Empirical Bayes Rebiasing

Wanyi Ling[1] **Sida Li**[2] **Junming Guan**[1] **Nikolaos Ignatiadis**[1,2]

[1]Department of Statistics, University of Chicago
[2]Data Science Institute, University of Chicago

{wanyiling, listar2000, junmingguan, ignat}@uchicago.edu

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 $\mathbb{P}\left[\theta \in \mathcal{I}\right] \approx 95\%$ under three conditions:

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

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, when available, 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 knowledge about the bias, it is prudent to be robust against non-negligible bias. 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.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

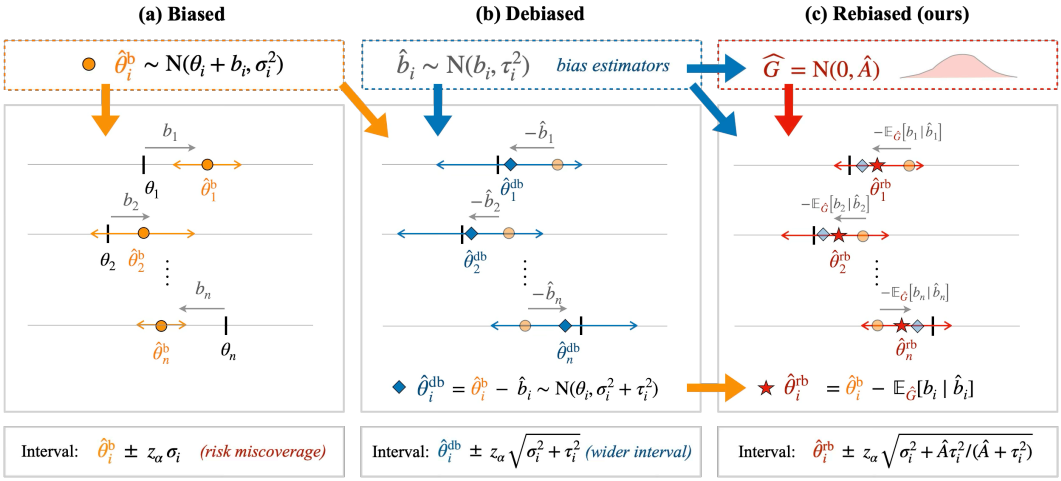


Figure 1: **From biased to debiased to rebiased.** Vertical black ticks mark the true θ_i. **(a)** Across n parallel tasks, each biased estimator $\hat{\theta}_i^{\mathrm{b}}$ is offset from its target θ_i by a draw $b_i \sim G$; the small variance σ_i^2 makes $\hat{\theta}_i^{\mathrm{b}}$ precise, but the resulting interval may undercover. **(b)** Subtracting an unbiased but noisy estimator $\hat{b}_i$ centers the estimator at θ_i 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.

In a range of settings, 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 distribution of 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 (i.e., to "rebias") 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 interval $\mathcal{I}_i^{\mathrm{rb}} = \hat{\theta}_i^{\mathrm{rb}} \pm z_\alpha(\sigma_i^2 + \widehat{A}\tau_i^2/\{\widehat{A} + \tau_i^2\})^{1/2}$, 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 $\hat{\theta}_i^{\mathrm{b}}$, $\widehat{A} = \infty$ recovers $\hat{\theta}_i^{\mathrm{db}}$, and intermediate $\widehat{A}$ rebiases, i.e., partially undoes the bias correction. As Fig. 1 illustrates, $\mathcal{I}_i^{\mathrm{rb}}$ is shorter than the Wald interval of the debiased estimator, with the gap governed by $\widehat{A}$. The interval $\mathcal{I}_i^{\mathrm{rb}}$ synthesizes three sources of information: the biased estimator $\hat{\theta}_i^{\mathrm{b}}$, 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 normal: $b_i \sim \mathrm{N}(0, A)$ with $A \geq 0$ (or uncentered, $b_i \sim \mathrm{N}(\mu, A)$). This normal component can capture, e.g., biases from unobserved confounding in observational studies, and A can be estimated from negative control studies. For non-control studies, one then reports the inflated interval $\hat{\theta}_i^{\mathrm{b}} \pm z_\alpha(\sigma_i^2 + \widehat{A})^{1/2}$, where $\sigma_i^2 = \mathrm{Var}_{\theta_i}[\hat{\theta}_i^{\mathrm{b}}]$. 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), and for inference with proxy outcomes (Wilkins-Reeves et al., 2026). 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 biases as drawn from a distribution G, but synthesize this distributional information with direct per-task estimators $\hat{b}_i$ to obtain shorter intervals, and allow for nonparametric G. 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 one unbiased experimental estimator, and Zhao et al. (2020) use 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 distribution 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), 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 $\hat{\theta}_i^{\mathrm{b}}$ 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,

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

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 G does not depend on $\theta_i, \sigma_i, \rho_i, \tau_i$. 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 distribution 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).

Our procedure below also admits a (fully) Bayes interpretation by placing a flat prior on θ_i, with θ_i independent of $b_i \sim G$.[1] We prefer the partially Bayes interpretation because it makes explicit that we assume shared structure in the biases, without imposing such structure on the θ_i.

[1] One can approximately interpret the flat prior as follows: start with a fixed smooth proper prior on θ_i, with θ_i independent of b_i. As the sample size within each task grows, rescale the bias distribution G so that biases remain of the same order as the standard errors. The prior on θ_i then becomes locally flat, while G remains informative (Hjort, 1993, Section 5E).

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 distribution G. When $G = \mathrm{N}(\mu, A)$, these expressions simplify, as we saw for $\mu = 0$ in §1; see Appendix A.1 for general μ. 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,[2]

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

We also interpret this procedure via the right-hand side of (4) as rebiasing the debiased estimator $\hat{\theta}_i^{\mathrm{db}}$. In the posterior for b_i, we condition only on $\hat{b}_i$. Explicitly, for any measurable set $B \subset \mathbb{R}$, we let

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

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 $b_i \mid \hat{b}_i$),

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

To derive the above conditional density, observe that $\hat{\theta}_i^{\mathrm{db}} - \theta_i = \{\hat{\theta}_i^{\mathrm{b}} - (\theta_i + b_i)\} + \{b_i - \hat{b}_i\}$ and so $\hat{\theta}_i^{\mathrm{db}} - \theta_i \mid b_i, \hat{b}_i \sim \mathrm{N}(b_i - \hat{b}_i, \sigma_i^2)$; then integrate against the posterior of b_i given $\hat{b}_i$.

Next, let us 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 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 distributions 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

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

et al., 2025) 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 fit or the NPMLE), we then compute the rebiased intervals using the plug-in principle, that is, pretending that $\widehat{G}$ is the true distribution 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 $\widehat{P}^{\mathrm{rb}}_{\widehat{G},i}$, obtained by replacing G with $\widehat{G}$ in (7). All the computations can be carried out readily; e.g., for the NPMLE we use the fact that it yields a discrete $\widehat{G}$ 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 access to $\hat{\theta}^{\mathrm{b}}_i$ 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 $\hat{\theta}^{\mathrm{db}}_i = \hat{\theta}^{\mathrm{ub}}_i$). Although formally equivalent, the two cases arise differently in practice. 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), combining exposed-only with difference estimates in sham-controlled experiments (Gelman and Vákár, 2021), and combining estimates from a narrow parametric model with those from a wider model that nests it (Hjort, 1993).

Prediction-powered inference (PPI). PPI (Angelopoulos et al., 2023, 2026) 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 $\theta_i = \mathbb{E}[Y_{ij}]$.[3] 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:

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

Setting $\hat{\theta}^{\mathrm{b}}_i = \hat{\theta}^{\mathrm{ML}}_i$ as the biased estimator, the PPI estimator is exactly the debiased estimator $\hat{\theta}^{\mathrm{db}}_i$ in our framework. Vanilla PPI is asymptotically dominated by the PPI++/PT (power tuned) estimator of Angelopoulos et al. (2026) that takes the form

$$\hat{\theta}^{\mathrm{PT}}_i = (1-\lambda^*_i)\bar{Y}_i + \lambda^*_i\hat{\theta}^{\mathrm{PPI}}_i \quad \text{for a choice of } \lambda^*_i.$$

To rebias PT, we treat $\hat{\theta}^{\mathrm{PT}}_i$ as the unbiased estimator $\hat{\theta}^{\mathrm{ub}}_i$ and $\hat{\theta}^{\mathrm{ML}}_i$ as the biased estimator $\hat{\theta}^{\mathrm{b}}_i$. Our experiments use PT throughout. We provide more details in Appendix B.1, including expressions for ρ_i, σ_i, τ_i in (3) and the plug-in estimates 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 $\hat{\theta}^{\mathrm{ML}}_i$, yielding shorter intervals; when biases are large, it stays close to PPI/PT. The prediction-powered adaptive shrinkage estimator of Li and Ignatiadis (2025) also shares information across tasks, but aims to reduce the mean squared error of PT rather than provide calibrated inference.

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.

[3] Our framework extends to general estimands beyond the mean via the predict-then-debias strategy described in Kluger et al. (2025); see also Miao and Lu (2024). Rebiasing also applies to batch active statistical inference (Zrnic and Candès, 2024), an extension of PPI in which labels are sampled with covariate-dependent probabilities (Appendix B.2).

The regression coefficient, denoted $\hat{\theta}_i^{\mathrm{b}}$ in our framework, and often termed the population-effect estimate in the literature, is biased as 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 i-th 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 with the 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.3 for details) and so we can apply our empirical Bayes rebiasing strategy that learns the bias distribution across SNPs.

5 Theoretical results

For our first 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). Later in this section, we also treat θ_i and b_i as both random (Theorem 5) and as both fixed (Theorem 6). We write $\mathbb{P}_{\boldsymbol{\theta},G}$ and $\mathbb{E}_{\boldsymbol{\theta},G}$ for probabilities and expectations taken jointly over the n tasks under the partially Bayes model with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ fixed. 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. 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 $\tilde{\sigma}_i^2 = \mathrm{Var}_{\theta_i}[\hat{\theta}_i^{\mathrm{db}}] = \sigma_i^2 + \tau_i^2 - 2\rho_i\sigma_i\tau_i$, $\gamma_i = \mathrm{Corr}_{\theta_i,b_i}[\hat{\theta}_i^{\mathrm{db}}, \hat{b}_i] = (\rho_i\sigma_i - \tau_i)/\tilde{\sigma}_i$, and $\check{\gamma} = \max_i |\gamma_i|$. 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 as 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).

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

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

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), $\check{\gamma} \in \{0.752, 0.891, 0.991\}$. 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

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

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,

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

A classical empirical Bayes analysis would instead estimate the distribution of θ_i using the NPMLE. The NPMLE's discreteness can lead to poor 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.

Hierarchical sampling model. A hierarchical Bayes model would typically posit $(\theta_i, b_i) \sim H$ for a (proper) joint distribution H. Under the additional assumption of independence between θ_i and b_i, we could apply Theorem 4 conditional on the θ_i. Our next theorem avoids this independence assumption, which may be implausible in some applications (Schuemie et al., 2018), by positing $(\theta_i, b_i) \sim H_i$, where each H_i has the same bias marginal, $b_i \sim G$, but can vary otherwise (that is, the marginal distribution of θ_i and the dependence of (θ_i, b_i) can vary with i).[4] We also assume that the tasks are independent and that, conditional on (θ_i, b_i), the observations $(\hat{\theta}_i^{\mathrm{b}}, \hat{b}_i)$ follow the normal distribution in (3). The coverage statement is averaged over both θ_i and b_i.

Theorem 5. Under the hierarchical sampling model above and Assumption 1, the coverage bound in Theorem 4 holds with $\mathbb{P}_{\boldsymbol{\theta},G}$ replaced by $\mathbb{P}_{\boldsymbol{H}}$, the probability law under $\boldsymbol{H} = (H_1, \ldots, H_n)$.

Compound decisions. We also obtain a purely frequentist guarantee with both the $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and $\boldsymbol{b} = (b_1, \ldots, b_n)$ fixed. The guarantee, which builds on compound decision theory (Robbins, 1951; Zhang, 2003), concerns average coverage across the θ_i (Wahba, 1983; Armstrong et al., 2022). We write $\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{b}}$ to denote the probability with $\boldsymbol{\theta}$ and $\boldsymbol{b}$ fixed and the pairs $(\hat{\theta}_i^{\mathrm{b}}, \hat{b}_i)$ jointly independent across tasks with the normal distribution in (3).

Theorem 6. Suppose that $\tilde{\sigma}_i^2, \tau_i^2, \gamma_i$ do not depend on i (and drop the subscript). Let $\check{\gamma} = |\gamma| \in (0, 1)$, and $0 < \tilde{\sigma}^2, \tau^2 < \infty$. Suppose that $|b_i| \le M$ for all i. Then there exists $C' = C'(M, \gamma, \tilde{\sigma}^2, \tau^2)$ with

$$\sup_{\alpha \in (0,1)} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{b}} \left[\theta_i \in \mathcal{I}^{\mathrm{rb}}_{\widehat{G},i}(1-\alpha) \right] - (1-\alpha) \right| \;\le\; C' \frac{(\log n)^{3/4}}{n^{(1-\check{\gamma}^2)/4}} \quad \text{for all } n \in \mathbb{N}_{\ge 2}.$$

Further robustness to assumptions. The preceding results assume exact normality in (3), known second moments, and independence across tasks. In practice, normality is an approximation motivated by the CLT, we replace σ_i, τ_i, ρ_i by plug-in estimates, and there may be cross-task dependence. Appendix C extends the coverage guarantees of Theorems 4, 5, and 6 to approximate normality, estimated second moments, and weak dependence across tasks. These extensions establish consistency without quantitative rates, and build on results of Chen et al. (2026). Our robustness results still assume that the distribution of b_i does not depend on $\theta_i, \sigma_i, \tau_i, \rho_i$ (for Theorem 4) or on σ_i, τ_i, ρ_i (for Theorem 5), or that all second moments are identical across tasks (for Theorem 6). Even with these robustness results, it is prudent to consider the suitability of the assumptions in practice. For each of our applications in §6, we provide diagnostics and a discussion of these assumptions in Appendix F.

6 Numerical results[5]

6.1 Experiments for prediction-powered inference

Intervals. Recall the PPI setup of §4. We compare our rebiased intervals against five 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); the power-tuned PPI *PT* interval of Angelopoulos et al. (2026) centered around the unbiased $\hat{\theta}_i^{\mathrm{PT}}$; the empirical Bayes confidence intervals of Armstrong et al. (2022), applied to the classical and PT point estimators to obtain the *EBCI Classical* and *EBCI PT* intervals, respectively. Our rebiased intervals (*RB*) are built on top of PT by fitting the bias distribution G on b_i, either as normal $\mathrm{N}(\mu, A)$ (suffix *Normal*) or nonparametrically (suffix *NPMLE*).

Metrics. We report the empirical *miscoverage rate* at nominal level $1-\alpha$, 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. Instead, we follow common practice in the PPI literature (Angelopoulos et al., 2023): we start with a fully labeled dataset, and treat the empirical mean over it as pseudo-ground truth for θ_i. (We acknowledge that this pseudo-ground truth is itself estimated and therefore introduces additional uncertainty.) Then, we randomly partition the data into $20/80$ labeled/unlabeled subsets by (artificially) masking labels in the "unlabeled" subset, imitating a situation with scarcer access to labels. We report all metrics averaged over $K = 200$ replications of this random partitioning, with Monte Carlo standard errors reported in Appendix E.

[4] This gives another interpretation of our method: we posit a common distribution G for the biases while leaving the distributions of the θ_i unrestricted across tasks.

[5] We provide code to reproduce all numerical results: https://github.com/wyling01/eb-rebiasing.

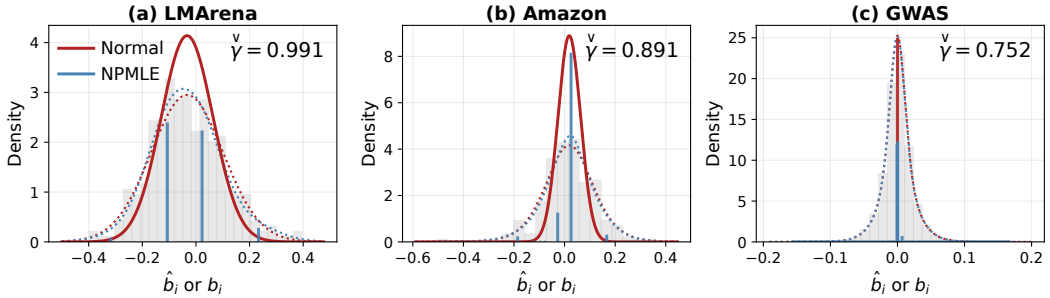


Figure 2: **Estimated bias distributions in the three applications.** Histogram of the estimator of bias $\hat{b}_i$ with the distribution estimated for the bias b_i (solid line) and the average marginal density of $\hat{b}_i$ implied by the estimated bias distribution (dotted line) overlaid. Two approaches to estimating G are shown: *Normal* and the *NPMLE*. The fitted estimated bias 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 average 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). The i-th 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 primary 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 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 E.1 for details). Fig. 2a shows the fitted normal and NPMLE distributions for b_i used to construct the rebiased intervals, together with their implied marginal densities on $\hat{b}_i$.

The predictor is biased enough that naïve intervals built from its predictions alone fail to cover the truth 65% of the time at $\alpha = 0.10$. *RB-Normal* can still effectively leverage this predictor: it has a much smaller miscoverage rate of 11% while reducing average interval width by approximately 25% relative to *Classical*, 23% relative to *PT*, and 14% relative to *EBCI PT*. *RB-NPMLE* further reduces width by approximately 29%, 27%, and 19%, respectively, but has a miscoverage rate of 15%; see Fig. 3a, which also shows results at other values of α. The more severe undercoverage of *RB-NPMLE*[6] is in line with Proposition 3: the lower bound there implies a slow estimation rate as $\hat{\gamma} \nearrow 1$, and here $\hat{\gamma} = 0.991$. The better performance of *RB-Normal* in the same regime is consistent with the remark following Theorem 2: when $\hat{\gamma}$ is close to one, committing to a well-motivated parametric family for G is a defensible alternative to fully nonparametric estimation.

Amazon Fine Food. Following Li and Ignatiadis (2025), we consider parallel PPI tasks using the *Amazon Fine Food Reviews* dataset (SNAP, 2014; McAuley and Leskovec, 2013), where the goal is to recover the average customer rating for each of $n = 200$ food 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. The predictor h is fine-tuned BERT (Devlin et al., 2019) (with fine-tuning on a disjoint pool of reviews). This application is meant to mimic situations in which annotation of text by trained human coders can be supplemented by ML predictions (Fan et al., 2024; Mozer and Miratrix, 2025); here, the star rating plays the role of the human-coded label. We leave additional details of our dataset and predictor to Appendix E.2.

On the Amazon benchmark, the predictor is fairly informative; this is evident in the concentrated estimated bias distributions (Fig. 2b). Intervals built from the prediction mean alone are about half the width of the *Classical* interval, but they exhibit undercoverage (11.3% miscoverage at $\alpha = 0.1$). Even so, our rebiasing approach further improves coverage and delivers shorter intervals than competing methods (see Fig. 3b). At $\alpha = 0.1$, *RB-NPMLE* attains a miscoverage rate of 3.7% with width 0.286, roughly 48% shorter than *Classical*, 17% shorter than *PT*, and 15% shorter than *EBCI PT*.

[6]The miscoverage of *RB-NPMLE* may also reflect departures from our assumptions. As discussed in Appendix F, the small labeled samples of binary outcomes per task may lead to a poor task-level normal approximation and imprecise plug-in second-moment estimates. Fig. S3a also suggests that the distribution of b_i may vary with the task-specific second moments.

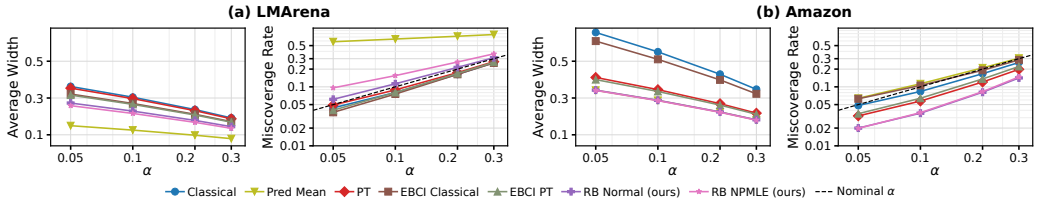


Figure 3: **Interval width and miscoverage rate for LM Arena and Amazon Fine Food.** Average width (left) and miscoverage rate (right) for two PPI applications. **(a)** Pairwise LLM win-rate tasks ($n = 298$) in LM Arena dataset. Both rebiasing methods shorten the intervals, but *RB-NPMLE* has higher miscoverage. **(b)** Product rating problems ($n = 200$) in the Amazon dataset. Both *RB-Normal* and *RB-NPMLE* produce the shortest intervals while attaining miscoverage rates below the nominal level. Standard errors are reported in Tables S1, S2.

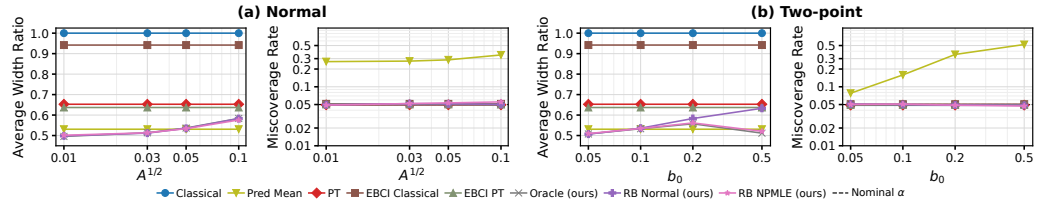


Figure 4: **Interval width ratio and miscoverage rate in the synthetic Amazon simulations.** Average width ratio (left) and miscoverage rate (right) for $n = 200$ simulated tasks in the synthetic Amazon dataset under two choices of bias distribution G: **(a)** normal and **(b)** two-point. Across both settings, our rebiased methods yield substantially shorter intervals with near-nominal miscoverage. *RB-Normal* and *RB-NPMLE* are nearly indistinguishable under the well-specified normal bias distribution, while *RB-NPMLE* is more robust under the two-point distribution, where the *RB-Normal* is misspecified and therefore loses efficiency. Standard errors are reported in Tables S3 and S4.

