Empirical Bayes Rebiasing

Wanyi Ling
wanyiling@uchicago.edu

Sida Li
listar2000@uchicago.edu

Junming Guan
junmingguan@uchicago.edu

Nikolaos Ignatiadis
ignat@uchicago.edu

Draft Manuscript, May 2026

Abstract

We study methods for simultaneous analysis of many noisy and biased estimates, each paired with an even noisier estimate of its own bias. The analyst's goal is to construct short calibrated intervals for each parameter. The standard debiasing approach, which subtracts the bias estimate from each biased estimate, inflates variance and yields long intervals. In this paper, we propose an empirical Bayes rebiasing strategy that starts from the fully debiased estimates and learns from data how much bias to reintroduce by estimating the unknown bias distribution. We provide convergence rates for the coverage of our intervals when the bias distribution is estimated using nonparametric maximum likelihood. Furthermore, we demonstrate substantial precision gains in prediction-powered inference, including pairwise LLM win-rate evaluations, as well as for inference of direct genetic effects in family-based GWAS.

1 Introduction

One of the simplest, yet most widely used statistical constructions is the Wald interval. Given a parameter θ and an estimator $\hat{\theta}$, we report the interval $\mathcal{I} = \hat{\theta} \pm 1.96\widehat{\mathrm{Var}}[\hat{\theta}]^{1/2}$, where $\widehat{\mathrm{Var}}[\cdot]$ denotes the estimated variance. This interval satisfies $\mathbb{P}[\theta \in \mathcal{I}] \approx 95\%$ under three conditions:

$$\text{(i)} \quad \hat{\theta} \approx \mathrm{N}\big(\mathbb{E}_\theta[\hat{\theta}],\, \mathrm{Var}_\theta[\hat{\theta}]\big), \quad \text{(ii)} \quad \frac{\widehat{\mathrm{Var}}[\hat{\theta}]}{\mathrm{Var}_\theta[\hat{\theta}]} \approx 1, \quad \text{(iii)} \quad \frac{|\,\mathrm{Bias}_\theta[\hat{\theta}]|^2}{\mathrm{Var}_\theta[\hat{\theta}]} \approx 0, \tag{1}$$

where $\mathrm{Bias}_\theta[\hat{\theta}] = \mathbb{E}_\theta[\hat{\theta}] - \theta$. In this paper, we are concerned with (iii); namely the requirement that $\hat{\theta}$ be (nearly) unbiased for θ, and we propose methods that relax this condition. The approximate normality in (i) can be justified by the central limit theorem, and the variance in (ii) can be estimated accurately, e.g., via the bootstrap.

Our starting point is that the analyst may prefer a potentially biased estimator $\hat{\theta}^{\mathrm{b}}$ of θ (say, an ML-based estimator computed on a large unlabeled corpus) because it has substantially lower variance than unbiased alternatives, and because the analyst has reason to believe the bias to be small (Fig. 1a). In terms of overall mean squared error, $\hat{\theta}^{\mathrm{b}}$ may incur a favorable bias-variance tradeoff. At the same time, the analyst may hesitate to report Wald intervals centered at $\hat{\theta}^{\mathrm{b}}$, since the coverage guarantee breaks down when (iii) is violated. A standard remedy is to debias: with access to a noisy unbiased estimator $\hat{b}$ of the bias $b = \mathrm{Bias}_\theta[\hat{\theta}^{\mathrm{b}}]$, one forms the debiased estimator $\hat{\theta}^{\mathrm{db}} = \hat{\theta}^{\mathrm{b}} - \hat{b}$, which is unbiased for θ (Fig. 1b). The debiased estimator is protected against arbitrarily large bias in $\hat{\theta}^{\mathrm{b}}$: it is unbiased no matter how badly biased $\hat{\theta}^{\mathrm{b}}$ happens to be. Such robustness is well-motivated, since the naïve alternative of trusting $\hat{\theta}^{\mathrm{b}}$ fails to deliver coverage unless the bias is negligible, and without prior

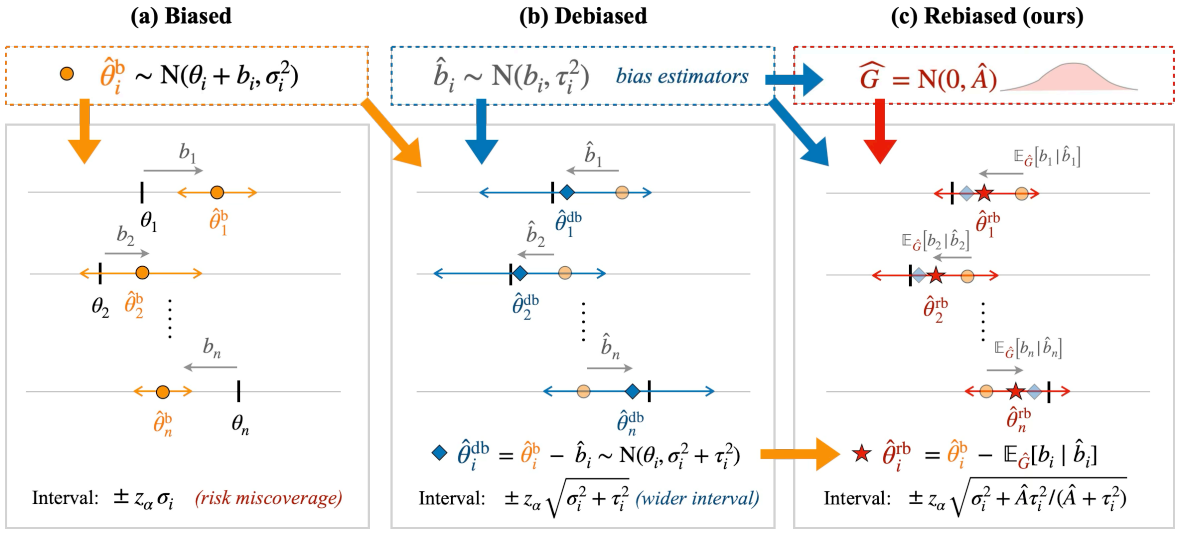

Figure 1: **From biased to debiased to rebiased.** Vertical **black** ticks mark the true θ_i. **(a)** Across n parallel tasks, each biased estimator $\hat{\theta}_i^{\mathrm{b}}$ is offset from its target θ_i by a draw $b_i \sim G$; the small variance σ_i^2 makes $\hat{\theta}_i^{\mathrm{b}}$ precise, but the resulting interval may undercover. **(b)** Subtracting an unbiased but noisy estimator $\hat{b}_i$ centers the estimator at θ_i *on average* but each individual $\hat{\theta}_i^{\mathrm{db}}$ is itself noisy, with a wider confidence interval. **(c)** Empirical Bayes rebiasing estimates the task-level bias distribution $\widehat{G}$ from $\hat{b}_1, \ldots, \hat{b}_n$ (shown here as $\widehat{G} = \mathrm{N}(0, \widehat{A})$ for illustration) and uses its posterior mean to partially undo the bias correction and yield a substantially shorter calibrated interval.

knowledge about the bias, it is prudent to be robust against bias of any magnitude. The cost is that the variance of $\hat{\theta}^{\mathrm{db}}$ is generally larger than that of $\hat{\theta}^{\mathrm{b}}$ (by $\mathrm{Var}_b[\hat{b}]$ when $\hat{\theta}^{\mathrm{b}}$ and $\hat{b}$ are uncorrelated), undoing the efficiency gains that motivated $\hat{\theta}^{\mathrm{b}}$ in the first place.

In many settings, however, the analyst faces not a single inference task but many. We consider parameters $\theta_1, \ldots, \theta_n$, where i indexes the task; for instance, the θ_i may be win-rates for n LLM pairs, or effect sizes for n genetic variants. Each task comes with its own pair $(\hat{\theta}_i^{\mathrm{b}}, \hat{b}_i)$. Across these tasks, the bias estimators $\hat{b}_1, \ldots, \hat{b}_n$ carry rich indirect information about the typical magnitude of the bias. The central premise of this paper is that bias should not always be treated as either zero, as in the naïve approach of reporting $\hat{\theta}_i^{\mathrm{b}}$ directly, or as arbitrary, as in full debiasing. Instead, we propose to model the bias, and we develop new calibrated inference methods that use empirical Bayes (EB) ideas [Robbins, 1956, Efron, 2010, Stephens, 2017] to learn how much bias correction to apply. Concretely, we place a distribution on task-level biases, $b_i \sim G$ for an unknown G, estimate G as $\widehat{G}$ via empirical Bayes, and use $\widehat{G}$ to partially undo the full bias correction when the evidence supports doing so. We refer to this approach as empirical Bayes rebiasing (Fig. 1c).

To preview the methodology, suppose for now that $G = \mathrm{N}(0, A)$, where $A \geq 0$ is the variance of the bias, and write $\sigma_i^2 = \mathrm{Var}_{\theta_i}[\hat{\theta}_i^{\mathrm{b}}]$ and $\tau_i^2 = \mathrm{Var}_{b_i}[\hat{b}_i]$. We model $\hat{b}_i \mid b_i \sim \mathrm{N}(b_i, \tau_i^2)$ with τ_i^2 known, in the spirit of (i) and (ii) in (1). Given access to the unbiased estimators of bias $\hat{b}_1, \ldots, \hat{b}_n$, we learn A via marginal maximum likelihood: if biases are small across tasks, then $\widehat{A} \approx 0$, while if they are large and heterogeneous, then $\widehat{A} \gg 0$. We then combine the prior information $b_i \sim \mathrm{N}(0, \widehat{A})$ with the unbiased estimator $\hat{b}_i$ via Bayes' rule, yielding the posterior $b_i \mid \hat{b}_i \sim \mathrm{N}(\widehat{A}/\{\widehat{A} + \tau_i^2\}\hat{b}_i,\ \widehat{A}\tau_i^2/\{\widehat{A} + \tau_i^2\})$, and subtract the posterior mean of the bias rather than $\hat{b}_i$ itself. The resulting rebiased estimator takes three equivalent forms,

$$\hat{\theta}_i^{\mathrm{rb}} \;=\; \hat{\theta}_i^{\mathrm{b}} - \frac{\widehat{A}}{\widehat{A} + \tau_i^2}\hat{b}_i \;=\; \hat{\theta}_i^{\mathrm{db}} + \frac{\tau_i^2}{\widehat{A} + \tau_i^2}\hat{b}_i \;=\; \frac{\widehat{A}}{\widehat{A} + \tau_i^2}\hat{\theta}_i^{\mathrm{db}} + \frac{\tau_i^2}{\widehat{A} + \tau_i^2}\hat{\theta}_i^{\mathrm{b}}, \tag{2}$$

and satisfies $\hat{\theta}_i^{\mathrm{rb}} - \theta_i \sim \mathrm{N}(0,\ \sigma_i^2 + \widehat{A}\tau_i^2/\{\widehat{A} + \tau_i^2\})$ (with b_i integrated out), yielding the rebiased inter-

val $\mathcal{I}_i^{\text{rb}} = \hat{\theta}_i^{\text{rb}} \pm z_\alpha(\sigma_i^2 + \widehat{A}\tau_i^2/\{\widehat{A} + \tau_i^2\})^{1/2}$, where z_α is the $(1-\alpha/2)$-standard normal quantile. The estimator interpolates between the two endpoints the analyst was previously stuck choosing between: $\widehat{A} = 0$ recovers $\hat{\theta}_i^{\text{b}}$, $\widehat{A} = \infty$ recovers $\hat{\theta}_i^{\text{db}}$, and intermediate $\widehat{A}$ partially undoes the bias correction. As Fig. 1 illustrates, $\mathcal{I}_i^{\text{rb}}$ is shorter than the Wald interval of the debiased estimator, with the gap governed by $\widehat{A}$. The interval $\mathcal{I}_i^{\text{rb}}$ synthesizes three sources of information: the biased estimator $\hat{\theta}_i^{\text{b}}$, the bias correction $\hat{b}_i$, and the estimated task-level distribution of the bias $\widehat{G}$.

2 Background and related work

A common modeling assumption is that biases across tasks are exchangeable draws from a distribution, often a normal $b_i \sim \text{N}(0, A)$ with $A \geq 0$ (or uncentered, $b_i \sim \text{N}(\mu, A)$). Such a normal component can capture, e.g., biases from unobserved confounding in observational studies, and A can be estimated from negative control studies. For other studies, one then reports the inflated interval $\hat{\theta}_i^{\text{b}} \pm z_\alpha(\sigma_i^2 + \widehat{A})^{1/2}$, where $\sigma_i^2 = \text{Var}_{\theta_i}[\hat{\theta}_i^{\text{b}}]$. This construction [Schuemie et al., 2014, 2016] is routinely used before reporting OHDSI (Observational Health Data Sciences and Informatics) results [Hripcsak et al., 2015] and has recently been advocated in economics [Bernard et al., 2024]. As Efron [2022] notes, treating bias as a further random component is akin to the physicist's "propagation of uncertainty"; we observe that it is also closely related to empirical null modeling [Efron, 2004]. We likewise model bias as drawn from a distribution G, but synthesize this distributional information with the direct per-task estimators $\hat{b}_i$ to obtain shorter intervals, and we allow G to be estimated nonparametrically. The normality assumption on biases also appears in related but distinct settings: Wu et al. [2026] use it for meta-analysis combining several biased observational estimators with a single unbiased experimental estimators, and Zhao et al. [2020] uses it to model bias from invalid genetic instruments (systematic pleiotropy) in Mendelian randomization.

Several authors have proposed taking task-wise convex combinations of unbiased and biased estimators using empirical Bayes ideas [Green and Strawderman, 1991, Green et al., 2005, Chen et al., 2015, Ignatiadis and Wager, 2019, Rosenman et al., 2023a, Li and Ignatiadis, 2025], which is precisely the form of the rebiased estimator in (2) (last equality). Building on the seminal work of James and Stein [1961], these papers seek estimators with smaller frequentist mean squared error, but focus on point estimation rather than interval coverage. Such methods admit an empirical Bayes interpretation under a normal prior for b_i and a flat improper prior for θ_i [Green and Strawderman, 1991]. A fully hierarchical approach with proper priors on both b_i and θ_i is taken in Gelman and Vákár [2021], Rosenman et al. [2023b]. We instead adopt a partially Bayes analysis [Brown, 1965, Cox, 1975] that places a prior only on b_i, while treating θ_i as fixed; in particular, we do not assume exchangeability of the θ_i across tasks. Our framing lets us state coverage guarantees in the empirical-Bayes-coverage tradition [Morris, 1983, Rubin, 1984], with formal results in §5.

A different line of work assumes a deterministic upper bound on the bias, $|b_i| \leq \Delta_i$, and combines $\hat{\theta}_i^{\text{b}}$ and $\hat{b}_i$ so as to (nearly) minimize the worst-case risk over $\{|b_i| \leq \Delta_i\}$ [Donoho, 1994, Armstrong and Kolesár, 2018, Lin et al., 2026]. Adaptation results [Cai and Low, 2004, Armstrong and Kolesár, 2021] show that Δ_i cannot be learned adaptively in a way that yields shorter intervals. These impossibility results concern a single task in isolation, whereas our setting offers a different opportunity: that of having many related tasks and the empirical Bayes paradigm [Robbins, 1956, Efron, 2010, Stephens, 2017]. By replacing the deterministic constraint $|b_i| \leq \Delta_i$ with a distributional assumption $b_i \sim G$, we can learn G from the data across tasks.

3 Statistical setting and proposed methods

Our statistical model for the i-th task is as follows,

$$\begin{pmatrix} \hat{\theta}_i^{\text{b}} \\ \hat{b}_i \end{pmatrix} \Bigg| \begin{pmatrix} \theta_i \\ b_i \end{pmatrix} \sim \text{N}\left\{ \begin{pmatrix} \theta_i + b_i \\ b_i \end{pmatrix}, \begin{pmatrix} \sigma_i^2 & \rho_i\sigma_i\tau_i \\ \rho_i\sigma_i\tau_i & \tau_i^2 \end{pmatrix} \right\}, \qquad b_i \sim G, \qquad \theta_i \text{ is fixed.} \tag{3}$$

To focus on the issue of the bias, we treat normality as holding exactly and assume that $\rho_i \in (-1, 1)$ and $\sigma_i^2, \tau_i^2 > 0$ are known (these pertain to (i) and (ii) in (1)). Meanwhile, θ_i, b_i, and G are unknown. We posit that all biases b_i are drawn from the same distribution G, that is, we assume bias is exchangeable across tasks and that G does not depend on θ_i. We treat the θ_i as fixed (as in a frequentist analysis). An analysis in which we treat the primary parameters (θ_i) as fixed, while the nuisance parameters (b_i) as drawn from a prior is called a partially Bayes analysis [Cox, 1975], which dates back to a proposal by John Tukey [Brown, 1965]; also see Ignatiadis and Sen [2025].

Oracle partially Bayes rebiasing. We start by explaining our approach to inference when the bias distribution G in (3) is known. (Later, we will explain how to estimate G.) The goal is to combine the distribution of the biases (G) with $\hat{b}_i$ via (a partial version of) Bayes' rule. We state our constructions in a general way that accommodates *any* fixed prior G (when $G = \mathrm{N}(\mu, A)$ expressions simplify as previewed in §1 and further explained in Appendix A.1). To streamline exposition, below we assume $\rho_i = 0$; the general case (relevant for our applications) is treated in Appendix A.2.

Although point estimation is not our focus, for intuition, we first state a generalization of the estimator $\hat{\theta}_i^{\mathrm{rb}}$ in (2). Rather than subtracting $\hat{b}_i$ from $\hat{\theta}_i^{\mathrm{b}}$, we subtract the posterior mean of the bias,[1]

$$\hat{\theta}_i^{\mathrm{rb}} = \hat{\theta}_i^{\mathrm{b}} - \mathbb{E}_G[b_i \mid \hat{b}_i] = \hat{\theta}_i^{\mathrm{db}} + \left\{\hat{b}_i - \mathbb{E}_G[b_i \mid \hat{b}_i]\right\}. \tag{4}$$

We also interpret this procedure via the right-hand side of (4) as rebiasing the debiased estimator $\hat{\theta}_i^{\mathrm{db}}$. To be more explicit about the bias-only posterior, for any measurable set $A \subset \mathbb{R}$, we write

$$\mathbb{P}_G[b_i \in A \,|\, \hat{b}_i] = \frac{\int_A \varphi(\hat{b}_i - b;\, \tau_i^2) G(\mathrm{d}b)}{\int \varphi(\hat{b}_i - b;\, \tau_i^2) G(\mathrm{d}b)}, \quad \mathbb{E}_G[b_i \,|\, \hat{b}_i] = \frac{\int b\, \varphi(\hat{b}_i - b;\, \tau_i^2) G(\mathrm{d}b)}{\int \varphi(\hat{b}_i - b;\, \tau_i^2) G(\mathrm{d}b)},$$

where $\varphi(\cdot; \tau_i^2)$ is the density of the $\mathrm{N}(0, \tau_i^2)$ distribution. For our purposes, the relevant object is the conditional density of $\hat{\theta}_i^{\mathrm{db}} - \theta_i$ conditional on $\hat{b}_i$ (that is, integrating over the posterior $b_i \mid \hat{b}_i$),

$$f_{G,i}(t \mid \hat{b}_i) \;=\; \frac{\int \varphi(t - b + \hat{b}_i;\, \sigma_i^2)\, \varphi(\hat{b}_i - b;\, \tau_i^2)\, G(\mathrm{d}b)}{\int \varphi(\hat{b}_i - b;\, \tau_i^2)\, G(\mathrm{d}b)}. \tag{5}$$

We also write $F_{G,i}(\cdot \mid \hat{b}_i)$ for the distribution function with density $f_{G,i}(\cdot \mid \hat{b}_i)$ and $q_{G,i,\alpha}(\hat{b}_i)$ for its α-quantile. The oracle rebiased equal-tailed $(1-\alpha)$-interval is then defined as

$$\mathcal{I}_{G,i}^{\mathrm{rb}}(1-\alpha) = \left[\hat{\theta}_i^{\mathrm{db}} - q_{G,i,1-\alpha/2}(\hat{b}_i),\, \hat{\theta}_i^{\mathrm{db}} - q_{G,i,\alpha/2}(\hat{b}_i)\right]. \tag{6}$$

Beyond intervals, we also consider the corresponding testing problem. For testing $H_{0i} : \theta_i = \theta_{i0}$ for pre-specified θ_{i0}, we define the oracle rebiased p-value as,

$$P_{G,i}^{\mathrm{rb}} = P_{G,\theta_{i0}}^{(i)}(\hat{\theta}_i^{\mathrm{db}}, \hat{b}_i), \;\text{ with }\; P_{G,\theta_{i0}}^{(i)}(z, l) = 2\min\{F_{G,i}(z - \theta_{i0} \mid l), 1 - F_{G,i}(z - \theta_{i0} \mid l)\}. \tag{7}$$

Our proposed oracle rebiased interval in (6) can be derived by test inversion of the oracle p-values,

$$\mathcal{I}_{G,i}^{\mathrm{rb}}(1-\alpha) = \{\theta_{i0} \in \mathbb{R} \,:\, \alpha/2 \le F_{G,i}(\hat{\theta}_i^{\mathrm{db}} - \theta_{i0} \mid \hat{b}_i) \le 1 - \alpha/2\}. \tag{8}$$

Empirical Bayes rebiasing. We next turn to the estimation of G in (3). As mentioned in §2, a common modeling choice for the bias distribution is that $G = \mathrm{N}(\mu, A)$. We can estimate the unknown μ and A using maximum marginal likelihood; see Appendix A.3. We emphasize, that unlike normality statements for $\hat{\theta}_i^{\mathrm{b}}$ and $\hat{b}_i$ in (3), normality of b_i does not follow from the central limit theorem. For this reason, when normal G is imposed, it is prudent to check the fit of the model. In what follows

[1] A similar estimator for general G appears in Kwon and Roth [2024, Proposition 1] based on a flat prior for θ_i.

we pursue an alternative avenue: estimating G in a fully nonparametric way, without imposing a parametric form.

