Nonparametric Regression on Fixed-Cardinality Subsets: Minimax Risk and Random-Design Effects

Gérard Biau

Sorbonne Université, Institut universitaire de France

gerard.biau@sorbonne-universite.fr

Abstract

We study regression with subsets as covariates. The response is an unknown function of the input subset, and observations consist of noisy evaluations at uniformly sampled subsets, each containing exactly k items from a ground set of size d. This problem arises in combination screening, bundle preference modeling, and other settings in which outcomes depend on collections of prescribed size. The fixed-cardinality constraint couples the membership coordinates, so standard product-domain notions of interaction and smoothness cannot be imported unchanged. We use the harmonic decomposition of the Johnson graph to define an intrinsic Johnson–Sobolev scale and determine the exact dimensions of low-order interaction spaces. We derive finite-sample risk bounds for harmonic projection and establish a nonasymptotic minimax characterization when the Johnson–Sobolev energy budget is controlled relative to the noise variance. We also show that stable least squares removes the signal-dependent fluctuation of empirical projection when the empirical Gram matrix is sufficiently well conditioned. For arbitrary signal-to-noise ratios, a rank-deficiency analysis quantifies the components left unidentified by the random design. Together, the rank-deficiency analysis and a centered completion estimator yield minimax bounds that match up to constants whenever $\min(k, d-k)$ and the smoothness order are fixed. Numerical experiments reported in the Supplementary Material illustrate estimation under different interaction profiles and distinguish the effects of observation noise, random-design fluctuation, and incomplete coverage.

Keywords. Combinatorial regression; fixed-cardinality subsets; harmonic analysis; Johnson graph; minimax estimation; random design; Sobolev ellipsoid.

1 Introduction

Suppose that a pharmacologist wants to estimate the expected efficacy of every three-compound combination drawn from a library of twenty candidates. The unknown response function maps each combination to its expected efficacy. Since there are $\binom{20}{3} = 1140$ possible combinations, an experiment typically provides noisy evaluations for only a fraction of them. The statistical goal is to learn the response function over the entire collection of combinations, including those that were not observed. This is a regression problem in which the input is a set rather than a vector of unconstrained covariates.

Regression problems of this kind arise in biological combination screens, preference studies for bundles, portfolio evaluation, configuration testing, and experimental designs involving selected groups of items. In each case, the response may depend not only on the individual items but also on their interactions within the observed set. In this paper, every input is a subset containing exactly k items selected from a ground set of d available items. The covariate domain is therefore the collection of all such k-subsets, and the statistical objective is to estimate the response associated with each of them. We assume that the observed subsets are sampled uniformly at random, which provides a symmetric benchmark when no subset is favored in advance. Nonuniform and adaptive sampling designs are beyond the scope of the present work.

Let $d \geqslant 2$ and $1 \leqslant k \leqslant d-1$, and define

$$\mathcal{S}_{d,k} = \{S \subseteq [d] : |S| = k\}, \qquad [d] = \{1, \ldots, d\}, \qquad N = |\mathcal{S}_{d,k}| = \binom{d}{k}.$$

We observe independent pairs (S_j, Y_j) satisfying

$$S_j \sim \nu_{d,k}, \qquad Y_j = f^\star(S_j) + \varepsilon_j, \qquad j = 1, \ldots, n,$$

where $\nu_{d,k}$ is the uniform distribution on $\mathcal{S}_{d,k}$. The noise variables are i.i.d., centered, independent of the design, and sub-Gaussian with variance proxy $\sigma^2 > 0$. We evaluate an estimator $\widehat{f}$ using the global squared $L^2(\nu_{d,k})$-loss

$$\|\widehat{f} - f^\star\|_2^2 = \frac{1}{N} \sum_{S \in \mathcal{S}_{d,k}} \{\widehat{f}(S) - f^\star(S)\}^2,$$

and define its risk as the expectation of this loss with respect to both the random design and the noise. Because sampling is performed with replacement, many subsets may remain unobserved, even when n is comparable to N. Thus, without a structural assumption linking the response values across subsets, the function cannot be recovered on unobserved inputs.

Johnson geometry and interaction structure. To make such a structural assumption useful, we first need to specify when two subsets should be regarded as close. Let $S \in \mathcal{S}_{d,k}$ denote a generic input and represent it by its membership indicators

$$X_i(S) = \mathbf{1}_{\{i \in S\}}, \qquad i = 1, \ldots, d.$$

These coordinates satisfy $\sum_{i=1}^d X_i(S) = k$ and are therefore dependent under the uniform distribution on $\mathcal{S}_{d,k}$. More precisely, for $i \neq i'$,

$$\operatorname{Cov}\{X_i(S), X_{i'}(S)\} = -\frac{k(d-k)}{d^2(d-1)}.$$

Thus, the covariate domain is not a product space, and the usual coordinatewise perturbations of the Boolean cube do not remain admissible: flipping a single membership indicator adds or removes an item and produces a subset whose cardinality is no longer k.

The smallest modification that remains inside $\mathcal{S}_{d,k}$ consists in replacing one selected item by one unselected item. The natural mathematical object for recording these cardinality-preserving moves is the Johnson graph $J(d,k)$ (Brouwer et al., 1989). Its vertices are the subsets in $\mathcal{S}_{d,k}$, and two vertices are connected when one can be obtained from the other by a single exchange. The corresponding graph distance is

$$\operatorname{dist}_J(S, S') = k - |S \cap S'|, \qquad S, S' \in \mathcal{S}_{d,k},$$

the number of exchanges needed to transform S into S'. The graph therefore provides both a notion of neighboring inputs and the analytical tools used below to describe regularity and interactions on the fixed-cardinality domain.

The Johnson graph is used here to analyze the response function, not as an object to be estimated. Its associated random walk moves from one k-subset to another by choosing a selected item and exchanging it with an unselected one. The spectral decomposition of this walk induces an orthogonal decomposition of every function on $\mathcal{S}_{d,k}$ into components of increasing interaction order (Delsarte, 1973; Bannai and Ito, 1984; Filmus, 2016). We refer to these components as the Johnson harmonic levels. Level zero contains the constant functions; level one contains centered additive item effects; level two, when present, contains pairwise effects that cannot be represented by constant or additive terms; and higher levels describe genuinely higher-order interactions.

As developed formally in Section 3, the decomposition has levels $0, \ldots, m$, where

$$m = \min(k, d-k).$$

The upper limit m reflects complementation: describing a k-subset through its selected items is equivalent to describing it through the $d-k$ items that are excluded. Our analysis concerns the nonconstant part of the response. We therefore separate the intercept and work with functions centered under the uniform design, i.e., $\mathbb{E}_{S\sim\nu_{d,k}}[f(S)] = 0$. This is a normalization rather than a substantive restriction, since an unknown constant component can be estimated separately at the parametric rate. Accordingly, for a cutoff $D \in \{0, \ldots, m\}$, the space containing all nonconstant interaction levels up to order D has

$$p_D = \binom{d}{D} - 1$$

independent degrees of freedom. The subtraction of one accounts for the removed one-dimensional space of constant functions. Hence, p_D is the number of coefficients to be estimated when interactions up to order D are included, with $p_0 = 0$.

The same random walk defines a normalized Johnson Laplacian, which measures how strongly a function varies across neighboring subsets. On harmonic level $r \in \{0, \ldots, m\}$, this operator acts by multiplication by

$$\gamma_{r,d} = \frac{r(d-r+1)}{d}.$$

Here, $\gamma_{0,d} = 0$ corresponds to the constant functions, while the positive eigenvalues $\gamma_{1,d}, \ldots, \gamma_{m,d}$ increase with the interaction order. The zero eigenvalue means that the Johnson Laplacian does not distinguish functions that differ only by a constant. This is why the smoothness constraint is imposed on the centered response. Section 4 uses the positive weights $\gamma_{r,d}^s$ to define a centered fractional Johnson–Sobolev class of smoothness order $s > 0$ and energy budget $B > 0$. This class is the parameter space for our minimax analysis.

Minimax risk and random-design uncertainty. Truncating the harmonic decomposition at level D creates a familiar statistical tradeoff. Increasing D reduces the approximation error by retaining more interactions, but it also increases the dimension p_D of the fitted model. For $0 \leqslant D < m$, the squared error contributed by the discarded levels is bounded by

$$b_D(s, B) = \frac{B}{\gamma_{D+1,d}^s},$$

because $D+1$ is the first omitted level. At the full cutoff $D = m$, no level is omitted, and we set $b_m(s, B) = 0$. Estimating the p_D retained coefficients from n observations has the usual noise-dependent cost $\sigma^2 p_D/n$, where σ^2 is the sub-Gaussian variance proxy. Our first main result shows that this approximation–estimation tradeoff determines the minimax risk over the Johnson–Sobolev class when the ratio B/σ^2 is uniformly bounded. More precisely, for every fixed $C_{\mathrm{snr}} > 0$, uniformly over all admissible d, k, n, s, B, σ^2 satisfying

$$\frac{B}{\sigma^2} \leqslant C_{\mathrm{snr}},$$

the minimax risk is comparable, up to constants depending only on C_{snr}, to

$$\min_{0 \leqslant D \leqslant m} \left\{ \frac{\sigma^2 p_D}{n} + b_D(s, B) \right\}.$$

Thus, the optimal cutoff balances the noise incurred by estimating the retained interaction levels against the approximation error created by omitting the remaining ones. An equivalent max–min formulation, used to establish the minimax lower bound, is given in Section 4.

The condition $B/\sigma^2 \leqslant C_{\mathrm{snr}}$ deserves further explanation. It is not imposed by the definition of the Johnson–Sobolev class, nor is it required for the minimax lower bound. It enters through the risk analysis of empirical projection. This estimator replaces each population projection coefficient by an average over the sampled subsets and is therefore affected by two sources of randomness. Observation noise perturbs the recorded responses, while the random design perturbs the empirical averages themselves. Thus, even when the responses are observed without noise, the estimated coefficients continue to fluctuate with the sampled subsets. The resulting variance depends on the size of the response function and produces the additional signal-dependent term Bp_D/n. The condition $B/\sigma^2 \leqslant C_{\mathrm{snr}}$ ensures that this term is bounded by a constant multiple of the ordinary noise contribution $\sigma^2 p_D/n$.

