The Limits of Experimental Design: Covariate Balance Beyond Low Dimension

Max Cytrynbaum*

July 2026

Abstract

We study how quickly experimental designs can reduce the gap between the variance of an unadjusted estimator of the average treatment effect and the semiparametric variance bound. We show an impossibility theorem: under weak conditions, no design can approach the variance bound uniformly over smooth outcome models unless covariate dimension $d = o(\log n)$. For experiments with thousands of units, this very-low dimensional regime allows only a handful of covariates. Motivated by this, we propose restricted efficiency targets for balancing lower-complexity nonparametric function classes. We construct new discrepancy minimizing designs that approach these restricted variance targets at fast rates. For example, we can allow approximately $d = o(n)$ covariates in an additive nonparametric specification. These designs can be further robustified by combining them with matching. In simulations calibrated to 12 published experiments, our designs reduce variance relative to matched pairs in every experiment.

Keywords: Semiparametric Efficiency, Kernel Methods, Discrepancy Minimization.
JEL Codes: C14, C21, C90.

*Department of Economics, Yale University.

1 Introduction

Balancing covariates is a central task of experimental design. Available methods range from stratified randomization, in use at least since Fisher (1926), to rerandomization and newer discrepancy minimization methods (Li et al., 2018; Harshaw et al., 2024). Effective covariate balance can improve precision, reducing the size and cost of the experiment required to study a given causal effect. For example, a survey of 200 randomized experiments in recent NBER working papers reported in Cytrynbaum (2026) finds that 43% used some form of stratification.

Recent theoretical work provides a strong justification for finely stratified designs like matched-pairs randomization. Under appropriate conditions, such designs make simple unadjusted estimators semiparametrically efficient, automatically attaining the Hahn (1998) variance bound for estimating the average treatment effect (Bai et al., 2022, 2026). However, these guarantees rest on strong conditions, requiring covariate differences within matched groups to vanish asymptotically. This may not be possible in experiments with a realistic number of covariates. Nearest-neighbor distances for n samples in d dimensions typically decrease at rate $n^{-1/d}$, so halving the average distance between matches requires 2^d times as many observations.

To see the practical implications of this curse of dimensionality, consider the OpenResearch Unconditional Income Study, a prominent experiment that randomized 3,000 participants to a universal basic income program, distributing \$40 million in unconditional cash transfers (Broockman et al., 2024). To randomize treatments, researchers first formed matched triples using several dozen covariates, then assigned one participant in each triple to receive \$1,000 per month for three years. For the semiparametric efficiency results above to be relevant here, we would need to obtain near-perfect matches across several dozen covariates, using only $n = 3000$ samples.

Motivated by this gap between theory and practice, we develop a new finite-sample efficiency theory for experimental design. For an unadjusted estimator $\widehat{\theta}$ and efficiency bound V^*, we study the *excess variance* $\mathcal{V}_n \equiv n \operatorname{Var}(\widehat{\theta}) - V^*$. Let ψ_i denote the covariates of unit i and define the outcome model $m(\psi) \equiv E[(Y(1) + Y(0))/2|\psi]$. Also let $Z_i \in \{-1, 1\}$ be the centered treatment assignment with $E[Z] = 0$. A calculation shows that excess variance is determined by in-sample imbalances in the outcome model:

$$\mathcal{V}_n = 4n \cdot E\left[\left(\frac{1}{n}\sum_{i=1}^{n} Z_i m(\psi_i)\right)^2\right]. \tag{1.1}$$

By studying such imbalances, we derive finite sample upper bounds on the excess variance $\mathcal{V}_n$, showing how rapidly different designs can approach V^* for rough, smooth, and restricted-complexity nonparametric outcome models $m(\psi)$. In several cases, we also obtain lower bounds establishing that these rates are minimax optimal.

More fundamentally, we prove an impossibility theorem showing that the curse of dimensionality illustrated above for matching actually affects every experimental design seeking to attain the classic variance bound V^*. Under mild conditions, our results imply that no design can guarantee uniform asymptotic efficiency, even for smooth outcome models, beyond the very-low dimensional regime $d = o(\log n)$. For experiments with thousands of units, this regime permits only a few covariates.

This suggests a new design principle. Rather than pursuing the classical efficiency bound V^*, we suggest balancing lower-complexity nonparametric classes $\mathcal{G}$ that may still approximate the true outcome model well. We develop experimental designs based on this principle using new nonparametric versions of the Gram-Schmidt walk (GSW) of Bansal et al. (2019) and Harshaw et al. (2024).

To measure their performance, we define the efficiency target $V^* + \Delta_{\mathcal{G}}$, where $\Delta_{\mathcal{G}}$ records outcome variation not captured by a restricted function class $\mathcal{G}$. The lower complexity of such classes allows our designs to approach this target at much faster uniform rates, allowing us to accommodate many more covariates, far beyond the $d = o(\log n)$ barrier above. For example, designs targeting an additive nonparametric specification can attain restricted efficiency with $d = o(n)$ covariates, up to log factors.

We show these designs can be combined with matching to preserve the fast rates above, while also providing some protection against unmodeled outcome variation. In an empirical application based on 12 randomized experiments recently published in economics, our designs reduce estimator variance relative to classical matched pairs across every specification and in every simulated empirical setting.

We make the following main contributions:

1. In Section 3, we study the finite-sample properties of matched pairs. We show that no matching scheme can attain an excess variance rate faster than $n^{-2/d}$, even over linear, hence infinitely smooth, outcome models. Despite this negative result, matched pairs is minimax optimal in fixed dimension for rough outcome models, attaining the rate $\mathcal{V}_n \asymp n^{-2\beta/d}$ over β-Hölder classes for $0 < \beta \leq 1$. Its slow convergence is therefore the price of robustness under very weak regularity.

2. Section 4 introduces new methods designed to exploit smoothness. To do so, we develop a nonparametric Gram-Schmidt Walk that controls worst-case imbalances over the unit ball of a chosen reproducing kernel Hilbert space. This design adapts without prior knowledge to every Sobolev smoothness order $s > 0$ and attains excess variance rate $\mathcal{V}_n = n^{-(2s/d)\wedge 1}$, up to logarithmic factors. This rate is minimax optimal when $2s \leq d$ and nearly parametric when $2s > d$.

3. Section 4 also establishes our main impossibility result: under weak conditions, no design can guarantee excess variance uniformly smaller than order $n^{-2s/d}$ over full-dimensional outcome models for any fixed smoothness $s > 0$. As above, this precludes uniform asymptotic efficiency for any experimental design outside the very-low dimensional regime $d = o(\log n)$.

4. In view of this impossibility result, Section 5 develops GSW designs using kernels tailored to lower-complexity nonparametric classes, including additive models of the form $g(\psi) = a + \sum_{j=1}^{d} g_j(\psi_j)$ and richer specifications with nonparametric bivariate interactions. The additive and bivariate designs approach their restricted efficiency targets while allowing $d = o(n)$ and $d = o(\sqrt{n})$, respectively, up to logarithmic factors.

5. Finally, Section 5 shows how to combine structured GSW with matching, retaining the faster rates for balancing low-complexity functions while also providing protection against unmodeled variation. We develop a conservative design-based variance estimator for such methods that performs well in our simulations and empirical application.

1.1 Related Literature

Our work is motivated by recent results showing that, under fine stratification, simple estimators automatically attain the Hahn (1998) variance bound for the ATE. Early contributions include Bai et al. (2022), Bai (2022), and Cytrynbaum (2026). Building on Armstrong (2022), Bai et al. (2026) further show that this classical efficiency bound remains valid for more general experimental designs in a GMM setting. Related work on stratified experiments includes Imai (2008), Fogarty (2018), Pashley and Miratrix (2021), and Kapelner et al. (2022), among others.

Our inefficiency result for matched pairs can be viewed as an experimental analogue of the classical result of Abadie and Imbens (2006), who show that slow nearest-neighbor convergence rates prevent matching from effectively debiasing observational estimates beyond a very low-dimensional setting. Wang and Li (2022) show a similar $d = o(\log n)$ threshold for balancing linear functions using rerandomization.

Our finite-sample perspective is related to that in Kallus (2018), who uses an analogue of Equation (1.1) to study imbalances over different function spaces. His analysis leads to semi-deterministic designs, some of which are NP-hard to compute. By contrast, we develop a full finite sample efficiency theory with explicit convergence rates to both classical and newly introduced variance bounds. We show that fast convergence rates can be achieved over nonparametric spaces by computationally tractable randomized designs. See the remarks below for a detailed comparison.

Our designs in Section 4 fundamentally rely on the Gram-Schmidt walk vector balancing algorithm of Bansal et al. (2019). Harshaw et al. (2024) first studied this algorithm in an experimental setting, proving a design-based oracle inequality for the mean squared error of an unadjusted estimator, using the algorithm to balance linear functions. Harshaw (2021) first observed that the design can be kernelized, extending the original oracle inequality to this setting. Chen and Zhang (2026) also suggest using GSW to balance nonlinear functions of the covariates.

To the best of our knowledge, we are the first to suggest finite sample minimax analysis of the excess variance $\mathcal{V}_n$ as an efficiency benchmark. The behavior of minimax rates over smoothness and complexity classes is a central organizing theme in nonparametric statistics (Stone, 1982; Yang and Barron, 1999; Giné and Nickl, 2016). Our use of Sobolev norms to quantify smoothness also follows a long tradition in this literature. For example, van der Vaart and van Zanten (2011) show that convergence rates for Gaussian process regression with Matérn priors are driven by a match between prior regularity and the Sobolev smoothness of the truth. Fischer and Steinwart (2020) derive learning rates for kernel ridge regression in Sobolev norms. See also Kennedy et al. (2024) for minimax rates for heterogeneous treatment effect estimation in observational studies.

The designs in Section 5 target balance over low-complexity function spaces. Our main examples involve additive and bivariate nonparametric models motivated by classical functional ANOVA (Stone, 1985; Hastie and Tibshirani, 1986; Sobol, 1993).

Other prominent approaches to covariate balance include rerandomization (Morgan and Rubin, 2012; Li et al., 2018) and optimization-based designs (Kasy, 2016; Krieger et al., 2019). A complementary literature studies regression adjustment with moderately high-dimensional covariates (Bloniarz et al., 2016; Wager et al., 2016; Lei and Ding, 2021; Lu et al., 2025). Our work can be viewed as a randomization-based analogue of such results.

Finally, our variance estimator combines the design-based quadratic bound approach of Harshaw et al. (2026) with the collapsed-strata and pairs-of-pairs methods of Hansen et al. (1953), Abadie and Imbens (2008), and Bai et al. (2022).

2 Setup

Consider an experiment with n units. Denote potential outcomes $Y_i(a)$ for $a \in \{0,1\}$ and $i \in [n]$, denoting $[r] = \{1, \ldots, r\}$ for any positive integer r.

Assumption 2.1. *Observations* $\big(\psi_i, Y_i(0), Y_i(1)\big) \overset{\text{iid}}{\sim} P$ *for* $i \in [n]$*. The covariate* ψ *is supported on* $[0,1]^d$ *and has a density* p *satisfying* $0 < \underline{p} \leq p(\psi) \leq \overline{p} < \infty$ *for every* $\psi \in [0,1]^d$*. For* $a \in \{0,1\}$*,* $E_P[Y(a)^2] < \infty$*.*

Unless stated otherwise, every data-generating law P in every model class $\mathcal{P}$ satisfies Assumption 2.1. We are interested in estimating the average treatment effect $\theta(P) = E_P[Y(1) - Y(0)]$. Write $\psi_{1:n} = (\psi_1, \ldots, \psi_n)$ for the realized covariates. For a realized treatment assignment $A_i \in \{0,1\}$, the observed outcome is $Y_i = Y_i(A_i)$. For convenience, in what follows, we work with $Z_i = 2A_i - 1 \in \{-1, 1\}$ and let $Z = (Z_1, \ldots, Z_n)$ be the full assignment vector.

Definition 2.2 (Admissible Designs)**.** A conditional assignment law $\sigma = Z|\psi_{1:n}$ is admissible $\sigma \in \mathcal{D}_n$ if $Z \perp\!\!\!\perp (Y_i(0), Y_i(1))_{i=1}^n | \psi_{1:n}$ and $E[Z_i|\psi_{1:n}] = 0$ for every $i \in [n]$.

For example, admissible designs include iid assignment, matched-pairs and other forms of stratified randomization, rerandomization with symmetric acceptance rules, as well as the Gram-Schmidt walk designs introduced in Sections 4 and 5. We use Horvitz-Thompson estimation

$$\widehat{\theta} = \frac{2}{n} \sum_{i=1}^{n} Z_i Y_i. \tag{2.1}$$

When exactly $n/2$ units are treated, as in matched pairs with even n, Equation (2.1) is the usual difference in means. Denote $\tau = Y(1) - Y(0)$ for the individual treatment effect and let $v_a^2(\psi) = \mathrm{Var}(Y(a)|\psi)$ for $a \in \{0,1\}$. Let $\tau(\psi) = E[\tau|\psi]$. At treatment probability one half, the semiparametric efficiency bound of Hahn (1998) for $\theta(P)$ is

$$V^*(P) = \mathrm{Var}\big(\tau(\psi)\big) + 2E\big[v_1^2(\psi)\big] + 2E\big[v_0^2(\psi)\big]. \tag{2.2}$$

This bound was originally derived for superpopulation settings with iid data. Building on Armstrong (2022), Bai et al. (2026) show it holds over a broad class of experimental designs with dependent treatment assignments.

In what follows, we study the gap between the finite sample variance $n\,\mathrm{Var}_{P,\sigma}(\widehat{\theta})$ for $\sigma \in \mathcal{D}_n$ and the asymptotic variance bound $V^*(P)$. To that end, define the *excess variance*

$$\mathcal{V}_n(\sigma, P) \equiv n\,\mathrm{Var}_{P,\sigma}(\widehat{\theta}) - V^*(P). \tag{2.3}$$

A sequence $\sigma_n \in \mathcal{D}_n$ is pointwise asymptotically efficient at P if the excess variance $\mathcal{V}_n(\sigma_n, P) \to 0$ as $n \to \infty$. Previous results have established pointwise asymptotic efficiency for general forms of fine stratification (Bai et al., 2022, 2026; Cytrynbaum, 2026). In what follows, we derive finite-sample lower bounds on the excess variance showing that such results can be a poor guide to finite-sample performance.

Let $\bar{Y} \equiv (Y(1)+Y(0))/2$ denote the outcome level. We call the predictable part of $\bar{Y}$ the *outcome model*, denoting $m(\psi) \equiv E[\bar{Y}|\psi]$. The next identity exactly identifies the excess variance with imbalances in the predictable part $m(\psi)$ of outcome levels.

Proposition 2.3 (Excess Variance)**.** *For every $\sigma \in \mathcal{D}_n$, the excess variance is*

$$\mathcal{V}_n(\sigma, P) = 4n \cdot E\left[\left(n^{-1}\sum_{i=1}^n Z_i m(\psi_i)\right)^2\right]. \tag{2.4}$$

This result recasts finite-sample efficiency as a function-balancing problem, with performance determined by the mean squared imbalance $n^{-1}\sum_{i=1}^n Z_i m(\psi_i)$. We also note that this equivalence was previously established in a slightly different form by Kallus (2018) for designs with exact treatment balance $\sum_i Z_i = 0$.

Since the distribution P is unknown at design time, a design cannot target $m(\psi)$ directly and must instead seek to control imbalance over a class of plausible models $P \in \mathcal{P}$. Protecting against too broad of a class may cause excess variance $\mathcal{V}_n(\sigma, P)$ to vanish too slowly for meaningful efficiency gains in finite samples. Next, we show how matched pairs exhibits an extreme case of this tradeoff.

3 The Limits of Matched Pairs Designs

Matched pairs protects against broad classes of weakly regular outcome models, but this robustness comes at a cost. In particular, no matched-pairs design can guarantee uniform asymptotic efficiency, even over linear outcome models, beyond the very-low dimensional regime $d = o(\log n)$.

3.1 A Lower Bound for Matching

Recall that a matched-pairs design first partitions the n units into $n/2$ unordered pairs (i_p, j_p), then randomly assigns one unit in each pair to treatment and the other to control with equal probability, independently across pairs. The pairing may be any measurable function of the realized covariates $\psi_{1:n}$ and exogenous randomness U_M. We write $\mathcal{M}_n$ for the resulting class of designs, with a given instance $M \in \mathcal{M}_n$. Matched pairs is clearly admissible, $\mathcal{M}_n \subseteq \mathcal{D}_n$. This results in $Z_{i_p} = -Z_{j_p}$ within each pair, so independence of the pair orientations and Proposition 2.3 imply

$$\mathcal{V}_n(M, P) = \frac{4}{n} E\left[\sum_{p=1}^{n/2} \big(m(\psi_{i_p}) - m(\psi_{j_p})\big)^2\right]. \tag{3.1}$$

Under Lipschitz continuity of $m(\psi)$, the within-pair differences $|m(\psi_{i_p}) - m(\psi_{j_p})|$ are bounded by a constant multiple of the corresponding covariate distances $\|\psi_{i_p} - \psi_{j_p}\|_2$. Pointwise asymptotic efficiency results therefore typically show $\mathcal{V}_n(M, P) \to 0$ for each fixed law P under a tight-matching condition of the form (Bai et al., 2022):

$$\frac{1}{n} \sum_{p=1}^{n/2} \|\psi_{i_p} - \psi_{j_p}\|_2^2 \xrightarrow{p} 0. \tag{3.2}$$

However, such asymptotic results do not characterize how finite-sample efficiency degrades with increasing covariate dimension because of the curse of dimensionality in matching. We next provide a finite-sample result showing that, even when excess variance vanishes asymptotically, it may converge too slowly to provide meaningful efficiency gains in finite samples. This limitation persists even over linear outcome models. To state the result, let $\mathcal{P}_{\text{lin}}$ contain the laws satisfying Assumption 2.1 whose outcome model has the form $m(\psi) = a + \gamma'\psi$, with $\|\gamma\|_2 \leq 1$.

Theorem 3.1 (Matching Slow Rate)**.** *For every even n and every $d \geq 1$,*

$$\inf_{M \in \mathcal{M}_n} \sup_{P \in \mathcal{P}_{\text{lin}}} \mathcal{V}_n(M, P) \geq \frac{1}{2\pi e} n^{-2/d}. \tag{3.3}$$

Theorem 3.1 applies to every pairing rule, not only optimal squared Euclidean distance matching as studied in Bai et al. (2022). Changing the matching criterion therefore cannot improve this slow excess variance rate.

We next use this finite-sample result to obtain alternative asymptotics for matched-pairs randomization in which $d = d_n$ may vary with n. This allows us to formalize the sense in which matching is asymptotically inefficient beyond the very low-dimensional regime $d_n \ll \log n$.

Corollary 3.2 (Asymptotic Inefficiency)**.** *If $d_n \geq c \log n$ for some fixed $c > 0$, then*

$$\liminf_{n \to \infty} \inf_{M \in \mathcal{M}_n} \sup_{P \in \mathcal{P}_{\text{lin}}} \mathcal{V}_n(M, P) \geq \frac{1}{2\pi e} e^{-2/c} > 0. \tag{3.4}$$

Equivalently, sample size n may be exponentially larger than covariate dimension d_n, yet the worst-case excess variance still remains bounded away from zero. The same curse of dimensionality drives the non-vanishing bias of matching estimators in observational studies (Abadie and Imbens, 2006) and slow convergence rates for nearest-neighbor methods in nonparametric regression (Györfi et al., 2002).

Figure 1 illustrates the finite-sample matching reversal behind Corollary 3.2 and the role of outcome-model complexity. Although recording more covariates steadily lowers the corresponding efficiency bound, matched-pairs variance eventually rises in both panels. Linear GSW avoids this reversal under a linear outcome model and rerandomization delays it, but both offer almost no gain under an adversarial nonlinear additive model. The nonparametric GSW design developed in Sections 4 and 5 below performs well in both settings.

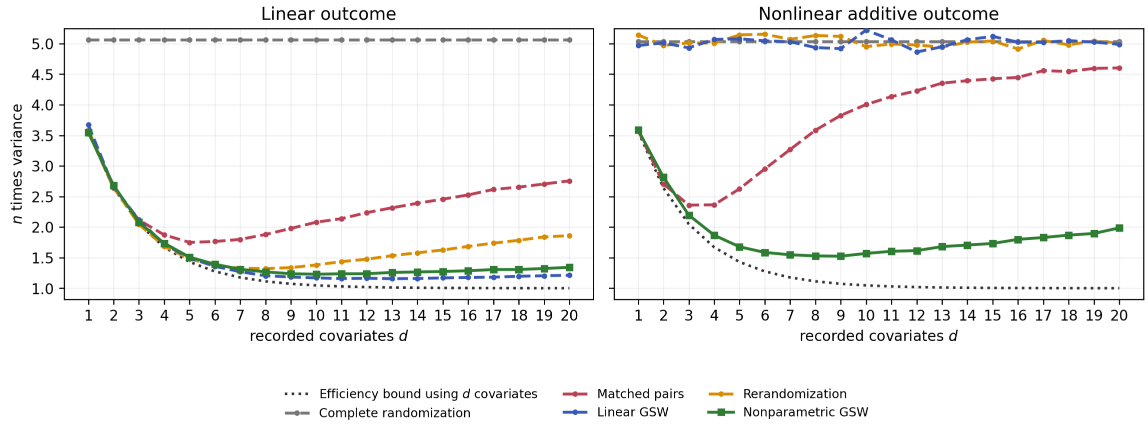

Figure 1: Variance for $n = 240$ units with independent uniform covariates. The left panel uses a normalized linear outcome model. The right panel replaces each linear coordinate with a nonlinear main effect while preserving the same coordinate-level variance profile. In both panels, the coordinate coefficients are proportional to 0.8^{j-1} and the dotted line is the efficiency bound using the first d covariates. Appendix A.4 gives the exact models and simulation protocol.

3.2 Matching Targets Rough Outcome Classes

For $0 < \beta \leq 1$, recall the Hölder coefficient $[m]_\beta = \sup_{x \neq y} |m(x) - m(y)|/\|x - y\|_2^\beta$. In this section, we show that slow excess variance rates of form $n^{-2\beta/d}$ are minimax optimal for fixed $d < \infty$ when the outcome model is β-Hölder continuous. We prove that a large family pairwise stable matchings achieve these rates adaptively for $\beta \in (0, 1]$, but fail to exploit additional smoothness beyond the Lipschitz endpoint $\beta = 1$.

Optimal Matching. We next describe a broad class of pairing rules that attain the excess variance rate $n^{-2\beta/d}$ for fixed d. Squared Euclidean matching, as in Bai et al. (2022), minimizes $\sum_{p=1}^{n/2} \|\psi_{i_p} - \psi_{j_p}\|_2^2$ over all pairings. More generally, an experimenter may specify any symmetric matching cost $c(x, y)$ and use Derigs (1988) algorithm to compute optimal pairs

$$\min_{(i_p, j_p)_{p=1}^{n/2}} \sum_{p=1}^{n/2} c(\psi_{i_p}, \psi_{j_p}). \tag{3.5}$$

Stable Matching. In fact, global optimality is stronger than necessary. A weaker notion is 2-swap stability, which requires that no two pairs can be rematched to reduce their combined cost. For any two matched pairs $\{i, j\}$ and $\{k, l\}$, we require that

$$c(\psi_i, \psi_j) + c(\psi_k, \psi_l) \leq \min \left\{ c(\psi_i, \psi_k) + c(\psi_j, \psi_l), c(\psi_i, \psi_l) + c(\psi_j, \psi_k) \right\}. \tag{3.6}$$

We show that every 2-swap-stable pairing attains the excess variance rate $n^{-2\beta/d}$ for fixed d whenever the matching cost satisfies the following weak geometric condition, which requires $c(x, y)$ to be comparable above and below to Euclidean distance.

Assumption 3.3 (Matching Cost)**.** *For some $q > 0$ and constants $0 < \underline{\lambda} \leq \overline{\lambda} < \infty$, the symmetric cost c satisfies $\underline{\lambda}\|x - y\|_2^q \leq c(x, y) \leq \overline{\lambda}\|x - y\|_2^q$ for every $x, y \in [0, 1]^d$.*

We define $\Lambda = (\overline{\lambda}/\underline{\lambda})^{1/q}$ as the condition number of the matching cost, which may depend on d. It measures the loss from translating control of the matching cost $c(x, y)$ into control of Euclidean distance. Thus $\Lambda = 1$ for any power of Euclidean distance, while $\Lambda = \sqrt{d}$ for $c(x, y) = \|x - y\|_1^q$ or $c(x, y) = \|x - y\|_\infty^q$ for any $q > 0$, for example.

Let $\mathcal{P}_\beta$ contain the laws satisfying Assumption 2.1 whose outcome model has Hölder coefficient $[m]_\beta \leq 1$. At $\beta = 1$, this is the ordinary Lipschitz class, while smaller β allow progressively rougher outcome functions.

Theorem 3.4 (Stable Matching)**.** *Fix an even $n \geq 2$ and $0 < \beta \leq 1$, and suppose $d \geq 3$. Impose Assumption 3.3 and let M_c be the matched-pairs design for any pairing that is 2-swap-stable for $c(x, y)$. There is $C > 0$ depending only on β such that*

$$\sup_{P \in \mathcal{P}_\beta} \mathcal{V}_n(M_c, P) \leq C(d\Lambda^2)^{2\beta} \cdot n^{-2\beta/d}. \tag{3.7}$$

This is a finite-sample bound, but we can also interpret it asymptotically. For example, in the low-dimensional regime with $d < \infty$ fixed while $n \to \infty$, any stable matching as above attains the excess variance rate $n^{-2\beta/d}$. This bound above holds simultaneously for every $\beta \in (0, 1]$, showing that matching adapts to the unknown Hölder smoothness of $m(\psi)$. We next establish a corresponding finite sample lower bound over all admissible designs, showing that no experimental design can improve this rate uniformly over the Hölder class in the low-dimensional regime.

Theorem 3.5 (Hölder Lower Bound)**.** *For every $0 < \beta \leq 1$, there is a constant $C > 0$ depending only on (β, d) such that, for every $n \geq 2$ and every $d \geq 1$,*

$$\inf_{\sigma \in \mathcal{D}_n} \sup_{P \in \mathcal{P}_\beta} \mathcal{V}_n(\sigma, P) \geq C n^{-2\beta/d}. \tag{3.8}$$

The lower bound in Theorem 3.5 is design agnostic. It applies not only to existing procedures, such as stratified randomization, symmetric rerandomization (Li et al., 2018), and optimization-based balancing (Kallus, 2018), but to every possible experimental design under mild admissibility requirements. Combining Theorems 3.4 and 3.5 immediately shows stable matched pairs is minimax rate optimal over rough Hölder classes in fixed low dimension $d < \infty$.

Corollary 3.6 (Minimaxity Over Rough Spaces)**.** *Let $0 < \beta \leq 1$ and fix dimension $3 \leq d < \infty$. Then as $n \to \infty$, stable matching achieves minimax rate*

$$\inf_{\sigma \in \mathcal{D}_n} \sup_{P \in \mathcal{P}_\beta} \mathcal{V}_n(\sigma, P) \asymp n^{-2\beta/d}. \tag{3.9}$$

Matching therefore provides minimax protection adaptively over the rough Hölder scale $\beta \in (0, 1]$. However, Theorem 3.1 above shows that no matched-pairs design can improve upon the $\beta = 1$ rate $n^{-2/d}$ even over the infinitely smooth linear class $\mathcal{P}_{\text{lin}}$. Thus matching cannot exploit smoothness beyond Lipschitz. The next section develops new designs that can adaptively exploit such higher order smoothness restrictions, while also retaining good performance for rough outcome models.

Remark 3.7 (Design Agnostic Bounds)**.** To the best of our knowledge, Corollary 3.6 provides the first design-agnostic minimax rate for covariate-adaptive randomization. To prove the lower bound, we place a prior on a hard finite subfamily of $\mathcal{P}_\beta$. We show that, for every treatment allocation Z, the squared imbalance $(n^{-1} \sum_{i=1}^n Z_i m(\psi_i))^2$, averaged over this family, is bounded below by unavoidable local contributions for a constant fraction of sampled units. Since this bound holds pointwise in Z, it remains valid after averaging over any admissible assignment distribution $\sigma \in \mathcal{D}_n$.

Remark 3.8. Kallus (2018) proves that optimal matched pairs randomization minimizes a related conditional imbalance criterion uniformly over Lipschitz outcome models. In his framework, the least-favorable outcome model m may depend on both realized covariates $\psi_{1:n}$ and the designer's candidate treatment allocation Z. By contrast, here we study the usual statistical minimax risk over a fixed superpopulation law. He does not study upper or lower bounds on the convergence rate of the resulting excess variance, as we do here.

4 Balancing Smooth Functions

In this section, we develop nonparametric discrepancy minimization designs that can exploit higher-order smoothness, building on the Gram-Schmidt walk of Bansal et al. (2019) and Harshaw et al. (2024). We show that kernelized GSW can achieve the excess variance rate $n^{-2s/d}$ for $s \leq d/2$ and n^{-1} otherwise, up to logarithmic factors. Unlike matched pairs designs, such global balancing schemes continue to improve under every additional order of Sobolev smoothness.

Despite these efficiency gains, we establish a hard dimensionality barrier, showing that no design can guarantee uniform convergence to the full efficiency bound, even over smooth outcome models, outside the very-low dimensional regime $d = o(\log n)$.

4.1 Sobolev Spaces and RKHS

Section 3 parameterized regularity using Hölder classes, which bound local changes in the outcome model $m(\psi)$ uniformly over the covariate space. To study outcome models with smoothness beyond Lipschitz order, here we instead work with Sobolev spaces, which provide a continuous scale of smoothness $s > 0$.

For a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d)$, write $|\alpha| = \sum_{j=1}^d \alpha_j$ and let $D^\alpha F$ denote the weak derivative of a function F on $\mathbb{R}^d$. Weak derivatives extend ordinary differentiation to functions whose rates of change exist only in an integrated sense, agreeing with ordinary derivatives whenever they exist. For integers $k \geq 1$, define the Sobolev space

$$H^k(\mathbb{R}^d) \equiv \left\{ F \in L^2(\mathbb{R}^d) : D^\alpha F \in L^2(\mathbb{R}^d) \text{ for every } |\alpha| \leq k \right\}. \tag{4.1}$$

At integer orders, this norm is equivalent to the integrated-derivative norm

$$\|F\|^2_{H^k(\mathbb{R}^d)} \asymp \sum_{|\alpha| \leq k} \int_{\mathbb{R}^d} |D^\alpha F(\psi)|^2 d\psi. \tag{4.2}$$

Sobolev spaces thus impose an integrated global budget on local fluctuations. Increasing the Sobolev order k requires more weak derivatives to exist and to be square integrable, producing a progressively more regular class. The definition extends to every real order $s > 0$ through the Fourier transform $\widehat{F}(\omega) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-i\omega' x} F(x) dx$.

The Sobolev space $H^s(\mathbb{R}^d)$ consists of functions $F \in L^2(\mathbb{R}^d)$ for which

$$\|F\|_{H^s(\mathbb{R}^d)}^2 \equiv \int_{\mathbb{R}^d} |\widehat{F}(\omega)|^2 (1 + \|\omega\|_2^2)^s d\omega < \infty. \tag{4.3}$$

At integer orders $s = k$, this yields the same space as the derivative-based definition above. Increasing the smoothness parameter s requires the high-frequency components of F to decay faster.