We propose to estimate G based on $(\hat{b}_1, ..., \hat{b}_n)$ by maximizing the marginal log-likelihood,

$$\widehat{G} \in \operatorname*{argmax}_{G'} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log \left(\int \varphi(\hat{b}_i - b;\, \tau_i^2)\, G'(\mathrm{d}b) \right) \right\}, \tag{9}$$

over *all* possible priors G'. This is the nonparametric maximum likelihood estimator (NPMLE) of Robbins [1950], Kiefer and Wolfowitz [1956]. The NPMLE is tuning-free and has strong regret guarantees for empirical Bayes problems [Jiang and Zhang, 2009]. Although the feasible set is infinite-dimensional, the optimizer is a discrete distribution with at most n atoms, and typically only $O(\log n)$, supported in $[\min_i\{\hat{b}_i\}, \max_i\{\hat{b}_i\}]$ [Lindsay, 1995, Polyanskiy and Wu, 2020]. The problem can be discretized without sacrificing statistical guarantees [Dicker and Zhao, 2016, Soloff et al., 2024] and recast as a conic program [Koenker and Mizera, 2014, Koenker and Gu, 2017], which we solve with MOSEK [MOSEK ApS, 2024]. Further details appear in Appendix A.3.

Given our estimate of $\widehat{G}$ (which could be the parametric normal prior or the NPMLE), we then compute the rebiased intervals using the plug-in principle, that is, pretending that $\widehat{G}$ is the true prior G. The rebiased $(1-\alpha)$-interval is constructed as

$$\mathcal{I}^{\mathrm{rb}}_{\widehat{G},i}(1-\alpha) = \left[\hat{\theta}^{\mathrm{db}}_i - q_{\widehat{G},i,1-\alpha/2}(\hat{b}_i),\ \hat{\theta}^{\mathrm{db}}_i - q_{\widehat{G},i,\alpha/2}(\hat{b}_i) \right]. \tag{10}$$

We can also estimate the oracle rebiased p-value $P^{\mathrm{rb}}_{G,i}$ by the empirical Bayes rebiased p-value $\widehat{P}^{\mathrm{rb}}_{\widehat{G},i}$, where the estimators are computed using (7) by replacing G with $\widehat{G}$. All of these computations can be carried out readily; e.g., for the NPMLE we can leverage the fact that it yields a discrete prior and so integrals reduce to sums.

4 Applications of empirical Bayes rebiasing

Our framework can be applied in two ways, depending on the auxiliary estimator paired with $\hat{\theta}^{\mathrm{b}}_i$. In the first, the researcher has the biased estimator $\hat{\theta}^{\mathrm{b}}_i$ and a direct bias-correction estimator $\hat{b}_i$, as in the setup so far. In the second, the researcher has $\hat{\theta}^{\mathrm{b}}_i$ together with an unbiased estimator $\hat{\theta}^{\mathrm{ub}}_i$, in which case the framework applies with $\hat{b}_i := \hat{\theta}^{\mathrm{b}}_i - \hat{\theta}^{\mathrm{ub}}_i$ (so that $\hat{\theta}^{\mathrm{db}}_i = \hat{\theta}^{\mathrm{ub}}_i$). The two are formally equivalent but arise differently in practice: in some applications a bias estimate is computed directly, while in others it is only implicit in the difference between two estimators. Here, we consider two prototypical applications: prediction-powered inference (PPI) and family-based genome-wide association studies (GWAS). The framework is broadly applicable beyond these examples; for instance, it covers combining observational and experimental causal estimates [Rosenman et al., 2023a,b] and combining exposed-only with difference estimates in sham-controlled experiments [Gelman and Vákár, 2021].

Prediction-powered inference (PPI). PPI [Angelopoulos et al., 2023, 2024] is a framework for constructing more accurate estimates by incorporating predictions from a black-box ML model h. For task i, we observe a small labeled sample $\{(X_{ij}, Y_{ij})\}_{j=1}^{m_i}$ with $Y_{ij} \in \mathbb{R}$ and a larger unlabeled sample $\{\tilde{X}_{ij}\}_{j=1}^{M_i}$ (with $m_i \ll M_i$ and $\tilde{X}_{ij} \overset{d}{=} X_{ij}$); the goal is to estimate $\theta_i = \mathbb{E}[Y_{ij}]$.[2] The classical mean $\bar{Y}_i = m_i^{-1} \sum_j Y_{ij}$ is unbiased but has variance $\propto 1/m_i$, while the ML-only estimate $\hat{\theta}^{\mathrm{ML}}_i = M_i^{-1} \sum_j h(\tilde{X}_{ij})$ has small variance ($\propto 1/M_i$) but bias $b_i = \mathbb{E}_{\theta_i}[h(X_{ij})] - \theta_i$ that depends on the quality of h. The PPI estimator bridges the two by debiasing $\hat{\theta}^{\mathrm{ML}}_i$ with the labeled sample:

$$\hat{\theta}^{\mathrm{PPI}}_i = \hat{\theta}^{\mathrm{ML}}_i - \hat{b}_i, \quad \hat{b}_i = m_i^{-1} \textstyle\sum_{j=1}^{m_i} \{h(X_{ij}) - Y_{ij}\}. \tag{11}$$

[2] Our framework extends to general estimands beyond the mean via the predict-then-debias strategy of Kluger et al. [2025].

Setting $\hat{\theta}_i^{\mathrm{b}} := \hat{\theta}_i^{\mathrm{ML}}$ as the biased estimator, the PPI estimator is exactly the debiased estimator $\hat{\theta}_i^{\mathrm{db}}$ in our framework. The vanilla PPI is asymptotically dominated by the PPI++/PT (power tuned) estimator of Angelopoulos et al. [2024] that takes the form $\hat{\theta}_i^{\mathrm{PT}} := (1-\lambda_i^*)\bar{Y}_i + \lambda_i^* \hat{\theta}_i^{\mathrm{PPI}}$ for a choice of λ_i^*. In this case, the vantage point to apply our rebiasing is to treat $\hat{\theta}_i^{\mathrm{PT}}$ as the unbiased estimator and $\hat{\theta}_i^{\mathrm{ML}}$ as the biased estimator. In our experiments, we only focus on PT. We provide more details in Appendix B.1 including expressions and plug-in estimates for ρ_i, σ_i, τ_i in (3) that we use in practice.

PPI/PT benefit from the ML predictor without requiring it to be accurate, but they consider only unbiased estimators of θ_i. Our rebiasing goes further by learning the bias distribution across tasks from the observed bias estimates. When the data reveal that biases are small, i.e., h is approximately calibrated (which is a milder requirement than accurately predicting Y_{ij}), the rebiased estimator moves back toward $\hat{\theta}_i^{\mathrm{ML}}$, yielding shorter intervals; when biases are large, it stays close to PPI/PT. A related approach is the prediction-powered adaptive shrinkage (PAS) estimator of Li and Ignatiadis [2025] that shares information across tasks to reduce mean squared error of PT; our rebiasing pools across tasks to deliver calibrated inference instead.

Inference of direct genetic effects in family-based GWAS. A standard genome-wide association study (GWAS) regresses a trait on each SNP (single nucleotide polymorphism) across the genome. The regression coefficient, denoted $\hat{\theta}_i^{\mathrm{b}}$ in our framework, and often termed the population-effect estimate in the literature, is biased in the sense that it captures not only the direct genetic effect θ_i, but also a bias component b_i, which may reflect contributions from population stratification, indirect genetic effects, and assortative mating [Kong et al., 2018]. The variance of $\hat{\theta}_i^{\mathrm{b}}$ typically scales as $1/m_i$, where m_i is the regression sample size for the ith SNP. When parental genotypes are observed or unbiasedly imputed [Young et al., 2022, Guan et al., 2025], family-based GWAS includes them as covariates in the SNP-level regression; the resulting offspring-genotype coefficient $\hat{\theta}_i^{\mathrm{ub}}$ is unbiased for θ_i, but has the larger variance $\propto 1/\{m_i(1-r_i^2)\}$, where r_i is the offspring-parental genotype correlation at SNP i. The pair $(\hat{\theta}_i^{\mathrm{b}}, \hat{\theta}_i^{\mathrm{ub}})$ fits our framework directly, with bias estimator $\hat{b}_i = \hat{\theta}_i^{\mathrm{b}} - \hat{\theta}_i^{\mathrm{ub}}$, which coincides with the parental-genotype regression coefficient. Summary statistics provided for family-based GWAS allow us to also recover $\rho_i, \sigma_i^2, \tau_i^2$ in (3) (see Appendix B.2 for details) and so we can apply our empirical Bayes rebiasing strategy that learns the bias distribution across SNPs.

5 Theoretical results

We work under the partially Bayes model (3), with θ_i a fixed parameter, $b_i \sim G$, and $(\hat{\theta}_i^{\mathrm{db}}, \hat{b}_i)$ jointly normal given (θ_i, b_i); all probabilities and expectations below are taken jointly over the n tasks under this model, with $\theta_1, \ldots, \theta_n$ fixed and we only make the dependence on G explicit (via a subscript). Our goal is to show coverage guarantees in this setting, in the spirit of the empirical Bayes coverage tradition [Morris, 1983], except that the classical version also averages over the parameters of interest θ_i.[3] Throughout, $\widehat{G}$ denotes the NPMLE in (9); we focus on the nonparametric case, which both reveals the statistical structure of the problem and the cost of fully nonparametric modeling.

Let $\tilde{\sigma}_i^2 = \mathrm{Var}[\hat{\theta}_i^{\mathrm{db}}] = \sigma_i^2 + \tau_i^2 - 2\rho_i\sigma_i\tau_i$, $\gamma_i = \mathrm{Corr}[\hat{\theta}_i^{\mathrm{db}}, \hat{b}_i] = (\rho_i\sigma_i - \tau_i)/\tilde{\sigma}_i$, and $\check{\gamma} = \max_i |\gamma_i|$. Moreover, for $\Gamma > 0$ we write $\mathcal{G}_\Gamma = \{G' : \mathbb{E}_{G'}[\exp\{\lambda(b - \mathbb{E}_{G'}[b])\}] \le \exp(\Gamma\lambda^2/2) \text{ for all } \lambda \in \mathbb{R}\}$ for the class of Γ-sub-Gaussian distributions. Our main assumption in this section is follows.

Assumption 1. We assume that the true bias distribution satisfies $G \in \mathcal{G}_\Gamma$ for some $\Gamma > 0$ and that the triples $(b_i, \hat{\theta}_i^{\mathrm{b}}, \hat{b}_i)$ for $i = 1, \ldots, n$ are jointly independent. Finally, we assume that for all i, $0 < \underline{\gamma} \le |\gamma_i| \le \bar{\gamma} < 1$ and $0 < \underline{\sigma}^2 \le \tilde{\sigma}_i^2, \tau_i^2 \le \bar{\sigma}^2 < \infty$ for some constants $\underline{\gamma}, \bar{\gamma}, \underline{\sigma}, \bar{\sigma} > 0$.

The next result establishes the rate at which we estimate the oracle rebiasing p-values in (7).

[3] One could formally obtain our intervals by imposing a flat prior on θ_i (cf. §2). However, such an improper prior does not define a sampling distribution, so the frequency interpretation of any coverage statement is unclear in that formulation.

Theorem 2. Under Assumption 1, there exists $C = C(\Gamma, \underline{\gamma}, \bar{\gamma}, \underline{\sigma}, \bar{\sigma})$ such that,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_G\left[\left|P^{(i)}_{G,\theta_{i0}}(\hat{\theta}_i^{\mathrm{db}}, \hat{b}_i) - P^{(i)}_{\widehat{G},\theta_{i0}}(\hat{\theta}_i^{\mathrm{db}}, \hat{b}_i)\right|\right] \le C\frac{(\log n)^{3/2}}{n^{(1-\check{\gamma}^2)/2}} \quad \text{for all } n \in \mathbb{N}_{\ge 2}.$$

For moderate $\check{\gamma}$ and n, this rate suggests little cost to pursuing the fully nonparametric strategy; as $\check{\gamma} \nearrow 1$, however, the rate slows, and stronger (e.g., parametric) modeling assumptions on G may be worthwhile. The value of $\check{\gamma}$ is known to the analyst; in our three applications (§6), $\check{\gamma} \in \{0.752, 0.892, 0.992\}$. The dependence on $\check{\gamma}$ is nearly sharp in a minimax sense.

Proposition 3. Suppose that $\tilde{\sigma}_i^2, \tau_i^2, \gamma_i$ do not depend on i and drop the subscript i. Let $\check{\gamma} = |\gamma| \in (0,1)$. Fix $t \in \mathbb{R}$. Then, for any $\beta > (1-\check{\gamma}^2)/2$ there exists $\Gamma > 0$ and $c > 0$ such that

$$\inf_{\widehat{\psi}(t)} \sup_{G \in \mathcal{G}_\Gamma} \mathbb{E}_G\left[\left|\widehat{\psi}(t) - F_{G,i}(t \mid \hat{b}_i)\right|\right] \ge c\frac{1}{n^{\beta}}, \text{ where } \widehat{\psi}(t) \text{ is any measurable function of } \hat{b}_1, \ldots, \hat{b}_n.$$

Building upon Theorem 2, we have the following convergence rate for the coverage of $\mathcal{I}^{\mathrm{rb}}_{\widehat{G},i}(1-\alpha)$.

Theorem 4. Under Assumption 1, there exists $C' = C'(\Gamma, \underline{\gamma}, \bar{\gamma}, \underline{\sigma}, \bar{\sigma})$ such that,

$$\sup_{\alpha \in (0,1)} \frac{1}{n}\sum_{i=1}^{n} \left|\mathbb{P}_G\left[\theta_i \in \mathcal{I}^{\mathrm{rb}}_{\widehat{G},i}(1-\alpha)\right] - (1-\alpha)\right| \le C'\frac{(\log n)^{3/2}}{n^{(1-\check{\gamma}^2)/2}} \quad \text{for all } n \in \mathbb{N}_{\ge 2}.$$

A classical empirical Bayes analysis would instead place an NPMLE prior directly on θ_i, in which case no analogous coverage guarantee is available. The challenge lies in the discrete nature of the NPMLE, and is reflected in bad coverage in finite samples [Jiang, 2019, Koenker, 2020]. The partially Bayes formulation sidesteps this issue, since the NPMLE is applied to b_i and integrated out when forming intervals on θ_i, so its discreteness does not propagate to the resulting inference.

6 Numerical results

6.1 Experiments for prediction-powered inference

Estimators. Recall the PPI setup of §4. We compare our rebiased intervals against three baselines: the *Classical* interval $(\bar{Y}_i \pm z_\alpha \operatorname{Var}[\bar{Y}_i]^{1/2}$; not using ML information); the *Pred Mean* interval $(\hat{\theta}_i^{\mathrm{ML}} \pm z_\alpha \operatorname{Var}[\hat{\theta}_i^{\mathrm{ML}}]^{1/2}$; without correcting for bias); and the power-tuned PPI *PT* interval [Angelopoulos et al., 2024] centered around the unbiased $\hat{\theta}_i^{\mathrm{PT}}$. Our rebiased intervals are built on top of PT by fitting the bias prior G on b_i, either as a Normal $G = \mathrm{N}(\mu, A)$ (suffix *Normal*) or nonparametrically (suffix *NPMLE*). This yields two rebiased intervals: *RB-Normal*, and *RB-NPMLE*. See Fig. 2 for the fitted priors in our applications.

Metrics. We report empirical *coverage* at level $1-\alpha$ (sometimes the miscoverage rate $1-$coverage instead), the average *width*, and the average *width-ratio* relative to the *Classical* interval (smaller is better). For PPI applications involving real-world datasets, the true θ_i values are not directly observed; following standard practice in the PPI literature [Angelopoulos et al., 2023, Li and Ignatiadis, 2025], we treat the empirical mean over the full labeled corpus as a pseudo-ground truth and use it as the target θ_i. In each Monte Carlo replicate we randomly partition the data into a labeled and an unlabeled subset by masking labels (with the per-application split ratio specified below); we report all metrics averaged over $K = 200$ such replicates, with Monte Carlo standard errors included in the full table.

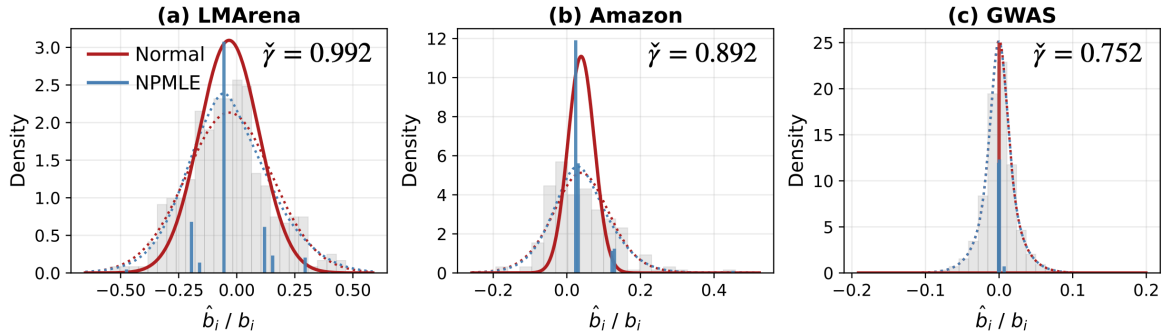

Figure 2: Histogram of the estimator of bias $\hat{b}_i$ with the prior fitted for the bias b_i (solid line) and the average marginal density of $\hat{b}_i$ implied by the prior (dotted line) overlaid. Two choices of prior shown: *Normal* prior and the *NPMLE* prior. The fitted prior distributions are more concentrated than the empirical distribution of $\hat{b}_i$, as the latter is further dispersed due to noise in the bias estimators. The implied marginal densities align well with the observed histograms.

LM Arena. We first consider the problem of estimating $n = 298$ average win rates between pairs of LLMs on the LM Arena platform [Chiang et al., 2024]. For each task i, the task corresponds to a comparison between two LLMs (LLM A versus LLM B), where the covariates X_{ij} consist of the two models' responses to a prompt, and $Y_{ij} \in \{0, 1\}$ denotes the human preference outcome (with $Y_{ij} = 1$ if LLM A is preferred, and 0 otherwise). The target parameter θ_i therefore represents the probability that humans prefer LLM A over LLM B across prompts. Recently, applying PPI to LLM evaluation and ranking has attracted growing interest; see, for example, Chatzi et al. [2024]. We use a 10/90 labeled/unlabeled split ratio for each Monte Carlo replicate, and generate predictions from the raw scores of the `Skywork-reward-v2` reward model [Liu et al., 2026] followed by a Bradley-Terry transformation (see Appendix D.1 for details).

The predictor is bad enough that naïve intervals built from its predictions alone cover the truth only 37% of the time at $\alpha = 0.10$. Despite this, rebiasing produces calibrated intervals that are roughly 23% shorter than *Classical* and 19% shorter than *PT* at nominal 90% coverage (Fig. 3).

At more stringent nominal levels (smaller α), all methods under-cover, the unbiased *Classical* and *PT* included. Two features of the setup plausibly contribute, and do so symmetrically across methods: the per-task labeled samples are small enough that the normal approximation underlying the intervals remains loose, and the pseudo-ground truth θ_i used to assess coverage are themselves estimated rather than directly observed. Within this regime, *RB-Normal* tracks *Classical* and *PT* in coverage while delivering markedly shorter intervals. *RB-NPMLE* sits slightly below in overage, in line with Proposition 3: the lower bound there implies a slow estimation rate as $\check{\gamma} \nearrow 1$, and here $\check{\gamma} = 0.992$. The better performance of *RB-Normal* in the same regime is consistent with the remark following Theorem 2: when $\check{\gamma}$ is close to one, committing to a well-motivated parametric family for G is a defensible alternative to fully nonparametric estimation.

Amazon data. Following the experimental design of Li and Ignatiadis [2025], we consider PPI problems in which the goal is to recover the average customer rating for each of $n = 200$ Amazon products. For product i, the parameter $\theta_i = \mathbb{E}[Y_{ij}]$ is the population mean rating, with $Y_{ij} \in \{1, \ldots, 5\}$ the star score assigned by reviewer j and X_{ij} the concatenation of that review's title and body text. We use a 20/80 labeled/unlabeled split per replicate, mimicking the regime in which expert ratings are scarce relative to the volume of available text [Fan et al., 2024, Mozer and Miratrix, 2025]. The predictor h is fine-tuned BERT [Devlin et al., 2019] (with fine-tuning on a disjoint pool of reviews). We leave additional details of our dataset and predictor to Appendix D.2.