Least squares can remove the part of this signal-dependent fluctuation generated by the component already contained in the fitted space. Let W_D denote the space containing the nonconstant interaction levels up to order D. If the true response belongs to W_D, noiseless least squares recovers it exactly whenever its values at the sampled subsets uniquely determine a function in W_D, or equivalently whenever the empirical Gram matrix is invertible. With observation noise, invertibility still guarantees a unique least-squares solution, but it does not control its accuracy: small eigenvalues of the empirical Gram matrix can strongly amplify the noise. Stable estimation therefore requires this matrix to be well conditioned. The leverage score at S measures the largest possible value of $g(S)^2$ among functions $g \in W_D$ with $\|g\|_2 = 1$. We prove that it equals p_D at every subset S. This uniform bound controls the contribution of each sampled subset to the empirical Gram matrix, and matrix concentration then shows that this matrix is well conditioned with high probability once n is of order $p_D \log(2p_D)$. In this regime, the leading variance is $\sigma^2 p_D/n$, and the signal-dependent projection variance disappears.

This stability result also has a limitation. Constant leverage controls the contribution of each individual observation to the empirical Gram matrix, but it does not guarantee that the realized evaluation matrix has full rank. This distinction becomes important as D increases and the fitted space approaches the full function space. Because sampling is performed with replacement, some subsets may remain unobserved, making the empirical Gram matrix rank-deficient and leaving some components of the response unidentified even without noise. We quantify this information loss by the normalized expected rank deficiency

$$\rho_{D,n} = \frac{1}{p_D}\mathbb{E}[p_D - \operatorname{rank}(\mathbf{X}_D)],$$

where $\mathbf{X}_D$ is the evaluation matrix on W_D. Thus, $\rho_{D,n}$ is the expected proportion of the p_D dimensions of W_D that remain unidentified in the noiseless experiment.

At the full level, this description of the unidentified components suggests replacing Gram inversion by direct completion of the response table. Repeated observations are averaged at each sampled subset, and the unobserved responses are assigned a common value chosen to enforce the zero-mean constraint. The resulting centered completion estimator controls the error caused by incomplete coverage while also averaging the observation noise. Combining its risk bound with that of harmonic projection yields lower and upper minimax bounds without any restriction on B/σ^2. These bounds match up to a factor proportional to $\gamma_{m,d}^s$. Since $\gamma_{m,d} \leqslant m$, this factor remains constant when the maximal interaction order $m = \min(k, d-k)$ and the smoothness order s are fixed. In particular, when k is fixed, the minimax risk is characterized up to constants depending only on k and s, uniformly over d, n, B, and σ^2. When m is allowed to grow, the remaining factor identifies the gap between our lower and upper bounds.

Contributions and scope. Taken together, our results turn classical Johnson harmonic analysis into a nonparametric framework for regression with subsets as covariates. The framework links smoothness under local exchanges to interaction spaces of explicit dimension and yields risk bounds that separate harmonic approximation, observation noise, and incomplete design coverage. As a constructive complement, Section B.2 of the Supplementary Material derives an explicit unit-norm tight frame at each harmonic level. This representation is not needed for the statistical guarantees, which depend only on the harmonic subspaces.

To the best of our knowledge, this is the first nonasymptotic minimax analysis of random-design regression on fixed-cardinality subsets under a smoothness condition controlling variation between subsets that differ by a single exchange. We are not aware of prior work that combines the exact interaction dimensions of this domain, a fractional Sobolev scale generated by the Johnson Laplacian, noisy random sampling, and risk bounds that account explicitly for incomplete coverage.

The remainder of the paper proceeds from the harmonic structure to the statistical results. Section 2 reviews the closest literature. Section 3 formalizes the statistical model and the Johnson harmonic decomposition. Section 4 introduces the smoothness classes and develops projection estimation and the first minimax bounds. Section 5 then studies stable least squares, Gram concentration, rank deficiency, and the coverage-sensitive minimax characterization. Section 6 concludes. The Supplementary Material reports the numerical experiments and contains the canonical frame construction and all proofs.

2 Related work

Learning functions on subsets. The closest learning-theoretic literature studies functions defined on the full power set $2^{[d]}$, whose standard harmonic representation is the Fourier–Walsh expansion (O'Donnell, 2014). Under sparsity assumptions on this representation, Stobbe and

Krause (2012) establish recovery guarantees from randomly sampled function values. Other work analyzes the learnability of valuation classes from values or pairwise comparisons, or imposes submodularity through explicit or neural parameterizations (Balcan et al., 2012, 2016; Dolhansky and Bilmes, 2016; De and Chakrabarti, 2022). These approaches differ from the present setting in both assumptions and objective. We observe only one fixed-cardinality layer, impose neither sparsity nor a shape constraint such as monotonicity or submodularity, and study global L^2-risk over a smoothness class rather than sparse identification or multiplicative approximation.

Permutation-invariant neural architectures provide general representations for functions of sets (Zaheer et al., 2017; Wagstaff et al., 2019). They can be applied here after encoding the identities of the items, but their principal results concern representation and approximation rather than smoothness-based statistical guarantees under noisy random sampling. Likewise, Gaussian-process models for choice functions learn the possibly set-valued choice made from a presented collection (Benavoli et al., 2023); this differs from estimating a scalar response attached to every subset.

Regression and signal recovery on graphs. Laplacian regularity is widely used to estimate functions observed on graph vertices. Kirichenko and van Zanten (2017) develop Bayesian Laplacian regularization on growing graphs, and Kirichenko and van Zanten (2018) establish corresponding minimax lower bounds. For neighborhood graphs approximating an underlying Euclidean domain, Laplacian-based estimators can recover continuum Sobolev rates (Green et al., 2021; Shi et al., 2024). The Johnson graph plays a different role here: its vertex set is exactly the finite covariate domain, and both its spectral weights and their multiplicities vary explicitly with (d, k).

Graph signal processing also considers the recovery of bandlimited signals from randomly sampled vertices. Puy et al. (2018) derive stable-sampling conditions governed by the dimension of the bandlimited space and an appropriate graph-coherence parameter; under sampling adapted to the local graph coherence, the resulting scale is of order $p \log p$ for a p-dimensional space. Our target functions need not be bandlimited. Johnson–Sobolev regularity instead controls the L^2-mass outside each low-order approximation space, and our risk analysis accounts jointly for approximation error, observation noise, and directions left unresolved by the random design.

Random least squares. The stability of least-squares approximation from random function evaluations is governed by the supremum of the diagonal of the projection kernel, often called the Christoffel or leverage function (Cohen et al., 2013); see also the correction (Cohen et al., 2019). Matrix Chernoff inequalities provide the corresponding concentration tool (Tropp, 2012). We specialize this general mechanism to the Johnson interaction spaces. Permutation invariance of these spaces, together with transitivity of the action on the Johnson domain, forces the projection-kernel diagonal to be constant and equal to the dimension, so uniform sampling is already leverage-optimal. We then incorporate this stability property into a smoothness-based risk analysis and separate observation noise from variation caused by the random design.

Noisy combinatorial optimization. Noisy evaluations of subsets also arise in subset selection, Bayesian optimization, and combinatorial bandits (Qian et al., 2017; Oh et al., 2019; Liang et al., 2024; Tajdini et al., 2024). These methods use sequentially chosen evaluations to locate or approximate an optimal subset. Their performance is consequently measured through optimization accuracy, regret, or query complexity. Our design is instead i.i.d. and nonadaptive, and the objective is to reconstruct the response function over the entire fixed-cardinality domain. Optimization guarantees therefore do not directly control the global prediction risk considered here.

Harmonic analysis of subset data. The Johnson association scheme, its eigenspaces, and its zonal functions are classical (Delsarte, 1973; Bannai and Ito, 1984). Statistical spectral analysis based on representations of the symmetric group was developed by Diaconis (1988, 1989), including the analysis of data indexed by combinatorial objects, while Diaconis and Rockmore (1993) study the computation of the associated isotypic projections. More recently, Filmus (2016) constructed an explicit orthogonal basis for functions on a fixed-cardinality slice, and Iglesias and Natale (2022) developed a fast full-table Fourier transform on the Johnson graph.

These results provide the harmonic foundation used throughout the paper. Section B.2 of the Supplementary Material gives an explicit unit-norm normalization of the canonical zonal frame and the resulting permutation-equivariant representation of the harmonic projections.

3 Model and Johnson harmonic geometry

3.1 Statistical experiment and risk

Throughout the paper, $d \geqslant 2$, $1 \leqslant k \leqslant d-1$, and $n \geqslant 2$. Set

$$[d] = \{1, \ldots, d\}, \qquad \mathcal{S}_{d,k} = \{S \subseteq [d] : |S| = k\}, \qquad N = |\mathcal{S}_{d,k}| = \binom{d}{k}.$$

The excluded cases $k = 0$ and $k = d$ are degenerate: the domain then contains a single subset, and no cardinality-preserving exchange is possible. We equip $\mathcal{S}_{d,k}$ with the uniform probability measure $\nu_{d,k}$. For real-valued functions f, g on $\mathcal{S}_{d,k}$, write

$$\langle f, g \rangle = \mathbb{E}_{S \sim \nu_{d,k}}[f(S)g(S)], \qquad \|f\|_2^2 = \langle f, f \rangle .$$

Thus,

$$\|f - g\|_2^2 = \frac{1}{N} \sum_{S \in \mathcal{S}_{d,k}} \{f(S) - g(S)\}^2.$$

The design variables $S_1, \ldots, S_n$ are i.i.d. with law $\nu_{d,k}$, and the responses satisfy

$$Y_j = f^\star(S_j) + \varepsilon_j, \qquad j = 1, \ldots, n.$$

The noise variables are i.i.d. with a common law P and are independent of the design. For $\sigma^2 > 0$, let $\mathcal{P}_{\sigma^2}$ denote the class of probability laws P on $\mathbb{R}$ such that, for $\varepsilon \sim P$, $\mathbb{E}_P[\varepsilon] = 0$ and

$$\mathbb{E}_P[\exp(t\varepsilon)] \leqslant \exp\left(\frac{\sigma^2 t^2}{2}\right), \qquad t \in \mathbb{R}.$$

In particular, every $P \in \mathcal{P}_{\sigma^2}$ satisfies $\mathbb{E}_P[\varepsilon^2] \leqslant \sigma^2$.