Synthetic data using covariance structure in Amazon data. To further evaluate our rebiased intervals in a setting with ground truth, we conduct simulation studies that generate synthetic estimators using the estimated task-level second moments from the Amazon data, while independently generating the underlying θ_i and b_i. We generate biases b_i from two types of bias distribution G: normal distribution $\mathrm{N}(-0.1, A)$ and two-point distribution $(\delta_0 + \delta_{b_0})/2$, 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 E.3. As a benchmark, we also include an *oracle* variant that plugs in the true data-generating distribution G 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 normal model is already sufficient, so the extra flexibility of NPMLE yields no efficiency gain, but no efficiency loss either. Under the two-point distribution, the effect of model misspecification becomes apparent. 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 a discrete, nonnormal bias distribution by a normal bias distribution. The *RB-NPMLE* interval is more adaptive in this setting: when b_0 is small, the bias distribution is nearly degenerate, and both rebiased methods perform well; as the two-point structure becomes more pronounced, *RB-NPMLE* better accommodates this structure, producing intervals closer in width to the oracle while maintaining coverage.

We additionally vary the dispersion of the primary parameter by generating $\theta_i \sim \mathrm{N}(4, \sigma_0^2)$ with different values of σ_0, while fixing the bias distribution, with results shown in Fig. S1. When primary parameters are highly concentrated, EBCI exploits the strong across-task similarity to produce shorter intervals than our rebiasing procedures. As θ_i dispersion increases, however, this advantage quickly diminishes, with EBCI intervals approaching their non-Bayes counterparts, while the widths and coverage of our procedures remain essentially unchanged. This highlights the complementary regimes of the two approaches: EBCI borrows information across θ_i, whereas our partially Bayes procedure borrows only through the biases and is therefore insensitive to the primary parameter heterogeneity.

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 E.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 ± SD.

	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

6.2 Family-based GWAS

Height GWAS. We illustrate rebiased p-values for SNP discovery in family-based GWAS. We analyze human height direct-effect estimates $\hat{\theta}_i^{\mathrm{ub}}$ and population-effect estimates $\hat{\theta}_i^{\mathrm{b}}$ for $n = 572{,}911$ SNPs from Guan et al. (2025), estimating the bias distribution G via NPMLE and via the normal parametric model. From each estimated bias distribution, we compute empirical Bayes rebiasing p-values and apply Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) at a nominal false discovery rate (FDR) of 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. Since we lack individual-level data, we cannot reuse the PPI evaluation strategy. Instead we compare discoveries against a separate GWAS with a similar family design (Howe et al., 2022) as preliminary supporting evidence.

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 bias distributions concentrate near zero (Fig. 2c), consistent with existing biological evidence specific to height (Appendix E.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 E.4 for further discussion).

Synthetic GWAS study. In Appendix E.5, we conduct a simulation study to check type-I error control of our proposed methods in a setting where we know the ground truth. In our simulation design, we mimic realistic genomic dependence structure and consider a setting with zero direct genetic effects and weak parental effects. We find that RB-NPMLE p-values used alongside BH are calibrated, while RB-Normal has slightly inflated type-I error.

7 Conclusion

For point estimation, forfeiting unbiasedness in exchange for variance savings is an uncontroversial tradeoff. For interval inference, it is common for the analyst to trust a biased estimator outright, hoping the bias is negligible, or to fully debias it and absorb the associated variance cost. 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 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 and Disclosure of Funding

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). The authors declare no conflicts of interest.

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. *The Annals of Applied Statistics*, 20(3):2235–2252, 2026.

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.

T. Berisa and J. K. Pickrell. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, 32(2):283–285, Jan. 2016.

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, 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 *Advances in Neural Information Processing Systems*, volume 37, pages 113096–113133, 2024.

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

B. Chen, Z. Zhou, L. Peng, and Z. Wang. Balanced active inference. *Advances in Neural Information Processing Systems*, 38:51742–51763, 2025.

J. Chen, N. Deb, and N. Ignatiadis. Normal approximations in nonparametric empirical Bayes. *arXiv preprint arXiv:2605.31599*, 2026.

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*, volume 235, pages 8359–8388. PMLR, 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.

A. Dimitromanolakis, J. Xu, A. Krol, and L. Briollais. sim1000G: a user-friendly genetic variant simulator in R for unrelated individuals and family-based designs. *BMC Bioinformatics*, 20(1):26, 2019.

D. L. Donoho. Statistical estimation and optimal recovery. *The Annals of Statistics*, 22(1):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 preprint*, 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.

S. Ghatan, J. de Vries, J.-B. Pingault, V. W. Jaddoe, C. Cecil, J. F. Felix, F. Rivadeneira, and C. Medina-Gomez. Genetic nurture: estimating the direct genetic effects of pediatric anthropometric traits. *Human Molecular Genetics*, 34(20):1744–1752, 2025.

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.

N. L. Hjort. Estimation in moderately misspecified models. Statistical Research Report, Department of Mathematics, University of Oslo and Department of Statistics, University of Oxford, 1993.

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*, volume 216, pages 574–578. 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 https://rkoenker.github.io/www/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):1–26, 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.

P. Li, T. Zrnic, and E. Candès. Robust sampling for active statistical inference. In *Advances in Neural Information Processing Systems*, volume 38, pages 68727–68756, 2025.

S. Li and N. Ignatiadis. Prediction-powered adaptive shrinkage estimation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 34836–34875. PMLR, 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.

J. McAuley and J. Leskovec. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, pages 897–908. ACM, 2013.

J. Miao and Q. Lu. Task-agnostic machine-learning-assisted inference. In *Advances in Neural Information Processing Systems*, volume 37, pages 106162–106189, 2024.

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. Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, volume 2, pages 131–149. University of California Press, 1951.

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. R. Rosenman, G. Basse, A. B. Owen, and M. Baiocchi. Combining observational and experimental datasets using shrinkage estimators. *Biometrics*, 79(4):2961–2973, 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.

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.

SNAP. Amazon fine food reviews. Kaggle Dataset, 2014. URL https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews/data. Stanford Network Analysis Project.

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*, 87(1):1–32, 2025.

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

T. Tan, H. Jayashankar, J. Guan, S. M. Nehzati, M. Mir, M. Bennett, E. Agerbo, R. Ahlskog, V. Pinto de Andrade Anapaz, B. O. Åsvold, S. Benonisdottir, L. Bhatta, D. I. Boomsma, B. Brumpton, A. Campbell, C. F. Chabris, R. Cheesman, Z. Chen, China Kadoorie Biobank Collaborative Group, E. de Geus, E. A. Ehli, A. G. Elnahas, Estonian Biobank Research Team, Finngen, A. Ganna, A. Giannelis, L. Hakaste, A. F. Hansen, A. Havdahl, C. Hayward, J.-J. Hottenga, M. A. Houmark, K. Hveem, J. Kaprio, A. Langhammer, A. Latvala, J. J. Lee, M. Lehtovirta, L. Li, LifeLines Cohort Study, K. Lin, R. K. Linnér, S. Lombardi, N. G. Martin, M. McGue, S. E. Medland, A. Metspalu, B. L. Mitchell, G. Ni, I. M. Nolte, M. T. Oetjens, S. Oskarsson, T. Palviainen, R. B. Prasad, A. Reigo, K. Reis, J. Sidorenko, K. Silventoinen, H. Snieder, T. Tuomi, B. J. Vilhjálmsson, R. G. Walters, E. A. Willoughby, B. S. Winsvold, E. Ystrom, J. Flint, L. Yengo, P. M. Visscher, A. Kong, E. M. Tucker-Drob, R. Border, D. Cesarini, P. Turley, A. Okbay, D. J. Benjamin, and A. S. Young. Family-GWAS reveals effects of environment and mating on genetic associations. *medRxiv*, Oct. 2025.

G. Wahba. Bayesian "confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society: Series B*, 45(1):133–150, 1983.

S. Wilkins-Reeves, A. N. M. Darmon, and D. Sinha. Estimate level adjustment for inference with proxies under random distribution shifts. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, 2026.

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.

C.-H. Zhang. Compound decision theory and empirical Bayes methods: Invited paper. *The Annals of Statistics*, 31(2):379–390, 2003.

C.-H. Zhang. Generalized maximum likelihood estimation of normal mixture densities. *Statistica Sinica*, 19(3):1297–1318, 2009.

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.

T. Zrnic and E. Candès. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 62993–63010. PMLR, 2024.

♣ Appendix: Table of Contents

A. Details on methods (Section 3) . . . . . Page 16

B. Details on applications (Section 4) . . . . . Page 17

C. Robustness results (Section 5) . . . . . Page 19

D. Proof of theoretical results (Section 5) . . . . . Page 21

E. Details on numerical studies (Section 6) . . . . . Page 50

F. Diagnostics for the theoretical assumptions . . . . . Page 58

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

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

The unknown parameters (μ, A) from 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 Rebiasing for general ρ

Consider a more generic setting of this problem. We now allow nonzero conditional correlation ρ_i between $(\hat{\theta}_i^{\mathrm{b}}, \hat{b}_i)$ given b_i 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 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 $\mathcal{I}_{G,i}^{\mathrm{rb}}(1-\alpha) = [\hat{\theta}_i^{\mathrm{db}} - q_{G,i,1-\alpha/2},\ \hat{\theta}_i^{\mathrm{db}} - q_{G,i,\alpha/2}]$.

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

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

After estimating the 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 define the maximum marginal-likelihood estimator by

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

For implementation, we parameterize $A = \exp(\eta)$ and jointly optimize over $(\mu, \eta) \in \mathbb{R}^2$ using BFGS. The numerical optimization therefore searches over the interior $A > 0$. This is equivalent to the constrained optimization whenever the maximizer satisfies $A > 0$, and approximates the boundary $A = 0$ as $\eta \to -\infty$.

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 programming 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. (2026) 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 $\hat{\theta}_i^{\text{PPI}}$ ($\lambda_i = 1$). 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

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

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 $\hat{\theta}_i^{\text{PT}} := \hat{\theta}_{i,\lambda_i^*}$.

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 $\hat{\theta}_i^{\text{ML}}$. Following §4, we therefore identify

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

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 $\hat{\theta}_i^{\text{b}}$. The pair $(\hat{\theta}_i^{\text{b}}, \hat{b}_i)$ thus fits the model in (3) with $\hat{\theta}_i^{\text{db}} = \hat{\theta}_i^{\text{b}} - \hat{b}_i = \hat{\theta}_i^{\text{PT}}$, and the rebiased estimator and intervals from §3 apply directly. Setting $\lambda_i^* = 1$ recovers the standard vanilla-PPI debiasing ($\hat{b}_i = \bar{Z}_i^h - \bar{Y}_i$); 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

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

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

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

When $\lambda_i^* = 0$, $\hat{\theta}_i^{\text{PT}}$ collapses to the classical mean $\bar{Y}_i$. When $\lambda_i^* = 1$, $\hat{\theta}_i^{\text{PT}}$ coincides with vanilla PPI and $\hat{b}_i = \bar{Z}_i^h - \bar{Y}_i$; the unlabeled-only $\hat{\theta}_i^{\text{b}}$ 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 the unknown second moments (v_i^2, w_i^2, c_i). For the tuning coefficient in (S7), we replace c_i by the sample covariance between Y_{ij} and $h(X_{ij})$ in the labeled sample, and v_i^2 by the sample variance of the predictions pooled across the labeled and unlabeled samples. We clip the resulting plug-in estimate to $[0, 1]$ to obtain the feasible coefficient $\hat{\lambda}_i^*$. For estimating the covariance parameters in (3), we estimate the terms v_i^2/M_i and v_i^2/m_i using the unlabeled- and labeled-sample prediction variances, respectively, while w_i^2 and c_i are estimated from the labeled sample. Substituting these estimates into (S9)–(S11) yields $(\hat{\sigma}_i^2, \hat{\tau}_i^2, \hat{\rho}_i)$. Following standard practice in PPI (Angelopoulos et al., 2026), we treat these plug-ins as known when forming Wald and rebiased intervals; see Appendix F for a discussion of whether this plug-in approximation is reasonable in each application.

B.2 Active statistical inference

A closely related setting to PPI is active statistical inference, which aims to improve statistical efficiency under a limited labeling budget by selectively acquiring informative labels. Zrnic and Candès (2024) consider a setting where one starts with unlabeled covariates $\{X_{ij}\}_{j=1}^{K_i}$, then based on X_{ij}, the analyst observes Y_{ij} with probability $\pi(X_{ij})$. Effectively, one stochastically prioritizes

measuring labels that will be more informative. Let $\xi_{ij} \in \{0, 1\}$ be the indicator of whether Y_{ij} is observed. Given a further predictive model $h(\cdot)$ as in PPI, their batch active mean estimator takes the following form:

$$\widehat{\theta}_i^{\mathrm{AS}} = \underbrace{\frac{1}{K_i}\sum_{j=1}^{K_i} h(X_{ij})}_{\hat{\theta}_i^{\mathrm{b}}} + \underbrace{\frac{1}{K_i}\sum_{j=1}^{K_i}(Y_{ij} - h(X_{ij}))\frac{\xi_{ij}}{\pi(X_{ij})}}_{-\hat{b}_i}.$$

The first term is a prediction-based, potentially biased estimator, while the second estimates its prediction bias correction. Thus, the batch active mean estimator can naturally be incorporated into our empirical Bayes rebiasing framework to borrow information across related tasks.

Recent work has further explored robust and balanced sampling strategies to improve the statistical efficiency of active inference (Li et al., 2025; Chen et al., 2025). More generally, combining active inference with empirical Bayes rebiasing raises additional challenges. In particular, with a limited labeling budget across many related tasks, one needs to jointly determine which observations to label within each task and how to allocate the budget across tasks. Developing such an adaptive labeling and budget-allocation strategy together with our empirical Bayes framework represents an interesting direction for future work.

B.3 Family-based GWAS

The summary statistics provide the regression coefficient estimates $(\hat{\theta}_i^{\mathrm{ub}}, \hat{b}_i)$, with variances $\tilde{\sigma}_i^2$ and τ_i^2 and correlation γ_i obtained from the estimated covariance matrix of the coefficient estimates. 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 Robustness results (Section 5)

We extend the coverage guarantees of Theorems 4, 5, and 6 to allow for approximate normality, estimated second moments, and weak dependence across tasks. Following the robustness framework of Chen et al. (2026), throughout this section we impose the bounded-support condition $b_i \in [-M, M]$; in the exchangeable setting this is encoded by compactly supported bias distributions, $G \in \mathcal{P}([-M, M])$ (for the robustness extensions of Theorems 4 and 5).

Throughout this section, we work in a triangular-array asymptotic framework, allowing the task-specific parameters, sampling laws, and common bias distribution G (in the partially Bayes and hierarchical settings) to depend on n, while maintaining $G \in \mathcal{P}([-M, M])$ for a fixed $M < \infty$. More precisely, for each n, we have n tasks with primary parameters $\theta_{1,n}, \ldots, \theta_{n,n}$, biased estimators $\hat{\theta}_{1,n}^{\mathrm{b}}, \ldots, \hat{\theta}_{n,n}^{\mathrm{b}}$, biases $b_{1,n}, \ldots, b_{n,n}$ (that are drawn from a distribution $G^{(n)}$ in the partially Bayes and hierarchical settings), bias estimators $\hat{b}_{1,n}, \ldots, \hat{b}_{n,n}$, and true and estimated second moments. In words, for each n, we imagine that we receive n new tasks. This allows us to study sequences of models in which departures from normality diminish on average across tasks. We refer to Chen et al. (2026) for detailed discussion and for examples where these asymptotics are applicable with a particular emphasis on asymptotic regimes that both make the normality assumption more plausible and still benefit from empirical Bayes shrinkage. We suppress the dependence on n in the notation and e.g., write $\theta_1, \ldots, \theta_n$ instead of $\theta_{1,n}, \ldots, \theta_{n,n}$.

To formulate the robustness conditions uniformly across three settings, we use $\mathbb{P}_\star$, $\mathbb{E}_\star$, and $\mathcal{L}_\star$ to denote probability, expectation, and distribution under the sampling law specified in each theorem below. For each task i, let

$$\Sigma_i := \begin{pmatrix} \tilde{\sigma}_i^2 & \gamma_i\tilde{\sigma}_i\tau_i \\ \gamma_i\tilde{\sigma}_i\tau_i & \tau_i^2 \end{pmatrix}$$

denote the reference sampling covariance matrix for $(\hat{\theta}_i^{\mathrm{db}} - \theta_i, \hat{b}_i - b_i)$, and let $\widehat{\Sigma}_i$ be its estimator. We take $\widehat{G}$ to be a global maximizer of the normal working marginal likelihood over $\mathcal{P}([-M, M])$, using the estimated second moments $\widehat{\Sigma}_{i,22}$.

To quantify departures from the normal working model underlying Theorems 4 to 6, we compare the actual bivariate experiment of $(\hat{\theta}_i^{\mathrm{db}} - \theta_i, \hat{b}_i)$ with its normal counterpart, $\mathrm{N}_2((0, b_i)^\top, \Sigma_i)$, through the averaged Wasserstein-2 discrepancy

$$\mathcal{W}_n^2 := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\star\left[W_2^2\left(\mathcal{L}_\star\left(\begin{pmatrix}\hat{\theta}_i^{\mathrm{db}} - \theta_i \\ \hat{b}_i\end{pmatrix}\middle| b_i\right), \mathrm{N}_2\left(\begin{pmatrix}0 \\ b_i\end{pmatrix}, \Sigma_i\right)\right)\right].$$

where W_2 denotes the Wasserstein-2 distance on $\mathbb{R}^2$.

Assumption 7 (Approximate normality)**.** We assume that $\mathcal{W}_n \to 0$.

Assumption 8 (Estimated second moments)**.** We assume that

$$\frac{1}{n}\sum_{i=1}^{n} \|\widehat{\Sigma}_i - \Sigma_i\| \xrightarrow{\mathbb{P}_\star} 0.$$

We further assume that there exist constants $0 < \underline{\lambda} < \overline{\lambda} < \infty$ such that $\underline{\lambda} I_2 \preceq \Sigma_i, \widehat{\Sigma}_i \preceq \overline{\lambda} I_2$ for all $1 \le i \le n$.

The bounds on $\widehat{\Sigma}_i$ are technical regularity conditions and can, in principle, be enforced by eigenvalue clipping. Our numerical implementation uses the unmodified plug-in estimates.

Finally, we replace cross-task independence with the following weak empirical averaging condition.

Assumption 9 (Approximate independence)**.** For every fixed triangular sequence of functions $f_i : \mathbb{R} \to \mathbb{R}$, $1 \le i \le n$, such that $|f_i'(0)| \le 1$ and $\max_i \|f_i''\|_\infty \le 1$, we assume that

$$\frac{1}{n}\sum_{i=1}^{n}\Big(f_i(\hat{b}_i) - \mathbb{E}_\star[f_i(\hat{b}_i)]\Big) \xrightarrow{\mathbb{P}_\star} 0.$$

We first consider the partially Bayes setting of Theorem 4, where θ_i is fixed and $b_i \overset{\mathrm{iid}}{\sim} G$.

Theorem 10 (Robustness extension for Theorem 4)**.** Suppose that Assumptions 7, 8, and 9 hold with $\mathbb{P}_\star = \mathbb{P}_{\boldsymbol{\theta},G}$, $\mathbb{E}_\star = \mathbb{E}_{\boldsymbol{\theta},G}$, and $\mathcal{L}_\star = \mathcal{L}_{\boldsymbol{\theta},G}$. Then for every fixed $\alpha \in (0,1)$,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{P}_{\boldsymbol{\theta},G}\left[\theta_i \in \mathcal{I}^{\mathrm{rb}}_{\widehat{G},i}(1-\alpha)\right] - (1-\alpha)\right| = 0.$$

We next consider the hierarchical setting of Theorem 5, where $(\theta_i, b_i) \sim H_i$, with a common marginal bias distribution $b_i \sim G$. In the definition of $\mathcal{W}_n$, the conditional law marginalizes over θ_i given b_i, while the outer expectation is taken over $b_i \sim G$, thereby allowing dependence between the primary parameter and the latent bias.

Theorem 11 (Robustness extension for Theorem 5)**.** Suppose that Assumptions 7, 8, and 9 hold with $\mathbb{P}_\star = \mathbb{P}_{\boldsymbol{H}}$, $\mathbb{E}_\star = \mathbb{E}_{\boldsymbol{H}}$, and $\mathcal{L}_\star = \mathcal{L}_{\boldsymbol{H}}$. Then for every fixed $\alpha \in (0,1)$,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{P}_{\boldsymbol{H}}\left[\theta_i \in \mathcal{I}^{\mathrm{rb}}_{\widehat{G},i}(1-\alpha)\right] - (1-\alpha)\right| = 0.$$

Finally, we consider the compound decision setting of Theorem 6, where (θ_i, b_i) are fixed, $b_i \in [-M, M]$, and $\Sigma_i = \Sigma$ for all i. Since b_i is fixed, the conditional law in $\mathcal{W}_n$ reduces to the unconditional law, and the outer expectation can be omitted.

