^*(\epsilon);\epsilon) - W(\beta_0;0)| = \left|\frac{d}{d\epsilon}W(\beta^*(\epsilon);\epsilon)\big|_{\epsilon=\epsilon_1}\right| |\epsilon - 0| \\
&= \left|\frac{1}{n}\sum_{i=1}^{n}(\mu_i - K_i)\,\varphi\left(\frac{\mu_i - \delta(\beta^*(\epsilon_1), Z_i)}{\sqrt{1+\epsilon_1^2}\sigma_i}\right)\frac{\mu_i - \delta(\beta^*(\epsilon_1), Z_i)}{\sigma_i}\right| \frac{1}{(1+\epsilon_1^2)^{3/2}}\epsilon_1\epsilon \\
&\lesssim \|D\|_{n,2}\frac{1}{(1+\epsilon_1^2)^{3/2}}\epsilon_1\epsilon \leq \epsilon^2\|D\|_{n,2}.
\end{aligned}
$$

□

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

$$
\mathbb{E}\left[V_n(\hat{\beta};\epsilon_n)\right] = W(\beta^*(\epsilon_n);\epsilon_n) - \max\left\{O\left(\frac{1}{\sqrt{n}}\right), O\left(\frac{\epsilon_n^{-\frac{1}{2}}(-\log\epsilon_n)^{\frac{1}{2}}}{\sqrt{n}}\right), O\left(\frac{\epsilon_n^{-1}(-\log\epsilon_n)}{n}\right)\right\}.
$$

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

$$
\begin{aligned}
&W(\beta^*(\epsilon_n);\epsilon_n) - \mathbb{E}\left[V_n(\hat{\beta};\epsilon_n)\right] \leq 2\mathbb{E}\left[\sup_\beta\left|\widehat{W}_{\mathrm{CB}}(\beta;\epsilon_n) - V_n(\beta;\epsilon_n)\right|\right] \\
&= 2\mathbb{E}\left[\sup_\beta\left|\left(\widehat{W}_{\mathrm{CB}}(\beta;\epsilon_n) - W(\beta;\epsilon_n)\right) - \left(V_n(\beta;\epsilon_n) - W(\beta;\epsilon_n)\right)\right|\right] \\
&= 2\mathbb{E}\left[\sup_\beta\left|\left(\widehat{W}_{\mathrm{CB}}(\beta;\epsilon_n) - \mathbb{E}\left[\widehat{W}_{\mathrm{CB}}(\beta;\epsilon_n)\right]\right) - \left(V_n(\beta;\epsilon_n) - \mathbb{E}\left[V_n(\beta;\epsilon_n)\right]\right)\right|\right] \\
&\leq 2\mathbb{E}\left[\sup_\beta\left|\widehat{W}_{\mathrm{CB}}(\beta;\epsilon_n) - \mathbb{E}\left[\widehat{W}_{\mathrm{CB}}(\beta;\epsilon_n)\right]\right|\right] + 2\mathbb{E}\left[\sup_\beta\left|V_n(\beta;\epsilon_n) - \mathbb{E}\left[V_n(\beta;\epsilon_n)\right]\right|\right] \\
&\leq 2\mathbb{E}\left[\sup_\beta\left|\frac{1}{n}\sum_{i=1}^{n}(Y_i - K_i)\,\Phi\left(\frac{Y_i - \delta(Z_i;\beta)}{\epsilon_n\sigma_i}\right) - \mathbb{E}\left[(Y_i - K_i)\,\Phi\left(\frac{Y_i - \delta(Z_i;\beta)}{\epsilon_n\sigma_i}\right)\right]\right|\right] \\
&\quad + 2\frac{1}{\epsilon_n}\mathbb{E}\left[\sup_\beta\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\varphi\left(\frac{Y_i - \delta(Z_i;\beta)}{\epsilon_n\sigma_i}\right) - \mathbb{E}\left[\sigma_i\varphi\left(\frac{Y_i - \delta(Z_i;\beta)}{\epsilon_n\sigma_i}\right)\right]\right|\right] \\
&\quad + 2\mathbb{E}\left[\sup_\beta\left|V_n(\beta;\epsilon_n) - \mathbb{E}\left[V_n(\beta;\epsilon_n)\right]\right|\right] \\
&= 2\left(T_{1,n} + \frac{1}{\epsilon_n}T_{2,n} + T_{3,n}\right).
\end{aligned}
$$

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 Vaart and Wellner (1996) from multiplicity. Since $t \mapsto \Phi(t)$ is a monotone function, by Lemma 2.6.18.viii of Vaart and Wellner (1996) and by Lemma 2.6.18.vi of 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: $|f_\beta(y,z)| \le |y-k|$, from Theorem 2.6.7 of Vaart and Wellner (1996), we have for the covering number

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

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:

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

where s_k, m_k 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, $\left\{\left|\frac{y-\delta(\beta,z)}{\epsilon_n\sigma}\right|\right\}$ is a VC subgraph class with dimension $V(\mathcal{D})$ by multiplicity (Lemma 2.6.18.vi of Vaart and Wellner (1996)) and part 5 of Lemma B.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 Vaart and Wellner (1996) and by Lemma 2.6.18.vi of Vaart and Wellner (1996) with multiplicity, $\mathcal{F}_2$ is a VC-subgraph class with dimension $V(\mathcal{D})$.

Since $\mathcal{F}_1$ is bounded $|f_\beta(y,z)| \leq M$, from Theorem 2.6.7 of Vaart and Wellner (1996), we have for the covering number

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

where K is a universal constant. For the entropy

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

we have $J(b, \mathcal{G}, L_2) = O(b\sqrt{\log(1/b)})$ as $b \downarrow 0$ (van der Vaart and Wellner, 2023, Before Theorem 2.14.1, P. 330)

Since Lemma A.3 shows that $\mathbb{E}f^2 \leq b_n^2 M^2$, with $b_n \sim \epsilon_n^{1/2}$ for $\forall f \in \mathcal{F}_2$, applying Theorem 2.14.2 in van der Vaart and Wellner (2023), we have

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

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

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

By the additive, multiplicative and monotone composition properties of VC-subgraph classes (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: $|f_\beta(y,z,\mu,\omega)| \leq |\mu - k|$, from Theorem 2.6.7 of Vaart and Wellner (1996) and the maximal inequality similar to Theorem 2.14.1 of van der Vaart and Wellner (2023), we have

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

where s_k, m_k are defined before Theorem 1, and A_1, A_2, A_3 are universal constants.

As a conclusion, we have the bound:

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

□

Lemma A.3. *Under the same assumption as Theorem 1, for $\epsilon_n \to 0$, we have*

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

Proof. We have

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

□

Appendix B. Proofs for Section 3

Proof of Thm 1. Starting from (3.1), we use the bias bound in Proposition 1 to bound:

$$(3.1) \lesssim \mathbb{E}\left[\sup_\beta |u(\beta) - W(\beta)| \right] + \mathbb{E}\left[\sup_\beta |\widehat{W}(\beta) - \mathbb{E}\widehat{W}(\beta)| \right] + \underbrace{\frac{m_1}{n^2 \log n}}_{\text{by Proposition 1}}.$$

Letting

$$A_1(\beta) := u(\beta) - W(\beta)$$
$$A_2(\beta) := \widehat{W}(\beta) - \mathbb{E}\widehat{W}(\beta),$$

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 the next two propositions (Propositions B.1 and B.2), which prove that

$$\mathbb{E}\sup_\beta |A_1(\beta)| \lesssim m_2 \frac{\sqrt{V(\mathcal{D})}}{\sqrt{n}}$$
$$\mathbb{E}\sup_\beta |A_2(\beta)| \lesssim \frac{\log n}{\sqrt{n}}(s_2 + m_2 + (\nu_4+1)\sqrt{V(\mathcal{D})}(\mathbb{E}\|D\|_n + (\mathbb{E}\|D\|_{4,n}^2)^{1/2})).$$

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

$$\mathsf{Regret}_n \leq (3.1) \lesssim M\frac{\log n}{\sqrt{n}}$$

with $M := \sqrt{V(\mathcal{D})}(s_2 + m_2 + (1+\nu_4)(\mathbb{E}\|D\|_n + (\mathbb{E}\|D\|_{4,n}^2)^{1/2}))$ as desired. □

Proposition B.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 INID 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 B.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 B.1(2) gives the claim. Next, we apply Theorem B.1 and Lemma B.2

$$\begin{aligned}
\mathbb{E}\sup_\beta |A_1(\beta)| &= \frac{1}{\sqrt{n}}\mathbb{E}\sup_\beta \frac{n}{\sqrt{n}}|A_1(\beta)| = \frac{1}{\sqrt{n}}\mathbb{E}\sup_\beta \frac{1}{\sqrt{n}}\sum_{i=1}^n \{f_{Y_i,\mu_i,Z_i} - \mathbb{E}[f(Y_i,\mu_i,Z_i)]\} \\
&\leq \frac{1}{\sqrt{n}} J(1, \mathcal{F}_1 \mid F, \mathbb{L}^2)\|F\|_P && \text{(Theorem B.1)} \\
&\lesssim \frac{1}{\sqrt{n}}\left(\frac{1}{n}\sum_{i=1}^n |\mu_i - K_i|^2\right)^{1/2} \int_0^1 \sqrt{1 + V(\mathcal{D})\ln\frac{1}{\epsilon}}d\epsilon && \text{(Lemma B.2)} \\
&\lesssim \sqrt{V(\mathcal{D})}\left(\frac{1}{n}\sum_{i=1}^n (\mu_i - K_i)^2\right)^{1/2}\frac{1}{\sqrt{n}}.
\end{aligned}$$

□

Proposition B.2. *We have*

$$\mathbb{E}\left|\sup_\beta A_2(\beta)\right| \lesssim \frac{\log n}{\sqrt{n}}(s_2 + m_2 + (\nu_4+1)\sqrt{V(\mathcal{D})}(\mathbb{E}\|D\|_n + (\mathbb{E}\|D\|_{4,n}^2)^{1/2})).$$

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

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

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 INID empirical process theory. The function class $\mathcal{G}$ has a uniform envelope function $G = c|y-K|$ for some absolute constant c, since the $\mathrm{Si}(\cdot)$ function is bounded. Following the proof of Theorem B.1, we may apply symmetrization and a maximal inequality to obtain

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

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

Now, we claim that for some constant C,

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

where $\widehat{\nu}_4 := \left(\frac{1}{n}\sum_{i=1}^n\frac{(Y_i-K_i)^4}{\sigma_i^4}\right)^{1/4}$. 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,

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

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 $C\lambda_n\widehat{\nu}_4\epsilon$-cover in the $\mathbb{L}^2(\mathbb{P}_n)$ seminorm.