On the Amazon benchmark, the predictor is fairly informative; this is evident in the concentrated estimated bias distributions (Fig. 2 b). Intervals built from the prediction mean alone are about half

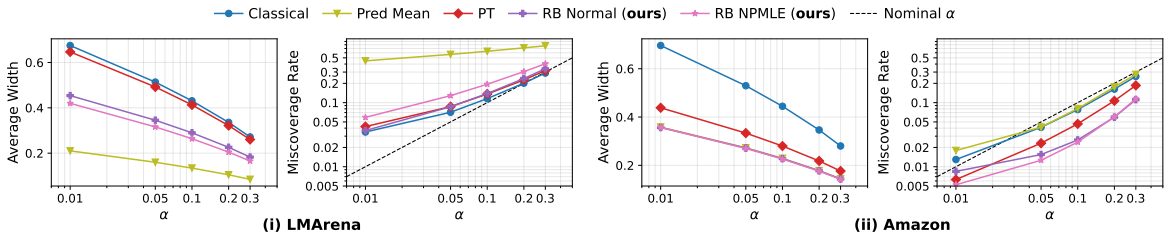

Figure 3: Average width (left) and miscoverage rate (right) for **(i)** $n = 298$ pairwise LLM win-rate estimation problems in LMArena dataset and **(ii)** $n = 200$ product rating estimation problems in the Amazon dataset. *RB-Normal* and *NPMLE* achieve a favorable trade-off, producing shorter intervals while maintaining similar coverage to the unbiased methods. Standard errors are reported in Tables S1, S2.

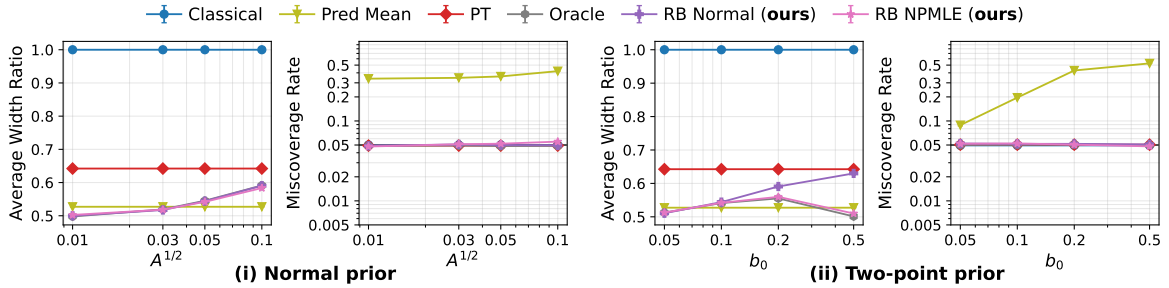

Figure 4: Average width ratio (left) and miscoverage rate (right) for $n = 200$ simulated tasks in synthetic Amazon dataset with two choices of prior on b_i: **(i)** normal prior, and **(ii)** two-point prior. *RB-Normal* and *RB-NPMLE* are nearly indistinguishable under the well-specified normal prior, while *RB-NPMLE* is more robust under the two-point prior, where the *RB-Normal* is misspecified and therefore loses efficiency. In Table S3 and S4 we include the standard errors for the metrics.

the width of the *Classical* interval. Even so, our rebiasing approach further stabilizes coverage and delivers the most efficient calibrated intervals (see Fig. 3). At nominal 90% coverage, *RB-NPMLE* attains 97.6% coverage with width 0.226, roughly 49% shorter than *Classical*, 19% shorter than *PT*.

Synthetic data using covariance structure in Amazon data. So far, the analysis is based on real data, where the true biases b_i and their underlying distribution are not directly observed. To further evaluate our proposed rebiased interval in a setting with access to the true bias distribution G, we conduct a synthetic simulation study based on the Amazon data that preserves the empirical structure of a real prediction-powered inference problem. Specifically, we generate bias b_i from two types of prior G: normal prior $\mathrm{N}(-0.1, A)$ and two-point prior $(\delta_0 + \delta_{b_0})/2$, where δ_u denotes a Dirac point mass at u. See details for the choices of A and b_0, and the construction of the synthetic dataset in Appendix D.3. As a benchmark, we also include an *oracle* variant that plugs in the true data-generating prior for b_i rather than the estimate $\widehat{G}$ when forming rebiased intervals. Results are displayed in Fig. 4. *RB-Normal* and *RB-NPMLE* perform similarly, as well as the oracle interval, when the bias distribution is normal. They maintain close-to-nominal coverage while achieving substantially shorter average widths than the fully debiased *PT* interval. Under this correctly specified setting, the parametric normal model is already sufficient, so the extra flexibility of NPMLE yields no efficiency gain, but, importantly, no efficiency loss either.

Under the two-point prior, the effect of prior misspecification becomes more visible. Although the *RB-Normal* intervals still maintain approximately nominal coverage, their average widths increase as b_0 grows and eventually become close to *PT*. This reflects the efficiency loss from approximating

Table 1: Summary of discoveries from the family-based GWAS summary statistics. We report the numbers of SNPs discovered from BH at FDR 0.05 applied to our *RB-NPMLE* and *RB-Normal* p-values, the (unbiased) direct-effect p-values, and the (biased) population-effect p-values, together with their overlaps (defined in Appendix D.4) with Howe et al. [2022] signals. We also include overlaps for a random set of 5,000 SNPs, averaged over 50 repetitions and reported as mean ± SE.

	RB-NPMLE (ours)	**RB-Normal (ours)**	**Direct-effect (unbiased)**	**Population-effect (biased)**	**Random**
# **SNPs**	1,547	1,700	272	1,594	5,000
# **overlaps**	731	743	234	658	10.84 ± 2.37

a discrete, nonnormal bias distribution by a single normal prior. The *RB-NPMLE* interval is more adaptive in this setting: when b_0 is small, the bias distribution is close to a nearly degenerate case, and both rebiased methods perform well; as the two-point structure becomes more pronounced, the *RB-NPMLE* is better able to capture this structure and therefore keeps the interval width closer to the oracle while preserving coverage.

6.2 Family-based GWAS

We illustrate rebiased p-values for SNP discovery in family-based GWAS, applying Benjamini-Hochberg (BH) [Benjamini and Hochberg, 1995] to control the false discovery rate (FDR). We analyze human height direct-effect estimates $\hat{\theta}_i^{\mathrm{ub}}$ and population-effect estimates $\hat{\theta}_i^{\mathrm{b}}$ for $n = 572{,}912$ SNPs from Guan et al. [2025], estimating the bias prior G via NPMLE and via a normal parametric model. From each estimated prior we compute empirical Bayes rebiasing p-values and apply BH at FDR 0.05; baselines are BH on the p-values of (i) the unbiased direct-effect $\hat{\theta}_i^{\mathrm{ub}}$, and (ii) the biased population-effect $\hat{\theta}_i^{\mathrm{b}}$. We also compare against a random set of 5,000 SNPs as a negative control. Lacking individual-level data, we cannot reuse the PPI evaluation strategy and instead compare discoveries against a separate GWAS with a similar family design [Howe et al., 2022] as preliminary external evidence, so results should be interpreted as exploratory. See Appendix D.4 for details.

A priori, one might trust the population-effect estimator when confounding is believed to be small and otherwise fall back to the much noisier direct-effect estimator. Rebiasing automates this choice: both estimated priors concentrate near zero (Fig. 2c), consistent with existing biological evidence specific to height (Appendix D.4), and the rebiased estimators are similar (but not identical) to the population-effect estimator. Table 1 shows that rebiasing recovers more Howe et al. [2022] signals than the baselines, while the random control shows essentially no overlap. Notably, RB-NPMLE makes slightly fewer total discoveries than the population-effect baseline, yet has more overlaps (see Appendix D.4 for further discussion). The reported $\check{\gamma} = 0.752$ falls in a regime where our theory (§5) gives faster rates, supporting the use of the NPMLE here.[4]

7 Conclusion

For point estimation, forfeiting unbiasedness in exchange for variance savings is an uncontroversial tradeoff. For interval inference, the analyst typically either trusts a biased estimator outright, hoping the bias is negligible, or fully debiases and absorbs the variance cost, with little principled middle ground. This paper proposes empirical Bayes rebiasing as a way to navigate the analogous tradeoff for calibrated inference, by modeling task-level bias as drawn from a distribution learned across many related tasks rather than treated as either zero or arbitrary. Our construction is deliberately lightweight: only the biases are modeled as exchangeable; the θ_i are held fixed. The analyst therefore

[4]Since nearby SNPs are correlated, the independence assumption of §5 does not strictly hold here. We nevertheless expect our estimated p-values to remain close to their oracle counterparts due to weak dependence across the genome.

need not commit to any pooling structure on the θ_i. Two modeling assumptions are worth flagging as limitations. First, exchangeability of the biases, in particular, that their distribution does not depend on the θ_i, is common in the literature (§2) but may fail in practice [Schuemie et al., 2018]. Second, motivated by the central limit theorem, we model the estimators as exactly normal with known second moments; our theory does not propagate the error from the CLT approximation or the uncertainty from plug-in estimates of σ_i, τ_i, ρ_i in (3).

Our framework applies wherever a biased estimator is paired with a bias correction (equivalently, an unbiased estimator) across many parallel tasks; PPI and family-based GWAS as in this paper, or combining observational with experimental causal estimates, and so forth. Calibrated inference in empirical Bayes problems is much less developed than the corresponding point estimation theory despite its practical importance, see e.g., Armstrong et al. [2022] for recent progress; the empirical partially Bayes formulation pursued here seems a promising avenue in this regard.

Acknowledgments Part of the computing for this project was conducted on UChicago's Data Science Institute cluster. N.I. gratefully acknowledges support from NSF (DMS 2443410).

References

1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

A. N. Angelopoulos, J. C. Duchi, and T. Zrnic. PPI++: Efficient prediction-powered inference. *arXiv preprint*, arXiv:2311.01453, 2024.

T. B. Armstrong and M. Kolesár. Optimal inference in a class of regression models. *Econometrica*, 86(2):655–683, 2018.

T. B. Armstrong and M. Kolesár. Sensitivity analysis using approximate moment condition models. *Quantitative Economics*, 12(1):77–108, 2021.

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

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

D. R. Bernard, G. Bryan, S. Chabé-Ferret, J. De Quidt, J. C. Fliegner, and R. Rathelot. How much should we trust observational estimates? Accumulating evidence using RCTs with imperfect compliance. CEPR Discussion Paper DP18794, Centre for Economic Policy Research, 2024.

M. B. Brown. *A Secondarily Bayes Approach to the Two-Means Problem.* PhD thesis, Princeton University, 1965.

C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, and J. Marchini. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, Oct. 2018.

T. T. Cai and M. G. Low. An adaptation theory for nonparametric confidence intervals. *The Annals of Statistics*, 32(5):1805–1840, 2004.

C. C. Chang, C. C. Chow, L. C. A. M. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4:7, 2015.

I. Chatzi, E. Straitouri, S. Thejaswi, and M. G. Rodriguez. Prediction-powered ranking of large language models. In *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*, 2024.

A. Chen, A. B. Owen, and M. Shi. Data enriched linear regression. *Electronic Journal of Statistics*, 9(1):1078–1112, 2015.

W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. I. Jordan, J. E. Gonzalez, et al. Chatbot arena: an open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, pages 8359–8388, 2024.

D. R. Cox. A note on partially Bayes inference and the linear model. *Biometrika*, 62(3):651–654, 1975.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186, 2019.

L. H. Dicker and S. D. Zhao. High-dimensional classification via nonparametric empirical Bayes and maximum likelihood inference. *Biometrika*, 103(1):21–34, 2016.

D. L. Donoho. Statistical estimation and optimal recovery. *The Annals of Statistics*, pages 238–270, 1994.

B. Efron. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004.

B. Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Institute of Mathematical Statistics Monographs. Cambridge University Press, Cambridge, 2010.

B. Efron. Discussion of "Confidence intervals for nonparametric empirical Bayes analysis" by Nikolaos Ignatiadis and Stefan Wager. *Journal of the American Statistical Association*, 117(539):1179–1180, 2022.

B. Elsworth, M. Lyon, T. Alexander, Y. Liu, P. Matthews, J. Hallett, P. Bates, T. Palmer, V. Haberland, G. D. Smith, J. Zheng, P. Haycock, T. R. Gaunt, and G. Hemani. The MRC IEU OpenGWAS data infrastructure. *bioRxiv*, 2020.

S. Fan, A. Visokay, K. Hoffman, S. Salerno, L. Liu, J. T. Leek, and T. McCormick. From narratives to numbers: Valid inference using language model predictions from verbal autopsies. In *First Conference on Language Modeling*, 2024.

A. Gelman and M. Vákár. Slamming the sham: A Bayesian model for adaptive adjustment with noisy control data. *Statistics in Medicine*, 40(15):3403–3424, 2021.

E. J. Green and W. E. Strawderman. A James-Stein type estimator for combining unbiased and possibly biased estimators. *Journal of the American Statistical Association*, 86(416):1001–1006, 1991.

E. J. Green, W. E. Strawderman, R. L. Amateis, and G. A. Reams. Improved estimation for multiple means with heterogeneous variances. *Forest Science*, 51(1):1–6, 2005.

J. Guan, T. Tan, S. M. Nehzati, M. Bennett, P. Turley, D. J. Benjamin, and A. S. Young. Family-based genome-wide association study designs for increased power and robustness. *Nature Genetics*, 57(4):1044–1052, 2025.

G. Hemani, B. Elsworth, T. Palmer, and R. Rasteiro. *ieugwasr: Interface to the 'OpenGWAS' Database API*, 2025. URL https://github.com/MRCIEU/ieugwasr. R package version 1.1.0.

L. J. Howe, M. G. Nivard, T. T. Morris, A. F. Hansen, H. Rasheed, Y. Cho, G. Chittoor, R. Ahlskog, P. A. Lind, T. Palviainen, M. D. van der Zee, R. Cheesman, M. Mangino, Y. Wang, S. Li, L. Klaric, S. M. Ratliff, L. F. Bielak, M. Nygaard, A. Giannelis, E. A. Willoughby, C. A. Reynolds, J. V. Balbona, O. A. Andreassen, H. Ask, A. Baras, C. R. Bauer, D. I. Boomsma, A. Campbell, H. Campbell, Z. Chen, P. Christofidou, E. Corfield, C. C. Dahm, D. R. Dokuru, L. M. Evans, E. J. C. de Geus, S. Giddaluru, S. D. Gordon, K. P. Harden, W. D. Hill, A. Hughes, S. M. Kerr, Y. Kim, H. Kweon, A. Latvala, D. A. Lawlor, L. Li, K. Lin, P. Magnus, P. K. E. Magnusson, T. T. Mallard, P. Martikainen, M. C. Mills, P. R. Njølstad, J. D. Overton, N. L. Pedersen, D. J. Porteous, J. Reid, K. Silventoinen, M. C. Southey, C. Stoltenberg, E. M. Tucker-Drob, M. J. Wright, Social Science Genetic Association Consortium, Within Family Consortium, J. K. Hewitt, M. C. Keller, M. C. Stallings, J. J. Lee, K. Christensen, S. L. R. Kardia, P. A. Peyser, J. A. Smith, J. F. Wilson, J. L. Hopper, S. Hägg, T. D. Spector, J.-B. Pingault, R. Plomin, A. Havdahl, M. Bartels, N. G. Martin, S. Oskarsson, A. E. Justice, I. Y. Millwood, K. Hveem, Ø. Naess, C. J. Willer, B. O. Åsvold, P. D. Koellinger, J. Kaprio, S. E. Medland, R. G. Walters, D. J. Benjamin, P. Turley, D. M. Evans, G. Davey Smith, C. Hayward, B. Brumpton, G. Hemani, and N. M. Davies. Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. *Nature Genetics*, 54 (5):581–592, 2022.

G. Hripcsak, J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, M. A. Suchard, R. W. Park, I. C. K. Wong, P. R. Rijnbeek, J. van der Lei, N. Pratt, N. Norén, Y.-C. Li, P. E. Stang, D. Madigan, and P. B. Ryan. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for observational researchers. In *Studies in Health Technology and Informatics*. IOS Press, 2015.

N. Ignatiadis and B. Sen. Empirical partially Bayes multiple testing and compound χ^2 decisions. *The Annals of Statistics*, 53(1):1–36, 2025.

N. Ignatiadis and S. Wager. Covariate-powered empirical Bayes estimation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 361–379, 1961.

W. Jiang. Comment: Empirical Bayes interval estimation. *Statistical Science*, 34(2):219–223, 2019.

W. Jiang. On general maximum likelihood empirical bayes estimation of heteroscedastic iid normal means. *Electronic Journal of Statistics*, 14(1):2272–2297, 2020.

W. Jiang and C.-H. Zhang. General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647–1684, 2009.

J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 27(4):887 – 906, 1956.

D. M. Kluger, K. Lu, T. Zrnic, S. Wang, and S. Bates. Prediction-powered inference with imputed covariates and nonuniform sampling. *arXiv preprint*, arXiv:2501.18577, 2025.

R. Koenker. Empirical Bayes confidence intervals: An R vinaigrette. Technical report, 2020. URL http://www.econ.uiuc.edu/~roger/research/vinaigrettes/cieb.pdf.

R. Koenker and J. Gu. REBayes: An R package for empirical Bayes mixture methods. *Journal of Statistical Software*, 82(8), 2017.

R. Koenker and I. Mizera. Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *Journal of the American Statistical Association*, 109(506):674–685, 2014.

A. Kong, G. Thorleifsson, M. L. Frigge, B. J. Vilhjalmsson, A. I. Young, T. E. Thorgeirsson, S. Benonisdottir, A. Oddsson, B. V. Halldorsson, G. Masson, D. F. Gudbjartsson, A. Helgason, G. Bjornsdottir, U. Thorsteinsdottir, and K. Stefansson. The nature of nurture: Effects of parental genotypes. *Science*, 359(6374):424–428, Jan. 2018.

S. Kwon and J. Roth. (Empirical) Bayes approaches to parallel trends. *AEA Papers and Proceedings*, 114:606–609, 2024.

S. Li and N. Ignatiadis. Prediction-powered adaptive shrinkage estimation. In *Forty-Second International Conference on Machine Learning*, 2025.

Z. Lin, P. J. Bickel, and P. Ding. Introducing the b-value: Combining unbiased and biased estimators from a sensitivity analysis perspective. *arXiv preprint*, arXiv:2602.16310, 2026.

B. Lindsay. *Mixture Models: Theory, Geometry, and Applications*. Conference Board of the Mathematical Sciences: NSF-CBMS Regional Conference Series in Probability and Statistics. Institute of Mathematical Statistics, 1995.

C. Y. Liu, L. Zeng, Y. Xiao, J. He, J. Liu, C. Wang, R. Yan, W. Shen, F. Zhang, J. Xu, and Y. Liu. Skywork-reward-V2: Scaling preference data curation via human-AI synergy. In *The Fourteenth International Conference on Learning Representations*, 2026.

C. N. Morris. Parametric empirical Bayes confidence intervals. In G. Box, T. Leonard, and C.-F. Wu, editors, *Scientific Inference, Data Analysis, and Robustness*, pages 25–50. Academic Press, 1983.

MOSEK ApS. *The MOSEK Optimization Suite Manual, Version 11.0*, 2024.

R. Mozer and L. Miratrix. More power to you: Using machine learning to augment human coding for more efficient inference in text-based randomized trials. *The Annals of Applied Statistics*, 19(1): 440–464, 2025.

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

A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38:904–909, 2006.

S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. Plink: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3):559–575, 2007.

H. Robbins. A generalization of the method of maximum likelihood: Estimating a mixing distribution (abstract). *The Annals of Mathematical Statistics*, 21:314–315, 1950.

H. Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 157–163. The Regents of the University of California, 1956.

E. T. Rosenman, G. Basse, A. B. Owen, and M. Baiocchi. Combining observational and experimental datasets using shrinkage estimators. *Biometrics*, page biom.13827, 2023a.

E. T. R. Rosenman, F. Dominici, and L. Miratrix. Empirical Bayes double shrinkage for combining biased and unbiased causal estimates. *arXiv preprint*, arXiv:2309.06727, 2023b.

D. B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172, 1984.

M. J. Schuemie, P. B. Ryan, W. DuMouchel, M. A. Suchard, and D. Madigan. Interpreting observational studies: Why empirical calibration is needed to correct p-values. *Statistics in Medicine*, 33 (2):209–218, 2014.

M. J. Schuemie, G. Hripcsak, P. B. Ryan, D. Madigan, and M. A. Suchard. Robust empirical calibration of p-values using observational data. *Statistics in Medicine*, 35(22):3883–3888, 2016.

M. J. Schuemie, G. Hripcsak, P. B. Ryan, D. Madigan, and M. A. Suchard. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proceedings of the National Academy of Sciences*, 115(11):2571–2577, 2018.

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

M. Stephens. False discovery rates: A new deal. *Biostatistics*, 18(2):275–294, 2017.

B. Wu, S. Salazar, D. P. Green, and D. M. Blei. The Illusion of learning from observational data: An empirical Bayes perspective. *arXiv preprint*, arXiv:2604.08853, 2026.

L. Yengo, M. R. Robinson, M. C. Keller, K. E. Kemper, Y. Yang, M. Trzaskowski, J. Gratten, P. Turley, D. Cesarini, D. J. Benjamin, N. R. Wray, M. E. Goddard, J. Yang, and P. M. Visscher. Imprint of assortative mating on the human genome. *Nature Human Behaviour*, 2(12):948–954, 2018.

A. I. Young, S. M. Nehzati, S. Benonisdottir, A. Okbay, H. Jayashankar, C. Lee, D. Cesarini, D. J. Benjamin, P. Turley, and A. Kong. Mendelian imputation of parental genotypes improves estimates of direct genetic effects. *Nature Genetics*, 54(6):897–905, 2022.

Q. Zhao, J. Wang, G. Hemani, J. Bowden, and D. S. Small. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *The Annals of Statistics*, 48(3):1742–1769, 2020.