For a function $f : \mathcal{S}_{d,k} \to \mathbb{R}$, regarded as a possible value of the unknown response $f^\star$, and a noise law $P \in \mathcal{P}_{\sigma^2}$, denote by $\mathbb{P}_{f,P}$ the joint law of $(S_j, Y_j)_{j=1}^n$ and by $\mathbb{E}_{f,P}$ the corresponding expectation. Expectations involving only a uniform subset of $\mathcal{S}_{d,k}$ are written explicitly as $\mathbb{E}_{S\sim\nu_{d,k}}$. For a parameter class $\mathcal{A}$, define the minimax risk

$$R_n^\star(\mathcal{A}) = \inf_{\widehat{f}} \sup_{f\in\mathcal{A}} \sup_{P\in\mathcal{P}_{\sigma^2}} \mathbb{E}_{f,P}\left[\left\|\widehat{f} - f\right\|_2^2\right],$$

where the infimum ranges over all measurable estimators, based on $(S_j, Y_j)_{j=1}^n$, taking values in the space of real-valued functions on $\mathcal{S}_{d,k}$. Upper bounds will hold uniformly over $\mathcal{P}_{\sigma^2}$, whereas lower bounds may be established by restricting to the Gaussian submodel $N(0,\sigma^2)$. All results below are nonasymptotic and uniform in d, k, and n. They therefore also apply to sequences of experiments in which $d = d_n$ and $k = k_n$ vary with the sample size.

3.2 Inclusion filtration and harmonic levels

For $T \subseteq [d]$, define the inclusion indicator

$$h_T(S) = \mathbf{1}_{\{T\subseteq S\}}, \qquad S \in \mathcal{S}_{d,k}.$$

Thus, $h_T(S)$ records whether all items in T are present in the input subset S. Set $m = \min(k, d-k)$, and, for $0 \leqslant r \leqslant m$, define

$$U_r = \operatorname{span}\{h_T : T \subseteq [d],\ |T| = r\}.$$

The space U_r therefore contains the functions representable by inclusion effects of order r, together with all lower-order effects. To see the latter point, let $1 \leqslant r \leqslant m$ and $|T| = r-1$. For every $S \in \mathcal{S}_{d,k}$,

$$h_T(S) = \frac{1}{k-r+1} \sum_{i\in[d]\setminus T} h_{T\cup\{i\}}(S).$$

Indeed, if $T \not\subseteq S$, both sides vanish. If $T \subseteq S$, exactly $k-r+1$ elements $i \in S\setminus T$ make $T\cup\{i\}$ a subset of S. Consequently,

$$U_0 \subseteq U_1 \subseteq \cdots \subseteq U_m.$$

The dimensions of these spaces follow from the rank of the inclusion matrix. Let $M_{r,k}$ be the matrix whose rows are indexed by the r-subsets T, whose columns are indexed by the k-subsets S, and whose entries are

$$(M_{r,k})_{T,S} = h_T(S) = \mathbf{1}_{\{T\subseteq S\}}.$$

Each row is the vector of values of h_T over the whole domain $\mathcal{S}_{d,k}$. Hence, the row space of $M_{r,k}$ is naturally identified with U_r, and

$$\dim(U_r) = \operatorname{rank}(M_{r,k}).$$

Gottlieb's rank theorem states that $M_{r,k}$ has full row rank whenever $r \leqslant k$ and $r+k \leqslant d$ (Gottlieb, 1966). Both conditions hold for $0 \leqslant r \leqslant m$, and therefore

$$\dim(U_r) = \binom{d}{r}, \qquad 0 \leqslant r \leqslant m.$$

At the final level,

$$\dim(U_m) = \binom{d}{m} = \binom{d}{k} = N,$$

where the second equality follows from $m \in \{k, d-k\}$. Since a function on $\mathcal{S}_{d,k}$ is determined by its N values, the full space $L^2(\mathcal{S}_{d,k}, \nu_{d,k})$ also has dimension N. Hence,

$$U_m = L^2(\mathcal{S}_{d,k}, \nu_{d,k}).$$

To isolate the new component appearing at each order, set $U_{-1} = \{0\}$ and define

$$V_r = U_r \cap U_{r-1}^{\perp}, \qquad 0 \leqslant r \leqslant m.$$

Because the inclusion spaces are nested,

$$U_r = U_{r-1} \oplus^{\perp} V_r.$$

Iterating this identity and using $U_m = L^2(\mathcal{S}_{d,k}, \nu_{d,k})$ gives

$$L^2(\mathcal{S}_{d,k}, \nu_{d,k}) = V_0 \oplus^{\perp} \cdots \oplus^{\perp} V_m, \qquad \dim(V_r) = \binom{d}{r} - \binom{d}{r-1}, \tag{1}$$

with the convention $\binom{d}{-1} = 0$. Let P_r denote the orthogonal projector onto V_r. Since $V_r = U_r \cap U_{r-1}^{\perp}$, the function $P_r f$ is the part of f that lies in U_r and is orthogonal to all lower-order inclusion effects. We call $P_r f$ the order-r harmonic component of f. As an orthogonal projection, it is uniquely determined by f and does not depend on any choice of coefficients in a representation using the indicators h_T.

We next connect this decomposition with the one-exchange geometry of the Johnson graph. For $S \in \mathcal{S}_{d,k}$, $a \in S$, and $b \notin S$, write

$$S^{a \to b} = (S \setminus \{a\}) \cup \{b\}.$$

Let S_0 denote the current subset. Conditional on $S_0 = S$, choose A uniformly from the k elements of S and, independently, choose B uniformly from the $d-k$ elements of $[d] \setminus S$. Set $S_1 = S^{A \to B}$. The Markov operator of this one-exchange walk is therefore

$$Kf(S) := \mathbb{E}[f(S_1) \mid S_0 = S] = \frac{1}{k(d-k)} \sum_{a \in S} \sum_{b \notin S} f(S^{a \to b}).$$

Equivalently, K is the normalized adjacency operator of the Johnson graph $J(d,k)$ (Brouwer et al., 1989).

For $S, S' \in \mathcal{S}_{d,k}$, let

$$q(S, S') = \mathbb{P}(S_1 = S' \mid S_0 = S) = \frac{1}{k(d-k)} \mathbf{1}_{\{|S \cap S'| = k-1\}}.$$

Thus, $q(S, S')$ is $1/\{k(d-k)\}$ when S' is obtained from S by one exchange and is zero otherwise. The Markov operator can equivalently be written as

$$Kf(S) = \sum_{S' \in \mathcal{S}_{d,k}} q(S, S') f(S').$$

The uniform measure $\nu_{d,k}$ is reversible for this walk. Indeed, every exchange $S \to S'$ has the equally likely reverse exchange $S' \to S$, so $q(S, S') = q(S', S)$. Since $\nu_{d,k}$ is uniform, it follows that

$$\nu_{d,k}(S) q(S, S') = \nu_{d,k}(S') q(S', S), \qquad S, S' \in \mathcal{S}_{d,k}.$$

This detailed-balance identity is precisely the reversibility condition. It implies that K is self-adjoint with respect to $\langle \cdot, \cdot \rangle$:

$$\langle f, Kg \rangle = \langle Kf, g \rangle, \qquad f, g \in L^2(\mathcal{S}_{d,k}, \nu_{d,k}).$$

The classical spectrum of the Johnson graph shows that the harmonic spaces in (1) diagonalize this operator (Delsarte, 1973; Bannai and Ito, 1984; Brouwer et al., 1989). The precise correspondence, together with the associated Dirichlet form identity, is recorded in the following proposition.

Proposition 3.1 (Johnson spectrum)**.** *For $0 \leqslant r \leqslant m$ and $f \in V_r$,*

$$Kf = \theta_{r,d,k} f, \qquad \theta_{r,d,k} = 1 - \frac{r(d-r+1)}{k(d-k)}. \tag{2}$$

Moreover, every $f \in L^2(\mathcal{S}_{d,k}, \nu_{d,k})$ satisfies

$$\langle f, (I-K)f \rangle = \frac{1}{2k(d-k)} \mathbb{E}_{S \sim \nu_{d,k}} \left[\sum_{a \in S} \sum_{b \notin S} \{ f(S) - f(S^{a \to b}) \}^2 \right]. \tag{3}$$

Since $r \mapsto r(d-r+1)$ is strictly increasing on $\{0, \dots, m\}$, the eigenvalues $\theta_{r,d,k}$ of K are strictly decreasing with r, whereas the eigenvalues

$$1 - \theta_{r,d,k} = \frac{r(d-r+1)}{k(d-k)}$$

of $I - K$ are strictly increasing. Together with the complete decomposition (1), the proposition therefore shows that V_r is precisely the eigenspace of K associated with $\theta_{r,d,k}$. Equation (3) gives this spectral ordering a local interpretation: the quadratic form $\langle f, (I-K)f \rangle$ is one half of the mean squared change in f under a uniformly chosen exchange. Among functions of unit L^2-norm, higher harmonic levels correspond to larger eigenvalues of $I - K$ and hence to greater mean squared variation across neighboring subsets.

The harmonic projections also admit a canonical subset-indexed tight frame that respects permutations of the ground set. This basis-free, permutation-equivariant representation may be of independent interest. Since it is not needed for the statistical risk analysis, its explicit construction and proof are deferred to Section B.2 of the Supplementary Material.