Theorem 12 (Robustness extension for Theorem 6)**.** Suppose that Assumptions 7, 8, and 9 hold with $\mathbb{P}_\star = \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{b}}$, $\mathbb{E}_\star = \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{b}}$, and $\mathcal{L}_\star = \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{b}}$. Then, for every fixed $\alpha \in (0,1)$,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{b}}\left[\theta_i \in \mathcal{I}^{\mathrm{rb}}_{\widehat{G},i}(1-\alpha)\right] = 1-\alpha.$$

D 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 distribution 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

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

Throughout the proofs in Appendix D.1–D.6, we may assume without loss of generality that $\mathbb{E}_G[b] = 0$. To see this, let $m = \mathbb{E}_G[b]$ and define

$$b_i' = b_i - m, \qquad \hat{b}_i' = \hat{b}_i - m, \qquad (\hat{\theta}_i^{\mathrm{b}})' = \hat{\theta}_i^{\mathrm{b}} - m.$$

Let G' denote the distribution of b_i'. Then G' remains in $\mathcal{G}_\Gamma$ and satisfies $\mathbb{E}_{G'}[b] = 0$. The debiased estimator $\hat{\theta}_i^{\mathrm{db}}$ and the covariance parameters are unchanged under this translation. Moreover, for any mixing distribution H, letting H' denote its translation by $-m$, we have

$$f_{H'}^{(i)}(l - m) = f_H^{(i)}(l), \qquad F_{H',i}(z \mid l - m) = F_{H,i}(z \mid l).$$

Consequently, the marginal likelihood and its maximizers transform accordingly, while the conditional CDFs, resulting p-values, and relevant Hellinger distances are invariant under the corresponding translations. Since any selected NPMLE for the original data corresponds to a valid maximizer for the centered data, it suffices to establish the following bounds under $\mathbb{E}_G[b] = 0$.

D.1 Preliminary lemmata and propositions

We state the following results from Jiang (2020).

Lemma 13 (Lemma 4 of Jiang (2020))**.** For $0 < \varepsilon < 1/\sqrt{2\pi}$, $U > 0$,

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

Here, $\mathcal{N}(\varepsilon, \mathcal{F}, d)$ denotes the ε-covering number of $\mathcal{F}$ with respect to distance metric d.

Lemma 14 (Theorem 4 of Jiang (2020))**.** Under Assumption 1, fix $c_0 \ge 2$, then there exists a constant C_0 depending only on $c_0, \Gamma, \underline{\sigma}, \bar{\sigma}$ 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 14. 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 $\mu_p(G) \le c\sqrt{\Gamma p}$ for all $p \ge 1$, where $c > 0$ is a universal constant.

For $n \ge 3$, we set $p = \log n \ge 1$ in (S13). For $A_n := \left[n^{1/p}\sqrt{\log n} \cdot \mu_p(G)\right]^{p/(2p+2)}$, we have

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

Then

$$A_n \le \max\left\{1, \sqrt{ec\sqrt{\Gamma}\log n}\right\} \le C_1(\Gamma)\sqrt{\log n},$$

where one may take $C_1(\Gamma) = \max\left\{1, \sqrt{ec}\,\Gamma^{1/4}\right\}$. Plugging back to (S13),

$$\varepsilon_n(n, G, \log n) \le C_2(\Gamma)\frac{\log n}{\sqrt{n}}, \qquad C_2(\Gamma) = \max\left\{\sqrt{2}, C_1(\Gamma)\right\}, \qquad n \ge 3$$

Since $\overline{d}$ is uniformly bounded, the remaining case $n = 2$ can be covered by enlarging the constant.

Therefore, with Theorem 4 of Jiang (2020) and a fixed $c_0 \ge 2$, there exists C_0 depending only on $c_0, \Gamma, \underline{\sigma}, \bar{\sigma}$ 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 15. 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,

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

with $C(\Sigma_i) := (2\pi|\Sigma_i|\Omega_{i,22})^{-1/2}$. Denote $p_G(u, l) = \int p_i(u, l \mid b_i)\,\mathrm{d}G(b_i)$ 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

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

□

D.2 Proof of Theorem 2

Since the map $x \mapsto 2\min\{x, 1-x\}$ is 2-Lipschitz on $[0,1]$, we have that,

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

Therefore, it suffices to prove

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

Let $\mathcal{A}$ be the event with the constant C_0 in Lemma 14:

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

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

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

By Lemma 14, $\mathbb{P}_G[\mathcal{A}^c] \le \exp(-c_0 \log n)$ with $c_0 \ge 2$. We consider the following two lemmas that bound terms I and II.

Lemma 16. 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 17. 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 16, we consider the following covering lemma.

Lemma 18. 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

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

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 $l \in \mathbb{R}$,

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

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 a normal tail bound and $\Sigma_{i,11} \le C$, we have

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

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 normal mixture densities

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

has $2n$ normal 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 13 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 $j \in \{1, \dots, J\}$ satisfying

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

for a sufficiently small constant $c > 0$.

We claim that the same $\{G_j\}$ also approximates $N^{(i)}_{\widetilde{G}}$ uniformly on $\mathbb{R} \times [-U, U]$. Indeed, by Proposition 15, for $|l| \le U$,

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

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 16. 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 14. By applying Lemma 18 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

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

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 $\mathcal{A}$,

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

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

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

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

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

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

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

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

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

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

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

where the second term is justified by

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

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

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

It remains to bound S_{ij}. By Proposition 15,

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

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,

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

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,

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

For I_{ij1},

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

For I_{ij2}, first noticing that

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

then applying the normal tail bound,

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

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

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

By using the inequality $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ 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 $j \in \{1, \ldots, J\}$,

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

Combining (S21) and (S24), we get

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

Also, combined with the Hoeffding term,

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

Plugging these bounds into (S19),

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

This completes the proof.

□

Proof of Lemma 17. 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 14. 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 13 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 13 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

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

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

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

For the remaining term,

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

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 13 and Lemma 14, 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 results into (S25), we have

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

□

D.3 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 15 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 $\cos(As\omega) - 1 \le 0$ and $\Gamma > A$. Hence $G_1 \in \mathcal{G}_\Gamma$.

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

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

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

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

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 normal 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 $l \in I$,

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

Here the boundary terms at $-\infty$ vanish because W_l and W_l' have normal 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,

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

Therefore,

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

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: $\mathcal{E}_{n,i} = \{(l_1, \ldots, l_n) \in \mathbb{R}^n : l_i \in S_n\}$. 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 $(\hat{b}_1, \ldots, \hat{b}_n)$. 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 two-point argument,

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

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

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

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

D.4 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

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

By definition of $F_{G,i}(\cdot \mid \hat{b}_i)$, 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

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

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

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

Hence

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

Then by taking the expectation on both sides, we have

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

Averaging over i, we get

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

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:

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

D.5 Proof of Theorem 5

The argument follows the proof of Theorem 4. Under the generalized model, conditional on (θ_i, b_i), the joint distribution of $(\hat{\theta}_i^{\mathrm{db}} - \theta_i, \hat{b}_i)$ depends on (θ_i, b_i) only through b_i. Since the marginal distribution of b_i under H_i is the common G, the marginal joint distribution of $(\hat{\theta}_i^{\mathrm{db}} - \theta_i, \hat{b}_i)$ is identical to that under the original partially Bayes model.

Consequently, the oracle conditional probability integral transform remains uniform given $\mathcal{L}_n = \sigma(\hat{b}_1, \ldots, \hat{b}_n)$. Moreover, the joint distribution of the bias estimates, and hence the distribution of Δ_i defined in the proof of Theorem 4, is unchanged. Applying the same argument and the conditional CDF estimation bound established in the proof of Theorem 2 yields

$$\sup_{\alpha\in(0,1)} \frac{1}{n}\sum_{i=1}^{n} \left| \mathbb{P}_{\boldsymbol{H}}\left[\theta_i \in \mathcal{I}^{\mathrm{rb}}_{\widehat{G},i}(1-\alpha) \right] - (1-\alpha) \right| \le C'\frac{(\log n)^{3/2}}{n^{(1-\tilde{\gamma}^2)/2}},$$

as claimed.

D.6 Proof of Theorem 6

For notational simplicity, we denote

$$\Sigma = (\Sigma_{jq}) := \begin{pmatrix} \widetilde{\sigma}^2 & \gamma\widetilde{\sigma}\tau \\ \gamma\widetilde{\sigma}\tau & \tau^2 \end{pmatrix}, \quad \Omega = (\Omega_{jq}) := \Sigma^{-1}.$$

Define the empirical distribution of the fixed biases $\boldsymbol{b} = (b_1, ..., b_n)$ by

$$G_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{b_i}.$$

Under the homogeneous model, we define

$$d(G_n, \widehat{G}) := \mathfrak{H}\left(f_{G_n}(\cdot; \Sigma_{22}), f_{\widehat{G}}(\cdot; \Sigma_{22})\right).$$

Lemma 19 (Theorem 1 of Zhang (2009))**.** Suppose that $\hat{b}_1, ..., \hat{b}_n$ are independent, $\hat{b}_i \sim \mathrm{N}(b_i, \tau^2)$, and $|b_i| \le M$. Fix $c_0 \ge 2$, then there exists a constant C_0 depending only on M, τ, c_0 such that,

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

Proof. We verify the rate obtained from Theorem 1 of Zhang (2009). In the notation of that theorem, the observations X_i, conditional means θ_i, and common standard deviation σ correspond to $\hat{b}_i$, b_i, and τ, respectively. Since the $\hat{b}_i$ are independent with $\hat{b}_i \sim \mathrm{N}(b_i, \tau^2)$, the sampling model satisfies (2.1) of Zhang (2009), with the average mixing distribution corresponding to G_n. Since $|b_i| \le M$ for all i, we have $G_n([-M, M]) = 1$. Thus, applying the bounded-support case of Theorem 1 of Zhang (2009), with $M_n = M/\tau$ after standardization, yields

$$\epsilon_n \lesssim n^{-1/2}(\log n)^{3/4}\left\{\left(\frac{M}{\tau}\right)^{1/2} \vee (\log n)^{1/4}\right\} \le C_1(M, \tau)\frac{\log n}{\sqrt{n}}.$$

Therefore, by the concentration inequality (2.11) of Zhang (2009), for any fixed $c_0 \ge 2$, choosing $t \ge \max\{t_*, \sqrt{c_0}\}$, there exist a constant C_0 that depends only on M, τ, c_0 such that

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

The difference in the normalization of the Hellinger distance is absorbed into the constant C_0. □

Proof of Theorem 6. As in the proof of Theorem 2, we should first try to bound

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{b}}[\Delta_i], \quad \Delta_i := \sup_{z \in \mathbb{R}} \left|F_{\widehat{G}}(z \mid \hat{b}_i) - F_{G_n}(z \mid \hat{b}_i)\right|.$$

Following the definition in Proposition 15, under the homogeneous covariance parameters, for any probability measure $\widetilde{G}$,

$$F_{\widetilde{G}}(z \mid l) := \frac{N_{\widetilde{G}}(z, l)}{D_{\widetilde{G}}(l)}, \qquad D_{\widetilde{G}}(l) := f_{\widetilde{G}}(l; \Sigma_{22}),$$

where

$$N_{\widetilde{G}}(z, l) = C(\Sigma)\int_{-\infty}^{z} \exp\left(-\frac{u^2}{2\Sigma_{11}}\right) f_{\widetilde{G}}\left(\frac{\Omega_{12}}{\Omega_{22}}u + l; \Omega_{22}^{-1}\right) \mathrm{d}u, \tag{S26}$$

where $C(\Sigma) = (2\pi|\Sigma|\Omega_{22})^{-1/2}$. Also note that $0 \le N_{\widetilde{G}}(z, l) \le D_{\widetilde{G}}(l)$. For all $z, l \in \mathbb{R}$,

$$\begin{aligned}
&|N_G(z, l) - N_{G_n}(z, l)| \\
&\le C(\Sigma)\int_{-\infty}^{z} \exp\left(-\frac{u^2}{2\Sigma_{11}}\right)\left|f_G\left(\frac{\Omega_{12}}{\Omega_{22}}u + l; \Omega_{22}^{-1}\right) - f_{G_n}\left(\frac{\Omega_{12}}{\Omega_{22}}u + l; \Omega_{22}^{-1}\right)\right| \mathrm{d}u \\
&\le C(\Sigma)\left(\int_{\mathbb{R}} \exp\left(-\frac{u^2}{\Sigma_{11}}\right)\mathrm{d}u\right)^{1/2}\left(\int_{\mathbb{R}}\left|f_G\left(\frac{\Omega_{12}}{\Omega_{22}}u + l; \Omega_{22}^{-1}\right) - f_{G_n}\left(\frac{\Omega_{12}}{\Omega_{22}}u + l; \Omega_{22}^{-1}\right)\right|^2 \mathrm{d}u\right)^{1/2} \\
&\lesssim_{\Sigma} \left\|f_G(\cdot; \Omega_{22}^{-1}) - f_{G_n}(\cdot; \Omega_{22}^{-1})\right\|_2,
\end{aligned}$$

where the last inequality follows from a change of variables, using $\Omega_{12} \neq 0$. Similarly, proceeding as in the proof of Lemma 16, for every $T > 0$, by the normal convolution identity and Plancherel's theorem

$$\begin{aligned}
&\left\|f_G(\cdot;\Omega_{22}^{-1}) - f_{G_n}(\cdot;\Omega_{22}^{-1})\right\|_2^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left|f_G^*(\omega;\Omega_{22}^{-1}) - f_{G_n}^*(\omega;\Omega_{22}^{-1})\right|^2 \mathrm{d}\omega \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{\Sigma_{22}\gamma^2\omega^2}\left|f_G^*(\omega;\Sigma_{22}) - f_{G_n}^*(\omega;\Sigma_{22})\right|^2 \mathrm{d}\omega \\
&\leq \frac{e^{\Sigma_{22}\gamma^2 T^2}}{2\pi}\int_{\mathbb{R}}\left|f_G^*(\omega;\Sigma_{22}) - f_{G_n}^*(\omega;\Sigma_{22})\right|^2 \mathrm{d}\omega + \frac{1}{2\pi}\int_{[-T,T]^c} e^{\Sigma_{22}\gamma^2\omega^2 - \Sigma_{22}\omega^2}|G^*(\omega) - G_n^*(\omega)|^2 \mathrm{d}\omega \\
&\leq e^{\Sigma_{22}\gamma^2 T^2}\left\|f_G(\cdot;\Sigma_{22}) - f_{G_n}(\cdot;\Sigma_{22})\right\|_2^2 + \frac{4}{2\pi}\int_{[-T,T]^c} e^{-\Sigma_{22}(1-\gamma^2)\omega^2}\mathrm{d}\omega,
\end{aligned}$$

since $|G^*(\omega)| \leq 1$ and $|G_n^*(\omega)| \leq 1$. Also, we have

$$\|f_G(\cdot;\Sigma_{22}) - f_{G_n}(\cdot;\Sigma_{22})\|_2^2 \lesssim_\Sigma \mathfrak{H}^2(f_G(\cdot;\Sigma_{22}), f_{G_n}(\cdot;\Sigma_{22})),$$

and

$$\int_{[-T,T]^c} e^{-\Sigma_{22}(1-\gamma^2)\omega^2}\,\mathrm{d}\omega \leq \frac{e^{-\Sigma_{22}(1-\gamma^2)T^2}}{\Sigma_{22}(1-\gamma^2)T}.$$

Taking $T = (\log n/\Sigma_{22})^{1/2}$, then

$$\sup_{z,l\in\mathbb{R}}|N_G(z,l) - N_{G_n}(z,l)| \lesssim_\Sigma n^{\gamma^2/2}\mathfrak{H}(f_G(\cdot;\Sigma_{22}), f_{G_n}(\cdot;\Sigma_{22})) + \frac{n^{-(1-\gamma^2)/2}}{(\log n)^{1/4}},$$

and it holds uniformly over G. Define $\mathcal{A} = \{d(G_n, \widehat{G}) \leq C_0 \log n/\sqrt{n}\}$, by Lemma 19, then on $\mathcal{A}$,

$$\sup_{z,l\in\mathbb{R}}|N_{\widehat{G}}(z,l) - N_{G_n}(z,l)| \lesssim_\Sigma \frac{\log n}{n^{(1-\gamma^2)/2}}.$$

Since $N_G(+\infty, l) = D_G(l)$, then for any fixed l,

$$|D_{\widehat{G}}(l) - D_{G_n}(l)| = \lim_{z\to+\infty}|N_{\widehat{G}}(z,l) - N_{G_n}(z,l)| \leq \sup_{z\in\mathbb{R}}|N_{\widehat{G}}(z,l) - N_{G_n}(z,l)|,$$

we also obtain

$$\sup_{l\in\mathbb{R}}|D_{\widehat{G}}(l) - D_{G_n}(l)| \lesssim_\Sigma \frac{\log n}{n^{(1-\gamma^2)/2}}.$$

Following the decomposition in Appendix D.2, define $\widehat{G}_* = (\widehat{G} + G_n)/2$, and

$$\left|F_{\widehat{G}}(z \mid l) - F_{G_n}(z \mid l)\right| \leq \frac{|N_{\widehat{G}}(z,l) - N_{G_n}(z,l)| + |D_{\widehat{G}}(l) - D_{G_n}(l)|}{D_{\widehat{G}_*}(l)}.$$

Since $D_{\widehat{G}_*}(l) = (D_{\widehat{G}}(l) + f_{G_n}(l;\Sigma_{22}))/2 \geq f_{G_n}(l;\Sigma_{22})/2$, on the event $\mathcal{A}$,

$$\sup_{z\in\mathbb{R}}|F_{\widehat{G}}(z \mid l) - F_{G_n}(z \mid l)| \lesssim_\Sigma \frac{\log n}{n^{(1-\gamma^2)/2}}\frac{1}{f_{G_n}(l;\Sigma_{22})}.$$

Taking $B_n = M + \sqrt{2\Sigma_{22}\log n}$, we have

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{b}}[\Delta_i] \lesssim_\Sigma \mathbb{P}_{\boldsymbol{b}}\left[\mathcal{A}^c\right] + \frac{1}{n}\sum_{i=1}^n \mathbb{P}_{\boldsymbol{b}}\left[|\hat{b}_i| > B_n\right] + \frac{\log n}{n^{(1-\gamma^2)/2}}\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{b}}\left[\frac{\mathbb{1}(|\hat{b}_i| \leq B_n)}{f_{G_n}(\hat{b}_i;\Sigma_{22})}\right].$$

Notice that the average marginal density of $\hat{b}_i$ is $f_{G_n}(\cdot;\Sigma_{22})$, then

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{b}}\left[\frac{\mathbb{1}(|\hat{b}_i| \leq B_n)}{f_{G_n}(\hat{b}_i;\Sigma_{22})}\right] = \int_{-B_n}^{B_n}\frac{n^{-1}\sum_{i=1}^n \varphi(l - b_i;\Sigma_{22})}{f_{G_n}(l;\Sigma_{22})}\,\mathrm{d}l = 2B_n.$$

Moreover, since $|b_i| \le M$, $\mathbb{P}_{\boldsymbol{b}}[|\hat{b}_i| > B_n] \le 2/n$. Hence,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{b}}[\Delta_i] \lesssim_{\Sigma,M} \frac{1}{n^{c_0}} + \frac{2}{n} + \frac{B_n \log n}{n^{(1-\gamma^2)/2}} \lesssim_{\Sigma,M} \frac{(\log n)^{3/2}}{n^{(1-\gamma^2)/2}}.$$

Now we need to move to the coverage result. Following the proof of Theorem 4, we define

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

By definition, $|\widehat{U}_i - U_i| \le \Delta_i$. For any $t \in [0,1]$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{b}}\left[U_i \le t\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{\theta_i,b_i}\left[F_{G_n}(\hat{\theta}_i^{\mathrm{db}} - \theta_i \mid \hat{b}_i) \le t\right] \\ &= \int \mathbb{P}_{\theta,b}\left[F_{G_n}(\hat{\theta}^{\mathrm{db}} - \theta \mid \hat{b}) \le t\right] G_n(\mathrm{d}b) = t,\end{aligned}$$

where the last equality follows from the probability integral transform under the mixture model with mixing distribution G_n. Similar to the set inclusions in the proof of Theorem 4, for any fixed $\delta > 0$, we have

$$\begin{aligned}&\left\{\frac{\alpha}{2} + \delta \le U_i \le 1 - \frac{\alpha}{2} - \delta\right\} \cap \{\Delta_i \le \delta\} \\ &\quad\subseteq \left\{\theta_i \in \mathcal{I}^{\mathrm{rb}}_{\widehat{G},i}(1-\alpha)\right\} \\ &\quad\subseteq \left\{\frac{\alpha}{2} - \delta \le U_i \le 1 - \frac{\alpha}{2} + \delta\right\} \cup \{\Delta_i > \delta\}.\end{aligned}$$

Taking probabilities, averaging over i, and using the average uniformity, we have

$$\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{b}}\left[\theta_i \in \mathcal{I}^{\mathrm{rb}}_{\widehat{G},i}(1-\alpha)\right] - (1-\alpha)\right| \le 2\delta + \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{\boldsymbol{b}}\left[\Delta_i > \delta\right].$$

Consequently, by Markov's inequality,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{b}}\left[\theta_i \in \mathcal{I}^{\mathrm{rb}}_{\widehat{G},i}(1-\alpha)\right] - (1-\alpha)\right| \le 2\delta + \frac{1}{n\delta}\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{b}}\left[\Delta_i\right] \lesssim_{\Sigma,M} 2\delta + \frac{1}{\delta}\frac{(\log n)^{3/2}}{n^{(1-\gamma^2)/2}}.$$

Taking $\delta = (\log n)^{3/4}/n^{(1-\gamma^2)/4}$, we obtain

$$\sup_{\alpha\in(0,1)}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{b}}\left[\theta_i \in \mathcal{I}^{\mathrm{rb}}_{\widehat{G},i}(1-\alpha)\right] - (1-\alpha)\right| \le C'\frac{(\log n)^{3/4}}{n^{(1-\gamma^2)/4}}.$$

□

D.7 Proof of Theorem 10

For notation simplicity, we denote $Y_i = \hat{\theta}_i^{\mathrm{db}} - \theta_i$, $A_i = \hat{b}_i$. For $\widetilde{G} \in \mathcal{P}([-M,M])$ and an admissible covariance matrix S, the *Normal working model* is

$$b_i \sim \widetilde{G}, \qquad (Y_i^G, A_i^G) \mid b_i \sim \mathrm{N}_2((0,b_i)^\top, S).$$

In particular, taking $\widetilde{G} = G$ and $S = \Sigma_i$ yields the matched normal experiment corresponding to the reference distribution in $\mathcal{W}_n$. Let $F_{\widetilde{G}}(y \mid a; S)$ denote the conditional CDF of $Y_i^G \mid A_i^G = a$ under the normal working model with mixing distribution $\widetilde{G}$ and covariance matrix S. We write $q_{\widetilde{G},\alpha}(a;S)$ for its α-quantile, then the rebiased interval proposed in (10) can be expressed as

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

We first collect several direct consequences of the assumptions that will be used repeatedly in the proof.