♣ Appendix: Table of Contents

A Details on methods (Section 3)

A.1 Normal bias distribution when $\rho = 0$

Consider the parametric case where we are willing to assume that the bias distribution is normal,

$$b_i \sim \mathrm{N}(\mu, A).$$

Combining with model (3), Bayes' rule gives

$$b_i \mid \hat{b}_i \sim \mathrm{N}\left(\mu + \frac{A}{A+\tau_i^2}(\hat{b}_i - \mu), \frac{A\tau_i^2}{A+\tau_i^2}\right),$$

and the oracle rebiased estimator takes the form

$$\hat{\theta}^{\mathrm{rb}} = \hat{\theta}_i^{\mathrm{b}} - \left\{\mu + \frac{A}{A+\tau_i^2}(\hat{b}_i - \mu)\right\} = \hat{\theta}_i^{\mathrm{db}} + \frac{\tau_i^2}{A+\tau_i^2}(\hat{b}_i - \mu).$$

This expression shows explicitly how the normal model interpolates between the biased and fully debiased estimators. When $A = 0$, the correction is shrunk completely toward μ; in particular, if $\mu = 0$, $\hat{\theta}^{\mathrm{rb}} = \hat{\theta}_i^{\mathrm{b}}$, while for $A = \infty$, $\hat{\theta}^{\mathrm{rb}} = \hat{\theta}_i^{\mathrm{db}}$.

Moreover, the oracle rebiased $(1-\alpha)$-interval reduces to the symmetric interval

$$\mathcal{I}_{G,i}^{\mathrm{rb}}(1-\alpha) = \hat{\theta}_i^{\mathrm{rb}} \pm z_{1-\alpha/2}\left(\sigma_i^2 + \frac{A\tau_i^2}{A+\tau_i^2}\right)^{1/2}. \tag{S1}$$

The unknown parameters (μ, A) from the prior G can be estimated via maximum marginal likelihood (see A.3). With $\widehat{G} = \mathrm{N}(\widehat{\mu}, \widehat{A})$ in hand, we can then plug in (S1) and obtain the rebiased $(1-\alpha)$-intervals $\mathcal{I}_{\widehat{G},i}^{\mathrm{rb}}(1-\alpha)$.

Similarly, we have the following form of the oracle rebiased p-value $P_i^{\mathrm{rb}} = P_{G,\theta_{i0}}^{(i)}(\hat{\theta}_i^{\mathrm{db}}, \hat{b}_i)$, with

$$P_{G,\theta_{i0}}^{(i)}(z,l) = 2\Phi\left(-\left|z + \frac{\tau_i^2}{A+\tau_i^2}(l-\mu) - \theta_{i0}\right| \Big/ \left(\sigma_i^2 + \frac{A\tau_i^2}{A+\tau_i^2}\right)^{1/2}\right), \tag{S2}$$

where Φ denotes the cumulative distribution function of a standard normal random variable. Plugging the estimated $\widehat{G}$, we can obtain the empirical Bayes rebiased p-value $\widehat{P}_i^{\mathrm{rb}} = P_{\widehat{G},\theta_{i0}}^{(i)}(\hat{\theta}_i^{\mathrm{db}}, \hat{b}_i)$ by replacing (μ, A) with $(\widehat{\mu}, \widehat{A})$ in (S2).

A.2 Generalization of ρ

Consider a more generic setting of this problem. We no longer assume independence between $(\hat{\theta}_i^{\mathrm{b}}, \hat{b}_i)$ ($\rho_i = 0$ in (3)). We can consider the following conditional distribution of $\hat{\theta}_i^{\mathrm{db}}$ given $\hat{b}_i$ that integrates out b_i over its posterior $\Pi_i(\cdot \mid \hat{b}_i)$ under the prior G,

$$\hat{\theta}_i^{\mathrm{db}} \mid \theta_i, \hat{b}_i \sim \int \mathrm{N}\left(\theta_i + c_i(\hat{b}_i - b),\ \sigma_i^2(1-\rho_i^2)\right)\Pi_i(\mathrm{d}b \mid \hat{b}_i), \tag{S3}$$

with $c_i = \rho_i\sigma_i/\tau_i - 1$. Writing $m_i \equiv m_i(\hat{b}_i) = \mathbb{E}_G[b_i \mid \hat{b}_i]$, the conditional mean of $\hat{\theta}_i^{\mathrm{db}}$ is $\mathbb{E}[\hat{\theta}_i^{\mathrm{db}} \mid \theta_i, \hat{b}_i] = \theta_i + c_i(\hat{b}_i - m_i)$, and we construct

$$\hat{\theta}_i^{\mathrm{rb}} = \hat{\theta}_i^{\mathrm{db}} - c_i(\hat{b}_i - m_i).$$

Let $q_{G,i,\alpha}(\hat{b}_i)$ denote the α-quantile of the conditional distribution of $\hat{\theta}_i^{\mathrm{db}} - \theta_i$ given $\hat{b}_i$, and we could report the $(1-\alpha)$-level oracle rebiased interval $\mathcal{I}_{G,i}^{\mathrm{rb}}(1-\alpha) = [\hat{\theta}_i^{\mathrm{rb}} - q_{G,i,1-\alpha/2},\ \hat{\theta}_i^{\mathrm{rb}} - q_{G,i,\alpha/2}]$.

In particular, if we assume $G = \mathrm{N}(\mu, A)$, and then (S3) can be written as

$$\hat{\theta}_i^{\mathrm{db}} \mid \theta_i, \hat{b}_i \sim \mathrm{N}\left(\theta_i + \frac{\rho_i\sigma_i\tau_i - \tau_i^2}{\tau_i^2 + A}(\hat{b}_i - \mu),\ \sigma_i^2 + \tau_i^2 - 2\rho_i\sigma_i\tau_i - \frac{(\rho_i\sigma_i\tau_i - \tau_i^2)^2}{\tau_i^2 + A}\right).$$

Obtain the estimated prior parameters $(\widehat{\mu}, \widehat{A})$ via maximum marginal likelihood, we can plug back and form the following rebiased $(1-\alpha)$-intervals for each i:

$$\hat{\theta}_i^{\mathrm{rb}} \pm z_\alpha \left[\sigma_i^2 + \tau_i^2 - 2\rho_i\sigma_i\tau_i - \frac{\left(\rho_i\sigma_i\tau_i - \tau_i^2\right)^2}{\tau_i^2 + \widehat{A}}\right]^{1/2}, \quad \hat{\theta}_i^{\mathrm{rb}} := \hat{\theta}_i^{\mathrm{db}} - \frac{\rho_i\sigma_i\tau_i - \tau_i^2}{\tau_i^2 + \widehat{A}}(\hat{b}_i - \widehat{\mu}).$$

A.3 Estimation for G

Parametric Normal Assume $G = \mathrm{N}(\mu, A)$, and we use the maximum marginal likelihood to estimate (μ, A). In particular, combining with the model (3), the marginal distribution of $\hat{b}_i$ is

$$\hat{b}_i \sim \mathrm{N}(\mu, A + \tau_i^2).$$

Therefore, we estimate (μ, A) by maximizing the marginal likelihood:

$$(\widehat{\mu}, \widehat{A}) = \arg\max_{\mu\in\mathbb{R},\, A>0} \sum_{i=1}^{n} \log\varphi\left(\hat{b}_i;\ \mu,\ A + \tau_i^2\right).$$

For implementation, we optimize over $(\mu, \log A)$ to enforce the constraint $A > 0$.

NPMLE As observed before, the optimization program in (9) is convex. We use the discretization technique proposed in Koenker and Mizera [2014] to solve this problem. In particular, we choose $B = 50$ grid points (For GWAS we choose $B = 300$) that are equally spaced between the smallest and largest value of $\{\hat{b}_1, ..., \hat{b}_n\}$, and we optimize (9) over all possible distributions supported on this finite grid, which is a conic programing problem and we solve it with the interior point convex programming solver MOSEK MOSEK ApS [2024].

B Details on applications (Section 4)

B.1 Prediction powered inference

Here we expand on our description of the PPI/PT estimators in §4, using the per-task quantities and the vanilla PPI estimator already introduced in (11). Throughout this subsection, we assume that the labeled samples $\{(X_{ij}, Y_{ij})\}_{j=1}^{m_i}$ and the unlabeled samples $\{\tilde{X}_{ij}\}_{j=1}^{M_i}$ are jointly independent both across the labeled/unlabeled split and across tasks $i = 1, \ldots, n$. The ML predictor h is considered fixed and independent of all of them. To compress notation, write the second moments for the samples in the i-th task as

$$v_i^2 := \mathrm{Var}[h(X_{ij})], \qquad w_i^2 := \mathrm{Var}[Y_{ij}], \qquad c_i := \mathrm{Cov}\left[h(X_{ij}),\, Y_{ij}\right], \tag{S4}$$

and we denote $\mu_i := \mathbb{E}[h(X_{ij})]$ so the true bias is $b_i = \mu_i - \theta_i$.

Power-tuning derivation. Angelopoulos et al. [2024] introduce a one-parameter family of estimators with a tuning parameter $\lambda_i \in \mathbb{R}$,

$$\hat{\theta}_{i,\lambda_i} = \bar{Y}_i + \lambda_i\big(\tilde{Z}_i^h - \bar{Z}_i^h\big), \qquad \lambda_i \in \mathbb{R}, \tag{S5}$$

which interpolates between the classical estimator $\bar{Y}_i$ ($\lambda_i = 0$) and $\hat{\theta}_i^{\mathrm{PPI}}$ ($\lambda_i = 1$). For every choice of λ_i, $\hat{\theta}_{i,\lambda_i}$ is unbiased since $\mathbb{E}[\tilde{Z}_i^h] = \mathbb{E}[\bar{Z}_i^h] = \mu_i$. A direct computation gives

$$\mathrm{Var}[\hat{\theta}_{i,\lambda_i}] = \frac{w_i^2}{m_i} + \lambda_i^2\, v_i^2\left(\frac{1}{M_i} + \frac{1}{m_i}\right) - \frac{2\lambda_i\, c_i}{m_i}. \tag{S6}$$

Minimizing this variance with respect to λ_i yields the optimal power tuning parameter

$$\lambda_i^* = \frac{M_i}{m_i + M_i}\,\frac{c_i}{v_i^2}, \tag{S7}$$

and the corresponding power-tuned (PT) estimator is defined as $\hat{\theta}_i^{\mathrm{PT}} := \hat{\theta}_{i,\lambda_i^*}$.

PT in the rebiasing framework. The PT estimator is unbiased for θ_i, but its variance (S6) at λ_i^* is generally larger than v_i^2/M_i, the variance of $\hat{\theta}_i^{\mathrm{ML}}$. Following §4, we therefore identify

$$\hat{\theta}_i^{\mathrm{b}} := \hat{\theta}_i^{\mathrm{ML}} = \tilde{Z}_i^h, \qquad \hat{b}_i := \hat{\theta}_i^{\mathrm{b}} - \hat{\theta}_i^{\mathrm{PT}} = (1-\lambda_i^*)\,\tilde{Z}_i^h + \lambda_i^*\,\bar{Z}_i^h - \bar{Y}_i. \tag{S8}$$

Then $\mathbb{E}[\hat{b}_i] = (1-\lambda_i^*)\mu_i + \lambda_i^*\mu_i - \theta_i = b_i$, so $\hat{b}_i$ is unbiased for the bias of $\hat{\theta}_i^{\mathrm{b}}$. The pair $(\hat{\theta}_i^{\mathrm{b}}, \hat{b}_i)$ thus fits the model in (3) with $\hat{\theta}_i^{\mathrm{db}} = \hat{\theta}_i^{\mathrm{b}} - \hat{b}_i = \hat{\theta}_i^{\mathrm{PT}}$, and the rebiased estimator and intervals from §3 apply directly. Setting $\lambda_i^* = 1$ recovers the standard vanilla-PPI debiasing ($\hat{b}_i = \tilde{Z}_i^h - \bar{Y}_i$); the rebiasing perspective therefore subsumes both PPI and PT under a single framework.

Expressions for $\sigma_i^2, \tau_i^2, \rho_i$. A short computation gives

$$\sigma_i^2 = \mathrm{Var}[\hat{\theta}_i^{\mathrm{b}}] = \frac{v_i^2}{M_i}, \tag{S9}$$

$$\tau_i^2 = \mathrm{Var}[\hat{b}_i] = (1-\lambda_i^*)^2\,\frac{v_i^2}{M_i} + \frac{(\lambda_i^*)^2\, v_i^2 - 2\lambda_i^*\, c_i + w_i^2}{m_i}, \tag{S10}$$

$$\rho_i\sigma_i\tau_i = \mathrm{Cov}\left[\hat{\theta}_i^{\mathrm{b}}, \hat{b}_i\right] = (1-\lambda_i^*)\,\frac{v_i^2}{M_i}, \qquad \rho_i = \frac{(1-\lambda_i^*)\, v_i/\sqrt{M_i}}{\tau_i}. \tag{S11}$$

Two limits are instructive. When $\lambda_i^* = 0$, $\hat{\theta}_i^{\mathrm{PT}}$ collapses to the classical mean $\bar{Y}_i$, and ρ_i is maximal because $\hat{\theta}_i^{\mathrm{b}}$ enters $\hat{b}_i$ undampened. When $\lambda_i^* = 1$, $\hat{\theta}_i^{\mathrm{PT}}$ coincides with vanilla PPI and $\hat{b}_i = \bar{Z}_i^h - \bar{Y}_i$; the unlabeled-only $\hat{\theta}_i^{\mathrm{b}}$ and the labeled-only $\hat{b}_i$ are then independent, giving $\rho_i = 0$ and matching the simplified development in §3.

Plug-in estimation. The expressions above involve (v_i^2, w_i^2, c_i), which are not known in practice. We replace each by its sample analog from the labeled set (the sample variance of Y_{ij} and $h(X_{ij})$, as well as the sample covariance of $(h(X_{ij}), Y_{ij})$); when $M_i \gg m_i$, v_i^2 may also be estimated from the larger unlabeled set $\{h(\tilde{X}_{ij})\}_{j=1}^{M_i}$. Plugging these estimates into (S7) yields a feasible $\hat{\lambda}_i^*$, and into (S9)–(S11) yields $(\hat{\sigma}_i^2, \hat{\tau}_i^2, \hat{\rho}_i)$. Following standard practice in PPI [Angelopoulos et al., 2024], we treat these plug-ins as known when forming Wald and rebiased intervals.

B.2 Family-based GWAS

The summary statistics we analyzed provide the pair $(\hat{\theta}_i^{\mathrm{ub}}, \hat{b}_i)$, and their respective variances $\tilde{\sigma}_i^2$ and τ_i^2, and correlation γ_i are also reported. The representation above is equivalent to the pair $(\hat{\theta}_i^{\mathrm{b}}, \hat{\theta}_i^{\mathrm{ub}})$: indeed, the biased estimator can be written as $\hat{\theta}_i^{\mathrm{b}} = \hat{\theta}_i^{\mathrm{ub}} + \hat{b}_i$, with variance $\sigma_i^2 = \tilde{\sigma}_i^2 + \tau_i^2 + 2\gamma_i\tilde{\sigma}_i\tau_i$, and its correlation with $\hat{b}_i$ is $\rho_i = (\tau_i + \gamma_i\tilde{\sigma}_i)/(\tilde{\sigma}_i^2 + \tau_i^2 + 2\gamma_i\tilde{\sigma}_i\tau_i)^{1/2}$.

C Proof of theoretical results (Section 5)

For notation simplicity, we denote

$$\Sigma_i = (\Sigma_{i,jq}) := \begin{pmatrix} \tilde{\sigma}_i^2 & \gamma_i \tilde{\sigma}_i \tau_i \\ \gamma_i \tilde{\sigma}_i \tau_i & \tau_i^2 \end{pmatrix}, \quad \Omega_i = (\Omega_{i,jq}) := \Sigma_i^{-1}.$$

Start by defining the class of marginal densities of $\hat{b}_i$ across i,

$$\mathcal{F}_n := \left\{ \left(f_{G'}^{(1)}, ..., f_{G'}^{(n)} \right) : G' \in \mathcal{G} \right\}, \tag{S12}$$

where $\mathcal{G}$ is the set of all possible priors G', and

$$f_{G'}^{(i)}(l) \equiv f_{G'}(l; \Sigma_{i,22}) := \int \varphi(l - b; \Sigma_{i,22}) \, \mathrm{d}G'(b).$$

Define the supremum norm in bounded intervals

$$\|\boldsymbol{h}\|_{\infty,U} = \max_{1 \le i \le n} \|h_i\|_{\infty,U} = \max_{1 \le i \le n} \sup_{|x| \le U} |h_i(x)|,$$

where $\boldsymbol{h} = (h_1, ..., h_n)$, and also define the average Hellinger distance

$$\bar{d}(G, \widehat{G}) = \left(\frac{1}{n} \sum_{i=1}^{n} \mathfrak{H}^2 \left(f_G^{(i)}, f_{\widehat{G}}^{(i)} \right) \right)^{\frac{1}{2}}, \quad \text{where } \mathfrak{H}^2(g,h) = \frac{1}{2} \int \left(\sqrt{g(t)} - \sqrt{h(t)} \right)^2 \mathrm{d}t.$$

C.1 Preliminary lemmata and propositions

We state the following results from Jiang [2020].

Lemma 5 (Lemma 4 of Jiang [2020])**.** For $0 < \varepsilon < 1/\sqrt{2\pi}$, $U > 0$,

$$\log \left(\mathcal{N}(\varepsilon, \mathcal{F}_n, \|\cdot\|_{\infty,U}) \right) \lesssim_{\underline{\sigma}, \bar{\sigma}} \left(\log \left(\frac{1}{\varepsilon} \right) \right)^2 \max \left\{ \frac{U}{\{\sqrt{|\log \epsilon|}}, 1 \right\}.$$

Lemma 6 (Theorem 4 of Jiang [2020])**.** Under Assumption 1, fix $c_0 \ge 2$, then there exists constants C_0 depending only on $\Gamma, \underline{\sigma}, \bar{\sigma}$ such that

$$\mathbb{P}_G \left[\bar{d}(G, \widehat{G}) \ge C_0 \frac{\log n}{\sqrt{n}} \right] \le \exp \left(-c_0 \log n \right) \quad \text{for all } n \in \mathbb{N}_{\ge 2}.$$

Proof of Lemma 6. We verify the rate obtained from Theorem 4 of Jiang [2020]. In the notation of that theorem, the Hellinger rate is of the form

$$\varepsilon_n(n, G, p) = \max \left\{ \sqrt{2 \log n}, \left[n^{1/p} \sqrt{\log n} \cdot \mu_p(G) \right]^{p/(2p+2)} \right\} \sqrt{\frac{\log n}{n}}, \tag{S13}$$

where G is the true bias distribution, and $\mu_p(G) := \left(\int |b|^p \, \mathrm{d}G(b) \right)^{1/p}$. Since $G \in \mathcal{G}_\Gamma$, the standard sub-Gaussian moment bound gives $\mu_p(G) \le c\Gamma\sqrt{p}$ for all $p \ge 1$, where $c > 0$ is a universal constant.

We set $p = \log n$ in (S13). For $A_n := \left[n^{1/p} \sqrt{\log n} \cdot \mu_p(G) \right]^{p/(2p+2)}$, we have

$$n^{1/p} \sqrt{\log n} \cdot \mu_p(G) \le e \cdot \sqrt{\log n} \cdot c\Gamma \sqrt{\log n} = ec\Gamma \, \log n, \qquad \frac{p}{2p+2} = \frac{\log n}{2 \log n + 2} \le \frac{1}{2}.$$

Then $A_n \le (ec\Gamma \log n)^{1/2} = \sqrt{ec\Gamma} \cdot \sqrt{\log n}$. Plugging back to (S13),

$$\varepsilon_n(n, G, \log n) = \max\Big\{\sqrt{2\log n},\, A_n\Big\}\sqrt{\frac{\log n}{n}} \le C_1(\Gamma)\frac{\log n}{\sqrt{n}} \quad \text{with} \quad C_1(\Gamma) = \max\{\sqrt{2}, \sqrt{ec\Gamma}\}.$$

Therefore, with Theorem 4 of Jiang [2020] and a fixed $c_0 \ge 2$, there exists C_0 depending only on $\Gamma, \underline{\sigma}, \bar{\sigma}$ such that

$$\mathbb{P}_G\left[\bar{d}(G, \widehat{G}) \ge C_0\frac{\log n}{\sqrt{n}}\right] \le \exp\left(-c_0 \log n\right) \quad \text{for all } n \in \mathbb{N}_{\ge 2}.$$

□

In the following proposition, we consider to express $F_{G,i}(z \mid l)$ in terms of $f_G^{(i)}$.

Proposition 7. It holds that

$$F_{G,i}(z \mid l) = \frac{C(\Sigma_i)}{f_G^{(i)}(l)}\int_{-\infty}^{z} \exp\left(-\frac{u^2}{2\Sigma_{i,11}}\right) f_G\left(\frac{\Omega_{i,12}}{\Omega_{i,22}}u + l; \Omega_{i,22}^{-1}\right) \mathrm{d}u,$$

where $C(\Sigma_i) = (2\pi|\Sigma_i|\Omega_{i,22})^{-1/2}$.