RKHSs. In what follows, we use reproducing kernel Hilbert space (RKHS) methods to construct a feasible balancing criterion that automatically adapts to the unknown Sobolev order of the outcome model $m(\psi)$. To connect RKHS methods to the Sobolev spaces introduced above, recall that W is a continuous positive semidefinite kernel if for every $r \geq 1$ and any $(x_i)_{i=1}^r \subseteq \mathbb{R}^d$, the Gram matrix $[W(x_i, x_j)]_{i,j=1}^r$ is positive semidefinite. Its RKHS $\mathcal{H}_W$ is a Hilbert space of functions with the reproducing property $h(\psi) = \langle h, W(\psi, \cdot)\rangle_{\mathcal{H}_W}$ for any $h \in \mathcal{H}_W$. By the Cauchy–Schwarz inequality, the reproducing property implies $|h(\psi)| \leq W(\psi, \psi)^{1/2}\|h\|_{\mathcal{H}_W}$, so point evaluation is a continuous linear functional on $\mathcal{H}_W$.

The Matérn family $W_d^{\mathrm{Mat},\nu}$ is indexed by the smoothness parameter $\nu > 0$. Its simplest member, obtained at $\nu = 1/2$, is the exponential kernel $W_d^{\mathrm{Mat},1/2}(\psi, \psi') = \exp(-\|\psi - \psi'\|_2)$. A formula for general ν is given in Appendix B.3. Write $\mathcal{H}_d^{\mathrm{Mat},\nu} \equiv \mathcal{H}_{W_d^{\mathrm{Mat},\nu}}$ for the associated RKHS. The following standard correspondence links these RKHSs to the Sobolev spaces defined above.

Proposition 4.1 (Matérn-Sobolev Correspondence)**.** *For every $d \geq 1$ and $\nu > 0$,*

$$\mathcal{H}_d^{\mathrm{Mat},\nu} = H^{d/2+\nu}(\mathbb{R}^d), \qquad \|f\|_{\mathcal{H}_d^{\mathrm{Mat},\nu}} \asymp_{d,\nu} \|f\|_{H^{d/2+\nu}(\mathbb{R}^d)}. \tag{4.4}$$

In particular, every Sobolev space $H^s(\mathbb{R}^d)$ with $s > d/2$ is a Matérn RKHS, obtained by taking $\nu = s - d/2$. This correspondence is standard (Wendland, 2005; Kanagawa et al., 2018). In Appendix B.3 we rederive it and record a uniform bound on the norm-equivalence constant as ν approaches zero, required for our later results.

4.2 Controlling Imbalances for Smooth Functions

Proposition 2.3 shows that excess variance is determined by the expected square of the outcome-model imbalance $n^{-1}\sum_{i=1}^n Z_i m(\psi_i)$. Since $m(\psi)$ is unknown at design time, the design cannot balance it directly. We can instead form a feasible criterion by controlling the worst-case imbalance over the unit ball of a chosen RKHS $\mathcal{H}_W$. Consider an allocation $Z \in \{-1, 1\}^n$ and let $W_n = [W(\psi_i, \psi_j)]_{i,j=1}^n$ be the corresponding kernel matrix. By the reproducing property,

$$\mathcal{I}_W(Z) \equiv \sup_{\|g\|_{\mathcal{H}_W} \leq 1} \left(\frac{1}{n} \sum_{i=1}^n Z_i g(\psi_i) \right)^2 = \frac{1}{n^2} Z' W_n Z. \tag{4.5}$$

Note that for $W = W_d^{\mathrm{Mat},\nu}$, Proposition 4.1 identifies this criterion with the worst-case squared imbalance over a ball in $H^{d/2+\nu}(\mathbb{R}^d)$. We therefore seek an admissible assignment law $\sigma \in \mathcal{D}_n$ under which the quadratic form $\mathcal{I}_W(Z)$ is small.

We use the Gram-Schmidt walk (GSW) algorithm of Bansal et al. (2019), as adapted to experimental design by Harshaw et al. (2024), to construct a computationally tractable assignment law that keeps $\mathcal{I}_W(Z)$ small while retaining the randomization needed for robustness and inference.

GSW Algorithm. The Gram-Schmidt walk algorithm of Bansal et al. (2019) takes as input an $n \times n$ positive semidefinite matrix Γ_n with diagonal entries at most one. It begins with the infeasible fractional allocation $z^{(0)} = 0 \in [-1, 1]^n$. At each step $t \geq 1$, it selects a fractional pivot p_t, which it retains until it reaches -1 or 1, and chooses an update direction u minimizing $u'\Gamma_n u$, subject to $u_{p_t} = 1$ and $u_j = 0$ for coordinates that are already integral. It then randomizes between the largest feasible positive and negative steps in direction u, with probabilities chosen so the update has conditional mean zero. At least one fractional coordinate reaches the boundary after each update, so the procedure terminates in finitely many steps in a feasible treatment allocation $Z \in \{\pm 1\}^n$. See Appendix A.5 for a more detailed discussion.

Kernelization. Harshaw (2021) originally observed that GSW can be kernelized by constructing its input matrix Γ_n from a kernel matrix. We use the *design kernel* $K = 1+W$, where W is a continuous positive semidefinite kernel satisfying $\sup_\psi W(\psi, \psi) \leq 1$. The added constant ensures that the associated RKHS can represent arbitrary constant levels of the outcome model m. By the sums-of-kernels theorem (Paulsen and Raghupathi, 2016), $\mathcal{H}_K$ consists of functions $g = a+w$, where $a \in \mathbb{R}$ and $w \in \mathcal{H}_W$, with norm $\|g\|_{\mathcal{H}_K}^2 = \min_{g=a+w}(a^2 + \|w\|_{\mathcal{H}_W}^2)$. Given covariates $\psi_{1:n}$ and a robustness parameter $\varphi \in (0,1)$, define the positive definite matrix

$$\Gamma_n = \varphi I_n + \frac{1-\varphi}{2}\big[K(\psi_i,\psi_j)\big]_{i,j=1}^n. \tag{4.6}$$

Because $K(\psi,\psi) \le 2$, the matrix Γ_n is positive definite with diagonal entries at most one and therefore satisfies the GSW input conditions above. The second term in Equation (4.6) rewards balance over $\mathcal{H}_K$, while the identity term supplies robustness in outcome directions that are not represented by this RKHS.

Write $\mathrm{GSW}(K,\varphi)$ for the resulting assignment law, suppressing its dependence on n. The mean-zero updates imply $E[Z_i|\psi_{1:n}] = 0$ for every unit (Harshaw et al., 2024). The update directions and step sizes depend only on the covariates and auxiliary randomness independent of the potential outcomes, so $Z \perp\!\!\!\perp (Y_i(0), Y_i(1))_{i=1}^n | \psi_{1:n}$. Thus, $\mathrm{GSW}(K,\varphi)$ is admissible.

Remark 4.2 (Kernel Allocation)**.** The RKHS imbalance objective in Equation (4.5) also underlies the kernel allocation design of Kallus (2018), which randomizes between z^* and $-z^*$ for the optimal vector $z^* \in \arg\min z'W_n z$ with $z \in \{-1,1\}^n$. Computing z^* requires solving an NP-hard mixed integer program, and the resulting design has support of size two. This limited randomization can have poor robustness properties and preclude standard inference (Kallus, 2021). Kernelized GSW avoids these issues through computationally tractable updates and broad randomization support.

4.3 Oracle Inequality and Smoothness Gains

Harshaw et al. (2024) show that the mean squared error of the Horvitz-Thompson estimator under GSW is bounded above by the loss of an implicit ridge regression of the outcome levels on the covariates. Harshaw (2021) extends this inequality to kernelized GSW, yielding an implicit kernel-ridge objective. Extending these results to our current superpopulation setting and combining with Proposition 2.3 yields the following oracle inequality for excess variance.

Proposition 4.3 (Oracle Inequality)**.** *Let $\sigma = \mathrm{GSW}(K,\varphi)$ as above. Then*

$$\mathcal{V}_n(\sigma,P) \le \inf_{g\in\mathcal{H}_K}\left\{\frac{4}{\varphi}E\left[(m(\psi)-g(\psi))^2\right] + \frac{8}{(1-\varphi)n}\|g\|_{\mathcal{H}_K}^2\right\}. \tag{4.7}$$

The kernelized GSW criterion is designed to control worst-case imbalance over the design RKHS $\mathcal{H}_K$. However, Proposition 4.3 shows that the outcome model need not itself belong to $\mathcal{H}_K$ for the design to yield efficiency gains. For any $g \in \mathcal{H}_K$, the first term controls the residual $m - g$ not targeted by the kernel, while the second reflects the difficulty of balancing g, as measured by its RKHS norm. Thus, the guarantee is strongest when m can be approximated accurately by functions of moderate norm, which we show below occurs under sufficient smoothness restrictions.

Recall that a sequence of designs σ_n is pointwise asymptotically efficient at P if $\mathcal{V}_n(\sigma_n, P) \to 0$. Previously, this property has been established only for matched pairs and other finely stratified designs (Bai et al., 2022, 2026). Equation (4.7) can be used to extend this result to any kernelized GSW design whose RKHS is dense in $L^2(P_\psi)$.

Corollary 4.4 (Pointwise Efficiency)**.** *Fix a design kernel K and $\varphi \in (0,1)$, and, for each n, let $\sigma_n = \mathrm{GSW}(K, \varphi)$. Suppose $\mathcal{H}_K$ is dense in $L^2(P_\psi)$. As $n \to \infty$, excess variance*

$$\mathcal{V}_n(\sigma_n, P) \longrightarrow 0.$$

For every $\nu > 0$, the Matérn design RKHS is dense in $L^2(P_\psi)$ under every Borel law on $[0,1]^d$. The Gaussian radial basis function kernel with length scale $\ell > 0$ and $W_{\mathrm{RBF},\ell}(\psi, \psi') = \exp\{-\|\psi - \psi'\|_2^2/(2\ell^2)\}$ has the same density property (Schölkopf and Smola, 2002). Thus, Corollary 4.4 applies to either kernel family.

However, as we argued in Section 3, such pointwise results can fail to yield meaningful guarantees on finite-sample efficiency. This suggests instead using the oracle inequality to produce uniform rates for excess variance over suitable regularity classes. The simplest example of such a bound follows easily when the outcome model belongs to a bounded ball in the design RKHS. For a design kernel K and $B > 0$, define this class by $\mathcal{P}_K(B) = \{P : m \in \mathcal{H}_K,\ \|m\|_{\mathcal{H}_K} \leq B\}$.

Corollary 4.5 (RKHS Fast Rate)**.** *Let $\sigma = \mathrm{GSW}(K, \varphi)$ as above. Then*

$$\sup_{P \in \mathcal{P}_K(B)} \mathcal{V}_n(\sigma, P) \leq \frac{8B^2}{(1-\varphi)n}. \tag{4.8}$$

This shows kernelized GSW attains the fast parametric rate n^{-1} uniformly over any fixed RKHS ball. Proposition 4.1 and the nesting of Sobolev spaces imply that, whenever $d/2 + \nu \leq s$, every bounded ball in $H^s(\mathbb{R}^d)$ is contained in a bounded ball of the RKHS induced by $K = 1 + W_d^{\mathrm{Mat},\nu}$. Corollary 4.5 therefore gives an immediate uniform excess-variance bound in the supercritical regime $s > d/2$ with sufficient smoothness. For comparisons across dimensions, let $H^s_{\mathrm{av}}(\mathbb{R}^d)$ denote $H^s(\mathbb{R}^d)$ equipped with the dimension-stable norm

$$\|F\|^2_{H^s_{\mathrm{av}}(\mathbb{R}^d)} \equiv \int_{\mathbb{R}^d} |\widehat{F}(\omega)|^2 \left(1 + \frac{\|\omega\|_2^2}{d}\right)^s d\omega. \tag{4.9}$$

For fixed d, this is equivalent to the ordinary Sobolev norm and hence defines the same function space. To understand this normalization, suppose that ψ is uniform on $[0,1]^d$. Lemma B.8 shows that the largest variance of $m(\psi)$ over $\|m\|_{H^s_{\mathrm{av}}(\mathbb{R}^d)} \le 1$ remains of constant order as d varies. Define the probability model

$$\mathcal{P}^s_d \equiv \left\{P : m \in H^s_{\mathrm{av}}(\mathbb{R}^d),\ \|m\|_{H^s_{\mathrm{av}}(\mathbb{R}^d)} \le 1\right\}. \tag{4.10}$$

Note that we let $B = 1$ without loss. A fixed radius B scales the bound by B^2.

Corollary 4.6 (Non-adaptive Fast Rate)**.** *Let $s > d/2$ and $\varphi \in (0,1)$. Choose $\nu > 0$ such that $d/2 + \nu \le s$, and let $K = 1 + W_d^{\mathrm{Mat},\nu}$. For design $\sigma = \mathrm{GSW}(K, \varphi)$, there is a constant C depending only on (d, ν, φ) such that*

$$\sup_{P \in \mathcal{P}^s_d} \mathcal{V}_n(\sigma, P) \le C n^{-1}. \tag{4.11}$$

A fixed Matérn kernelized GSW design therefore attains the parametric excess-variance rate n^{-1} uniformly over Sobolev outcome models with smoothness $s > d/2$. This result, however, requires both $s > d/2$ and enough prior knowledge of s to choose the Matérn tuning parameter ν. For fixed s, the condition $s > d/2$ eventually fails along any sequence $d = d_n \to \infty$, so this simple result is unsuitable for studying asymptotics beyond low, fixed dimension.

The next subsection removes both restrictions by constructing a single sequence of kernelized GSW designs that adapts to every $s > 0$ without prior knowledge of the smoothness order.

4.4 Adaptation to Unknown Smoothness

For any fixed $c > 0$, set $\nu_n = c/\log n$, so the Sobolev order $d/2 + \nu_n$ of the Matérn design RKHS approaches the critical boundary $d/2$ from above, independent of the unknown smoothness s. Next we show that this calibration yields a design sequence that adapts to every $s > 0$.

Theorem 4.7 (Adaptive Kernel GSW)**.** *Fix $c > 0$ and $\varphi \in (0,1)$, and set $\nu_n = c/\log n$, $K_n = 1 + W_d^{\mathrm{Mat},\nu_n}$, and $\sigma_n = \mathrm{GSW}(K_n, \varphi)$. For every $s > 0$, there is a constant C depending only on $(s, c, \varphi, \overline{p})$ such that, for every $d \geq 1$ and $n \geq 2$,*

$$\sup_{P \in \mathcal{P}_d^s} \mathcal{V}_n(\sigma_n, P) \leq C \left(\frac{\log n}{n} \right)^{(2s/d) \wedge 1}. \tag{4.12}$$

For fixed d, the exponent $2s/d$ increases continuously with smoothness until the rate reaches $\log n/n$ at $s = d/2$. When $s > d/2$, this is only a logarithmic factor slower than the n^{-1} rate in Corollary 4.6. Thus, a single design sequence adapts to every $s > 0$ without prior knowledge of s, paying only a logarithmic factor for adaptivity.

Remark 4.8 (Comparison to Matched Pairs)**.** In Appendix A.2, we show that, for fixed $d \geq 3$, the same design attains the minimax rate $n^{-2\beta/d}$, up to logarithmic factors, uniformly over β-Hölder outcome models for every $0 < \beta \leq 1$. Unlike matching, however, its rate continues to improve under Sobolev smoothness above order one and becomes nearly parametric once $s \geq d/2$.

Figure 2 illustrates the finite-sample efficiency gains from smoothness. The improvement of Matérn GSW over matched pairs grows as the outcome model becomes smoother. In particular, a design using the fixed parameter $\nu_0 = 1$ captures much of the gain achieved by Oracle Matérn GSW, which uses the true smoothness parameter, and its advantage increases with the sample size.

For comparison, the figure also includes a kernelized version of best-of-m rerandomization (Wang and Li, 2025). It selects, out of $m = 10{,}000$ independent allocations Z, the one minimizing the worst-case imbalance $\mathcal{I}_W(Z)$ in Equation (4.5), for the Matérn kernel W with the oracle value of ν. In contrast to its strong performance for the linear outcome model in Figure 1, here rerandomization offers essentially no improvement over matched pairs, reflecting the difficulty of balancing a nonparametric function space through random search.

4.5 The Dimensionality Barrier

The adaptive rate in Theorem 4.7 still has the basic form $n^{-2s/d}$, which becomes vacuous when s is fixed and $d = d_n \geq a \log n$ for some $a > 0$. Similar to the Hölder lower bound in Theorem 3.5, the next theorem shows that this is not a limitation of kernelized GSW: every admissible design faces the same dimensionality barrier.

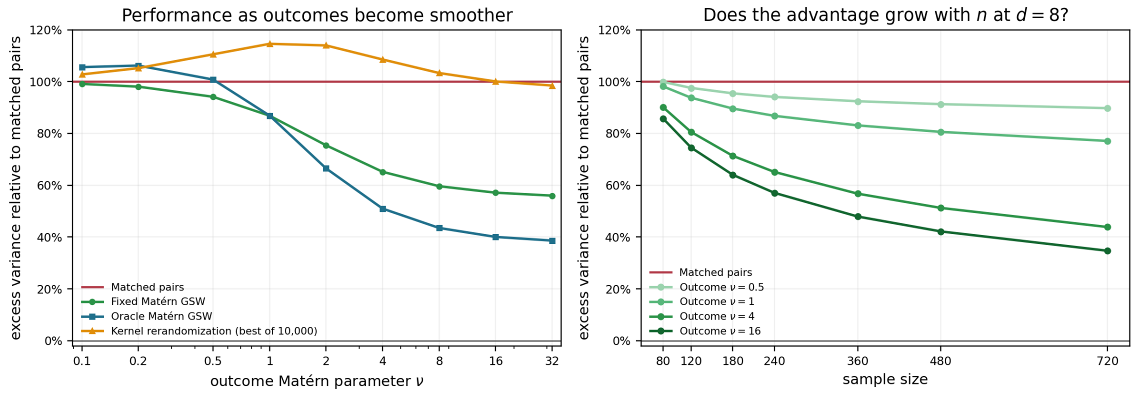

Figure 2: Prior-averaged excess variance for outcome models $m(\psi)$ drawn from a Matérn Gaussian process with parameter ν. Covariates are uniform on $[0, 1]^d$. The left panel fixes $n = 240$ and $d = 8$. Fixed Matérn GSW uses $\nu_0 = 1$ independently of the true ν, while Oracle Matérn GSW uses $\nu_0 = \nu$. Kernel rerandomization selects, among 10,000 draws, the allocation minimizing $Z'W_nZ$, where W_n is the Matérn kernel matrix constructed using the true ν. The right panel fixes $d = 8$ and varies n for $\nu \in \{0.5, 1, 4, 16\}$. Appendix A.4 gives the exact model and simulation protocol.

Theorem 4.9 (Design-Agnostic Lower Bound)**.** *Fix $s > 0$. There is a constant $c_0 > 0$ depending only on s such that, for every $d \geq 1$ and every $n \geq 2$,*

$$\inf_{\sigma \in \mathcal{D}_n} \sup_{P \in \mathcal{P}_d^s} \mathcal{V}_n(\sigma, P) \geq c_0 n^{-2s/d}. \tag{4.13}$$

The central implication is the dimensionality barrier. For fixed s and any sequence $d_n \geq a \log n$, Theorem 4.9 gives the design-agnostic lower bound

$$\liminf_{n \to \infty} \inf_{\sigma \in \mathcal{D}_n} \sup_{P \in \mathcal{P}_{d_n}^s} \mathcal{V}_n(\sigma, P) \geq c_0 e^{-2s/a} > 0. \tag{4.14}$$

Hence, no admissible design can guarantee convergence to the full efficiency bound beyond the low-dimensional regime $d_n = o(\log n)$. Motivated by this impossibility, the next section relaxes the pursuit of full efficiency by balancing lower-complexity nonparametric working models. When such a working model closely approximates the true outcome model, its variance benchmark remains close to the Hahn bound but can be approached much more rapidly in finite samples.

Remark 4.10 (Fixed-Dimensional Minimaxity)**.** For fixed d and $s \leq d/2$, combining Theorems 4.7 and 4.9 gives the following finite-sample minimax comparison. There are constants $0 < c_1 < C_1 < \infty$, depending on $(d, s, c, \varphi, \overline{p})$, such that, for every $n \geq 2$,

$$c_1 n^{-2s/d} \leq \inf_{\sigma \in \mathcal{D}_n} \sup_{P \in \mathcal{P}_d^s} \mathcal{V}_n(\sigma, P) \leq C_1 \left(\frac{n}{\log n} \right)^{-2s/d}. \tag{4.15}$$

In low-dimensional asymptotics with $d < \infty$ fixed while $n \to \infty$, kernelized GSW is thus minimax rate optimal for every $s \leq d/2$, up to logarithmic factors.

5 Efficient Designs Beyond Low Dimensions

In this section, we construct kernelized GSW designs that balance lower-complexity nonparametric working models chosen to approximate the outcome model $m(\psi)$, rather than pursuing uniform convergence to the full efficiency bound. By imposing additive or low-order interaction structure, these designs can accommodate many more covariates than matching or full-dimensional kernelized GSW, while still providing substantial nonparametric efficiency gains.

5.1 Structured Nonparametric Balance

Kernelized GSW controls the worst-case imbalance $n^{-1} \sum_{i=1}^n Z_i g(\psi_i)$ over the unit ball of the RKHS supplied to the algorithm. Section 4.2 used a full-dimensional Matérn space. Here we instead use structured sum RKHSs, beginning with two examples.

Example 5.1 (Nonparametric Main Effects)**.** We can control each covariate nonparametrically by targeting additive functions of the form $g(\psi) = a + \sum_{j=1}^d g_j(\psi_j)$.

Example 5.2 (Bivariate Nonparametric Effects)**.** Alternatively, we can control every bivariate interaction nonparametrically by balancing functions

$$g(\psi) = a + \sum_{j=1}^{d} g_j(\psi_j) + \sum_{j<\ell} g_{j\ell}(\psi_j, \psi_\ell). \tag{5.1}$$

Both classes can be implemented directly with kernelized GSW through kernel addition. Let $W_1, \ldots, W_B$ be continuous positive semidefinite kernels satisfying $W_b(\psi, \psi) \leq 1$ and set $K(\psi, \psi') = 1 + B^{-1} \sum_{b=1}^B W_b(\psi, \psi')$. Then the RKHS $\mathcal{H}_K$ consists of functions of the form $g = a + \sum_{b=1}^B f_b$ with $f_b \in \mathcal{H}_{W_b}$. For instance, one kernel corresponding to Example 5.1 is $K_1(\psi, \psi') = 1 + d^{-1} \sum_{j=1}^d W_1^{\mathrm{Mat},\nu}(\psi_j, \psi_j')$.

Variance Gap. The sum RKHS provides a general formula for lower-complexity nonparametric approximations to m. Let $\mathcal{G}$ denote the $L^2(P_\psi)$ closure of $\mathcal{H}_K$ and let $m_{\mathcal{G}}$ be the $L^2(P_\psi)$ projection of m onto $\mathcal{G}$. We view this as a working model for m, with variance gap

$$\Delta_{\mathcal{G}}(P) \equiv 4E\big((m(\psi) - m_{\mathcal{G}}(\psi))^2\big). \tag{5.2}$$

The corresponding variance target is $V^*(P) + \Delta_{\mathcal{G}}(P)$. When $\mathcal{G} = L^2(P_\psi)$, we have $m_{\mathcal{G}} = m$ and $\Delta_{\mathcal{G}}(P) = 0$, recovering the full efficiency bound. More generally, when $\mathcal{G}$ closely approximates m, this target remains close to $V^*(P)$.

Regularity Conditions. As in Section 4, obtaining uniform rates requires regularity of the structured approximation $m_{\mathcal{G}}$. The component representations in Examples 5.1 and 5.2 are not unique. For example, main effects can be absorbed into the bivariate components. Because of this, we impose regularity conditions on the canonical functional ANOVA decomposition (Sobol, 1993). In particular, expand

$$m_{\mathcal{G}}(\psi) = a + \sum_{j=1}^{d} m_j(\psi_j) + \sum_{j<\ell} m_{j\ell}(\psi_j, \psi_\ell). \tag{5.3}$$

We require $\int_0^1 m_j(u)du = 0$ as well as $\int_0^1 m_{j\ell}(u,v)du = 0$ for every v and vice-versa. These centering conditions make the corresponding component subspaces mutually orthogonal in $L^2([0,1]^d)$, so the decomposition is unique. For smoothness $s > 0$, we require the canonical components in Equation (5.3) to satisfy the pooled Sobolev budget

$$\sum_{j=1}^{d} \|m_j\|^2_{H^s_{\text{av}}(\mathbb{R})} + \sum_{j<\ell} \|m_{j\ell}\|^2_{H^s_{\text{av}}(\mathbb{R}^2)} \le 1. \tag{5.4}$$

The pooled bound prevents the aggregate magnitude and roughness of the structured projection from growing mechanically with the number of recorded covariates. For the main-effects working model, the bivariate terms are omitted.

We state the main result for the richer bivariate design in Example 5.2, leaving the main-effects case to Theorem B.10 in the appendix. For this result, take $\mathcal{G}_2$ to be the $L^2(P_\psi)$ closure of the class in Equation (5.1) and let $\mathcal{P}^s_{d,2}$ contain the laws satisfying Assumption 2.1, $E[m(\psi)^2] \le 1$, and Equation (5.4). Fix $c > 0$ and set $\nu_n = c/\log n$, so each component RKHS lies $c/\log n$ above its critical Sobolev order.

Writing $\psi_{j\ell} = (\psi_j, \psi_\ell)$ and similarly for ψ', define

$$K_n(\psi, \psi') = 1 + \binom{d}{2}^{-1} \sum_{j<\ell} W_2^{\mathrm{Mat},\nu_n}(\psi_{j\ell}, \psi'_{j\ell}).$$

For every n, the $L^2(P_\psi)$ closure of $\mathcal{H}_{K_n}$ is $\mathcal{G}_2$.

Theorem 5.3 (Bivariate Effects)**.** *Fix $n \geq 2$, $d \geq 2$, and smoothness $s > 0$. Let $\sigma_n = \mathrm{GSW}(K_n, \varphi_n)$, where $1 - \varphi_n = 1/2 \wedge (n^{-1} d^2 \log n)^{1/2}$. Then, for a constant $C > 0$ independent of n and d, uniformly over $P \in \mathcal{P}^s_{d,2}$,*

$$\mathcal{V}_n(\sigma_n, P) \leq \Delta_{\mathcal{G}_2}(P) + C \left(\frac{d^2 \log n}{n} \right)^{(s/2) \wedge (1/2)}. \tag{5.5}$$

Theorem B.10 in the appendix extends this result to allow a fixed block of priority covariates, such as baseline outcomes, to enter jointly and have a different smoothness order. The extension adds the corresponding fixed-dimensional approximation rate while retaining the bivariate rate above for the remaining covariates.

When the working model is accurate, $\Delta_{\mathcal{G}_2}(P)$ is small, so the design approaches a variance target close to the Hahn bound $V^*(P)$ much more rapidly than the $n^{-2s/d}$ rate obtained by the full-dimensional Matérn designs in Section 4, up to logarithmic factors. The lower-complexity specification also accommodates many more covariates than those designs can. Whereas all of the preceding designs, including matching and full-dimensional GSW, require the very low-dimensional regime $d_n = o(\log n)$, Equation (5.5) allows the number of covariates to grow nearly as fast as $\sqrt{n}$:

$$\mathcal{V}_n(\sigma_n, P) \leq \Delta_{\mathcal{G}_2}(P) + o(1) \quad \text{if} \quad d_n = o(\sqrt{n/\log n}).$$

Note also that, as in Section 4, the design does not require prior knowledge of smoothness since the near-critical kernel sequence adapts automatically. The calibration above uses $\varphi_n \to 1$, but the theorem does not suggest that this is necessary. In simulations, small fixed values of φ continue to perform well, and we conjecture that this technical requirement is an artifact of the current analysis.

Well Specification. In Theorem B.10 in the appendix, we provide a sharper rate under correct specification of the working model $m_{\mathcal{G}}$. In particular, when $m = m_{\mathcal{G}_2}$, any fixed $\varphi \in (0, 1)$ gives a design $\sigma_n = \mathrm{GSW}(K_n, \varphi)$ satisfying $\mathcal{V}_n(\sigma_n, P) \lesssim$

$(d^2 \log n/n)^{s\wedge 1}$.

Main effects. The additive nonparametric design in Example 5.1 permits still faster dimension growth. Let $\mathcal{G}_1$ denote the $L^2(P_\psi)$ closure of the main-effects models in Example 5.1. Writing $\sigma_{n,1}$ for the corresponding main-effects GSW design, Theorem B.10 shows that $\mathcal{V}_n(\sigma_{n,1}, P) \le \Delta_{\mathcal{G}_1}(P) + o(1)$ whenever $d_n = o(n/\log n)$. Thus, replacing bivariate interactions by additive main effects increases the allowable dimension from $o(\sqrt{n/\log n})$ to $o(n/\log n)$, at the cost of the potentially larger variance gap $\Delta_{\mathcal{G}_1}(P)$.

5.2 Combining Matching with Structured Balance

Section 5.1 showed that balancing a lower-complexity working model can rapidly reduce excess variance far beyond the low-dimensional regime, at the cost of a variance gap from outcome variation the working model does not capture. We now combine these designs with matching, allowing kernelized GSW to balance the low complexity working model globally, while within-pair differences attenuate the unmodeled component locally. This hybrid is our preferred design, and Section 6 shows it has very strong performance relative to classic matched pairs.

Matched GSW. Let $M = \{(i_p, j_p) : p \in [n/2]\}$ be a matching as in Section 3. Any assignment treating exactly one unit in each pair can be written as $Z_{i_p} = T_p$ and $Z_{j_p} = -T_p$ for a *pair orientation* $T_p \in \{-1, 1\}$. In classic matched-pairs, the T_p are iid random signs. Here, we propose to use kernelized GSW to generate them.

The imbalance can be written $n^{-1}\sum_{i=1}^n Z_i g(\psi_i) = n^{-1}\sum_{p=1}^{n/2} T_p(g(\psi_{i_p}) - g(\psi_{j_p}))$. Using this formula and a reproducing-property calculation as in Section 4.2, one can show that the squared worst-case imbalance over the RKHS unit ball is $n^{-2}T'G_n^M T$, where $G_n^M \in \mathbb{R}^{(n/2)\times(n/2)}$ is the pair-difference matrix

$$(G_n^M)_{pq} = K(\psi_{i_p}, \psi_{i_q}) - K(\psi_{i_p}, \psi_{j_q}) - K(\psi_{j_p}, \psi_{i_q}) + K(\psi_{j_p}, \psi_{j_q}). \tag{5.6}$$

As before, we seek to randomize in a way that makes $T'G_n^M T$ small. This suggests normalizing G_n^M and using it as the input to the Gram-Schmidt walk to generate the pair orientations T. Let $\widehat{\kappa}_n = \max_p (G_n^M)_{pp} = \max_p \|K(\psi_{i_p}, \cdot) - K(\psi_{j_p}, \cdot)\|^2_{\mathcal{H}_K}$ be the largest squared RKHS distance within a matched pair and apply GSW to

$$\Gamma_n^M = \varphi I_{n/2} + \frac{1-\varphi}{\widehat{\kappa}_n} G_n^M. \tag{5.7}$$

If $\widehat{\kappa}_n = 0$, set $\Gamma_n^M = \varphi I_{n/2}$. The matrix G_n^M is positive semidefinite, and the diagonal entries of Γ_n^M are at most one, as required by the algorithm. Write $\mathrm{GSW}_M(K, \varphi)$ for the resulting admissible assignment law, which treats exactly half the units.