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

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

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 $\epsilon \ge 1$ is equal to 1, so we may bound the integral for $\epsilon \in [1, \eta_n/C\lambda_n\widehat{\nu}_4\|D\|_{4,n})$ by $(\eta_n/C\lambda_n\widehat{\nu}_4\|D\|_{4,n} - 1)_+$. Applying the envelope function bound for $\mathcal{G}$ and Lemma B.2,

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

Taking expectations over the data and applying Cauchy–Schwarz, we obtain

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

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

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

$$\mathbb{E}\widehat{\nu}_4^2 = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \frac{(Y_i - K_i)^4}{\sigma_i^4}\right)^{1/2} \lesssim \left(1 + \frac{1}{n}\sum_{i=1}^n (\mu_i - K_i)^4/\sigma_i^4\right)^{1/2} \lesssim 1 + \nu_4^2.$$

Thus

$$\begin{aligned} &\mathbb{E}\sup_\beta \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n g_\beta(Y_i, \mu_i, Z_i) - \mathbb{E}g_\beta(Y_i, \mu_i, Z_i)\right| \\ &\lesssim \left(s_2 + m_2 + \log n\sqrt{V(\mathcal{D})}(1+\nu_4)(\mathbb{E}\|D\|_{4,n}^2)^{1/2}\right). \end{aligned}$$

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 B.1 again, we obtain

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

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

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

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

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

Combining the pieces, we have

$$\begin{aligned} \mathbb{E}\left|\sup_\beta A_2(\beta)\right| &\lesssim \left(s_2 + m_2 + \log n\sqrt{V(\mathcal{D})}(1+\nu_4)(\mathbb{E}\|D\|_{4,n}^2)^{1/2}\right) \\ &\quad + \log n\left(s_2 + \sqrt{V(\mathcal{D})}\mathbb{E}\|D\|_n\right) + s_2 \\ &\lesssim (\log n)(s_2 + m_2 + (\nu_4 + 1)\sqrt{V(\mathcal{D})}(\mathbb{E}\|D\|_n + (\mathbb{E}\|D\|_{4,n}^2)^{1/2})). \end{aligned}$$

□

B.1. **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 B.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 $g(y, y) : X \times Y \to \mathbb{R}$ such that $g(y, y) = f(x)$. 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 intersection of the subgraphs of $f, -f$. Combining Lemma 2.6.20 parts ii and iv completes the proof. □

Lemma B.2 (VC Function Covering, Thm 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 $0 < \epsilon < 1$.

Write $\mathbb{G}_n$ to be the empirical process $\sqrt{n}(\mathbb{P}_n - P)$ so that $\mathbb{G}_n f = \frac{1}{\sqrt{n}} \sum_{i=1}^n (f(Y_i) - Pf)$. The next theorem records a uniform entropy bound from (van der Vaart and Wellner, 2023); we claim that it also applies to the i.n.i.d setting.

Theorem B.1 (Theorem 2.14.1 of van der Vaart and Wellner (2023))**.** *Let $\mathcal{F}$ be a P-measurable class of functions with measurable envelope function F. Then for $p \geq 1$,*

$$\| \|\mathbb{G}_n\|_{\mathcal{F}}^* \|_{P,p} \lesssim \|J(z_n, \mathcal{F} \mid F, L_2) \|F\|_n\|_{P,p} \lesssim J(1, \mathcal{F} \mid F, L_2) \|F\|_{P,2\vee p}$$

where $\|g\|_n^2 = \frac{1}{n}\sum_{i=1}^n g(Y_i)^2$ *is the* $\mathbb{L}^2(\mathbb{P}_n)$*-seminorm and* $z_n = \||\|f\|_n\|_{\mathcal{F}}/\|F\|_n$.

Proof. The proof is verbatim from van der Vaart and Wellner (2023). Denote by $\mathbb{G}_n^o$ the symmetrized process given by

$$\mathbb{G}_n^o f := \frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(Z_i).$$

The same argumentation as Theorem 2.14.1 of van der Vaart and Wellner (2023) shows that conditional on $Z_1, \ldots, Z_n$, the sub-Gaussian norm of the symmetrized process satisfies

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

where $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(Z_i)$ and

$$\eta_n := \|\|f\|_n\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}}\sqrt{\frac{1}{n}\sum_{i=1}^n f(Z_i)^2}.$$

A change of variable in the integral by $\epsilon \leftarrow \epsilon\|F\|_n$ gives

$$\begin{aligned}\|\|\mathbb{G}_n^o\|_{\mathcal{F}}\|_{\psi_2|Z} &\lesssim \|F\|_n \int_0^{\eta_n/\|F\|_n} \sqrt{1 + \log N(\epsilon\|F\|_n, \mathcal{F}, L_2(\mathbb{P}_n))}d\epsilon \\ &\le \|F\|_n J(\theta_n, \mathcal{F} \mid F, L_2)\end{aligned}$$

The rest of the proof follows as is. □

Proposition B.3 (Orlicz norm bounds, INID analogue of Theorem 2.14.23 van der Vaart and Wellner (2023))**.** *Let* $\mathcal{F}$ *be a class of measurable functions with measurable envelope function* F*. Then*

$$\|\|\mathbb{G}_n\|_{\mathcal{F}}^*\|_{P,\psi_2} \lesssim \mathbb{E}^*\|\mathbb{G}_n\|_{\mathcal{F}} + \left(\frac{1}{n}\sum_{i=1}^n \|F(Z_i)\|_{\psi_2}^2\right)^{1/2} \qquad (1 < p \le 2)$$

Proof. We recall Proposition A.1.6 of van der Vaart and Wellner (2023). Let $X_1, \ldots, X_n$ be independent mean zero stochastic processes indexed by an arbitrary set T. Set $S_n = \sum_{i=1}^n Y_i$. Then we have

$$\|\|S_n\|^*\|_{\psi_2} \le K\left[\|\|S_n\|^*\|_{P,1} + \left(\sum_{i=1}^n \|\|Y_i\|^*\|_{\psi_2}^2\right)^{1/2}\right],$$

where $\|X\|^*$ is shorthand for $\sup_t |X_t|$. Take $Y_i = \frac{1}{\sqrt{n}}(f(Z_i) - \mathbb{E}f(Z_i))$, indexed by $f \in \mathcal{F}$ so that $\|S_n\|^* = \|\mathbb{G}_n\|_{\mathcal{F}}$. Plugging this into the proposition, we obtain

$$\| \|\mathbb{G}_n\|^*_{\mathcal{F}} \|_{\psi_2} \lesssim E^* \|\mathbb{G}_n\|_{\mathcal{F}} + n^{-1/2} \left(\sum_{i=1}^n \left\| \sup_{f \in \mathcal{F}} |f(Z_i) - \mathbb{E}[f(Z_i)]| \right\|^2_{\psi_2} \right)^{1/2} \tag{B.8}$$

$$\leq E^* \|\mathbb{G}_n\|_{\mathcal{F}} + n^{-1/2} \left(\sum_{i=1}^n \|F(Z_i) + \mathbb{E}[F(Z_i)]\|^2_{\psi_2} \right)^{1/2} \tag{B.9}$$

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

$$\begin{aligned} \|F(Z_i) + \mathbb{E}[F(Z_i)]\|_{\psi_2} &\leq \|F(Z_i)\|_{\psi_2} + \|\mathbb{E}[F(Z_i)]\|_{\psi_2} \\ &\leq \|F(Z_i)\|_{\psi_2} + \mathbb{E}[F(Z_i)]/\sqrt{\ln 2} \\ &\lesssim \|F(Z_i)\|_{\psi_2}. \end{aligned}$$

This gives a final bound by a constant multiple of

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

from which the desired claim holds. □

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

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, 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. □

B.2. **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

$$u(\beta^*) - u(\widehat{\beta}).$$

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 Section B.1.

Theorem B.2. *Fix $\delta > 0$. Under the assumptions above, with probability at least $1-\delta$*

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

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

Proof. The optimality of $\hat{\beta}$ for $u(\cdot)$ shows that $u(\beta^*) - u(\widehat{\beta}) \leq \sup_\beta |u(\beta) - \widehat{W}(\beta)|$. Combined with the uniform bias bound on $\widehat{W}$ from Proposition 1, we have

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

Let

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

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 $z = (K, \sigma, X)$. 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 B.3, 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 $\mathbb{P}(|Y| > t) \leq 2\exp\left(-x^2/\|Y\|_{\psi_2}^2\right)$. Equivalently, with probability $\geq 1 - \delta$,

$$|Y| < \|Y\|_{\psi_2} \sqrt{\log \frac{2}{\delta}}.$$

Combining this bound with Propositions B.1 and B.2 completes the proof. The analogous result follows for $W(\beta^*) - W(\widehat{\beta})$ by first noting that $W(\beta^*) - W(\widehat{\beta}) \leq 2\sup_\beta |\widehat{W}(\beta) - W(\beta)|$ and then repeating the same steps of the proof. □

B.3. **Proofs for Section 3.3.**

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

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

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

$$\frac{h}{\sqrt{n}}\left(\Phi\left(\frac{h}{\sqrt{n}} + M\right) - 1 + \max(1 - g_+, g_-)\right) \tag{B.10}$$

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

$$\sup_{\mu_{1:n}\in[-1,1]^n}\left\{\sup_{|\beta|<M} W(\beta; \mu_{1:n}) - \frac{1}{n}\sum_{i=1}^n \mu_i \mathbb{E}_{Y_{1:n}\sim\mathsf{N}(\mu_{1:n}, I_n)}[a_i(Y_1, \ldots, Y_n)]\right\} \geq \text{(B.10)}.$$

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

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

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

$$\text{(B.10)} \geq \frac{h}{\sqrt{n}}\left(\Phi\left(\frac{h}{\sqrt{n}} + M\right) - 1 + \frac{1}{2} - \frac{h}{2}\right) \geq \frac{h}{\sqrt{n}}\left(c_M - \frac{h}{2}\right) \gtrsim_M \frac{1}{\sqrt{n}}$$