4 Smoothness, projection, and minimax risk

4.1 The intrinsic Johnson–Sobolev scale

Proposition 3.1 shows that the quadratic form associated with $I - K$ measures the variation of a function under a single admissible exchange. Thus, $I - K$ is the normalized random-walk Laplacian of the Johnson graph. We rescale it by defining

$$L_J = \frac{k(d-k)}{d}(I - K). \tag{4}$$

The normalization is chosen so that the smallest nonzero eigenvalue of L_J is equal to one. The scaling factor $k(d-k)/d$ is unchanged when k is replaced by $d-k$, consistently with the symmetry between a subset and its complement.

By Proposition 3.1, L_J acts on V_r as multiplication by

$$\gamma_{r,d} = \frac{r(d-r+1)}{d}, \qquad 0 \leqslant r \leqslant m.$$

In particular,

$$\gamma_{0,d} = 0, \qquad \gamma_{1,d} = 1.$$

Moreover,

$$\gamma_{r+1,d} - \gamma_{r,d} = \frac{d-2r}{d} > 0, \qquad 0 \leqslant r < m,$$

because $m \leqslant d/2$. Hence, the eigenvalues increase strictly with the harmonic order.

Since L_J acts on V_r by multiplication by $\gamma_{r,d}$, the harmonic decomposition of f gives

$$L_J f = \sum_{r=0}^{m} \gamma_{r,d} P_r f = \sum_{r=1}^{m} \gamma_{r,d} P_r f,$$

where the second equality uses $\gamma_{0,d} = 0$. More generally, the spectral functional calculus defines, for $s > 0$,

$$L_J^s f = \sum_{r=1}^{m} \gamma_{r,d}^s P_r f.$$

Equivalently, L_J^s vanishes on V_0 and acts by multiplication by $\gamma_{r,d}^s$ on V_r, for $1 \leqslant r \leqslant m$.

Definition 4.1 (Johnson–Sobolev energy)**.** For $s > 0$ and $f \in L^2(\mathcal{S}_{d,k}, \nu_{d,k})$, define

$$I_{J,d,k}^{(2,s)}(f) := \langle f, L_J^s f \rangle . \tag{5}$$

Equivalently,

$$I_{J,d,k}^{(2,s)}(f) = \sum_{r=1}^{m} \gamma_{r,d}^s \, \|P_r f\|_2^2 .$$

The superscript 2 indicates that the energy is quadratic in the $L^2(\nu_{d,k})$ norm, while s is the spectral smoothness order. Since $\gamma_{1,d} = 1$, the first nonconstant harmonic level has unit weight for every $s > 0$. At $s = 1$, the energy has the local exchange representation

$$I_{J,d,k}^{(2,1)}(f) = \frac{1}{2d} \mathbb{E}_{S \sim \nu_{d,k}} \left[\sum_{a \in S} \sum_{b \notin S} \{ f(S) - f(S^{a \to b}) \}^2 \right] .$$

Indeed, this identity follows by combining (4) with the Dirichlet representation in Proposition 3.1. Thus, at smoothness order $s = 1$, the Johnson–Sobolev energy measures the aggregate squared variation of f across admissible exchanges, averaged over the initial subset and normalized by $1/(2d)$.

The entire scale is invariant under complementation. More precisely, if

$$f^c(A) = f([d] \setminus A), \qquad A \in \mathcal{S}_{d,d-k},$$

then complementation is a measure-preserving bijection between $\mathcal{S}_{d,k}$ and $\mathcal{S}_{d,d-k}$ that maps admissible exchanges bijectively and hence intertwines the two Johnson walks. Since the scaling factor $k(d-k)/d$ is unchanged, the same relation holds for the corresponding operators L_J and their spectral powers. Thus,

$$I^{(2,s)}_{J,d,d-k}(f^{c}) = I^{(2,s)}_{J,d,k}(f).$$

Because the constant component $P_0 f$ does not appear in (5), the energy is a squared seminorm rather than a squared norm. Its square root is unchanged when a constant is added to f, and it vanishes exactly when f is constant.

Finally, for $0 \leqslant D < m$, the monotonicity of the spectral weights gives

$$\begin{aligned} I^{(2,s)}_{J,d,k}(f) &\geqslant \sum_{r>D} \gamma^{s}_{r,d} \, \|P_r f\|_2^2 \\ &\geqslant \gamma^{s}_{D+1,d} \sum_{r>D} \|P_r f\|_2^2 . \end{aligned}$$

Therefore,

$$\sum_{r>D} \|P_r f\|_2^2 \leqslant \frac{I^{(2,s)}_{J,d,k}(f)}{\gamma^{s}_{D+1,d}}. \tag{6}$$

This is the approximation inequality used in the statistical analysis below.

4.2 Projection estimation

For $s, B > 0$, consider the centered ellipsoid

$$\mathcal{E}^{s}_{d,k}(B) = \left\{ f \in L^2(\mathcal{S}_{d,k}, \nu_{d,k}) : P_0 f = 0,\ I^{(2,s)}_{J,d,k}(f) \leqslant B \right\}. \tag{7}$$

Because $\gamma_{r,d} \geqslant 1$ for $r \geqslant 1$,

$$\|f\|_2^2 \leqslant B, \qquad f \in \mathcal{E}^{s}_{d,k}(B). \tag{8}$$

Since $V_0 = \operatorname{span}\{\mathbf{1}\}$, the constant component of any function f is

$$P_0 f = \mathbb{E}_{S \sim \nu_{d,k}}[f(S)]\, \mathbf{1}.$$

Centering therefore isolates the nonconstant component controlled by the energy. More precisely, suppose that the response can be written as

$$f = c + g, \qquad c = \mathbb{E}_{S \sim \nu_{d,k}}[f(S)], \qquad g \in \mathcal{E}^{s}_{d,k}(B).$$

Since g and the noise are centered, $\overline{Y} = n^{-1} \sum_{j=1}^{n} Y_j$ is unbiased for c, and

$$\mathbb{E}_{f,P}[(\overline{Y} - c)^2] = \frac{\|g\|_2^2 + \operatorname{Var}_P(\varepsilon)}{n} \leqslant \frac{B + \sigma^2}{n}.$$

Thus an unknown intercept can be handled at the parametric rate, for example on an independent sample split. No intercept is included in the minimax class (7).

For $0 \leqslant D \leqslant m$, let

$$W_D = V_1 \oplus^{\perp} \cdots \oplus^{\perp} V_D, \qquad P^{\circ}_{\leqslant D} = P_1 + \cdots + P_D,$$

with $W_0 = \{0\}$ and $P^\circ_{\leqslant 0} = 0$. The constant level V_0 is omitted because the response class $\mathcal{E}^s_{d,k}(B)$ is centered. The dimension of W_D is

$$p_D := \dim(W_D) = \sum_{r=1}^{D} \dim(V_r) = \binom{d}{D} - 1,$$

including $p_0 = 0$.

Choose any orthonormal basis $e_1, \ldots, e_{p_D}$ of W_D. The coefficient of the population projection $P^\circ_{\leqslant D} f$ in direction e_ℓ is

$$\theta_\ell = \langle f, e_\ell \rangle = \mathbb{E}_{S \sim \nu_{d,k}}[f(S) e_\ell(S)].$$

Since the noise is centered and independent of the design,

$$\mathbb{E}_{f,P}[Y_j e_\ell(S_j)] = \theta_\ell.$$

The sample average

$$\widehat{\theta}_\ell = \frac{1}{n} \sum_{j=1}^{n} Y_j e_\ell(S_j)$$

is therefore an unbiased estimator of the population projection coefficient θ_ℓ. This leads to the empirical projection estimator

$$\widehat{f}_D^{\text{proj}} = \sum_{\ell=1}^{p_D} \widehat{\theta}_\ell e_\ell. \tag{9}$$

When $D = 0$, the sum in (9) is empty, and hence $\widehat{f}_0^{\text{proj}} = 0$. Although this definition uses an orthonormal basis, the resulting estimator depends only on W_D. Indeed, substituting the definition of $\widehat{\theta}_\ell$ gives

$$\begin{aligned}
\widehat{f}_D^{\text{proj}}(S) &= \sum_{\ell=1}^{p_D} \left\{ \frac{1}{n} \sum_{j=1}^{n} Y_j e_\ell(S_j) \right\} e_\ell(S) \\
&= \frac{1}{n} \sum_{j=1}^{n} Y_j \mathcal{K}^\circ_D(S_j, S),
\end{aligned}$$

where

$$\mathcal{K}^\circ_D(S, S') = \sum_{\ell=1}^{p_D} e_\ell(S) e_\ell(S')$$

is the orthogonal projection kernel of W_D. This kernel is basis-independent. More precisely, if $\widetilde{e}_1, \ldots, \widetilde{e}_{p_D}$ is another orthonormal basis, then the two bases are related by an orthogonal matrix O. Writing

$$e(S) = \big(e_1(S), \ldots, e_{p_D}(S)\big)^\top,$$

we have $\widetilde{e}(S) = O e(S)$, and hence

$$\widetilde{e}(S)^\top \widetilde{e}(S') = e(S)^\top O^\top O e(S') = e(S)^\top e(S').$$

Thus both the kernel and the estimator are determined by the space W_D, rather than by the basis used to represent it.

Remark 4.2. Section B.2 of the Supplementary Material gives an alternative permutation-equivariant representation of the same estimator in terms of a canonical tight frame indexed by subsets.

The full-table Johnson transform of Iglesias and Natale (2022) computes harmonic projections from an N-dimensional array indexed by all vertices of $\mathcal{S}_{d,k}$. The empirical coefficients can be obtained by applying this transform to the N-vector whose entry at S is $\frac{N}{n}\sum_{j:S_j=S} Y_j$, with value zero when S is unobserved. Applying the full-table transform to this vector still requires computation on the entire domain and therefore scales with $N = \binom{d}{k}$, not with the sample size n or the retained dimension p_D. Developing a sample-based low-order transform is a separate computational question.