Let $\kappa_n = E[\widehat{\kappa}_n]$ be the expected largest squared RKHS distance within a matched pair. Impose the Section 4.2 normalization $K = 1 + W$, with $\sup_\psi W(\psi, \psi) \leq 1$, so that $0 \leq \kappa_n \leq 4$. For a square-integrable function $r(\psi)$, define its expected pair energy by

$$Q_n^M(r) = n^{-1} E\Bigg[\sum_{(i,j)\in M} \big(r(\psi_i) - r(\psi_j)\big)^2\Bigg]. \tag{5.8}$$

The following paired analogue of Proposition 4.3 formalizes the local-global division of labor: matching controls within-pair variation locally, while kernelized GSW controls structured imbalance globally.

Theorem 5.4 (Matched Oracle Inequality). *Fix $\varphi \in (0,1)$ and let $\sigma = \mathrm{GSW}_M(K, \varphi)$.*

$$\mathcal{V}_n(\sigma, P) \leq \inf_{f \in \mathcal{H}_K} \left\{ \frac{4}{\varphi} Q_n^M(m - f) + \frac{4\kappa_n}{n(1-\varphi)} \|f\|_{\mathcal{H}_K}^2 \right\}. \tag{5.9}$$

Relative to the unit-level oracle in Equation (4.7), matching replaces the global approximation error $E[(m(\psi) - f(\psi))^2]$ by the within-pair energy $Q_n^M(m - f)$ and replaces the fixed RKHS penalty coefficient 8 by $4\kappa_n$. This can significantly attenuate both terms when the residual varies little within pairs and matched units are close in the kernel geometry.

Example 5.5 (Smooth Plus Rough Decomposition). Suppose we expand $m = a+g+r$ for some $g \in \mathcal{H}_K$ with $\|g\|_{\mathcal{H}_K} \leq B$ and a residual r. Because pair energy is invariant to constants, taking $f = g$ in Theorem 5.4 gives

$$\mathcal{V}_n(\sigma, P) \leq \frac{4}{\varphi} Q_n^M(r) + \frac{4\kappa_n B^2}{n(1-\varphi)}. \tag{5.10}$$

The first term asks matching to control the residual r, while the second asks GSW to balance the structured component g. One can show $\kappa_n = o(1)$ under suitable regularity conditions, providing a further rate enhancement.

We now instantiate this oracle inequality with the main-effects design in Example 5.1. Recall that $\mathcal{G}_1$ is the $L^2(P_\psi)$ closure of the class in Example 5.1. Let $\mathcal{P}^s_{d,1}$ be defined as $\mathcal{P}^s_{d,2}$ with $\mathcal{G}_2$ replaced by $\mathcal{G}_1$ and the bivariate components omitted from Equations (5.3) and (5.4). Using the same $\nu_n = c/\log n$ as above, define $K_{n,1}(\psi, \psi') = 1 + d^{-1} \sum_{j=1}^d W_1^{\mathrm{Mat},\nu_n}(\psi_j, \psi'_j)$.

Theorem 5.6 (Matched Main-Effects)**.** *Fix $n \geq 2$, d, smoothness $s > 0$, and $\varphi \in (0,1)$. Let $\sigma_n = \mathrm{GSW}_M(K_{n,1}, \varphi)$. For each $P \in \mathcal{P}^s_{d,1}$, define the residual $r = m - m_{\mathcal{G}_1}$. Then, uniformly over $P \in \mathcal{P}^s_{d,1}$, for a constant independent of n and d,*

$$\mathcal{V}_n(\sigma_n, P) \lesssim Q_n^M(r) + \left(\frac{d \log n}{n}\right)^{(2s)\wedge 1}. \tag{5.11}$$

Unlike Theorem 5.3, this bound contains no fixed variance gap: the residual enters through the within-pair energy $Q_n^M(r)$. If $d_n = o(n/\log n)$, the structured term vanishes and

$$\mathcal{V}_n(\sigma_n, P) \lesssim Q_n^M(r) + o(1).$$

If $Q_n^M(r) \to 0$ as well, then $\mathcal{V}_n(\sigma_n, P) \to 0$, and the design approaches the full Hahn bound. If $d \geq 3$, the matching M is 2-swap-stable for a cost satisfying Assumption 3.3, and r is β-Hölder with coefficient L_r for some $0 < \beta \leq 1$, then also

$$\mathcal{V}_n(\sigma_n, P) \lesssim L_r^2 (d\Lambda^2)^{2\beta} n^{-2\beta/d} + \left(\frac{d \log n}{n}\right)^{(2s)\wedge 1}.$$

Thus only the residual pays the rough-class matching rate from Section 3, while GSW attains the faster main-effects rate for the structured component.

Figure 3 illustrates the finite-sample gains from targeting a lower-complexity nonparametric function class when many covariates are recorded. When the nonparametric main-effects working model is well specified, so that $m = m_{\mathcal{G}_1}$, both main-effects designs strongly outperform matched pairs and full-dimensional GSW. When the true outcome model additionally contains interactions, $m \neq m_{\mathcal{G}_1}$ and matched pairs and full-dimensional GSW initially outperform main-effects GSW but deteriorate as the number of covariates increases and are eventually overtaken by it.

Matched main-effects GSW performs the best overall. It interpolates between the two regimes, attenuating the omitted interaction terms when the dimension is small but also preserving global balance when the dimension is large.

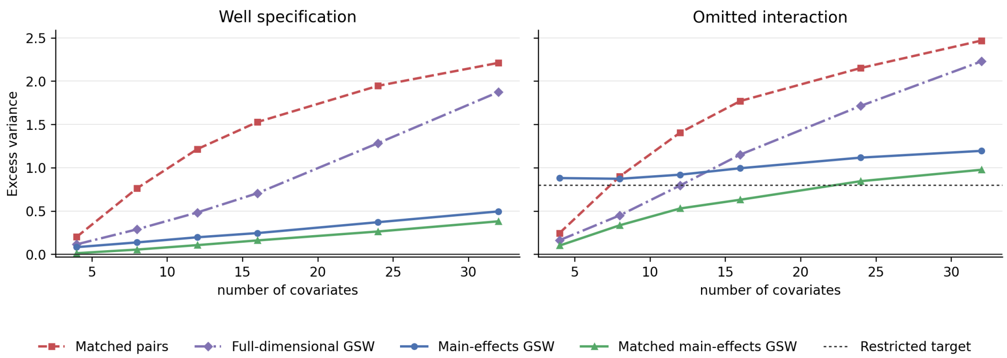

Figure 3: Excess variance at $n = 240$ for the main-effects design in Example 5.1. The panels correspond to well specification and additional interactions between the covariates, respectively. The dotted line is the working-model excess-variance target $\Delta_{\mathcal{G}_1}(P)$. The main-effects designs use the additive kernel K_1 from Example 5.1, an average of univariate Matérn kernels. All GSW designs use $\varphi = 0.03$ and $\nu = 1$. Appendix A.4 gives the exact model and simulation protocol.

6 Empirical Application

In this section, we evaluate our new designs using Monte Carlo simulations calibrated to 12 randomized experiments in recently published or forthcoming economics papers. We compare classical matched-pairs randomization with the robust GSW designs from Section 5, with kernels ranging from the full-dimensional Matérn to the reduced complexity nonparametric main effects specification.

The nonparametric additive and bivariate GSW designs in Section 5.1 reduce estimator variance relative to classical matched pairs in every experiment, while full-dimensional GSW as in Section 4 performs similarly to matched pairs on average. The matched versions of GSW in Section 5.2 provide further robustness: every design reduces estimator variance relative to classical matched pairs in every experiment. The main-effects specification in Example 5.1 delivers the largest variance reduction overall, which is reflected in shorter confidence intervals using our inference methods.

We selected 12 experiments from public replication data made available through the AEA, J-PAL, the World Bank, the *Journal of Development Economics*, and the *Quarterly Journal of Economics*. The screening considered data availability, sample size, and the existence of meaningful pretreatment covariates and was fixed before any design comparisons. Appendix A.3 describes the selection procedure in detail. The resulting experiments have sizes $n \in [96, 1{,}200]$ and covariate dimensions $d \in [7, 51]$.

Data Calibration. For each experiment, we calibrate the Monte Carlo population to its empirical covariate distribution and estimated conditional potential-outcome laws. For continuous outcomes, we model $Y_i(a) = m_a(\psi_i) + \sigma_a(\psi_i)\epsilon_{ia}$, $a \in \{0, 1\}$, where each ϵ_{ia} has a normal distribution. For binary outcomes, the conditional law is Bernoulli with success probability $m_a(\psi_i)$. We use cross-validation to select a regression model from linear regression, elastic net with interactions, random forests, and gradient boosting, while a shallow random forest estimates σ_a^2 from squared cross-fitted residuals. Clustered experiments are aggregated to the level of randomization using study-specific rules. We then draw covariates $\psi_{1:n}$ from a smoothed bootstrap and potential outcomes from the fitted conditional laws. See Appendix A.3 for details.

Designs and Estimators. We compare classical matched-pairs randomization with both unit-level and matched versions of the full-dimensional, bivariate, and additive GSW designs from Section 5. All GSW designs set $\varphi = 0.03$ and use the adaptive near-critical Matérn calibration $\nu_n = 1/\log n$. We form a common set of pairs using the matching algorithm in Cytrynbaum (2026). Under classical matched-pairs, the pair orientations $T_p \in \{-1, 1\}$ are independent, whereas matched GSW chooses them jointly using the Gram-Schmidt walk. Thus, in the simulation the matched designs differ only in how they jointly orient a common set of pairs. All designs use difference in means as the point estimator.

Inference. Matched GSW makes the pair orientations dependent, so the usual variance estimators available for classical matched pairs are not valid. Because of this, in Appendix A.1 we develop a new covariance corrected estimator for these designs. Our main result in Theorem A.1 establishes design-based finite-sample conservativeness for inference on the sample average treatment effect (SATE), while Proposition A.2 calibrates it for asymptotically non-conservative inference on the ATE. For each matched GSW design, we report the stabilized estimator $\widehat{\mathcal{V}}_{\text{stab}}$, implemented using an allocation bank of size $2J = 1024$. See the appendix for details.

For each simulated empirical environment, we draw 400 independent covariate populations and, within each population, 120 fresh potential outcome schedules and assignments. We report estimator variance reduction relative to matched pairs, coverage of nominal 95% ATE confidence intervals, and reduction in CI width.

Results. Figure 4 documents broad efficiency gains: aside from unit-level full-dimensional GSW, every GSW design reduces estimator variance relative to classical matched pairs in every experiment, and all three matched designs do so uniformly.

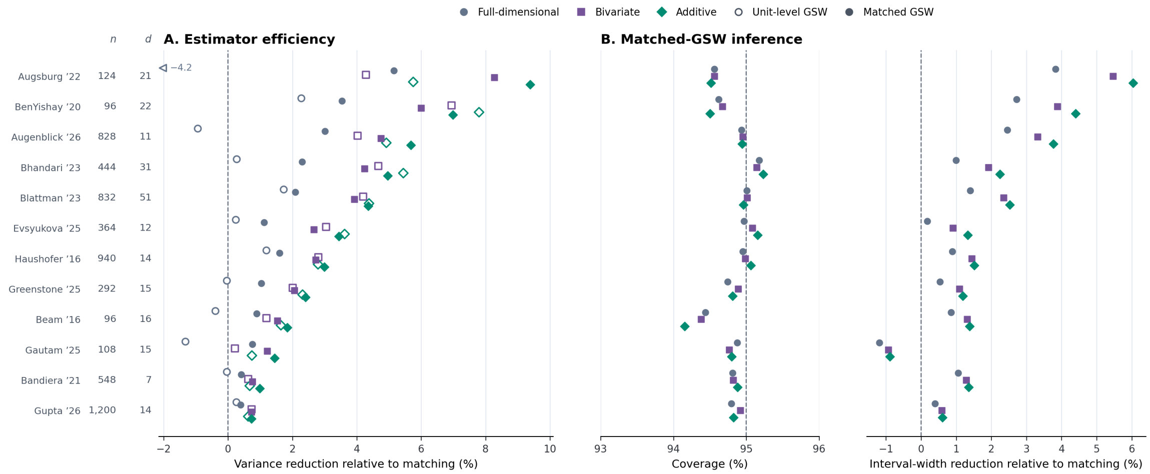

Figure 4: GSW relative to matched pairs across 12 simulated experiments, with sample size and covariate dimension shown at left. Panel A reports reduction in estimator variance for unit-level and matched GSW. Panel B reports coverage of nominal 95% ATE confidence intervals and the reduction in mean interval width for matched GSW.

The ranking among the unit-level designs illustrates the benefits from balancing lower-complexity nonparametric function spaces. Recall that full-dimensional GSW targets the true semiparametric efficiency bound $V^*(P)$, whereas the bivariate and additive designs target the potentially larger efficiency bounds $V^*(P) + \Delta_{\mathcal{G}}(P)$.

These lower-complexity spaces are much easier to balance in finite samples. As suggested by our rate theory, this advantage more than offsets the higher asymptotic variance target. In particular, the nonparametric main effects design performs best in all settings, despite having the largest asymptotic variance target.

Panel B focuses on inference for matched GSW. We develop new inference methods for these designs in Appendix A.1, which exploit the variance reduction to provide shorter confidence intervals. Average coverage is 94.8% for each matched design, with experiment-level coverage between 94.2% and 95.2%.

Taken together, these simulations show how the theoretical efficiency gains studied in the previous sections persist at the sample sizes and under the covariate distributions and potential-outcome relationships in actual experiments. They support matched additive GSW as our preferred empirical design, since it combines the strongest efficiency gains with shorter confidence intervals and coverage close to nominal across all 12 experiments.

References

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Abadie, A. and Imbens, G. W. (2008). Estimation of the conditional variance in paired experiments. *Annales d'Economie et de Statistique*, (91–92):175–187.

Adams, R. A. and Fournier, J. J. F. (2003). *Sobolev Spaces*. Academic Press, Amsterdam, 2 edition.

Armstrong, T. B. (2022). Asymptotic efficiency bounds for a class of experimental designs. arXiv:2205.02726.

Augenblick, N., Jack, B. K., Kaur, S., Masiye, F., and Swanson, N. (2026). Retrieval failures and consumption smoothing: A field experiment on seasonal poverty. Technical Report 35430, National Bureau of Economic Research.

Augsburg, B., Bancalari, A., Durrani, Z., Vaidyanathan, M., and White, Z. (2022). When nature calls back: Sustaining behavioral change in rural pakistan. *Journal of Development Economics*, 158:102933.

Bai, Y. (2022). Optimality of matched-pair designs in randomized controlled trials. *American Economic Review*, 112(12):3911–3940.

Bai, Y., Liu, J., Shaikh, A. M., and Tabord-Meehan, M. (2026). On the efficiency of highly stratified experiments. Forthcoming, The Annals of Statistics. arXiv:2307.15181.

Bai, Y., Romano, J. P., and Shaikh, A. M. (2022). Inference in experiments with matched pairs. *Journal of the American Statistical Association*, 117(540):1726–1737.

Bandiera, O., Best, M. C., Khan, A. Q., and Prat, A. (2021). The allocation of authority in organizations: A field experiment with bureaucrats. *The Quarterly Journal of Economics*, 136(4):2195–2242.

Bansal, N., Dadush, D., Garg, S., and Lovett, S. (2019). The Gram–Schmidt walk: A cure for the Banaszczyk blues. *Theory of Computing*, 15(21):1–27.

Beam, E. A. (2016). Do job fairs matter? experimental evidence on the impact of job-fair attendance. *Journal of Development Economics*, 120:32–40.

BenYishay, A., Jones, M., Kondylis, F., and Mobarak, A. M. (2020). Gender gaps in technology diffusion. *Journal of Development Economics*, 143:102380.

Bhandari, A., Larreguy, H., and Marshall, J. (2023). Able and mostly willing: An empirical anatomy of information's effect on voter-driven accountability in senegal. *American Journal of Political Science*, 67(4):1040–1066.

Blattman, C., Chaskel, S., Jamison, J. C., and Sheridan, M. (2023). Cognitive behavioral therapy reduces crime and violence over ten years: Experimental evidence. *American Economic Review: Insights*, 5(4):527–545.

Bloniarz, A., Liu, H., Zhang, C.-H., Sekhon, J. S., and Yu, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(27):7383–7390.

Broockman, D. E., Rhodes, E., Bartik, A. W., Dotson, K., Miller, S., Krause, P. K., and Vivalt, E. (2024). The causal effects of income on political attitudes and behavior: A randomized field experiment. NBER Working Paper No. 33214.

Chen, Q. and Zhang, P. (2026). Nonlinear covariate balance in experimental design. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*.

Cytrynbaum, M. (2026). Fine stratification of survey experiments. arXiv:2111.08157.

Derigs, U. (1988). Solving non-bipartite matching problems via shortest path techniques. *Annals of Operations Research*, 13(1):225–261.

Evsyukova, Y., Rusche, F., and Mill, W. (2025). Linkedout? a field experiment on discrimination in job network formation. *The Quarterly Journal of Economics*, 140(1):283–334.

Fischer, S. and Steinwart, I. (2020). Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21:1–38.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33:503–513.

Fogarty, C. B. (2018). On mitigating the analytical limitations of finely stratified experiments. *Journal of the Royal Statistical Society: Series B*, 80(5).

Gautam, S., Gechter, M., Guiteras, R. P., and Mobarak, A. M. (2025). To use financial incentives or not? insights from experiments in encouraging sanitation investments in four countries. *World Development*, 187:106791.

Giné, E. and Nickl, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*, volume 40 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

Greenstone, M., Pande, R., Ryan, N., and Sudarshan, A. (2025). Can pollution markets work in developing countries? experimental evidence from india. *The Quarterly Journal of Economics*, 140(2):1003–1060.

Gupta, S. (2026). Perils of the paperwork: The impact of information and application assistance on social benefit take-up in india. *American Economic Journal: Economic Policy*. Forthcoming.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953). *Sample Survey Methods and Theory*. Wiley.

Harshaw, C., Middleton, J., and Sävje, F. (2026). Optimized variance estimation under interference and complex experimental designs. *Journal of the American Statistical Association*.

Harshaw, C., Sävje, F., Spielman, D. A., and Zhang, P. (2024). Balancing covariates in randomized experiments with the Gram–Schmidt walk design. *Journal of the American Statistical Association*, 119(548):2934–2946.

Harshaw, C. R. (2021). *Algorithmic Advances for the Design and Analysis of Randomized Experiments*. PhD thesis, Yale University.

Hastie, T. J. and Tibshirani, R. J. (1986). Generalized additive models. *Statistical Science*, 1(3):297–310.

Haushofer, J. and Shapiro, J. (2016). The short-term impact of unconditional cash transfers to the poor: Experimental evidence from kenya. *The Quarterly Journal of Economics*, 131(4):1973–2042.

Imai, K. (2008). Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Statistics in Medicine*, 27:4857–4873.

Kallus, N. (2018). Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society: Series B*, 80(1):85–112.

Kallus, N. (2021). On the optimality of randomization in experimental design: How to randomize for minimax variance and design-based inference. *Journal of the Royal Statistical Society: Series B*, 83(2):404–409.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. (2018). Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*.

Kapelner, A., Krieger, A., and Azriel, D. (2022). The role of pairwise matching in experimental design for an incidence outcome. arXiv:2209.00490.

Kasy, M. (2016). Why experimenters might not always want to randomize, and what they could do instead. *Political Analysis*, 24(3):324–338.

Kennedy, E. H., Balakrishnan, S., Robins, J. M., and Wasserman, L. (2024). Minimax rates for heterogeneous causal effect estimation. *The Annals of Statistics*, 52(2):793–816.

Krieger, A., Azriel, D., and Kapelner, A. (2019). Nearly random designs with greatly improved balance. *Biometrika*.

Lei, L. and Ding, P. (2021). Regression adjustment in completely randomized experiments with a diverging number of covariates. *Biometrika*, 108(4):815–828.

Li, X., Ding, P., and Rubin, D. B. (2018). Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, 115(37):9157–9162.

Lu, X., Yang, F., and Wang, Y. (2025). Debiased regression adjustment in completely randomized experiments with moderately high-dimensional covariates. *The Annals of Statistics*, 53(4):1535–1558.

Morgan, K. L. and Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *Annals of Statistics*, 40(2).

Pashley, N. E. and Miratrix, L. W. (2021). Insights on variance estimation for blocked and matched pairs designs. *Journal of Educational and Behavioral Statistics*.

Paulsen, V. I. and Raghupathi, M. (2016). *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*, volume 152 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA.

Sobol, I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Mathematical Modeling and Computational Experiment*, 1(4):407–414.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2):689–705.

Taylor, M. E. (2011). *Partial Differential Equations I: Basic Theory*, volume 115 of *Applied Mathematical Sciences*. Springer, New York, 2 edition.

van der Vaart, A. W. and van Zanten, J. H. (2011). Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research*, 12:2095–2119.

Wager, S., Du, W., Taylor, J., and Tibshirani, R. J. (2016). High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(45):12673–12678.

Wang, Y. and Li, X. (2022). Rerandomization with diminishing covariate imbalance and diverging number of covariates. *The Annals of Statistics*, 50(6):3439–3465.

Wang, Y. and Li, X. (2025). Asymptotic theory of the best-choice rerandomization using the mahalanobis distance. *Journal of Econometrics*, 251:106049.

Wendland, H. (2005). *Scattered Data Approximation*. Cambridge University Press, Cambridge.

Yang, Y. and Barron, A. R. (1999). Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599.

A Appendix

A.1 Inference for Matched Gram-Schmidt Walk Designs

Matched GSW makes the orientations of different pairs dependent, so the usual matched-pairs variance estimator does not apply. In this section, we construct a design-based variance estimator for matched GSW that is exactly conservative in finite samples for inference on the sample average treatment effect. We also adapt the construction for inference on the population average treatment effect.

Suppose that $n = 2N$, where N is even, and write $M = \{(i_p, j_p) : p \in [N]\}$ for the original matching. Let $T_p = 1$ when i_p is treated and $T_p = -1$ otherwise. Recall $\tau_i = Y_i(1) - Y_i(0)$ and $\bar{Y}_i = (Y_i(1) + Y_i(0))/2$, and define $A_p = (\tau_{i_p} + \tau_{j_p})/2$ and $B_p = \bar{Y}_{i_p} - \bar{Y}_{j_p}$. The two potential treated-minus-control contrasts in pair p are $C_p(1) = Y_{i_p}(1) - Y_{j_p}(0)$ and $C_p(-1) = Y_{j_p}(1) - Y_{i_p}(0)$. Only the contrast selected by T_p is observed, and it satisfies $C_p = C_p(T_p) = T_p(Y_{i_p} - Y_{j_p}) = A_p + T_p B_p$. Then

$$\widehat{\theta} = \frac{1}{N}\sum_{p=1}^{N} A_p + \frac{1}{N} T' B. \tag{A.1}$$

Orientation bank. Conditional on $\psi_{1:n}$ and the matching M, we draw independent matched-GSW orientation vectors $T^{(1)}, \ldots, T^{(J)}$ and form the sign-symmetric bank of allocations $\mathcal{B}_J = \big(T^{(1)}, -T^{(1)}, \ldots, T^{(J)}, -T^{(J)}\big)$. We select the implemented orientation uniformly from the $2J$ indexed entries of $\mathcal{B}_J$. Because the matched-GSW law is sign symmetric, this leaves its marginal assignment law unchanged. Recall $\mathcal{W}_n = \sigma((\psi_i, Y_i(0), Y_i(1))_{i=1}^n)$. Let U_n be any exogenous randomness used during matching and set $\mathcal{W}_{n,J} = \sigma(\mathcal{W}_n, U_n, \mathcal{B}_J)$. By sign symmetry, $E[T|\mathcal{W}_{n,J}] = 0$. The conditional covariance matrix is $R \equiv E[TT'|\mathcal{W}_{n,J}] = J^{-1}\sum_{j=1}^{J} T^{(j)} T^{(j)\prime}$. Then

$$n \operatorname{Var}(\widehat{\theta}|\mathcal{W}_{n,J}) = \frac{2}{N} B' R B. \tag{A.2}$$

Following the quadratic-bound approach to design-based variance estimation of Harshaw et al. (2026), let L be any outcome-blind positive semidefinite $N \times N$ matrix with $L_{pp} = 1$ that is measurable with respect to the realized covariates, matchings, and bank. If $|R_{pq}| < 1$ for every $p \neq q$, define variance estimator

$$\widehat{\mathcal{V}}(L) \equiv \frac{2}{N}\left(\sum_{p=1}^{N} C_p^2 + \sum_{p \neq q} \frac{T_p T_q R_{pq} + L_{pq}}{1 + T_p T_q R_{pq}} C_p C_q\right). \tag{A.3}$$

Define the sample average treatment effect $\mathrm{SATE}_n \equiv n^{-1}\sum_{i=1}^{n} \tau_i = N^{-1}\sum_{p=1}^{N} A_p$. Sign symmetry gives $E[\widehat{\theta}|\mathcal{W}_{n,J}] = \mathrm{SATE}_n$ and hence $E[\widehat{\theta}|\mathcal{W}_n] = \mathrm{SATE}_n$. The proof of the following result is given in Appendix B.5.

Theorem A.1 (Design-Based Conservativeness for SATE). *For every matrix L as above, $E[\widehat{\mathcal{V}}(L)|\mathcal{W}_{n,J}] = n\,\mathrm{Var}(\widehat{\theta}|\mathcal{W}_{n,J}) + 2A'LA/N$. In particular, $\widehat{\mathcal{V}}(L)$ is conservative for design-based variance of $\widehat{\theta}$ about the* SATE_n*:*

$$E[\widehat{\mathcal{V}}(L)|\mathcal{W}_n] \geq n\,\mathrm{Var}(\widehat{\theta}|\mathcal{W}_n). \tag{A.4}$$

Inference on the SATE. Theorem A.1 establishes finite-sample conservatism for inference on SATE_n, with the choice of L determining its degree. To limit this conservativeness, apply the fine-stratification matcher of Cytrynbaum (2026) to the covariate centroids of the N original pairs, and let Π denote the resulting outer matching. For each $p \in [N]$, write $\pi(p)$ for its outer mate and define the swap matrix S_Π by $(S_\Pi x)_p = x_{\pi(p)}$. The matrix $L_\Pi = I_N - S_\Pi$ is PSD with unit diagonal, and we have

$$E[\widehat{\mathcal{V}}(L_\Pi)|\mathcal{W}_{n,J}] - n\,\mathrm{Var}(\widehat{\theta}|\mathcal{W}_{n,J}) = \frac{2}{N}A'L_\Pi A = \frac{2}{N}\sum_{\{p,q\}\in\Pi}(A_p - A_q)^2. \tag{A.5}$$

Thus, $\widehat{\mathcal{V}}(L_\Pi)$ is exactly conservative for SATE_n, with smaller slack when the outer matching groups pairs with similar pair-average treatment effects.

Inference on the ATE. The law of total variance gives $n\,\mathrm{Var}(\widehat{\theta}) = E[n\,\mathrm{Var}(\widehat{\theta}|\mathcal{W}_{n,J})] + \mathrm{Var}(\tau)$. Then for inference on the ATE, we choose L so the term $2A'LA/N$ converges to $\mathrm{Var}(\tau)$. Set $L_0 = N(I_N - 11'/N)/(N-1)$ and $L_{\mathrm{pop}} = (L_\Pi + L_0)/2$, which are PSD with unit diagonal. Writing $\bar{A} = N^{-1}\sum_{p=1}^{N} A_p$, the new component satisfies $N^{-1}A'L_0A = (N-1)^{-1}\sum_{p=1}^{N}(A_p - \bar{A})^2$ and measures the global dispersion of pair-average treatment effects. The following result formalizes this calibration. Its proof is given in Appendix B.5.

Proposition A.2 (ATE Calibration)**.** *Under Assumption 2.1, suppose the matching satisfies Equation (3.2) and construct* Π *as above. Then* $2A'L_{\text{pop}}A/N \xrightarrow{p} \mathrm{Var}(\tau)$.

The covariance-corrected estimator can be noisy. By the Schur product theorem, $L_S = R \circ R$ is positive semidefinite with unit diagonal, and substituting it into Equation (A.3) gives the always-nonnegative estimator $\widehat{\mathcal{V}}_S = 2(D_T C)'R(D_T C)/N$, where D_T is diagonal with entries T_p. Take $\lambda_N = N^{-1/2}$, using $L = L_\Pi$ for SATE inference and $L = L_{\text{pop}}$ for ATE inference. When $|R_{pq}| < 1$ for every $p \neq q$, we report the nominal interval $\widehat{\theta} \pm z_{1-\alpha/2}(\widehat{\mathcal{V}}_{\text{stab}}(L)/n)^{1/2}$, where

$$\widehat{\mathcal{V}}_{\text{stab}}(L) \equiv \max\bigl((1-\lambda_N)\widehat{\mathcal{V}}(L) + \lambda_N \widehat{\mathcal{V}}_S, \lambda_N^2 \widehat{\mathcal{V}}_S\bigr). \tag{A.6}$$

Otherwise, we use $\widehat{\mathcal{V}}_S$ in place of $\widehat{\mathcal{V}}_{\text{stab}}(L)$.

A.2 Hölder-to-Sobolev Transfer

For $0 < \beta \leq 1$, recall $[f]_\beta = \sup_{x \neq y} |f(x) - f(y)|/\|x-y\|_2^\beta$ and let $\mathcal{P}_{d,\beta}^{\text{Hol}}$ contain the laws for which $\|m\|_\infty + [m]_\beta \leq 1$. We use the following embedding lemma to transfer Theorem 4.7 to these classes.

Lemma A.3 (Hölder-to-Sobolev Embedding)**.** *Fix* $d \geq 1$ *and* $0 < \beta < 1$*. There is a constant* $C_{d,\beta} < \infty$ *such that every bounded* $f : [0,1]^d \to \mathbb{R}$ *with* $[f]_\beta < \infty$ *and every* $s \in [\beta/2, \beta)$ *admit* $F_s \in H^s(\mathbb{R}^d)$ *satisfying* $F_s = f$ *on* $[0,1]^d$ *and*

$$\|F_s\|_{H^s(\mathbb{R}^d)}^2 \leq \frac{C_{d,\beta}}{\beta - s}\left(\|f\|_\infty + [f]_\beta\right)^2. \tag{A.7}$$

Every bounded Lipschitz function $f : [0,1]^d \to \mathbb{R}$ *admits* $F_1 \in H^1(\mathbb{R}^d)$ *satisfying* $F_1 = f$ *on* $[0,1]^d$ *and* $\|F_1\|_{H^1(\mathbb{R}^d)} \leq C_d(\|f\|_\infty + [f]_1)$.