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

B.4. **Proofs for Section 3.2.**

Proposition B.4 (VC-subgraphness in Example 4)**.** *The decision threshold classes in Equation (3.2), Equation (3.3), and Equation (3.4) are all VC-subgraph.*

Proof of Proposition B.4. Let us consider the first claim. Notice that the decision class (3.2) is 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 B.1.I establishes that $\mathcal{D}_{\text{linear-shrink}}$ is VC subgraph. For the second claim, notice that Equation (3.3) is contained in the function class

$$d_{b_0,\beta}(k, \sigma^2, x) := k + b_0 \sigma^2 k - x^\top \beta$$

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 a textbook reference on this topic. Consider the decision class in (3.4) 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. □

B.5. **Proofs for Section 3.4.**

Proof of Theorem 3. The proof proceeds as follows. Decompose the regret as

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

Theorem B.4 will show that the first term is $O((\log n)^5 n^{-1})$. To handle the second term $\mathbb{E}\left[W(\widehat{\beta}) - u(\widehat{\beta})\right]$, we use a leave-one-out stability argument. Define Y_{-i} to be the vector of observations without Y_i and

$$\widehat{\beta}^{(-i)} = \underset{\beta}{\operatorname{argmax}} \sum_{j \neq i} g_n(Y_j, Z_j, \beta), \tag{B.11}$$

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, $\widehat{\beta}^{(-i)}$ only differs from $\widehat{\beta}$ by $\tilde{O}_P(1/n)$. To explain why this will suffice, first write

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

Thus

$$\mathbb{E}\left[W(\widehat{\beta}) - u(\widehat{\beta})\right] = \frac{1}{n}\sum_{i=1}^{n}(\mu_i - K_i)\mathbb{E}\left[\Phi\left(\frac{\delta(\widehat{\beta}, Z_i) - \mu_i}{\sigma_i}\right) - \Phi\left(\frac{\delta(\widehat{\beta}^{(-i)}, Z_i) - \mu_i}{\sigma_i}\right)\right] \tag{B.12}$$

$$+ \frac{1}{n}\sum_{i=1}^{n}(\mu_i - K_i)\left(\mathbb{P}(Y_i \geq \delta(\widehat{\beta}^{(-i)}, Z_i)) - \mathbb{P}(Y_i \geq \delta(\widehat{\beta}, Z_i))\right). \tag{B.13}$$

Using Lipschitzness of Φ and $\delta(\cdot, Z_i)$, we will upper bound (B.12) by

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

for some constants L_i. To handle (B.13), we truncate. Fix a truncation threshold $E_n := C(\log n)^{\gamma}/n$ for C, γ constants to be chosen later.

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

Using Lipschitzness again, we find that (B.13) is bounded above by

$$\frac{1}{n}\sum_{i=1}^{n}\frac{L_i|\mu_i-K_i|}{\sigma_i}E_n+\frac{1}{n}\sum_{i=1}^{n}(\mu_i-K_i)\mathbb{P}(|\widehat{\beta}^{(-i)}-\widehat{\beta}|>E_n). \tag{B.14}$$

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

$$\mathbb{E}[|\widehat{\beta}^{(-i)}-\widehat{\beta}|]\le\mathbb{P}(|\widehat{\beta}^{(-i)}-\widehat{\beta}|>E_n)+ME_n$$

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

$$\frac{1}{n}\sum_{i=1}^{n}\left(1+\frac{L_i}{\sigma_i}\right)|\mu_i-K_i|\left(E_n+\mathbb{P}(|\widehat{\beta}^{(-i)}-\widehat{\beta}|>E_n)\right). \tag{B.15}$$

Theorem B.5 will show that

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

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

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

Combining this with the conclusion of Theorem B.4 gives the result.

□

B.5.1. *Auxiliary results for Theorem 3.*

Theorem B.3 (Uniform bounds on the Derivative Bias)**.** *We have*

$$|W'(\beta)-\mathbb{E}\widehat{W}_h'(\beta)|\lesssim\left(\frac{1}{n}\sum_{i=1}^{n}\frac{|\mu_i-K_i|}{\sigma_i}|\delta'(\beta,Z_i)|\right)he^{-1/2h^2}$$

$$|W''(\beta)-\mathbb{E}\widehat{W}''(\beta)|\lesssim$$

$$\left[h\left(\frac{1}{n}\sum_{i=1}^{n}\frac{|\mu_i-K_i|}{\sigma_i}|\delta''(\beta,Z_i)|\right)+\left(\frac{1}{n}\sum_{i=1}^{n}\frac{|\mu_i-K_i|}{\sigma_i^2}\delta'(\beta,Z_i)^2\right)\right]e^{-1/2h^2}$$

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 B.3. In the proof of Theorem 1, we showed

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

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

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

$$-\frac{\mu - K}{\sigma}\varphi\left(\frac{\mu - C}{\sigma}\right) = -\frac{1}{\pi}\frac{\mu - K}{\sigma}\int_0^\infty e^{-\frac{1}{2}\omega^2}\cos(\omega(\mu - C)/\sigma)d\omega$$

$$\frac{d}{dC}\mathbb{E}\Psi_h(Y, Z, C) = -\frac{1}{\pi}\frac{\mu - K}{\sigma}\int_0^{1/h} e^{-\frac{1}{2}\omega^2}\cos(\omega(\mu - C)/\sigma)d\omega.$$

By the chain rule, we have

$$\widehat{W}_h'(\beta) = \frac{1}{n}\sum_{i=1}^n \frac{d\Psi_h}{dC}(Y_i, Z_i, \delta(Z_i; \beta))\delta'(\beta, Z_i)$$

By the formulas for derivatives of Ψ in Lemma E.3 and Lemma E.2 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,

$$|W'(\beta) - \mathbb{E}\widehat{W}_h'(\beta)| \leq \frac{1}{\pi}\left(\frac{1}{n}\sum_{i=1}^n \frac{|\mu_i - K_i|}{\sigma_i}\delta'(\beta, Z_i)\right)\int_{1/h}^\infty e^{-\frac{1}{2}\omega^2}\cos(\omega(\mu - C)/\sigma)d\omega$$

$$\lesssim \left(\frac{1}{n}\sum_{i=1}^n \frac{|\mu_i - K_i|}{\sigma_i}\delta'(\beta, Z_i)\right)he^{-1/2h^2}$$

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

$$\widehat{W}_h''(\beta) = \frac{1}{n}\sum_{i=1}^n \Psi_{CC}(Y_i, Z_i, \delta(Z_i;\beta))\delta'(\beta, Z_i)^2 + \Psi_C(Y_i, Z_i, \delta(Z_i;\beta))\delta''(\beta, Z_i). \quad \text{(B.16)}$$

Thus,

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

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

Theorem B.4. *Under the assumptions of Theorem 3, for n large enough, we have*

$$\widehat{\beta} - \beta^* = \frac{2C}{\kappa}\frac{(\log n)^{5/2}}{\sqrt{n}}$$

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,

$$\mathbb{E}\left[W(\widehat{\beta}) - W(\beta^*)\right] \leq C\frac{(\log n)^5}{n}$$

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 B.4. Using (A1) and the fact that φ and its derivatives are uniformly bounded, $W''(\beta), W'''(\beta) \lesssim B^4$. By the results of Theorem B.2 we have with probability at least $1 - \delta$

$$W(\widehat{\beta}) - W(\beta^*) \leq \frac{m_1}{n\log n} + \frac{m_2 + (M + s_2)\log n}{\sqrt{n}}\sqrt{\log\frac{2}{\delta}}.$$

where M is the constant in the statement of Theorem 1. Substituting $\delta = 1/n$ and using the boundedness assumption (A1), we have

$$W(\widehat{\beta}) - W(\beta^*) \leq CB\frac{(\log n)^{3/2}}{\sqrt{n}} \quad \text{(B.17)}$$

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. (B.17) is less than ξ. Using (A4), we conclude that on A_n, $|\beta^* - \widehat{\beta}| \leq \kappa/B$. Now, Taylor expand W' around

β^* to obtain

$$W'(\beta^*) - W'(\widehat{\beta}) = -W''(\beta^*)(\beta^* - \widehat{\beta}) - \frac{W'''(\tilde{\beta})}{2}(\beta^* - \widehat{\beta})^2 \tag{B.18}$$

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

$$\begin{aligned}\left|W'(\beta^*) - W'(\widehat{\beta})\right| &\geq |W''(\beta^*)| \left|\beta^* - \widehat{\beta}\right| - \frac{|W'''(\tilde{\beta})|}{2}(\beta^* - \widehat{\beta})^2 \\ &\geq \kappa|\beta^* - \widehat{\beta}| - \frac{B^4}{2}(\beta^* - \widehat{\beta})^2,\end{aligned}$$

where the last line follows from (A3). Combining this with the Taylor expansion in Eq. (B.18), for n large enough, on the event A_n,

$$\frac{\kappa}{2}\left|\beta^* - \widehat{\beta}\right| \leq \left|W'(\beta^*) - W'(\widehat{\beta})\right|. \tag{B.19}$$

Next, by Theorem B.6, with probability at least $1 - 1/n$

$$\left|W'(\widehat{\beta}) - W'(\beta^*)\right| \leq \frac{C}{n\sqrt{\log n}} + \frac{C(\log n)^2}{\sqrt{n}}\sqrt{\log(2n)} \tag{B.20}$$

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 (B.19) and (B.20), we conclude that on $A_n \cap \tilde{A}_n$, which has probability at least $1 - \frac{2}{n}$, that

$$\left|\beta^* - \widehat{\beta}\right| \leq \frac{2C}{\kappa}\frac{(\log n)^{5/2}}{\sqrt{n}}$$

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

$$W(\beta^*) - W(\widehat{\beta}) \lesssim B^4(\widehat{\beta} - \beta^*)^2 + B^4(\widehat{\beta} - \beta^*)^3.$$

Then for n large enough,

$$\begin{aligned}\mathbb{E}\left[W(\beta^*) - W(\widehat{\beta})\right] &\leq \mathbb{E}\left[W(\beta^*) - W(\widehat{\beta}); A_n \cap \tilde{A}_n\right] + \mathbb{E}\left[W(\beta^*) - W(\widehat{\beta}); (A_n \cap \tilde{A}_n)^c\right] \\ &\lesssim B^4\mathbb{E}\left[(\beta^* - \widehat{\beta})^2; A_n \cap \tilde{A}_n\right] + B\mathbb{P}\left((A_n \cap \tilde{A}_n)^c\right) \\ &\lesssim \frac{C^2B^4}{\kappa^2}\frac{(\log n)^5}{n} + \frac{B}{n}.\end{aligned}$$