Lemma 20. Let $\mu_i = \mathcal{L}_{\boldsymbol{\theta},G}(A_i)$, and let ν_i have density $f_G(\cdot;\Sigma_{i,22})$. Note that $\tau_i^2 = \Sigma_{i,22}$, and let $\hat{\tau}_i^2 = \widehat{\Sigma}_{i,22}$. Then under assumptions of Theorem 10,

$$V_n := \frac{1}{n}\sum_{i=1}^{n} W_1(\mu_i,\nu_i) \le \mathcal{W}_n \to 0, \tag{S27}$$

$$\lim_{T\to\infty}\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{\theta},G}\left[A_i^2\mathbb{1}(|A_i|>T)\right] = 0, \tag{S28}$$

$$\frac{1}{n}\sum_{i=1}^{n}|\widehat{\tau}_i-\tau_i| \xrightarrow{\mathbb{P}_{\boldsymbol{\theta},G}} 0. \tag{S29}$$

Here W_p denotes the Wasserstein-p distance.

Proof. For each i, conditionally on $b_i = b$, take an optimal bivariate W_2 coupling between $\mathcal{L}_{\boldsymbol{\theta},G}((Y_i,A_i) \mid b_i = b)$ and $\mathrm{N}_2((0,b)^\top,\Sigma_i)$. Let (Y_i^G,A_i^G) denote the normal random vector in this coupling, and let $d_i(b)$ be the corresponding Wasserstein-2 distance. Projection gives

$$|A_i - A_i^G|^2 \le \|(Y_i,A_i)-(Y_i^G,A_i^G)\|^2,$$
$$W_1(\mu_i,\nu_i) \le \int W_1\big(\mathcal{L}(A_i \mid b_i=b),\mathrm{N}(b,\tau_i^2)\big)\,G(\mathrm{d}b) \le \int d_i(b)\,G(\mathrm{d}b),$$

since $\mathbb{E}\left[|A_i-A_i^G|\right] \le \sqrt{\mathbb{E}\left[|A_i-A_i^G|^2\right]} \le d_i(b)$. Cauchy–Schwarz now yields

$$V_n \le \sqrt{\frac{1}{n}\sum_{i=1}^{n}\int d_i(b)^2\,G(\mathrm{d}b)} = \mathcal{W}_n.$$

Under the same individual couplings, $n^{-1}\sum_{i=1}^n \mathbb{E}\left[|A_i-A_i^G|^2\right] \le \mathcal{W}_n^2$. Since $A_i^G \mid b_i \sim \mathrm{N}(b_i,\tau_i^2)$, its marginal distribution admits the representation $A_i^G = b_i+\tau_i Z_i$, where $Z_i \sim \mathrm{N}(0,1)$ is independent of b_i, and $|b_i| \le M$ almost surely. Hence

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[A_i^2\mathbb{1}(|A_i|>T)\right] &\le \frac{4}{n}\sum_{i=1}^{n}\mathbb{E}\left[|A_i-A_i^G|^2\right] + \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left[(A_i^G)^2\mathbb{1}(|A_i^G|>T/2)\right]\\ &\le 4\mathcal{W}_n^2 + \frac{8}{T^2}\sup_i\mathbb{E}\left[(A_i^G)^4\right] \le 4\mathcal{W}_n^2 + \frac{64(M^4+3\bar{\lambda}^2)}{T^2},\end{aligned}$$

where the first inequality follows from $x^2\mathbb{1}(|x|>T) \le 4(x-z)^2 + 2z^2\mathbb{1}(|z|>T/2)$. The last uses $(u+v)^4 \le 8(u^4+v^4)$, therefore $\mathbb{E}\left[(A_i^G)^4\right] \le 8(\mathbb{E}\left[b_i^4\right]+\tau_i^4\mathbb{E}\left[Z^4\right]) \le 8(M^4+3\bar{\lambda}^2)$. Taking $\limsup_n$, then $T\to\infty$, proves (S28). Finally,

$$\frac{1}{n}\sum_{i=1}^{n}|\widehat{\tau}_i-\tau_i| = \frac{1}{n}\sum_{i=1}^{n}\frac{|\widehat{\Sigma}_{i,22}-\Sigma_{i,22}|}{\widehat{\tau}_i+\tau_i} \le \frac{1}{2\sqrt{\underline{\lambda}}\,n}\sum_{i=1}^{n}\|\widehat{\Sigma}_i-\Sigma_i\| \xrightarrow{\mathbb{P}_{\boldsymbol{\theta},G}} 0.$$

□

We next establish the mixing-distribution consistency needed for the coverage proof.

Lemma 21. Under assumptions of Theorem 10, we have $W_1(\widehat{G},G) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta},G}} 0$. Here W_1 denotes the Wasserstein-1 distance on $\mathcal{P}([-M,M])$.

Proof. Let

$$L_n^{\mathrm{true}}(\widetilde{G}) = \frac{1}{n}\sum_{i=1}^{n}\log f_{\widetilde{G}}(A_i;\Sigma_{i,22}), \qquad L_n^{\mathrm{est}}(\widetilde{G}) = \frac{1}{n}\sum_{i=1}^{n}\log f_{\widetilde{G}}(A_i;\widehat{\Sigma}_{i,22}).$$

By the global optimality of $\widehat{G}$ and since $G \in \mathcal{P}([-M, M])$, we have

$$L_n^{\text{est}}(\widehat{G}) \geq L_n^{\text{est}}(G).$$

For $v \in [\underline{\lambda}, \overline{\lambda}]$,

$$\partial_v \log f_{\widetilde{G}}(a; v) = \mathbb{E}_{b \sim \widetilde{G}_{a,v}} \left[-\frac{1}{2v} + \frac{(a-b)^2}{2v^2} \right], \quad \widetilde{G}_{a,v}(\mathrm{d}b) = \frac{\varphi(a-b; v)\widetilde{G}(\mathrm{d}b)}{f_{\widetilde{G}}(a; v)}.$$

Since $\widetilde{G} \in \mathcal{P}([-M, M])$ and $v \in [\underline{\lambda}, \overline{\lambda}]$, we have $\left|\partial_v \log f_{\widetilde{G}}(a; v)\right| \leq C(1 + a^2)$. For every realized $(A_i, \widehat{\Sigma}_{i,22})$ and $T > 1$,

$$\begin{aligned}
R_n &:= \sup_{\widetilde{G} \in \mathcal{P}([-M,M])} |L_n^{\text{est}}(\widetilde{G}) - L_n^{\text{true}}(\widetilde{G})| \\
&\leq \frac{C}{n} \sum_{i=1}^n (1 + A_i^2) |\widehat{\Sigma}_{i,22} - \Sigma_{i,22}| \\
&\leq C(1 + T^2) \frac{1}{n} \sum_{i=1}^n |\widehat{\Sigma}_{i,22} - \Sigma_{i,22}| + \frac{2C\overline{\lambda}}{n} \sum_{i=1}^n (1 + A_i^2) \mathbb{1}(|A_i| > T).
\end{aligned}$$

By Lemma 20, for every fixed $T > 1$, the first term is $o_{\mathbb{P}}(1)$. For the second term, notice that $(1 + A_i^2)\mathbb{1}(|A_i| > T) \leq 2A_i^2 \mathbb{1}(|A_i| > T)$ and use Markov's inequality and (S28), we have

$$\lim_{T \to \infty} \limsup_{n \to \infty} \mathbb{P}\left[\frac{1}{n} \sum_i (1 + A_i^2) \mathbb{1}(|A_i| > T) > \varepsilon \right] = 0$$

for every $\varepsilon > 0$. Choosing T sufficiently large and then letting $n \to \infty$ yields $R_n = o_{\mathbb{P}_{\boldsymbol{\theta},G}}(1)$. Consequently, by the global optimality of $\widehat{G}$, $L_n^{\text{true}}(\widehat{G}) - L_n^{\text{true}}(G) \geq -2R_n = -o_{\mathbb{P}_{\boldsymbol{\theta},G}}(1)$.

Set $g_{i,\widetilde{G}}(a) = \log\{f_{\widetilde{G}}(a; \Sigma_{i,22}) / f_G(a; \Sigma_{i,22})\}$ and $m_{\widetilde{G},a,v} = \mathbb{E}_{\widetilde{G}_{a,v}}[b]$. Posterior differentiation gives

$$\partial_a \log f_{\widetilde{G}}(a; v) = \frac{m_{\widetilde{G},a,v} - a}{v}, \qquad \partial_a^2 \log f_{\widetilde{G}}(a; v) = \frac{\mathrm{Var}_{\widetilde{G}_{a,v}}(b)}{v^2} - \frac{1}{v}.$$

Since $\widetilde{G}_{a,v}$ is supported on $[-M, M]$, we have $|m_{\widetilde{G},a,v}| \leq M$, $\mathrm{Var}_{\widetilde{G}_{a,v}}(b) \leq M^2$. Therefore, for every $\widetilde{G} \in \mathcal{P}([-M, M])$,

$$|g'_{i,\widetilde{G}}(0)| \leq \frac{2M}{\underline{\lambda}}, \qquad \|g'_{i,\widetilde{G}}\|_\infty \leq \frac{2M}{\underline{\lambda}}, \qquad \|g''_{i,\widetilde{G}}\|_\infty \leq \frac{2M^2}{\underline{\lambda}^2}.$$

Since Assumption 9 allows deterministic triangular arrays of functions, it applies to $g_{i,\widetilde{G}}/C_0$ for each fixed $\widetilde{G}$, with $C_0 = \max\{1, 2M/\underline{\lambda}, 2M^2/\underline{\lambda}^2\}$, and it gives

$$Z_n(\widetilde{G}) := \frac{1}{n} \sum_{i=1}^n \left(g_{i,\widetilde{G}}(A_i) - \mathbb{E}_{\boldsymbol{\theta},G}\left[g_{i,\widetilde{G}}(A_i) \right] \right) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta},G}} 0. \tag{S30}$$

We next upgrade the pointwise convergence in (S30) to uniform convergence over $\widetilde{G} \in \mathcal{P}([-M, M])$, we separately control the compact region $|A_i| \leq T$ and the tails. On the compact region $|a| \leq T$, when $|b| \leq M$, $v \in [\underline{\lambda}, \overline{\lambda}]$, set $d_T = \inf \varphi(a - b; v) > 0$ and $C_T = \sup |\partial_b \varphi(a - b; v)| < \infty$. Then, by the Kantorovich–Rubinstein duality for W_1, $|f_{\widetilde{G}}(a; v) - f_{\widetilde{G}'}(a; v)| \leq C_T W_1(\widetilde{G}, \widetilde{G}')$. Since $f_{\widetilde{G}}(a; v), f_{\widetilde{G}'}(a; v) \geq d_T$,

$$\left|\log f_{\widetilde{G}}(a; v) - \log f_{\widetilde{G}'}(a; v)\right| \leq \frac{1}{d_T} \left| f_{\widetilde{G}}(a; v) - f_{\widetilde{G}'}(a; v) \right| \leq \frac{C_T}{d_T} W_1(\widetilde{G}, \widetilde{G}').$$

Hence

$$\sup_{i, |a| \leq T} |g_{i,\widetilde{G}}(a) - g_{i,\widetilde{G}'}(a)| \leq \frac{C_T}{d_T} W_1(\widetilde{G}, \widetilde{G}').$$

For the tail region, since $\Sigma_{i,22} \in [\underline{\lambda}, \overline{\lambda}]$ and $G, \widetilde{G} \in \mathcal{P}([-M, M])$, we have $\sup_{i,\widetilde{G}} |g_{i,\widetilde{G}}(0)| < \infty$, and $\sup_{i,\widetilde{G}} \|g'_{i,\widetilde{G}}\|_\infty \le 2M/\underline{\lambda}$. Hence, for some constant $C_1 < \infty$,

$$|g_{i,\widetilde{G}}(a)| \le |g_{i,\widetilde{G}}(0)| + \|g'_{i,\widetilde{G}}\|_\infty |a| \le C_1(1 + |a|)$$

uniformly in i, $\widetilde{G}$, and a. Therefore, on $\{|A_i| > T\}$, $|g_{i,\widetilde{G}}(A_i)| \le C_1(1 + |A_i|)$, and for $T \ge 1$,

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{\theta},G}[(1 + |A_i|)\mathbb{1}(|A_i| > T)] \le \left(\frac{1}{T^2} + \frac{1}{T}\right)\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{\theta},G}\left[A_i^2 \mathbb{1}(|A_i| > T)\right],$$

combining with Lemma 20,

$$\lim_{T\to\infty} \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{\theta},G}[(1 + |A_i|)\mathbb{1}(|A_i| > T)] = 0. \tag{S31}$$

We can now combine the compact-region continuity bound and the tail bound with a finite W_1-net of $\mathcal{P}([-M, M])$. Let $\widetilde{G}_1, \ldots, \widetilde{G}_J$ be a finite W_1-δ-net of $\mathcal{P}([-M, M])$, i.e. for any $\widetilde{G} \in \mathcal{P}([-M, M])$ we can find a $j \le J$ such that $W_1(\widetilde{G}, \widetilde{G}_j) \le \delta$, and we have

$$|Z_n(\widetilde{G})| \le |Z_n(\widetilde{G}_j)| + |Z_n(\widetilde{G}) - Z_n(\widetilde{G}_j)|.$$

By taking supremum over all possible $\widetilde{G}$,

$$\sup_{\widetilde{G}\in\mathcal{P}([-M,M])} |Z_n(\widetilde{G})| \le \max_{j\le J} |Z_n(\widetilde{G}_j)| + \sup_{\widetilde{G}\in\mathcal{P}([-M,M])} |Z_n(\widetilde{G}) - Z_n(\widetilde{G}_j)|.$$

For the second term,

$$\begin{aligned}
&\left|Z_n(\widetilde{G}) - Z_n(\widetilde{G}_j)\right| \\
&\le \frac{1}{n}\sum_{i=1}^n \left|g_{i,\widetilde{G}}(A_i) - g_{i,\widetilde{G}_j}(A_i)\right| + \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{\theta},G}\left[\left|g_{i,\widetilde{G}}(A_i) - g_{i,\widetilde{G}_j}(A_i)\right|\right] \\
&\le \frac{1}{n}\sum_{i=1}^n \left|g_{i,\widetilde{G}}(A_i) - g_{i,\widetilde{G}_j}(A_i)\right| \mathbb{1}(|A_i| \le T) + \frac{1}{n}\sum_{i=1}^n \left|g_{i,\widetilde{G}}(A_i) - g_{i,\widetilde{G}_j}(A_i)\right| \mathbb{1}(|A_i| > T) \\
&\quad + \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{\theta},G}\left[\left|g_{i,\widetilde{G}}(A_i) - g_{i,\widetilde{G}_j}(A_i)\right| \mathbb{1}(|A_i| \le T)\right] + \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{\theta},G}\left[\left|g_{i,\widetilde{G}}(A_i) - g_{i,\widetilde{G}_j}(A_i)\right| \mathbb{1}(|A_i| > T)\right] \\
&\le 2\frac{C_T}{d_T}\delta + \frac{2C_1}{n}\sum_{i=1}^n (1 + |A_i|)\mathbb{1}(|A_i| > T) + \frac{2C_1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{\theta},G}[(1 + |A_i|)\mathbb{1}(|A_i| > T)].
\end{aligned}$$

For $\epsilon > 0$, choose T large enough such that the last deterministic term is at most $\epsilon/4$; then choose $\delta > 0$ such that $2(C_T/d_T)\delta < \epsilon/4$. For these fixed T, δ, and all sufficiently large n,

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\theta},G}\left[\sup_{\widetilde{G}\in\mathcal{P}([-M,M])} |Z_n(\widetilde{G})| > \epsilon\right] &\le \mathbb{P}_{\boldsymbol{\theta},G}\left[\max_{j\le J} |Z_n(\widetilde{G}_j)| > \epsilon/4\right] \\
&\quad + \mathbb{P}_{\boldsymbol{\theta},G}\left[\frac{2C_1}{n}\sum_{i=1}^n (1 + |A_i|)\mathbb{1}(|A_i| > T) > \epsilon/4\right].
\end{aligned}$$

For the first probability term, since the W_1-net is finite, $J < \infty$, and $Z_n(\widetilde{G}_j) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta},G}} 0$ for each fixed j by (S30), a finite union bound gives

$$\mathbb{P}_{\boldsymbol{\theta},G}\left[\max_{j\le J} |Z_n(\widetilde{G}_j)| > \varepsilon/4\right] \le \sum_{j=1}^J \mathbb{P}_{\boldsymbol{\theta},G}\left[|Z_n(\widetilde{G}_j)| > \varepsilon/4\right] \longrightarrow 0.$$

For the second probability term, Markov's inequality gives

$$\mathbb{P}_{\boldsymbol{\theta},G}\left[\frac{2C_1}{n}\sum_{i=1}^n (1 + |A_i|)\mathbb{1}(|A_i| > T) > \frac{\varepsilon}{4}\right] \le \frac{8C_1}{\varepsilon}\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{\theta},G}\left[(1 + |A_i|)\mathbb{1}(|A_i| > T)\right],$$

and then combine it with (S31), we have

$$\lim_{T\to\infty}\limsup_{n\to\infty}\mathbb{P}_{\boldsymbol{\theta},G}\left(\frac{2C_1}{n}\sum_{i=1}^{n}(1+|A_i|)\mathbb{1}(|A_i|>T)>\frac{\varepsilon}{4}\right)=0.$$

Combining the two bounds yields

$$\sup_{\widetilde{G}\in\mathcal{P}([-M,M])}\left|\frac{1}{n}\sum_{i=1}^{n}\left(g_{i,\widetilde{G}}(A_i)-\mathbb{E}_{\boldsymbol{\theta},G}\left[g_{i,\widetilde{G}}(A_i)\right]\right)\right|\xrightarrow{\mathbb{P}_{\boldsymbol{\theta},G}}0. \tag{S32}$$

With μ_i,ν_i from Lemma 20, and $\|g'_{i,\widetilde{G}}\|_\infty\le 2M/\underline{\lambda}$, by the Kantorovich–Rubinstein duality for W_1, we have $\left|\mathbb{E}_{\mu_i}g_{i,\widetilde{G}}-\mathbb{E}_{\nu_i}g_{i,\widetilde{G}}\right|\le(2M/\underline{\lambda})\cdot W_1(\mu_i,\nu_i)$, and consequently

$$\sup_{\widetilde{G}\in\mathcal{P}([-M,M])}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}_{\mu_i}g_{i,\widetilde{G}}-\mathbb{E}_{\nu_i}g_{i,\widetilde{G}}\right)\right|\le\frac{2M}{\underline{\lambda}}\frac{1}{n}\sum_{i=1}^{n}W_1(\mu_i,\nu_i)=\frac{2M}{\underline{\lambda}}V_n=o(1). \tag{S33}$$

Also notice that $\mathbb{E}_{\nu_i}g_{i,\widetilde{G}}=-\operatorname{KL}\big(f_G(\cdot;\Sigma_{i,22})\|f_{\widetilde{G}}(\cdot;\Sigma_{i,22})\big)$. Defining

$$K_n(\widetilde{G})=\frac{1}{n}\sum_{i=1}^{n}\operatorname{KL}\big(f_G(\cdot;\Sigma_{i,22})\|f_{\widetilde{G}}(\cdot;\Sigma_{i,22})\big)=-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\nu_i}g_{i,\widetilde{G}},$$

we have

$$\begin{aligned}L_n^{\text{true}}(\widetilde{G})-L_n^{\text{true}}(G)+K_n(\widetilde{G})=&\left[L_n^{\text{true}}(\widetilde{G})-L_n^{\text{true}}(G)-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mu_i}g_{i,\widetilde{G}}\right]\\&+\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}_{\mu_i}g_{i,\widetilde{G}}-\mathbb{E}_{\nu_i}g_{i,\widetilde{G}}\right),\end{aligned}$$

then combining (S32) and (S33), we obtain

$$\sup_{\widetilde{G}\in\mathcal{P}([-M,M])}\left|L_n^{\text{true}}(\widetilde{G})-L_n^{\text{true}}(G)+K_n(\widetilde{G})\right|\xrightarrow{\mathbb{P}_{\boldsymbol{\theta},G}}0.$$

Since $\widehat{G}\in\mathcal{P}([-M,M])$, and $L_n^{\text{true}}(\widehat{G})-L_n^{\text{true}}(G)\ge-2R_n=-o_{\mathbb{P}_{\boldsymbol{\theta},G}}(1)$, it follows that

$$0\le K_n(\widehat{G})\le\sup_{\widetilde{G}\in\mathcal{P}([-M,M])}\left|L_n^{\text{true}}(\widetilde{G})-L_n^{\text{true}}(G)+K_n(\widetilde{G})\right|+2R_n=o_{\mathbb{P}_{\boldsymbol{\theta},G}}(1).$$

Therefore, since $\mathfrak{H}^2(p,q)\le\operatorname{KL}(p\|q)$,

$$\frac{1}{n}\sum_{i=1}^{n}\mathfrak{H}^2\left(f_{\widehat{G}}(\cdot;\Sigma_{i,22}),f_G(\cdot;\Sigma_{i,22})\right)\le K_n(\widehat{G})=o_{\mathbb{P}_{\boldsymbol{\theta},G}}(1).$$

It remains to translate the mixture Hellinger consistency into consistency of the mixing distribution. If $\varepsilon>2M$, then $W_1(\widehat{G},G)<\varepsilon$ automatically. Otherwise, for $0<\varepsilon\le 2M$, define

$$c_\varepsilon:=\inf_{\substack{G_1,G_2\in\mathcal{P}([-M,M]),\,W_1(G_1,G_2)\ge\varepsilon,\\ v\in[\underline{\lambda},\overline{\lambda}]}}\mathfrak{H}^2\left(f_{G_1}(\cdot;v),f_{G_2}(\cdot;v)\right).$$

The index set $\{(G_1,G_2,v):G_1,G_2\in\mathcal{P}([-M,M]),W_1(G_1,G_2)\ge\varepsilon,\,v\in[\underline{\lambda},\overline{\lambda}]\}$ is compact, and the map

$$(G_1,G_2,v)\mapsto\mathfrak{H}^2\left(f_{G_1}(\cdot;v),f_{G_2}(\cdot;v)\right)$$

is continuous. Suppose for contradiction that $c_\varepsilon=0$. Then the infimum is attained at some $(G_1^\star,G_2^\star,v^\star)$ such that $\mathfrak{H}^2\left(f_{G_1^\star}(\cdot;v^\star),f_{G_2^\star}(\cdot;v^\star)\right)=0$, therefore $f_{G_1^\star}(\cdot;v^\star)=f_{G_2^\star}(\cdot;v^\star)$. Taking characteristic functions gives

$$\left(\int e^{itb}\,\mathrm{d}G_1^\star(b)\right)e^{-v^\star t^2/2}=\left(\int e^{itb}\,\mathrm{d}G_2^\star(b)\right)e^{-v^\star t^2/2},$$

and hence $G_1^\star = G_2^\star$, contradicting $W_1(G_1^\star, G_2^\star) \geq \varepsilon$. Therefore, $c_\varepsilon > 0$ and it is independent of n. Since $\widehat{G}$ and G belong to $\mathcal{P}([-M, M])$, it follows that

$$c_\varepsilon \mathbb{1}(W_1(\widehat{G}, G) \geq \varepsilon) \leq \frac{1}{n} \sum_{i=1}^{n} \mathfrak{H}^2 \left(f_{\widehat{G}}(\cdot; \Sigma_{i,22}), f_G(\cdot; \Sigma_{i,22}) \right) = o_{\mathbb{P}_{\boldsymbol{\theta},G}}(1),$$

which implies $W_1(\widehat{G}, G) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta},G}} 0$. □

The following lemma gives a uniform Lipschitz bound for the normal working-model conditional CDF with respect to the observation pair.