Proof. We start by rewriting the distribution of $(\hat{\theta}_i^{\mathrm{db}} - \theta_i, \hat{b}_i) = (z, l)$ given b_i,

$$\begin{aligned}
p_i(z, l \mid b_i) &= (2\pi)^{-1}|\Sigma_i|^{-1/2}\exp\left(-\frac{1}{2}\left[\Omega_{i,11}z^2 + 2\Omega_{i,12}z(l - b_i) + \Omega_{i,22}(l - b_i)^2\right]\right) \\
&= C'\exp\left(-\frac{1}{2}\left[\Omega_{i,22}b_i^2 - 2(\Omega_{i,12}z + \Omega_{i,22}l)b_i\right] - \frac{1}{2}\left[\Omega_{i,11}z^2 + 2\Omega_{i,12}zl + \Omega_{i,22}l^2\right]\right) \\
&= C'\exp\left(-\frac{\Omega_{i,22}}{2}\left(b_i - \frac{\Omega_{i,12}z + \Omega_{22}l}{\Omega_{i,22}}\right)^2\right)\exp\left(-\frac{1}{2}\left(\Omega_{i,11} - \frac{\Omega_{i,12}^2}{\Omega_{i,22}}\right)z^2\right) \\
&= C(\Sigma_i)\varphi\left(\frac{\Omega_{i,12}z + \Omega_{i,22}l}{\Omega_{i,22}} - b_i; \Omega_{i,22}^{-1}\right)\exp\left(-\frac{1}{2}\left(\Omega_{i,11} - \frac{\Omega_{i,12}^2}{\Omega_{i,22}}\right)z^2\right),
\end{aligned}$$

with $C(\Sigma_i) := (2\pi|\Sigma_i|\Omega_{i,22})^{-1/2}$. Denote $p_G(u, l) = \int p_i(u, l \mid b_i)\, dG(b_i)$ as the joint marginal density of $(\hat{\theta}_i^{\mathrm{db}} - \theta_i, \hat{b}_i)$ under G, then we can express $F_{G,i}(z \mid l)$ as

$$\begin{aligned}
F_{G,i}(z \mid l) &= \mathbb{P}_G(\hat{\theta}_i^{\mathrm{db}} - \theta_i \le z \mid \hat{b}_i = l) \\
&= \int_{-\infty}^{z}\frac{p_G(u, l)}{f_G^{(i)}(l)}\,\mathrm{d}u \\
&= \frac{1}{f_G^{(i)}(l)}\int_{-\infty}^{z}\int p_i(u, l \mid b_i)\,\mathrm{d}G(b_i)\,\mathrm{d}u \\
&= \frac{C(\Sigma_i)}{f_G^{(i)}(l)}\int_{-\infty}^{z}\exp\left(-\frac{u^2}{2\Sigma_{i,11}}\right)\int\varphi\left(\frac{\Omega_{i,12}}{\Omega_{i,22}}u + l - b_i; \Omega_{i,22}^{-1}\right)\mathrm{d}G(b_i)\,\mathrm{d}u \\
&= \frac{C(\Sigma_i)}{f_G^{(i)}(l)}\int_{-\infty}^{z}\exp\left(-\frac{u^2}{2\Sigma_{i,11}}\right) f_G\left(\frac{\Omega_{i,12}}{\Omega_{i,22}}u + l; \Omega_{i,22}^{-1}\right)\mathrm{d}u.
\end{aligned}$$

□

C.2 Proof of Theorem 2

Since the map $x \mapsto 2\min\{x, 1-x\}$ is 2-Lipschitz on $[0,1]$, we have that under the null $H_{0i} : \theta_i = \theta_{i0}$,

$$\begin{aligned}
&\left|P^{(i)}_{G,\theta_{i0}}(\hat{\theta}^{\mathrm{db}}_i, \hat{b}_i) - P^{(i)}_{\widehat{G},\theta_{i0}}(\hat{\theta}^{\mathrm{db}}_i, \hat{b}_i)\right| \\
&\quad = \left|2\min\left\{F_{G,i}(\hat{\theta}^{\mathrm{db}}_i - \theta_{i0} \mid \hat{b}_i), 1 - F_{G,i}(\hat{\theta}^{\mathrm{db}}_i - \theta_{i0} \mid \hat{b}_i)\right\}\right. \\
&\qquad \left. -2\min\left\{F_{\widehat{G},i}(\hat{\theta}^{\mathrm{db}}_i - \theta_{i0} \mid \hat{b}_i), 1 - F_{\widehat{G},i}(\hat{\theta}^{\mathrm{db}}_i - \theta_{i0} \mid \hat{b}_i)\right\}\right| \\
&\quad \le 2\left|F_{G,i}(\hat{\theta}^{\mathrm{db}}_i - \theta_{i0} \mid \hat{b}_i) - F_{\widehat{G},i}(\hat{\theta}^{\mathrm{db}}_i - \theta_{i0} \mid \hat{b}_i)\right| \\
&\quad \le 2\sup_{z\in\mathbb{R}}\left|F_{G,i}(z \mid \hat{b}_i) - F_{\widehat{G},i}(z \mid \hat{b}_i)\right|.
\end{aligned}$$

Therefore, it suffices to prove

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_G\left[\sup_{z\in\mathbb{R}}\left|F_{G,i}(z \mid \hat{b}_i) - F_{\widehat{G},i}(z \mid \hat{b}_i)\right|\right] \lesssim \frac{(\log n)^{3/2}}{n^{(1-\tilde{\gamma}^2)/2}}.$$

Let $\mathcal{A}$ be the event with the constant C_0 in Lemma 6:

$$\mathcal{A} := \left\{\bar{d}(G, \widehat{G}) < C_0 \log n/\sqrt{n}\right\}.$$

and for any distribution $\widetilde{G}$ supported on $\mathbb{R}$, define

$$\begin{aligned}
N^{(i)}_{\widetilde{G}}(z, l) &:= F_{\widetilde{G},i}(z \mid l) \cdot f^{(i)}_{\widetilde{G}}(l), \\
N_i(z, \widetilde{G}) &:= N^{(i)}_{\widetilde{G}}(z, \hat{b}_i), \text{ and} \\
D_i(\widetilde{G}) &:= f^{(i)}_{\widetilde{G}}(\hat{b}_i).
\end{aligned}$$

By these definitions, it holds that $F_{\widetilde{G},i}(z \mid \hat{b}_i) = N_i(z, \widetilde{G})/D_i(\widetilde{G})$. Let $\widehat{G}_* = (\widehat{G} + G)/2$, then

$$\begin{aligned}
&\left|F_{G,i}(z \mid \hat{b}_i) - F_{\widehat{G},i}(z \mid \hat{b}_i)\right| \\
&\quad = \left|\frac{N_i(z,G)}{D_i(G)} - \frac{N_i(z,\widehat{G})}{D_i(\widehat{G})}\right| \\
&\quad = \left|\frac{N_i(z,G)}{D_i(G)} - \frac{N_i(z,G)}{D_i(\widehat{G}_*)} + \frac{N_i(z,G)}{D_i(\widehat{G}_*)} - \frac{N_i(z,\widehat{G})}{D_i(\widehat{G}_*)} + \frac{N_i(z,\widehat{G})}{D_i(\widehat{G}_*)} - \frac{N_i(z,\widehat{G})}{D_i(\widehat{G})}\right| \\
&\quad \le \frac{N_i(z,G)}{D_i(G)}\frac{|D_i(G) - D_i(\widehat{G}_*)|}{D_i(\widehat{G}_*)} + \frac{|N_i(z,G) - N_i(z,\widehat{G})|}{D_i(\widehat{G}_*)} + \frac{N_i(z,\widehat{G})}{D_i(\widehat{G})}\frac{|D_i(\widehat{G}_*) - D_i(\widehat{G})|}{D_i(\widehat{G}_*)} \\
&\quad \le \frac{|N_i(z,G) - N_i(z,\widehat{G})|}{D_i(\widehat{G}_*)} + \frac{|D_i(G) - D_i(\widehat{G})|}{D_i(\widehat{G}_*)}.
\end{aligned}$$

In the last step, we used two facts: first, it holds that $N_i(z, \widetilde{G})/D_i(\widetilde{G}) \in [0,1]$ for all $\widetilde{G}$ since they are conditional CDFs, and second, the map $\widetilde{G} \mapsto D_i(\widetilde{G})$ is linear, which implies that

$$D_i(G) - D_i(\widehat{G}_*) = (D_i(G) - D_i(\widehat{G}))/2, \quad D_i(\widehat{G}_*) - D_i(\widehat{G}) = (D_i(G) - D_i(\widehat{G}))/2.$$

Combining all of the above results, we get

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_G\left[\sup_{z\in\mathbb{R}}\left|F_{G,i}(z\mid\hat{b}_i)-F_{\widehat{G},i}(z\mid\hat{b}_i)\right|\right]\\
&\quad\le\mathbb{P}_G\left[\mathcal{A}^c\right]+\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_G\left[\sup_{z\in\mathbb{R}}\left|\frac{N_i(z,G)-N_i(z,\widehat{G})}{D_i(\widehat{G}_*)}\right|\mathbb{1}(\mathcal{A})\right]+\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_G\left[\left|\frac{D_i(G)-D_i(\widehat{G})}{D_i(\widehat{G}_*)}\right|\mathbb{1}(\mathcal{A})\right]\\
&\quad=:\mathbb{P}_G\left[\mathcal{A}^c\right]+\mathrm{I}+\mathrm{II}.
\end{aligned}\tag{S14}$$

By Lemma 6, $\mathbb{P}_G\left[\mathcal{A}^c\right]\le\exp\left(-c_0\log n\right)$ with $c_0\ge 2$. We consider the following two lemmas that bound terms I and II.

Lemma 8. It holds that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_G\left[\sup_{z\in\mathbb{R}}\left|\frac{N_i(z,G)-N_i(z,\widehat{G})}{D_i(\widehat{G}_*)}\right|\mathbb{1}(\mathcal{A})\right]\lesssim_{\Gamma,\underline{\sigma},\bar{\sigma},\underline{\gamma},\bar{\gamma}}\frac{(\log n)^{3/2}}{n^{(1-\bar{\gamma}^2)/2}}.$$

Lemma 9. It holds that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_G\left[\left|\frac{D_i(G)-D_i(\widehat{G})}{D_i(\widehat{G}_*)}\right|\mathbb{1}(\mathcal{A})\right]\lesssim_{\Gamma,\underline{\sigma},\bar{\sigma}}\frac{\log n}{\sqrt{n}}.$$

Plugging in the conclusions of the above two lemmas in (S14), the assertion of the theorem follows. To prove Lemma 8, we consider the following covering lemma.

Lemma 10. For $U>0$, define the distance

$$d_{N,U}(G_1,G_2):=\max_{1\le i\le n}\left[\left\|f^{(i)}_{G_1}-f^{(i)}_{G_2}\right\|_{\infty,U}+\sup_{z\in\mathbb{R}}\sup_{|l|\le U}\left|N^{(i)}_{G_1}(z,l)-N^{(i)}_{G_2}(z,l)\right|\right].\tag{S15}$$

Then for any $0<\varepsilon<1/\sqrt{2\pi}$ and $U>0$, it holds that

$$\log\mathcal{N}\left(\varepsilon,\mathcal{G},d_{N,U}\right)\lesssim_{\underline{\sigma},\bar{\sigma},\bar{\gamma}}\left(\log\left(\frac{1}{\varepsilon}\right)\right)^2\max\left\{\frac{U}{\sqrt{|\log\varepsilon|}},1\right\}.$$

Moreover, the same bound holds for a proper cover of any subset of $\mathcal{G}$, up to changing the covering radius by a universal constant.

Proof. Denote $a_i:=\Omega_{i,12}/\Omega_{i,22}$. Under Assumption 1, there exists a constant $C=C(\underline{\sigma},\bar{\sigma},\bar{\gamma})$ such that

$$|a_i|\le C,\qquad C(\Sigma_i)\le C,\qquad \Sigma_{i,11}\le C,\qquad \Sigma_{i,22}\in[C^{-1},C],\qquad \Omega^{-1}_{i,22}\in[C^{-1},C],$$

for all $i\in\{1,\dots,n\}$. Let $R_\varepsilon=C_1\sqrt{|\log\varepsilon|}$, with C_1 sufficiently large, and define

$$U_\varepsilon:=U+CR_\varepsilon.$$

We claim that for all $l\in\mathbb{R}$,

$$\sup_{\widetilde{G}\in\mathcal{G}}\sup_{1\le i\le n}C(\Sigma_i)\int_{|u|>R_\varepsilon}\exp\left(-\frac{u^2}{2\Sigma_{i,11}}\right)f_{\widetilde{G}}\left(a_iu+l;\Omega^{-1}_{i,22}\right)\mathrm{d}u\le\frac{\varepsilon}{4}.\tag{S16}$$

In particular, since $\Omega^{-1}_{i,22}\ge C^{-1}$ and $\|\varphi(\cdot;\sigma^2)\|_\infty\le 1/\sqrt{2\pi\sigma^2}$, we have $f_{\widetilde{G}}(t;\Omega^{-1}_{i,22})\le 1/\sqrt{2\pi C^{-1}}$. Then there is a constant $K_1=K_1(\underline{\sigma},\bar{\sigma},\bar{\gamma})$ such that $\sup_{\widetilde{G}}\sup_i\sup_{t\in\mathbb{R}}f_{\widetilde{G}}(t;\Omega^{-1}_{i,22})\le K_1$. Combining

with standard Gaussian tail bound and $\Sigma_{i,11} \le C$, we have

$$\int_{|u|>R_\varepsilon} \exp\left(-\frac{u^2}{2\Sigma_{i,11}}\right) f_{\widetilde{G}}\left(a_i u + l; \Omega_{i,22}^{-1}\right) \mathrm{d}u \le K_1 \int_{|u|>R_\varepsilon} \exp\left(-\frac{u^2}{2\Sigma_{i,11}}\right) \mathrm{d}u \le 2K_1\sqrt{2\pi C} \exp\left(-\frac{C_1^2|\log\varepsilon|}{2C}\right).$$

Since $C(\Sigma_i) \le C$, choose $K_2 = K_2(\underline{\sigma}, \bar{\sigma}, \bar{\gamma}) := 2CK_1\sqrt{2\pi C}$, and choose C_1 large enough so that $K_2\varepsilon^{C_1^2/(2C)-1} \le 1/4$ uniformly for $\varepsilon \in (0, 1/\sqrt{2\pi})$; this is possible by taking $C_1^2/(2C)$ as large as needed. Then we plug back and prove (S16).

Now consider the augmented class of Gaussian mixture densities

$$\mathcal{F}_n^{\mathrm{aug}} := \left\{ \left(f_{\widetilde{G}}(\cdot;\Sigma_{1,22}), \ldots, f_{\widetilde{G}}(\cdot;\Sigma_{n,22}), f_{\widetilde{G}}(\cdot;\Omega_{1,22}^{-1}), \ldots, f_{\widetilde{G}}(\cdot;\Omega_{n,22}^{-1}) \right) : \widetilde{G} \in \mathcal{G} \right\}.$$

has $2n$ Gaussian convolutions of $\widetilde{G}$, all with variances bounded above and below by constants depending only on $\underline{\sigma}, \bar{\sigma}, \bar{\gamma}$. Hence, the proof of Lemma 5 applies verbatim to this augmented class. Therefore, there exist $G_1, ..., G_J$ with

$$\log J \lesssim_{\underline{\sigma},\bar{\sigma},\bar{\gamma}} \left(\log\left(\frac{1}{\varepsilon}\right)\right)^2 \max\left\{\frac{U_\varepsilon}{\sqrt{|\log\varepsilon|}}, 1\right\} \lesssim \left(\log\left(\frac{1}{\varepsilon}\right)\right)^2 \max\left\{\frac{U}{\sqrt{|\log\varepsilon|}}, 1\right\},$$

such that for every $\widetilde{G} \in \mathcal{G}$, there is some $j \in \{1, \ldots, J\}$ satisfying

$$\max_{1\le i\le n} \left\| f_{\widetilde{G}}(\cdot;\Sigma_{i,22}) - f_{G_j}(\cdot;\Sigma_{i,22}) \right\|_{\infty,U} + \max_{1\le i\le n} \left\| f_{\widetilde{G}}(\cdot;\Omega_{i,22}^{-1}) - f_{G_j}(\cdot;\Omega_{i,22}^{-1}) \right\|_{\infty,U_\varepsilon} \le c\varepsilon,$$

for a sufficiently small constant $c > 0$.

We claim that the same $\{G_j\}$ also approximates $N_{\widetilde{G}}^{(i)}$ uniformly on $\mathbb{R} \times [-U, U]$. Indeed, by Proposition 7, for $|l| \le U$,

$$\begin{aligned}
&\sup_{z\in\mathbb{R}} \left| N_{\widetilde{G}}^{(i)}(z,l) - N_{G_j}^{(i)}(z,l) \right| \\
&\quad \le C(\Sigma_i) \int_{\mathbb{R}} \exp\left(-\frac{u^2}{2\Sigma_{i,11}}\right) \left| f_{\widetilde{G}}\left(a_i u + l; \Omega_{i,22}^{-1}\right) - f_{G_j}\left(a_i u + l; \Omega_{i,22}^{-1}\right) \right| \mathrm{d}u \\
&\quad \le Cc\varepsilon + C(\Sigma_i) \int_{|u|>R_\varepsilon} \exp\left(-\frac{u^2}{2\Sigma_{i,11}}\right) \left[f_{\widetilde{G}}\left(a_i u + l; \Omega_{i,22}^{-1}\right) + f_{G_j}\left(a_i u + l; \Omega_{i,22}^{-1}\right) \right] \mathrm{d}u \\
&\quad \le Cc\varepsilon + \epsilon/2 \le \varepsilon,
\end{aligned}$$

where we use the fact that $|a_i u + l| \le U_\varepsilon$ whenever $|u| \le R_\varepsilon$ and $|l| \le U$, combined with (S16), and by choosing a $c > 0$ sufficiently small. Thus, the same cover is an ε-cover under $d_{N,U}$, after adjusting constants.

Indeed, if we restrict to any subset of $\mathcal{G}$, an improper $\varepsilon/2$-cover can be converted into a proper ε-cover by replacing each center whose ball intersects the subset by one point in that intersection. This proves the last claim. □

Proof of Lemma 8. Since $b_i \sim G \in \mathcal{G}_\Gamma$, and $\hat{b}_i = b_i + \varepsilon_i$ with $\varepsilon_i \sim \mathrm{N}(0, \tau_i^2)$, then $\hat{b}_i$ is a sub-Gaussian with a proxy constant depending only on $(\Gamma, \bar{\sigma})$ (since $\tau_i^2 \le \bar{\sigma}^2$). In particular, there exists a constant $c > 0$, depending only on $(\Gamma, \bar{\sigma})$, such that $\mathbb{P}(|\hat{b}_i| > t) \le 2e^{-ct^2}$. Hence, we can choose $B_n := C_B\sqrt{\log n}$ with C_B sufficiently large such that

$$\mathbb{P}_G\left[|\hat{b}_i| > B_n\right] \le \frac{1}{n}.$$

Also let $\eta := n^{-2}$. Define the local class

$$\mathcal{G}_n := \left\{ \widetilde{G} \in \mathcal{G} : \ \bar{d}(G, \widetilde{G}) < C_0 \frac{\log n}{\sqrt{n}} \right\},$$

where C_0 is the constant in Lemma 6. By applying Lemma 10 with $U = B_n$, $\varepsilon = \eta$, there exists a proper η-cover $\{G_j : j = 1, \dots, J\} \subseteq \mathcal{G}_n$ of $\mathcal{G}_n$ under d_{N,B_n} satisfying

$$\log J \lesssim_{\underline{\sigma},\bar{\sigma},\bar{\gamma}} (\log n)^2 \max\left\{ \frac{B_n}{\sqrt{\log n}}, 1 \right\} \lesssim (\log n)^2.$$

On the event $\mathcal{A}$, we have $\widehat{G} \in \mathcal{G}_n$, and hence there exists a random index $\widehat{j} \in \{1, \dots, J\}$ such that $d_{N,B_n}(\widehat{G}, G_{\widehat{j}}) \le \eta$, that is

$$\sup_{z \in \mathbb{R}} \left| N^{(i)}_{\widehat{G}}(z,l) - N^{(i)}_{G_{\widehat{j}}}(z,l) \right| \le \eta, \qquad \left| f^{(i)}_{\widehat{G}}(l) - f^{(i)}_{G_{\widehat{j}}}(l) \right| \le \eta \qquad \text{for all } |l| \le B_n,\ i \le n. \tag{S17}$$

For each deterministic G_j, define

$$V_{ij}(l) := \frac{\sup_{z \in \mathbb{R}} \left| N^{(i)}_{G}(z,l) - N^{(i)}_{G_j}(z,l) \right|}{f^{(i)}_{G}(l) + f^{(i)}_{G_j}(l)},$$

and $V_{ij}(l) \in [0,1]$ since $0 \le N^{(i)}_{\widetilde{G}}(z,l) \le f^{(i)}_{\widetilde{G}}(l)$ for any $\widetilde{G}$.