This concludes the proof. ∎

Theorem B.5. *In the setting and notation of Theorem 3, we have*

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

Proof of Theorem B.5. Recall the function Ψ in the ASSURE estimator of Eq. (2.6). Let Ψ_C, Ψ_{CC} denote the first and second partial derivatives in C, given in Lemma E.3. To control $\widehat{\beta} - \widehat{\beta}^{(-i)}$, we will apply a standard Taylor expansion using the first-order conditions

$$\sum_{i=1}^{n} \Psi_C(Y_i, Z_i, \delta(\widehat{\beta}, Z_i))\delta'(\widehat{\beta}, Z_i) = 0 \tag{B.21}$$

$$\sum_{j\neq i} \Psi_C(Y_j, Z_j, \delta(\widehat{\beta}^{(-i)}, Z_j))\delta'(\widehat{\beta}^{(-i)}, Z_j) = 0, \quad \forall i = 1, \ldots, n. \tag{B.22}$$

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

$$\begin{aligned}
0 &= \frac{1}{n}\sum_{i=1}^{n} \Psi_C(Y_i, Z_i, \delta(\widehat{\beta}, Z_i))\delta'(\widehat{\beta}, Z_i) \\
&= \frac{1}{n}\sum_{j=1}^{n} \Psi_C(Y_j, Z_j, \delta(\widehat{\beta}^{(-i)}, Z_j))\delta'(\widehat{\beta}^{(-i)}, Z_j) + \Delta_n(\widehat{\beta} - \widehat{\beta}^{(-i)}) \\
&= \frac{1}{n}\Psi_C(Y_i, Z_i, \delta(\widehat{\beta}^{(-i)}, Z_i))\delta'(\widehat{\beta}^{(-i)}, Z_i) + \Delta_n(\widehat{\beta} - \widehat{\beta}^{(-i)}).
\end{aligned}$$

where

$$\Delta_n := \left(\frac{1}{n}\sum_{j=1}^{n} \Psi_{CC}(Y_j, Z_j, \delta(\tilde{\beta}_i, Z_j))\delta'(\tilde{\beta}_i, Z_j)^2 + \Psi_C(Y_j, \tilde{\beta}_i)\delta''(\tilde{\beta}_i, Z_j)\right) \tag{B.23}$$

$$= \widehat{W}''(\tilde{\beta}_i). \tag{B.24}$$

for some $\tilde{\beta}_i$ between $\widehat{\beta}$ and $\widehat{\beta}^{(-i)}$. By assumption, $|\delta'(\widehat{\beta}^{(-i)}, Z_i)| \leq B$ and it is not difficult to show that

$$\Psi_C(Y_i, Z_i, \delta(\widehat{\beta}^{(-i)}, Z_i)) \lesssim \lambda_n^2 + \lambda_n \delta(\widehat{\beta}^{(-i)}, Z_i) \lesssim B\lambda_n^2.$$

using (A1). Thus,

$$\left|\widehat{\beta} - \widehat{\beta}^{(-i)}\right| \lesssim \frac{B\lambda_n^2}{n\Delta_n}. \tag{B.25}$$

Thus, if we can show that with probability greater than $1 - \frac{1}{n}, \Delta_n \geq \xi_1$ for some absolute constant $\xi_1 > 0$, the proof is finished.

To handle this term, observe that

$$\begin{aligned}\widehat{W}''(\tilde{\beta}_i) &= \widehat{W}''(\tilde{\beta}_i) - \mathbb{E}\left[\widehat{W}''(\tilde{\beta}_i)\right] \\ &\quad + \mathbb{E}\left[\widehat{W}''(\tilde{\beta}_i)\right] - W''(\tilde{\beta}_i) \\ &\quad + W''(\tilde{\beta}_i) - W''(\beta^*) \\ &\quad + W''(\beta^*).\end{aligned}$$

Since W''' is uniformly bounded by B^2, we have $W''(\tilde{\beta}_i) - W''(\beta^*) \leq B^2|\beta^* - \tilde{\beta}_i| \leq B^2 \max\left(|\beta^* - \beta^{(-i)}|, |\beta^* - \widehat{\beta}|\right)$. Thus by the reverse triangle inequality and (A3),

$$\begin{aligned}|\widehat{W}''(\tilde{\beta}_i)| \geq \kappa &- \sup_\beta \left|\widehat{W}''(\beta) - \mathbb{E}\left[\widehat{W}''(\beta)\right]\right| \\ &- \sup_\beta \left|\mathbb{E}\left[\widehat{W}''(\beta)\right] - W''(\beta)\right| \\ &- B^2 \max\left(|\beta^* - \beta^{(-i)}|, |\beta^* - \widehat{\beta}|\right).\end{aligned}$$

By (A1) and Theorem B.3,

$$\sup_\beta \left|\mathbb{E}\left[\widehat{W}''(\beta)\right] - W''(\beta)\right| \leq \left(\frac{B^3}{\sqrt{2\log n}} + B^5\right)\frac{1}{n} = O(n^{-1}). \tag{B.26}$$

Theorem B.4 shows that with probability at least $1-2/n$, $|\beta^* - \widehat{\beta}| \leq \frac{2C}{\kappa}(\log n)^{5/2} n^{-1/2}$. Applying this to both $\widehat{\beta}$ and $\widehat{\beta}^{(-i)}$, we find that

$$\max\left(|\beta^* - \beta^{(-i)}|, |\beta^* - \widehat{\beta}|\right) \leq \frac{4C}{\kappa}(\log n)^{5/2} n^{-1/2}, \tag{B.27}$$

say, with probability at least $1 - 4/(n-1)$. Finally, the result of Theorem B.7 states that with probability at least $1 - 1/n$,

$$\sup_\beta \left|\widehat{W}''(\beta) - \mathbb{E}\left[\widehat{W}''(\beta)\right]\right| \leq C\frac{(\log n)^4}{\sqrt{n}} \tag{B.28}$$

Now for n large enough such that the sum of the right hand sides of Equations (B.26), (B.27), (B.28) is less than $\kappa/2$, we conclude that $\Delta_n \geq \kappa/2$ with probability at least $1 - O(n^{-1})$. Plugging this into Eq. (B.25) we find

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

for some absolute constant M. □

The next lemma records the derivatives of the average welfare. Since the computation is straightforward, we omit the proof.

Lemma B.4 (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*

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

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

and

$$\begin{aligned} W'''(\beta) &:= -\frac{1}{n}\sum_{i=1}^{n}(\mu_i - K_i)A_i \\ A_i &:= \varphi\left(T_i(\beta)\right)\delta'''(\beta, Z_i) - 3\varphi'\left(T_i(\beta)\right)\delta''(\beta, Z_i)\delta'(\beta, Z_i) \\ &\quad + \varphi''\left(T_i(\beta)\right)\delta'(\beta, Z_i)^3 \end{aligned}$$

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

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

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

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

Proof of Theorem B.6. By (A2), $W'(\beta^*) = 0$ and by definition we have $\widehat{W}'(\widehat{\beta}) = 0$. Therefore,

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

which can in turn be bounded above by

$$\sup_{\beta}|W'(\beta) - \mathbb{E}\widehat{W}'(\beta)| + \sup_{\beta}|\widehat{W}'(\beta) - \mathbb{E}\widehat{W}'(\beta)|. \tag{B.29}$$

The first term represents the bias of the derivative of the ASSURE estimator, which we bounded in Theorem B.3 by

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

We will focus on bounding the second term $\sup_\beta |\widehat{W}'(\beta) - \mathbb{E}\widehat{W}'(\beta)|$ using INID empirical process theory. The proof structure follows exactly the proof of Theorem 1. Recalling the notation in Eq. (E.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

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

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

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

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

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

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 $\Psi_C(y, z, \delta(z;\beta))\delta'(\beta, z)$.

Using a uniform bound for sinc and sinc$'$ alongside assumption (A1), envelope functions for $\mathcal{F}^{(1)}, \mathcal{F}^{(2)}$ are given by

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

for some absolute constant L. Propositions B.5 and B.6 bound these terms, yielding

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

Combining these propositions and the bias bound, we obtain

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

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

To get the second claim, we will apply Proposition B.3 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

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

We conclude by the definition of the Orlicz norm. □

Theorem B.7. *Under the same assumptions of Theorem 3,*

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

for an absolute constant C depending 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

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

Proof of Theorem B.7. Recall that

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

Let

$$\begin{aligned}A(\beta)&:=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi_{CC}(Y_i,Z_i,\delta(Z_i;\beta))\delta'(\beta,Z_j)^2\\ B(\beta)&:=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi_{C}(Y_i,Z_i,\delta(Z_i;\beta))\delta''(\beta,Z_j).\end{aligned}$$

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 B.6, replacing δ' with δ'', establishes the bound

$$\mathbb{E}\left[\sup_{\beta}|B(\beta)-\mathbb{E}B(\beta)|\right]\leq C(\log n)^2 \tag{B.30}$$

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 E.3, Ψ_{CC} is given by

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

We emulate the proofs of Theorems 1 and B.6. Throughout this proof, define the following functions classes indexed by β:

$$\begin{aligned}
\mathcal{F}^{(0)} &:= \left\{f_\beta^{(0)}(y,z)=\frac{\lambda_n^2}{\pi}\frac{(Y_i-K_i)}{\sigma_i^2}\operatorname{sinc}'\left(\lambda_n\left(\frac{Y_i-\delta(z;\beta)}{\sigma_i}\right)\right)\right\}\\
\mathcal{F}^{(1)} &:= \left\{f_\beta^{(1)}(y,z)=\frac{\lambda_n^2}{\pi}\frac{(Y_i-K_i)}{\sigma_i^2}\operatorname{sinc}'\left(\lambda_n\left(\frac{Y_i-\delta(z;\beta)}{\sigma_i}\right)\right)\delta'(\beta,z)^2\right\}\\
\mathcal{F}^{(2)} &:= \left\{f_\beta^{(2)}(y,z)=\frac{\lambda_n^3}{\sigma_i}\operatorname{sinc}''\left(\lambda_n\left(\frac{Y_i-\delta(z;\beta)}{\sigma_i}\right)\right)\delta'(\beta,z)^2\right\}\\
\mathcal{D}'^2 &:= \left\{\delta'(\beta,z)^2\right\}.
\end{aligned}$$

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 (A1), envelope functions for $\mathcal{F}^{(1)},\mathcal{F}^{(2)}$ are given by

$$F^{(1)}:=CB^3\left(\lambda_n^2\frac{|y-k|}{\sigma}\right) \tag{B.31}$$

$$F^{(2)}:=CB^3\lambda_n^3. \tag{B.32}$$

for some absolute constant C. Furthermore notice that $\mathcal{D}'^2$ is a VC subgraph class, by composing Lemma B.1.V and Lemma 2.6.20.viii of van der Vaart and Wellner

(2023). To proceed, write

$$\sup_{\beta} |A(\beta) - \mathbb{E}A(\beta)| \leq \|\mathbb{G}\|_{\mathcal{F}^{(1)}} + \|\mathbb{G}\|_{\mathcal{F}^{(2)}}.$$

Propositions B.7 and B.8 show that

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

for constants C that depend on B and the VC indexes of $\mathcal{D}, \mathcal{D}'^2$. Combining this with Eq. (B.30) we obtain the first claim:

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

To obtain the second claim, apply Proposition B.3 with the envelope function $F^{(1)} + F^{(2)} = CB^3\left(\lambda_n^2 \frac{|y-k|}{\sigma}\lambda_n^3\right)$. 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 Prop. B.3 plus the definition of the Orlicz norm yields

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

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

B.5.2. *Proofs for Uniform Bounds for* ASSURE *Derivatives.* The structure of the following results are all analogous to Theorem 1.