Thus, a β-Hölder outcome model can be treated as a Sobolev model of any order $s < \beta$. Choosing s just below β trades the norm inflation $(\beta - s)^{-1}$ against the rate in Theorem 4.7. Taking $s = \beta - 1/\log n$ limits this inflation to a logarithmic factor while preserving the exponent $2\beta/d$ up to constants. When $\beta = 1$, the endpoint embedding applies directly without this additional logarithmic loss. The following corollary records the resulting guarantee.

Corollary A.4 (Hölder Adaptation). *Fix $d \geq 3$, $0 < \beta \leq 1$, $c > 0$, and $\varphi \in (0,1)$. Let σ_n be the design in Theorem 4.7. There is a constant C depending only on $(d, \beta, c, \varphi, \overline{p})$ such that, for all sufficiently large n,*

$$\sup_{P \in \mathcal{P}^{\mathrm{Hol}}_{d,\beta}} \left\{ n \operatorname{Var}_{P,\sigma_n}(\widehat{\theta}) - V^*(P) \right\} \leq C \begin{cases} \log n \left(\dfrac{n}{\log n} \right)^{-2\beta/d}, & 0 < \beta < 1, \\ \left(\dfrac{n}{\log n} \right)^{-2/d}, & \beta = 1. \end{cases} \tag{A.8}$$

Proof. Suppose first that $0 < \beta < 1$ and set $s_n = \beta - 1/\log n$. For all sufficiently large n, $s_n \in [\beta/2, \beta)$. Because $d \geq 3$ and $\beta < 1$, we have $s_n < 1 < d/2$, so only the subcritical calculation in Lemma B.7 is required. Its constants are uniform over $s_n \in [\beta/2, \beta)$ because s_n is bounded away from zero and the s_n-dependent exponential factor in its proof is bounded by $e^{4\beta c}$. Set $\lambda_n = \max\{2\varphi/((1-\varphi)n), n^{-1}\}$. For fixed φ, $n^{-1} \leq \lambda_n \leq C_\varphi/n$, so $\lambda_n \leq c_0/\log n$ for all sufficiently large n and the penalty condition in Lemma B.7 holds. Fix $P \in \mathcal{P}^{\mathrm{Hol}}_{d,\beta}$ and let F_{s_n} be the extension of m supplied by Lemma A.3. Since $F_{s_n} = m$ on the unit cube and the dimension-normalized spectral weight is no larger than the ordinary Sobolev weight, the oracle calculation in the proof of Theorem 4.7 and Lemma B.7 give

$$\|F_{s_n}\|^2_{H^{s_n}_{\mathrm{av}}(\mathbb{R}^d)} \leq \|F_{s_n}\|^2_{H^{s_n}(\mathbb{R}^d)} \leq \frac{C_{d,\beta}}{\beta - s_n} (\|m\|_\infty + [m]_\beta)^2 \leq C_{d,\beta} \log n,$$
$$\mathcal{V}_n(P) \leq C \|F_{s_n}\|^2_{H^{s_n}_{\mathrm{av}}(\mathbb{R}^d)} (\lambda_n \log n)^{2s_n/d} \leq C \log n \left(\frac{\log n}{n} \right)^{2s_n/d}.$$

Moreover,

$$\left(\frac{n}{\log n} \right)^{-2s_n/d} \leq e^{2/d} \left(\frac{n}{\log n} \right)^{-2\beta/d}. \tag{A.9}$$

This proves the first branch uniformly over $P \in \mathcal{P}^{\mathrm{Hol}}_{d,\beta}$. When $\beta = 1$, the Lipschitz conclusion of Lemma A.3 supplies an extension F_1 whose $H^1(\mathbb{R}^d)$ norm is bounded by a constant depending only on d. The spectral comparison above also bounds its $H^1_{\mathrm{av}}(\mathbb{R}^d)$ norm. Since $1 < d/2$, the same oracle and approximation calculation at $s = 1$ gives $\mathcal{V}_n(P) \leq C(\log n/n)^{2/d}$, proving the second branch. □

A.3 Empirical Application Details

Study selection. We constructed a fixed candidate frame of 100 randomized studies from 247 records in five public replication repositories. We sourced 30 studies from AEA/openICPSR, 20 each from the J-PAL Dataverse, World Bank Impact Evaluation Surveys, and the *Journal of Development Economics* Mendeley collection, and 10 from the *Quarterly Journal of Economics* Dataverse. Within each source, we verified randomized assignment and selected the most recent qualifying studies.

We screened the replication packages for data containing a treatment, outcome, at least five meaningful pretreatment covariates, and a recoverable randomization unit. We also required an adequate number of randomization units and legally usable files. Thirty-five studies passed this initial screen. We ranked these studies by their baseline-covariate counts, subject to a cap of six studies from any source, and retained 20 for a detailed file-level audit.

Neither treatment-effect estimates, balance statistics, estimator variances, nor the performance of any design entered the screening or ranking decisions. One selected package could not be fitted because its raw microdata remained behind an access agreement, leaving 19 fitted experimental populations.

The final design-independent screen removed populations with insufficient covariate signal. We retained fitted populations satisfying

$$\frac{4\,\mathrm{Var}(m(\psi))}{V^*(P) + 4\,\mathrm{Var}(m(\psi))} \geq 0.05.$$

This quantity is the share of complete-randomization variance that any covariate-adaptive design could potentially remove. When it is small, there is too little scope for evaluating the relative performance of different covariate-balancing designs. The rule was fixed before any design comparisons and yielded 13 eligible fitted populations, of which Figure 4 reports 12, with one excluded due to Monte Carlo time considerations.

Table 1: Calibrated experimental populations in Figure 4. The sample size is the Monte Carlo panel size after rounding to the nearest multiple of four. RF denotes a shallow random forest and ENet-2 denotes an elastic net with degree-two terms.

Study	Outcome	Randomization unit	n	d	Mean model
Gupta (2026)	Administrative pension receipt	Eligible woman	1,200	14	RF
Blattman et al. (2023)	Antisocial-behavior index	Individual	832	51	RF
Gautam et al. (2025)	Sanitation ownership rate	Village	108	15	RF
Bhandari et al. (2023)	Evaluation of the incumbent	Village	444	31	RF
Augsburg et al. (2022)	Open-defecation rate	Village	124	21	ENet-2
BenYishay et al. (2020)	Farmer knowledge score	Village	96	22	ENet-2
Beam (2016)	Employment rate	Neighborhood	96	16	RF
Evsyukova et al. (2025)	LinkedIn connections	Profile	364	12	RF
Augenblick et al. (2026)	Savings in maize-equivalent kilograms	Household	828	11	OLS
Greenstone et al. (2025)	Log particulate emissions	Plant	292	15	RF
Bandiera et al. (2021)	Item-adjusted log purchase price	Office	548	7	RF
Haushofer and Shapiro (2016)	Nondurable consumption	Household	940	14	RF

Fitted populations. For each study, we reproduce the paper's treatment contrast and outcome sample, applying documented study-specific sample restrictions and missing-data rules. When treatment is assigned by cluster, we average outcomes and baseline covariates within each randomization cluster. We verify the resulting sample by reproducing a treatment-effect estimate reported in the original paper.

For each arm $a \in \{0, 1\}$, let $\widehat{m}_a(\psi)$ and $\widehat{v}_a(\psi)$ denote the fitted conditional mean and variance. Five-fold cross-validation selects one mean-model family for each study from OLS, an elastic net with degree-two terms, a shallow random forest, and gradient-boosted trees. For continuous outcomes, the score is out-of-sample R^2. For binary outcomes, it is the improvement in Brier loss over the arm mean. The family with the highest score averaged across the two arms is then fit separately within each arm on all observed units. For continuous outcomes, a shallow random forest estimates the conditional variance from squared five-fold cross-fitted residuals. For binary outcomes, the conditional variance is $\widehat{m}_a(\psi)(1-\widehat{m}_a(\psi))$.

To simulate covariates, we draw an observed covariate vector uniformly and add independent Gaussian jitter to continuous coordinates having at least 15 distinct values. The jitter standard deviation is the mean nearest-neighbor spacing, capped at one quarter of the coordinate's sample standard deviation, and perturbed values are clipped to the observed range. The fitted outcome models use these perturbed covariates on their original scale. After perturbation, a separate copy of the design covariates is mapped to $[0,1]^d$ before constructing the kernels, using the empirical CDF for higher-cardinality coordinates and min-max scaling otherwise, with unordered categories represented by normalized indicator blocks.

At each Monte Carlo iteration, we evaluate the fitted conditional laws at the simulated covariates and draw both potential outcomes. For continuous outcomes, $Y(a)=\widehat{m}_a(\psi)+\widehat{v}_a(\psi)^{1/2}\epsilon_a$ with Gaussian innovations, followed by clipping to any declared outcome bounds. For binary outcomes, we draw Bernoulli potential outcomes with success probability $\widehat{m}_a(\psi)$. The main specification takes the two innovations to be conditionally independent. The sensitivity specifications use a Gaussian copula with correlation 0.5 or a common latent rank, including the analogous latent-Gaussian construction for binary outcomes. These alternatives change the across-study average variance reduction by at most 0.04 percentage points for any GSW design. The fitted models are held fixed, so the reported Monte Carlo uncertainty does not include model-estimation uncertainty.

For each study, we draw 400 independent covariate populations. Within each population, we construct the matching and a sign-symmetric assignment bank, as in Appendix A.1, for each of the three unit-level and three matched GSW designs. Each bank contains 512 independently generated walk allocations and their sign reversals, giving 1,024 possible allocations. We then draw 120 potential-outcome schedules and one implemented allocation from each design for every schedule.

A.4 Simulation Designs

This subsection records the models and Monte Carlo protocols for Figures 1–3. The calibrated empirical populations in Figure 4 are described in Section 6. At every parameter configuration, we average over 100 independent covariate samples $\psi_{1:n}^{(r)}$, $r \in [100]$, drawn from the uniform distribution on the relevant unit cube. Within each sample, design expectations that are not available in closed form are approximated using 100 assignment vectors. The matched designs use minimum-total squared-Euclidean matching constructed separately within each covariate sample.

Matching reversal. For Figure 1, let $D = 20$, define $v(x) = \sqrt{12}(x - 1/2)$ and $w(x) = \sqrt{2}\cos(2\pi x)$, and set $a_j = 0.8^{j-1}/(\sum_{\ell=1}^{D} 0.8^{2(\ell-1)})^{1/2}$. The linear and nonlinear outcome models are $m_{\text{lin}}(\psi) = \sum_{j=1}^{D} a_j v(\psi_j)$ and $m_{\text{add}}(\psi) = \sum_{j=1}^{D} a_j w(\psi_j)$, respectively. Both have unit variance and the nonlinear main effect is orthogonal to a linear function of its coordinate. Each design observes only the first d coordinates, although its variance is evaluated under the corresponding full outcome model. The dotted line is the efficiency bound using the observed coordinates, $V_d = 1 + 4\sum_{j>d} a_j^2$.

Let $b_d(Z) = n^{-1}\sum_{i=1}^{n} Z_i \psi_{i,1:d}$ and let $\widehat{\Sigma}_d$ be the sample covariance matrix of the observed covariates. Linear GSW applies the walk to their Mahalanobis-standardized values and uses $\varphi = 0.1$. Best-of-M rerandomization selects, from $M = 10{,}000$ independent randomizations, the allocation minimizing $b_d(Z)'\widehat{\Sigma}_d^{-1} b_d(Z)$. Nonparametric GSW uses the additive kernel K_1 from Example 5.1 with $\nu = 1$ and $\varphi = 0.03$. All five designs are evaluated under both outcome models.

Smooth outcomes. For Figure 2, let W_ν be the unit-length-scale Matérn kernel with parameter ν and let ψ and ψ' be independent and uniform on $[0,1]^d$. Define $c_\nu = 1 - E[W_\nu(\psi, \psi')]$ and let $m_\nu \sim \mathrm{GP}(0, W_\nu/c_\nu)$. This normalization gives $E[\mathrm{Var}(m_\nu(\psi)|m_\nu)] = 1$. The figure reports prior-averaged excess variance relative to matched pairs. Equivalently, the plotted value for a design σ is $E_\sigma[Z'W_\nu Z]/E_{\text{MP}}[Z'W_\nu Z]$. In the left panel, $n = 240$, $d = 8$, and ν ranges over $\{0.1, 0.2, 0.5, 1, 2, 4, 8, 16, 32\}$. Fixed Matérn GSW uses W_1, while Oracle Matérn GSW and kernel rerandomization use W_ν. In the right panel, $d = 8$, fixed Matérn GSW uses W_1, and n ranges over $\{80, 120, 180, 240, 360, 480, 720\}$ for $\nu \in \{0.5, 1, 4, 16\}$. All GSW designs use $\varphi = 0.03$.

Structured balance. For Figure 3, $u_j(x) = \sqrt{2}\cos(\pi x)$, $g_d(\psi) = d^{-1/2}\sum_{j=1}^{d} u_j(\psi_j)$, and $r(\psi) = u_1(\psi_1)u_2(\psi_2)$. Under the uniform covariate law, g_d and r have unit variance and are orthogonal. For $\rho \geq 0$, the outcome model is

$$m_\rho(\psi) = \frac{g_d(\psi) + \rho r(\psi)}{\sqrt{1+\rho^2}}. \tag{A.10}$$

The left panel uses $\rho = 0$, so $m = m_{\mathcal{G}_1}$. The right panel uses $\rho = 0.5$, so the omitted interaction contributes one fifth of the predictable variance and $V^*(P) + \Delta_{\mathcal{G}_1}(P) = V^*(P) + 0.8$. Full-dimensional GSW uses the Matérn kernel on $[0,1]^d$, main-effects GSW uses the additive kernel K_1, and matched main-effects GSW applies the same kernel to pair orientations. All kernels use $\nu = 1$, and all GSW designs use $\varphi = 0.03$.

A.5 The Gram-Schmidt Walk

This subsection gives the formal description of the assignment algorithm used in Section 4.2. The construction is the Gram-Schmidt walk of Bansal et al. (2019), with the zero initialization and randomized pivot order of Harshaw et al. (2024) and the Gram-matrix formulation in Harshaw (2021, Section 2.7).

Let Γ be an $n \times n$ positive semidefinite matrix with diagonal entries at most one. The coordinates of the walk index units. Starting from $z^{(0)} = 0$, define the active set

$$\mathcal{A}_t = \left\{ i \in [n] : |z_i^{(t)}| < 1 \right\}. \tag{A.11}$$

Its elements are the units whose assignments remain fractional. The algorithm selects a pivot $p_t \in \mathcal{A}_t$ uniformly at random and retains that pivot until it reaches -1 or 1. When the current pivot leaves $\mathcal{A}_t$, the algorithm selects a new pivot uniformly from the remaining active units. Given $\mathcal{A}_t$ and p_t, choose

$$u^{(t)} \in \operatorname{argmin}\left\{ u'\Gamma u : u \in \mathbb{R}^n, u_{p_t} = 1, u_i = 0 \text{ for every } i \notin \mathcal{A}_t \right\}. \tag{A.12}$$

Let δ_t^+ and δ_t^- be the largest nonnegative values such that $z^{(t)} + \delta_t^+ u^{(t)}$ and $z^{(t)} - \delta_t^- u^{(t)}$ remain in $[-1,1]^n$. The next assignment is

$$z^{(t+1)} = \begin{cases} z^{(t)} + \delta_t^+ u^{(t)}, & \text{with probability } \dfrac{\delta_t^-}{\delta_t^+ + \delta_t^-}, \\ z^{(t)} - \delta_t^- u^{(t)}, & \text{with probability } \dfrac{\delta_t^+}{\delta_t^+ + \delta_t^-}. \end{cases} \tag{A.13}$$

The probabilities in Equation (A.13) make $E[z^{(t+1)}|z^{(t)}, \mathcal{A}_t, p_t] = z^{(t)}$. At least one active coordinate reaches the boundary after every update. The algorithm therefore terminates at an assignment $Z \in \{-1, 1\}^n$ after at most n updates.

To connect this formulation to the ordinary vector algorithm, take any factor C satisfying $C'C = \Gamma$ and let c_i be its ith column. Then $\|\sum_{i=1}^n u_i c_i\|_2^2 = u'\Gamma u$, so Equation (A.12) is exactly the direction chosen by ordinary GSW applied to the columns of C. The direction can also be computed directly from Γ, so an explicit factorization and an explicit feature map are not part of the definition.

For the kernelized design, Equation (4.6) supplies the input matrix. The covariance guarantee $\mathrm{Cov}(Z|\psi_{1:n}) \preceq \Gamma_n^{-1} \preceq \varphi^{-1} I_n$ prevents balance of the selected kernel features from creating arbitrarily large variance in directions not described by the kernel. See Harshaw et al. (2024) for a detailed design-based interpretation.

B Proofs

B.1 Setup

For each unit, write $W_i = (\psi_i, Y_i(0), Y_i(1))$.

Proof of Proposition 2.3. Fix $\sigma \in \mathcal{D}_n$. Let $\mathcal{F}_n = \sigma(\psi_{1:n})$ and $\mathcal{W}_n = \sigma(W_{1:n})$ denote the corresponding sigma-algebras, with $\mathcal{F}_n \subseteq \mathcal{W}_n$. All expectations and variances below are taken under the joint law (P, σ) unless stated otherwise.

Using Equation (2.1), write $\widehat{\theta} - \theta(P) = A_n + B_n$, where $A_n = n^{-1} \sum_{i=1}^n \tau_i - \theta(P)$ and $B_n = 2n^{-1} \sum_{i=1}^n Z_i \bar{Y}_i$. Definition 2.2 gives $Z \perp\!\!\!\perp W_{1:n} | \mathcal{F}_n$ and $E[Z_i|\mathcal{F}_n] = 0$. It follows that $E[Z_i|\mathcal{W}_n] = E[Z_i|\mathcal{F}_n] = 0$. Hence $E[B_n|\mathcal{W}_n] = 0$, while iid sampling gives $E[A_n] = 0$. It follows that $\widehat{\theta}$ is unbiased. Moreover, A_n is $\mathcal{W}_n$-measurable, so tower law gives $E[A_n B_n] = E[A_n E[B_n|\mathcal{W}_n]] = 0$. Then we have

$$n\,\mathrm{Var}(\widehat{\theta}) = nE[(A_n + B_n)^2] = \mathrm{Var}(\tau) + \frac{4}{n} E\left[\left(\sum_{i=1}^n Z_i \bar{Y}_i\right)^2\right]. \tag{B.1}$$

Write $\varepsilon_i = \bar{Y}_i - m(\psi_i)$ and $\bar{\sigma}^2(x) = \mathrm{Var}(\bar{Y}|\psi = x)$. For a random element A, write $\mathcal{L}(A|B)$ for its conditional law given B. By iid sampling, the conditional law factorizes as $\mathcal{L}(\varepsilon_{1:n}|\mathcal{F}_n) = \Pi_{i=1}^n \mathcal{L}(\varepsilon_i|\psi_i)$. The conditional independence in Definition 2.2 gives $\varepsilon_{1:n} \perp\!\!\!\perp Z|\mathcal{F}_n$, and therefore $\mathcal{L}(\varepsilon_{1:n}|\mathcal{F}_n, Z) = \mathcal{L}(\varepsilon_{1:n}|\mathcal{F}_n)$. Taking the ith marginal of this law gives $E[\varepsilon_i|\mathcal{F}_n, Z] = E[\varepsilon_i|\mathcal{F}_n] = E[\varepsilon_i|\psi_i] = 0$. For $i \neq j$, the product law gives $E[\varepsilon_i\varepsilon_j|\mathcal{F}_n, Z] = E[\varepsilon_i|\psi_i]E[\varepsilon_j|\psi_j] = 0$, while its ith marginal gives $E[\varepsilon_i^2|\mathcal{F}_n, Z] = E[\varepsilon_i^2|\psi_i] = \bar{\sigma}^2(\psi_i)$. Set $C_n = \sum_{i=1}^n Z_i m(\psi_i)$ and $R_n = \sum_{i=1}^n Z_i\varepsilon_i$. The preceding moment identities give $E[R_n|\mathcal{F}_n, Z] = \sum_{i=1}^n Z_i E[\varepsilon_i|\mathcal{F}_n, Z] = 0$. They also give

$$E[R_n^2|\mathcal{F}_n, Z] = \sum_{i,j=1}^n Z_i Z_j E[\varepsilon_i\varepsilon_j|\mathcal{F}_n, Z] = \sum_{i=1}^n \bar{\sigma}^2(\psi_i).$$

The last equality uses $Z_i^2 = 1$. Since C_n is measurable with respect to $\sigma(\mathcal{F}_n, Z)$,

$$E\left[\left(\sum_{i=1}^n Z_i\bar{Y}_i\right)^2\middle|\mathcal{F}_n, Z\right] = C_n^2 + 2C_nE[R_n|\mathcal{F}_n, Z] + E[R_n^2|\mathcal{F}_n, Z] = C_n^2 + \sum_{i=1}^n \bar{\sigma}^2(\psi_i).$$

Taking expectations and using identical sampling gives

$$E\left[\left(\sum_{i=1}^n Z_i\bar{Y}_i\right)^2\right] = E\left[\left(\sum_{i=1}^n Z_i m(\psi_i)\right)^2\right] + nE[\bar{\sigma}^2(\psi)]. \tag{B.2}$$

It remains to identify the design-invariant terms. Let $\rho(x) = \mathrm{Cov}(Y(1), Y(0)|\psi = x)$. The law of total variance gives $\mathrm{Var}(\tau) = \mathrm{Var}\big(\tau(\psi)\big) + E[v_1^2(\psi) + v_0^2(\psi) - 2\rho(\psi)]$. The conditional variance formula for $\bar{Y}$ gives $4E[\bar{\sigma}^2(\psi)] = E[v_1^2(\psi) + v_0^2(\psi) + 2\rho(\psi)]$. Adding these identities cancels $\rho(\psi)$ and gives $\mathrm{Var}(\tau) + 4E[\bar{\sigma}^2(\psi)] = V^*(P)$. Substituting Equation (B.2) and this identity into Equation (B.1) proves Equation (2.4). □

B.2 Matching as a Conservative Design

We first record the exact form of the excess variance under matched-pairs randomization.

Lemma B.1 (Matched-Pairs Excess Variance)**.** *Let $n \geq 2$ be even and let M be a matching with pairs (i_p, j_p) for $p \in [n/2]$. Then*

$$\mathcal{V}_n(M, P) = \frac{4}{n}E\left[\sum_{p=1}^{n/2}\Big(m(\psi_{i_p}) - m(\psi_{j_p})\Big)^2\right]. \tag{B.3}$$

Proof. By Jensen's inequality and Assumption 2.1, $E[m(\psi)^2] \leq E[\bar{Y}^2] < \infty$, so all second moments below are finite. Let $U \perp\!\!\!\perp W_{1:n}$ denote exogenous randomness that may be used to break ties in the pairing and set $\mathcal{F}_n = \sigma(\psi_{1:n}, U)$. Choose an $\mathcal{F}_n$-measurable enumeration $(i_p, j_p)_{p=1}^{n/2}$ of the pairs. Then the within-pair treatments are $Z_{i_p} = -Z_{j_p} = T_p$ for iid Rademacher pair orientations $(T_p)_{p=1}^{n/2} \perp\!\!\!\perp \mathcal{W}_n$. Set within-pair gap $d_p = m(\psi_{i_p}) - m(\psi_{j_p})$. Then $\sum_{i=1}^n Z_i m(\psi_i) = \sum_{p=1}^{n/2} T_p d_p$, and we may calculate

$$E\left[\left(\sum_{i=1}^n Z_i m(\psi_i)\right)^2 \middle| \mathcal{F}_n\right] = E\left[\left(\sum_{p=1}^{n/2} T_p d_p\right)^2 \middle| \mathcal{F}_n\right] = \sum_{p,q=1}^{n/2} d_p d_q E[T_p T_q | \mathcal{F}_n] = \sum_{p=1}^{n/2} d_p^2.$$

Taking expectations and applying Proposition 2.3 proves the result. □

B.2.1 A Lower Bound for Matching

The next lemma gives a geometric lower bound that holds for every pairing rule.

Lemma B.2 (Uniform Pair-Distance Bound)**.** *Let $n \geq 2$ be even, let $\psi_1, \ldots, \psi_n$ be iid uniform on $[0,1]^d$, let $U \perp\!\!\!\perp \psi_{1:n}$, and let $(i_p, j_p)_{p=1}^{n/2}$ be any perfect pairing measurable with respect to $\sigma(\psi_{1:n}, U)$. If ω_d is the volume of the unit Euclidean ball in $\mathbb{R}^d$, then*

$$E\left[\frac{2}{n}\sum_{p=1}^{n/2} \|\psi_{i_p} - \psi_{j_p}\|_2^2\right] \geq \frac{1}{2}\left(2\omega_d n\right)^{-2/d}. \tag{B.4}$$

Proof. For each unit, let $N_i = \min_{j \neq i} \|\psi_i - \psi_j\|_2$ denote its nearest-neighbor distance. For every realization of $\psi_{1:n}$ and every pairing, a pair $\{i, j\}$ satisfies $\|\psi_i - \psi_j\|_2^2 \geq \max\{N_i^2, N_j^2\} \geq (N_i^2 + N_j^2)/2$. Since $\sqcup_{p=1}^{n/2}\{i_p, j_p\} = [n]$, summing this inequality over the pairs gives

$$\sum_{p=1}^{n/2} \|\psi_{i_p} - \psi_{j_p}\|_2^2 \geq \sum_{p=1}^{n/2} \frac{N_{i_p}^2 + N_{j_p}^2}{2} = \frac{1}{2}\sum_{i=1}^n N_i^2.$$

Let $B(x, r)$ denote the Euclidean ball of radius r centered at x. Conditional on $\psi_i = x$, the remaining covariates are independent and uniform on $[0,1]^d$. Hence, for every $r > 0$, the union bound gives

$$\begin{aligned} P(N_i \leq r | \psi_i = x) &\leq \sum_{j \neq i} P\big(\psi_j \in B(x, r) \cap [0,1]^d | \psi_i = x\big) \\ &= (n-1)\mathrm{vol}\big(B(x,r) \cap [0,1]^d\big) \leq (n-1)\omega_d r^d < n\omega_d r^d. \end{aligned}$$

Set $r_0 = (2\omega_d n)^{-1/d}$. The preceding inequality gives $P(N_i > r_0) \geq 1/2$. Since $N_i^2 \geq$

$r_0^2 \mathbb{1}\{N_i > r_0\}$, it follows that $E[N_i^2] \geq r_0^2 P(N_i > r_0) \geq r_0^2/2$. Taking expectations in the pathwise bound and using identical distributions gives

$$E\left[\frac{2}{n}\sum_{p=1}^{n/2} \|\psi_{i_p} - \psi_{j_p}\|_2^2\right] \geq \frac{1}{n}\sum_{i=1}^{n} E[N_i^2] \geq \frac{r_0^2}{2}.$$

Substituting the definition of r_0 proves Equation (B.4). $\square$

Proof of Theorem 3.1. Fix $M \in \mathcal{M}_n$. Let $U \perp\!\!\!\perp W_{1:n}$ be its exogenous tie-breaking seed and set $\mathcal{F}_n = \sigma(\psi_{1:n}, U)$, so the resulting pairs are $\mathcal{F}_n$-measurable. Draw linear coefficient $\gamma \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ with $\gamma \perp\!\!\!\perp (\psi_{1:n}, U)$. For each γ, let P_γ be the data law under which $\psi \sim \mathrm{Unif}([0,1]^d)$ and $Y(1) = Y(0) = m_\gamma(\psi) = \gamma'(\psi - \mathbb{1}_d/2)$. Since $\|\gamma\|_2 = 1$, we have $P_\gamma \in \mathcal{P}_{\text{lin}}$. Under P_γ, $\tau = 0$ and $v_a^2(\psi) = 0$ for $a \in \{0,1\}$, so $V^*(P_\gamma) = \mathrm{Var}\big(\tau(\psi)\big) + 2E[v_1^2(\psi)] + 2E[v_0^2(\psi)] = 0$.

Write $A = E_\gamma[\gamma\gamma']$. The standard isotropy calculation for the uniform law on $\mathbb{S}^{d-1}$ gives $A = d^{-1}I_d$, and hence $E_\gamma[(\gamma' u)^2] = u'Au = d^{-1}\|u\|_2^2$ for every $u \in \mathbb{R}^d$.

Set $\Delta_p = \psi_{i_p} - \psi_{j_p}$. Since $P_\gamma \in \mathcal{P}_{\text{lin}}$ for every $\gamma \in \mathbb{S}^{d-1}$,

$$\begin{aligned}\sup_{P \in \mathcal{P}_{\text{lin}}} \mathcal{V}_n(M, P) \geq E_\gamma[\mathcal{V}_n(M, P_\gamma)] &= \frac{4}{n} E\left[\sum_{p=1}^{n/2} E_\gamma\big[(\gamma'\Delta_p)^2 | \mathcal{F}_n\big]\right] \\ &= \frac{4}{nd} E\left[\sum_{p=1}^{n/2} \|\Delta_p\|_2^2\right] \geq \frac{1}{d}\big(2\omega_d\big)^{-2/d} n^{-2/d}. \end{aligned} \tag{B.5}$$

The first inequality uses $P_\gamma \in \mathcal{P}_{\text{lin}}$ for every $\gamma \in \mathbb{S}^{d-1}$. Lemma B.1 and Tonelli's theorem give the first equality. Independence of γ from $\mathcal{F}_n$ and the preceding isotropy identity give the equality on the second line. The final inequality is Lemma B.2.

It remains lower bound $d^{-1}(2\omega_d)^{-2/d}$ uniformly in d. For $d = 1$, $\omega_1 = 2$, so this term equals $1/16 > 1/(2\pi e)$. For $d \geq 2$, let $r = d/2$. Using $\omega_d = \pi^{d/2}/\Gamma(d/2+1)$ gives $d^{-1}(2\omega_d)^{-2/d} = (2r)^{-1}\Gamma(r+1)^{1/r}\pi^{-1}2^{-1/r}$. Since $r \geq 1$, the Stirling lower bound gives $\Gamma(r+1) \geq \sqrt{2\pi r}(r/e)^r \geq 2(r/e)^r$, so that $\Gamma(r+1)^{1/r} \geq 2^{1/r} r/e$. Substitution into the preceding expression gives $(2r)^{-1}\Gamma(r+1)^{1/r}\pi^{-1}2^{-1/r} \geq (2r)^{-1}(2^{1/r}r/e)\pi^{-1}2^{-1/r} = 1/(2\pi e)$. The bound holds for the arbitrary design M fixed at the start of the proof. Taking the infimum over $M \in \mathcal{M}_n$ completes the proof. $\square$

Proof of Corollary 3.2. Let $R_n \equiv \inf_{M \in \mathcal{M}_n} \sup_{P \in \mathcal{P}_{\text{lin}}} \mathcal{V}_n(M, P)$. By Theorem 3.1,

$R_n \geq (2\pi e)^{-1} n^{-2/d_n}$. If $d_n \geq c \log n$, then $n^{-2/d_n} = \exp(-2\log n/d_n) \geq e^{-2/c}$. This proves Equation (3.4). $\square$

B.2.2 Matching over Rough Outcome Classes

The following geometric lemma turns 2-swap stability into a bound on the total within-pair distance.