Lemma 22. For $\widetilde{G} \in \mathcal{P}([-M, M])$ and an admissible covariance matrix S defined below, we define the following

$$S = \begin{pmatrix} v_Y & c \\ c & v_A \end{pmatrix},$$

$$\widetilde{G}_{a,S}(\mathrm{d}b) = \frac{\varphi(a-b; v_A)\widetilde{G}(\mathrm{d}b)}{f_{\widetilde{G}}(a; v_A)}, \quad F_{\widetilde{G}}(y \mid a; S) = \int \Phi\left(\frac{y - \frac{c}{v_A}(a-b)}{\sqrt{v_Y - c^2/v_A}} \right) \widetilde{G}_{a,S}(\mathrm{d}b),$$

where $\underline{\lambda} I_2 \preceq S \preceq \bar{\lambda} I_2$. Then there is $L < \infty$, depending only on $M, \underline{\lambda}, \overline{\lambda}$, such that uniformly in $\widetilde{G}, S, y, a, y', a'$,

$$|F_{\widetilde{G}}(y \mid a; S) - F_{\widetilde{G}}(y' \mid a'; S)| \leq L\sqrt{(y-y')^2 + (a-a')^2}. \tag{S34}$$

Proof. Conditional on b, normal conditioning gives

$$Y^G \mid (A^G = a, b) \sim \mathrm{N}\left(\frac{c}{v_A}(a-b), v_Y - \frac{c^2}{v_A} \right).$$

Under the normal working model, Bayes' rule gives the conditional distribution of b given $A^G = a$ is $\widetilde{G}_{a,S}$. Hence, integrating the preceding normal CDF with respect to $\widetilde{G}_{a,S}$ gives the stated expression for $F_{\widetilde{G}}(y \mid a; S)$. To establish the uniform Lipschitz bound, note that

$$v_Y - \frac{c^2}{v_A} = \left(1, -\frac{c}{v_A}\right) S \left(1, -\frac{c}{v_A}\right)^\top \geq \underline{\lambda}\left(1 + \frac{c^2}{v_A^2}\right) \geq \underline{\lambda}, \qquad \left|\frac{c}{v_A}\right| \leq \frac{\bar{\lambda}}{\underline{\lambda}}.$$

For brevity, set

$$q(y, a, b) = \Phi\left(\frac{y - \frac{c}{v_A}(a-b)}{\sqrt{v_Y - c^2/v_A}} \right).$$

Differentiating under the posterior integral gives

$$\partial_y F_{\widetilde{G}}(y \mid a; S) = \int \varphi\left(y - \frac{c}{v_A}(a-b); v_Y - \frac{c^2}{v_A} \right) \widetilde{G}_{a,S}(\mathrm{d}b),$$

and

$$\partial_a F_{\widetilde{G}}(y \mid a; S) = -\frac{c}{v_A} \int \varphi\left(y - \frac{c}{v_A}(a-b); v_Y - \frac{c^2}{v_A} \right) \widetilde{G}_{a,S}(\mathrm{d}b) + \frac{1}{v_A} \operatorname{Cov}_{\widetilde{G}_{a,S}}\left(q(y, a, b), b \right).$$

Since $\varphi(x; v) \leq 1/\sqrt{2\pi v}$ and $v_Y - c^2/v_A \geq \underline{\lambda}$, we have $\left|\partial_y F_{\widetilde{G}}(y \mid a; S)\right| \leq 1/\sqrt{2\pi\underline{\lambda}}$. For the second term, since $0 \leq q(y, a, b) \leq 1$ and $|b| \leq M$, it holds that $\left|\operatorname{Cov}_{\widetilde{G}_{a,S}}(q(y, a, b), b)\right| \leq 2M$. Therefore,

$$\left|\partial_a F_{\widetilde{G}}(y \mid a; S)\right| \leq \frac{\overline{\lambda}}{\sqrt{2\pi}\,\underline{\lambda}^{3/2}} + \frac{2M}{\underline{\lambda}}.$$

Thus the Euclidean norm of the gradient of $(y, a) \mapsto F_{\widetilde{G}}(y \mid a; S)$ is bounded uniformly over all admissible S and $\widetilde{G} \in \mathcal{P}([-M, M])$. Integrating the gradient along the line segment joining (y, a) and (y', a') yields

$$\left|F_{\widetilde{G}}(y \mid a; S) - F_{\widetilde{G}}(y' \mid a'; S)\right| \leq L\sqrt{(y-y')^2 + (a-a')^2},$$

for some $L < \infty$ depending only on $M, \underline{\lambda}, \overline{\lambda}$. □

We next complement the preceding Lipschitz bound with stability with respect to the mixing distribution and covariance matrix.

Lemma 23. For every $K > 0$ there is a finite C_K such that, for admissible S, S' and $\widetilde{G}, \widetilde{G}' \in \mathcal{P}([-M, M])$,

$$\sup_{|a|\le K, y\in\mathbb{R}} |F_{\widetilde{G}}(y \mid a; S) - F_{\widetilde{G}'}(y \mid a; S)| \le C_K W_1(\widetilde{G}, \widetilde{G}'), \tag{S35}$$

$$\sup_{\widetilde{G}, |a|\le K, y\in\mathbb{R}} |F_{\widetilde{G}}(y \mid a; S) - F_{\widetilde{G}}(y \mid a; S')| \le C_K \|S - S'\|. \tag{S36}$$

Proof. Constants denoted by C_K below depend only on $K, M, \underline{\lambda}, \overline{\lambda}$. For fixed y, a, S (S defined in Lemma 22), define

$$N_{\widetilde{G}}(y, a; S) := F_{\widetilde{G}}(y \mid a; S) f_{\widetilde{G}}(a; v_A), \qquad D_{\widetilde{G}}(a; S) := f_{\widetilde{G}}(a; v_A).$$

Equivalently,

$$N_{\widetilde{G}}(y, a; S) = \int \Phi\left(\frac{y - \frac{c}{v_A}(a-b)}{\sqrt{v_Y - c^2/v_A}}\right) \varphi(a-b; v_A)\, \widetilde{G}(\mathrm{d}b), \quad D_{\widetilde{G}}(a; S) = \int \varphi(a-b; v_A)\, \widetilde{G}(\mathrm{d}b).$$

For $|a| \le K$, define

$$d_K := \inf_{\substack{|a|\le K,\ |b|\le M,\\ v\in[\underline{\lambda},\overline{\lambda}]}} \varphi(a-b; v) > 0.$$

Then, uniformly in admissible S, it holds that $D_{\widetilde{G}}(a; S) \ge d_K$ and $D_{\widetilde{G}'}(a; S) \ge d_K$. We first have on $|a| \le K$ and $|b| \le M$,

$$|\partial_b \varphi(a-b; v_A)| = \left|\frac{a-b}{v_A}\varphi(a-b; v_A)\right| \le \frac{K+M}{\underline{\lambda}}(2\pi\underline{\lambda})^{-1/2} =: C_{K,1},$$

and

$$|\partial_b q(y, a, b)| = \left|\frac{c}{v_A}\right| \varphi\left(y - \frac{c}{v_A}(a-b); v_Y - \frac{c^2}{v_A}\right) \le \frac{\overline{\lambda}}{\sqrt{2\pi}\,\underline{\lambda}^{3/2}} =: C_2,$$

uniformly in y and admissible S. Since

$$\begin{aligned}|\partial_b\{q(y, a, b)\varphi(a-b; v_A)\}| &\le |\partial_b q(y, a, b)|\, \varphi(a-b; v_A) + |q(y, a, b)|\, |\partial_b \varphi(a-b; v_A)| \\ &\le C_2(2\pi\underline{\lambda})^{-1/2} + C_{K,1} \\ &=: C_{K,2},\end{aligned}$$

then both

$$b \mapsto \varphi(a-b; v_A), \quad b \mapsto q(y, a, b)\varphi(a-b; v_A)$$

are uniformly Lipschitz on $|a| \le K$, $|b| \le M$, over admissible S and all $y \in \mathbb{R}$. Therefore, by Kantorovich–Rubinstein duality, there exists a constant $C_K > 0$ such that

$$|D_{\widetilde{G}}(a; S) - D_{\widetilde{G}'}(a; S)| + |N_{\widetilde{G}}(y, a; S) - N_{\widetilde{G}'}(y, a; S)| \le C_K W_1(\widetilde{G}, \widetilde{G}').$$

Since $0 \le N_{\widetilde{G}'}(y, a; S)/D_{\widetilde{G}'}(a; S) = F_{\widetilde{G}'}(y \mid a; S) \le 1$,

$$\begin{aligned}&\left|F_{\widetilde{G}}(y \mid a; S) - F_{\widetilde{G}'}(y \mid a; S)\right| = \left|\frac{N_{\widetilde{G}}(y, a; S)}{D_{\widetilde{G}}(a; S)} - \frac{N_{\widetilde{G}'}(y, a; S)}{D_{\widetilde{G}'}(a; S)}\right| \\ &= \left|\frac{N_{\widetilde{G}}(y, a; S)}{D_{\widetilde{G}}(a; S)} - \frac{N_{\widetilde{G}'}(y, a; S)}{D_{\widetilde{G}}(a; S)} + \frac{N_{\widetilde{G}'}(y, a; S)}{D_{\widetilde{G}}(a; S)} - \frac{N_{\widetilde{G}'}(y, a; S)}{D_{\widetilde{G}'}(a; S)}\right| \\ &\le \frac{\left|N_{\widetilde{G}}(y, a; S) - N_{\widetilde{G}'}(y, a; S)\right|}{D_{\widetilde{G}}(a; S)} + \frac{N_{\widetilde{G}'}(y, a; S)}{D_{\widetilde{G}'}(a; S)} \frac{\left|D_{\widetilde{G}}(a; S) - D_{\widetilde{G}'}(a; S)\right|}{D_{\widetilde{G}}(a; S)} \\ &\le \frac{\left|N_{\widetilde{G}}(y, a; S) - N_{\widetilde{G}'}(y, a; S)\right| + \left|D_{\widetilde{G}}(a; S) - D_{\widetilde{G}'}(a; S)\right|}{d_K} \\ &\le C_K W_1(\widetilde{G}, \widetilde{G}'),\end{aligned}$$

which proves (S35). For (S36), let $S_t = (1-t)S + tS', t \in [0,1]$, and write

$$S_t = \begin{pmatrix} v_{Y,t} & c_t \\ c_t & v_{A,t} \end{pmatrix},$$

and the class of admissible covariance matrices is convex, $\underline{\lambda} I_2 \preceq S_t \preceq \overline{\lambda} I_2$ for every $t \in [0,1]$. In particular,

$$v_{A,t} \geq \underline{\lambda}, \qquad v_{Y,t} - \frac{c_t^2}{v_{A,t}} \geq \underline{\lambda}, \qquad |c_t| \leq \overline{\lambda}, \qquad |\partial_t v_{Y,t}|,\ |\partial_t v_{A,t}|,\ |\partial_t c_t| \leq \|S' - S\|.$$

Define

$$z_t(y,a,b) = \frac{y - \frac{c_t}{v_{A,t}}(a-b)}{\sqrt{v_{Y,t} - c_t^2/v_{A,t}}},$$

then on $|a| \leq K$ and $|b| \leq M$, we have

$$\begin{aligned} |\partial_t z_t(y,a,b)| &\leq \frac{|a-b|}{\sqrt{v_{Y,t} - c_t^2/v_{A,t}}} \left| \partial_t \left(\frac{c_t}{v_{A,t}} \right) \right| + |z_t(y,a,b)| \left| \partial_t \log \sqrt{v_{Y,t} - \frac{c_t^2}{v_{A,t}}} \right| \\ &\leq C_K (1 + |z_t(y,a,b)|) \|S - S'\|, \end{aligned}$$

and $|\partial_t \Phi(z_t(y,a,b))| = \varphi(z_t(y,a,b);1) |\partial_t z_t(y,a,b)| \leq C_K \|S - S'\|$, since $\sup_z \varphi(z;1) < \infty$, $\sup_z |z| \varphi(z;1) < \infty$. For the normal marginal factor,

$$\partial_t \varphi(a-b; v_{A,t}) = \varphi(a-b; v_{A,t}) \left[-\frac{1}{2v_{A,t}} + \frac{(a-b)^2}{2v_{A,t}^2} \right] \partial_t v_{A,t},$$

since $v_{A,t} \geq \underline{\lambda}$, then on $|a| \leq K$ and $|b| \leq M$, it holds that $|\partial_t \varphi(a-b; v_{A,t})| \leq C_K \|S - S'\|$, and then

$$\begin{aligned} \left| \partial_t N_{\widetilde{G}}(y,a;S_t) \right| &\leq \int |\partial_t \Phi(z_t(y,a,b))|\, \varphi(a-b; v_{A,t})\, \widetilde{G}(\mathrm{d}b) \\ &\quad + \int \Phi(z_t(y,a,b))\, |\partial_t \varphi(a-b; v_{A,t})|\, \widetilde{G}(\mathrm{d}b) \\ &\leq C_K \|S - S'\|. \end{aligned}$$

Similarly, $\left| \partial_t D_{\widetilde{G}}(a;S_t) \right| \leq C_K \|S - S'\|$. Then, integrating over $t \in [0,1]$ gives

$$|N_{\widetilde{G}}(y,a;S) - N_{\widetilde{G}}(y,a;S')| + |D_{\widetilde{G}}(a;S) - D_{\widetilde{G}}(a;S')| \leq C_K \|S - S'\|.$$

Finally, since for $|a| \leq K$, $D_{\widetilde{G}}(a;S),\ D_{\widetilde{G}}(a;S') \geq d_K > 0$, the similar quotient bound used above yields

$$|F_{\widetilde{G}}(y \mid a;S) - F_{\widetilde{G}}(y \mid a;S')| \leq C_K \|S - S'\|,$$

uniformly over $\widetilde{G} \in \mathcal{P}([-M,M])$, $|a| \leq K$, and $y \in \mathbb{R}$. □

The final lemma combines the preceding bounds to compare the estimated conditional CDF with its oracle and matched-normal counterparts.

Lemma 24. Let $V_i = F_G(Y_i \mid A_i; \Sigma_i)$, $U_i^G = F_G(Y_i^G \mid A_i^G; \Sigma_i)$, and $\widehat{U}_i = F_{\widehat{G}}(Y_i \mid A_i; \widehat{\Sigma}_i)$. On the actual data space, $n^{-1} \sum_{i=1}^n \mathbb{E}_{\boldsymbol{\theta},G} \left[|\widehat{U}_i - V_i| \right] \to 0$; under separate per-task normal couplings, $n^{-1} \sum_{i=1}^n \mathbb{E} \left[|V_i - U_i^G| \right] \to 0$.

Proof. For each i, conditionally on b_i, couple the actual pair (Y_i, A_i) with its matched normal pair (Y_i^G, A_i^G) by an optimal W_2 coupling. By Lemma 22,

$$|V_i - U_i^G| \leq L \|(Y_i, A_i) - (Y_i^G, A_i^G)\|.$$

Under the optimal W_2 coupling,

$$\mathbb{E} \left[\|(Y_i, A_i) - (Y_i^G, A_i^G)\|^2 \mid b_i \right] = W_2^2 \left(\mathcal{L}_{\boldsymbol{\theta},G}((Y_i, A_i) \mid b_i), \mathrm{N}_2((0, b_i)^\top, \Sigma_i) \right),$$

and then averaging first over $b_i \sim G$ and then over i, and applying Cauchy–Schwarz,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|V_i - U_i^G| &\le \frac{L}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|(Y_i, A_i) - (Y_i^G, A_i^G)\right\|\right] \\
&\le L\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|(Y_i, A_i) - (Y_i^G, A_i^G)\right\|^2\right]\right\}^{1/2} \\
&= L\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{b_i\sim G}\left[W_2^2\big(\mathcal{L}((Y_i, A_i) \mid b_i), \mathrm{N}_2((0, b_i)^\top, \Sigma_i)\big)\right]\right\}^{1/2} \\
&= L\mathcal{W}_n \longrightarrow 0.
\end{aligned}$$

For $K > 0$, it holds that

$$\begin{aligned}
D_n &:= \frac{1}{n}\sum_{i=1}^{n}\left|F_{\widehat{G}}(Y_i \mid A_i; \widehat{\Sigma}_i) - F_G(Y_i \mid A_i; \Sigma_i)\right| \\
&\le \frac{1}{n}\sum_{i=1}^{n}\left|F_{\widehat{G}}(Y_i \mid A_i; \widehat{\Sigma}_i) - F_G(Y_i \mid A_i; \widehat{\Sigma}_i)\right|\mathbb{1}(|A_i| \le K) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\left|F_G(Y_i \mid A_i; \widehat{\Sigma}_i) - F_G(Y_i \mid A_i; \Sigma_i)\right|\mathbb{1}(|A_i| \le K) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\left|F_{\widehat{G}}(Y_i \mid A_i; \widehat{\Sigma}_i) - F_G(Y_i \mid A_i; \Sigma_i)\right|\mathbb{1}(|A_i| > K) \\
&\le C_K W_1(\widehat{G}, G) + \frac{C_K}{n}\sum_{i=1}^{n}\|\widehat{\Sigma}_i - \Sigma_i\| + \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(|A_i| > K),
\end{aligned}$$

where on $\{|A_i| \le K\}$ we apply Lemma 23, on $\{|A_i| > K\}$ we use $|F_{\widehat{G}} - F_G| \le 1$. By Lemma 20, 21, Assumption 8, we have that

$$W_1(\widehat{G}, G) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta},G}} 0, \quad \frac{1}{n}\sum_{i=1}^{n}\|\widehat{\Sigma}_i - \Sigma_i\| \xrightarrow{\mathbb{P}_{\boldsymbol{\theta},G}} 0, \quad \lim_{K\to\infty}\limsup_n \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{\boldsymbol{\theta},G}\left[|A_i| > K\right] = 0.$$

Hence, for every $\eta > 0$,

$$\lim_{K\to\infty}\limsup_{n\to\infty}\mathbb{P}_{\boldsymbol{\theta},G}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(|A_i| > K) > \eta\right] \le \lim_{K\to\infty}\limsup_{n\to\infty}\frac{1}{\eta n}\sum_{i=1}^{n}\mathbb{P}_{\boldsymbol{\theta},G}\left[|A_i| > K\right] = 0.$$

Therefore, choosing K sufficiently large and then letting $n \to \infty$, we obtain $D_n \xrightarrow{\mathbb{P}_{\boldsymbol{\theta},G}} 0$ and $\mathbb{E}_{\boldsymbol{\theta},G}[D_n] \to 0$, since $0 \le D_n \le 1$. □

We now combine the preceding results to provide coverage guarantee.