We claim that on $\mathcal{A}$,

$$\sup_{z \in \mathbb{R}} \left| \frac{N_i(z,G) - N_i(z,\widehat{G})}{D_i(\widehat{G}_*)} \right| \lesssim \mathbb{1}(|\hat{b}_i| > B_n) + V_{i\widehat{j}}(\hat{b}_i) + \eta \frac{\mathbb{1}(|\hat{b}_i| \le B_n)}{f^{(i)}_{G}(\hat{b}_i)}. \tag{S18}$$

We prove this by case analysis. For every $z \in \mathbb{R}$, write

$$\frac{|N_i(z,G) - N_i(z,\widehat{G})|}{D_i(\widehat{G}_*)} \le \frac{|N_i(z,G) - N_i(z,G_{\widehat{j}})|}{D_i(\widehat{G}_*)} + \frac{|N_i(z,G_{\widehat{j}}) - N_i(z,\widehat{G})|}{D_i(\widehat{G}_*)}.$$

For the second term, by (S17) applied at $l = \hat{b}_i$ when $|\hat{b}_i| \le B_n$, the numerator is $\le \eta$. The denominator satisfies $D_i(\widehat{G}_*) = (f^{(i)}_{G}(\hat{b}_i) + f^{(i)}_{\widehat{G}}(\hat{b}_i))/2 \ge f^{(i)}_{G}(\hat{b}_i)/2$, hence

$$\frac{|N_i(z,G_{\widehat{j}}) - N_i(z,\widehat{G})|}{D_i(\widehat{G}_*)} \le \frac{2\eta}{f^{(i)}_{G}(\hat{b}_i)} \quad \text{on } \{|\hat{b}_i| \le B_n\}.$$

For the first term, for $|l| \le B_n$. If $f^{(i)}_{G}(l) + f^{(i)}_{G_{\widehat{j}}}(l) \ge 2\eta$, then $D_i(\widehat{G}_*) \ge \{f^{(i)}_{G}(l) + f^{(i)}_{G_{\widehat{j}}}(l)\}/4$, and hence

$$\frac{\sup_z |N^{(i)}_{G}(z,l) - N^{(i)}_{G_{\widehat{j}}}(z,l)|}{D_i(\widehat{G}_*)} \le 4 V_{i\widehat{j}}(l).$$

On the other hand, if $f^{(i)}_{G}(l) + f^{(i)}_{G_{\widehat{j}}}(l) < 2\eta$, then both $f^{(i)}_{G}(l) < 2\eta$ and $f^{(i)}_{G_{\widehat{j}}}(l) < 2\eta$, then $\sup_z |N^{(i)}_{G}(z,l) - N^{(i)}_{G_{\widehat{j}}}(z,l)| \le 2\eta$. Combined with $D_i(\widehat{G}_*) \ge f^{(i)}_{G}(l)/2$, then

$$\frac{|N_i(z,G) - N_i(z,G_{\widehat{j}})|}{D_i(\widehat{G}_*)} \le \frac{4\eta}{f^{(i)}_{G}(l)}.$$

For $|\hat{b}_i| > B_n$, then $|N_i(z,G) - N_i(z,\widehat{G})| \le f_G^{(i)}(\hat{b}_i) + f_{\widehat{G}}^{(i)}(\hat{b}_i) = 2D_i(\widehat{G}_*)$, so

$$\frac{|N_i(z,G) - N_i(z,\widehat{G})|}{D_i(\widehat{G}_*)} \le 2\,\mathbb{1}(|\hat{b}_i| > B_n).$$

Combining all together claims (S18). We further take expectations in (S18),

$$\begin{aligned} \mathrm{I} &\lesssim \frac{1}{n} + \eta \cdot \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_G\left[\frac{\mathbb{1}(|\hat{b}_i| \le B_n)}{f_G^{(i)}(\hat{b}_i)}\right] + \mathbb{E}_G\left[\max_{1\le j\le J}\frac{1}{n}\sum_{i=1}^{n} V_{ij}(\hat{b}_i)\right] \\ &= \frac{1}{n} + 2B_n\eta + \mathbb{E}_G\left[\max_{1\le j\le J}\frac{1}{n}\sum_{i=1}^{n} V_{ij}(\hat{b}_i)\right], \end{aligned} \tag{S19}$$

where the second term is justified by

$$\mathbb{E}_G\left[\frac{\mathbb{1}(|\hat{b}_i| \le B_n)}{f_G^{(i)}(\hat{b}_i)}\right] = \int_{-B_n}^{B_n} \frac{1}{f_G^{(i)}(t)} f_G^{(i)}(t)\,\mathrm{d}t = 2B_n.$$

For a fixed j, the random variables $V_{ij}(\hat{b}_i)$ are independent and take values in $[0,1]$. Hence, by Hoeffding's lemma and a union bound,

$$\begin{aligned} &\mathbb{E}_G\left[\max_{1\le j\le J}\frac{1}{n}\sum_{i=1}^{n} V_{ij}(\hat{b}_i)\right] \\ &\le \max_{1\le j\le J}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_G\left[V_{ij}(\hat{b}_i)\right] + \mathbb{E}_G\left[\max_{1\le j\le J}\frac{1}{n}\sum_{i=1}^{n}\left\{V_{ij}(\hat{b}_i) - \mathbb{E}_G[V_{ij}(\hat{b}_i)]\right\}\right] \\ &\le \max_{1\le j\le J}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_G\left[V_{ij}(\hat{b}_i)\right] + \inf_{\lambda>0}\left\{\frac{\log J}{\lambda} + \frac{\lambda}{8n}\right\} \\ &\le \max_{1\le j\le J}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_G\left[V_{ij}(\hat{b}_i)\right] + C\sqrt{\frac{\log J}{n}}. \end{aligned} \tag{S20}$$

We next control the deterministic means in (S20). For fixed i and j, set

$$S_{ij} := \sup_{z\in\mathbb{R}}\sup_{|l|\le B_n}\left|N_G^{(i)}(z,l) - N_{G_j}^{(i)}(z,l)\right|.$$

Since $V_{ij} \le 1$ and $f_G^{(i)}(l)/\{f_G^{(i)}(l) + f_{G_j}^{(i)}(l)\} \le 1$, we have

$$\begin{aligned} \mathbb{E}_G\left[V_{ij}(\hat{b}_i)\right] &\le \mathbb{P}_G\left[|\hat{b}_i| > B_n\right] + \int_{-B_n}^{B_n}\sup_{z\in\mathbb{R}}\left|N_G^{(i)}(z,l) - N_{G_j}^{(i)}(z,l)\right|\mathrm{d}l \\ &\le \frac{1}{n} + 2B_nS_{ij}. \end{aligned} \tag{S21}$$

It remains to bound S_{ij}. By Proposition 7,

$$\begin{aligned} &\left|N_G^{(i)}(z,l) - N_{G_j}^{(i)}(z,l)\right| / C(\Sigma_i) \\ &= \left|\int_{-\infty}^{z}\exp\left(-\frac{u^2}{2\Sigma_{i,11}}\right)\left(f_G\left(\frac{\Omega_{i,12}}{\Omega_{i,22}}u + l; \Omega_{i,22}^{-1}\right) - f_{G_j}\left(\frac{\Omega_{i,12}}{\Omega_{i,22}}u + l; \Omega_{i,22}^{-1}\right)\right)\,\mathrm{d}u\right| \\ &\le \left[\int_{-\infty}^{z}\exp\left(-\frac{u^2}{\Sigma_{i,11}}\right)\mathrm{d}u\right]^{\frac{1}{2}}\left[\int_{-\infty}^{z}\left(f_G\left(\frac{\Omega_{i,12}}{\Omega_{i,22}}u + l; \Omega_{i,22}^{-1}\right) - f_{G_j}\left(\frac{\Omega_{i,12}}{\Omega_{i,22}}u + l; \Omega_{i,22}^{-1}\right)\right)^2\mathrm{d}u\right]^{\frac{1}{2}}. \end{aligned}$$

Let

$$h_{ij}(t) := f_G(t;\Omega_{i,22}^{-1}) - f_{G_j}(t;\Omega_{i,22}^{-1}), \qquad a_i := \frac{\Omega_{i,12}}{\Omega_{i,22}} = -\gamma_i \frac{\tau_i}{\tilde{\sigma}_i},$$

and by Assumption 1, $|a_i|$ is bounded away from zero,

$$\begin{aligned}\int_{-\infty}^{z} \left(f_G\left(\frac{\Omega_{i.12}}{\Omega_{i,22}}u + l; \Omega_{i,22}^{-1}\right) - f_{G_j}\left(\frac{\Omega_{i,12}}{\Omega_{i,22}}u + l; \Omega_{i,22}^{-1}\right)\right)^2 \mathrm{d}u &= \int_{-\infty}^{z} h_{ij}(a_i u + l)^2 \mathrm{d}u \\ &\leq \frac{1}{|a_i|}\int_{-\infty}^{\infty} h_{ij}(t)^2 \mathrm{d}t,\end{aligned}$$

Therefore

$$\sup_{z\in\mathbb{R}}\sup_{l\in\mathbb{R}} \left| N_G^{(i)}(z,l) - N_{G_j}^{(i)}(z,l)\right| \lesssim_{\underline{\sigma},\bar{\sigma},\underline{\gamma},\bar{\gamma}} \left(\int_{\mathbb{R}} h_{ij}(t)^2 \mathrm{d}t\right)^{1/2}. \tag{S22}$$

Notice that $\Omega_{i,22}^{-1} = \Sigma_{i,22}(1-\gamma_i^2)$, and denote f^* as the Fourier transform of f, then

$$f_G^*(\omega;\Omega_{i,22}^{-1}) = G^*(\omega)e^{-\frac{1}{2}\Omega_{i,22}^{-1}\omega^2} = \left(G^*(\omega)e^{-\frac{1}{2}\Sigma_{i,22}\omega^2}\right)e^{\frac{1}{2}\Sigma_{i,22}\gamma_i^2\omega^2} = f_G^*(\omega;\Sigma_{i,22})e^{\frac{1}{2}\Sigma_{i,22}\gamma_i^2\omega^2}.$$

Similar results also hold for $\{G_j\}$. By Plancherel's identity, and a non-random $U_i > 0$ to be chosen later,

$$\begin{aligned}\int_{-\infty}^{\infty} h_{ij}(t)^2 \mathrm{d}t &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \left| f_G^*(\omega;\Omega_{i,22}^{-1}) - f_{G_j}^*(\omega;\Omega_{i,22}^{-1})\right|^2 d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{\Sigma_{i,22}\gamma_i^2\omega^2}\left| f_G^*(\omega;\Sigma_{i,22}) - f_{G_j}^*(\omega;\Sigma_{i,22})\right|^2 d\omega \\ &= \frac{1}{2\pi}\int_{[-U_i,U_i]} e^{\Sigma_{i,22}\gamma_i^2\omega^2}\left| f_G^*(\omega;\Sigma_{i,22}) - f_{G_j}^*(\omega;\Sigma_{i,22})\right|^2 d\omega \\ &\quad + \frac{1}{2\pi}\int_{[-U_i,U_i]^c} e^{\Sigma_{i,22}\gamma_i^2\omega^2}\left| f_G^*(\omega;\Sigma_{i,22}) - f_{G_j}^*(\omega;\Sigma_{i,22})\right|^2 d\omega \\ &:= I_{ij1} + I_{ij2}.\end{aligned}$$

For I_{ij1},

$$\begin{aligned}I_{ij1} &\leq e^{\Sigma_{i,22}\gamma_i^2 U_i^2}\cdot\frac{1}{2\pi}\int_{-\infty}^{\infty}\left| f_G^*(\omega;\Sigma_{i,22}) - f_{G_j}^*(\omega;\Sigma_{i,22})\right|^2 d\omega \\ &= e^{\Sigma_{i,22}\gamma_i^2 U_i^2}\int_{-\infty}^{\infty}\left(f_G(t;\Sigma_{i,22}) - f_{G_j}(t;\Sigma_{i,22})\right)^2 dt \\ &= e^{\Sigma_{i,22}\gamma_i^2 U_i^2}\int_{-\infty}^{\infty}\left(\sqrt{f_G(t;\Sigma_{i,22})} - \sqrt{f_{G_j}(t;\Sigma_{i,22})}\right)^2\left(\sqrt{f_G(t;\Sigma_{i,22})} + \sqrt{f_{G_j}(t;\Sigma_{i,22})}\right)^2 dt \\ &\leq 2e^{\Sigma_{i,22}\gamma_i^2 U_i^2}\left\|\sqrt{f_G(t;\Sigma_{i,22})} + \sqrt{f_{G_j}(t;\Sigma_{i,22})}\right\|_\infty^2 \mathfrak{H}^2(f_G(\cdot;\Sigma_{i,22}), f_{G_j}(\cdot;\Sigma_{i,22})) \\ &\lesssim e^{\Sigma_{i,22}\gamma_i^2 U_i^2}\mathfrak{H}^2(f_G(\cdot;\Sigma_{i,22}), f_{G_j}(\cdot;\Sigma_{i,22})).\end{aligned}$$

For I_{ij2}, first noticing that

$$\left| f_G^*(\omega;\Sigma_{i,22}) - f_{G_j}^*(\omega;\Sigma_{i,22})\right|^2 = \left|G^*(\omega) - G_j^*(\omega)\right|^2 e^{-\Sigma_{i,22}\omega^2} \leq 4e^{-\Sigma_{i,22}\omega^2},$$

then applying the standard Gaussian-tail bound,

$$I_{i2} \leq \frac{4}{2\pi}\int_{[-U_i,U_i]^c} e^{-\Sigma_{i,22}\omega^2}e^{\Sigma_{i,22}\gamma_i^2\omega^2}\, d\omega \lesssim \frac{1}{\Sigma_{i,22}(1-\gamma_i^2)U_i}e^{-\Sigma_{i,22}(1-\gamma_i^2)U_i^2}.$$

Take $U_i = (\log n/\Sigma_{i,22})^{1/2}$, then we have

$$\int_{-\infty}^{\infty} h_{ij}(t)^2 dt \lesssim n^{\gamma_i^2}\mathfrak{H}^2(f_G^{(i)}, f_{G_j}^{(i)}) + \frac{n^{-(1-\gamma_i^2)}}{\sqrt{\log n}}.$$

By using the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b > 0$, and combined with (S22), we have that

$$S_{ij} \lesssim_{\underline{\sigma},\bar{\sigma}\underline{\gamma},\bar{\gamma}} n^{\gamma_i^2/2}\mathfrak{H}(f_G^{(i)}, f_{G_j}^{(i)}) + n^{-(1-\gamma_i^2)/2}(\log n)^{-1/4}. \tag{S23}$$

Since the cover is proper, each G_j belongs to $\mathcal{G}_n$. Hence, uniformly over $j \in \{1, \ldots, J\}$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} S_{ij} &\lesssim n^{\tilde{\gamma}^2/2}\frac{1}{n}\sum_{i=1}^{n}\mathfrak{H}(f_G^{(i)}, f_{G_j}^{(i)}) + n^{-(1-\tilde{\gamma}^2)/2}(\log n)^{-1/4} \\ &\leq n^{\tilde{\gamma}^2/2}\bar{d}(G, G_j) + n^{-(1-\tilde{\gamma}^2)/2}(\log n)^{-1/4} \\ &\lesssim \frac{\log n}{n^{(1-\tilde{\gamma}^2)/2}}. \end{aligned} \tag{S24}$$

Combining (S21) and (S24), we get

$$\max_{1\leq j\leq J}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_G\left[V_{ij}(\hat{b}_i)\right] \lesssim \frac{1}{n} + \frac{B_n\log n}{n^{(1-\tilde{\gamma}^2)/2}} \lesssim \frac{(\log n)^{3/2}}{n^{(1-\tilde{\gamma}^2)/2}}.$$

Also, combined with the Hoeffding term,

$$\sqrt{\frac{\log J}{n}} \lesssim \frac{\log n}{\sqrt{n}} \lesssim \frac{(\log n)^{3/2}}{n^{(1-\tilde{\gamma}^2)/2}}.$$

Plugging these bounds into (S19),

$$\mathrm{I} \lesssim_{\Gamma,\underline{\sigma},\bar{\sigma},\underline{\gamma},\bar{\gamma}} \frac{(\log n)^{3/2}}{n^{(1-\tilde{\gamma}^2)/2}}.$$

This completes the proof.

□

Proof of Lemma 9. Let $\eta = 1/n$. Consider the following class of densities

$$\mathcal{F}_n^{-1} := \left\{\left(f_{\widetilde{G}}^{(1)}, ..., f_{\widetilde{G}}^{(n)}\right) : \widetilde{G} \in \mathcal{G},\ \bar{d}(G, \widetilde{G}) < \frac{C_0\log n}{\sqrt{n}}\right\}.$$

The constant C_0 is in Lemma 6. Again, we take $B_n := C_B\sqrt{\log n}$. Let $\mathcal{S} = \{(f_{G_j}^{(1)}, ..., f_{G_j}^{(n)}) : j \in \mathcal{J}\} \subseteq \mathcal{F}_n^{-1}$, $\mathcal{J} = \{1, ..., J\}$, $J = \#S$ be a proper $(\|\cdot\|_{\infty,B_n}, \eta)$-cover of $\mathcal{F}_n^{-1}$. Here, a proper cover means that the centers of the cover are themselves elements of $\mathcal{F}_n^{-1}$. Lemma 5 provides a cover for a larger class of functions, however, it is not a proper cover for $\mathcal{F}_n^{-1}$. By a standard argument, an $\eta/2$-cover in Lemma 5 yields a proper η-cover for $\mathcal{F}_n^{-1}$. Therefore, we get that

$$\log J \lesssim_{\underline{\sigma},\bar{\sigma}} (\log n)^2 \max\left\{\frac{B_n}{\sqrt{\log n}}, 1\right\}.$$

By the definition of B_n, we further conclude that $\log J \lesssim_{\Gamma,\underline{\sigma},\bar{\sigma}} (\log n)^2$. Now consider the main argument of the lemma. On the event $\mathcal{A}$, there must exist a (random) index $\hat{j}$ such that

$$\max_{1\leq i\leq n}\max_{|t|\leq B_n}\left|f_{\widehat{G}}^{(i)}(t) - f_{G_{\hat{j}}}^{(i)}(t)\right| \leq \eta.$$

In particular, on $\{|\hat{b}_i| \le B_n\}$, $\left|D_i(\widehat{G}) - D_i(G_{\hat{j}})\right| \le \eta$. Also, we have $\mathbb{P}_G\left[|\hat{b}_i| > B_n\right] \le 1/n$ for $n \ge 1$. Then,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_G\left[\left|\frac{D_i(G) - D_i(\widehat{G})}{D_i(\hat{G}_*)}\right|\mathbb{1}(\mathcal{A})\right] \\
&= \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}_G\left[\left|\frac{D_i(G) - D_i(\widehat{G})}{D_i(G) + D_i(\widehat{G})}\right|\mathbb{1}(\mathcal{A})\right] \\
&\le \frac{2}{n}\sum_{i=1}^{n}\left\{\mathbb{E}_G\left[\left|\frac{D_i(G) - D_i(\widehat{G})}{D_i(G) + D_i(\widehat{G})}\right|\mathbb{1}(\mathcal{A})\mathbb{1}(|\hat{b}_i| \le B_n)\right] + \mathbb{P}\left[|\hat{b}_i| > B_n)\right]\right\} \\
&\le \frac{2}{n} + \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}_G\left[\left|\frac{D_i(G) - D_i(G_{\hat{j}})}{D_i(G) + D_i(G_{\hat{j}})}\right|\mathbb{1}(\mathcal{A})\mathbb{1}(|\hat{b}_i| \le B_n)\right] \\
&\qquad + \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}_G\left[\left|\frac{D_i(G) - D_i(\widehat{G})}{D_i(G) + D_i(\widehat{G})} - \frac{D_i(G) - D_i(G_{\hat{j}})}{D_i(G) + D_i(G_{\hat{j}})}\right|\mathbb{1}(\mathcal{A})\mathbb{1}(|\hat{b}_i| \le B_n)\right]. \qquad \text{(S25)}
\end{aligned}$$

Proceeding similarly as in the proof of Lemma S10 of Ignatiadis and Sen [2025], we can show that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_G\left[\left|\frac{D_i(G) - D_i(\widehat{G})}{D_i(G) + D_i(\widehat{G})} - \frac{D_i(G) - D_i(G_{\hat{j}})}{D_i(G) + D_i(G_{\hat{j}})}\right|\mathbb{1}(\mathcal{A})\mathbb{1}(|\hat{b}_i| \le B_n)\right] \\
&\le 2\eta\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_G\left[\frac{\mathbb{1}(|\hat{b}_i| \le B_n)}{D_i(G)}\right] \le 4\eta B_n.
\end{aligned}$$

For the remaining term,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_G\left[\left|\frac{D_i(G) - D_i(G_{\hat{j}})}{D_i(G) + D_i(G_{\hat{j}})}\right|\mathbb{1}(\mathcal{A})\mathbb{1}(|\hat{b}_i| \le B_n)\right] \\
&\le \mathbb{E}_G\left[\sup_{j\in\mathcal{J}}\left\{\frac{1}{n}\sum_{i=1}^{n}\left|\frac{D_i(G) - D_i(G_j)}{D_i(G) + D_i(G_j)}\right|\right\}\right] \\
&\le \mathbb{E}_G\left[\sup_{j\in\mathcal{J}}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(\left|\frac{D_i(G) - D_i(G_j)}{D_i(G) + D_i(G_j)}\right| - \mathbb{E}_G\left[\left|\frac{D_i(G) - D_i(G_j)}{D_i(G) + D_i(G_j)}\right|\right]\right)\right\}\right] \\
&\qquad + \sup_{j\in\mathcal{J}}\left\{\mathbb{E}_G\left[\frac{1}{n}\sum_{i=1}^{n}\left|\frac{D_i(G) - D_i(G_j)}{D_i(G) + D_i(G_j)}\right|\right]\right\}.
\end{aligned}$$

Both terms in the last step can be controlled using the same techniques applied in the proof of Lemma S10 of Ignatiadis and Sen [2025]. In fact, using Lemma 5 and Lemma 6, we can show that

$$\sup_{j\in\mathcal{J}}\left\{\mathbb{E}_G\left[\frac{1}{n}\sum_{i=1}^{n}\left|\frac{D_i(G) - D_i(G_j)}{D_i(G) + D_i(G_j)}\right|\right]\right\} \lesssim_{\Gamma,\underline{\sigma},\bar{\sigma}} \frac{\log n}{\sqrt{n}},$$

and

$$\mathbb{E}_G\left[\sup_{j\in\mathcal{J}}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(\left|\frac{D_i(G) - D_i(G_j)}{D_i(G) + D_i(G_j)}\right| - \mathbb{E}_G\left[\left|\frac{D_i(G) - D_i(G_j)}{D_i(G) + D_i(G_j)}\right|\right]\right)\right\}\right] \lesssim_{\Gamma,\underline{\sigma},\bar{\sigma}} \frac{\log n}{\sqrt{n}}.$$

Therefore, combining above result into (S25), we have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_G\left[\left|\frac{D_i(G) - D_i(\widehat{G})}{D_i(\widehat{G}_*)}\right|\mathbb{1}(\mathcal{A})\right] \lesssim_{\Gamma,\underline{\sigma},\bar{\sigma}} \frac{\log n}{\sqrt{n}}.$$

□

C.3 Proof of Theorem 4

Fix $i \in \{1, ..., n\}$. We consider the testing problem $H_{0i} : \theta_i = \theta_{i0}$, and take $\theta_{i0} = \theta_i$ to be the true value. Under H_{0i}, we have

$$\begin{pmatrix} \hat{\theta}_i^{\mathrm{db}} - \theta_i \\ \hat{b}_i \end{pmatrix} \Bigg| \, b_i \sim N\left\{ \begin{pmatrix} 0 \\ b_i \end{pmatrix}, \Sigma_i \right\}.$$

By definition of $F_{G,i}(\cdot \mid L_i)$, the conditional CDF of $\hat{\theta}_i^{\mathrm{db}} - \theta_i$ given $\hat{b}_i$ under the oracle model is $F_{G,i}(\cdot \mid \hat{b}_i)$. Therefore, by the probability integral transform,

$$U_i := F_{G,i}(\hat{\theta}_i^{\mathrm{db}} - \theta_i \mid \hat{b}_i) \sim \mathrm{Unif}(0, 1).$$

Since $\widehat{G}$ is computed from $\{\hat{b}_1, \ldots, \hat{b}_n\}$, we may also condition $\mathcal{L}_n := \sigma(\hat{b}_1, \ldots, \hat{b}_n)$, and $U_i \mid \mathcal{L}_n \sim \mathrm{Unif}(0, 1)$.