Define the worst-case approximation bound

$$b_D(s,B) = \begin{cases} \dfrac{B}{\gamma^s_{D+1,d}}, & 0 \leqslant D < m, \\ 0, & D = m. \end{cases}$$

Proposition 4.3 (Projection risk)**.** *For every $s, B > 0$ and $0 \leqslant D \leqslant m$,*

$$\sup_{f \in \mathcal{E}^s_{d,k}(B)} \sup_{P \in \mathcal{P}_{\sigma^2}} \mathbb{E}_{f,P}\left[\left\|\widehat{f}^{\mathrm{proj}}_D - f\right\|_2^2\right] \leqslant \left(\sigma^2 + B\right)\frac{p_D}{n} + b_D(s,B). \tag{10}$$

The two terms on the right-hand side of (10) bound distinct contributions: $(\sigma^2+B)p_D/n$ bounds the coefficient-estimation term, whereas $b_D(s,B)$ bounds the approximation term contributed by the omitted harmonic levels. The key fact in the variance calculation is that the projection kernel has constant diagonal:

$$\mathcal{K}^\circ_D(S,S) = p_D, \qquad S \in \mathcal{S}_{d,k}. \tag{11}$$

This identity follows from the permutation symmetry of W_D; its proof is given together with that of Proposition 4.3. It yields

$$\mathbb{E}_{f,P}\left[\left\|\widehat{f}^{\mathrm{proj}}_D - P^\circ_{\leqslant D} f\right\|_2^2\right] \leqslant (\sigma^2+B)\frac{p_D}{n}.$$

The estimation component $\widehat{f}^{\mathrm{proj}}_D - P^\circ_{\leqslant D} f$ belongs to W_D, whereas the approximation component $f - P^\circ_{\leqslant D} f$ belongs to $W_D^\perp$. Their squared norms therefore add exactly, and (6) bounds the latter by $b_D(s,B)$.

The contribution Bp_D/n is not observation noise. It arises because the population projection coefficients are estimated by averages over randomly sampled subsets, which fluctuate with the design even when the responses are noiseless.

4.3 Nonasymptotic minimax risk

For $1 \leqslant D \leqslant m$, define

$$a_D(s,B) = \frac{B}{\gamma^s_{D,d}}, \qquad v_D(n,\sigma^2) = \frac{\sigma^2 p_D}{n}.$$

The quantity $a_D(s, B)$ is the squared radius of an L^2-ball centered at zero in W_D and contained in $\mathcal{E}^s_{d,k}(B)$. Indeed, for every $g \in W_D$,

$$\begin{aligned} I^{(2,s)}_{J,d,k}(g) &= \sum_{r=1}^{D} \gamma^s_{r,d} \, \|P_r g\|_2^2 \\ &\leqslant \gamma^s_{D,d} \sum_{r=1}^{D} \|P_r g\|_2^2 = \gamma^s_{D,d} \, \|g\|_2^2 . \end{aligned}$$

Consequently,

$$\{g \in W_D : \|g\|_2^2 \leqslant a_D(s, B)\} \subseteq \mathcal{E}^s_{d,k}(B).$$

The second quantity, $v_D(n, \sigma^2)$, is the usual parametric risk scale for estimating p_D coefficients from n observations under Gaussian noise with variance σ^2. A standard finite-dimensional testing argument compares the squared radius $a_D(s, B)$ with the p_D-dimensional noise cost $v_D(n, \sigma^2)$ and leads to the scale

$$\min\{a_D(s, B), v_D(n, \sigma^2)\}.$$

We therefore define the spectral minimax scale by

$$\mathfrak{R}_{n,d,k}(s, B, \sigma^2) = \max_{1 \leqslant D \leqslant m} \min\left\{a_D(s, B), v_D(n, \sigma^2)\right\}. \tag{12}$$

With the convention $v_0 = 0$, a deterministic crossing argument, given in the proof of Theorem 4.4, yields

$$\begin{aligned} \mathfrak{R}_{n,d,k}(s, B, \sigma^2) &\leqslant \min_{0 \leqslant D \leqslant m} \left\{v_D(n, \sigma^2) + b_D(s, B)\right\} \\ &\leqslant \min_{0 \leqslant D \leqslant m} \left\{(\sigma^2 + B)\frac{p_D}{n} + b_D(s, B)\right\} \\ &\leqslant 2\left(1 + \frac{B}{\sigma^2}\right) \mathfrak{R}_{n,d,k}(s, B, \sigma^2). \end{aligned}$$

The first minimum is the ideal approximation–observation-noise tradeoff, whereas the second is the optimized right-hand side of the empirical-projection bound. Thus empirical projection matches the spectral scale with constants uniform in the parameters whenever B/σ^2 is bounded.

Theorem 4.4 (Uniform nonasymptotic minimax bounds)**.** *There exists a universal constant $c > 0$ such that, for every $d \geqslant 2$, $1 \leqslant k \leqslant d-1$, $n \geqslant 2$, and $s, B, \sigma^2 > 0$,*

$$c\,\mathfrak{R}_{n,d,k}(s, B, \sigma^2) \leqslant R^\star_n(\mathcal{E}^s_{d,k}(B)) \leqslant 2\left(1 + \frac{B}{\sigma^2}\right) \mathfrak{R}_{n,d,k}(s, B, \sigma^2). \tag{13}$$

Consequently, for every fixed $C_{\mathrm{snr}} > 0$, the condition

$$\frac{B}{\sigma^2} \leqslant C_{\mathrm{snr}}$$

implies

$$c\,\mathfrak{R}_{n,d,k}(s, B, \sigma^2) \leqslant R^\star_n(\mathcal{E}^s_{d,k}(B)) \leqslant 2(1 + C_{\mathrm{snr}})\mathfrak{R}_{n,d,k}(s, B, \sigma^2).$$

This comparison is nonasymptotic and holds uniformly over all d, k, n, s, B, σ^2 satisfying the displayed condition.

For the lower bound, we restrict the experiment to the p_D-dimensional ball in W_D described above and apply a finite-dimensional testing argument in the Gaussian noise submodel. Maximizing the resulting bound over D gives the spectral scale $\mathfrak{R}_{n,d,k}(s, B, \sigma^2)$.

For the upper bound, Proposition 4.3 yields a bias–variance sum indexed by the cutoff D. A deterministic crossing argument compares its minimum with the max–min expression in (12). The upper bound is an oracle minimax statement. For each fixed D, the estimator $\widehat{f}_D^{\mathrm{proj}}$ itself does not depend on s, B, or σ^2. However, the cutoff used to obtain the bound may, for example, be chosen as

$$D_{\mathrm{or}} \in \arg\min_{0 \leqslant D \leqslant m} \left\{ \frac{\sigma^2 p_D}{n} + b_D(s, B) \right\},$$

and therefore depends on the parameters defining the response and noise classes. As usual in the definition of minimax risk, these class parameters are treated as known. Data-driven adaptation to unknown s, B, or σ^2 is not addressed here.

The condition $B/\sigma^2 \leqslant C_{\mathrm{snr}}$ is not needed for either inequality in (13). It is used only to make the upper and lower bounds comparable with constants that are uniform over the stated parameter range. The next section examines the random-design effects that become relevant when the signal scale is large relative to the observation-noise variance.

It is worth emphasizing that the nonasymptotic formulation permits d and k themselves to vary with n. There is therefore no single rate as a function of n alone without specifying their joint growth. Along any such sequence, Proposition 4.3 shows that, for fixed s, B, and σ^2, the worst-case risk of $\widehat{f}_{D_n}^{\mathrm{proj}}$ over $\mathcal{E}_{d,k}^s(B)$ and $\mathcal{P}_{\sigma^2}$ tends to zero if one can choose cutoffs $0 \leqslant D_n \leqslant m$ such that

$$\frac{p_{D_n}}{n} \longrightarrow 0 \qquad \text{and} \qquad b_{D_n}(s, B) \longrightarrow 0.$$

The first condition requires the sample size to dominate the dimension of the retained interaction space, while the second requires its omitted harmonic tail to vanish.

5 Random-design uncertainty and stable least squares

The preceding section revealed that random sampling affects estimation through more than observation noise. We now study this design uncertainty directly. We first establish conditions under which the empirical Gram matrix is well conditioned and least squares is stable. We then quantify the information lost when sampled evaluations leave part of the fitted space unidentified, construct a centered completion estimator, and combine these results in rank-aware minimax bounds.

5.1 Constant leverage and Gram concentration

To remove the signal-dependent fluctuation of empirical projection, least squares must be stable over the whole fitted space W_D. We therefore study whether the empirical squared norm induced by the design is uniformly comparable to the population $L^2(\nu_{d,k})$-norm on W_D.

Fix $1 \leqslant D \leqslant m$, and let $e_1, \ldots, e_{p_D}$ be an orthonormal basis of W_D. For $S \in \mathcal{S}_{d,k}$, define

$$x_D(S) = (e_1(S), \ldots, e_{p_D}(S))^\top,$$

and define the empirical Gram matrix by

$$\widehat{G}_D = \frac{1}{n}\sum_{j=1}^{n} x_D(S_j)x_D(S_j)^\top.$$

Orthonormality gives

$$\mathbb{E}_{S\sim\nu_{d,k}}[x_D(S)x_D(S)^\top] = I_{p_D},$$

so the population Gram matrix is the identity.