Proof of Theorem 10. Take $C_\alpha = [\alpha/2, 1 - \alpha/2]$, and by definition

$$\theta_i \in \mathcal{I}^{\mathrm{rb}}_{\widehat{G},i}(1-\alpha) \iff \frac{\alpha}{2} \le F_{\widehat{G}}\left(\widehat{\theta}_i^{\mathrm{db}} - \theta_i \mid \widehat{b}_i; \widehat{\Sigma}_i\right) \le 1 - \frac{\alpha}{2} \iff \widehat{U}_i \in C_\alpha. \tag{S37}$$

It remains to show

$$\frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{P}_{\boldsymbol{\theta},G}\left[\widehat{U}_i \in C_\alpha\right] - (1-\alpha)\right| \to 0.$$

First, since $F_G(\cdot \mid A_i^G; \Sigma_i)$ is the conditional CDF of $Y_i^G \mid A_i^G$ under the matched normal model, the probability integral transform gives

$$U_i^G \mid A_i^G \sim \mathrm{Unif}(0,1), \qquad U_i^G \sim \mathrm{Unif}(0,1).$$

For $\varepsilon > 0$, define

$$C_\alpha^{+\varepsilon} = \left[\frac{\alpha}{2} - \varepsilon, 1 - \frac{\alpha}{2} + \varepsilon\right] \cap [0,1], \quad C_\alpha^{-\varepsilon} = \left[\frac{\alpha}{2} + \varepsilon, 1 - \frac{\alpha}{2} - \varepsilon\right],$$

with $C_\alpha^{-\varepsilon} = \varnothing$ if the left endpoint exceeds the right endpoint. On the actual data space,

$$\{\widehat{U}_i \in C_\alpha\} \subseteq \left\{V_i \in C_\alpha^{+\varepsilon/2}\right\} \cup \left\{|\widehat{U}_i - V_i| > \frac{\varepsilon}{2}\right\},$$
$$\left\{V_i \in C_\alpha^{-\varepsilon/2}\right\} \subseteq \{\widehat{U}_i \in C_\alpha\} \cup \left\{|\widehat{U}_i - V_i| > \frac{\varepsilon}{2}\right\}.$$

Since the marginal law of V_i is preserved under the chosen per-task coupling,

$$\mathbb{P}_{\boldsymbol{\theta},G}\left[V_i \in C_\alpha^{+\varepsilon/2}\right] = \mathbb{P}\left[V_i \in C_\alpha^{+\varepsilon/2}\right],$$

and similarly for $C_\alpha^{-\varepsilon/2}$. Under the per-task coupling,

$$\left\{V_i \in C_\alpha^{+\varepsilon/2}\right\} \subseteq \{U_i^G \in C_\alpha^{+\varepsilon}\} \cup \left\{|V_i - U_i^G| > \frac{\varepsilon}{2}\right\}$$
$$\{U_i^G \in C_\alpha^{-\varepsilon}\} \subseteq \left\{V_i \in C_\alpha^{-\varepsilon/2}\right\} \cup \left\{|V_i - U_i^G| > \frac{\varepsilon}{2}\right\}.$$

Hence

$$\mathbb{P}_{\boldsymbol{\theta},G}\left[\widehat{U}_i \in C_\alpha\right] \le \mathbb{P}\left[U_i^G \in C_\alpha^{+\varepsilon}\right] + \mathbb{P}_{\boldsymbol{\theta},G}\left[|\widehat{U}_i - V_i| > \frac{\varepsilon}{2}\right] + \mathbb{P}\left[|V_i - U_i^G| > \frac{\varepsilon}{2}\right],$$
$$\mathbb{P}_{\boldsymbol{\theta},G}\left[\widehat{U}_i \in C_\alpha\right] \ge \mathbb{P}\left[U_i^G \in C_\alpha^{-\varepsilon}\right] - \mathbb{P}_{\boldsymbol{\theta},G}\left[|\widehat{U}_i - V_i| > \frac{\varepsilon}{2}\right] - \mathbb{P}\left[|V_i - U_i^G| > \frac{\varepsilon}{2}\right].$$

Since $U_i^G \sim \mathrm{Unif}(0,1)$,

$$\mathbb{P}\left[U_i^G \in C_\alpha^{+\varepsilon}\right] \le 1 - \alpha + 2\varepsilon, \quad \mathbb{P}\left[U_i^G \in C_\alpha^{-\varepsilon}\right] \ge 1 - \alpha - 2\varepsilon.$$

Therefore,

$$\left|\mathbb{P}_{\boldsymbol{\theta},G}\left[\widehat{U}_i \in C_\alpha\right] - (1-\alpha)\right| \le \mathbb{P}_{\boldsymbol{\theta},G}\left[|\widehat{U}_i - V_i| > \frac{\varepsilon}{2}\right] + \mathbb{P}\left[|V_i - U_i^G| > \frac{\varepsilon}{2}\right] + 2\varepsilon.$$

Averaging over i and applying Markov's inequality gives

$$\frac{1}{n}\sum_{i=1}^n \left|\mathbb{P}_{\boldsymbol{\theta},G}\left[\widehat{U}_i \in C_\alpha\right] - (1-\alpha)\right| \le \frac{2}{n\varepsilon}\sum_{i=1}^n \left\{\mathbb{E}_{\boldsymbol{\theta},G}|\widehat{U}_i - V_i| + \mathbb{E}|V_i - U_i^G|\right\} + 2\varepsilon.$$

By Lemma 24,

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{\theta},G}|\widehat{U}_i - V_i| \longrightarrow 0, \quad \frac{1}{n}\sum_{i=1}^n \mathbb{E}|V_i - U_i^G| \longrightarrow 0.$$

Thus, for every fixed $\varepsilon > 0$,

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \left|\mathbb{P}_{\boldsymbol{\theta},G}\left[\widehat{U}_i \in C_\alpha\right] - (1-\alpha)\right| \le 2\varepsilon.$$

Letting $\varepsilon \downarrow 0$ and combining with (S37), we conclude that

$$\frac{1}{n}\sum_{i=1}^n \left|\mathbb{P}_{\boldsymbol{\theta},G}\left[\theta_i \in \mathcal{I}_{\widehat{G},i}^{\mathrm{rb}}(1-\alpha)\right] - (1-\alpha)\right| \longrightarrow 0.$$

□

D.8 Proof of Theorem 11

The argument follows the proof of Theorem 10, with all expectations and convergence statements taken under the hierarchical sampling law $\mathbb{P}_{\boldsymbol{H}}$. Since each b_i has marginal distribution G, and Assumption 7 controls the conditional law of (Y_i, A_i) given b_i, the proof of Lemma 20 carries over under $\mathbb{P}_{\boldsymbol{H}}$ using the same second-moment conditions. The proof of Lemma 21 then applies under Assumptions 7–9, yielding $W_1(\widehat{G}, G) \xrightarrow{\mathbb{P}_{\boldsymbol{H}}} 0$, which is used in the conditional CDF approximation below. Lemma 22 and 23 can be applied directly. It remains to verify the conditional CDF approximation used in Lemma 24. For each i, conditionally on b_i, couple the actual pair (Y_i, A_i), whose conditional law marginalizes over $\theta_i \mid b_i$, with its matched normal pair (Y_i^G, A_i^G) using an optimal W_2 coupling. By Lemma 22 and Cauchy-Schwarz,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[|V_i - U_i^G|\right] &\le \frac{L}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|(Y_i, A_i) - (Y_i^G, A_i^G)\right\|\right] \\ &\le L\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{b_i\sim G}\left[W_2^2(\mathcal{L}_{\boldsymbol{H}}((Y_i, A_i) \mid b_i), \mathrm{N}_2((0, b_i)^\top, \Sigma_i))\right]\right\}^{1/2} \\ &= L\mathcal{W}_n \longrightarrow 0,\end{aligned}$$

Moreover, Lemma 20, 21, and 23, together with Assumption 8, yield

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{H}}\left[|\widehat{U}_i - V_i|\right] \longrightarrow 0$$

by the same argument as in Lemma 24. Finally, the interval inversion identity (S37) holds pointwise, regardless of whether θ_i is fixed or random. Also, the probability integral transform under the matched normal working model gives $U_i^G \sim \mathrm{Unif}(0, 1)$, then the same coverage-transfer argument as in the proof of Theorem 10 yields, for every fixed $\varepsilon > 0$,

$$\frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{P}_{\boldsymbol{H}}\left[\widehat{U}_i \in C_\alpha\right] - (1-\alpha)\right| \le \frac{2}{n\varepsilon}\sum_{i=1}^{n}\left\{\mathbb{E}_{\boldsymbol{H}}|\widehat{U}_i - V_i| + \mathbb{E}|V_i - U_i^G|\right\} + 2\varepsilon.$$

The two averaged error terms converge to zero by the preceding arguments. Consequently,

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{P}_{\boldsymbol{H}}\left[\widehat{U}_i \in C_\alpha\right] - (1-\alpha)\right| \le 2\varepsilon.$$

Letting $\varepsilon \downarrow 0$ and applying the interval inversion identity, we conclude that

$$\frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{P}_{\boldsymbol{H}}\left[\theta_i \in \mathcal{I}^{\mathrm{rb}}_{\widehat{G},i}(1-\alpha)\right] - (1-\alpha)\right| \longrightarrow 0.$$

D.9 Proof of Theorem 12

Let $G_n = n^{-1}\sum_{i=1}^{n}\delta_{b_i}$ denote the empirical distribution of the fixed biases. The proof largely follows the structure of Theorem 10. Lemmas 22 and 23 apply directly with $\widetilde{G} = G_n$,and the arguments of Lemmas 20, 21, and 24 are adapted to the compound setting as follows.

Lemma 25. Let $\mu_i = \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{b}}(A_i)$ and $\nu_i = \mathrm{N}(b_i, \Sigma_{22})$. Under the assumptions of Theorem 12,

$$\frac{1}{n}\sum_{i=1}^{n}W_1(\mu_i, \nu_i) \le \mathcal{W}_n \to 0, \qquad \lim_{T\to\infty}\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{b}}[A_i^2\mathbf{1}\{|A_i| > T\}] = 0,$$

and, on writing $\widehat{\tau}_i^2 = \widehat{\Sigma}_{i,22}$, $\tau^2 = \Sigma_{22}$,

$$\frac{1}{n}\sum_{i=1}^{n}|\widehat{\tau}_i - \tau| \xrightarrow{\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{b}}} 0.$$

Proof. Let $d_i = W_2\left(\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{b}}((Y_i, A_i)), \mathrm{N}_2((0, b_i)^\top, \Sigma)\right)$. Projection gives $W_1(\mu_i, \nu_i) \le d_i$, and hence, by Cauchy–Schwarz,

$$\frac{1}{n}\sum_i W_1(\mu_i, \nu_i) \le \left(\frac{1}{n}\sum_i d_i^2\right)^{1/2} = \mathcal{W}_n.$$

For the tail bound, the proof of Lemma 20 applies directly. In the present setting, b_i is deterministic and $A_i^G = b_i + \tau Z_i$, where $Z_i \sim \mathrm{N}(0,1)$, $|b_i| \le M$, and $\tau^2 \le \overline{\lambda}$. Hence the same uniform bound as in Lemma 20 holds, and therefore

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{b}}\left[A_i^2 \mathbf{1}\{|A_i| > T\}\right] \le 4\mathcal{W}_n^2 + \frac{C}{T^2}.$$

Letting $n \to \infty$ and then $T \to \infty$ proves the second claim. The final claim follows exactly as in Lemma 20 with τ_i replaced by the common τ and Σ_i by Σ. □

Lemma 26. Under the assumptions of Theorem 12, we have $W_1(\widehat{G}, G_n) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{b}}} 0$.

Proof. The proof follows that of Lemma 21. For $\widetilde{G} \in \mathcal{P}([-M, M])$, define

$$L_n^{\mathrm{true}}(\widetilde{G}) = \frac{1}{n}\sum_{i=1}^n \log f_{\widetilde{G}}(A_i; \Sigma_{22}), \qquad L_n^{\mathrm{est}}(\widetilde{G}) = \frac{1}{n}\sum_{i=1}^n \log f_{\widetilde{G}}(A_i; \widehat{\Sigma}_{i,22}).$$

We record only the changes specific to the compound setting. As in Lemma 21, the variance-derivative bound together with Lemma 25 gives

$$R_n := \sup_{\widetilde{G}\in\mathcal{P}([-M,M])} \left|L_n^{\mathrm{est}}(\widetilde{G}) - L_n^{\mathrm{true}}(\widetilde{G})\right| = o_{\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{b}}}(1).$$

Since $\widehat{G}$ is a global maximizer of L_n^{est} over $\mathcal{P}([-M, M])$ and $G_n \in \mathcal{P}([-M, M])$, we have $L_n^{\mathrm{est}}(\widehat{G}) \ge L_n^{\mathrm{est}}(G_n)$. Consequently,

$$L_n^{\mathrm{true}}(\widehat{G}) - L_n^{\mathrm{true}}(G_n) \ge -2R_n = -o_{\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{b}}}(1).$$

Set $g_{n,\widetilde{G}}(a) = \log\{f_{\widetilde{G}}(a; \Sigma_{22})/f_{G_n}(a; \Sigma_{22})\}$. Since G_n is deterministic and both G_n and $\widetilde{G}$ belong to $\mathcal{P}([-M, M])$, the derivative bounds, truncation argument, and finite-W_1-net argument from Lemma 21 carry over. In particular, Assumption 9 applies to the triangular sequence $g_{n,\widetilde{G}}$ for each fixed $\widetilde{G}$, yielding

$$\sup_{\widetilde{G}} \left|\frac{1}{n}\sum_{i=1}^n \left(g_{n,\widetilde{G}}(A_i) - \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{b}}\left[g_{n,\widetilde{G}}(A_i)\right]\right)\right| = o_{\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{b}}}(1).$$

Likewise, using the same Lipschitz bound as in Lemma 21 together with Lemma 25,

$$\sup_{\widetilde{G}\in\mathcal{P}([-M,M])} \left|\frac{1}{n}\sum_{i=1}^n \left(\mathbb{E}_{\mu_i} g_{n,\widetilde{G}} - \mathbb{E}_{\nu_i} g_{n,\widetilde{G}}\right)\right| \le \frac{2M}{\underline{\lambda}}\frac{1}{n}\sum_{i=1}^n W_1(\mu_i, \nu_i) = o(1).$$

Since $\nu_i = \mathrm{N}(b_i, \Sigma_{22})$ and $n^{-1}\sum_{i=1}^n \nu_i$ has density $f_{G_n}(\cdot; \Sigma_{22})$,

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\nu_i} g_{n,\widetilde{G}} = \int f_{G_n}(a; \tau^2) \log \frac{f_{\widetilde{G}}(a; \tau^2)}{f_{G_n}(a; \tau^2)}\, \mathrm{d}a = -\mathrm{KL}\left(f_{G_n}(\cdot; \tau^2) \| f_{\widetilde{G}}(\cdot; \tau^2)\right).$$

Together with $L_n^{\mathrm{true}}(\widehat{G}) - L_n^{\mathrm{true}}(G_n) \ge -2R_n$, this yields

$$\begin{aligned} 0 &\le \mathrm{KL}\left(f_{G_n}(\cdot; \tau^2) \| f_{\widehat{G}}(\cdot; \tau^2)\right) \\ &\le \sup_{\widetilde{G}\in\mathcal{P}([-M,M])} \left|L_n^{\mathrm{true}}(\widetilde{G}) - L_n^{\mathrm{true}}(G_n) + \mathrm{KL}\left(f_{G_n}(\cdot; \tau^2) \| f_{\widetilde{G}}(\cdot; \tau^2)\right)\right| + 2R_n = o_{\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{b}}}(1), \end{aligned}$$

and therefore

$$\mathfrak{H}^2\big(f_{\widehat{G}}(\cdot;\tau^2), f_{G_n}(\cdot;\tau^2)\big) \le \mathrm{KL}\big(f_{G_n}(\cdot;\tau^2)\|f_{\widehat{G}}(\cdot;\tau^2)\big) = o_{\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{b}}}(1).$$

To account for the moving target G_n, define

$$c_\varepsilon = \inf_{\substack{G,\widetilde{G}\in\mathcal{P}([-M,M]) \\ W_1(G,\widetilde{G})\ge\varepsilon,\ v\in[\underline{\lambda},\overline{\lambda}]}} \mathfrak{H}^2\big(f_G(\cdot;v), f_{\widetilde{G}}(\cdot;v)\big).$$

Compactness, continuity, and identifiability of normal convolution imply $c_\varepsilon > 0$ (similar to Lemma 21). Hence

$$c_\varepsilon \mathbb{1}(W_1(\widehat{G}, G_n) \ge \varepsilon) \le \mathfrak{H}^2\big(f_{\widehat{G}}(\cdot;\tau^2), f_{G_n}(\cdot;\tau^2)\big) = o_{\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{b}}}(1),$$

which proves

$$W_1(\widehat{G}, G_n) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{b}}} 0.$$

□

Lemma 27. Define $V_i = F_{G_n}(Y_i \mid A_i;\Sigma)$, $U_i^{G_n} = F_{G_n}(Y_i^G \mid A_i^G;\Sigma)$, and $\widehat{U}_i = F_{\widehat{G}}(Y_i \mid A_i;\widehat{\Sigma}_i)$. Then, on the actual data space, $n^{-1}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{b}}\left[|\widehat{U}_i - V_i|\right] \to 0$; under separate per-task optimal normal couplings, $n^{-1}\sum_{i=1}^n \mathbb{E}\left[|V_i - U_i^{G_n}|\right] \to 0$.

Proof. The second statement follows exactly as in Lemma 24 by applying Lemma 22 with $\widetilde{G} = G_n$ and $S = \Sigma$:

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}|V_i - U_i^{G_n}| \le \frac{L}{n}\sum_{i=1}^n \mathbb{E}\left[\|(Y_i, A_i) - (Y_i^G, A_i^G)\|\right] \le L\mathcal{W}_n \to 0.$$

For the first statement, the same truncation argument as in Lemma 24 together with Lemma 23 gives, for every $K > 0$,

$$D_n := \frac{1}{n}\sum_{i=1}^n |\widehat{U}_i - V_i| \le C_K W_1(\widehat{G}, G_n) + \frac{C_K}{n}\sum_{i=1}^n \|\widehat{\Sigma}_i - \Sigma\| + \frac{1}{n}\sum_{i=1}^n 1\{|A_i| > K\}.$$

By Lemmas 25 and 26, together with Assumption 8, the right-hand side converges to zero in $\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{b}}$-probability after first letting $n \to \infty$ and then $K \to \infty$. Hence $D_n \to 0$ in $\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{b}}$-probability. Since $0 \le D_n \le 1$, uniform integrability gives $\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{b}}[D_n] \to 0$. □

Proof of Theorem 12. Fix $\alpha \in (0,1)$ and let $C_\alpha = [\alpha/2, 1-\alpha/2]$. As in the proof of Theorem 10,

$$\theta_i \in \mathcal{I}^{\mathrm{rb}}_{\widehat{G},i}(1-\alpha) \quad \Longleftrightarrow \quad \widehat{U}_i \in C_\alpha.$$

It remains to establish the oracle calibration appropriate for the compound setting. Let

$$J \sim \mathrm{Unif}\{1,\dots,n\}, \qquad (Y_J^G, A_J^G) \mid (J = i) \sim \mathrm{N}_2\left(\begin{pmatrix}0\\ b_i\end{pmatrix}, \Sigma\right).$$

Since J is uniform, $b_J \sim G_n$, the marginal law of (Y_J^G, A_J^G) is the normal working model with mixing distribution G_n and covariance Σ. Hence the probability integral transform gives

$$U_J^{G_n} = F_{G_n}(Y_J^G \mid A_J^G;\Sigma) \sim \mathrm{Unif}(0,1).$$

Equivalently, for every measurable $C \subseteq [0,1]$,

$$\frac{1}{n}\sum_{i=1}^n \mathbb{P}\left[U_i^{G_n} \in C\right] = \mathrm{Leb}(C),$$

where $\mathrm{Leb}(C)$ denotes the Lebesgue measure of C. Applying the same enlargement and shrinkage argument as in the proof of Theorem 10, with $V_i = F_{G_n}(Y_i \mid A_i;\Sigma)$, for every sufficiently small $\varepsilon > 0$,

$$\left|\frac{1}{n}\sum_{i=1}^n \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{b}}\left[\widehat{U}_i \in C_\alpha\right] - (1-\alpha)\right| \le 2\varepsilon + \frac{2}{n\varepsilon}\sum_{i=1}^n \left\{\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{b}}|\widehat{U}_i - V_i| + \mathbb{E}|V_i - U_i^{G_n}|\right\}.$$

By Lemma 27,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{b}}|\widehat{U}_i - V_i| \to 0, \qquad \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|V_i - U_i^{G_n}| \to 0.$$

Therefore, letting $n \to \infty$ and then $\varepsilon \downarrow 0$,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{b}}\left[\theta_i \in \mathcal{I}^{\mathrm{rb}}_{\widehat{G},i}(1-\alpha)\right] \to 1-\alpha.$$

□

E Details on numerical studies (Section 6)

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

Computational resources. We run the downstream numerical studies for the PPI setting (LM Arena, Amazon Fine Food Reviews, and Synthetic Amazon) on an HPC cluster with 16-core Intel Xeon CPUs and 32GB memory. For the PPI setting (LM Arena, Amazon Fine Food Reviews, and Synthetic Amazon), we used $K = 200$ random splits for the two real-data experiments and $K = 500$ Monte Carlo replications for the synthetic Amazon experiments. Each downstream interval experiment was completed within 10 minutes. The height GWAS analysis and realistic-LD null simulation were run on a MacBook Pro with a 15-core Apple M5 Pro CPU and 24 GB of memory, taking approximately 6 minutes and 4.2 hours, respectively.