Proposition B.5. *Under the assumptions and notation of Theorem B.6, we have*

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

Proof. Following the proof of Theorem B.1 as before, we may apply symmetrization and a maximal inequality to obtain

$$\mathbb{E}\left[\|\mathbb{G}\|_{\mathcal{F}^{(1)}}\right] \leq \mathbb{E}\left[\int_0^{\eta_n} \sqrt{1 + \log N(\epsilon, \mathcal{F}^{(1)}, \mathbb{L}_2(\mathbb{P}_n))}d\epsilon\right], \tag{B.33}$$

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

$$\mathcal{F}^{(0)} := \left\{ f_\beta^{(0)}(y,z) := \frac{\lambda_n}{\pi} \left(\frac{y-k}{\sigma} \right) \operatorname{sinc}\left(\lambda_n \left(y - \delta(z;\beta) \right) / \sigma \right) : \beta \in \mathbb{R} \right\} \tag{B.34}$$

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:

$$N(C\left(\lambda_n \widehat{\nu}_4 + B\right)\epsilon, \mathcal{F}^{(1)}, \mathbb{L}_2(\mathbb{P}_n)) \leq N(\epsilon, \mathcal{D}', \mathbb{L}_4(\mathbb{P}_n)) \cdot N(\epsilon, \mathcal{F}^{(0)}, \mathbb{L}_2(\mathbb{P}_n)). \tag{B.35}$$

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

$$N(C\epsilon\lambda_n^2 \widehat{\zeta}_4, \mathcal{F}^{(0)}, \mathbb{L}_2(\mathbb{P}_n)) \leq N(\epsilon, \mathcal{D}, \mathbb{L}_4(\mathbb{P}_n)). \tag{B.36}$$

Putting these bounds into Eq. (B.41), we have

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

Following the change of variables argument used to obtain Equation (B.5) and the uniform envelope function B for $\mathcal{D}, \mathcal{D}'$ we then have

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

Using Cauchy Schwarz repeatedly and (A1), 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 B.6. *Under the assumptions and notation of Theorem B.6, we have*

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

for an absolute constant depending on bound B.

Proof. Following the proof of Theorem B.1 as before, we may apply symmetrization and a maximal inequality to obtain

$$\mathbb{E}\left[\|\mathbb{G}\|_{\mathcal{F}^{(2)}}\right] \leq \mathbb{E}\left[\int_0^{\eta_n} \sqrt{1 + \log N(\epsilon, \mathcal{F}^{(2)}, \mathbb{L}_2(\mathbb{P}_n))} d\epsilon\right], \tag{B.37}$$

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

$$\mathcal{S} := \left\{s_\beta(y,z) := \lambda_n^2 \text{sinc}'\left(\lambda_n\left(y - \delta(z;\beta)\right)/\sigma\right) : \beta \in \mathbb{R}\right\} \tag{B.38}$$

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

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

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

$$N(C\left(\lambda_n + B\right)\epsilon, \mathcal{F}^{(2)}, \mathbb{L}_2(\mathbb{P}_n)) \leq N(\epsilon, \mathcal{D}', \mathbb{L}_2(\mathbb{P}_n)) \cdot N(\epsilon, \mathcal{S}, \mathbb{L}_2(\mathbb{P}_n)) \tag{B.39}$$

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

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

Putting these bounds into Eq. (B.44), we have

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

Following the argument of Prop. B.2 again we then have

$$\mathbb{E}\|\mathbb{G}\|_{\mathcal{F}^{(2)}} \lesssim \mathbb{E}\left\|F^{(2)}\right\|_n + \lambda_n^3(\lambda_n + B)\sqrt{V(\mathcal{D})} + (\lambda_n + B)\sqrt{V(\mathcal{D}')}. \tag{B.40}$$

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

□

Proposition B.7. *Under the setting and assumptions of Theorem B.7, we have*

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

for a constant C that depends on B.

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 B.1 as before, we may apply symmetrization and a maximal inequality to obtain

$$\mathbb{E}\left[\|\mathbb{G}\|_{\mathcal{F}^{(1)}}\right] \leq \mathbb{E}\left[\int_0^{\eta_n} \sqrt{1 + \log N(\epsilon, \mathcal{F}^{(1)}, \mathbb{L}_2(\mathbb{P}_n))} d\epsilon\right], \tag{B.41}$$

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 $\mathcal{F}^{(0)}$ 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:

$$N(CB\left(\lambda_n^2\widehat{\nu}_4 + B\right)\epsilon, \mathcal{F}^{(1)}, \mathbb{L}_2(\mathbb{P}_n)) \leq N(\epsilon, \mathcal{D}'^2, \mathbb{L}_4(\mathbb{P}_n)) \cdot N(\epsilon, \mathcal{F}^{(0)}, \mathbb{L}_2(\mathbb{P}_n)). \quad \text{(B.42)}$$

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

$$N(CB^2\lambda_n^3\widehat{\nu}_4\epsilon, \mathcal{F}^{(0)}, \mathbb{L}_2(\mathbb{P}_n)) \leq N(\epsilon, \mathcal{D}, \mathbb{L}_4(\mathbb{P}_n)). \quad \text{(B.43)}$$

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,

$$\begin{aligned}
&\left\|f_\beta^{(0)}(y, z) - \frac{\lambda_n^2}{\pi}\frac{(y-K)}{\sigma_i^2}\text{sinc}'(\lambda_n(y - \tilde{d})/\sigma)\right\|_{\mathbb{L}^2(\mathbb{P}_n)} \\
&\leq C\lambda_n^2 \left(\frac{1}{n}\sum_{i=1}^n \frac{(Y_i - K_i)^2}{\sigma_i^4}\left(\delta(Z_i;\beta) - \tilde{d}\right)^2\right)^{1/2} \\
&\leq CB^2\lambda_n^2 \left(\frac{1}{n}\sum_{i=1}^n \frac{(Y_i - K_i)^4}{\sigma_i^4}\right)^{1/4} \left(\frac{1}{n}\sum_{i=1}^n \left(\delta(Z_i;\beta) - \tilde{d}\right)^4\right)^{1/4} \qquad \text{(Cauchy–Schwarz, (A1))} \\
&\leq CB^2\lambda_n^2\widehat{\nu}_4\epsilon.
\end{aligned}$$

Putting these bounds into Eq. (B.41), we have

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

Following the change of variables argument leading up to Equation (B.5) and using the uniform envelope functions B, B^2 for $\mathcal{D}, \mathcal{D}'^2$ we then have

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

where $F^{(1)}$ is the envelope function defined in (B.31). Using Cauchy Schwarz repeatedly and (A1), it is not difficult to observe the bound

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

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

Proposition B.8. *Under the setting and assumptions of Theorem B.7, 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 an absolute constant C depending on the bound B.

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,

$$\mathbb{E}\left[\|\mathbb{G}\|_{\mathcal{F}^{(2)}}\right] \leq \mathbb{E}\left[\int_0^{\eta_n} \sqrt{1+\log N(\epsilon, \mathcal{F}^{(2)}, \mathbb{L}_2(\mathbb{P}_n))} d\epsilon\right], \tag{B.44}$$

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

$$\mathcal{S} := \left\{s_\beta(y,z) := \frac{\lambda_n^3}{\sigma}\text{sinc}''\left(\lambda_n\left(y-\delta(z;\beta)\right)/\sigma\right) : \beta \in \mathbb{R}\right\} \tag{B.45}$$

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 (A1), 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,

$$N(CB\left(\lambda_n^3 + B^2\right)\epsilon, \mathcal{F}^{(2)}, \mathbb{L}_2(\mathbb{P}_n)) \leq N(\epsilon, \mathcal{D}'^2, \mathbb{L}_2(\mathbb{P}_n)) \cdot N(\epsilon, \mathcal{S}, \mathbb{L}_2(\mathbb{P}_n)) \tag{B.46}$$

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. (B.44), 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 Prop. B.2 again we then have

$$\mathbb{E}\|\mathbb{G}\|_{\mathcal{F}^{(2)}} \lesssim \mathbb{E}\left\|F^{(2)}\right\|_n + \lambda_n^4(\lambda_n^3 + B^2)\sqrt{V(\mathcal{D})} + (\lambda_n^3 + B^2)\sqrt{V(\mathcal{D}'^2)}. \tag{B.47}$$

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

□

B.5.3. *Example decision rule class for Theorem 3.*

Example 5. 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 the assumptions (A1)-(A4). The following lower-level assumptions give such a class of $\mu_{1:n}$. Fix $\kappa_0 > 0, B > 0$ and let $\overline{\mu} := \frac{1}{n}\sum_{i=1}^n \mu_i, \overline{\sigma}^2 := \frac{1}{n}\sum_{i=1}^n(\mu_i - \overline{\mu})^2$.

(L1) $-B < \min_i \mu_i < 0$ and $0 < \max_i \mu_i < B$.