In general, the leverage score

$$\ell_D(S) := \|x_D(S)\|_2^2$$

may vary substantially with S, and concentration of the empirical Gram matrix is governed by its largest value

$$L_D := \max_{S\in\mathcal{S}_{d,k}} \ell_D(S).$$

In the present Johnson setting, however, permutation invariance of W_D and transitivity of the permutation action on $\mathcal{S}_{d,k}$ imply, as established in (11), that the leverage score is constant:

$$\ell_D(S) = p_D, \qquad S \in \mathcal{S}_{d,k}.$$

Consequently, $L_D = p_D$. This is the smallest possible value of the maximum leverage for a p_D-dimensional subspace of $L^2(\mathcal{S}_{d,k}, \nu_{d,k})$. Indeed, if $h_1, \ldots, h_{p_D}$ is any orthonormal basis of such a subspace, then its leverage function

$$\ell(S) = \sum_{\ell=1}^{p_D} h_\ell(S)^2$$

satisfies

$$\mathbb{E}_{S\sim\nu_{d,k}}[\ell(S)] = \sum_{\ell=1}^{p_D} \|h_\ell\|_2^2 = p_D.$$

Hence,

$$\max_{S\in\mathcal{S}_{d,k}} \ell(S) \geqslant \mathbb{E}_{S\sim\nu_{d,k}}[\ell(S)] = p_D.$$

Thus, in the present Johnson setting, the leverage is perfectly balanced over the fixed-cardinality domain. For a square matrix M, let $\|M\|_{\mathrm{op}}$ denote its Euclidean operator norm. Moreover, for every $S \in \mathcal{S}_{d,k}$,

$$\|x_D(S)x_D(S)^\top\|_{\mathrm{op}} = \|x_D(S)\|_2^2 = \ell_D(S) = p_D.$$

This uniform bound on the rank-one matrices entering $\widehat{G}_D$ is the key input to the matrix concentration argument below.

Define

$$\mathcal{A}_D = \left\{ \|\widehat{G}_D - I_{p_D}\|_{\mathrm{op}} \leqslant \frac{1}{2} \right\}.$$

Proposition 5.1 (Johnson Gram concentration)**.** *With*

$$c_0 = \frac{3}{2}\log\left(\frac{3}{2}\right) - \frac{1}{2} > 0,$$

one has

$$\mathbb{P}_{\nu_{d,k}^{\otimes n}}(\mathcal{A}_D^c) \leqslant 2p_D \exp\left(-c_0 \frac{n}{p_D}\right). \tag{14}$$

To interpret $\mathcal{A}_D$, write

$$g = \sum_{\ell=1}^{p_D} u_\ell e_\ell \qquad \text{for some } u \in \mathbb{R}^{p_D}.$$

Then

$$g(S) = u^\top x_D(S), \qquad \|g\|_2^2 = \|u\|_2^2, \qquad \frac{1}{n}\sum_{j=1}^{n} g(S_j)^2 = u^\top \widehat{G}_D u.$$

Since $\widehat{G}_D$ is symmetric, the definition of $\mathcal{A}_D$ implies

$$\max_{1 \leqslant \ell \leqslant p_D} |\lambda_\ell(\widehat{G}_D) - 1| \leqslant \frac{1}{2}.$$

Hence every eigenvalue of $\widehat{G}_D$ lies in $[1/2, 3/2]$, and the Rayleigh quotient gives, for every $g \in W_D$,

$$\frac{1}{2}\|g\|_2^2 \leqslant \frac{1}{n}\sum_{j=1}^{n} g(S_j)^2 \leqslant \frac{3}{2}\|g\|_2^2. \tag{15}$$

In particular, $\lambda_{\min}(\widehat{G}_D) \geqslant 1/2 > 0$, so $\widehat{G}_D$ is invertible and

$$\|\widehat{G}_D^{-1}\|_{\mathrm{op}} = \frac{1}{\lambda_{\min}(\widehat{G}_D)} \leqslant 2.$$

It follows from (14) that, for every $0 < \delta < 1$,

$$n \geqslant \frac{p_D}{c_0} \log\left(\frac{2p_D}{\delta}\right)$$

is sufficient for (15) to hold with probability at least $1 - \delta$. By contrast, $n \geqslant p_D$ is only a necessary algebraic condition for invertibility: if $n < p_D$, then $\widehat{G}_D$, being a sum of n rank-one matrices, has rank at most $n < p_D$. The event $\mathcal{A}_D$ provides the stronger quantitative control needed for the least-squares analysis that follows.

5.2 Stable least squares

Throughout this subsection, fix a deterministic cutoff $1 \leqslant D \leqslant m$. This cutoff determines both the fitted space W_D and the stability event $\mathcal{A}_D$; no data-driven selection of D is considered here. On $\mathcal{A}_D$, the empirical and population norms are uniformly comparable over W_D, and the empirical Gram matrix is well conditioned. Least squares can therefore be applied without the error amplification caused by a nearly singular Gram matrix. Define the stable least-squares estimator by

$$\widehat{f}_D^{\mathrm{LS}} = \begin{cases} \operatorname*{arg\,min}_{g \in W_D} \dfrac{1}{n}\displaystyle\sum_{j=1}^{n}\{Y_j - g(S_j)\}^2, & \text{on } \mathcal{A}_D, \\ 0, & \text{on } \mathcal{A}_D^c. \end{cases}$$

On $\mathcal{A}_D$, the minimizer is unique because $\lambda_{\min}(\widehat{G}_D) \geqslant 1/2$. Since $\mathcal{A}_D$ depends only on the observed design points, this fallback rule is data-driven and prevents an ill-conditioned Gram matrix from producing an arbitrarily large estimate.

In the orthonormal basis $e_1, \dots, e_{p_D}$, the coefficient vector of the estimator on $\mathcal{A}_D$ is

$$\widehat{\theta}_D^{\mathrm{LS}} = \widehat{G}_D^{-1} \frac{1}{n} \sum_{j=1}^{n} Y_j x_D(S_j). \tag{16}$$

Although this formula uses a basis, the resulting function is basis-independent: on $\mathcal{A}_D$, it is the unique minimizer of a criterion defined intrinsically over the space W_D.

For a general response f, write

$$g_D = P^{\circ}_{\leqslant D} f, \qquad h_D = f - g_D,$$

and let $\theta_D \in \mathbb{R}^{p_D}$ be the coefficient vector of g_D, so that

$$g_D(S) = x_D(S)^\top \theta_D.$$

Since $h_D \perp W_D$,

$$\mathbb{E}_{S\sim\nu_{d,k}}[x_D(S) h_D(S)] = 0.$$

Thus, the random vectors $x_D(S_j)h_D(S_j)$ are centered; this fact controls the contribution of the omitted component in the risk calculation. Substituting $Y_j = g_D(S_j) + h_D(S_j) + \varepsilon_j$ into (16) gives, on $\mathcal{A}_D$,

$$\widehat{\theta}_D^{\mathrm{LS}} - \theta_D = \widehat{G}_D^{-1} \left\{ \frac{1}{n} \sum_{j=1}^{n} x_D(S_j) h_D(S_j) + \frac{1}{n} \sum_{j=1}^{n} x_D(S_j) \varepsilon_j \right\}. \tag{17}$$

Thus the component g_D already contained in the fitted space cancels exactly from the normal equations. The coefficient error is generated only by observation noise and by the empirical correlation between the omitted component h_D and W_D.

Theorem 5.2 (Stable least-squares risk)**.** *For every $s, B > 0$ and $1 \leqslant D \leqslant m$,*

$$\begin{aligned} &\sup_{f\in\mathcal{E}^s_{d,k}(B)} \sup_{P\in\mathcal{P}_{\sigma^2}} \mathbb{E}_{f,P}\left[\|\widehat{f}_D^{\mathrm{LS}} - f\|_2^2\right] \\ &\quad\leqslant \left(1 + 4\frac{p_D}{n}\right) b_D(s,B) + 4\sigma^2 \frac{p_D}{n} + B \min\left\{1,\, 2p_D \exp\left(-c_0 \frac{n}{p_D}\right)\right\}. \end{aligned} \tag{18}$$

The bound holds for every deterministic cutoff D; optimizing its right-hand side over D would again constitute an oracle choice.

The three terms in (18) have distinct origins. The quantity $b_D(s,B)$ bounds the approximation error $\|h_D\|_2^2$, while the additional factor $4p_D/n$ accounts for the empirical correlation of h_D with the fitted space. The term $4\sigma^2 p_D/n$ is the variance caused by observation noise. Finally, on $\mathcal{A}_D^c$ the estimator is zero, so its loss is bounded by $\|f\|_2^2 \leqslant B$; Gram concentration produces the last term of the bound.

Equation (17) shows that, on $\mathcal{A}_D$, the estimation error does not involve the fitted component g_D: only the omitted component h_D and the observation noise remain. Thus, unlike empirical projection, least squares pays no random-design variance for the component already contained in W_D. In particular, if $h_D = 0$ and the observations are noiseless, then $\widehat{f}_D^{\mathrm{LS}} = f$ on $\mathcal{A}_D$.

To make the resulting bound explicit, suppose that $n \geqslant p_D$ and, for some $a > 0$,

$$n \geqslant \frac{a+1}{c_0}\, p_D \log(2p_D).$$

Then

$$1 + 4\frac{p_D}{n} \leqslant 5$$

and

$$2p_D \exp\left(-c_0 \frac{n}{p_D}\right) \leqslant p_D^{-a}.$$

Consequently, (18) becomes

$$\sup_{f \in \mathcal{E}^s_{d,k}(B)} \sup_{P \in \mathcal{P}_{\sigma^2}} \mathbb{E}_{f,P}\left[\|\widehat{f}_D^{\mathrm{LS}} - f\|_2^2\right] \leqslant 5b_D(s, B) + 4\sigma^2 \frac{p_D}{n} + Bp_D^{-a}.$$

Thus, in the stable window, the signal-dependent variance Bp_D/n appearing in the empirical-projection bound is absent. The signal contributes only through approximation error and the Gram-instability term.

The preceding least-squares guarantee is useful only when $\mathcal{A}_D$ has high probability. When the fitted space is too large relative to the sampled design, the empirical Gram matrix may even lose rank. Distinct functions in W_D can then agree at every sampled subset, so part of the fitted response is not identifiable from the observations. We now quantify this intrinsic information loss. At the full level, rank deficiency admits a concrete interpretation in terms of unobserved subsets, which will motivate the centered completion estimator.