These timings exclude upstream predictor training and scoring, whose hardware requirements and execution instructions are documented in the released code repository[7].

Empirical Bayes confidence intervals. For the EBCI baselines (Armstrong et al., 2022) used in Section 6.1, we use the implementation in the `ebci` R package with its default fourth-moment adjustment. Given task-level estimates Y_i with standard errors σ_i, the procedure estimates the across-task mean δ, residual second moment μ_2, and kurtosis κ, and forms the empirical Bayes estimator $\hat{\theta}_i^{\mathrm{EB}} = \delta + w_i(Y_i - \delta)$, where $w_i = \mu_2/(\mu_2 + \sigma_i^2)$. Thus, when the primary parameters are tightly concentrated across tasks, the procedure shrinks more strongly toward the common mean, while the shrinkage diminishes as across-task heterogeneity increases. EBCIs are then constructed around the shrunken estimates using critical values that incorporate the estimated kurtosis to account for possible non-normality of the primary parameter distribution. In our implementation, we use an intercept-only specification with inverse-variance weights $1/\sigma_i^2$ and the PMT finite-sample correction. We apply the procedure to $(\bar{Y}_i, \widehat{\mathrm{se}}(\bar{Y}_i))$ for Classical EBCI and to $(\hat{\theta}_i^{\mathrm{PT}}, \widehat{\mathrm{se}}(\hat{\theta}_i^{\mathrm{PT}}))$ for PT EBCI.

We additionally provide a Python implementation of the same EBCI construction, including the moment estimators and robust critical-value calculations, for use in our Python-based experiments. In both this Python implementation and the supplementary wrapper around the R `ebci` implementation, tasks with exactly zero plug-in standard error are excluded from the inverse-variance-weighted cross-task moment fit, since $1/\sigma_i^2$ is undefined at $\sigma_i = 0$. Their intervals are instead left as the corresponding unshrunk Wald intervals, which reduce to the point intervals $[\bar{Y}_i, \bar{Y}_i]$ for Classical EBCI and $[\hat{\theta}_i^{\mathrm{PT}}, \hat{\theta}_i^{\mathrm{PT}}]$ for EBCI PT when $\sigma_i = 0$.

E.1 LMArena

Dataset construction. We begin with the public dataset https://huggingface.co/datasets/lmarena-ai/arena-human-preference-140k, which contains 140k pairwise LLM comparison records, where each row consists of two LLMs' responses to a given prompt, and a human's decision of which response is better. Our data cleaning pipeline filters out samples that (i) involve a multi-turn conversation (i.e., the user asks a follow-up question), (ii) have a tied human decision, or (iii) could not be successfully parsed or scored by the reward-model pipeline. 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),

[7] https://github.com/wyling01/eb-rebiasing.

this gives us $n = 298$ LLM pairs that have at least 100 data points (i.e. $M_i + m_i \geq 100$) for our numerical experiments (covering a total of ~41k pairwise comparisons).

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 the Bradley–Terry probabilities $\hat{p}_j$ over the unlabeled comparisons for model pair i. Across all comparisons, the MSE of these probabilities against the observed binary preferences is 0.337.

Table S1: LMArena results across $\alpha \in \{0.05, 0.10, 0.20, 0.30\}$, averaged over $K = 200$ random labeled/unlabeled splits with $n = 298$ pairwise LLM problems. Each cell reports mean $\pm$ 1 Monte-Carlo SE. *Mis-cov* denotes the miscoverage rate. *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 the bias distribution estimated by normal and NPMLE respectively. *EBCI Classical* and *EBCI PT* apply the robust empirical Bayes intervals of Armstrong et al. (2022) to the Classical and PT estimators, respectively. The width-ratio is computed taskwise relative to the Classical interval within each split, excluding task/split pairs with 0 classical width, and is then averaged across tasks within each split and across Monte-Carlo splits.

α		Classical	Pred Mean	PT	Classical EBCI	PT EBCI	RB Normal	RB NPMLE
0.05	mis-cov (%)	4.4 ± 0.1	58.1 ± 0.1	4.9 ± 0.1	3.7 ± 0.1	4.0 ± 0.1	6.1 ± 0.1	9.6 ± 0.2
	width	0.363 ± 0.001	0.150 ± 0.000	0.354 ± 0.001	0.322 ± 0.001	0.316 ± 0.001	0.273 ± 0.001	0.258 ± 0.001
	width-ratio	1.000 ± 0.000	0.416 ± 0.001	0.975 ± 0.001	0.891 ± 0.001	0.875 ± 0.001	0.758 ± 0.001	0.717 ± 0.002
0.10	mis-cov (%)	8.0 ± 0.1	64.6 ± 0.1	8.9 ± 0.1	7.5 ± 0.1	8.0 ± 0.1	11.3 ± 0.2	15.5 ± 0.2
	width	0.304 ± 0.000	0.125 ± 0.000	0.297 ± 0.000	0.270 ± 0.001	0.265 ± 0.001	0.229 ± 0.001	0.216 ± 0.001
	width-ratio	1.000 ± 0.000	0.416 ± 0.001	0.975 ± 0.001	0.892 ± 0.001	0.876 ± 0.001	0.758 ± 0.001	0.714 ± 0.002
0.20	mis-cov (%)	16.3 ± 0.2	72.0 ± 0.1	17.4 ± 0.2	16.3 ± 0.2	16.8 ± 0.2	21.4 ± 0.2	26.4 ± 0.3
	width	0.237 ± 0.000	0.098 ± 0.000	0.231 ± 0.000	0.212 ± 0.000	0.208 ± 0.000	0.178 ± 0.001	0.168 ± 0.001
	width-ratio	1.000 ± 0.000	0.416 ± 0.001	0.975 ± 0.001	0.896 ± 0.001	0.879 ± 0.001	0.758 ± 0.001	0.710 ± 0.002
0.30	mis-cov (%)	25.7 ± 0.2	77.0 ± 0.1	26.7 ± 0.2	25.4 ± 0.2	26.2 ± 0.2	31.2 ± 0.2	36.2 ± 0.3
	width	0.192 ± 0.000	0.079 ± 0.000	0.187 ± 0.000	0.172 ± 0.000	0.168 ± 0.000	0.144 ± 0.001	0.135 ± 0.001
	width-ratio	1.000 ± 0.000	0.416 ± 0.001	0.975 ± 0.001	0.899 ± 0.001	0.882 ± 0.001	0.758 ± 0.001	0.708 ± 0.002

E.2 Amazon Fine Food reviews

Dataset construction. We use the publicly available *Amazon Fine Food Reviews* dataset (McAuley and Leskovec, 2013), hosted by the Stanford Network Analysis Project (SNAP) on Kaggle.[8] Duplicate reviews, defined as reviews sharing the same `(title, text, rating, UserID)` quadruplet, can appear under different `ProductID`s. We therefore deduplicate in three steps: (i) we merge products whose reviews are all identical, since this usually means several `ProductID`s point to the same underlying product; (ii) if two products share only some of their reviews and the overlap is at most 10, we remove the shared reviews from both products; and (iii) if the overlap exceeds 10, we drop the product with fewer total reviews. To avoid data contamination when training our predictor, after selecting the held-out products (the 200 products with most reviews; see below), we further remove 18 pairs of reviews from the training and held-out splits whose tokenized model inputs overlap (i.e. reviews whose title plus text look the same up to capitalization or white spaces). After filtering, 67,619 products with 393,818 reviews remain.

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 (every product has at least 151 reviews, totaling 52,706 reviews). Selecting the top-reviewed products mitigates extreme heteroscedasticity across the per-task variances σ_i^2 that would otherwise

[8] https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews/data

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.

Table S2: Amazon results across $\alpha \in \{0.05, 0.10, 0.20, 0.30\}$, 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	EBCI Classical	EBCI PT	RB Normal	RB NPMLE
0.05	mis-cov (%)	4.8 ± 0.1	6.4 ± 0.1	3.2 ± 0.1	6.3 ± 0.1	3.5 ± 0.1	2.0 ± 0.1	2.0 ± 0.1
	width	0.656 ± 0.001	0.342 ± 0.000	0.412 ± 0.001	0.609 ± 0.001	0.400 ± 0.001	0.343 ± 0.001	0.341 ± 0.001
	width-ratio	1.000 ± 0.000	0.537 ± 0.001	0.646 ± 0.001	0.940 ± 0.001	0.630 ± 0.001	0.538 ± 0.001	0.533 ± 0.001
0.10	mis-cov (%)	8.4 ± 0.1	11.3 ± 0.1	5.7 ± 0.1	10.6 ± 0.1	6.4 ± 0.1	3.6 ± 0.1	3.7 ± 0.1
	width	0.551 ± 0.001	0.287 ± 0.000	0.346 ± 0.001	0.510 ± 0.001	0.336 ± 0.001	0.288 ± 0.001	0.286 ± 0.001
	width-ratio	1.000 ± 0.000	0.537 ± 0.001	0.646 ± 0.001	0.938 ± 0.001	0.630 ± 0.001	0.538 ± 0.001	0.532 ± 0.001
0.20	mis-cov (%)	16.7 ± 0.2	21.0 ± 0.1	12.1 ± 0.2	19.1 ± 0.2	13.6 ± 0.2	8.1 ± 0.1	8.3 ± 0.1
	width	0.429 ± 0.001	0.224 ± 0.000	0.269 ± 0.001	0.398 ± 0.001	0.262 ± 0.001	0.224 ± 0.001	0.222 ± 0.001
	width-ratio	1.000 ± 0.000	0.537 ± 0.001	0.646 ± 0.001	0.940 ± 0.001	0.630 ± 0.001	0.538 ± 0.001	0.531 ± 0.001
0.30	mis-cov (%)	25.9 ± 0.2	30.5 ± 0.2	19.8 ± 0.2	28.6 ± 0.2	22.1 ± 0.2	14.2 ± 0.2	14.7 ± 0.2
	width	0.347 ± 0.001	0.181 ± 0.000	0.218 ± 0.001	0.323 ± 0.001	0.212 ± 0.001	0.181 ± 0.001	0.179 ± 0.001
	width-ratio	1.000 ± 0.000	0.537 ± 0.001	0.646 ± 0.001	0.941 ± 0.001	0.630 ± 0.001	0.538 ± 0.001	0.530 ± 0.001

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[9], 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 our test set (top-200 products) from 66.6% (off-the-shelf checkpoint) to 76.6% (fine-tuned after 2 epochs). We thus use this fine-tuned version for our model h. Fine-tuning uses 341,112 reviews from 67,419 products outside the held-out top 200 ones; none of the held-out review rows is used for fine-tuning. 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.

E.3 Synthetic study

Data generation process. For each task $i \in \{1, \ldots, n\}$ with $n = 200$, we draw a true primary parameter θ_i and a prediction bias b_i from the distributions specified for each setting below. 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.

E.3.1 Sensitivity to the bias distribution

In this setting, we fix the true primary parameter distribution at $\theta_i \sim \mathrm{N}(4, 0.5^2)$ and vary the distribution of b_i (normal or two-point) as described below. The pseudo-ground-truth parameters θ_i in the Amazon data have the sample standard deviation of approximately 0.48, motivating $\sigma_0 = 0.5$ as a comparable reference scale.

Choice of A for the normal bias distribution. 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 bias distribution fitted from the real Amazon split in Section E.2, whose marginal-MLE estimate has standard deviation ≈ 0.045. The smallest value ($A^{1/2} = 0.01$) corresponds to a regime in which the biases are tightly concentrated around their nonzero common mean of -0.1; therefore, pooling across tasks provides strong information about each task-specific bias, enabling stronger shrinkage and shorter rebiased intervals. The largest ($A^{1/2} = 0.1$) is roughly $2\times$ wider than the empirical estimate and allows substantially

[9] https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

greater variation in the biases, weakening the benefit of shrinkage and making the rebiased procedure behave more like the fully debiased procedure. We also choose two intermediate values.

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

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

$A^{1/2}$		**Oracle**	**Classical**	**Pred Mean**	**PT**
0.01	mis-cov (%)	5.0 ± 0.1	5.0 ± 0.1	26.7 ± 0.1	4.9 ± 0.1
	width	0.321 ± 0.001	0.660 ± 0.001	0.341 ± 0.001	0.419 ± 0.001
	width-ratio	0.496 ± 0.001	1.000 ± 0.000	0.531 ± 0.001	0.653 ± 0.001
0.03	mis-cov (%)	5.0 ± 0.1	5.0 ± 0.1	27.2 ± 0.1	4.9 ± 0.1
	width	0.331 ± 0.001	0.660 ± 0.001	0.341 ± 0.001	0.419 ± 0.001
	width-ratio	0.513 ± 0.001	1.000 ± 0.000	0.531 ± 0.001	0.653 ± 0.001
0.05	mis-cov (%)	5.0 ± 0.1	5.0 ± 0.1	28.6 ± 0.1	4.9 ± 0.1
	width	0.344 ± 0.001	0.660 ± 0.001	0.341 ± 0.001	0.419 ± 0.001
	width-ratio	0.537 ± 0.001	1.000 ± 0.000	0.531 ± 0.001	0.653 ± 0.001
0.10	mis-cov (%)	4.9 ± 0.1	5.0 ± 0.1	34.6 ± 0.1	4.9 ± 0.1
	width	0.372 ± 0.001	0.660 ± 0.001	0.341 ± 0.001	0.419 ± 0.001
	width-ratio	0.584 ± 0.001	1.000 ± 0.000	0.531 ± 0.001	0.653 ± 0.001

$A^{1/2}$		**EBCI Classical**	**EBCI PT**	**RB Normal**	**RB NPMLE**
0.01	mis-cov (%)	5.1 ± 0.1	5.0 ± 0.1	5.0 ± 0.1	4.9 ± 0.1
	width	0.614 ± 0.001	0.407 ± 0.001	0.323 ± 0.001	0.325 ± 0.001
	width-ratio	0.942 ± 0.001	0.637 ± 0.001	0.499 ± 0.001	0.501 ± 0.001
0.03	mis-cov (%)	5.1 ± 0.1	5.0 ± 0.1	5.2 ± 0.1	5.2 ± 0.1
	width	0.614 ± 0.001	0.407 ± 0.001	0.330 ± 0.001	0.331 ± 0.001
	width-ratio	0.942 ± 0.001	0.637 ± 0.001	0.512 ± 0.001	0.513 ± 0.001
0.05	mis-cov (%)	5.1 ± 0.1	5.0 ± 0.1	5.2 ± 0.1	5.3 ± 0.1
	width	0.614 ± 0.001	0.407 ± 0.001	0.343 ± 0.001	0.343 ± 0.001
	width-ratio	0.942 ± 0.001	0.637 ± 0.001	0.535 ± 0.001	0.534 ± 0.001
0.10	mis-cov (%)	5.1 ± 0.1	5.0 ± 0.1	5.0 ± 0.1	5.5 ± 0.1
	width	0.614 ± 0.001	0.407 ± 0.001	0.372 ± 0.001	0.369 ± 0.001
	width-ratio	0.942 ± 0.001	0.637 ± 0.001	0.583 ± 0.001	0.576 ± 0.001

Choice of b_0 for the two-point bias distribution. For the misspecification case, 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 model. When $b_0 = 0.05$, the support points are close enough that the induced distribution of the observed bias estimates can be difficult to distinguish from that under a tightly concentrated normal working prior centered near $b_0/2$. As b_0 increases, the support-point separation and prior variance increase, making the discrete two-point structure more pronounced and the normal working model more substantially misspecified. The two intermediate values trace the transition from a nearly normal-approximable regime to a strongly non-normal regime.

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

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

b_0		**Oracle**	**Classical**	**Pred Mean**	**PT**
0.05	mis-cov (%)	4.9 ± 0.1	4.9 ± 0.1	7.8 ± 0.1	4.9 ± 0.1
	width	0.328 ± 0.001	0.660 ± 0.001	0.341 ± 0.001	0.419 ± 0.001
	width-ratio	0.508 ± 0.001	1.000 ± 0.000	0.531 ± 0.001	0.653 ± 0.001
0.10	mis-cov (%)	4.9 ± 0.1	4.9 ± 0.1	15.9 ± 0.1	4.9 ± 0.1
	width	0.342 ± 0.001	0.660 ± 0.001	0.341 ± 0.001	0.419 ± 0.001
	width-ratio	0.533 ± 0.001	1.000 ± 0.000	0.531 ± 0.001	0.653 ± 0.001
0.20	mis-cov (%)	4.9 ± 0.1	4.9 ± 0.1	35.3 ± 0.2	4.9 ± 0.1
	width	0.360 ± 0.001	0.660 ± 0.001	0.341 ± 0.001	0.419 ± 0.001
	width-ratio	0.557 ± 0.001	1.000 ± 0.000	0.531 ± 0.001	0.653 ± 0.001
0.50	mis-cov (%)	4.9 ± 0.1	4.9 ± 0.1	52.1 ± 0.2	4.9 ± 0.1
	width	0.335 ± 0.001	0.660 ± 0.001	0.341 ± 0.001	0.419 ± 0.001
	width-ratio	0.512 ± 0.001	1.000 ± 0.000	0.531 ± 0.001	0.653 ± 0.001

b_0		**EBCI Classical**	**EBCI PT**	**RB Normal**	**RB NPMLE**
0.05	mis-cov (%)	5.1 ± 0.1	4.9 ± 0.1	5.1 ± 0.1	5.0 ± 0.1
	width	0.614 ± 0.001	0.407 ± 0.001	0.327 ± 0.001	0.329 ± 0.001
	width-ratio	0.942 ± 0.001	0.637 ± 0.001	0.507 ± 0.001	0.510 ± 0.001
0.10	mis-cov (%)	5.1 ± 0.1	4.9 ± 0.1	5.0 ± 0.1	5.1 ± 0.1
	width	0.614 ± 0.001	0.407 ± 0.001	0.343 ± 0.001	0.343 ± 0.001
	width-ratio	0.942 ± 0.001	0.637 ± 0.001	0.536 ± 0.001	0.535 ± 0.001
0.20	mis-cov (%)	5.1 ± 0.1	4.9 ± 0.1	4.9 ± 0.1	4.9 ± 0.1
	width	0.614 ± 0.001	0.407 ± 0.001	0.372 ± 0.001	0.362 ± 0.001
	width-ratio	0.942 ± 0.001	0.637 ± 0.001	0.583 ± 0.001	0.562 ± 0.001
0.50	mis-cov (%)	5.1 ± 0.1	4.9 ± 0.1	4.8 ± 0.1	4.7 ± 0.1
	width	0.614 ± 0.001	0.407 ± 0.001	0.405 ± 0.001	0.340 ± 0.001
	width-ratio	0.942 ± 0.001	0.637 ± 0.001	0.633 ± 0.001	0.521 ± 0.001

E.3.2 Sensitivity to primary parameter dispersion

In this setting, we fix the bias distribution at $b_i \sim \mathrm{N}(-0.1, 0.01^2)$ and vary the dispersion of the primary parameter distribution by taking $\theta_i \sim \mathrm{N}(4, \sigma_0^2)$ with $\sigma_0 \in \{0.01, 0.1, 0.5, 1\}$. The range of σ_0 spans primary parameter distributions from highly concentrated to substantially dispersed across tasks. When σ_0 is small, EBCI can benefit substantially from borrowing information across the θ_i's, while this advantage is expected to diminish as σ_0 increases. Table S5 reports average coverage, average width, and average width-ratio for each estimator under different choices of σ_0.

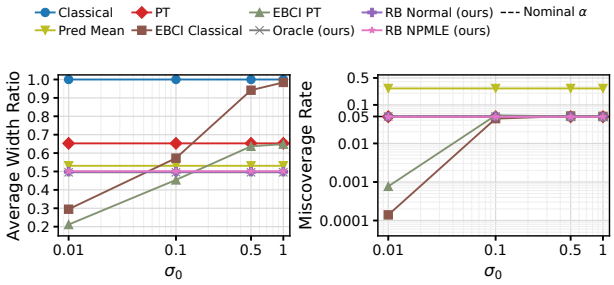


Figure S1: **Interval width ratio and miscoverage rate in the simulation test for primary parameter dispersion.** Average width ratio (left) and miscoverage rate (right) for $n = 200$ simulated tasks in the synthetic Amazon dataset with different levels of primary parameter dispersion. In Table S5 we include the standard errors for the metrics.

Table S5: Simulation results under $\theta_i \sim \mathrm{N}(4, \sigma_0^2)$ with covariance structure adopted from the Amazon data ($n = 200$, $K = 500$ replications). Entries are reported as mean $\pm$ Monte Carlo SE.