Lemma B.3 (Pair Distances). *Suppose $n \geq 2$ is even, $d \geq 3$, $0 < \beta \leq 1$, and Assumption 3.3 holds. For any $\psi_{1:n} \subseteq [0,1]^d$ and every 2-swap-stable pairing,*

$$\frac{4}{n} \sum_{p=1}^{n/2} \|\psi_{i_p} - \psi_{j_p}\|_2^{2\beta} \leq 24\big(44d\Lambda^2\big)^{2\beta} n^{-2\beta/d}. \tag{B.6}$$

Proof. For each pair p, write $L_p \equiv \|\psi_{i_p} - \psi_{j_p}\|_2$. For $t > 0$, let $N(t) \equiv |\{p : L_p \geq t\}|$. We first show

$$N(t) \leq 2\big(44d\Lambda^2\big)^d t^{-d} \quad \text{for every } t > 0. \tag{B.7}$$

Every pair distance is at most $\sqrt{d}$, so $N(t) = 0$ and the claim is immediate when $t > \sqrt{d}$. Then fix $t \leq \sqrt{d}$ and partition the pairs with distance at least t into the dyadic bands $\mathcal{B}_j \equiv \{p : 2^j t \leq \|\psi_{i_p} - \psi_{j_p}\|_2 < 2^{j+1} t\}$ for $j \geq 0$.

Fix a nonempty band, write $r = 2^j t$, and set $h = r/(2\sqrt{d}\Lambda)$. The band's nonemptiness and the diameter of $[0,1]^d$ give $r \leq \sqrt{d}$, so $h \leq 1/2$ because $\Lambda \geq 1$. For $v \in \mathbb{Z}^d$, let cell $C_v \equiv h(v + [0,1)^d)$. These half-open cells form a pairwise disjoint partition of $\mathbb{R}^d$. Let $\mathcal{V}_h \equiv \{v : C_v \cap [0,1]^d \neq \varnothing\}$. Let $v(i) \in \mathcal{V}_h$ be the unique index satisfying $\psi_i \in C_{v(i)}$. Define G_j to be the multigraph with vertex set $\mathcal{V}_h$ and one edge $e_p \equiv \{v(i_p), v(j_p)\}$ for each pair $p \in \mathcal{B}_j$. We will show that G_j is simple and has maximum degree at most $D_d \equiv (11\sqrt{d}\Lambda)^d$. The edges of G_j are in one-to-one correspondence with pairs $p \in \mathcal{B}_j$, so the degree-sum identity will then give $|\mathcal{B}_j| \leq |\mathcal{V}_h| D_d/2$.

We begin with simplicity of G_j. Each cell has diameter $h\sqrt{d} = r/(2\Lambda) < r$. If $p \in \mathcal{B}_j$, then $\|\psi_{i_p} - \psi_{j_p}\|_2 \geq r$, so $v(i_p) \neq v(j_p)$ by this cell-diameter bound. Thus G_j has no loops. Next, suppose for contradiction that two edges join the same vertices, $e_p = e_q = \{u, v\}$ for distinct $u, v \in \mathcal{V}_h$. Relabeling endpoints if necessary, $\psi_{i_p}, \psi_{i_q} \in C_u$ and $\psi_{j_p}, \psi_{j_q} \in C_v$. The alternative pairs $\{i_p, i_q\}$ and $\{j_p, j_q\}$ have lengths at most $r/(2\Lambda)$. By Assumption 3.3, their combined cost satisfies $c(\psi_{i_p}, \psi_{i_q}) + c(\psi_{j_p}, \psi_{j_q}) \leq 2\overline{\lambda}(r/(2\Lambda))^q = 2^{1-q}\underline{\lambda} r^q < 2\underline{\lambda} r^q$, where the equality uses $\Lambda^q = \overline{\lambda}/\underline{\lambda}$. By Assumption 3.3

again, the original pairs satisfy $c(\psi_{i_p}, \psi_{j_p}) + c(\psi_{i_q}, \psi_{j_q}) \geq 2\underline{\lambda} r^q$. This 2-swap would then strictly reduce their combined cost, contradicting (3.6). Thus distinct pairs in $\mathcal{B}_j$ induce distinct edges, so G_j is simple.

We next bound the maximum degree of G_j. Suppose an edge joins C_u and C_v, and write its endpoints as $x \in C_u$ and $y \in C_v$. Its length is less than $2r$ by the definition of $\mathcal{B}_j$. Because $x_\ell \in [hu_\ell, h(u_\ell+1))$ and $y_\ell \in [hv_\ell, h(v_\ell+1))$ for every index $\ell \in [d]$, we have $2r > \|x-y\|_2 \geq |x_\ell - y_\ell| \geq h(|u_\ell - v_\ell| - 1)$. Thus $|u_\ell - v_\ell| < 2r/h + 1 = 4\sqrt{d}\Lambda + 1$. By this bound, for fixed u_ℓ the number of possible integer values of v_ℓ is at most $2(4\sqrt{d}\Lambda + 1) + 1 = 8\sqrt{d}\Lambda + 3 \leq 11\sqrt{d}\Lambda$. For fixed $u \in \mathcal{V}_h$, the number of vertices $v \in \mathcal{V}_h$ satisfying these d coordinatewise restrictions is therefore at most $\Pi_{\ell=1}^d (11\sqrt{d}\Lambda) = (11\sqrt{d}\Lambda)^d = D_d$. G_j is simple, so at most one edge joins u to each such v. Then G_j has maximum degree at most D_d.

It remains to count the graph's vertices. If $v \in \mathcal{V}_h$, then for each coordinate ℓ the interval $[hv_\ell, h(v_\ell+1))$ must intersect $[0,1]$ for $v_\ell \in \mathbb{Z}$. This implies $0 \leq v_\ell \leq \lfloor h^{-1} \rfloor$, and hence $|\mathcal{V}_h| \leq (\lfloor h^{-1} \rfloor + 1)^d \leq (h^{-1}+1)^d \leq (2/h)^d$, since $h \leq 1/2$. Then the degree-sum identity shows the number of edges $|\mathcal{B}_j|$ of G_j is

$$|\mathcal{B}_j| = \frac{1}{2} \sum_{u \in \mathcal{V}_h} \deg_j(u) \leq \frac{1}{2} |\mathcal{V}_h| D_d \leq \frac{1}{2} \left(\frac{2}{h}\right)^d \left(11\sqrt{d}\Lambda\right)^d \leq \left(44 d \Lambda^2\right)^d (2^j t)^{-d}. \tag{B.8}$$

The bands partition the pairs counted by $N(t)$, so summing Equation (B.8) over $j \geq 0$ proves Equation (B.7) because $\sum_{j \geq 0} 2^{-jd} = (1-2^{-d})^{-1} \leq 2$. Set $C \equiv 2(44 d\Lambda^2)^d$, so Equation (B.7) becomes $N(t) \leq Ct^{-d}$ for every $t > 0$. Since $L_p \leq \sqrt{d}$ for every pair,

$$\begin{aligned}
\sum_{p=1}^{n/2} L_p^{2\beta} &= 2\beta \sum_{p=1}^{n/2} \int_0^{\sqrt{d}} t^{2\beta-1} \mathbb{1}\{L_p \geq t\} dt \\
&= 2\beta \int_0^{\sqrt{d}} t^{2\beta-1} \sum_{p=1}^{n/2} \mathbb{1}\{L_p \geq t\} dt = 2\beta \int_0^{\sqrt{d}} t^{2\beta-1} N(t) dt.
\end{aligned}$$

The bound $N(t) \leq Ct^{-d}$ diverges as $t \to 0$, whereas the trivial count gives $N(t) \leq n/2$ for any t. Let $t_0 \equiv (2C/n)^{1/d}$ be the unique point at which $Ct_0^{-d} = n/2$. Then $n/2$ is the sharper bound for $t \leq t_0$, while Ct^{-d} is the sharper bound for $t \geq t_0$. Suppose first that $t_0 \le \sqrt{d}$. Because $d \ge 3$ and $\beta \le 1$, we have $2\beta < d$. Splitting the integral,

$$\sum_{p=1}^{n/2} L_p^{2\beta} \le 2\beta \int_0^{t_0} t^{2\beta-1} \frac{n}{2} dt + 2\beta C \int_{t_0}^{\sqrt{d}} t^{2\beta-d-1} dt$$
$$\le \frac{n}{2} t_0^{2\beta} + \frac{2\beta}{d-2\beta} C t_0^{2\beta-d} \le \frac{3n}{2} t_0^{2\beta}. \tag{B.9}$$

The final inequality uses $2\beta/(d-2\beta) \le 2$ and $Ct_0^{-d} = n/2$. If $t_0 > \sqrt{d}$, then $L_p < t_0$ for every pair, so the final bound in Equation (B.9) continues to hold. Consequently,

$$\frac{4}{n} \sum_{p=1}^{n/2} \|\psi_{i_p} - \psi_{j_p}\|_2^{2\beta} \le 6 t_0^{2\beta} = 6 \cdot 4^{2\beta/d} \big(44 d \Lambda^2\big)^{2\beta} n^{-2\beta/d} \le 24 \big(44 d \Lambda^2\big)^{2\beta} n^{-2\beta/d}.$$

The final inequality follows from $2\beta/d \le 2/3 < 1$. $\square$

Proof of Theorem 3.4. Fix $P \in \mathcal{P}_\beta$. Then the excess variance $\mathcal{V}_n(M_c, P)$ is equal to

$$\frac{4}{n} E\left[\sum_{p=1}^{n/2} \big(m(\psi_{i_p}) - m(\psi_{j_p})\big)^2\right] \le \frac{4}{n} E\left[\sum_{p=1}^{n/2} \|\psi_{i_p} - \psi_{j_p}\|_2^{2\beta}\right] \le 24(44 d \Lambda^2)^{2\beta} n^{-2\beta/d}.$$

The first expression follows from Lemma B.1, and the first inequality from the Hölder condition $[m]_\beta \le 1$. The second inequality follows from Lemma B.3. The right hand side is independent of P. Taking the supremum over $P \in \mathcal{P}_\beta$ proves the result. $\square$

The all-design lower bound uses signed functions that are constant on the interior of small cells and vanish near their boundaries.

Lemma B.4 (Signed Tabletop Functions)**.** *Fix $0 < \beta \le 1$, integers $d, q \ge 1$, and $h = q^{-1}$. Partition $[0,1]^d$ into q^d cells C of side length h. There exist functions $F_C : [0,1]^d \to [0,1]$ and tabletop subsets $B_C \subseteq C$ such that* $\operatorname{supp}(F_C) \subseteq C^\circ$ *and $F_C|_{B_C} = 1$. Then $|B_C| \ge |C|/2 = 1/(2q^d)$. For $(\omega_C)_C \in \{-1,1\}^{q^d}$, $[m_\omega]_\beta \le 1$ for*

$$m_\omega(x) \equiv a \sum_C \omega_C F_C(x), \qquad a \equiv 2^{-1-2\beta} h^\beta d^{-\beta}. \tag{B.10}$$

Proof. For a cell C, define $\delta_C(x) \equiv \inf_{y \in [0,1]^d \setminus C^\circ} \|x-y\|_2$. Then $\delta_C(x)$ is the distance to the nearest face of C when $x \in C^\circ$, and $\delta_C(x) = 0$ otherwise. Let $T(t) \equiv ((t-1) \vee 0) \wedge 1$. Define $F_C(x) = T(\delta_C(x)/b)$ for buffer $b = h/(8d)$. Then $F_C(x) = 0$ when $\delta_C(x) \le b$,

$F_C(x) = (\delta_C(x) - b)/b$ when $b < \delta_C(x) < 2b$, and $F_C(x) = 1$, the "tabletop", when $\delta_C(x) \geq 2b$. In particular, F_C takes values in $[0, 1]$ and is supported in C°.

The triangle inequality gives $|\delta_C(x) - \delta_C(y)| \leq \|x - y\|_2$. Each of the maps $t \mapsto t - 1$, $t \mapsto t \vee 0$, and $t \mapsto t \wedge 1$ is one-Lipschitz, so T is one-Lipschitz as their composition. Therefore, $|F_C(x) - F_C(y)| \leq b^{-1}|\delta_C(x) - \delta_C(y)| \leq b^{-1}\|x - y\|_2$.

For a sign vector ω, write $F_\omega \equiv \sum_C \omega_C F_C$. We next show that F_ω is b^{-1} Lipschitz on the full cube. If $x, y \in C$, then $F_\omega(x) = \omega_C F_C(x)$ and $F_\omega(y) = \omega_C F_C(y)$, so the previous bound proves the claim. Now let $x \in C$ and $y \in C'$ for distinct cells C and C'. Parameterize $\gamma(t) = (1 - t)x + ty$ for $t \in [0, 1]$. Set $t_x = 0$ when $x \in \partial C$, and otherwise set $t_x = \sup\{t \in [0, 1] : \gamma(t) \in C^\circ\}$. Set $t_y = 1$ when $y \in \partial C'$, and otherwise set $t_y = \inf\{t \in [0, 1] : \gamma(t) \in (C')^\circ\}$. Because γ is affine and the cell interiors are convex, the two preimages are intervals. Their disjointness gives $0 \leq t_x \leq t_y \leq 1$, while continuity gives $z_x = \gamma(t_x) \in \partial C$ and $z_y = \gamma(t_y) \in \partial C'$. Since every F_C is supported in a cell interior, $F_\omega(z_x) = F_\omega(z_y) = 0$. Moreover x and z_x both lie in C, while z_y and y both lie in C', so the same-cell bound above applies to each of these two pairs. Then calculate

$$
\begin{aligned}
&|F_\omega(x) - F_\omega(y)| \leq |F_\omega(x) - F_\omega(z_x)| + |F_\omega(z_y) - F_\omega(y)| \\
&\leq b^{-1}\big(\|x - z_x\|_2 + \|z_y - y\|_2\big) = b^{-1}(t_x + 1 - t_y)\|x - y\|_2 \leq b^{-1}\|x - y\|_2.
\end{aligned}
$$

This proves the claim. By disjoint support, $\|F_\omega\|_\infty \leq 1$, so $\|m_\omega\|_\infty \leq a$ and $|m_\omega(x) - m_\omega(y)| \leq 2a$. The preceding Lipschitz bound is $|m_\omega(x) - m_\omega(y)| \leq (a/b)\|x - y\|_2$. Interpolating between these two bounds using $\min\{u, v\} \leq u^{1-\beta}v^\beta$ for $u, v \geq 0$,

$$
|m_\omega(x) - m_\omega(y)| \leq \min\left\{2a, \frac{a}{b}\|x - y\|_2\right\} \leq (2a)^{1-\beta}\left(\frac{a}{b}\|x - y\|_2\right)^\beta = \|x - y\|_2^\beta.
$$

For the final equality, note that $b = h/(8d)$ and Equation (B.10) give $a = 2^{\beta-1}b^\beta$, so $(2a)^{1-\beta}(a/b)^\beta = 2^{1-\beta}ab^{-\beta} = 2^{1-\beta}2^{\beta-1} = 1$. Hence $[m_\omega]_\beta \leq 1$.

Let $B_C \equiv \{x \in C : \delta_C(x) \geq 2b\}$, so $F_C = 1$ on B_C. Write $C = \prod_{\ell=1}^d [c_\ell, c_\ell + h]$ and set $C^- = \prod_{\ell=1}^d [c_\ell + 2b, c_\ell + h - 2b]$. If $x \in C^-$ and $y \in [0, 1]^d \setminus C^\circ$, then some ℓ satisfies $y_\ell \leq c_\ell$ or $y_\ell \geq c_\ell + h$, and hence $\|x - y\|_2 \geq |x_\ell - y_\ell| \geq 2b$. Taking the infimum over y gives $\delta_C(x) \geq 2b$, so $C^- \subseteq B_C$ and

$$
|B_C| \geq (h - 4b)^d = h^d\left(1 - \frac{1}{2d}\right)^d \geq \frac{h^d}{2} = \frac{|C|}{2}.
$$

The final inequality is Bernoulli's inequality. $\square$

Proof of Theorem 3.5. Let $(F_C, B_C)_C$ be the functions and sets from Lemma B.4 for cells C of side length $h = Q^{-1}$, where $Q = \lceil n^{1/d} \rceil$. The number of cells is $K = Q^d$. For each sign vector $\omega = (\omega_C)_C \in \{-1, 1\}^K$, define a law P_ω by $\psi \sim \mathrm{Unif}([0,1]^d)$ and $Y(1) = Y(0) = m_\omega(\psi)$, as in Equation (B.10). Then P_ω satisfies Assumption 2.1, and Lemma B.4 gives $[m_\omega]_\beta \le 1$. Thus $P_\omega \in \mathcal{P}_\beta$. Moreover, $\tau(\psi) = 0$ and $v_1^2(\psi) = v_0^2(\psi) = 0$, so Equation (2.2) gives $V^*(P_\omega) = 0$.

Now fix $\sigma \in \mathcal{D}_n$. Let E_ψ denote expectation under $\psi_i \overset{\text{iid}}{\sim} \mathrm{Unif}([0,1]^d)$, and let $E_{\psi,\sigma}$ denote expectation under the joint law of $(\psi_{1:n}, Z)$. For each cell, define imbalance $G_C = \sum_{i=1}^n Z_i F_C(\psi_i)$. Let ω be distributed uniformly on $\{-1,1\}^K$ independently of $(\psi_{1:n}, Z)$, and let E_ω denote expectation over ω. Then

$$
\begin{aligned}
\sup_{P \in \mathcal{P}_\beta} \mathcal{V}_n(\sigma, P) \ge E_\omega[\mathcal{V}_n(\sigma, P_\omega)] &= \frac{4}{n} E_{\psi,\sigma} E_\omega \left[\left(\sum_{i=1}^n Z_i m_\omega(\psi_i) \right)^2 \right] \\
&= \frac{4a^2}{n} E_{\psi,\sigma} E_\omega \left[\left(\sum_C \omega_C G_C \right)^2 \right] = \frac{4a^2}{n} E_{\psi,\sigma} \left[\sum_C G_C^2 \right].
\end{aligned} \tag{B.11}
$$

The inequality follows since $P_\omega \in \mathcal{P}_\beta$ for every ω. The first equality follows from Proposition 2.3 and Tonelli's theorem applied to the independent law of ω and the joint law of $(\psi_{1:n}, Z)$. The second equality uses Equation (B.10) and the definition of G_C. The final equality follows from $E_\omega[\omega_C \omega_{C'}] = \mathbb{1}\{C = C'\}$.

For each cell, set $I_C = \{i \in [n] : \psi_i \in C\}$ and let $\mathcal{S} = \{C : I_C = \{i\}$ for some $i \in [n]$ with $\psi_i \in B_C\}$ be the set of cells containing exactly one sampled unit, with that unit on the tabletop B_C. If $C \in \mathcal{S}$ and $I_C = \{i\}$, then $F_C(\psi_j) = 0$ for $j \ne i$ and $F_C(\psi_i) = 1$. Therefore, $G_C^2 = (Z_i F_C(\psi_i))^2 = Z_i^2 F_C(\psi_i)^2 = 1$, so $\sum_C G_C^2 \ge |\mathcal{S}|$ for every realization of Z. Taking expectations, since $\mathcal{S}$ is measurable with respect to $\psi_{1:n}$, we have $E_{\psi,\sigma}[\sum_C G_C^2] \ge E_{\psi,\sigma}[|\mathcal{S}|] = E_\psi[|\mathcal{S}|]$. It remains to lower bound $E_\psi[|\mathcal{S}|]$.

$$
\begin{aligned}
P(C \in \mathcal{S}) &= \sum_{i=1}^n P\big(\psi_i \in B_C,\ \psi_j \notin C \text{ for every } j \ne i\big) \\
&= \sum_{i=1}^n |B_C| \left(1 - \frac{1}{K}\right)^{n-1} = n|B_C| \left(1 - \frac{1}{K}\right)^{n-1} \ge \frac{n}{2K} \left(1 - \frac{1}{K}\right)^{n-1}.
\end{aligned}
$$

The first equality partitions the event $\{C \in \mathcal{S}\}$ according to the unique sampled unit in C. The second equality follows by uniformity and independence, since $|C| = 1/K$.

The final inequality follows from Lemma B.4, which gives $|B_C| \geq 1/(2K)$. Therefore,

$$E_\psi[|\mathcal{S}|] = E_\psi\left[\sum_C \mathbb{1}\{C \in \mathcal{S}\}\right] = \sum_C P(C \in \mathcal{S}) \geq \frac{n}{2}\left(1 - \frac{1}{K}\right)^{n-1} \geq \frac{n}{2e}.$$

The first inequality sums the preceding bound over the K total cells. For the final inequality, note that $K \geq n$, so $(1 - 1/K)^{n-1} \geq (1 - 1/K)^{K-1}$. Then calculate $(1 - 1/K)^{K-1} = ((K-1)/K)^{K-1} = (1 + 1/(K-1))^{-(K-1)} \geq e^{-1}$, where the final inequality uses $(1 + 1/m)^m \leq e$. Equation (B.11) and the work above give

$$\sup_{P \in \mathcal{P}_\beta} \mathcal{V}_n(\sigma, P) \geq \frac{4a^2}{n} E_\psi[|\mathcal{S}|] \geq \frac{2a^2}{e} = \frac{2^{-1-4\beta}}{e} h^{2\beta} d^{-2\beta} \geq \frac{2^{-1-6\beta}}{e} d^{-2\beta} n^{-2\beta/d}.$$

The equality is from the definition of a in Equation (B.10). For the final inequality, $Q = \lceil n^{1/d} \rceil \leq 2n^{1/d}$, so $h = Q^{-1} \geq 2^{-1} n^{-1/d}$. Because $\sigma \in \mathcal{D}_n$ was arbitrary, taking the infimum over σ proves the theorem with constant $2^{-1-6\beta} d^{-2\beta}/e$. □

B.3 Proofs for Section 4

Recall $\mathcal{F}_n = \sigma(\psi_{1:n})$. For a continuous positive semidefinite kernel W satisfying $\sup_\psi W(\psi, \psi) \leq 1$, write $W_n = [W(\psi_i, \psi_j)]_{i,j=1}^n$ and $K_n = [K(\psi_i, \psi_j)]_{i,j=1}^n$ for $K = 1 + W$, and set $\Gamma_n \equiv \varphi I_n + (1-\varphi) K_n/2$. Then $\Gamma_n \succeq \varphi I_n$ and $\Gamma_{n,ii} = \varphi + (1-\varphi) K(\psi_i, \psi_i)/2 \leq 1$.

Let $B_n = \Gamma_n^{1/2}$ and let $b_{n,i}$ be its ith column. Since $B_n' B_n = \Gamma_n$, we have $\|b_{n,i}\|_2^2 = e_i' B_n' B_n e_i = \Gamma_{n,ii} \leq 1$. Theorem 6.3 of Harshaw et al. (2024) therefore gives $\mathrm{Cov}(B_n Z | \mathcal{F}_n) \preceq B_n \Gamma_n^{-1} B_n' = I_n$. Moreover, $\Gamma_n \succeq \varphi I_n$ makes B_n invertible, so premultiplying and postmultiplying by B_n^{-1} gives $\mathrm{Cov}(Z|\mathcal{F}_n) \preceq \Gamma_n^{-1}$. Lemma 6.1 of Harshaw et al. (2024) also gives $E[Z|\mathcal{F}_n] = 0$, and hence

$$E[Z|\mathcal{F}_n] = 0, \qquad \mathrm{Cov}(Z|\mathcal{F}_n) \preceq \Gamma_n^{-1}. \tag{B.12}$$

The next lemma records the algebraic form of the covariance bound.

Lemma B.5 (GSW Ridge Bound)**.** *Let $\mathcal{H}$ be a real Hilbert space and let $\mathcal{A}$ be a sigma-algebra. Let $x_1, \ldots, x_N \in \mathcal{H}$, $a, b > 0$, and $y \in \mathbb{R}^N$ be $\mathcal{A}$-measurable, and define $F : \mathcal{H} \to \mathbb{R}^N$ by $(Fh)_i = \langle h, x_i \rangle_{\mathcal{H}}$. Suppose a random vector $S \in \mathbb{R}^N$ satisfies*

$E[S|\mathcal{A}] = 0$ *and* $\mathrm{Cov}(S|\mathcal{A}) \preceq (aI_N + bFF^*)^{-1}$. *Then*

$$E[(S'y)^2|\mathcal{A}] \leq \min_{h\in\mathcal{H}} \left\{ \frac{1}{a}\|y - Fh\|_2^2 + \frac{1}{b}\|h\|_{\mathcal{H}}^2 \right\}. \tag{B.13}$$

Proof. The argument is pointwise in $\mathcal{A}$. Fix a realization, so a, b, F, and y are deterministic. Define $L(h) \equiv a^{-1}\|y - Fh\|_2^2 + b^{-1}\|h\|_{\mathcal{H}}^2$. For any $\delta \in \mathcal{H}$, its directional derivative is $(d/dt)L(h + t\delta)|_{t=0} = -2a^{-1}\langle y - Fh, F\delta\rangle + 2b^{-1}\langle h, \delta\rangle_{\mathcal{H}}$. A second calculation gives the second directional derivative $2a^{-1}\|F\delta\|_2^2 + 2b^{-1}\|\delta\|_{\mathcal{H}}^2 > 0$ for $\delta \neq 0$, so L is strictly convex. Writing $\langle y - Fh, F\delta\rangle = \langle F^*(y - Fh), \delta\rangle_{\mathcal{H}}$ and setting the derivative to zero for every δ gives $bF^*(y - Fh) = ah$, so the first-order condition is $(aI + bF^*F)h = bF^*y$. For every $u \in \mathcal{H}$, $\langle u, (aI + bF^*F)u\rangle_{\mathcal{H}} = a\|u\|_{\mathcal{H}}^2 + b\|Fu\|_2^2 \geq a\|u\|_{\mathcal{H}}^2$, so $aI + bF^*F$ is invertible. The first-order condition therefore has the unique solution $h_* = b(aI + bF^*F)^{-1}F^*y$, which is the unique minimizer of L.

To express the solution using the inverse in the covariance bound, write $R = (aI_N + bFF^*)^{-1}$. The push-through identity $(aI + bF^*F)^{-1}F^* = F^*R$ follows from $(aI + bF^*F)F^* = F^*(aI_N + bFF^*)$. Hence $h_* = bF^*Ry$. Moreover, $(aI_N + bFF^*)Ry = y$, so $y = aRy + bFF^*Ry$ and therefore $y - Fh_* = y - bFF^*Ry = aRy$. Then

$$\begin{aligned} L(h_*) &= \frac{1}{a}\|aRy\|_2^2 + \frac{1}{b}\|bF^*Ry\|_{\mathcal{H}}^2 \\ &= ay'R^2y + by'RFF^*Ry = y'R(aI_N + bFF^*)Ry = y'Ry. \end{aligned}$$

By the covariance bound and $E[S|\mathcal{A}] = 0$, $E[(S'y)^2|\mathcal{A}] \leq y'Ry = L(h_*) = \min_{h\in\mathcal{H}} L(h)$. This proves Equation (B.13). ☐

Matérn RKHS. Consider the Matérn kernel. For $d \geq 1$ and $\nu > 0$, put $r = d/2 + \nu$ and define the spectral density $\rho_{d,\nu}(\omega) = \pi^{-d/2}\Gamma(r)\Gamma(\nu)^{-1}(1 + \|\omega\|_2^2)^{-r}$. Then

$$W_d^{\mathrm{Mat},\nu}(x, x') = \int_{\mathbb{R}^d} e^{i\omega'(x-x')}\rho_{d,\nu}(\omega)d\omega. \tag{B.14}$$

This representation follows, for example, by setting $\ell = \sqrt{2\nu}$ in the Matérn spectral density of Rasmussen and Williams (2006, Equation (4.15)), substituting it into their Equation (4.6), and changing variables. A calculation shows that $W_d^{\mathrm{Mat},\nu}(x, x) = \int_{\mathbb{R}^d} \rho_{d,\nu}(\omega)d\omega = 1$. Because $\rho_{d,\nu}$ is even, Equation (B.14) also shows that the kernel is real-valued and symmetric. We next verify that this kernel is positive definite. For distinct $x_1, \ldots, x_N \in \mathbb{R}^d$ and nonzero $c \in \mathbb{R}^N$, Equation (B.14) gives

$$\sum_{i,j=1}^{N} c_i c_j W_d^{\mathrm{Mat},\nu}(x_i, x_j) = \int_{\mathbb{R}^d} \left| \sum_{i=1}^{N} c_i e^{i\omega' x_i} \right|^2 \rho_{d,\nu}(\omega) d\omega > 0. \tag{B.15}$$

The strict inequality follows because $\rho_{d,\nu}(\omega) > 0$ and the distinct complex exponentials are linearly independent. Thus, $W_d^{\mathrm{Mat},\nu}$ is positive definite.

Let $\mathcal{H}_W(\mathbb{R}^d)$ denote the RKHS generated by W on $\mathbb{R}^d$. Write $\mathcal{F}$ and $\mathcal{F}^{-1}$ for the Fourier transform and inverse Fourier transform under the convention $(\mathcal{F}h)(\omega) = \widehat{h}(\omega) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-i\omega' x} h(x) dx$. We invoke Theorem 10.12 of Wendland (2005), which characterizes the RKHS of any real-valued, symmetric, positive definite kernel $W(x, x') = \kappa(x - x')$ with $\kappa \in C(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$ through $\widehat{\kappa}$. In particular, When $\widehat{\kappa} > 0$, Wendland (2005) shows that $\|h\|^2_{\mathcal{H}_W(\mathbb{R}^d)} = (2\pi)^{-d/2} \int_{\mathbb{R}^d} |\widehat{h}(\omega)|^2 / \widehat{\kappa}(\omega) d\omega$ and

$$\mathcal{H}_W(\mathbb{R}^d) = \left\{ h \in C(\mathbb{R}^d) \cap L^2(\mathbb{R}^d) : \|h\|_{\mathcal{H}_W(\mathbb{R}^d)} < \infty \right\}. \tag{B.16}$$

The next lemma specializes this characterization to the Matérn kernel.

Lemma B.6 (Matérn RKHS Norm)**.** *Let $W = W_d^{\mathrm{Mat},\nu}$ and $\Lambda_{r,d} = (4\pi)^{-d/2}\Gamma(\nu)/\Gamma(r)$ for $r = d/2 + \nu$. Define $\Theta_{r,d} = d^r \Lambda_{r,d}$. Then $\mathcal{H}_W(\mathbb{R}^d) = H^r(\mathbb{R}^d)$. For $h \in \mathcal{H}_W(\mathbb{R}^d)$, the RKHS norm is*

$$\|h\|^2_{\mathcal{H}_W(\mathbb{R}^d)} = \Lambda_{r,d} \int_{\mathbb{R}^d} |\widehat{h}(\omega)|^2 (1 + \|\omega\|_2^2)^r d\omega. \tag{B.17}$$

Moreover, for every $\bar{\nu} > 0$, there is $C_{\bar{\nu}} < \infty$ such that, for all $d \geq 1$ and $0 < \nu \leq \bar{\nu}$, $\Theta_{d/2+\nu,d} \leq C_{\bar{\nu}}/\nu$.