5.3 Unidentified directions and incomplete coverage

Proposition 5.1 ensures that, on $\mathcal{A}_D$, the sampled evaluation map is injective and well conditioned on W_D. We now study the complementary situation in which the sampled evaluations do not uniquely identify functions in W_D, and quantify the part of this space that remains invisible to the design. This loss of information persists even in the absence of observation noise.

Let $\mathbf{X}_D$ be the $n \times p_D$ matrix whose j-th row is $x_D(S_j)^\top$. If

$$g = \sum_{\ell=1}^{p_D} u_\ell e_\ell \in W_D,$$

then the vector of its sampled values is

$$\big(g(S_1), \ldots, g(S_n)\big)^\top = \mathbf{X}_D u.$$

The sampled values depend on the coefficient vector u only through $\mathbf{X}_D u$. Consequently, two coefficient vectors that differ by an element of $\ker(\mathbf{X}_D)$ produce exactly the same noiseless observations. Thus, $\operatorname{rank}(\mathbf{X}_D)$ is the number of linearly independent measurements of the coefficient vector supplied by the design, whereas

$$p_D - \operatorname{rank}(\mathbf{X}_D) = \dim \ker(\mathbf{X}_D)$$

is the dimension of the subspace of coefficient perturbations that remain invisible. We measure the expected proportion of such unidentified directions by

$$\rho_{D,n} = \frac{1}{p_D} \mathbb{E}_{\nu_{d,k}^{\otimes n}} \left[p_D - \operatorname{rank}(\mathbf{X}_D)\right], \qquad 1 \leqslant D \leqslant m. \tag{19}$$

This quantity does not depend on the chosen orthonormal basis: a change of basis multiplies $\mathbf{X}_D$ on the right by an invertible orthogonal matrix and therefore leaves its rank unchanged.

Proposition 5.3 (Rank-deficiency lower bound)**.** *For every $s, B, \sigma^2 > 0$,*

$$R_n^\star(\mathcal{E}_{d,k}^s(B)) \geqslant \max_{1\leqslant D\leqslant m} \frac{B}{\gamma_{D,d}^s}\rho_{D,n}. \tag{20}$$

Moreover, for every $1 \leqslant D \leqslant m$,

$$\rho_{D,n} \geqslant \left(1 - \frac{n}{p_D}\right)_+. \tag{21}$$

At the full level $D = m$, let

$$\mathcal{O}_n = \{S_1, \ldots, S_n\} \subseteq \mathcal{S}_{d,k}$$

denote the set of distinct observed subsets, and let $M_n = N - |\mathcal{O}_n|$ be the number of unobserved subsets. Then

$$\rho_{m,n} = \frac{\mathbb{E}_{\nu_{d,k}^{\otimes n}}[(M_n - 1)_+]}{N-1}. \tag{22}$$

Equivalently,

$$\rho_{m,n} = \frac{N(1-1/N)^n - 1 + \mathbb{P}_{\nu_{d,k}^{\otimes n}}(M_n = 0)}{N-1}. \tag{23}$$

Here,

$$\mathbb{P}_{\nu_{d,k}^{\otimes n}}(M_n = 0) = \sum_{j=0}^{N}(-1)^j \binom{N}{j}\left(1 - \frac{j}{N}\right)^n$$

by inclusion–exclusion; in particular, this probability is zero when $n < N$.

The mechanism behind the lower bound can be seen by restricting the noise law to the degenerate distribution δ_0, under which $\varepsilon = 0$ almost surely. This noiseless law belongs to $\mathcal{P}_{\sigma^2}$ for every $\sigma^2 > 0$. For a fixed D, consider a response drawn uniformly from the sphere

$$\left\{ g \in W_D : \|g\|_2^2 = \frac{B}{\gamma_{D,d}^s} \right\}.$$

This sphere is contained in the Johnson–Sobolev ellipsoid. Recall that

$$\mathbb{R}^{p_D} = \mathrm{row}(\mathbf{X}_D) \oplus^\perp \ker(\mathbf{X}_D).$$

Conditional on the sampled subsets, the noiseless observations determine the component of the response's coefficient vector lying in the row space of $\mathbf{X}_D$, but they do not determine its component in $\ker(\mathbf{X}_D)$. Rotational symmetry of the spherical prior makes the conditional mean of this unidentified component equal to zero. Averaged over the spherical prior, its expected squared norm, conditional on the design, is

$$\frac{B}{\gamma_{D,d}^s}\,\frac{p_D - \mathrm{rank}(\mathbf{X}_D)}{p_D}.$$

Averaging over the design gives the term $B\rho_{D,n}/\gamma_{D,d}^s$ in (20).

At the full level $D = m$, this loss of information has the concrete coverage interpretation in (22). If M_n subsets are unobserved, then the centered functions supported on those subsets and vanishing at every observed input form a space of dimension $(M_n - 1)_+$. The subtraction of one comes from centering: once the response is known at all but one subset, its value at the last subset is determined by the zero-sum constraint. Hence the full-level rank deficiency measures the dimension of the entire unresolved part of the response, rather than the probability of a single incomplete-coverage event.

5.4 Centered completion

At the full level, the rank-deficiency obstruction suggests a natural estimator. Repeated observations of the same subset are first averaged, and the values at unobserved subsets are then chosen to satisfy the centering constraint.

For $S \in \mathcal{S}_{d,k}$, let

$$C_n(S) = \sum_{j=1}^{n} \mathbf{1}_{\{S_j = S\}}.$$

Equivalently, the set of observed subsets introduced above satisfies

$$\mathcal{O}_n = \{S \in \mathcal{S}_{d,k} : C_n(S) > 0\}.$$

For every $S \in \mathcal{O}_n$, define the corresponding cell mean by

$$\overline{Y}(S) = \frac{1}{C_n(S)} \sum_{j : S_j = S} Y_j.$$

Recall that

$$M_n = N - |\mathcal{O}_n| = \#\{S \in \mathcal{S}_{d,k} : C_n(S) = 0\}$$

is the number of subsets that are not observed.

If $M_n \geqslant 1$, define

$$\widehat{f}^{\,\mathrm{comp}}(S) = \begin{cases} \overline{Y}(S), & S \in \mathcal{O}_n, \\ -\dfrac{1}{M_n} \displaystyle\sum_{T \in \mathcal{O}_n} \overline{Y}(T), & S \notin \mathcal{O}_n. \end{cases} \tag{24}$$

Thus, all unobserved subsets receive the same value, chosen so that

$$\sum_{S \in \mathcal{S}_{d,k}} \widehat{f}^{\,\mathrm{comp}}(S) = 0.$$

Among all centered functions agreeing with the observed cell means, this choice has the smallest $L^2(\nu_{d,k})$-norm: subject to a fixed sum, the sum of the squared values assigned to the unobserved subsets is minimized when those values are equal.

If $M_n = 0$, every subset is observed, so there are no missing values that can be chosen to enforce centering. We instead subtract the average of the complete table of cell means:

$$\widehat{f}^{\,\mathrm{comp}}(S) = \overline{Y}(S) - \frac{1}{N} \sum_{T \in \mathcal{S}_{d,k}} \overline{Y}(T).$$

The resulting function is centered because its values sum to zero over $\mathcal{S}_{d,k}$. It is obtained from $S \mapsto \overline{Y}(S)$ by subtracting its uniform average and is therefore the closest centered function to $S \mapsto \overline{Y}(S)$ in $L^2(\nu_{d,k})$.

For every fixed $S \in \mathcal{S}_{d,k}$, the count $C_n(S)$ has distribution $\mathrm{Bin}(n, 1/N)$. Let $C \sim \mathrm{Bin}(n, 1/N)$, and set

$$\eta_{n,N} = \sum_{c=1}^{n} \frac{1}{c}\, \mathbb{P}(C = c). \tag{25}$$

Proposition 5.4 (Risk of centered completion)**.** *For every $s, B > 0$,*

$$\sup_{f \in \mathcal{E}^s_{d,k}(B)} \sup_{P \in \mathcal{P}_{\sigma^2}} \mathbb{E}_{f,P}\left[\left\|\widehat{f}^{\,\mathrm{comp}} - f\right\|_2^2\right] \leqslant B\rho_{m,n} + 2\sigma^2\eta_{n,N}, \tag{26}$$

where $\eta_{n,N}$ is defined in (25) and satisfies

$$\eta_{n,N} \leqslant \min\left\{1, \frac{2N}{n+1}\right\}. \tag{27}$$

To isolate the error caused by incomplete coverage, consider the noiseless version of the estimator. Let $\widehat{f}^{\,\mathrm{comp},0}$ denote the estimator obtained by applying the same completion rule to the exact observed values $f(S)$. If $M_n \geqslant 1$, then

$$\widehat{f}^{\,\mathrm{comp},0}(S) = \begin{cases} f(S), & S \in \mathcal{O}_n, \\ \dfrac{1}{M_n}\displaystyle\sum_{T \in \mathcal{U}_n} f(T), & S \in \mathcal{U}_n, \end{cases} \qquad \mathcal{U}_n = \mathcal{S}_{d,k} \setminus \mathcal{O}_n.$$

Indeed, since f is centered,

$$-\frac{1}{M_n}\sum_{T \in \mathcal{O}_n} f(T) = \frac{1}{M_n}\sum_{T \in \mathcal{U}_n} f(T).$$

If $M_n = 0$, every subset is observed and $\widehat{f}^{\,\mathrm{comp},0} = f$.