(L2) Letting $m_B := \inf_{[B,3B]} t\varphi(t)$,

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

(L3) $\frac{1}{n}\sum_{i=1}^n \mu_i^2 > \kappa_0/\varphi(3B)$.

Proposition B.9. *Fix constant B, κ_0 as above. Assume (L1)-(L3). Then (A1)-(A4) hold with $\kappa = \kappa_0$ and with some $\xi > 0$ depending only on B, κ_0.*

Proof. By (L1), it is easy to see that (A1) holds. Now without loss, assume that $\mu_1 \le \mu_2 \le \cdots \le \mu_n$. In this setting, $W(\beta) = \frac{1}{n}\sum_{i=1}^n \mu_i \Phi(\mu_i - \beta), W'(\beta) = -\frac{1}{n}\sum_{i=1}^n \mu_i \varphi(\mu_i - \beta)$, and $W''(\beta) = \frac{1}{n}\sum_{i=1}^n \mu_i(\beta - \mu_i)\varphi(\mu_i - \beta)$. A result of Schönberg (1948) shows that φ is a totally positive kernel and, for any sequence of $c_1 \le c_2 \le ... \le c_n$, the number of sign changes in the vector $(W'(c_1), \ldots, W'(c_n))$ is at most that in the vector $(\mu_1, \ldots, \mu_n)$. Note that since $(\mu_1, ..., \mu_n)$ changes sign only once, $(W'(c_1), \ldots, W'(c_n))$ changes sign at most once. This implies that there cannot be multiple local maximizers.

Next, we claim that (L2) implies that $W'(-2B) > 0$ and $W'(2B) < 0$. If so, this would show (A2) 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

$$-W'(2B) = \frac{1}{n}\sum_{i=1}^n \mu_i f_+(\mu_i) = \overline{\mu}\overline{f}_+ + \frac{1}{n}\sum_{i=1}^n (\mu_i - \overline{\mu})\left(f_+(\mu_i) - f_+(\overline{\mu})\right) \qquad \text{(B.48)}$$

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 $f_+'(\xi_i) \ge m_B = \inf_{t\in[B,3B]} t\varphi(t) > 0$. Moreover, $\overline{f}_+ \le \varphi(B)$. Thus,

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

We can repeat the argument in the other direction with $f_- := \varphi(x + 2B)$. Observe the analogous decomposition

$$\frac{1}{n}\sum_{i=1}^n \mu_i f_-(\mu_i) = \overline{\mu}\overline{f}_- + \frac{1}{n}\sum_{i=1}^n (\mu_i - \overline{\mu})\left(f_-(\mu_i) - f_-(\overline{\mu})\right) \qquad \text{(B.49)}$$

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 $f'_-(x) = -(x+2B)\varphi(x+2B) \le -m_B$. Moreover, $\overline{f}_- \le \varphi(B)$. Thus,

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

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 (L3) implies

$$W''(\beta^*) < -\frac{\kappa_0}{\varphi(3B)}\varphi(3B) = -\kappa_0.$$

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

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

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

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

Thus

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

Now by the intermediate value theorem, for some R small enough, less than κ_0/B^2, $W''(\beta) < -\kappa_0/2$ for all $|\beta - \beta^*| \le R$. 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 (A4).

□

B.6. **Proof of Theorem 4.**

Proof of Theorem 4. Throughout the proof let $a_{1:n}$ denote a vector of decision rules, where the ith index is $a_i(Y_1, \ldots, Y_n) \in \{0, 1\}$. We will first exhibit a class of priors that are supported on Θ. Consider the priors $\mathsf{Rad}_p^{\otimes n}$ for which $\mu_i = +1$ with probability p and -1 with probability $1-p$ and μ_i are i.i.d. Let $\mathsf{Rad}_p^{\otimes n}|A$ 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 $\mathsf{Rad}_p^{\otimes n}|A$ satisfies (L1) with $B = 1$ and, choosing κ small enough, it satisfies (L3) since $\frac{1}{n}\sum_{i=1}^n \mu_i^2 = 1$. Lastly, (L2) 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:

$$\sup_{\mu_{1:n}\in\Theta} \mathsf{Regret}_n(a_{1:n}) \geq \mathbb{E}_{\mathsf{Rad}_p^{\otimes n}|A}\left[\mathsf{Regret}_n(a_{1:n})\right] \tag{B.50}$$

$$= \frac{\mathbb{E}_{\mathsf{Rad}_p^{\otimes n}}\left[\mathsf{Regret}_n(a_{1:n})\mathbf{1}(A)\right]}{\mathbb{P}_{\mathsf{Rad}_p^{\otimes n}}(A)} \tag{B.51}$$

$$= \frac{\mathbb{E}_{\mathsf{Rad}_p^{\otimes n}}\left[\mathsf{Regret}_n(a_{1:n})\right] - \mathbb{E}_{\mathsf{Rad}_p^{\otimes n}}\left[\mathsf{Regret}_n(a_{1:n})\mathbf{1}(A^c)\right]}{\mathbb{P}_{\mathsf{Rad}_p^{\otimes n}}(A)} \tag{B.52}$$

$$\geq \frac{\mathbb{E}_{\mathsf{Rad}_p^{\otimes n}}\left[\mathsf{Regret}_n(a_{1:n})\right] - \mathbb{P}_{\mathsf{Rad}_p^{\otimes n}}(A^c)}{1 - \mathbb{P}_{\mathsf{Rad}_p^{\otimes n}}(A^c)}. \tag{B.53}$$

It is easy to see that $\mathbb{P}_{\mathsf{Rad}_p^{\otimes n}}(A^c)$ is exponentially small in n for all p in a small neighborhood of $\frac{1}{2}$. We will focus on analyzing the term $\mathbb{E}_{\mathsf{Rad}_p^{\otimes n}}\left[\mathsf{Regret}_n(a_{1:n})\right]$. We have

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

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

$$\sup_\beta \mathbb{E}_G W(\beta) = \mathbb{E}_G \mathbb{E}_X[\mu \mathbf{1}\left\{\mathbb{E}[\mu \mid Y] > 0\right\}],$$

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 prior $\mathsf{Rad}_p^{\otimes n}$ up to some additive and multiplicative factors which are exponentially small in n.

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

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

We claim that the Bayes regret can also be written as a weighted classification loss

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_G\left[\left|m_G(Y_i)\right| \mathbf{1}\left\{a^*_G(Y_i) \neq a_i(Y_{1:n})\right\}\right]. \tag{B.54}$$

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

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

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

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

Next, consider two priors $G_0 := \mathsf{Rad}^{\otimes n}_{1/2 - h/\sqrt{n}}$ and $G_1 := \mathsf{Rad}^{\otimes n}_{1/2 + h/\sqrt{n}}$ for h to be chosen later. Let $c^*_{G_i}$ be the zero of the posterior mean $m_{G_i}(y)$ for $i = 0, 1$. By monotonicity of the decision rule $a_i(Y_{1:n})$, observe that

$$\mathbf{1}\left\{a^*_{G_0}(Y_i) \neq a_i(Y_{1:n})\right\} + \mathbf{1}\left\{a^*_{G_1}(Y_i) \neq a_i(Y_{1:n})\right\} \geq \mathbf{1}\left\{c^*_{G_0} < Y_i < c^*_{G_1}\right\}. \tag{B.55}$$

Next, define $w(y) = \min\left(\left|m_{G_0}(y)\right|, \left|m_{G_1}(y)\right|\right)$ and $q(\mathbf{y}) = \min(p^{\otimes n}_{G_0}(\mathbf{y}), p^{\otimes n}_{G_1}(\mathbf{y}))$ where $p^{\otimes n}_{G_i}(\mathbf{y})$ is the joint density of the data $Y_1, \ldots, Y_n$ under the prior G_i. Let $p_{G_i}(y)$ be the marginal density of a single y under prior G_i. Finally, let $p^{(-j)}_{G_i}(\mathbf{y})$ be the joint density of the data $Y_1, \dots, Y_{j-1}, Y_{j+1}, \dots, Y_n$, under G_i. Then,

$$
\begin{aligned}
R(G_0, \delta) + R(G_1, \delta) &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{G_0} \left[|m_{G_0}(Y_i)| \, \mathbf{1} \left\{ a^*_{G_0}(Y_i) \neq a_i(Y_{1:n}) \right\} \right] \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{G_1} \left[|m_{G_1}(Y_i)| \, \mathbf{1} \left\{ a^*_{G_1}(Y_i) \neq a_i(Y_{1:n}) \right\} \right] \\
&\geq \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}^n} w(y_i) \mathbf{1} \left\{ c^*_{G_0} < y_i < c^*_{G_1} \right\} q(\mathbf{y}) dy_1 \dots dy_n.
\end{aligned}
$$

Inspecting the integral, first notice that

$$
q(\mathbf{y}) \geq \min \left(p_{G_0}(y_i), p_{G_1}(y_i) \right) \min \left(\prod_{j \neq i} p_{G_0}(y_j), \prod_{j \neq i} p_{G_1}(y_j) \right).
$$

Thus,

$$
R(G_0, \delta) + R(G_1, \delta) \geq \alpha_{n-1} \frac{1}{n} \sum_{i=1}^{n} \int_{c^*_{G_0}}^{c^*_{G_1}} w(y_i) \min \left(p_{G_0}(y_i), p_{G_1}(y_i) \right) dy_i \tag{B.56}
$$

where

$$
\alpha_{n-1} := \int_{\mathbb{R}^{n-1}} \min \left(p^{(-i)}_{G_0}(\mathbf{y}_{-i}), p^{(-i)}_{G_1}(\mathbf{y}_{-i}) \right) d\mathbf{y}_{-i}.
$$

Notice that $\alpha_{n-1} = 1 - \mathsf{TV}(p^{(-i)}_{G_0}, p^{(-i)}_{G_1}) \geq 1 - \sqrt{\frac{n-1}{2} \mathsf{KLD}(p_{G_0}, p_{G_1})}$ by Pinsker's and tensorization of KL divergence. By the data processing inequality for divergences,

$$
\mathsf{KLD}(p_{G_0}, p_{G_1}) \leq \mathsf{KLD}(G_0, G_1) = \frac{2h}{\sqrt{n}} \ln \frac{\frac{1}{2} + \frac{h}{\sqrt{n}}}{\frac{1}{2} - \frac{h}{\sqrt{n}}}.
$$