σ_0		**Oracle**	**Classical**	**Pred Mean**	**PT**
0.01	mis-cov (%)	5.0 ± 0.1	5.0 ± 0.1	26.7 ± 0.1	4.9 ± 0.1
	width	0.321 ± 0.001	0.660 ± 0.001	0.341 ± 0.001	0.419 ± 0.001
	width-ratio	0.496 ± 0.001	1.000 ± 0.000	0.531 ± 0.001	0.653 ± 0.001
0.10	mis-cov (%)	5.0 ± 0.1	5.0 ± 0.1	26.7 ± 0.1	4.9 ± 0.1
	width	0.321 ± 0.001	0.660 ± 0.001	0.341 ± 0.001	0.419 ± 0.001
	width-ratio	0.496 ± 0.001	1.000 ± 0.000	0.531 ± 0.001	0.653 ± 0.001
0.50	mis-cov (%)	5.0 ± 0.1	5.0 ± 0.1	26.7 ± 0.1	4.9 ± 0.1
	width	0.321 ± 0.001	0.660 ± 0.001	0.341 ± 0.001	0.419 ± 0.001
	width-ratio	0.496 ± 0.001	1.000 ± 0.000	0.531 ± 0.001	0.653 ± 0.001
1.00	mis-cov (%)	5.0 ± 0.1	5.0 ± 0.1	26.7 ± 0.1	4.9 ± 0.1
	width	0.321 ± 0.001	0.660 ± 0.001	0.341 ± 0.001	0.419 ± 0.001
	width-ratio	0.496 ± 0.001	1.000 ± 0.000	0.531 ± 0.001	0.653 ± 0.001

σ_0		**EBCI Classical**	**EBCI PT**	**RB Normal**	**RB NPMLE**
0.01	mis-cov (%)	0.0 ± 0.0	0.1 ± 0.1	5.0 ± 0.1	4.9 ± 0.1
	width	0.175 ± 0.004	0.125 ± 0.003	0.323 ± 0.001	0.325 ± 0.001
	width-ratio	0.295 ± 0.006	0.212 ± 0.004	0.499 ± 0.001	0.501 ± 0.001
0.10	mis-cov (%)	4.4 ± 0.1	5.4 ± 0.1	5.0 ± 0.1	4.9 ± 0.1
	width	0.352 ± 0.002	0.278 ± 0.001	0.323 ± 0.001	0.325 ± 0.001
	width-ratio	0.573 ± 0.003	0.454 ± 0.001	0.499 ± 0.001	0.501 ± 0.001
0.50	mis-cov (%)	5.1 ± 0.1	5.0 ± 0.1	5.0 ± 0.1	4.9 ± 0.1
	width	0.614 ± 0.001	0.407 ± 0.001	0.323 ± 0.001	0.325 ± 0.001
	width-ratio	0.942 ± 0.001	0.637 ± 0.001	0.499 ± 0.001	0.501 ± 0.001
1.00	mis-cov (%)	5.0 ± 0.1	5.0 ± 0.1	5.0 ± 0.1	4.9 ± 0.1
	width	0.647 ± 0.001	0.416 ± 0.001	0.323 ± 0.001	0.325 ± 0.001
	width-ratio	0.984 ± 0.001	0.649 ± 0.001	0.499 ± 0.001	0.501 ± 0.001

E.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

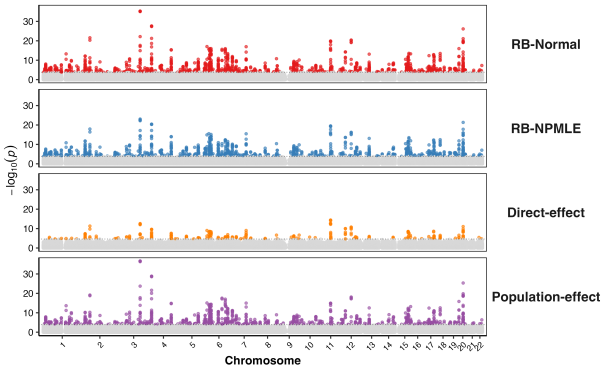


Figure S2: **SNP-level p-values in family-based GWAS.** Manhattan plots of SNP-level p-values, shown on the $-\log_{10}$ scale. Horizontal dashed lines indicate the corresponding B-H thresholds. SNPs passing the threshold are highlighted as significant. For visualization, values exceeding $-\log_{10}(10^{-50}) = 50$ are truncated at 50.

Table S6: 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	514	546	105	533
# overlaps	179	180	71	159

(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. We compare each method's height GWAS discoveries with putatively significant SNPs from the sibling-based direct-effect analysis of Howe et al. (2022), defined using the threshold p-value $< 1 \times 10^{-6}$. We first restrict our discoveries and Howe et al.'s putatively significant SNPs 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, we apply LD clumping (Purcell et al., 2007) to each method's BH discoveries, selecting representative SNPs in order of increasing method-specific unadjusted p-value and excluding nearby correlated SNPs (within 250 kb and $r^2 > 0.1$). We repeat the overlap analysis, with results summarized in Table S6. 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.

Table S7: Empirical FDR comparison across methods in simulation study with realistic LD. Entries report empirical FDR with exact 95% binomial confidence intervals, based on 400 Monte Carlo replicates.

Method	0.5% parental variance	1% parental variance
RB-NPMLE	0.0300 [0.0156, 0.0518]	0.0425 [0.0249, 0.0672]
RB-Normal	0.0600 [0.0388, 0.0880]	0.0800 [0.0554, 0.1111]
Direct-effect	0.0300 [0.0156, 0.0518]	0.0300 [0.0156, 0.0518]
Population-effect	0.0550 [0.0348, 0.0821]	0.1100 [0.0811, 0.1448]

E.5 Simulation study under realistic linkage disequilibrium

We used sim1000G (Dimitromanolakis et al., 2019) and phased 1000 Genomes Phase 3 EUR haplotypes to generate parental haplotypes, from which we simulated Mendelian transmission to 5,000 two-sibling families (10,000 analyzed offspring). We analyzed 6,000 SNPs across 173 approximately independent LDetect regions (Berisa and Pickrell, 2016) on chromosomes 18–22, selecting a contiguous window of 34–35 SNPs centered within each region. Parental genotype sums were imputed using sibling genotypes and IBD states, and association testing used a sibship mixed model (Young et al., 2022). We simulated phenotypes based on the following null model, with zero direct genetic effect ($\theta_i = 0$):

$$Y = \sum_{i=1}^{6000} b_i g_{\mathrm{par},i} + \varepsilon,\ b_i \sim \mathrm{N}(0, \sigma_b^2),\ \varepsilon \sim \mathrm{N}(0, \sigma_\varepsilon^2).$$

To represent a modest but non-negligible parental indirect genetic effect, we rescaled the parental effects b_j (the bias term), so that the parental component explained 0.5% and 1% of phenotypic variance. These choices are motivated by small parental polygenic score (PRS) contributions to height variation (Ghatan et al., 2025; Tan et al., 2025)[10].

We generated 400 phenotype replicates on the same genotype panel and applied BH at $q = 0.05$. Under this global null, every rejection is a false discovery. Therefore, if R denotes the number of rejections, $\mathrm{FDR} = \mathbb{P}(R \geq 1)$ (the family-wise error rate), which we estimated as the proportion of simulation replicates with at least one rejection. Empirical FDRs for our methods in both parental variance regimes are shown in Table S7.

These results suggest that RB-NPMLE remains well calibrated under realistic local LD and parental-effect sizes relevant to traits such as height. Although the bias terms were drawn from a single normal distribution, LD causes the single-SNP estimates to reflect marginal effects that aggregate signals from nearby SNPs. Variation in LD patterns across SNPs can induce heterogeneity in marginal effect variances, so their pooled distribution may depart from a single normal, which may explain the mild inflation observed for RB-Normal. However, for traits with stronger parental effects, LD may compromise calibration even for RB-NPMLE, which we acknowledge as a potential limitation of our methods when applied to family-based GWAS data.

E.6 Data and model licenses

We summarize the licenses and use conditions stated by the providers of the datasets and models used in our LMArena, Amazon and GWAS experiments.

LMArena data. The `arena-human-preference-140k` dataset card states that user prompts are licensed under Creative Commons Attribution 4.0 (CC BY 4.0), while model-generated responses are subject to the respective model providers' terms.[11] Our filtering and preprocessing are described in Appendix E.1.

[10] Using the reported parental PRS coefficients and correlations from the height analyses, we calculate combined parental component variances of approximately 0.02–0.20% for Ghatan et al. (2025) and 0.26–0.32% for Tan et al. (2025). Ghatan et al. (2025) studied childhood height, so we use their estimates as a rough guide for our simulations.

[11] `https://huggingface.co/datasets/lmarena-ai/arena-human-preference-140k`

Skywork reward model. We use the checkpoint `Skywork/Skywork-Reward-V2-Llama-3.1-8B`, whose model card specifies the Llama 3.1 Community License.[12]

Amazon Fine Food Reviews data. We use the Amazon Fine Food Reviews corpus associated with McAuley and Leskovec (2013), downloaded from SNAP.[13] The Kaggle distribution lists CC0 1.0 Universal; the original SNAP download page does not separately state an explicit dataset license.[14] Our deduplication and filtering procedures are described in Appendix E.2.

BERT model. We initialize our predictor from the multilingual BERT sentiment model released by NLP Town[15], whose model card declares the MIT License, and further fine-tune it as described in Appendix E.2.

GWAS data. The GWAS inputs used in our experiments are described in Appendix E.4. SSGAC requires each user to accept its terms and prohibits redistribution.[16] The 1000 Genomes reference data are publicly available and should be cited.[17] OpenGWAS datasets remain subject to the original providers' terms.[18] We release a SNP-selection file for reproducibility, but no GWAS summary statistics or reference genotypes.

F Diagnostics for the theoretical assumptions

As a complement to the robustness results in Appendix C, we assess the main approximation assumptions separately for each empirical application, including the plausibility of the normal approximation and plug-in second moments, as well as the extent of cross-task dependence. We also provide qualitative diagnostics for the exchangeability assumptions.

We split our discussion into two parts. Appendix F.1 considers assessments and diagnostics that can be carried out using the data available in practice. Appendix F.2 provides additional diagnostics for the partially Bayes assumption in two PPI applications, using also the labels that we artificially masked in our evaluation.

F.1 Assessment of assumptions in practice

The marginal fits in Figure 2 provide a basic diagnostic for the fitted bias model: a substantial discrepancy between the observed and implied distributions of $\hat{b}_i$ would indicate marginal model misspecification. The close agreement in Figure 2 provides no evidence of such a failure for our three applications.

Since the latent biases b_i are generally unobserved, we examine whether the estimated biases $\hat{b}_i$ exhibit systematic variation with task-specific second-moment quantities $q_i \in \{\rho_i, \tau_i^2, \sigma_i^2\}$. Specifically, for any such q_i, since $\hat{b}_i$ is unbiased for b_i, we have that $\mathbb{E}[\hat{b}_i \mid q_i] = \mathbb{E}[b_i \mid q_i]$. Thus, systematic variation in the conditional mean of $\hat{b}_i$ with q_i can indicate departures from exchangeability. Figure S3 shows the relationship between $\hat{b}_i$ and the task-specific covariance quantities $q_i \in \{\rho_i, \tau_i^2, \sigma_i^2\}$ across the three applications. In practice, the q_i used in Figure S3 are plug-in estimates computed from the same data as $\hat{b}_i$ and are treated as fixed. Since these estimates may share estimation noise with $\hat{b}_i$, we only view Figure S3 as a qualitative diagnostic.

As a possible extension, one could also examine second-moment departures from exchangeability using $\mathrm{Var}[b_i \mid q_i] = \mathrm{Var}[\hat{b}_i \mid q_i] - \mathbb{E}[\tau_i^2 \mid q_i]$, and assess whether the conditional variance of b_i varies systematically with q_i. We leave this and potential extensions to higher-order moments for future work.

[12] https://huggingface.co/Skywork/Skywork-Reward-V2-Llama-3.1-8B
[13] https://snap.stanford.edu/data/web-FineFoods.html
[14] https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews
[15] https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment
[16] https://thessgac.com/register/
[17] https://www.internationalgenome.org/IGSR_disclaimer/
[18] https://opengwas.io/policies/eula

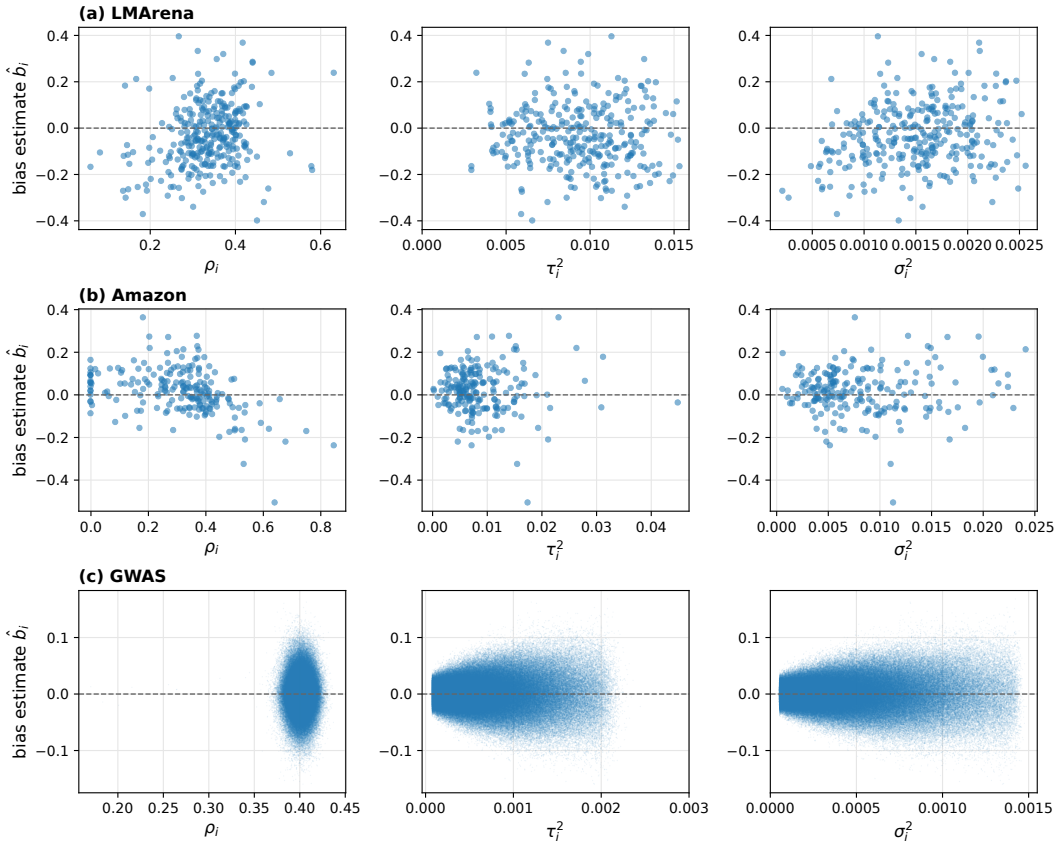


Figure S3: **Exchangeability diagnostic: Scatter plot of bias estimates vs. nuisance parameters in our three applications.**

LMArena. For a representative random labeled/unlabeled (20/80) split, the 298 model-pair tasks contain 20–55 labeled observations and 80–221 unlabeled observations. Given these small sample sizes and binary outcomes Y_i, the normal approximation may be poor, and the plug-in second moments may also be imprecisely estimated. Each comparison belongs to a single LLM-pair task, so substantial cross-task dependence due to overlapping observations is less of a concern, although some dependence may still arise when different model-pair tasks share a common model. Fig. S3a reveals some systematic variation in the estimated biases with the nuisance quantities, suggesting possible departures from exchangeability.

Amazon reviews. For a representative random labeled/unlabeled (20/80) split of the Amazon data, each task contains 30–182 labeled observations and 121–730 unlabeled observations. The normal approximation may be more plausible here than for LM Arena, given the larger labeled sample sizes and the use of five-point ratings for Y_i rather than binary outcomes. Since the product-level tasks are constructed from separate review sets, substantial cross-task dependence is less of a concern. In Fig. S3b, estimated biases $\hat{b}_i$ exhibit only weak associations with the nuisance quantities. Overall, these plots provide no clear evidence of systematic departures from exchangeability with respect to the task-specific covariance quantities.

Family-based GWAS. GWAS summary statistics are obtained from regressions with a sample size of 44,570, so normal approximation and plug-in second moments are expected to be reasonably accurate. Cross-task independence is not satisfied in this example due to LD. Fig. S3c shows no clear systematic relationship between the estimated biases and the nuisance quantities.

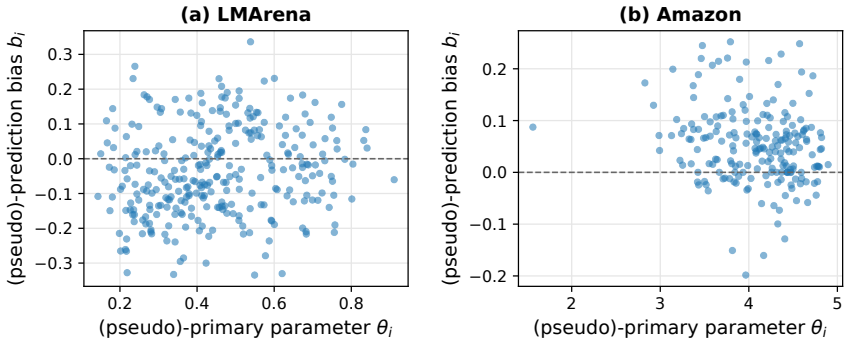


Figure S4: **Diagnostic for the partially Bayes assumption: Scatter plots of (pseudo)-prediction biases vs. (pseudo)-ground-truth primary parameters in our two PPI applications.**

F.2 Diagnostics for the partially Bayes assumption based on pseudo-ground-truth quantities

Here we seek to check whether biases b_i are exchangeable conditional on the primary parameters θ_i. This is required for the conclusion under our partially Bayes framing (e.g., Theorem 4), but not for our coverage results in the hierarchical setting and compound settings (Theorems 5 and 6).

For the two PPI applications (LM Arena and Amazon Reviews), we additionally have access to full-corpus observations, allowing us to construct pseudo-ground-truth for the primary parameters and biases. For each task i, we randomly partition the full corpus using a 50/50 split, with the two resulting subsets denoted by $\mathcal{I}_i^\theta$ and $\mathcal{I}_i^b$. We define the pseudo-primary parameter and pseudo-prediction bias as

$$\tilde{\theta}_i := \frac{1}{|\mathcal{I}_i^\theta|} \sum_{j \in \mathcal{I}_i^\theta} Y_{ij}, \qquad \tilde{b}_i := \frac{1}{|\mathcal{I}_i^b|} \sum_{j \in \mathcal{I}_i^b} (h(X_{ij}) - Y_{ij}).$$

This 50/50 split-sample construction prevents the same outcome noise from entering $\tilde{\theta}_i$ and $\tilde{b}_i$ with opposite signs, which could otherwise induce a mechanical association between them. Using disjoint samples avoids the dependence that would arise from reusing the same outcomes. We then plot $\tilde{b}_i$ against $\tilde{\theta}_i$ across tasks to assess whether prediction bias varies systematically with the primary parameter. If θ_i and b_i are independent, no systematic pattern should appear between their split-sample proxies.

LMArena. Figure S4a shows substantial variation in the pseudo-ground-truth biases, but no clear systematic relationship with the primary parameter θ_i.

Amazon reviews. Figure S4b shows a mild association between the pseudo-biases b_i and the pseudo-parameters θ_i, although the observations are substantially dispersed. This diagnostic suggests that the independence assumption underlying the partially Bayes formulation, $\theta_i \perp b_i$ with $b_i \sim G$ and θ_i fixed, may hold only approximately in this application (see Section 3). However, this pattern does not directly contradict our more general theoretical results, which allow dependence between the primary parameters and biases in the hierarchical setting and impose no independence assumption between the fixed (θ_i, b_i) sequences in the compound setting (see Section 5).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and §3 state the empirical Bayes rebiasing methodology with both Normal and NPMLE bias distributions. The theoretical claims are established in §5 and Appendix C, while the empirical claims are supported by the experiments in §6.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   The paper discusses limitations at the relevant points throughout the main text and appendices. These include possible departures from the task-level normal approximation and imprecise plug-in second-moment estimates for small samples with binary outcomes and the possibility that stronger parental effects may compromise calibration in family-based GWAS.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical results are numbered and cross-referenced, with their assumptions stated or explicitly referenced. The main and robustness results appear in §5 and Appendix C, respectively, and complete proofs are provided in Appendix D.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All experiments use publicly available datasets. See Appendix E for experimental details.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All datasets used in the experiments are publicly available (citations and URLs given in Appendix E). The code and instructions for reproducing the experiments and figures are available at https://github.com/wyling01/eb-rebiasing.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We have specified these experimental details either directly in the main §6 section, or in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All confidence-interval experiments report empirical miscoverage and interval width, averaged over Monte Carlo replications and we also provided Monte Carlo standard errors. We use $K = 500$ replications for the synthetic Amazon experiments and $K = 200$ for both real-data experiments (Amazon and LMArena). Thus, the reported values estimate the corresponding population quantities with Monte Carlo standard errors of order $1/\sqrt{K}$. The variability underlying these standard errors arises from the random labeled/unlabeled

splits in the Amazon and LMArena experiments and from the random draws of b_i and θ_i in the synthetic experiments. Monte Carlo standard errors are reported numerically in the corresponding Tables S1–S5. Error bars are omitted from the figures for visual clarity, since they are negligible at the scale shown.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the compute resources and runtimes for the downstream statistical analyses at the start of Appendix E. Hardware requirements and execution instructions for the upstream BERT fine-tuning and Skywork-Reward-V2 scoring pipelines are provided in the released code repository.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is methodological and uses only publicly released, de-identified datasets cited in Appendix E; we have reviewed the NeurIPS Code of Ethics and the research conforms to it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is foundational methodological research on calibrated frequentist inference; we expect the primary effect to be more reliable uncertainty quantification in downstream scientific applications (e.g. LLM evaluation, GWAS, prediction-powered inference). We do not foresee any direct path from the methods themselves to negative societal applications.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The released artifact contains methods code and processed analysis inputs, but no model weights, raw prompts or responses, or raw review text. We do not consider additional safeguards necessary for this release.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets, models, and third-party code are properly credited, with their licenses or provider-specific terms documented in Appendix E.6. Corresponding license and attribution notices are also included in the code repository.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new asset is the public reproduction-code repository, which contains the empirical Bayes rebiasing intervals implementation and the scripts used for all experiments and figures. The accompanying README files document the dependencies, data requirements, external software, and instructions for reproducing the results.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not run any crowdsourcing studies or original human-subject experiments.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not collect or interact with human subjects; so no IRB review of the present work is required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core methodology is statistical and does not involve LLMs. The LMArena experiment uses a publicly released reward model (`Skywork-Reward-V2` (Liu et al., 2026)) only as a black-box predictor to demonstrate prediction-powered inference; it is not a methodological component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.