Proof. We have $\kappa(u) = W(u, 0)$. Positive definiteness was established in Equation (B.15). It remains to verify that $\kappa \in C(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$ and $\widehat{\kappa} > 0$. Since $r > d/2$, $\rho_{d,\nu} \in L^1(\mathbb{R}^d)$, and Equation (B.14) gives $\kappa = \mathcal{F}^{-1}[(2\pi)^{d/2}\rho_{d,\nu}]$. Thus, $\kappa \in C(\mathbb{R}^d)$ as the inverse Fourier transform of an L^1 function. Equation (4.14) of Rasmussen and Williams (2006) with length scale $\sqrt{2\nu}$ gives $\kappa(u) = 2^{1-\nu}\Gamma(\nu)^{-1}\|u\|_2^\nu K_\nu(\|u\|_2)$, where K_ν is the modified Bessel function of the second kind. This expression is bounded near zero and decays exponentially as $\|u\|_2 \to \infty$, so $\kappa \in L^1(\mathbb{R}^d)$. Since $\rho_{d,\nu}$ is continuous, Fourier inversion gives $\widehat{\kappa} = \mathcal{F}\kappa = \mathcal{F}\left(\mathcal{F}^{-1}[(2\pi)^{d/2}\rho_{d,\nu}]\right) = (2\pi)^{d/2}\rho_{d,\nu} > 0$. Hence

Equation (B.16) applies. Substituting $\widehat{\kappa}$ into the formula above gives for $h \in \mathcal{H}_W(\mathbb{R}^d)$,

$$\|h\|^2_{\mathcal{H}_W(\mathbb{R}^d)} = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{|\widehat{h}(\omega)|^2}{\widehat{\kappa}(\omega)} d\omega = (4\pi)^{-d/2} \frac{\Gamma(\nu)}{\Gamma(r)} \int_{\mathbb{R}^d} |\widehat{h}(\omega)|^2 (1 + \|\omega\|_2^2)^r d\omega.$$

This proves Equation (B.17). It remains to identify $\mathcal{H}_W(\mathbb{R}^d) = H^r(\mathbb{R}^d)$. By Equation (4.3), $H^r(\mathbb{R}^d) = \{h \in L^2(\mathbb{R}^d) : \int_{\mathbb{R}^d} |\widehat{h}(\omega)|^2 (1 + \|\omega\|_2^2)^r d\omega < \infty\}$. Moreover, since $r > d/2$, the Sobolev embedding theorem gives $H^r(\mathbb{R}^d) \subseteq C(\mathbb{R}^d)$ (Taylor, 2011). Comparing these facts with Equation (B.16) proves the identification and, together with Equation (B.17), Proposition 4.1.

For the final statement, continuity of Γ on $[1, 1+\bar{\nu}]$ gives $M_{\bar{\nu}} \equiv \sup_{0 \le t \le \bar{\nu}} \Gamma(1+t) < \infty$, so $\Gamma(\nu) = \Gamma(1+\nu)/\nu \le M_{\bar{\nu}}/\nu$. For $d \ge 3$, monotonicity of Γ on $[3/2, \infty)$ and Stirling's lower bound give

$$\Theta_{d/2+\nu,d} \le \frac{M_{\bar{\nu}}}{\nu} d^{\bar{\nu}} \frac{(d/(4\pi))^{d/2}}{\Gamma(d/2)} \le \frac{C}{\nu} d^{\bar{\nu}+1/2} \left(\frac{e}{2\pi}\right)^{d/2} \le \frac{C_{\bar{\nu}}}{\nu}.$$

For $d = 1, 2$, $\nu \Theta_{d/2+\nu,d}$ extends continuously to $0 \le \nu \le \bar{\nu}$ and is therefore bounded. This proves the final statement. $\square$

The Fourier cutoff yields the following penalized approximation bound for near-critical Matérn kernels.

Lemma B.7 (Penalized Matérn Approximation)**.** *Fix $c, s > 0$. There are constants $c_0, C > 0$ and $n_0 \ge 3$, depending only on (c, s), such that the following holds for every integer $d \ge 1$ and $n \ge n_0$. Set $\nu_n = c/\log n$ and $W_n = W_d^{\mathrm{Mat},\nu_n}$. For every penalty $\lambda_n \in [n^{-1}, c_0/\log n]$ and $h \in H^s_{\mathrm{av}}(\mathbb{R}^d)$,*

$$\inf_{v \in \mathcal{H}_{W_n}} \left\{ \|h - v\|^2_{L^2([0,1]^d)} + \lambda_n \|v\|^2_{\mathcal{H}_{W_n}} \right\} \le C \|h\|^2_{H^s_{\mathrm{av}}(\mathbb{R}^d)} (\lambda_n \log n)^{(2s/d) \wedge 1}. \tag{B.18}$$

If X has a density on $[0,1]^d$ bounded above by $\overline{p}$, Equation B.18 holds with the first term inside the braces replaced by $E[(h(X) - v(X))^2]$, with C also depending on $\overline{p}$.

Proof. First suppose that $s \le d/2$. Recall the inverse Fourier transform $\mathcal{F}^{-1}$. For $R \ge 1$, define $\widehat{h}_R(\omega) = \widehat{h}(\omega) \mathbb{1}\{1 + \|\omega\|_2^2/d \le R\}$ and $h_R = \mathcal{F}^{-1}(\widehat{h}_R)$. Since $h \in H^s_{\mathrm{av}}(\mathbb{R}^d) \subseteq L^2(\mathbb{R}^d)$, Plancherel's theorem gives $\widehat{h} \in L^2(\mathbb{R}^d)$. Hence $\widehat{h}_R \in L^2(\mathbb{R}^d)$, and another application of Plancherel's theorem gives $h_R \in L^2(\mathbb{R}^d)$, so

$$\|h - h_R\|^2_{L^2([0,1]^d)} \le \int_{\mathbb{R}^d} |h(x) - h_R(x)|^2 dx = \int_{\mathbb{R}^d} |\widehat{h}(\omega) - \widehat{h}_R(\omega)|^2 d\omega$$
$$= \int_{1+\|\omega\|_2^2/d>R} |\widehat{h}(\omega)|^2 d\omega \le R^{-s} \int_{\mathbb{R}^d} |\widehat{h}(\omega)|^2 \left(1 + \frac{\|\omega\|_2^2}{d}\right)^s d\omega = R^{-s} \|h\|^2_{H^s_{\mathrm{av}}(\mathbb{R}^d)}.$$

The first equality is Plancherel's identity. For the second inequality, note that $(1 + \|\omega\|_2^2/d)^{-s} \le R^{-s}$ on the domain of integration. Because $\widehat{h}_R$ is supported on $\{1 + \|\omega\|_2^2/d \le R\}$, we have $h_R \in H^{d/2+\nu_n}(\mathbb{R}^d) = \mathcal{H}_{W_n}$. Write $\Lambda_n = \Lambda_{d/2+\nu_n,d}$ and $\Theta_n = \Theta_{d/2+\nu_n,d}$. Similarly, the RKHS norm of this approximation satisfies

$$\|h_R\|^2_{\mathcal{H}_{W_n}} = \Lambda_n \int_{1+\|\omega\|_2^2/d \le R} |\widehat{h}(\omega)|^2 (1 + \|\omega\|_2^2)^{d/2+\nu_n} d\omega$$
$$\le \Theta_n R^{d/2+\nu_n-s} \int_{\mathbb{R}^d} |\widehat{h}(\omega)|^2 \left(1 + \frac{\|\omega\|_2^2}{d}\right)^s d\omega = \Theta_n R^{d/2+\nu_n-s} \|h\|^2_{H^s_{\mathrm{av}}(\mathbb{R}^d)}.$$

The first equality follows from Lemma B.6, noting that fourier inversion gives $\mathcal{F}h_R = \mathcal{F}(\mathcal{F}^{-1}(\widehat{h}_R)) = \widehat{h}_R$ almost everywhere, and the definition of $\widehat{h}_R$ then restricts the integral to the displayed domain. Since $1 + \|\omega\|_2^2 \le d(1 + \|\omega\|_2^2/d)$ and $d/2 + \nu_n \ge d/2 \ge s$, we have $(1 + \|\omega\|_2^2/d)^{d/2+\nu_n-s} \le R^{d/2+\nu_n-s}$ on this domain, which gives the inequality.

Using h_R as a witness in the infimum on the left hand side of Equation (B.18) gives the upper bound $\|h\|^2_{H^s_{\mathrm{av}}(\mathbb{R}^d)}\{R^{-s} + \lambda_n \Theta_n R^{d/2+\nu_n-s}\}$. Apply the final statement of Lemma B.6 with $\bar{\nu} = c$, and denote the corresponding constant by C_c. For $n \ge 3$, $\nu_n \le c$, so $\Theta_n \le C_c/\nu_n = (C_c/c)\log n$ uniformly over $d \ge 1$. Set $y = \lambda_n \log n$, so $\log n/n \le y \le c_0 < 1$ and $\lambda_n \Theta_n \le (C_c/c)y$. Decrease c_0 if necessary so that $\lambda_n \Theta_n \le (C_c/c)c_0 \le 1$, and take $R = (\lambda_n \Theta_n)^{-1/(d/2+\nu_n)} \ge 1$. Since $s/(d/2 + \nu_n) \le 1$,

$$R^{-s} + \lambda_n \Theta_n R^{d/2+\nu_n-s} = 2(\lambda_n \Theta_n)^{s/(d/2+\nu_n)} \le C y^{s/(d/2+\nu_n)}$$
$$= C y^{2s/d} \exp\left(\left(\frac{2s}{d} - \frac{s}{d/2+\nu_n}\right) \log(1/y)\right) \le C y^{2s/d}.$$

To see the final inequality, note that

$$0 \le \frac{2s}{d} - \frac{s}{d/2+\nu_n} = \frac{4s\nu_n}{d(d+2\nu_n)} \le \frac{4sc}{\log n}, \quad \log(1/y) \le \log n.$$

Thus the exponential factor is at most e^{4sc}, which is absorbed into C. This proves Equation (B.18) when $s \leq d/2$.

Now suppose $s > d/2$. Only finitely many integer dimensions satisfy this inequality. If this set is nonempty, define $\gamma_s = \min_{d\in\mathbb{N}:1\leq d<2s}(s - d/2) > 0$. For all sufficiently large n, $\nu_n \leq \gamma_s$, so $d/2 + \nu_n \leq s$ simultaneously for these dimensions. Taking $v = h$ and using Lemma B.6 and monotonicity of the spectral weights gives zero approximation error and

$$\lambda_n \|h\|^2_{\mathcal{H}_{W_n}} \leq \lambda_n \Theta_n \|h\|^2_{H^s_{\text{av}}(\mathbb{R}^d)} \leq (C_c/c) y \|h\|^2_{H^s_{\text{av}}(\mathbb{R}^d)}.$$

This proves Equation (B.18) when $s > d/2$.

If X has density bounded above by $\overline{p}$, then $E[(h(X) - v(X))^2] \leq \overline{p}\|h - v\|^2_{L^2([0,1]^d)}$ for every v. Since $\overline{p} \geq 1$, the population criterion is at most $\overline{p}$ times the criterion on the left hand side of Equation (B.18). This proves the final statement. $\square$

Proof of Proposition 4.3. In the notation of Lemma B.5, let $\mathcal{H} = \mathcal{H}_K$, $x_i = K(\psi_i, \cdot)$, and $F = F_n$, where $F_n : \mathcal{H} \to \mathbb{R}^n$ is the evaluation operator $(F_n g)_i = g(\psi_i)$. First, we identify the adjoint $F_n^* : \mathbb{R}^n \to \mathcal{H}$. Fix $u \in \mathbb{R}^n$ and any $g \in \mathcal{H}$. By the reproducing property, $\langle g, F_n^* u\rangle_{\mathcal{H}} = \langle F_n g, u\rangle_2 = \sum_{i=1}^n u_i g(\psi_i) = \langle g, \sum_{i=1}^n u_i K(\psi_i, \cdot)\rangle_{\mathcal{H}}$. This holds for every $g \in \mathcal{H}$, so $F_n^* u = \sum_{i=1}^n u_i K(\psi_i, \cdot)$. Next, we claim that $F_n F_n^* = K_n$. Taking $u = e_j$ gives $(F_n F_n^*)_{ij} = (F_n^* e_j)(\psi_i) = K(\psi_j, \psi_i) = K(\psi_i, \psi_j)$, proving the claim.

Let $m_{1:n} = (m(\psi_1), \ldots, m(\psi_n))'$. Together with Equation (B.12), the identity $F_n F_n^* = K_n$ gives $E[Z|\mathcal{F}_n] = 0$ and $\mathrm{Cov}(Z|\mathcal{F}_n) \preceq (\varphi I_n + (1-\varphi) F_n F_n^*/2)^{-1}$. Applying Lemma B.5 with $\mathcal{A} = \mathcal{F}_n$, $S = Z$, $y = m_{1:n}$, $a = \varphi$, and $b = (1-\varphi)/2$ therefore gives, for every fixed $g \in \mathcal{H}_K$,

$$\begin{aligned} E\left[\left(\sum_{i=1}^n Z_i m(\psi_i)\right)^2 \Big| \mathcal{F}_n\right] &\leq \min_{h\in\mathcal{H}_K} \left\{\frac{1}{\varphi}\|m_{1:n} - F_n h\|_2^2 + \frac{2}{1-\varphi}\|h\|^2_{\mathcal{H}_K}\right\} \\ &\leq \frac{1}{\varphi}\sum_{i=1}^n (m(\psi_i) - g(\psi_i))^2 + \frac{2}{1-\varphi}\|g\|^2_{\mathcal{H}_K}. \end{aligned} \tag{B.19}$$

The second inequality evaluates the minimum at $h = g$. Take expectations in Equation (B.19), multiply by $4/n$, and apply Proposition 2.3. The resulting inequality holds for every fixed $g \in \mathcal{H}_K$. Taking the infimum over g proves Equation (4.7). $\square$

Proof of Corollary 4.4. Recall $m(\psi) = E[\bar{Y}|\psi]$. By conditional Jensen's inequality, tower law, and Assumption 2.1, $E[m(\psi)^2] \le E[\bar{Y}^2] \le \frac{1}{2}\sum_{a=0}^{1} E[Y(a)^2] < \infty$, so $m \in L^2(P_\psi)$. Fix $\epsilon > 0$. Density of $\mathcal{H}_K$ provides $g_\epsilon \in \mathcal{H}_K$ such that $E[(m(\psi) - g_\epsilon(\psi))^2] \le \epsilon$. Using g_ϵ as a witness in Proposition 4.3 gives $\mathcal{V}_n(P) \le 4\epsilon/\varphi + 8\|g_\epsilon\|^2_{\mathcal{H}_K}/\{(1-\varphi)n\}$. Hence $\limsup_{n\to\infty} \mathcal{V}_n(P) \le 4\epsilon/\varphi$. Since ϵ was arbitrary and $\mathcal{V}_n(P) \ge 0$ by Proposition 2.3, the conclusion follows. □

Proof of Corollary 4.5. For $P \in \mathcal{P}_K(B)$, take $g = m$ in Proposition 4.3. The approximation term vanishes and $\|m\|_{\mathcal{H}_K} \le B$, so $\mathcal{V}_n(P) \le 8B^2/\{(1-\varphi)n\}$. The right hand side is independent of P, and taking the supremum proves the result. □

Proof of Corollary 4.6. Put $r = d/2 + \nu \le s$ and fix $P \in \mathcal{P}^s_d$. Recall that $K = 1+W$. By the sums-of-kernels theorem (Paulsen and Raghupathi, 2016, Theorem 5.4), $\|f\|^2_{\mathcal{H}_K} = \min_{f=a+w}\{a^2 + \|w\|^2_{\mathcal{H}_W}\}$, where $a \in \mathbb{R}$ and $w \in \mathcal{H}_W$. Taking $a = 0$ and $w = g$ gives $\|g\|_{\mathcal{H}_K} \le \|g\|_{\mathcal{H}_W}$ for every $g \in \mathcal{H}_W$. Since $r \le s$, Lemma B.6 and the monotonicity of the spectral weights give

$$\begin{aligned}
\|m\|^2_{\mathcal{H}_W} &= \Lambda_{r,d} \int_{\mathbb{R}^d} |\widehat{m}(\omega)|^2 (1 + \|\omega\|_2^2)^r d\omega \\
&\le \Theta_{r,d} \int_{\mathbb{R}^d} |\widehat{m}(\omega)|^2 \left(1 + \frac{\|\omega\|_2^2}{d}\right)^s d\omega = \Theta_{r,d}\|m\|^2_{H^s_{\mathrm{av}}(\mathbb{R}^d)} \le \Theta_{r,d}.
\end{aligned}$$

Thus, $\|m\|^2_{\mathcal{H}_K} \le \Theta_{r,d}$. Apply Corollary 4.5 with $B^2 = \Theta_{r,d}$. The constant $8\Theta_{r,d}/(1-\varphi)$ depends only on (d, ν, φ), as claimed. □

Proof of Theorem 4.7. Put $W_n = W_d^{\mathrm{Mat},\nu_n}$ and recall that $K_n = 1 + W_n$, where $\nu_n = c/\log n$. Fix $P \in \mathcal{P}^s_d$ and set $\lambda_n = \max\{2\varphi/((1-\varphi)n), n^{-1}\}$. By the sums-of-kernels theorem (Paulsen and Raghupathi, 2016, Theorem 5.4), every $v \in \mathcal{H}_{W_n}$ belongs to $\mathcal{H}_{K_n}$ with $\|v\|_{\mathcal{H}_{K_n}} \le \|v\|_{\mathcal{H}_{W_n}}$. Proposition 4.3 therefore gives

$$\begin{aligned}
\mathcal{V}_n(P) &\le \frac{4}{\varphi} \inf_{v \in \mathcal{H}_{W_n}} \left\{ E[(m(\psi) - v(\psi))^2] + \frac{2\varphi}{(1-\varphi)n}\|v\|^2_{\mathcal{H}_{W_n}} \right\} \\
&\le \frac{4}{\varphi} \inf_{v \in \mathcal{H}_{W_n}} \left\{ E[(m(\psi) - v(\psi))^2] + \lambda_n \|v\|^2_{\mathcal{H}_{W_n}} \right\}.
\end{aligned}$$

For fixed φ, $n^{-1} \le \lambda_n \le C_\varphi/n$, so the penalty rate condition in Lemma B.7 holds for all sufficiently large n. Apply the population conclusion of that lemma in dimension

d at smoothness s, with $h = m$ and $X = \psi$. Since $\|m\|_{H^s_{\mathrm{av}}(\mathbb{R}^d)} \le 1$, we obtain

$$\mathcal{V}_n(P) \le C_{s,c,\varphi,\overline{p}}(\lambda_n \log n)^{(2s/d)\wedge 1} \le C_{s,c,\varphi,\overline{p}} \left(\frac{\log n}{n}\right)^{(2s/d)\wedge 1}.$$

Thus, there is $n_0 \ge 3$ depending only on (s, c, φ) such that the claimed bound holds for every $n \ge n_0$, uniformly over d and P. For $2 \le n < n_0$, Proposition 4.3 evaluated at $g = 0$ and the density upper bound give $\mathcal{V}_n(P) \le 4E[m(\psi)^2]/\varphi \le 4\overline{p}/\varphi$. Moreover, $(\log n/n)^{(2s/d)\wedge 1} \ge \log n/n$, whose minimum over the finite set $2 \le n < n_0$ is strictly positive. Enlarging $C_{s,c,\varphi,\overline{p}}$ therefore establishes the same bound for these remaining values of n. Taking the supremum over $P \in \mathcal{P}^s_d$ proves the theorem. $\square$

Proof of Theorem 4.9. Fix $\sigma \in \mathcal{D}_n$. Choose a smooth function $\eta : \mathbb{R} \to \mathbb{R}$ supported in a compact subinterval of $(0, 1)$ and normalized so that $\int_0^1 \eta(u)du = 0$ and $\int_0^1 \eta(u)^2 du = 1$. Set $q \equiv \lceil n^{1/d} \rceil$, $h \equiv q^{-1}$, and $K \equiv q^d$. Choose one cell C of the regular partition of $[0, 1]^d$ into K cubes of side length h, let b be its lower-left corner, and define the whole-space bump

$$g_h(x) \equiv \Pi_{\ell=1}^d \eta\left(\frac{x_\ell - b_\ell}{h}\right).$$

We claim that, for a constant $C_s < \infty$ depending only on s, $\|g_h\|^2_{H^s_{\mathrm{av}}(\mathbb{R}^d)} \le C_s h^{d-2s}$. The function g_h is smooth and supported in the interior of C. Under the Fourier convention above, $\widehat{g}_h(\omega) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-i\omega' x} \Pi_{\ell=1}^d \eta((x_\ell - b_\ell)/h)dx$. Setting $x = b + hz$,

$$\widehat{g}_h(\omega) = h^d e^{-i\omega' b}(2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-ih\omega' z} \Pi_{\ell=1}^d \eta(z_\ell)dz = h^d e^{-i\omega' b} \Pi_{\ell=1}^d \widehat{\eta}(h\omega_\ell).$$

The second equality is from $e^{-ih\omega' z} = \Pi_{\ell=1}^d e^{-ih\omega_\ell z_\ell}$ and Fubini. Since $|e^{-i\omega' b}| = 1$,

$$\begin{aligned} \|g_h\|^2_{H^s_{\mathrm{av}}(\mathbb{R}^d)} &= h^{2d} \int_{\mathbb{R}^d} \Pi_{\ell=1}^d |\widehat{\eta}(h\omega_\ell)|^2 \left(1 + \frac{\|\omega\|_2^2}{d}\right)^s d\omega \\ &= h^d \int_{\mathbb{R}^d} \Pi_{\ell=1}^d |\widehat{\eta}(u_\ell)|^2 \left(1 + \frac{h^{-2}\|u\|_2^2}{d}\right)^s du. \end{aligned} \tag{B.20}$$

The final equality uses $u = h\omega$, so $d\omega = h^{-d}du$ and $\|\omega\|_2^2 = h^{-2}\|u\|_2^2$. By Plancherel's identity, $\int_{\mathbb{R}} |\widehat{\eta}(u)|^2 du = \int_{\mathbb{R}} |\eta(u)|^2 du = \int_0^1 \eta(u)^2 du = 1$, where the second equality uses the support of η. Thus, $d\mu(u) = |\widehat{\eta}(u)|^2 du$ defines a probability measure.

Let $U_1, \dots, U_d$ be iid with law μ. Since $h \le 1$, Equation (B.20) gives

$$\|g_h\|^2_{H^s_{\mathrm{av}}(\mathbb{R}^d)} \le h^{d-2s} E\left[\left(1 + \frac{1}{d}\sum_{\ell=1}^{d} U_\ell^2\right)^s\right]. \tag{B.21}$$

Suppose first that $0 < s \le 1$. Jensen's inequality gives $E[(1 + d^{-1}\sum_{\ell=1}^d U_\ell^2)^s] \le (E[1 + d^{-1}\sum_{\ell=1}^d U_\ell^2])^s = (1 + E[U_1^2])^s$. Set $C_s = (1 + E[U_1^2])^s$. Then the display above gives $\|g_h\|^2_{H^s_{\mathrm{av}}(\mathbb{R}^d)} \le C_s h^{d-2s}$. It remains to show $E[U_1^2] < \infty$. Integration by parts gives $\widehat{\eta'}(u) = iu\widehat{\eta}(u)$, and hence $E[U_1^2] = \int_{\mathbb{R}} u^2 |\widehat{\eta}(u)|^2 du = \int_{\mathbb{R}} |\widehat{\eta'}(u)|^2 du = \|\eta'\|^2_{L^2(\mathbb{R})} < \infty$. The final equality is Plancherel's identity, and finiteness follows because η is smooth and compactly supported.

Suppose next that $s \ge 1$. By Jensen, $(1 + d^{-1}\sum_{\ell=1}^d U_\ell^2)^s = d^{-s}(\sum_{\ell=1}^d (1 + U_\ell^2))^s \le d^{-1}\sum_{\ell=1}^d (1 + U_\ell^2)^s$. Taking expectations and applying Equation (B.21) gives $\|g_h\|^2_{H^s_{\mathrm{av}}(\mathbb{R}^d)} \le E[(1 + U_1^2)^s] h^{d-2s}$. Thus the same bound holds with $C_s = E[(1 + U_1^2)^s]$, provided that $E[|U_1|^{2s}] < \infty$. To verify this, choose an integer $k \ge s$. Since $|u|^{2s} \le 1 + |u|^{2k}$ and repeated integration by parts gives $\widehat{\eta^{(k)}}(u) = (iu)^k \widehat{\eta}(u)$, Plancherel's identity yields $E[|U_1|^{2s}] \le 1 + \int_{\mathbb{R}} |u|^{2k} |\widehat{\eta}(u)|^2 du = 1 + \|\eta^{(k)}\|^2_{L^2(\mathbb{R})} < \infty$. The final inequality follows because $\eta \in C^\infty(\mathbb{R})$ is compactly supported.

Above we showed $\|g_h\|^2_{H^s_{\mathrm{av}}(\mathbb{R}^d)} \le C_s h^{d-2s}$. Then define $A = C_s^{-1/2} h^{s-d/2}$ and $m(x) = A g_h(x)$ so that $\|m\|_{H^s_{\mathrm{av}}(\mathbb{R}^d)} \le 1$. Let ψ be uniform on $[0,1]^d$ and set $Y(1) = Y(0) = m(\psi)$. This law belongs to $\mathcal{P}^s_d$ and has $V^*(P) = 0$.

Let $\mathcal{E}_C$ denote the event $\sum_{i=1}^n \mathbb{1}(\psi_i \in C) = 1$, and let $i(C)$ denote the unique index in C on this event. Since m vanishes outside C, on $\mathcal{E}_C$ we have $\sum_{i=1}^n Z_i m(\psi_i) = A Z_{i(C)} g_h(\psi_{i(C)})$ and hence $(\sum_{i=1}^n Z_i m(\psi_i))^2 = A^2 g_h(\psi_{i(C)})^2$. Then pointwise we have $\mathbb{1}(\mathcal{E}_C)(\sum_{i=1}^n Z_i m(\psi_i))^2 = A^2 \sum_{i=1}^n \mathbb{1}(\psi_i \in C)\prod_{j \ne i} \mathbb{1}(\psi_j \notin C) g_h(\psi_i)^2$. Proposition 2.3 therefore gives

$$\mathcal{V}_n(\sigma, P) \ge \frac{4}{n} E[\mathbb{1}(\mathcal{E}_C)(\sum_{i=1}^n Z_i m(\psi_i))^2] = \frac{4A^2}{n} \sum_{i=1}^n \int_C g_h(x)^2 dx P(\psi \notin C)^{n-1}.$$

The equality follows by taking expectations in the pointwise identity and using independence and uniformity of the sampled covariates. Uniformity gives $P(\psi \notin C) = 1 - 1/K$. Since $K \ge n$, $(1 - 1/K)^{n-1} \ge (1 - 1/K)^{K-1} = (1 + 1/(K-1))^{-(K-1)} \ge e^{-1}$, where the final inequality uses $(1 + 1/r)^r \le e$. Together with $\int_C g_h(x)^2 dx = h^d$, this shows $\mathcal{V}_n(\sigma, P) \ge 4A^2 h^d / e = 4h^{2s}/(eC_s)$. Finally, $q \le 2n^{1/d}$, so $h^{2s} \geq 2^{-2s} n^{-2s/d}$. Because the constructed law P does not depend on σ, the argument in fact gives $\sup_{P \in \mathcal{P}_d^s} \inf_{\sigma \in \mathcal{D}_n} \mathcal{V}_n(\sigma, P) \geq 4n^{-2s/d}/(eC_s 2^{2s})$. The theorem follows from $\inf_\sigma \sup_P \mathcal{V}_n(\sigma, P) \geq \sup_P \inf_\sigma \mathcal{V}_n(\sigma, P)$, with $c_0 = 4/(eC_s 2^{2s})$. □

Lemma B.8 (Sobolev Variance Calibration)**.** *Fix $s > 0$ and let U_d be uniform on $[0,1]^d$. There is a constant $c_s > 0$ depending only on s such that, for every $d \geq 1$,*

$$c_s \leq \sup_{\|m\|_{H^s_{\text{av}}(\mathbb{R}^d)} \leq 1} \operatorname{Var}(m(U_d)) \leq 1.$$

Proof. For the upper bound, every m in the unit ball satisfies

$$\operatorname{Var}(m(U_d)) \leq \int_{[0,1]^d} m(x)^2 dx \leq \|m\|^2_{L^2(\mathbb{R}^d)} \leq \|m\|^2_{H^s_{\text{av}}(\mathbb{R}^d)} \leq 1.$$

For the lower bound, use the function η from the proof above and set $g_d(x) = \Pi_{\ell=1}^d \eta(x_\ell)$. Because $\int_0^1 \eta(u)du = 0$ and $\int_0^1 \eta(u)^2 du = 1$, we have $E[g_d(U_d)] = 0$ and $\operatorname{Var}(g_d(U_d)) = 1$. The calculation in Equations (B.20)–(B.21), specialized to $h = 1$, gives $\|g_d\|^2_{H^s_{\text{av}}(\mathbb{R}^d)} \leq C_s$ uniformly over $d \geq 1$. Therefore, $m_d = C_s^{-1/2} g_d$ belongs to the unit ball and satisfies $\operatorname{Var}(m_d(U_d)) = C_s^{-1}$. The result follows with $c_s = C_s^{-1}$. □

We finish with the proof of the embedding used to obtain Corollary A.4.

Proof of Lemma A.3. Write $Q = [0,1]^d$, $M = \|f\|_\infty$, and $L = [f]_\beta$. Because $0 < \beta \leq 1$, $\rho_\beta(x, y) = \|x - y\|_2^\beta$ is a metric on $\mathbb{R}^d$. Define the McShane extension

$$\tilde{f}(x) = \inf_{z \in Q} \big(f(z) + L\|x - z\|_2^\beta \big). \tag{B.22}$$

For $x \in Q$, choosing $z = x$ gives $\tilde{f}(x) \leq f(x)$, while Hölder continuity gives $f(z) + L\|x - z\|_2^\beta \geq f(x)$ for every $z \in Q$. Thus $\tilde{f} = f$ on Q. The triangle inequality for ρ_β also gives $[\tilde{f}]_\beta \leq L$. Set $\bar{f}(x) = (-M) \vee (\tilde{f}(x) \wedge M)$. Because truncation is nonexpansive, $\bar{f} = f$ on Q, $\|\bar{f}\|_\infty \leq M$, and $[\bar{f}]_\beta \leq L$.

Fix a smooth function χ with compact support such that $0 \leq \chi \leq 1$ and $\chi = 1$ on Q, and define $F = \chi \bar{f}$. Then $F = f$ on Q and F has support in a fixed bounded set D. If $\|x - y\|_2 \leq 1$, then $|F(x) - F(y)| \leq |\bar{f}(x) - \bar{f}(y)| + M|\chi(x) - \chi(y)| \leq (L + M\|\nabla\chi\|_\infty)\|x - y\|_2^\beta$. If $\|x - y\|_2 > 1$, then $|F(x) - F(y)| \leq 2M \leq 2M\|x - y\|_2^\beta$. Consequently, $\|F\|_\infty + [F]_\beta \leq C_d(M + L)$. Let $A = M + L$ and fix $s \in [\beta/2, \beta)$.