The last equality follows from the formula for the Bernoulli KL divergence. Then for large enough n, $\mathsf{KLD}(p_{G_0}, p_{G_1}) \leq 16h^2/n$ and so

$$
\alpha_{n-1} \geq 1 - h\sqrt{\frac{8(n-1)}{n}}.
$$

Taking $h = 1/4$ we have $\alpha_{n-1} \geq \alpha := 1 - \sqrt{2}/2$. Next, observe that the posterior distribution $m_{\mathsf{Rad}_p}(y)$ is given by the explicit formula

$$
\frac{pe^{2y} - (1-p)}{pe^{2y} + (1-p)}
$$

with $c^*_{\mathsf{Rad}_p} = \frac{1}{2} \ln \frac{1-p}{p}$. Thus,

$$
c^*_{G_0} = -c^*_{G_1} = \frac{1}{2} \ln \left(1 + \frac{2\frac{h}{\sqrt{n}}}{\frac{1}{2} - \frac{h}{\sqrt{n}}} \right).
$$

Using the inequality $\ln(1+x) \geq x/2$ for small x, for n large enough we find that

$$c^*_{G_0} \geq \frac{1}{2} \ln\left(1 + \frac{8h}{\sqrt{n}}\right) \geq \frac{2}{\sqrt{n}}.$$

Furthermore, it is not difficult to see that on $[c^*_{G_1}, c^*_{G_0}]$, for n large enough, we have

$$w(y) \geq \frac{1}{2}|c^*_{G_0} - y|.$$

This follows by noting $m_{\mathsf{Rad}_p}(y) \geq \frac{1}{2}(y - c^*_{\mathsf{Rad}_p})$ on $[c^*_{\mathsf{Rad}_p}, c^*_{\mathsf{Rad}_p} + \epsilon]$ and $m_{\mathsf{Rad}_p}(y) \leq \frac{1}{2}(y - c^*_{\mathsf{Rad}_p})$ on $[c^*_{\mathsf{Rad}_p} - \epsilon, c^*_{\mathsf{Rad}_p}]$, for some small ϵ. Lastly, $\min(p_{G_0}(y_i), p_{G_1}(y_i))$ is bounded above by an absolute constant c on the interval $[c^*_{G_1}, c^*_{G_0}]$. Combining these items together, we have

$$\begin{aligned} R(G_0, \delta) + R(G_1, \delta) &\geq \frac{\alpha c}{n} \sum_{i=1}^{n} \int_{-c^*_{G_0}}^{c^*_{G_0}} w(y_i) dy_i \\ &\geq \frac{\alpha c}{2} \int_{-c^*_{G_0}}^{c^*_{G_0}} |c^*_{G_0} - y| dy \\ &\geq \frac{\alpha c}{2} \frac{c^{*2}_{G_0}}{4} = \Omega(n^{-1}). \end{aligned}$$

Recall from Eq. (B.53) that

$$\sup_{\mu_{1:n} \in \Theta} \mathsf{Regret}_n(a_{1:n}) \geq \max\left(\frac{R(G_0, \delta) - \mathbb{P}_{G_0}(A^c)}{1 - \mathbb{P}_{G_0}(A^c)}, \frac{R(G_1, \delta) - \mathbb{P}_{G_1}(A^c)}{1 - \mathbb{P}_{G_1}(A^c)}\right).$$

For n large enough, $\mathbb{P}_{G_0}(A^c), \mathbb{P}_{G_1}(A^c) = O(e^{-n})$. Thus,

$$\begin{aligned} \sup_{\mu_{1:n} \in \Theta} \mathsf{Regret}_n(a_{1:n}) &\geq \max(R(G_0, \delta), R(G_1, \delta)) - O(e^{-n}) \\ &\geq \frac{1}{2}(R(G_0, \delta) + R(G_1, \delta)) - O(e^{-n}) \\ &= \Omega(n^{-1}), \end{aligned}$$

as desired. $\square$

Appendix C. Proofs for Section 4

Proof of Proposition 2. Let $Y \sim \mathsf{Poi}(\mu)$. Then for any integer $c \geq 0$, we claim that $\mathbb{E}[Y\mathbf{1}\{Y \geq c+1\}] = \mu\mathbb{P}(Y \geq c)$. Indeed,

$$\begin{aligned}\mathbb{E}[Y\mathbf{1}\{Y \geq c+1\}] &= \sum_{k=c+1}^{\infty} k\frac{e^{-\mu}\mu^k}{k!} \\ &= \mu\sum_{k=c+1}^{\infty}\frac{e^{-\mu}\mu^{k-1}}{(k-1)!} \\ &= \mu\sum_{k=c}^{\infty}\frac{e^{-\mu}\mu^k}{k!} \\ &= \mu\mathbb{P}(Y \geq c).\end{aligned}$$

Taking $c = \delta(Z_i;\beta)$ and applying the above computation to each summand of Equation (4.1) □

Proof of Theorem 5. 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 Thm. 1 shows that the regret is bounded as

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

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

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

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 B.1, we may apply symmetrization and a maximal inequality to obtain

$$\mathbb{E}\sup_\beta\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n f_\beta(Y_i,\mu_i,Z_i)\right| \leq \mathbb{E}\left[\int_0^{\eta_n}\sqrt{1+\log N(\epsilon,\mathcal{F},\mathbb{L}_2(\mathbb{P}_n))}d\epsilon\right], \tag{C.2}$$

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 B.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

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

Next, by Jensen's inequality observe the bounds

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

$$\mathbb{E}\|F_2\|_n \le \left(\frac{1}{n}\sum_{i=1}^n \mu_i^2\right)^{1/2}$$

and $\mathbb{E}\|F\|_n \le \mathbb{E}\|F_1\|_n + \mathbb{E}\|F_2\|_n$. 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. □

Theorem 6 (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 (3.7) except with μ_i replaced with $\mu_i - K$.

Proof of Theorem 6. 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 ship all indices i (take $C_{\text{opt}} = 0$). In the second case, the optimal decision is to ship no indices (take $C_{\text{opt}} = \infty$). Then:

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

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

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

Now, applying Pinsker's inequality and tensorization yields

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

But

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

Thus

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

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 $O(n^{-1/2})$ as desired.

□

Appendix D. Additional Details for Section 5 and Section 6

D.1. Simulation Details.

D.1.1. *Calibrated OA Data.* In this simulation, $n \approx 10^4$ and $\sigma_{1:n}$ are taken from the Opportunity Atlas dataset. We generate $Y_i = \mu_i + \mathsf{N}(0, \sigma_i^2)$ for each i. In addition, we add constant costs $K_i = 0.361$ (motivated by the empirical application) and a one-dimensional covariate X_i generated by taking μ_i and adding scaled t-distributed noise with 10 degrees of freedom ($\rho \approx 0.7$). We compare 11 methods across 40 Monte Carlo runs, where $Y_{1:n}$ are re-randomized for each run.

Firstly, we consider the class of linear shrinkage rules from (3.2). For this class, we consider 3 competing methods: 1) the plug-in approach where μ_0, τ^2 are the grand mean and standard deviation estimated from the data, 2) ASSURE, and 3) coupled bootstrap. Recall that method 1) corresponds to computing the empirical Bayes posterior mean assuming a Gaussian prior, then implementing if the posterior mean is greater than the implementation cost.

Secondly, we consider the class of decisions from (3.3) based on the Fay-Herriot model. For this class, we consider 3 competing methods: 4) the empirical Bayes plug-in approach from (Fay III and Herriot, 1979) 5) ASSURE, and 6) coupled bootstrap. These models take advantage of a one-dimensional covariate X_i.

Thirdly, we consider the CLOSE-GAUSS model discussed in Example 4. We consider decision rules where the threshold is given by

$$\mathcal{D}_{\text{CLOSE-GAUSS}} = \left\{ K_i + \frac{\sigma_i^2}{s_0^2(Z_i)}(K_i - m_0(Z_i)) \right\},$$

where $s_0^2(Z_i), m_0(Z_i)$ depend only on σ_i and are obtained via nonparametric regression. We also consider the decision in the parametric model Eq. (3.4) chosen by ASSURE. This yields methods 7), 8) respectively.

Fourth, we consider 9) Gaussian Heteroskedastic NPMLE using the `REBayes` package (Koenker and Gu, 2017) and 10) CLOSE-NPMLE as implemented in Chen (2025), where a unit is selected if its posterior mean is greater than implementation cost K. Finally, we consider decision rules obtained by ensembling a predictive model and empirical Bayes estimate, as described in Section 4. Concretely, we consider rules such that

$$\alpha m_{\widehat{G}}(Y_i, \sigma_i) + (1-\alpha)\widehat{f}(X_i) \geq K_i,$$

where $m_{\widehat{G}}$ is an empirical Bayes posterior mean function estimated from the data, and $\widehat{f}$ is a model estimated from the data $(X_{1:n}, Y_{1:n})$. α is a parameter in $[0,1]$ tuned using ASSURE. Taking the CLOSE-GAUSS posterior mean function

$$m_{\widehat{G}}(Y_i, \sigma_i) = \frac{\sigma_i^2}{s_0^2(\sigma_i) + \sigma_i^2} m_0(\sigma_i) + \frac{s_0^2(\sigma_i)}{s_0^2(\sigma_i) + \sigma_i^2} Y_i$$

with m_0, s_0^2 estimated nonparametrically and the linear model $X_i^\top \beta$, the decision thresholds can be stated as

$$\mathcal{D}_{\text{ENSEMBLE}} = \left\{ \left(\frac{K_i - (1-\alpha)X_i^\top \beta}{\alpha} \right) \left[1 + \frac{\sigma_i^2}{s_0^2(\sigma_i)} \right] - m_0(\sigma_i) \frac{\sigma_i^2}{s_0^2(\sigma_i)} : \alpha \in \mathbb{R} \right\}. \quad \text{(D.1)}$$

We fit β with GLS and m_0, s_0^2 using CLOSE-GAUSS as in (Chen, 2025). This is method 11 of the comparison.

We also compare the same methods in an adverse setting where both the outcome distribution is misspecified and the covariate distribution is pure noise. We take $Y_i = \mu_i + \sigma_i T_i$ with $T_i \overset{i.i.d}{\sim} t_{10}$, the t-distribution with 10 degrees of freedom. Notice that Gaussianity only holds approximately here and the variance of the observation distribution is also misspecified. We take X to be t-distributed noise completely uncorrelated with of the true effects. The results are displayed in Figure 11. All methods perform slightly worse in the misspecified setting, with larger variability in performance. As expected, the Fay-Herriot class performs comparably to plug-in shrinkage as the X_i are not predictive of the μ_i. Interestingly, ASSURE still improves on the plug-in empirical Bayes methods in the first three groups of methods, while giving an ensemble method which is on par with CLOSE-NPMLE.