Now we define the plug-in conditional probability transform

$$\widehat{U}_i := F_{\widehat{G},i}(\hat{\theta}_i^{\mathrm{db}} - \theta_i \mid \hat{b}_i),$$

and we have that

$$|\widehat{U}_i - U_i| \le \sup_{z \in \mathbb{R}} \left| F_{\widehat{G},i}(z \mid \hat{b}_i) - F_{G,i}(z \mid \hat{b}_i) \right| =: \Delta_i.$$

By the definition of $\mathcal{I}_{\widehat{G},i}(1-\alpha)$,

$$\{\theta_i \in \mathcal{I}_{\widehat{G},i}(1-\alpha)\} = \left\{ \frac{\alpha}{2} \le F_{\widehat{G},i}(\hat{\theta}_i^{\mathrm{db}} - \theta_i \mid \hat{b}_i) \le 1 - \frac{\alpha}{2} \right\} = \left\{ \frac{\alpha}{2} \le \widehat{U}_i \le 1 - \frac{\alpha}{2} \right\}.$$

Combined with $|\widehat{U}_i - U_i| \le \Delta_i$, we have the following set inclusions

$$\left\{ \frac{\alpha}{2} + \Delta_i \le U_i \le 1 - \frac{\alpha}{2} - \Delta_i \right\} \subseteq \{\theta_i \in \mathcal{I}_{\widehat{G},i}(1-\alpha)\} \left\{ \frac{\alpha}{2} - \Delta_i \le U_i \le 1 - \frac{\alpha}{2} + \Delta_i \right\}.$$

Taking conditional probabilities given $\mathcal{L}_n$, and using $U_i \mid \mathcal{L}_n \sim \mathrm{Unif}(0, 1)$, we obtain

$$1 - \alpha - 2\Delta_i \le \mathbb{P}_G \left\{ \theta_i \in \mathcal{I}_{\widehat{G},i}(1-\alpha) \mid \mathcal{L}_n \right\} \le 1 - \alpha + 2\Delta_i.$$

Hence

$$\left| \mathbb{P}_G \left\{ \theta_i \in \mathcal{I}_{\widehat{G},i}(1-\alpha) \mid \mathcal{L}_n \right\} - (1-\alpha) \right| \le 2\Delta_i.$$

Then by taking the expectation on both sides, we have

$$\begin{aligned} &\left| \mathbb{P}_G \left\{ \theta_i \in \mathcal{I}_{\widehat{G},i}(1-\alpha) \right\} - (1-\alpha) \right| \\ &= \left| \mathbb{E}_G \left[P_G \left\{ \theta_i \in \mathcal{I}_{\widehat{G},i}(1-\alpha) \mid \mathcal{L}_n \right\} - (1-\alpha) \right] \right| \\ &\le \mathbb{E}_G \left[\left| \mathbb{P}_G \left\{ \theta_i \in \mathcal{I}_{\widehat{G},i}(1-\alpha) \mid \mathcal{L}_n \right\} - (1-\alpha) \right| \right] \le 2\mathbb{E}_G \left[\Delta_n \right]. \end{aligned}$$

Averaging over i, we get

$$\frac{1}{n} \sum_{i=1}^{n} \left| \mathbb{P}_G \left\{ \theta_i \in \mathcal{I}_{\widehat{G},i}(1-\alpha) \right\} - (1-\alpha) \right| \le \frac{2}{n} \sum_{i=1}^{n} \mathbb{E}_G[\Delta_i] \le 2C \frac{(\log n)^{3/2}}{n^{(1-\check{\gamma}^2)/2}},$$

where the last inequality comes from the proof of Theorem 2. This holds for all $\alpha \in (0, 1)$, then the following completes the proof:

$$\sup_{\alpha \in (0,1)} \frac{1}{n} \sum_{i=1}^{n} \left| \mathbb{P}_G \left\{ \theta_i \in \mathcal{I}_{\widehat{G},i}(1-\alpha) \right\} - (1-\alpha) \right| \le C' \frac{(\log n)^{3/2}}{n^{(1-\check{\gamma}^2)/2}}.$$

C.4 Proof of Proposition 3

Proof. Since $\tilde{\sigma}_i^2, \tau_i^2, \gamma_i$ do not depend on i, we drop the subscript i throughout the proof. Write $\nu = \tau^2(1-\gamma^2)$. Recall from Proposition 7 that, in this homoscedastic case, $F_G(z \mid l) = N_G(z,l)/D_G(l)$, where $D_G(l) = f_G(l;\tau^2)$ and

$$N_G(z,l) = C\int_{-\infty}^{z} \exp\left(-\frac{u^2}{2\tilde{\sigma}^2}\right) f_G\left(l - \frac{\gamma\tau}{\tilde{\sigma}}u;\nu\right)\,\mathrm{d}u, \qquad C = (2\pi\tilde{\sigma}^2)^{-1/2}.$$

Here we used that

$$\frac{\Omega_{12}}{\Omega_{22}} = -\frac{\gamma\tau}{\tilde{\sigma}}, \qquad \Omega_{22}^{-1} = \tau^2(1-\gamma^2) = \nu.$$

Our strategy is to use Le Cam's two-point lemma. Define

$$r_A = \frac{\nu(A+\tau^2)}{\tau^2(A+\nu)} = \frac{(1-\gamma^2)(A+\tau^2)}{A+\tau^2(1-\gamma^2)}.$$

Since $\gamma \neq 0$, we have $r_A < 1$, and also $r_A \to 1-\gamma^2$ as $A \to \infty$. Since $\beta > (1-\gamma^2)/2$, we may choose A large enough so that $r_A/2 < \beta$, and then choose $\Gamma > A$. Below we fix these values of A and Γ.

As the first point, take $G_0 = \mathrm{N}(0,A)$, and write g_0 for its density. Then G_0 is A-sub-Gaussian, and hence $G_0 \in \mathcal{G}_\Gamma$. For the second point, let

$$\omega^2 = \log n \cdot \frac{A+\tau^2}{A\tau^2}$$

and define

$$h(b) = g_0(b)\left\{\cos(\omega b) - \exp(-A\omega^2/2)\right\}.$$

For $\varepsilon \in (0, 1/4)$, to be chosen sufficiently small below, let G_1 have density

$$g_1(b) = g_0(b) + \varepsilon h(b).$$

We first check that g_1 is a density. Indeed,

$$g_1(b) \geq g_0(b)\left\{1 - \varepsilon\left(|\cos(\omega b)| + \exp(-A\omega^2/2)\right)\right\} \geq g_0(b)(1-2\varepsilon) > 0,$$

and

$$\int h(b)\,\mathrm{d}b = \int g_0(b)\left\{\cos(\omega b) - \exp(-A\omega^2/2)\right\}\,\mathrm{d}b = 0.$$

Thus g_1 is nonnegative and integrates to one.

We next check the sub-Gaussian property. Since g_1 is even, $\mathbb{E}_{G_1}[b] = 0$. Moreover, for any $s \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}_{G_1}\left[\exp(sb)\right] &= \exp(As^2/2) + \varepsilon\exp(As^2/2 - A\omega^2/2)\left\{\cos(As\omega) - 1\right\} \\ &\leq \exp(As^2/2) \leq \exp(\Gamma s^2/2),\end{aligned}$$

where we used $\cos(As\omega) - 1 \leq 0$ and $\Gamma > A$. Hence $G_1 \in \mathcal{G}_\Gamma$.

We will use the following elementary Gaussian convolution identity: for any $v > 0$,

$$\begin{aligned}&\int \varphi(x-b;v)\,\varphi(b;A)\left\{\cos(\omega b) - \exp(-A\omega^2/2)\right\}\,\mathrm{d}b \\ &\quad = \varphi(x;A+v)\left[\exp\left\{-\frac{1}{2}\omega^2\frac{Av}{A+v}\right\}\cos\left(\omega\frac{A}{A+v}x\right) - \exp(-A\omega^2/2)\right].\end{aligned}$$

Let $\Delta D(l) = D_{G_1}(l) - D_{G_0}(l)$ and $\Delta N(z,l) = N_{G_1}(z,l) - N_{G_0}(z,l)$. Applying the identity with $v = \tau^2$ gives

$$\frac{\Delta D(l)}{D_{G_0}(l)} = \varepsilon\left[n^{-1/2}\cos\left(\omega\frac{A}{A+\tau^2}l\right) - n^{-\frac{A+\tau^2}{2\tau^2}}\right].$$

In particular,

$$\sup_{l\in\mathbb{R}}\left|\frac{\Delta D(l)}{D_{G_0}(l)}\right| \le 2\varepsilon n^{-1/2}.$$

Applying the same identity with $v=\nu$ gives

$$\Delta N(t,l) = \varepsilon C\int_{-\infty}^{t}\exp\left(-\frac{u^2}{2\tilde\sigma^2}\right)\varphi\left(l-\frac{\gamma\tau}{\tilde\sigma}u; A+\nu\right) \times\left[n^{-r_A/2}\cos\left(\omega\frac{A}{A+\nu}\left(l-\frac{\gamma\tau}{\tilde\sigma}u\right)\right) - n^{-\frac{A+\tau^2}{2\tau^2}}\right]\mathrm{d}u.$$

Thus

$$\frac{\Delta N(t,l)}{D_{G_0}(l)} = \varepsilon n^{-r_A/2}B_n(l) - \varepsilon n^{-\frac{A+\tau^2}{2\tau^2}}F_{G_0}(t\mid l),$$

where

$$B_n(l) := \int_{-\infty}^{t} W_l(u)\cos\left(\omega\frac{A}{A+\nu}\left(l-\frac{\gamma\tau}{\tilde\sigma}u\right)\right)\mathrm{d}u$$

and

$$W_l(u) := \frac{C}{D_{G_0}(l)}\exp\left(-\frac{u^2}{2\tilde\sigma^2}\right)\varphi\left(l-\frac{\gamma\tau}{\tilde\sigma}u; A+\nu\right).$$

We now lower bound $B_n(l)$ on a set of l's that has probability bounded away from zero. Fix a compact interval $I=[-1,1]$. Uniformly over $l\in I$, the function $u\mapsto W_l(u)$ is smooth and has Gaussian tails; moreover, for constants $0<c<C<\infty$ not depending on n,

$$0<c\le W_l(t)\le C, \qquad |W_l'(t)| + \int_{-\infty}^{t}|W_l''(u)|\,\mathrm{d}u \le C.$$

Write, only for this calculation,

$$\phi_l(u) = \omega\frac{A}{A+\nu}\left(l-\frac{\gamma\tau}{\tilde\sigma}u\right).$$

Since

$$\frac{\mathrm{d}}{\mathrm{d}u}\sin\{\phi_l(u)\} = -\frac{\omega A\gamma\tau}{(A+\nu)\tilde\sigma}\cos\{\phi_l(u)\},$$

integration by parts gives

$$B_n(l) = -\frac{W_l(t)\sin\{\phi_l(t)\}}{\omega A\gamma\tau/\{(A+\nu)\tilde\sigma\}} + \frac{1}{\omega A\gamma\tau/\{(A+\nu)\tilde\sigma\}}\int_{-\infty}^{t}W_l'(u)\sin\{\phi_l(u)\}\,\mathrm{d}u.$$

Applying integration by parts once more to the remaining integral, using

$$\frac{\mathrm{d}}{\mathrm{d}u}\cos\{\phi_l(u)\} = \frac{\omega A\gamma\tau}{(A+\nu)\tilde\sigma}\sin\{\phi_l(u)\},$$

we get, uniformly over $l\in I$,

$$\begin{aligned} B_n(l) &= -\frac{W_l(t)\sin\{\phi_l(t)\}}{\omega A\gamma\tau/\{(A+\nu)\tilde\sigma\}} \\ &\quad + \frac{W_l'(t)\cos\{\phi_l(t)\} - \int_{-\infty}^{t}W_l''(u)\cos\{\phi_l(u)\}\,\mathrm{d}u}{\left(\omega A\gamma\tau/\{(A+\nu)\tilde\sigma\}\right)^2} \\ &= -\frac{W_l(t)\sin\{\phi_l(t)\}}{\omega A\gamma\tau/\{(A+\nu)\tilde\sigma\}} + O(\omega^{-2}). \end{aligned}$$

Here the boundary terms at $-\infty$ vanish because W_l and W_l' have Gaussian tails.

The sine is not uniformly large, so define

$$S_n = \left\{ l \in I : \left| \sin\left(\omega \frac{A}{A+\nu} \left(l - \frac{\gamma\tau}{\tilde{\sigma}} t \right) \right) \right| \geq \frac{1}{2} \right\}.$$

On S_n, the leading term has size at least a constant multiple of ω^{-1}, while the remainder is $O(\omega^{-2})$. Hence, for all sufficiently large n,

$$|B_n(l)| \gtrsim \omega^{-1} \gtrsim \frac{1}{\sqrt{\log n}}, \qquad l \in S_n.$$

Finally, as a function of l, the sine above has period of order ω^{-1}, and on each full period the set where its absolute value is at least $1/2$ occupies a fixed positive fraction of the period. Since $\omega \to \infty$, I contains many periods for large n. Since under G_0 we have $\hat{b}_i \sim \mathrm{N}(0, A+\tau^2)$ and this density is bounded below on I, there is a constant $p > 0$ such that

$$\mathbb{P}_{G_0}\left[\hat{b}_i \in S_n \right] \geq p$$

for all sufficiently large n.

Combining the expressions for $\Delta D(l)$ and $\Delta N(t,l)$, we get

$$\begin{aligned} F_{G_1}(t \mid l) - F_{G_0}(t \mid l) &= \frac{\Delta N(t,l)/D_{G_0}(l) - F_{G_0}(t \mid l)\Delta D(l)/D_{G_0}(l)}{1 + \Delta D(l)/D_{G_0}(l)} \\ &= \frac{\varepsilon n^{-r_A/2} B_n(l) - \varepsilon n^{-1/2} F_{G_0}(t \mid l) \cos\left(\omega \frac{A}{A+\tau^2} l \right)}{1 + \Delta D(l)/D_{G_0}(l)}. \end{aligned}$$

The denominator is bounded away from zero for all sufficiently large n, while the second term in the numerator is negligible relative to $n^{-r_A/2}/\sqrt{\log n}$ because $r_A < 1$. Hence, for all sufficiently large n and all $l \in S_n$,

$$|F_{G_1}(t \mid l) - F_{G_0}(t \mid l)| \gtrsim \varepsilon \frac{n^{-r_A/2}}{\sqrt{\log n}}.$$

It remains to verify that the two experiments are close. Let P_j denote the joint law of $(\hat{b}_1, \ldots, \hat{b}_n)$ under G_j, $j = 0, 1$, and let $P_{j,1}$ denote the corresponding one-coordinate marginal law. The density of $P_{j,1}$ is D_{G_j}. Therefore,

$$\chi^2(P_{1,1}, P_{0,1}) = \int \left(\frac{D_{G_1}(l)}{D_{G_0}(l)} - 1 \right)^2 D_{G_0}(l)\, \mathrm{d}l.$$

From the expression above for $\Delta D(l) = D_{G_1}(l) - D_{G_0}(l)$,

$$\frac{D_{G_1}(l)}{D_{G_0}(l)} - 1 = \varepsilon \left[n^{-1/2} \cos\left(\omega \frac{A}{A+\tau^2} l \right) - n^{-\frac{A+\tau^2}{2\tau^2}} \right].$$

Under $P_{0,1}$, we have $\hat{b}_i \sim \mathrm{N}(0, A+\tau^2)$. Hence, using $\mathbb{E}\left[\cos(sX)\right] = \exp(-s^2 \operatorname{Var}[X]/2)$ for a centered normal random variable X,

$$\mathbb{E}_{P_{0,1}}\left[\cos\left(\omega \frac{A}{A+\tau^2} \hat{b}_i \right) \right] = \exp\left\{ -\frac{1}{2}\omega^2 \frac{A^2}{A+\tau^2} \right\} = n^{-A/(2\tau^2)}.$$

Similarly,

$$\begin{aligned} \mathbb{E}_{P_{0,1}}\left[\cos^2\left(\omega \frac{A}{A+\tau^2} \hat{b}_i \right) \right] &= \frac{1}{2} + \frac{1}{2}\mathbb{E}_{P_{0,1}}\left[\cos\left(2\omega \frac{A}{A+\tau^2} \hat{b}_i \right) \right] \\ &= \frac{1}{2}\left(1 + n^{-2A/\tau^2} \right). \end{aligned}$$

Therefore,

$$\begin{aligned}\chi^2(P_{1,1},P_{0,1}) &= \varepsilon^2\mathbb{E}_{P_{0,1}}\left[\left[n^{-1/2}\cos\left(\omega\frac{A}{A+\tau^2}\hat{b}_i\right)-n^{-\frac{A+\tau^2}{2\tau^2}}\right]^2\right]\\ &= \varepsilon^2\left[\frac{1}{2n}\left(1+n^{-2A/\tau^2}\right)-2n^{-1/2-\frac{A+\tau^2}{2\tau^2}}n^{-A/(2\tau^2)}+n^{-\frac{A+\tau^2}{\tau^2}}\right]\\ &= \frac{\varepsilon^2}{2n}\left(1-n^{-A/\tau^2}\right)^2 \le \frac{\varepsilon^2}{2n}.\end{aligned}$$

Since the coordinates are independent under both P_0 and P_1,

$$1+\chi^2(P_1,P_0) = \left\{1+\chi^2(P_{1,1},P_{0,1})\right\}^n.$$

Thus

$$\chi^2(P_1,P_0) \le \left(1+\frac{\varepsilon^2}{2n}\right)^n - 1 \le \exp(\varepsilon^2/2)-1.$$

Consequently,

$$\mathrm{TV}(P_0,P_1) \le \left\{\chi^2(P_1,P_0)\right\}^{1/2} \le \left\{\exp(\varepsilon^2/2)-1\right\}^{1/2}.$$

Choosing $\varepsilon>0$ small enough ensures that $\mathrm{TV}(P_0,P_1)\le p/2$.

Now lift S_n to an event in the full n-dimensional sample space: $\mathcal{E}_{n,i}=\{(l_1,\dots,l_n)\in\mathbb{R}^n : l_i\in S_n\}$. Then $P_0(\mathcal{E}_{n,i})\ge p$. We write $P_0\wedge P_1$ for the common part of the two measures; if p_0,p_1 are their densities, then $P_0\wedge P_1$ has density $\min\{p_0,p_1\}$. Hence

$$(P_0\wedge P_1)(\mathcal{E}_{n,i}) \ge P_0(\mathcal{E}_{n,i}) - \mathrm{TV}(P_0,P_1) \ge p/2.$$

Let $\widehat{\psi}(t)$ be any measurable function of $(\hat{b}_1,\dots,\hat{b}_n)$. Pointwise,

$$\left|\widehat{\psi}(t)-F_{G_0}(t\mid l_i)\right| + \left|\widehat{\psi}(t)-F_{G_1}(t\mid l_i)\right| \ge \left|F_{G_1}(t\mid l_i)-F_{G_0}(t\mid l_i)\right|.$$

By Le Cam's argument,

$$\begin{aligned}&\max_{j\in\{0,1\}}\mathbb{E}_{G_j}\left[\left|\widehat{\psi}(t)-F_{G_j}(t\mid\hat{b}_i)\right|\right]\\ &\quad\ge \frac{1}{2}\int\left|F_{G_1}(t\mid l_i)-F_{G_0}(t\mid l_i)\right|\,\mathrm{d}(P_0\wedge P_1)(l_1,\dots,l_n)\\ &\quad\ge \frac{1}{2}\int_{\mathcal{E}_{n,i}}\left|F_{G_1}(t\mid l_i)-F_{G_0}(t\mid l_i)\right|\,\mathrm{d}(P_0\wedge P_1)(l_1,\dots,l_n)\\ &\quad\gtrsim \varepsilon\frac{n^{-r_A/2}}{\sqrt{\log n}}(P_0\wedge P_1)(\mathcal{E}_{n,i}) \gtrsim \frac{n^{-r_A/2}}{\sqrt{\log n}}.\end{aligned}$$

Since $r_A/2<\beta$,

$$\frac{n^{-r_A/2}}{\sqrt{\log n}} \gtrsim n^{-\beta}$$

for all sufficiently large n. Since $G_0,G_1\in\mathcal{G}_\Gamma$, this implies

$$\inf_{\widehat{\psi}(t)}\sup_{G\in\mathcal{G}_\Gamma}\mathbb{E}_G\left[\left|\widehat{\psi}(t)-F_{G,i}(t\mid\hat{b}_i)\right|\right] \ge cn^{-\beta}$$

for some $c>0$, after reducing c if necessary to handle finitely many small values of n. This proves the claim. □

D Details on numerical studies (Section 6)

This appendix collects the dataset construction, hyperparameter choices, and supplementary tables for the numerical studies in Section 6.