The whole-space increment characterization of $H^s(\mathbb{R}^d)$ (Adams and Fournier, 2003) applies uniformly over this range of s. If $F(x) - F(y) \neq 0$, at least one of x and y belongs to D. Since the integrand below is symmetric in x and y, we may therefore restrict one integral to D at the cost of a factor of two:

$$\begin{aligned}\|F\|_{H^s(\mathbb{R}^d)}^2 &\leq C_{d,\beta}\left(\|F\|_{L^2(\mathbb{R}^d)}^2 + \iint_{\mathbb{R}^d\times\mathbb{R}^d} \frac{|F(x)-F(y)|^2}{\|x-y\|_2^{d+2s}} dxdy\right) \\ &\leq C_{d,\beta}\left(\|F\|_{L^2(\mathbb{R}^d)}^2 + 2\int_D\int_{\mathbb{R}^d} \frac{|F(x)-F(y)|^2}{\|x-y\|_2^{d+2s}} dydx\right).\end{aligned}$$

A calculation using the increment bounds above and polar coordinates gives

$$\begin{aligned}\|F\|_{H^s(\mathbb{R}^d)}^2 &\leq C_{d,\beta}A^2\left(1 + \int_0^1 r^{2(\beta-s)-1}dr + \int_1^\infty r^{-2s-1}dr\right) \\ &= C_{d,\beta}A^2\left(1 + \frac{1}{2(\beta-s)} + \frac{1}{2s}\right) \leq \frac{C_{d,\beta}}{\beta-s}A^2.\end{aligned}$$

Setting $F_s = F$ proves Equation (A.7).

For a bounded Lipschitz f, repeat the same construction with $\beta = 1$ to obtain a compactly supported Lipschitz extension F_1 satisfying $\|F_1\|_\infty + [F_1]_1 \leq C_d(\|f\|_\infty + [f]_1)$. Rademacher's theorem bounds $\|\nabla F_1\|_\infty$ by the Lipschitz coefficient of F_1. Because F_1 has support in the fixed bounded set D, Equation (4.2) gives $\|F_1\|_{H^1(\mathbb{R}^d)} \leq C_d(\|f\|_\infty + [f]_1)$, proving the final claim. □

B.4 Proofs for Section 5

We first record an oracle inequality separating working-model approximation from design error.

Proposition B.9 (Approximation Oracle)**.** *Let $K = 1+W$ be a design kernel, where W is PSD, let $\mathcal{G}$ be the $L^2(P_\psi)$ closure of $\mathcal{H}_K$, and let $m_\mathcal{G}$ be the $L^2(P_\psi)$ projection of m onto $\mathcal{G}$. Define $\Delta_\mathcal{G}(P) = 4E[(m(\psi) - m_\mathcal{G}(\psi))^2]$. For $\sigma = \mathrm{GSW}(K, \varphi)$ with $\varphi \in (0,1)$,*

$$\mathcal{V}_n(\sigma, P) \leq \frac{\Delta_\mathcal{G}(P)}{\varphi} + \inf_{f\in\mathcal{H}_K}\left\{\frac{4}{\varphi}E[(m_\mathcal{G}(\psi) - f(\psi))^2] + \frac{8}{n(1-\varphi)}\|f\|_{\mathcal{H}_K}^2\right\}.$$

Proof. Set $A = E[(m(\psi) - m_\mathcal{G}(\psi))^2]$, so that $\Delta_\mathcal{G}(P) = 4A$. Because $m - m_\mathcal{G}$ is orthogonal in $L^2(P_\psi)$ to $m_{\mathcal{G}} - f$ for every $f \in \mathcal{H}_K$, we have $E[(m(\psi) - f(\psi))^2] = A + E[(m_{\mathcal{G}}(\psi) - f(\psi))^2]$. By Proposition 4.3,

$$\begin{aligned}\mathcal{V}_n(\sigma, P) &\le \inf_{f\in\mathcal{H}_K}\left\{\frac{4}{\varphi}\left(A + E[(m_{\mathcal{G}}(\psi) - f(\psi))^2]\right) + \frac{8}{n(1-\varphi)}\|f\|^2_{\mathcal{H}_K}\right\}\\ &= \frac{\Delta_{\mathcal{G}}(P)}{\varphi} + \inf_{f\in\mathcal{H}_K}\left\{\frac{4}{\varphi}E[(m_{\mathcal{G}}(\psi) - f(\psi))^2] + \frac{8}{n(1-\varphi)}\|f\|^2_{\mathcal{H}_K}\right\}.\end{aligned}$$

□

We next give the general result underlying the streamlined main-text theorem. In addition to main effects and bivariate interactions, it allows a fixed block $J \subseteq [d]$ of priority covariates, such as baseline outcomes, to enter jointly and have a different smoothness order. Write $d_0 = |J|$ and $q = d - d_0$. Let $\mathcal{G}_1$ be the $L^2(P_\psi)$ closure of functions $g(\psi) = a + g_J(\psi_J) + \sum_{j\notin J} g_j(\psi_j)$, and let $\mathcal{G}_2$ be the closure of functions $g(\psi) = a + g_J(\psi_J) + \sum_{\substack{j<\ell\\ j,\ell\notin J}} g_{j\ell}(\psi_j, \psi_\ell)$. For $k \in \{1, 2\}$, write the canonical projection as

$$m_{\mathcal{G}_k}(\psi) = a + m_J(\psi_J) + \sum_{j\notin J} m_j(\psi_j) + \sum_{\substack{j<\ell\\ j,\ell\notin J}} m_{j\ell}(\psi_j, \psi_\ell). \tag{B.23}$$

The bivariate terms are omitted for $k = 1$. The priority component and main effects have mean zero under Lebesgue measure, and each bivariate component integrates to zero in either argument. For $S = (s_J, s)$ with $s_J, s > 0$, let $\mathcal{P}^S_{d,k}$ contain the laws satisfying Assumption 2.1, $E[m(\psi)^2] \le 1$, and

$$\|m_J\|^2_{H^{s_J}_{\mathrm{av}}(\mathbb{R}^{d_0})} + \sum_{j\notin J}\|m_j\|^2_{H^s_{\mathrm{av}}(\mathbb{R})} + \sum_{\substack{j<\ell\\ j,\ell\notin J}}\|m_{j\ell}\|^2_{H^s_{\mathrm{av}}(\mathbb{R}^2)} \le 1. \tag{B.24}$$

Again, the bivariate terms are omitted for $k = 1$. When J is empty, every priority-block term is omitted. Fix $c > 0$, set $\nu_n = c/\log n$, and write $W_{r,n} = W_r^{\mathrm{Mat},\nu_n}$. When J is nonempty, define

$$\begin{aligned}K_{n,1}(\psi, \psi') &= 1 + \frac{1}{2}W_{d_0,n}(\psi_J, \psi'_J) + \frac{1}{2q}\sum_{j\notin J} W_{1,n}(\psi_j, \psi'_j),\\ K_{n,2}(\psi, \psi') &= 1 + \frac{1}{2}W_{d_0,n}(\psi_J, \psi'_J) + \frac{1}{2\binom{q}{2}}\sum_{\substack{j<\ell\\ j,\ell\notin J}} W_{2,n}(\psi_{j\ell}, \psi'_{j\ell}).\end{aligned}$$

When J is empty, omit the priority term and normalize the remaining nonconstant kernel weight to one. The canonical component spaces are mutually orthogonal and closed in Lebesgue L^2, and the density bounds make the Lebesgue and $L^2(P_\psi)$ norms equivalent. Since each Matérn RKHS is dense in its corresponding component L^2 space, $\mathcal{G}_k$ is exactly the corresponding structured function class written in canonical form. For $k \in \{1,2\}$, let $m_{\mathcal{G}_k}$ be the $L^2(P_\psi)$ projection of m onto $\mathcal{G}_k$ and write $\Delta_{\mathcal{G}_k}(P) = 4E[(m(\psi) - m_{\mathcal{G}_k}(\psi))^2]$.

Theorem B.10 (General Low-Order Nonparametric Balance)**.** *Fix $n \geq 2$, d, smoothness $S = (s_J, s)$ with $s_J, s > 0$, $c > 0$, and $k \in \{1,2\}$, where $d_0 = |J|$ and $d - d_0 \geq k$. Define $\alpha_k = (2s/k) \wedge 1$ and, when J is nonempty, define $\alpha_J = (2s_J/d_0) \wedge 1$. When J is empty, omit α_J and every priority-block term below.*

(i) *Let $\sigma_n = \mathrm{GSW}(K_{n,k}, \varphi_n)$, where $1 - \varphi_n = 1/2 \wedge (n^{-1}(d-d_0)^k \log n)^{1/2}$. Then, for some $C > 0$ depending on $(d_0, s_J, s, c, \underline{p}, \overline{p})$, uniformly over $P \in \mathcal{P}^S_{d,k}$,*

$$\mathcal{V}_n(\sigma_n, P) \leq \Delta_{\mathcal{G}_k}(P) + C\left(\left(\frac{\log n}{n}\right)^{\alpha_J} + \left(\frac{(d-d_0)^k \log n}{n}\right)^{\alpha_k} \right)^{1/2}. \qquad \text{(B.25)}$$

(ii) *On the subclass of $\mathcal{P}^S_{d,k}$ with $m = m_{\mathcal{G}_k}$, any fixed $\varphi \in (0,1)$ yields the same bound for $\sigma_n = \mathrm{GSW}(K_{n,k}, \varphi)$ with $\Delta_{\mathcal{G}_k}(P) = 0$ and the outer square root removed. The constant may additionally depend on φ.*

The next lemma collects the approximation calculation shared by the two structured designs. When J is empty, every priority-block term in the statement and proof is omitted.

Lemma B.11 (Low-Order Matérn Approximation)**.** *Fix $k \in \{1,2\}$ and smoothness $s_0, s > 0$. Let g have the decomposition in Equation (B.23), with canonical components g_J, g_j, and $g_{j\ell}$. Suppose these satisfy the pooled smoothness budget in Equation (B.24) and also that $E[g(\psi)^2] \leq 1$. Set $\alpha_k = (2s/k) \wedge 1$ and, when $d_0 > 0$, set $\alpha_J = (2s_0/d_0) \wedge 1$. There are constants $c_0, C > 0$ and $n_0 \geq 1$ depending only on $(d_0, s_0, s, c, \underline{p}, \overline{p})$ such that, for every $n \geq n_0$ satisfying $(d-d_0)^k \log n \leq c_0 n$ and every λ_n satisfying $n^{-1} \leq \lambda_n \leq c_0/((d-d_0)^k \log n)$,*

$$\inf_{f \in \mathcal{H}_{K_{n,k}}} \left\{ E[(g(\psi) - f(\psi))^2] + \lambda_n \|f\|^2_{\mathcal{H}_{K_{n,k}}} \right\} \leq C\big((\lambda_n \log n)^{\alpha_J} + ((d-d_0)^k \lambda_n \log n)^{\alpha_k}\big).$$

Proof. As in Equation (B.23), write $g = a + g_J + \sum_j g_j + \sum_{j<\ell} g_{j\ell}$, where every sum in the proof ranges over indices outside J. We construct $f = a + f_J + \sum_j f_j + \sum_{j<\ell} f_{j\ell} \in \mathcal{H}_{K_{n,k}}$ as a witness for the inequality above. The construction has three steps.

1. *Componentwise approximation.* For each component g_J, g_j, and $g_{j\ell}$, we construct a raw Matérn-RKHS approximation, denoted by v_J, v_j, or $v_{j\ell}$, satisfying the componentwise approximation profile in Lemma B.7.

2. *Canonicalization.* The initial approximations may not satisfy the same canonical centering restrictions as their targets. For example, $\int_0^1 g_{j\ell}(u, y)du = \int_0^1 g_{j\ell}(x, u)du = 0$ for every $x, y \in [0, 1]$, whereas these equalities may fail for $v_{j\ell}$. We therefore transform the initial approximations using appropriate L^2 orthogonal projections to construct f_J, f_j, and $f_{j\ell}$ satisfying these restrictions, while preserving each approximation profile up to a constant factor.

3. *Assembly.* We set $f = a + f_J + \sum_j f_j + \sum_{j<\ell} f_{j\ell}$. The component errors $g_J - f_J$, $g_j - f_j$, and $g_{j\ell} - f_{j\ell}$ lie in mutually orthogonal subspaces under Lebesgue measure. Hence the squared Lebesgue L^2 error is the sum of their squared component errors. The assumed density upper bound in Assumption 2.1 transfers this to $E[(g(\psi) - f(\psi))^2]$. The RKHS norm of f is also controlled by the corresponding component RKHS norms by the sum-kernel norm formula. Combining these bounds with the pooled Sobolev budget gives the required witness bound.

Step 1: Componentwise approximation. For a penalty λ_n and component kernel W, define the penalized approximation profile $\mathcal{A}_{\lambda_n,W}(h, v) = \|h - v\|_2^2 + \lambda_n \|v\|_{\mathcal{H}_W}^2$, where the first norm is Lebesgue L^2 on the appropriate cube. After decreasing c_0 if necessary, the restrictions on λ_n imply that both λ_n and $(d - d_0)^k \lambda_n$ belong to $[n^{-1}, c_0/\log n]$. Lemma B.7 therefore allows us to choose a raw priority block witness v_J and, when $k = 2$, raw interaction witnesses $v_{j\ell}$ satisfying

$$\mathcal{A}_{\lambda_n,W_{d_0,n}}(g_J, v_J) \leq C\|g_J\|^2_{H^{s_0}_{\mathrm{av}}(\mathbb{R}^{d_0})}(\lambda_n \log n)^{\alpha_J}$$
$$\mathcal{A}_{(d-d_0)^2\lambda_n,W_{2,n}}(g_{j\ell}, v_{j\ell}) \leq C\|g_{j\ell}\|^2_{H^{s}_{\mathrm{av}}(\mathbb{R}^2)}((d - d_0)^2\lambda_n \log n)^{\alpha_k}.$$

The first row is omitted when J is empty, and the second is used only when $k = 2$. When $k = 1$, applying the same lemma directly to each main effect gives

$$\mathcal{A}_{(d-d_0)\lambda_n,W_{1,n}}(g_j, v_j) \leq C\|g_j\|^2_{H^{s}_{\mathrm{av}}(\mathbb{R})}((d - d_0)\lambda_n \log n)^{\alpha_k}.$$

When $k = 2$, the canonical decomposition of g contains main effects $g_j \in H^s_{\mathrm{av}}(\mathbb{R})$, but the design RKHS only contains bivariate components $f_{j\ell} \in \mathcal{H}_{W_{2,n}}$. List the coordinates outside J as $j_1 < \cdots < j_q$, where $q = d - d_0$, and set $\pi(j_1) = \pi(j_2) = (j_1, j_2)$ and $\pi(j_r) = (j_1, j_r)$ for $r = 3, \ldots, q$. For each $j \notin J$, we use the bivariate kernel indexed by $\pi(j)$ to approximate g_j. Viewed as a function on $\mathbb{R}^2$ that is constant in the other coordinate, the main effect need not be square-integrable. We therefore fix a smooth compactly supported function χ equal to one on $[0,1]$ and define $g_j^\chi(x, y) = g_j(x)\chi(y)$ if j is the first coordinate of $\pi(j)$ and $g_j^\chi(x, y) = \chi(x)g_j(y)$ if it is the second. On the unit square, g_j^χ agrees with g_j evaluated at coordinate j. When j is the first coordinate, Fourier factorization gives $\widehat{g_j^\chi}(\omega_1, \omega_2) = \widehat{g}_j(\omega_1)\widehat{\chi}(\omega_2)$, while $1 + (\omega_1^2 + \omega_2^2)/2 \le (1 + \omega_1^2)(1 + \omega_2^2)$. Then

$$\begin{aligned}\|g_j^\chi\|^2_{H^s_{\mathrm{av}}(\mathbb{R}^2)} &= \int_{\mathbb{R}^2} |\widehat{g}_j(\omega_1)|^2 |\widehat{\chi}(\omega_2)|^2 \left(1 + \frac{\omega_1^2 + \omega_2^2}{2}\right)^s d\omega_1 d\omega_2 \\ &\le \|g_j\|^2_{H^s_{\mathrm{av}}(\mathbb{R})} \|\chi\|^2_{H^s(\mathbb{R})} \le C\|g_j\|^2_{H^s_{\mathrm{av}}(\mathbb{R})}.\end{aligned}$$

The display treats the case in which j is the first coordinate, and the other case is identical after exchanging ω_1 and ω_2. Then $g_j^\chi \in H^s_{\mathrm{av}}(\mathbb{R}^2)$, so once again Lemma B.7 supplies a raw witness v_j satisfying approximation rate $\mathcal{A}_{(d-d_0)^2\lambda_n, W_{2,n}}(g_j^\chi, v_j) \le C\|g_j\|^2_{H^s_{\mathrm{av}}(\mathbb{R})}((d - d_0)^2\lambda_n \log n)^{\alpha_k}$.

Step 2: Canonicalization. Fix one of the profiles displayed in Step 1, and let h, v, W, and $\lambda_{h,n}$ denote its target, raw witness, kernel, and penalty. We transform v into a witness f satisfying the canonical restrictions with

$$\mathcal{A}_{\lambda_{h,n}, W}(h, f) \le C\mathcal{A}_{\lambda_{h,n}, W}(h, v).$$

First, we impose the scalar centering restrictions. The priority block and main effect witnesses f_J and f_j must satisfy $\int_{[0,1]^{d_0}} f_J = 0$ and $\int_0^1 f_j = 0$ when $k = 1$. Consider the priority witness v_J. Simply demeaning it by setting $f_J = v_J - \int_{[0,1]^{d_0}} v_J$ is not feasible, because nonzero constant functions do not belong to the RKHS $\mathcal{H}_{W_{d_0,n}} = H^{d_0/2 + c/\log n}(\mathbb{R}^{d_0})$. To handle this technical complication, let χ_{d_0} be a fixed smooth compactly supported function equal to one on the cube, set $\mu_J = \int_{[0,1]^{d_0}} v_J$, and define

$f_J = v_J - \mu_J \chi_{d_0}$. Then $\int_{[0,1]^{d_0}} f_J = \int_{[0,1]^{d_0}} g_J = 0$, and

$$\|g_J - v_J\|_2^2 = \|g_J - f_J\|_2^2 + \mu_J^2 + 2\int_{[0,1]^{d_0}} (g_J - f_J)(f_J - v_J) = \|g_J - f_J\|_2^2 + \mu_J^2.$$

The cross term vanishes because $f_J - v_J = -\mu_J$ on the cube and $g_J - f_J$ has mean zero. In particular, $\mu_J^2 \le \|g_J - v_J\|_2^2$ and the approximation error contracts: $\|g_J - f_J\|_2^2 \le \|g_J - v_J\|_2^2$. The final statement of Lemma B.6 gives $\|\chi_{d_0}\|^2_{\mathcal{H}_{W_{d_0,n}}} \le C \log n$. Therefore,

$$\begin{aligned}\lambda_n \|f_J\|^2_{\mathcal{H}_{W_{d_0,n}}} &= \lambda_n \|v_J - \mu_J \chi_{d_0}\|^2_{\mathcal{H}_{W_{d_0,n}}} \le 2\lambda_n \|v_J\|^2_{\mathcal{H}_{W_{d_0,n}}} + 2\lambda_n \mu_J^2 \|\chi_{d_0}\|^2_{\mathcal{H}_{W_{d_0,n}}} \\ &\le 2\lambda_n \|v_J\|^2_{\mathcal{H}_{W_{d_0,n}}} + C\lambda_n \log n \|g_J - v_J\|_2^2.\end{aligned}$$

The first inequality is Young's and the second uses bounds above on μ_J and χ_{d_0}. Since $\lambda_n \log n \le c_0$, the L^2 contraction and RKHS-norm bound give

$$\begin{aligned}\mathcal{A}_{\lambda_n, W_{d_0,n}}(g_J, f_J) &= \|g_J - f_J\|_2^2 + \lambda_n \|f_J\|^2_{\mathcal{H}_{W_{d_0,n}}} \\ &\le (1 + C\lambda_n \log n)\|g_J - v_J\|_2^2 + 2\lambda_n \|v_J\|^2_{\mathcal{H}_{W_{d_0,n}}} \\ &\le C\big(\|g_J - v_J\|_2^2 + \lambda_n \|v_J\|^2_{\mathcal{H}_{W_{d_0,n}}}\big) = C\mathcal{A}_{\lambda_n, W_{d_0,n}}(g_J, v_J).\end{aligned}$$

When $k = 1$, the same construction with penalty $(d - d_0)\lambda_n$, produces centered f_j from v_j that preserve the approximation profile up to the same constant factor.

Interactions. We next canonicalize the bivariate interaction witnesses used when $k = 2$. Each $f_{j\ell}$ must satisfy $\int_0^1 f_{j\ell}(x, z)dz = \int_0^1 f_{j\ell}(z, y)dz = 0$ for every $x, y \in [0, 1]$. To describe the required projection, let $u \in L^2([0, 1]^2)$ and set $A_1 u(x) = \int_0^1 u(x, z)dz$, $A_2 u(y) = \int_0^1 u(z, y)dz$, and $A_0 u = \int_0^1 \int_0^1 u(z, w)dzdw$. The orthogonal projection of u onto the subspace of $L^2([0, 1]^2)$ satisfying the two canonical restrictions is

$$(Qu)(x, y) = u(x, y) - A_1 u(x) - A_2 u(y) + A_0 u. \tag{B.26}$$

Now let $v \in \mathcal{H}_{W_{2,n}}$ be one of the raw witnesses from Step 1. The same projection formula can be used to define an operator $(Qv)(x, y) = v(x, y) - A_1 v(x) - A_2 v(y) + A_0 v$ for $v \in L^2(\mathbb{R}^2)$. However, $Qv \notin L^2(\mathbb{R}^2)$ in general, let alone $\mathcal{H}_{W_{2,n}}$, because its marginal terms are constant in the unused coordinate. To fix this, we use the compactly supported smooth function χ above and define the modified projection

$$(\widetilde{Q}v)(x,y) = v(x,y) - A_1v(x)\chi(y) - A_2v(y)\chi(x) + (A_0v)\chi(x)\chi(y).$$

Because $\chi = 1$ on $[0,1]$, the restrictions of $\widetilde{Q}v$ and Qv to the unit square coincide. We first show that $\widetilde{Q}$ is a bounded linear operator on $H^t(\mathbb{R}^2)$, uniformly over $t \in [1,2]$. Fourier factorization and $(1+\omega_1^2+\omega_2^2)^t \le (1+\omega_1^2)^t(1+\omega_2^2)^t$ give $\|A_1v(x)\chi(y)\|_{H^t(\mathbb{R}^2)} \le \|A_1v\|_{H^t(\mathbb{R})}\|\chi\|_{H^t(\mathbb{R})}$, and similarly for the other product terms above. It therefore suffices to control the marginal Sobolev norms. For $v \in H^t(\mathbb{R}^2)$, we claim that $\|A_1v\|_{H^t(\mathbb{R})} \le \|v\|_{H^t(\mathbb{R}^2)}$. We first verify the claim for $v \in C_c^\infty(\mathbb{R}^2)$. Writing $\mathcal{F}_x$ for the Fourier transform in the first coordinate, Fubini's theorem gives $\widehat{A_1v}(\omega_1) = \int_0^1 \mathcal{F}_xv(\omega_1,z)dz$. Hence by Jensen's inequality

$$\|A_1v\|_{H^t(\mathbb{R})}^2 = \int_{\mathbb{R}} (1+\omega_1^2)^t \left| \int_0^1 \mathcal{F}_xv(\omega_1,z)dz \right|^2 d\omega_1 \le \int_{\mathbb{R}} \int_0^1 (1+\omega_1^2)^t |\mathcal{F}_xv(\omega_1,z)|^2 dz d\omega_1.$$

For each ω_1, Plancherel's identity in the second coordinate gives $\int_{\mathbb{R}} |\mathcal{F}_xv(\omega_1,z)|^2dz = \int_{\mathbb{R}} |\mathcal{F}_z\mathcal{F}_xv(\omega_1,\omega_2)|^2d\omega_2 = \int_{\mathbb{R}} |\widehat{v}(\omega_1,\omega_2)|^2d\omega_2$. Using this identity and extending the z-integral in the preceding display to $\mathbb{R}$, we have

$$\begin{aligned}\int_{\mathbb{R}} \int_0^1 (1+\omega_1^2)^t |\mathcal{F}_xv(\omega_1,z)|^2 dz d\omega_1 &\le \int_{\mathbb{R}^2} (1+\omega_1^2)^t |\widehat{v}(\omega_1,\omega_2)|^2 d\omega_1 d\omega_2 \\ &\le \int_{\mathbb{R}^2} (1+\omega_1^2+\omega_2^2)^t |\widehat{v}(\omega_1,\omega_2)|^2 d\omega_1 d\omega_2 = \|v\|_{H^t(\mathbb{R}^2)}^2.\end{aligned}$$

For general $v \in H^t(\mathbb{R}^2)$, choose $v_m \in C_c^\infty(\mathbb{R}^2)$ with $v_m \to v$ in $H^t(\mathbb{R}^2)$. The preceding bound makes A_1v_m Cauchy in $H^t(\mathbb{R})$, while Cauchy–Schwarz gives $\|A_1(v_m - v)\|_{L^2(\mathbb{R})} \le \|v_m - v\|_{L^2(\mathbb{R}^2)} \to 0$. Its $H^t(\mathbb{R})$ limit is therefore A_1v, proving the claim by density. The same bound holds for A_2v. Moreover, A_0v is the coefficient of the $L^2([0,1]^2)$ projection onto the constants, so $|A_0v| \le \|v\|_{L^2([0,1]^2)} \le \|v\|_{L^2(\mathbb{R}^2)} \le \|v\|_{H^t(\mathbb{R}^2)}$. The final inequality follows from Plancherel's identity and $(1+\|\omega\|_2^2)^t \ge 1$.

More explicitly, the same factorization gives $\|A_2v(y)\chi(x)\|_{H^t(\mathbb{R}^2)} \le \|A_2v\|_{H^t(\mathbb{R})}\|\chi\|_{H^t(\mathbb{R})}$ and $\|A_0v\chi(x)\chi(y)\|_{H^t(\mathbb{R}^2)} \le |A_0v|\|\chi\|_{H^t(\mathbb{R})}^2$. Since $\|\chi\|_{H^t(\mathbb{R})} \le \|\chi\|_{H^2(\mathbb{R})}$ for $t \in [1,2]$, the preceding bounds imply $\|\widetilde{Q}v\|_{H^t(\mathbb{R}^2)} \le C\|v\|_{H^t(\mathbb{R}^2)}$ uniformly over $t \in [1,2]$.

Set $r_n = 1 + c/\log n$, the Sobolev order of $\mathcal{H}_{W_{2,n}}$. For large enough n, $r_n \in [1,2]$.

Writing $\Lambda_n = \Lambda_{r_n,2}$, Lemma B.6 and the preceding bound at $t = r_n$ give for $v \in \mathcal{H}_{W_{2,n}}$

$$\|\widetilde{Q}v\|^2_{\mathcal{H}_{W_{2,n}}} = \Lambda_n \|\widetilde{Q}v\|^2_{H^{r_n}(\mathbb{R}^2)} \le C\Lambda_n \|v\|^2_{H^{r_n}(\mathbb{R}^2)} = C\|v\|^2_{\mathcal{H}_{W_{2,n}}} < \infty. \tag{B.27}$$

Thus $\widetilde{Q}v$ is a valid witness in $\mathcal{H}_{W_{2,n}}$, and its RKHS penalty is controlled by that of v. Because $Qg_{j\ell} = g_{j\ell}$, orthogonal projection gives the contraction property

$$\|g_{j\ell} - \widetilde{Q}v_{j\ell}\|_{L^2([0,1]^2)} = \|Q(g_{j\ell} - v_{j\ell})\|_{L^2([0,1]^2)} \le \|g_{j\ell} - v_{j\ell}\|_{L^2([0,1]^2)}.$$

Combining the L^2 and Sobolev bounds above, $f_{j\ell} = \widetilde{Q}v_{j\ell}$ satisfies the canonical restrictions and has profile $\mathcal{A}_{(d-d_0)^2\lambda_n, W_{2,n}}(g_{j\ell}, f_{j\ell}) \le C\mathcal{A}_{(d-d_0)^2\lambda_n, W_{2,n}}(g_{j\ell}, v_{j\ell})$.

Lifted main effects. Recall that Step 1 assigned g_j to the pair $\pi(j)$, lifted it to g_j^χ, and produced a raw bivariate witness $v_j \in \mathcal{H}_{W_{2,n}}$ for the kernel acting on that pair. Suppose first that j is the first coordinate of $\pi(j)$. For $(x, y) \in [0,1]^2$, $g_j^\chi(x, y) = g_j(x)$ and $\int_0^1 g_j(x)dx = 0$, so the target is a centered function of the first coordinate alone. The orthogonal projection onto this subspace and its smooth modification on $\mathbb{R}^2$ are

$$(Q_1 v_j)(x, y) = A_1 v_j(x) - A_0 v_j, \qquad (\widetilde{Q}_1 v_j)(x, y) = A_1 v_j(x)\chi(y) - A_0 v_j \chi(x)\chi(y).$$

Because $\chi = 1$ on $[0, 1]$, the restrictions of $\widetilde{Q}_1 v_j$ and $Q_1 v_j$ to the unit square coincide. The calculations above show that $\widetilde{Q}_1 v_j \in \mathcal{H}_{W_{2,n}}$ and $\|\widetilde{Q}_1 v_j\|_{\mathcal{H}_{W_{2,n}}} \le C\|v_j\|_{\mathcal{H}_{W_{2,n}}}$. Moreover, $Q_1 g_j^\chi = g_j^\chi$ on the unit square, so

$$\|g_j^\chi - \widetilde{Q}_1 v_j\|_{L^2([0,1]^2)} = \|Q_1(g_j^\chi - v_j)\|_{L^2([0,1]^2)} \le \|g_j^\chi - v_j\|_{L^2([0,1]^2)}.$$

If j is the second coordinate of $\pi(j)$, use instead $(Q_2 v_j)(x, y) = A_2 v_j(y) - A_0 v_j$ and $(\widetilde{Q}_2 v_j)(x, y) = A_2 v_j(y)\chi(x) - A_0 v_j \chi(x)\chi(y)$. Set $f_j = \widetilde{Q}_1 v_j$ in the first case and $f_j = \widetilde{Q}_2 v_j$ in the second. Then f_j satisfies the required centering restriction and $\mathcal{A}_{(d-d_0)^2\lambda_n, W_{2,n}}(g_j^\chi, f_j) \le C\mathcal{A}_{(d-d_0)^2\lambda_n, W_{2,n}}(g_j^\chi, v_j)$.

Step 3: Assembly. Recall that we must bound $E[(g(\psi) - f(\psi))^2] + \lambda_n \|f\|^2_{\mathcal{H}_{K_{n,k}}}$ for the approximation $f = a + f_J + \sum_j f_j + \sum_{j<\ell} f_{j\ell}$ constructed above. We first consider $k = 2$ and bound its RKHS norm. For each pair (j, ℓ), the interaction witness $f_{j\ell}$ belongs to $\mathcal{H}_{W_{2,n}}$ and approximates the component $g_{j\ell}$. Whenever $\pi(s) = (j, \ell)$, the main-effect witness f_s was constructed in the same bivariate component RKHS $\mathcal{H}_{W_{2,n}}$ and approximates g_s through the lifted target g_s^χ. By the sums-of-kernels theorem (Paulsen and Raghupathi, 2016, Theorem 5.4), $\|f\|^2_{\mathcal{H}_{K_{n,2}}}$ is the minimum of the weighted component norms over all representations $f = a' + h_J + \sum_{j<\ell} h_{j\ell}$ with one $h_{j\ell} \in \mathcal{H}_{W_{2,n}}$ for each pair.