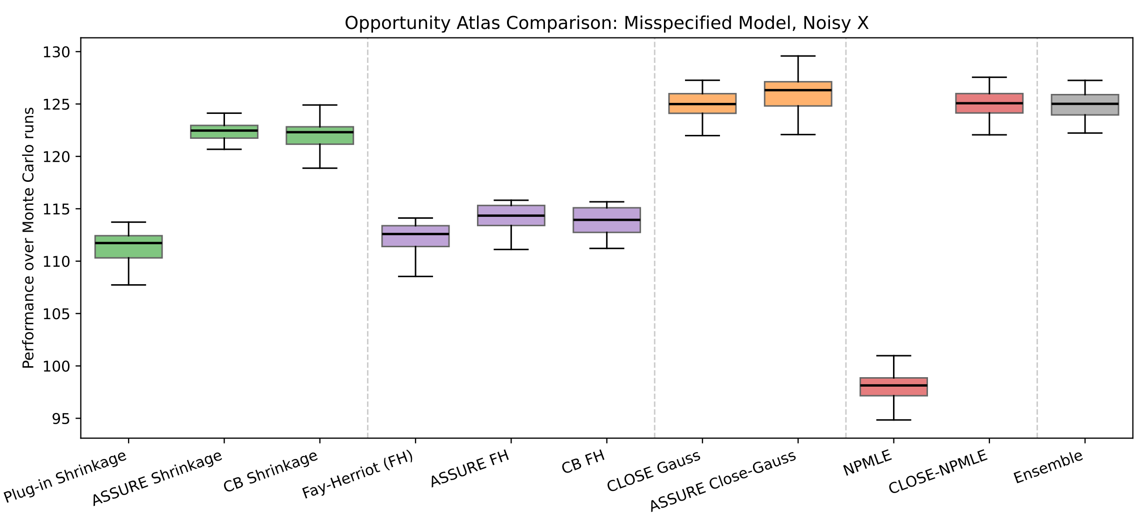


FIGURE 11. Same simulation setting as in Figure 2, but Y_i are drawn from the misspecified model described in the text. One-dimensional X_i covariates are a fixed draw $X_i \overset{i.i.d.}{\sim} 0.1t_{10}$ with constant costs $K = 0.361$.

We also consider two additional settings, a misspecified model with informative covariates and a well-specified model with noisy covariates. All the results are summarized as in Figure 12 and reinforce our previous claims. ASSURE is seen to produce slightly better decisions than coupled bootstrap. Both these methods allow for fine-tuning of the empirical Bayes plug-in decision, especially under misspecification. The only cases where ASSURE is inferior to the empirical Bayes plug-in is in the CLOSE-GAUSS class in the first and third rows, but the difference is only slight and can be attributed to the fact that the empirical Bayes procedure in this class estimates the conditional mean and variance using a nonparametric regression while the corresponding ASSURE method is restricted to a parametric specification. It comes as a nice surprise to see when the model is misspecified ASSURE makes a better decision than the nonparametric method.

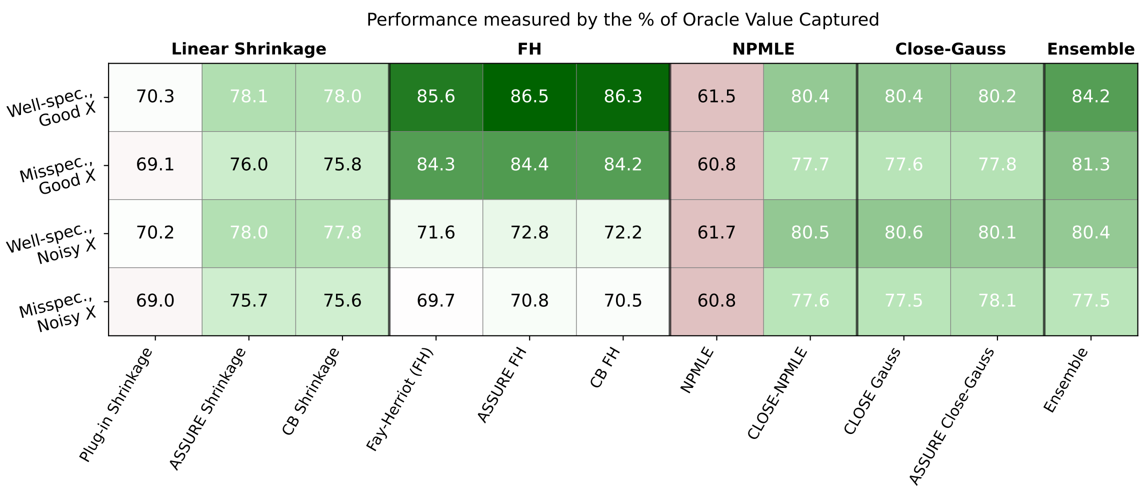


FIGURE 12. Comparison of all methods in four simulation settings for the Opportunity Atlas dataset.

D.1.2. *Calibrated Experimentation Program Data.* To generate the calibrated dataset on experimentation programs, we first take the dataset described in Section 6.2. Given $Y_{1:n}, \sigma_{1:n}$, we use heteroskedastic Gaussian NPMLE to obtain estimates $\tilde{\mu}_i$ of μ_i. The $\tilde{\mu}_i$ should only be interpreted as a sensible proxy for the unknown μ_i. We then treat these estimates $\tilde{\mu}_i$ as the true parameters, and regenerate a new synthetic dataset $\tilde{Y}_{1:n}, \sigma_{1:n}$ using $\tilde{Y}_i = \tilde{\mu}_i + \mathsf{N}(0, \sigma_i^2)$. In addition, we add costs $K_i = K$ for each decision problem i corresponding to the median of $\{|\tilde{\mu}_i|\}_{i=1}^n$.

D.2. **Empirical Application Details.**

D.2.1. *Additional Opportunity Atlas Details.* For our empirical application on data from the Opportunity Atlas, we consider $13,600$ census tracts in the top 20 Commuting Zones of the US. The metric of interest μ_i is the mean percentile rank of household income measured in adulthood for all Black children whose parents were at the 25th percentile of income. For each tract, an estimate Y_i and associated standard error σ_i is obtained from a particular regression specified in Chetty *et al.* (2018).

Regarding the cost specifications $\{0.2, 0.361, 0.369\}$ the first considers improvement over the baseline mean income of participants in the CMTO trial. Bergman *et al.* (2024) report that the mean income is \$20,000 for participants in the program and the program costs roughly \$2,670 per family issued a voucher. As a first attempt, we should select all tracts for which the dollar equivalent of μ_i is greater than \$22,670. This is roughly the 20th percentile rank of household income, so $K_i = 0.2$. To facilitate a closer comparison to choosing the top 1/3 of census tracts as in Bergman *et al.* (2024), we can also consider the costs which imply such a decision. If $K = 0.361$, then a plug-in empirical Bayes model chooses exactly 1/3 of units. Equivalently, the 67th percentile of the linear EB posterior means is 0.361. This is the second cost specification we consider. Finally, we consider the implied cost for the naive truncation decision rule $Y_i > K$ to select 1/3 of units. This corresponds to the 67th percentile of estimates Y_i.

Appendix E. Misc. Lemmas

Lemma E.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 $a_i(\cdot, Y_{1:n})$ is not almost surely zero or almost surely one. There exists a threshold rule $a_i^*(\cdot; Y_{-i}) = \mathbf{1}(Y_i > \delta_i(Y_{-i}))$, where $\delta_i(Y_{-i}) \in \mathbb{R}$, such that*

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

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 $a_i^*(y; Y_{-i}) = \mathbf{1}(y \ge c(Y_{-i}))$ such that

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

Note that $a_i^*(y; Y_{-i})$ is the uniformly most powerful test for $H_0 : \mu_i \le K_i$ 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

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

since a_i^* is weakly more powerful than a_i. Conversely, if $\mu_i < K_i$, then

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

since $1 - a^*$ is more powerful than $1 - a_i$. Thus

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

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

Lemma E.2 (Properties of 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 $|\operatorname{sinc}(y)| \le \pi^{-1}$ 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 E.3 (Derivatives of Ψ)**.** *Let Ψ_C, Ψ_{CC} denote the first and second partial derivatives in C. They are given by the following formulas.*

$$\Psi_C(Y_i, Z_i, C) = \frac{1}{\pi}\frac{(Y_i - K_i)}{h\sigma_i}\operatorname{sinc}\left(\frac{Y_i - C}{h\sigma_i}\right) - \frac{1}{h^2}\operatorname{sinc}'\left(\frac{Y_i - C}{h\sigma_i}\right) \tag{E.1}$$

$$\Psi_{CC}(Y_i, Z_i, C) = \frac{1}{\pi}\frac{(Y_i - K_i)}{h^2\sigma_i^2}\operatorname{sinc}'\left(\frac{Y_i - C}{h\sigma_i}\right) - \frac{1}{\sigma_i h^3}\operatorname{sinc}''\left(\frac{Y_i - C}{h\sigma_i}\right) \tag{E.2}$$

E.1. **Lack of an unbiased estimator.** We present an argument that is referenced in Stefanski (1989) without proof. By Lemma E.4, 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 E.4. *Let* $Y \sim \mathsf{N}(\mu, \sigma^2)$*. 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 E.4. Rewrite the expression as

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

□

Proposition E.1 (Stefanski (1989))**.** *Let* $f : \mathbb{R} \to \mathbb{R}$ *be some real-valued function where* $\mathbb{E}_{Y\sim\mathsf{N}(\mu,1)}[|f(Y)|] < \infty$*. 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 $t \in \mathbb{R}$,

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

Note that

$$\int |f(y)\varphi(y)|dy = \int |f(y)|\varphi(y)\,dy = \mathbb{E}_{Y\sim\mathsf{N}(0,1)}[|f(Y)|] < \infty.$$

Thus $f\varphi$ is L^1. By the Riemann–Lebesgue lemma, the Fourier transform vanishes at infinity:

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

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: $\mathbb{E}_{Y\sim\mathsf{N}(0,1+\epsilon^2)}[|f(Y)|] < \infty$. 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.