Computational resources. We run all the numerical studies on an HPC cluster with 16-core Intel Xeon CPUs and 32GB memory. With our choice of $K = 200$ Monte Carlo (in the synthetic Amazon case, $K = 500$) replicates, all our results can be obtained within 5 minutes.

D.1 LMArena

Table S1: LMArena results across $\alpha \in \{0.01, 0.05, 0.10, 0.20, 0.30\}$, averaged over $K = 200$ random labeled/unlabeled splits with $n = 298$ pairwise LLM problems. Each cell reports mean $\pm$ 1 Monte-Carlo SE. *Classical* is the interval without ML information; *Pred Mean* is the prediction-only interval; *PT* denotes the power-tuned PPI baseline; *RB Normal* and *RB NPMLE* are the rebiased PT estimators with Normal and NPMLE priors for bias. The width-ratio is normalized by the Classical (CLT) interval.

α		**Classical**	**Pred Mean**	**PT**	**RB Normal**	**RB NPMLE**
0.01	coverage (%)	96.5 ± 0.1	55.1 ± 0.1	95.8 ± 0.1	96.3 ± 0.1	94.1 ± 0.1
	width	0.675 ± 0.000	0.210 ± 0.000	0.647 ± 0.000	0.454 ± 0.001	0.420 ± 0.002
	width-ratio	1.000 ± 0.000	0.311 ± 0.000	0.958 ± 0.000	0.673 ± 0.002	0.622 ± 0.003
0.05	coverage (%)	92.9 ± 0.1	43.4 ± 0.1	91.3 ± 0.1	91.4 ± 0.1	87.1 ± 0.2
	width	0.513 ± 0.000	0.160 ± 0.000	0.492 ± 0.000	0.345 ± 0.001	0.316 ± 0.001
	width-ratio	1.000 ± 0.000	0.311 ± 0.000	0.958 ± 0.000	0.673 ± 0.002	0.615 ± 0.002
0.1	coverage (%)	88.4 ± 0.1	36.7 ± 0.1	86.4 ± 0.1	86.1 ± 0.2	80.6 ± 0.2
	width	0.431 ± 0.000	0.134 ± 0.000	0.413 ± 0.000	0.290 ± 0.001	0.264 ± 0.001
	width-ratio	1.000 ± 0.000	0.311 ± 0.000	0.958 ± 0.000	0.673 ± 0.002	0.612 ± 0.002
0.2	coverage (%)	79.9 ± 0.2	28.5 ± 0.1	77.4 ± 0.2	76.2 ± 0.2	69.5 ± 0.3
	width	0.336 ± 0.000	0.105 ± 0.000	0.322 ± 0.000	0.226 ± 0.001	0.205 ± 0.001
	width-ratio	1.000 ± 0.000	0.311 ± 0.000	0.958 ± 0.000	0.673 ± 0.002	0.610 ± 0.002
0.3	coverage (%)	71.1 ± 0.2	22.9 ± 0.1	68.3 ± 0.2	66.3 ± 0.2	59.6 ± 0.3
	width	0.272 ± 0.000	0.085 ± 0.000	0.260 ± 0.000	0.183 ± 0.000	0.165 ± 0.001
	width-ratio	1.000 ± 0.000	0.311 ± 0.000	0.958 ± 0.000	0.673 ± 0.002	0.609 ± 0.003

Dataset construction. We begin with the public dataset https://huggingface.co/datasets/lmarena-ai/arena-human-preference-140k, which contains 140k pairwise LLM data, where each row consists of two LLMs responses to a given prompt, and human's decision of which response is better. Our data cleaning pipeline then removes all rows containing non-English prompts and involving multi-turn conversation (i.e. the user asks a follow-up question). Then we group the rows by the pair of LLMs involved (we take LLM A to be the one whose model name is alphabetically smaller), this gives us $n = 298$ LLM pairs for our numerical experiments. All tasks have at least 100 data points (i.e. $M_i + m_i \geq 100$).

Reward model predictions. For each comparison X_{ij}, the reward model `Skywork-Reward-V2` directly outputs two scalar scores s_j^A and s_j^B for the responses of LLM A and B, respectively. We then convert them into a probabilistic prediction via the Bradley-Terry model

$$\hat{p}_j = \frac{1}{1 + \exp(s_j^B - s_j^A)}.$$

The biased estimator $\hat{\theta}_i^{\mathrm{ML}}$ is the average of $\hat{p}_j$ across all comparisons for the pair. The pooled MSE of these predictions against the true preference frequencies is 0.38, indicating substantial predictor mis-specification.

D.2 Amazon reviews

Dataset. We use the publicly available *Amazon Fine Food Reviews* dataset, hosted by the Stanford Network Analysis Project (SNAP) on Kaggle.[5] Reviews are grouped by their `ProductID`; for each review we form the covariate X_{ij} by concatenating the review title and body text, and we let the response $Y_{ij} \in \{1, 2, 3, 4, 5\}$ be the integer rating (higher is better) chosen by the j-th reviewer of this product. Following the experimental design of Li and Ignatiadis [2025], we restrict attention to the $n = 200$ products with the most reviews (totaling 74,913 reviews). Selecting the top-reviewed products mitigates extreme heteroscedasticity across the per-task variances σ_i^2 that would otherwise dominate any cross-task comparison. For each product we randomly split its reviews into a labeled and unlabeled partition with a 20/80 ratio, repeating the random split for $K = 200$ trials; the results reported in Fig. 3 and Table S2 are computed over these trials.

Predictor. The prediction model h is a fine-tuned BERT [Devlin et al., 2019] neural network. We start from the publicly available `bert-base-multilingual` checkpoint[6], which is pre-trained on general multilingual product reviews (not exclusive to Amazon) for the same 1–5 star prediction task, and we further fine-tune it for two full epochs on the reviews *outside* the top-200 products, using the Hugging Face `transformers` library. Fine-tuning improves prediction accuracy on a disjoint validation set of 100 products ($\sim$ 46k reviews) from 67.5% (off-the-shelf checkpoint) to 78.8% (fine-tuned). We thus use this fine-tuned version for our model h. Crucially, the per-product samples used to compute $\hat{\theta}_i^{\mathrm{b}}$ and $\hat{b}_i$ in (S8) are entirely disjoint from the corpus on which h was fine-tuned, so the predictor is independent of the samples that we use to construct estimators.

Table S2: Amazon results across $\alpha \in \{0.01, 0.05, 0.10, 0.20, 0.30\}$, averaged over 200 random splits with $n = 200$ tasks. width-ratio is normalized by the Classical interval. Entries are reported as mean $\pm$ Monte Carlo SE.

α		Classical	Pred Mean	PT	RB Normal	RB NPMLE
0.01	coverage (%)	98.7 ± 0.1	98.2 ± 0.1	99.4 ± 0.1	99.2 ± 0.1	99.5 ± 0.1
	width	0.696 ± 0.001	0.358 ± 0.000	0.438 ± 0.001	0.357 ± 0.001	0.357 ± 0.001
	width-ratio	1.000 ± 0.000	0.525 ± 0.001	0.640 ± 0.001	0.521 ± 0.001	0.520 ± 0.001
0.05	coverage (%)	95.9 ± 0.1	95.9 ± 0.1	97.7 ± 0.1	98.5 ± 0.1	98.7 ± 0.1
	width	0.530 ± 0.001	0.273 ± 0.000	0.334 ± 0.001	0.271 ± 0.001	0.270 ± 0.001
	width-ratio	1.000 ± 0.000	0.525 ± 0.001	0.640 ± 0.001	0.521 ± 0.001	0.517 ± 0.001
0.10	coverage (%)	92.2 ± 0.2	91.9 ± 0.2	95.3 ± 0.1	97.4 ± 0.1	97.6 ± 0.1
	width	0.445 ± 0.001	0.229 ± 0.000	0.280 ± 0.001	0.228 ± 0.001	0.226 ± 0.001
	width-ratio	1.000 ± 0.000	0.525 ± 0.001	0.640 ± 0.001	0.521 ± 0.001	0.516 ± 0.001
0.20	coverage (%)	83.7 ± 0.2	82.5 ± 0.2	89.3 ± 0.2	94.1 ± 0.1	94.0 ± 0.1
	width	0.346 ± 0.001	0.178 ± 0.000	0.218 ± 0.001	0.177 ± 0.001	0.176 ± 0.001
	width-ratio	1.000 ± 0.000	0.525 ± 0.001	0.640 ± 0.001	0.521 ± 0.001	0.515 ± 0.001
0.30	coverage (%)	74.3 ± 0.2	72.4 ± 0.2	81.4 ± 0.2	88.9 ± 0.2	88.6 ± 0.2
	width	0.280 ± 0.001	0.144 ± 0.000	0.176 ± 0.000	0.144 ± 0.001	0.142 ± 0.001
	width-ratio	1.000 ± 0.000	0.525 ± 0.001	0.640 ± 0.001	0.521 ± 0.001	0.514 ± 0.001

[5] https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews
[6] https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

D.3 Synthetic study

Data generation process. For each task $i \in \{1, \dots, n\}$ with $n = 200$, we draw a prediction mean $\theta_{0i} \sim \mathrm{N}(4,\, 0.01^2)$ and a prediction bias b_i according to the prior under study (Normal or two-point; see below). The true target parameter is $\theta_i = \theta_{0i} - b_i$. Synthetic summary-level observations $(\bar{Y}_i, \bar{Z}_i^h, \tilde{Z}_i^h)$ are then generated from an empirical covariance structure extracted from one task chosen at random from the Amazon dataset; from these we form the PT summary pair $(\hat{\theta}_i^{\mathrm{PT}},\, \hat{b}_{i,\hat{\lambda}_i})$ using the power-tuning constant as in the main text for each i. Each setting is repeated for $K = 500$ Monte Carlo replications, and we report average coverage, average length, and average length-ratio.

Choice of A for the Normal prior. For $b_i \sim \mathrm{N}(-0.1, A)$, we sweep $A^{1/2} \in \{0.01, 0.03, 0.05, 0.1\}$. The grid is anchored on the parametric prior fitted from the real Amazon split in Section D.2, whose marginal-MLE estimate has standard deviation ≈ 0.022. The smallest value ($\sigma = 0.01$) corresponds to a regime in which the bias is negligible and a rebiased estimator should behave like the biased one; the largest ($\sigma = 0.1$) is roughly $4\times$ wider than the empirical estimate and forces the bias to dominate the prior, so that a rebiased estimator should behave more like the debiased one. The two intermediate values trace the transition.

Two-point prior. For the misspecification probe, we draw b_i from a discrete two-point distribution $b_i \sim \frac{1}{2}\delta_0 + \frac{1}{2}\delta_{b_0}$, and we sweep $b_0 \in \{0.05, 0.1, 0.2, 0.5\}$. This grid controls the degree to which the bias distribution departs from the normal working prior. For $b_0 = 0.05$, the two support points are close enough that the distribution is difficult to distinguish from a nearly degenerate prior around zero, and the normal prior approximation is expected to work well. As b_0 increases, the discreteness and asymmetry of the bias distribution become more pronounced. At $b_0 = 0.5$, the two components are well separated, making the normal prior substantially misspecified. The two intermediate values trace the transition from a nearly normal-approximable regime to a strongly non-normal regime.

Per-A numerical results. Table S3 reports average coverage, average width, and average width-ratio for each estimator under the Gaussian-prior simulation, across the four values of $A^{1/2}$.

Table S3: Simulation results under normal prior with covariance structure adopted from the Amazon data ($n = 200$, $K = 500$ replications). Entries are reported as mean ± Monte Carlo SE.

$A^{1/2}$		**Oracle**	**Classical**	**Pred Mean**	**PT**	**RB Normal**	**RB NPMLE**
0.01	coverage (%)	95.1 ± 0.1	95.0 ± 0.1	66.2 ± 0.2	95.1 ± 0.1	95.1 ± 0.1	95.2 ± 0.1
	width	0.259 ± 0.001	0.525 ± 0.001	0.274 ± 0.001	0.331 ± 0.001	0.260 ± 0.001	0.261 ± 0.001
	width-ratio	0.498 ± 0.001	1.000 ± 0.000	0.527 ± 0.001	0.642 ± 0.001	0.500 ± 0.001	0.503 ± 0.001
0.03	coverage (%)	95.1 ± 0.1	95.0 ± 0.1	65.4 ± 0.2	95.1 ± 0.1	94.9 ± 0.1	94.9 ± 0.1
	width	0.269 ± 0.001	0.525 ± 0.001	0.274 ± 0.001	0.331 ± 0.001	0.268 ± 0.001	0.269 ± 0.001
	width-ratio	0.519 ± 0.001	1.000 ± 0.000	0.527 ± 0.001	0.642 ± 0.001	0.517 ± 0.001	0.518 ± 0.001
0.05	coverage (%)	95.1 ± 0.1	95.0 ± 0.1	63.9 ± 0.2	95.1 ± 0.1	95.0 ± 0.1	94.8 ± 0.1
	width	0.282 ± 0.001	0.525 ± 0.001	0.274 ± 0.001	0.331 ± 0.001	0.281 ± 0.001	0.280 ± 0.001
	width-ratio	0.545 ± 0.001	1.000 ± 0.000	0.527 ± 0.001	0.642 ± 0.001	0.544 ± 0.001	0.542 ± 0.001
0.10	coverage (%)	95.1 ± 0.1	95.0 ± 0.1	58.0 ± 0.2	95.1 ± 0.1	95.0 ± 0.1	94.5 ± 0.1
	width	0.305 ± 0.001	0.525 ± 0.001	0.274 ± 0.001	0.331 ± 0.001	0.304 ± 0.001	0.302 ± 0.001
	width-ratio	0.591 ± 0.001	1.000 ± 0.000	0.527 ± 0.001	0.642 ± 0.001	0.590 ± 0.001	0.584 ± 0.001

Per-b_0 numerical results. Table S4 reports average coverage, average width, and average width-ratio for each estimator under the two-point prior simulation, across the four values of b_0.

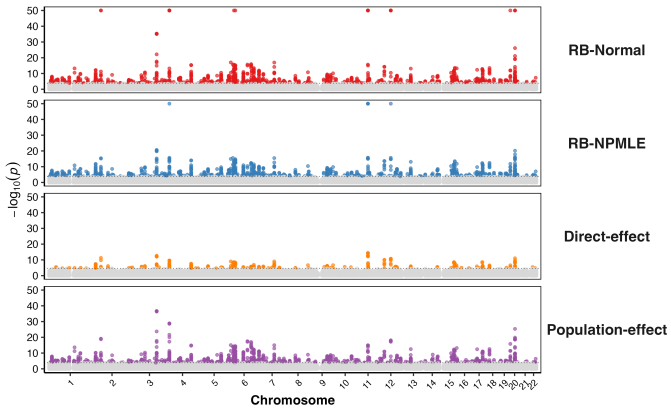

Figure S1: Manhattan plots of SNP-level p-values, shown on the $-\log_{10}$ scale. Horizontal dashed lines indicate the corresponding B-H thresholds. SNPs passing the threshold are highlighted as significant. For visualization, values exceeding $-\log_{10}(10^{-50}) = 50$ are truncated at 50.

Table S5: Summary of LD-clumped SNP discoveries from the height family-based GWAS summary statistics. We report the numbers of LD-clumped discoveries from BH at FDR 0.05 applied to our *RB-NPMLE* and *RB-Normal* p-values, the (unbiased) direct-effect p-values, and the (biased) population-effect p-values, together with their overlaps with Howe et al. [2022] signals.

	RB-NPMLE (ours)	**RB-Normal (ours)**	**Direct-effect (unbiased)**	**Population-effect (biased)**
# discoveries	544	743	111	560
# overlaps	176	175	77	162

Table S4: Simulation results under two-point prior bias with covariance structure adopted from the Amazon data ($n = 200$, $K = 500$ replications). Entries are reported as mean ± Monte Carlo SE.

b_0		**Oracle**	**Classical**	**Pred Mean**	**PT**	**RB Normal**	**RB NPMLE**
0.05	coverage (%)	95.0 ± 0.1	95.1 ± 0.1	91.1 ± 0.1	95.0 ± 0.1	94.8 ± 0.1	94.8 ± 0.1
	width	0.266 ± 0.001	0.525 ± 0.001	0.274 ± 0.001	0.332 ± 0.001	0.265 ± 0.001	0.266 ± 0.001
	width-ratio	0.513 ± 0.001	1.000 ± 0.000	0.527 ± 0.001	0.643 ± 0.001	0.511 ± 0.001	0.513 ± 0.001
0.10	coverage (%)	95.0 ± 0.1	95.1 ± 0.1	80.4 ± 0.2	95.0 ± 0.1	94.8 ± 0.1	94.8 ± 0.1
	width	0.280 ± 0.001	0.525 ± 0.001	0.274 ± 0.001	0.332 ± 0.001	0.281 ± 0.001	0.280 ± 0.001
	width-ratio	0.541 ± 0.001	1.000 ± 0.000	0.527 ± 0.001	0.643 ± 0.001	0.544 ± 0.001	0.542 ± 0.001
0.20	coverage (%)	94.9 ± 0.1	95.1 ± 0.1	57.2 ± 0.2	95.0 ± 0.1	94.8 ± 0.1	95.0 ± 0.1
	width	0.289 ± 0.001	0.525 ± 0.001	0.274 ± 0.001	0.332 ± 0.001	0.305 ± 0.001	0.291 ± 0.001
	width-ratio	0.555 ± 0.001	1.000 ± 0.000	0.527 ± 0.001	0.643 ± 0.001	0.591 ± 0.001	0.560 ± 0.001
0.50	coverage (%)	95.0 ± 0.1	95.1 ± 0.1	47.5 ± 0.2	95.0 ± 0.1	94.9 ± 0.1	95.2 ± 0.1
	width	0.261 ± 0.001	0.525 ± 0.001	0.274 ± 0.001	0.332 ± 0.001	0.325 ± 0.001	0.266 ± 0.001
	width-ratio	0.501 ± 0.001	1.000 ± 0.000	0.527 ± 0.001	0.643 ± 0.001	0.630 ± 0.001	0.510 ± 0.001

D.4 Family-based GWAS

Datasets. The height family-based GWAS summary statistics from Guan et al. [2025] are publicly available at `https://thessgac.com`. The estimates were obtained by running family-based SNP-wise regressions on 44,570 "white British" individuals in the UK Biobank [Bycroft et al., 2018], controlling for 40 genetic principal components and other covariates. Sibling (close to our target direct effect) estimate summary statistics from Howe et al. [2022] are accessible from OpenGWAS [Elsworth et al., 2020] through the `ieugwasr` R package [Hemani et al., 2025] with OpenGWAS ID ieu-b-4813. 1000 Genomes Phase 3 EUR reference panel [1000 Genomes Project Consortium et al., 2015] can be obtained from `http://fileserve.mrcieu.ac.uk/ld/1kg.v3.tgz`.

Details on overlap and LD-matching analysis. For each height analysis in Howe et al. [2022], we define putatively significant SNPs using their suggested liberal threshold, p-value $< 1 \times 10^{-6}$, and perform the comparison separately for the sibling-based direct-effect and population-effect estimates. We first restrict our discoveries to variants present in the 1000 Genomes European reference panel 1000 Genomes Project Consortium et al. [2015], which was used to estimate linkage disequilibrium (LD). We then count a discovery as matching Howe et al. if it is either identical to one of their putatively significant SNPs or is in LD with one, defined as being within 250 kb and having $r^2 \geq 0.8$. This LD-based comparison provides a stringent operational definition for treating nearby correlated variants as evidence for the same underlying association signal. We used `plink2` [Chang et al., 2015] to compute LD and retain variant pairs within 250 kb with $r^2 \geq 0.8$.

Overlap analysis on LD clumped discoveries One possible concern is that the larger number of overlaps from our rebiasing procedure, compared with the population-effect baseline, could simply reflect the discovery of more correlated SNPs. To investigate this further, for each discovery set, we identify independent significant SNPs via LD clumping [Purcell et al., 2007], so that discovered SNPs that are physically close to each other (within 250 kb) and are correlated ($r^2 > 0.1$) are reduced to one single representative SNP. We repeat the overlap analysis, with results summarized in Fig. S5. The observed patterns are similar to those shown in Table 1, with our RB-NPMLE-based procedure showing more overlaps than the biased population-effect baseline but with fewer total LD-clumped discoveries.

Comments on small bias assumptions for height. We expect the true bias b_i for height to be small for the following reasons: (1) the top 40 genetic principal components are included as controls in the regressions, so much of the confounding bias due to population stratification should be accounted for [Price et al., 2006]; and (2) previous analyses [Yengo et al., 2018, Kong et al., 2018, Young et al., 2022] have shown that parental indirect genetic effects are weak for height. Although confounding due to assortative mating may still be present, it is expected to be small.