To obtain a representation that we can insert into this minimum, we collect all witnesses constructed in the same pair-indexed bivariate RKHS. To that end, for every $j < \ell$, define the witness assigned to the pair (j, ℓ) by $F_{j\ell} = f_{j\ell} + \sum_{s \notin J} \mathbb{1}(\pi(s) = (j, \ell)) f_s$. Because every main effect is assigned once, $\sum_{j<\ell} F_{j\ell} = \sum_{j<\ell} f_{j\ell} + \sum_j f_j$. The triangle inequality followed by Jensen's inequality gives $\|h_1 + \cdots + h_r\|^2 \le r^2(r^{-1} \sum_{b=1}^r \|h_b\|)^2 \le r \sum_{b=1}^r \|h_b\|^2$. Since each $F_{j\ell}$ contains at most three terms, $\sum_{j<\ell} \|F_{j\ell}\|^2_{\mathcal{H}_{W_{2,n}}} \le 3(\sum_{j<\ell} \|f_{j\ell}\|^2_{\mathcal{H}_{W_{2,n}}} + \sum_j \|f_j\|^2_{\mathcal{H}_{W_{2,n}}})$.

Writing $q = d - d_0$, in the defining sum for $K_{n,2}$ each bivariate kernel has coefficient $1/(2\binom{q}{2})$ when J is nonempty and $1/\binom{q}{2}$ when J is empty. Thus its reciprocal coefficient is at most $2\binom{q}{2}$. Evaluating the minimum defining the sum-kernel RKHS norm at the approximation witnesses $a' = a$, $h_J = f_J$, and $h_{j\ell} = F_{j\ell}$ gives

$$
\begin{aligned}
\|f\|^2_{\mathcal{H}_{K_{n,2}}} &\le \min_{f = a' + h_J + \sum_{j<\ell} h_{j\ell}} \left\{ (a')^2 + 2\|h_J\|^2_{\mathcal{H}_{W_{d_0,n}}} + 2\binom{q}{2} \sum_{j<\ell} \|h_{j\ell}\|^2_{\mathcal{H}_{W_{2,n}}} \right\} \\
&\le a^2 + 2\|f_J\|^2_{\mathcal{H}_{W_{d_0,n}}} + 6\binom{q}{2} \left(\sum_j \|f_j\|^2_{\mathcal{H}_{W_{2,n}}} + \sum_{j<\ell} \|f_{j\ell}\|^2_{\mathcal{H}_{W_{2,n}}} \right).
\end{aligned}
$$

For $k = 1$, the same theorem applied to the representation $f = a + f_J + \sum_j f_j$ gives $\|f\|^2_{\mathcal{H}_{K_{n,1}}} \le a^2 + 2\|f_J\|^2_{\mathcal{H}_{W_{d_0,n}}} + 2q \sum_j \|f_j\|^2_{\mathcal{H}_{W_{1,n}}}$. Since $6\binom{q}{2} \le 3q^2$ and $2q \le 3q$, in either case

$$
\begin{aligned}
\|f\|^2_{\mathcal{H}_{K_{n,k}}} \le C\Bigg(& a^2 + \|f_J\|^2_{\mathcal{H}_{W_{d_0,n}}} \\
& + q^k \left(\sum_j \|f_j\|^2_{\mathcal{H}_{W_{k,n}}} + \sum_{j<\ell} \|f_{j\ell}\|^2_{\mathcal{H}_{W_{2,n}}} \right) \Bigg).
\end{aligned}
\tag{B.28}
$$

As throughout the proof, the priority term is omitted when J is empty and the bivariate sum is omitted when $k = 1$. The density upper bound in Assumption 2.1 gives $E[(g(\psi) - f(\psi))^2] \le \bar{p}\|g - f\|^2_{L^2([0,1]^d)}$. For $k = 2$, canonical orthogonality and the identity of g_j^χ and g_j on the assigned unit square give $\|g - f\|^2_{L^2([0,1]^d)} = \|g_J - f_J\|_2^2 + \sum_j \|g_j^\chi - f_j\|_2^2 + \sum_{j<\ell} \|g_{j\ell} - f_{j\ell}\|_2^2$. Combining these facts with Equation (B.28) collects the component errors and RKHS penalties into the profiles constructed above:

$$E[(g(\psi)-f(\psi))^2]+\lambda_n\|f\|^2_{\mathcal{H}_{K_{n,2}}}\le C\lambda_n a^2+C\big(\|g_J-f_J\|_2^2+\lambda_n\|f_J\|^2_{\mathcal{H}_{W_{d_0,n}}}\big)$$
$$+C\sum_j\big(\|g_j^{\chi}-f_j\|_2^2+q^2\lambda_n\|f_j\|^2_{\mathcal{H}_{W_{2,n}}}\big)+C\sum_{j<\ell}\big(\|g_{j\ell}-f_{j\ell}\|_2^2+q^2\lambda_n\|f_{j\ell}\|^2_{\mathcal{H}_{W_{2,n}}}\big).$$

The parentheticals above are, respectively, $\mathcal{A}_{\lambda_n,W_{d_0,n}}(g_J,f_J)$, $\mathcal{A}_{q^2\lambda_n,W_{2,n}}(g_j^{\chi},f_j)$, and $\mathcal{A}_{q^2\lambda_n,W_{2,n}}(g_{j\ell},f_{j\ell})$. Continuing, by step 1 this is bounded above by

$$C\lambda_n a^2+C(\lambda_n\log n)^{\alpha_J}\|g_J\|^2_{H^{s_0}_{\mathrm{av}}(\mathbb{R}^{d_0})}+C(q^2\lambda_n\log n)^{\alpha_2}\Bigg(\sum_j\|g_j\|^2_{H^s_{\mathrm{av}}(\mathbb{R})}+\sum_{j<\ell}\|g_{j\ell}\|^2_{H^s_{\mathrm{av}}(\mathbb{R}^2)}\Bigg).$$

For $k=1$, the same calculation omits the interaction sum and replaces $(g_j^{\chi},W_{2,n},q^2,\alpha_2)$ by $(g_j,W_{1,n},q,\alpha_1)$. The pooled Sobolev budget bounds the right hand side by $C\lambda_n a^2+C((\lambda_n\log n)^{\alpha_J}+((d-d_0)^k\lambda_n\log n)^{\alpha_k})$.

It remains to control the intercept. Because the canonical components are orthogonal, the density lower bound gives $a^2\le\|g\|^2_{L^2([0,1]^d)}\le\underline{p}^{-1}E[g(\psi)^2]\le\underline{p}^{-1}$. Set $x=(d-d_0)^k\lambda_n\log n$. After decreasing c_0 if necessary, $0<x\le c_0<1$, and $(d-d_0)^k\ge 1$ and $\alpha_k\le 1$ give $\lambda_n\le x\le x^{\alpha_k}$. Hence the nonpriority profile absorbs $\lambda_n a^2$. The componentwise bound and intercept calculation establish the required witness bound. Taking the infimum over $f\in\mathcal{H}_{K_{n,k}}$ proves the claim. □

Proof of Theorem B.10. Define $a_n=\log n/n$ and $b_n=(d-d_0)^k\log n/n$. We use the convention that the priority exponent and every priority term are omitted when J is empty. For part (i), put $\delta_n=1-\varphi_n=(1/2)\wedge b_n^{1/2}$. Let c_0 and n_0 be as in Lemma B.11, decrease c_0 so that $c_0\le 1/2$, and set $b_0=(1/4)\wedge c_0^2$. If $n<n_0$ or $b_n>b_0$, Proposition 4.3 evaluated at zero gives $\mathcal{V}_n(\sigma_n,P)\le 4E[m(\psi)^2]/\varphi_n\le 8$. In this case, b_n is bounded below by a positive constant depending only on the lemma constants: this follows from $b_n>b_0$ in the second case and from $b_n\ge\log n/n$ over the finite set $2\le n<n_0$ in the first. Enlarging C therefore gives the claimed bound. It remains to consider $n\ge n_0$ and $b_n\le b_0$, for which $\delta_n=b_n^{1/2}\le 1/2$. Because $0\in\mathcal{G}_k$ and $m_{\mathcal{G}_k}$ is a projection,

$$\Delta_{\mathcal{G}_k}(P)=4E[(m(\psi)-m_{\mathcal{G}_k}(\psi))^2]\le 4E[m(\psi)^2]\le 4. \tag{B.29}$$

Projection contraction also gives $E[m_{\mathcal{G}_k}(\psi)^2]\le E[m(\psi)^2]\le 1$, so the approximation lemma applies with $g = m_{\mathcal{G}_k}$. Since $1-\varphi_n = \delta_n$ and $\varphi_n = 1-\delta_n \geq 1/2$, the part of $\Delta_{\mathcal{G}_k}(P)/\varphi_n$ exceeding $\Delta_{\mathcal{G}_k}(P)$ is at most $4\delta_n/(1-\delta_n) \leq 8\delta_n$. Proposition B.9 and Lemma B.11 with $\lambda_n = (n\delta_n)^{-1}$ therefore yield

$$\begin{aligned}\mathcal{V}_n(\sigma_n, P) &\leq \Delta_{\mathcal{G}_k}(P) + 8\delta_n + 8 \inf_{f\in\mathcal{H}_{K_{n,k}}} \left\{ E[(m_{\mathcal{G}_k}(\psi) - f(\psi))^2] + \frac{\|f\|^2_{\mathcal{H}_{K_{n,k}}}}{n\delta_n} \right\} \\ &\leq \Delta_{\mathcal{G}_k}(P) + C\left\{ b_n^{1/2} + (a_n/\delta_n)^{\alpha_J} + (b_n/\delta_n)^{\alpha_k} \right\}. \end{aligned} \tag{B.30}$$

The lemma's penalty conditions hold because $\delta_n \leq 1$ gives $(n\delta_n)^{-1} \geq n^{-1}$, while $(d-d_0)^k \log n (n\delta_n)^{-1} = b_n/\delta_n = b_n^{1/2} \leq c_0$. Moreover, $b_n \geq a_n$, so $(a_n/\delta_n)^{\alpha_J} \leq a_n^{\alpha_J/2}$. Because $0 < b_n \leq 1$ and $\alpha_k \leq 1$, we have $b_n^{1/2} = b_n^{\alpha_k/2} b_n^{(1-\alpha_k)/2} \leq b_n^{\alpha_k/2}$, while $(b_n/\delta_n)^{\alpha_k} = b_n^{\alpha_k/2}$. Thus the remainder after $\Delta_{\mathcal{G}_k}(P)$ in Equation (B.30) is at most $C(a_n^{\alpha_J/2} + b_n^{\alpha_k/2})$. By Cauchy–Schwarz, $a_n^{\alpha_J/2} + b_n^{\alpha_k/2} \leq \sqrt{2}(a_n^{\alpha_J} + b_n^{\alpha_k})^{1/2}$, and absorbing $\sqrt{2}$ into C gives the claimed bound. The constant is uniform over $P \in \mathcal{P}^S_{d,k}$, which proves part (i).

For part (ii), fix $\varphi \in (0,1)$ and suppose $m = m_{\mathcal{G}_k}$. Proposition 4.3 gives

$$\mathcal{V}_n(\sigma_n, P) \leq \frac{4}{\varphi} \inf_{f\in\mathcal{H}_{K_{n,k}}} \left\{ E[(m(\psi) - f(\psi))^2] + \frac{2\varphi}{n(1-\varphi)} \|f\|^2_{\mathcal{H}_{K_{n,k}}} \right\}. \tag{B.31}$$

Let λ_n be the maximum of $2\varphi/\{n(1-\varphi)\}$ and $1/n$. The approximation functional is nondecreasing in its penalty, so the right hand side of Equation (B.31) is at most the same expression evaluated at λ_n. Write $L_\varphi = 1 \vee (2\varphi/(1-\varphi))$, so that $\lambda_n = L_\varphi/n$. If $n \geq n_0$ and $L_\varphi b_n \leq c_0$, Lemma B.11 applies and gives the claimed bound because $E[m(\psi)^2] \leq 1$. Otherwise, evaluating Equation (B.31) at zero gives $\mathcal{V}_n(\sigma_n, P) \leq 4/\varphi$. As above, b_n is then bounded below by a positive constant depending only on (n_0, c_0, φ), so enlarging the constant gives the same bound. The constant may additionally depend on the fixed φ, which proves part (ii). $\square$

Proof of Theorem 5.3. Apply Theorem B.10 with $J = \emptyset$ and $k = 2$. Then $d_0 = 0$, $K_{n,2} = K_n$, and $\alpha_2 = s \wedge 1$, so Equation (B.25) reduces to Equation (5.5). $\square$

We next turn to the paired design.

Proof of Theorem 5.4. Let U_M denote any exogenous randomness, independent of the data, used to break ties when constructing the matching, and set $\mathcal{A}_M = \sigma(\psi_{1:n}, U_M)$. Order the pairs and the two indices within each pair $\mathcal{A}_M$-measurably, and let D_M be the $(n/2) \times n$ matrix whose pth row is $e'_{i_p} - e'_{j_p}$. Write $A_n = [K(\psi_i, \psi_j)]_{i,j=1}^n$. Since A_n is positive semidefinite, $G_n^M = D_M A_n D_M'$ is also positive semidefinite. If $\widehat{\kappa}_n > 0$, then $\Gamma_n^M \succeq \varphi I_{n/2}$ and $(\Gamma_n^M)_{pp} = \varphi + (1-\varphi)(G_n^M)_{pp}/\widehat{\kappa}_n \leq 1$ for every p. If $\widehat{\kappa}_n = 0$, then $\Gamma_n^M = \varphi I_{n/2}$ has the same properties. Let $B_M = (\Gamma_n^M)^{1/2}$. Its pth column has squared norm $(\Gamma_n^M)_{pp} \leq 1$, so Lemma 6.1 of Harshaw et al. (2024) gives $E[T|\mathcal{A}_M] = 0$, while their Theorem 6.3 gives $\mathrm{Cov}(B_M T|\mathcal{A}_M) \preceq I_{n/2}$. Since B_M is invertible,

$$E[T|\mathcal{A}_M] = 0, \qquad E[TT'|\mathcal{A}_M] = \mathrm{Cov}(T|\mathcal{A}_M) \preceq (\Gamma_n^M)^{-1}. \tag{B.32}$$

Since D_M is $\mathcal{A}_M$-measurable and $Z = D_M' T$, the tower law gives

$$E[Z|\psi_{1:n}] = E\big[E[D_M' T|\mathcal{A}_M]\big|\psi_{1:n}\big] = E\big[D_M' E[T|\mathcal{A}_M]\big|\psi_{1:n}\big] = 0.$$

By definition, the auxiliary GSW randomization is independent of the data and U_M. Thus U_M and the GSW randomization are jointly independent of the data. Since Z is a measurable function of $\psi_{1:n}$ and these random variables, $Z \perp\!\!\!\perp (Y_i(0), Y_i(1))_{i=1}^n | \psi_{1:n}$. Therefore, $\mathrm{GSW}_M(K, \varphi)$ is admissible.

Let $F_n : \mathcal{H}_K \to \mathbb{R}^n$ be the evaluation operator $(F_n f)_i = f(\psi_i)$ and define $F_n^M = D_M F_n$, so $(F_n^M f)_p = f(\psi_{i_p}) - f(\psi_{j_p})$. As in the proof of Proposition 4.3, we first identify the adjoint $(F_n^M)^* : \mathbb{R}^{n/2} \to \mathcal{H}_K$. Fix $u \in \mathbb{R}^{n/2}$ and $g \in \mathcal{H}_K$. The reproducing property gives $\langle g, (F_n^M)^* u\rangle_{\mathcal{H}_K} = \langle F_n^M g, u\rangle_2 = \sum_{p=1}^{n/2} u_p(g(\psi_{i_p}) - g(\psi_{j_p})) = \langle g, \sum_{p=1}^{n/2} u_p(K(\psi_{i_p}, \cdot) - K(\psi_{j_p}, \cdot))\rangle_{\mathcal{H}_K}$. Since this holds for every $g \in \mathcal{H}_K$, $(F_n^M)^* u = \sum_{p=1}^{n/2} u_p(K(\psi_{i_p}, \cdot) - K(\psi_{j_p}, \cdot))$. Taking $u = e_q$ gives $(F_n^M (F_n^M)^*)_{pq} = K(\psi_{i_p}, \psi_{i_q}) - K(\psi_{i_p}, \psi_{j_q}) - K(\psi_{j_p}, \psi_{i_q}) + K(\psi_{j_p}, \psi_{j_q}) = (G_n^M)_{pq}$. Hence $F_n^M (F_n^M)^* = G_n^M$.

When $\widehat{\kappa}_n > 0$, Equation (B.32) and the definition of Γ_n^M give

$$\mathrm{Cov}(T|\mathcal{A}_M) \preceq \left(\varphi I_{n/2} + \frac{1-\varphi}{\widehat{\kappa}_n} F_n^M (F_n^M)^*\right)^{-1}.$$

Apply Lemma B.5 with $\mathcal{H} = \mathcal{H}_K$, $\mathcal{A} = \mathcal{A}_M$, $F = F_n^M$, $S = T$, $a = \varphi$, and $b = (1-\varphi)/\widehat{\kappa}_n$. Evaluating the resulting minimum at any $f \in \mathcal{H}_K$ gives, for every $\mathcal{A}_M$-measurable y,

$$E[(T'y)^2|\mathcal{A}_M] \leq \frac{1}{\varphi}\|y - F_n^M f\|_2^2 + \frac{\widehat{\kappa}_n}{1-\varphi}\|f\|_{\mathcal{H}_K}^2. \tag{B.33}$$

When $\widehat{\kappa}_n = 0$, the positive semidefinite matrix G_n^M has zero diagonal and hence equals zero. Therefore, $F_n^M (F_n^M)^* = 0$. For every $u \in \mathbb{R}^{n/2}$, this gives $\|(F_n^M)^* u\|_{\mathcal{H}_K}^2 =$

$u'F_n^M(F_n^M)^*u = 0$, so $(F_n^M)^* = 0$ and $F_n^M = 0$. Since $\Gamma_n^M = \varphi I_{n/2}$, Equation (B.32) gives

$$E[(T'y)^2|\mathcal{A}_M] \leq \frac{1}{\varphi}\|y\|_2^2 = \frac{1}{\varphi}\|y - F_n^M f\|_2^2 + \frac{\widehat{\kappa}_n}{1-\varphi}\|f\|_{\mathcal{H}_K}^2.$$

Thus, Equation (B.33) also holds in this case. Let $m_{1:n} = (m(\psi_1), \dots, m(\psi_n))'$. Proposition 2.3, the identity $Z = D_M' T$, and the tower law give

$$\mathcal{V}_n(\sigma, P) = \frac{4}{n}E[(T'D_M m_{1:n})^2]. \tag{B.34}$$

For $f \in \mathcal{H}_K$, write $f_{1:n} = (f(\psi_1), \dots, f(\psi_n))'$ and apply Equation (B.33) with $y = D_M m_{1:n}$. Taking expectations and then the infimum over deterministic f gives

$$\begin{aligned}\mathcal{V}_n(\sigma, P) &\leq \inf_{f \in \mathcal{H}_K} E\left[\frac{4}{\varphi n}\|D_M(m_{1:n} - f_{1:n})\|_2^2 + \frac{4\widehat{\kappa}_n}{n(1-\varphi)}\|f\|_{\mathcal{H}_K}^2\right] \\ &= \inf_{f \in \mathcal{H}_K}\left\{\frac{4}{\varphi}Q_n^M(m-f) + \frac{4\kappa_n}{n(1-\varphi)}\|f\|_{\mathcal{H}_K}^2\right\}. \end{aligned} \tag{B.35}$$

The last equality uses $\kappa_n = E[\widehat{\kappa}_n]$ and the definition of Q_n^M. This proves the theorem.

□

Proof of Theorem 5.6. Write $g = m_{\mathcal{G}_1}$ and $r = m - g$. For any square-integrable h, set $\mu_h = E[h(\psi)]$. For every pair $(i,j) \in M$, the inequality $(a-b)^2 \leq 2a^2 + 2b^2$ gives $(h(\psi_i) - h(\psi_j))^2 \leq 2(h(\psi_i) - \mu_h)^2 + 2(h(\psi_j) - \mu_h)^2$. Since every unit belongs to exactly one pair, summing this inequality and taking expectations gives

$$Q_n^M(h) = \frac{1}{n}E\left[\sum_{(i,j)\in M}(h(\psi_i) - h(\psi_j))^2\right] \leq \frac{2}{n}E\left[\sum_{i=1}^n (h(\psi_i) - \mu_h)^2\right] = 2\operatorname{Var}(h(\psi)).$$

Consequently, for every $f \in \mathcal{H}_{K_{n,1}}$,

$$Q_n^M(m-f) \leq 2Q_n^M(r) + 2Q_n^M(g-f) \leq 2Q_n^M(r) + 4\operatorname{Var}(g(\psi) - f(\psi)). \tag{B.36}$$

The first inequality applies the same two-term bound to $m - f = r + (g - f)$ inside each pair difference. The second applies the preceding bound with $h = g - f$.

Apply Theorem 5.4 with $K = K_{n,1}$ and use Equation (B.36). Set $\lambda_n =$

$\varphi\kappa_n/(4n(1-\varphi))$ and $\overline{\lambda}_n = \max\{\lambda_n, n^{-1}\}$. We obtain

$$\mathcal{V}_n(\sigma_n, P) \le \frac{8}{\varphi} Q_n^M(r) + \frac{16}{\varphi} \inf_{f \in \mathcal{H}_{K_{n,1}}} \left\{ \mathrm{Var}(g(\psi) - f(\psi)) + \lambda_n \|f\|^2_{\mathcal{H}_{K_{n,1}}} \right\}. \tag{B.37}$$

Since $0 \le \kappa_n \le 4$ and φ is fixed, $n^{-1} \le \overline{\lambda}_n \le C_\varphi n^{-1}$ for a constant C_φ depending only on φ. Let c_0 and n_0 be the constants in Lemma B.11, and set $b_n = d \log n / n$. If $n < n_0$ or $C_\varphi b_n > c_0$, Equation (B.37) evaluated at zero and projection contraction give $\mathcal{V}_n(\sigma_n, P) \le 8Q_n^M(r)/\varphi + 16/\varphi$. In this case, b_n is bounded below by a positive constant depending only on (n_0, c_0, φ), so enlarging C proves Equation (5.11). It remains to consider $n \ge n_0$ and $C_\varphi b_n \le c_0$. We then have $d \log n \le c_0 n$ and $d\overline{\lambda}_n \log n \le c_0$. Together with $\overline{\lambda}_n \ge n^{-1}$, these are exactly the lemma's penalty conditions for $k = 1$ and J empty. Projection contraction gives $E[g(\psi)^2] \le E[m(\psi)^2] \le 1$. Because $\mathrm{Var}(g(\psi) - f(\psi)) \le E[(g(\psi) - f(\psi))^2]$ and the approximation functional is nondecreasing in its penalty, Equation (B.37) and Lemma B.11, applied with $k = 1$, J empty, $g = m_{\mathcal{G}_1}$, and penalty $\overline{\lambda}_n$, give

$$\begin{aligned}
\mathcal{V}_n(\sigma_n, P) &\le \frac{8}{\varphi} Q_n^M(r) + \frac{16}{\varphi} \inf_{f \in \mathcal{H}_{K_{n,1}}} \left\{ E[(g(\psi) - f(\psi))^2] + \overline{\lambda}_n \|f\|^2_{\mathcal{H}_{K_{n,1}}} \right\} \\
&\le C \left(Q_n^M(r) + (d\overline{\lambda}_n \log n)^{(2s) \wedge 1} \right).
\end{aligned}$$

Finally, $\overline{\lambda}_n \le C_\varphi / n$ bounds the last line by the right hand side of Equation (5.11). This proves the theorem.

For the Hölder specialization, additionally suppose $d \ge 3$, M is 2-swap-stable for a cost satisfying Assumption 3.3, and r satisfies $|r(x) - r(y)| \le L_r \|x - y\|_2^\beta$. Lemma B.3 gives

$$Q_n^M(r) \le \frac{L_r^2}{n} E\left[\sum_{p=1}^{n/2} \|\psi_{i_p} - \psi_{j_p}\|_2^{2\beta} \right] \le 6L_r^2 (44d\Lambda^2)^{2\beta} n^{-2\beta/d}. \tag{B.38}$$

Substitution into Equation (5.11) gives the displayed bound following the theorem. □

B.5 Proofs for Appendix A.1

Proof of Theorem A.1. Sign symmetry gives $E[T_p | \mathcal{W}_{n,J}] = 0$. Moreover, $T_p T_q$ is $\{-1, 1\}$-valued with conditional mean R_{pq}, so $P(T_p T_q = s | \mathcal{W}_{n,J}) = (1 + sR_{pq})/2$

for $s \in \{-1, 1\}$. For $p \neq q$, let $E_s = \{T_pT_q = s\}$. Both the conditional assignment law and E_s are invariant under the global sign reversal $T \mapsto -T$. Hence $E[T_p|E_s, \mathcal{W}_{n,J}] = E[-T_p|E_s, \mathcal{W}_{n,J}]$, so $E[T_p|E_s, \mathcal{W}_{n,J}] = 0$, and the same argument applies to T_q. Since $C_pC_q = A_pA_q + A_pT_qB_q + A_qT_pB_p + T_pT_qB_pB_q$, it follows that $E[C_pC_q|E_s, \mathcal{W}_{n,J}] = A_pA_q + sB_pB_q$. For fixed $p \neq q$, abbreviate $S = T_pT_q$, $r = R_{pq}$, $\ell = L_{pq}$, $a = A_pA_q$, and $b = B_pB_q$, and let $H_{pq} = (Sr + \ell)C_pC_q/(1 + Sr)$ denote the corresponding summand in Equation (A.3). Iterated expectations give

$$
\begin{aligned}
E[H_{pq}|\mathcal{W}_{n,J}] &= \sum_{s=\pm 1} \frac{1 + sr}{2} \frac{sr + \ell}{1 + sr}(a + sb) = \frac{1}{2} \sum_{s=\pm 1} (sr + \ell)(a + sb) \\
&= \frac{1}{2}\big((r + \ell)(a + b) + (\ell - r)(a - b)\big) = \ell a + rb.
\end{aligned}
$$

Thus, the factor $1 + sr$ in the conditional probability cancels the denominator in the correction, leaving $E[H_{pq}|\mathcal{W}_{n,J}] = L_{pq}A_pA_q + R_{pq}B_pB_q$. On the diagonal, $E[C_p^2|\mathcal{W}_{n,J}] = A_p^2 + B_p^2 = L_{pp}A_p^2 + R_{pp}B_p^2$ because $E[T_p|\mathcal{W}_{n,J}] = 0$ and $L_{pp} = R_{pp} = 1$. Summing over p and q gives $2(A'LA + B'RB)/N$. Equation (A.2) and positive semidefiniteness of L prove the conditional identity in the theorem. Since $E[\widehat{\theta}|\mathcal{W}_{n,J}] = \mathrm{SATE}_n$ is $\mathcal{W}_n$-measurable, the conditional law of total variance gives $\mathrm{Var}(\widehat{\theta}|\mathcal{W}_n) = E[\mathrm{Var}(\widehat{\theta}|\mathcal{W}_{n,J})|\mathcal{W}_n]$. Taking conditional expectations of that identity given $\mathcal{W}_n$ and using positive semidefiniteness of L proves Equation (A.4). □

Proof of Proposition A.2. Write $c(\psi) = E[\tau|\psi]$ and $v(\psi) = \mathrm{Var}(\tau|\psi)$. Decompose

$$
\frac{2}{N} A' L_{\mathrm{pop}} A = \frac{1}{N} \sum_{p=1}^{N} A_p^2 - \frac{2}{N} \sum_{\{p,q\} \in \Pi} A_pA_q + \frac{1}{N-1} \sum_{p=1}^{N} (A_p - \bar{A})^2. \tag{B.39}
$$

Let $\mathcal{P}_{2,n} = \{\{i_p, j_p\} : p \in [N]\}$ be the collection of original pairs and let $\mathcal{P}_{4,n} = \{\{i_p, j_p, i_q, j_q\} : \{p, q\} \in \Pi\}$ be the collection of four-unit groups induced by the outer matching. Writing each inner sum below over ordered pairs, the first term in Equation (B.39) equals $(2n)^{-1}(\sum_i \tau_i^2 + \sum_{G \in \mathcal{P}_{2,n}} \sum_{k \neq \ell \in G} \tau_k\tau_\ell)$. The cross-product magnitude $2N^{-1} \sum_{\{p,q\} \in \Pi} A_pA_q$ equals $(2n)^{-1}(\sum_{G \in \mathcal{P}_{4,n}} \sum_{k \neq \ell \in G} \tau_k\tau_\ell - \sum_{G \in \mathcal{P}_{2,n}} \sum_{k \neq \ell \in G} \tau_k\tau_\ell)$.

We analyze these sums using the bilinear-form lemma of Cytrynbaum (2026, Lemma C.11). For $\mathcal{P}_{2,n}$, the lemma's hypothesis is equivalent to the tight-matching condition in Equation (3.2). For $\mathcal{P}_{4,n}$, the matching-for-unions result of Cytrynbaum (2026, Lemma D.23) establishes the same hypothesis. Apply Lemma C.11 first with

$K = 2$ and $\mathcal{P}_n = \mathcal{P}_{2,n}$, then with $K = 4$ and $\mathcal{P}_n = \mathcal{P}_{4,n}$. In both cases, set the lemma's generic variables equal to τ_i and sampling propensity equal to one. It gives $n^{-1}\sum_{G\in\mathcal{P}_{2,n}}\sum_{k\neq\ell\in G}\tau_k\tau_\ell \xrightarrow{p} E[c(\psi)^2]$ and $n^{-1}\sum_{G\in\mathcal{P}_{4,n}}\sum_{k\neq\ell\in G}\tau_k\tau_\ell \xrightarrow{p} 3E[c(\psi)^2]$.

The law of large numbers also implies $n^{-1}\sum_i \tau_i^2 \xrightarrow{p} E[\tau^2]$ and $\bar{A} = n^{-1}\sum_i \tau_i \xrightarrow{p} E[\tau]$. Consequently, the first term in Equation (B.39) and the cross-product magnitude converge to $(E[\tau^2] + E[c(\psi)^2])/2$ and $E[c(\psi)^2]$, respectively. Their difference is $N^{-1}A'L_\Pi A \xrightarrow{p} E[v(\psi)]/2$. For the third term, $(N-1)^{-1}\sum_p (A_p - \bar{A})^2 = N(N-1)^{-1}(N^{-1}\sum_p A_p^2 - \bar{A}^2) \xrightarrow{p} \mathrm{Var}(c(\psi)) + E[v(\psi)]/2$. Since $L_{\mathrm{pop}} = (L_\Pi + L_0)/2$, adding the limits gives $2A'L_{\mathrm{pop}}A/N \xrightarrow{p} \mathrm{Var}(\tau)$ by the law of total variance. $\square$