Let

$$\mathcal{H}_n = \{h \in W_m : h(S) = 0 \text{ for every } S \in \mathcal{O}_n\}.$$

Equivalently,

$$\mathcal{H}_n = \left\{h : \mathcal{S}_{d,k} \to \mathbb{R} : h(S) = 0 \text{ for } S \in \mathcal{O}_n, \quad \sum_{S \in \mathcal{U}_n} h(S) = 0\right\}.$$

This is the space of centered functions supported on the unobserved subsets. From the preceding formula,

$$\left(f - \widehat{f}^{\,\mathrm{comp},0}\right)(S) = \begin{cases} 0, & S \in \mathcal{O}_n, \\ f(S) - \dfrac{1}{M_n}\displaystyle\sum_{T \in \mathcal{U}_n} f(T), & S \in \mathcal{U}_n, \end{cases}$$

when $M_n \geqslant 1$. When $M_n = 0$, the difference is identically zero. Its values on $\mathcal{U}_n$ sum to zero, so

$$f - \widehat{f}^{\,\mathrm{comp},0} \in \mathcal{H}_n.$$

Suppose first that $M_n \geqslant 1$. For every $h \in \mathcal{H}_n$,

$$\begin{aligned} \left\langle \widehat{f}^{\,\mathrm{comp},0}, h \right\rangle &= \frac{1}{N}\sum_{S \in \mathcal{U}_n} \widehat{f}^{\,\mathrm{comp},0}(S) h(S) \\ &= \frac{1}{N}\left\{\frac{1}{M_n}\sum_{T \in \mathcal{U}_n} f(T)\right\}\sum_{S \in \mathcal{U}_n} h(S) = 0. \end{aligned}$$

When $M_n = 0$, one has $\mathcal{H}_n = \{0\}$, so the same orthogonality conclusion is immediate. Thus, $\widehat{f}^{\,\mathrm{comp},0} \in \mathcal{H}_n^{\perp}$, and the orthogonal decomposition

$$f = \widehat{f}^{\,\mathrm{comp},0} + \left(f - \widehat{f}^{\,\mathrm{comp},0}\right)$$

shows that

$$f - \widehat{f}^{\,\mathrm{comp},0} = P_{\mathcal{H}_n} f.$$

The space $\mathcal{H}_n$ has dimension $(M_n - 1)_+$: when $M_n \geqslant 1$, its values on the M_n unobserved subsets are subject to one linear constraint, and when $M_n = 0$ it reduces to $\{0\}$. The proof shows that

$$\mathbb{E}_{\nu_{d,k}^{\otimes n}}\left[\|P_{\mathcal{H}_n} f\|_2^2\right] = \rho_{m,n} \|f\|_2^2 \leqslant B\rho_{m,n},$$

which gives the first term in (26).

We now return to the original model with observation noise. The second term in (26) controls the variance of the observed cell means and of the centering correction. The factor $\eta_{n,N}$ is the expected reciprocal count for a fixed cell, with an empty cell contributing zero. By (27), the second term is bounded by $4\sigma^2 N/(n+1)$, and hence by $4\sigma^2 N/n$, for every $n \geqslant 1$.

5.5 A rank-aware minimax characterization

Define the design-deficiency scale

$$\mathfrak{D}_{n,d,k}(s, B) = \max_{1 \leqslant D \leqslant m} \frac{B}{\gamma_{D,d}^{s}} \rho_{D,n}.$$

This scale complements the noise-driven spectral scale $\mathfrak{R}_{n,d,k}(s, B, \sigma^2)$ defined in (12). The spectral scale measures the noise-driven cost of estimation over the harmonic approximation spaces. The design-deficiency scale measures the loss caused by directions that the sampled evaluations do not identify, a phenomenon that is already present in the noiseless submodel.

Theorem 5.5 (Rank-aware minimax bounds)**.** *There exist universal constants $c, C > 0$ such that, for every $d \geqslant 2$, $1 \leqslant k \leqslant d - 1$, $n \geqslant 2$, and $s, B, \sigma^2 > 0$,*

$$\begin{aligned} c\left\{\mathfrak{R}_{n,d,k}(s, B, \sigma^2) + \mathfrak{D}_{n,d,k}(s, B)\right\} &\leqslant R_n^{\star}(\mathcal{E}_{d,k}^{s}(B)) \\ &\leqslant C\gamma_{m,d}^{s}\left\{\mathfrak{R}_{n,d,k}(s, B, \sigma^2) + \mathfrak{D}_{n,d,k}(s, B)\right\}. \end{aligned} \tag{28}$$

The lower bound follows by combining the two separate lower bounds: Theorem 4.4 gives the spectral term, while Proposition 5.3 gives the design-deficiency term.

For the upper bound, we use the better of empirical projection and centered completion. The relevant comparison is between the smallest squared L^2-semi-axis of the Johnson–Sobolev ellipsoid

$$a_m(s, B) = \frac{B}{\gamma_{m,d}^{s}}$$

and the noise variance σ^2. If $a_m(s, B) \leqslant \sigma^2$, then $B/\sigma^2 \leqslant \gamma_{m,d}^{s}$, and the empirical projection bound gives the required result. If $a_m(s, B) > \sigma^2$, centered completion is more appropriate. Its signal term satisfies

$$B\rho_{m,n} \leqslant \gamma_{m,d}^{s} \mathfrak{D}_{n,d,k}(s, B),$$

while its noise term is controlled by the full-dimensional contribution to $\mathfrak{R}_{n,d,k}(s,B,\sigma^2)$. The proof gives the precise comparison in the two regimes.

Finally,

$$\gamma_{m,d} = \frac{m(d-m+1)}{d} \leqslant m.$$

Consequently, for fixed m and s, there exist a universal constant $c>0$ and a constant $C_{m,s}<\infty$, depending only on m and s, such that

$$\begin{aligned} c\left\{\mathfrak{R}_{n,d,k}(s,B,\sigma^2)+\mathfrak{D}_{n,d,k}(s,B)\right\} &\leqslant R_n^\star(\mathcal{E}^s_{d,k}(B)) \\ &\leqslant C_{m,s}\left\{\mathfrak{R}_{n,d,k}(s,B,\sigma^2)+\mathfrak{D}_{n,d,k}(s,B)\right\}. \end{aligned}$$

For every fixed m and s, these inequalities hold uniformly over all d, k, n, B, and σ^2 such that $\min(k,d-k)=m$, without any restriction on B/σ^2. In particular, they apply to sequences with fixed k and $d\to\infty$, since $m=k$ for all sufficiently large d.

A concrete rate can be obtained in this regime. Fix k, s, B, and σ^2, let $d=d_n\to\infty$, and write $N_n=\binom{d_n}{k}$. Suppose that $n/N_n\to\infty$. The lower inequality in Theorem 4.4, applied to the $D=m=k$ term in the definition (12), gives

$$R_n^\star(\mathcal{E}^s_{d_n,k}(B)) \geqslant c\min\left\{\frac{B}{\gamma^s_{k,d_n}},\frac{\sigma^2(N_n-1)}{n}\right\}.$$

Since $\gamma_{k,d_n}\to k$, the first term in the minimum stays bounded away from zero, whereas the second tends to zero; the minimum is therefore eventually of order $\sigma^2 N_n/n$. Conversely, (22), the inequality $(M_n-1)_+\leqslant M_n$, and $\mathbb{E}[M_n]=N_n(1-1/N_n)^n$ give

$$\begin{aligned} \rho_{m,n} &\leqslant \frac{\mathbb{E}[M_n]}{N_n-1} = \frac{N_n}{N_n-1}\left(1-\frac{1}{N_n}\right)^n \\ &\leqslant 2\exp\left(-\frac{n}{N_n}\right), \end{aligned}$$

where we used $N_n\geqslant 2$ and $1-x\leqslant e^{-x}$. Together with $\eta_{n,N_n}\leqslant 2N_n/(n+1)$, Proposition 5.4 gives

$$R_n^\star(\mathcal{E}^s_{d_n,k}(B)) \leqslant 2B\exp\left(-\frac{n}{N_n}\right)+4\sigma^2\frac{N_n}{n+1}.$$

Under the assumption $n/N_n\to\infty$,

$$\exp\left(-\frac{n}{N_n}\right)=o\left(\frac{N_n}{n}\right).$$

Combining the preceding lower and upper bounds therefore shows that the minimax risk is of order

$$\sigma^2\frac{N_n}{n}=\sigma^2\frac{\binom{d_n}{k}}{n},$$

or, equivalently, of order $\sigma^2 d_n^k/n$ when k is fixed. For example, if $d_n=k+\lceil\log n\rceil$, the resulting rate is of order $\sigma^2(\log n)^k/n$. Conversely, if n/N_n does not tend to infinity, the displayed spectral lower bound remains bounded away from zero along a subsequence. Hence $n/N_n\to\infty$ is necessary and sufficient for minimax consistency in the fixed-k regime.

When m is allowed to grow, the factor $\gamma^s_{m,d}$ leaves room for improvement. Removing it would require an estimator that exploits the Sobolev geometry when controlling unresolved components, whereas (24) uses only the centering constraint.

Numerical experiments reported in Section A of the Supplementary Material complement these theoretical results. They compare prediction under different interaction profiles and illustrate the signal-dependent fluctuation of empirical projection, Gram stability, and the loss of information caused by incomplete coverage.

6 Discussion

We have studied the problem of learning an unknown response function from noisy observations whose inputs are subsets of a fixed ground set. Each observed subset contains exactly k items among d, and the statistical objective is to estimate the response not only at the sampled subsets, but over the entire fixed-cardinality domain. This setting arises whenever the outcome of interest depends on a combination of prescribed size and interactions among the selected items may matter.

The fixed-cardinality constraint prevents the usual product-domain coordinatewise methods from being applied unchanged, because adding or removing a single item leaves the domain. We instead used the Johnson graph, whose edges correspond to replacing one selected item by one unselected item. Its harmonic decomposition organizes response functions into successive interaction levels. The dimensions of these levels determine how many coefficients must be estimated, while the Johnson eigenvalues measure variation under local exchanges and provide a natural notion of smoothness. This leads to explicit approximation spaces, a Johnson–Sobolev class, and projection estimators with finite-sample risk bounds.

The resulting spectral scale characterizes the minimax risk up to constants depending only on a fixed upper bound for the ratio B/σ^2 of the Johnson–Sobolev energy budget to the noise variance. This conclusion holds uniformly in d, k, and n, including when the sample size is smaller than, comparable to, or larger than the number of possible subsets. To cover arbitrary values of B/σ^2, we also incorporated the information lost when the sampled subsets fail to identify every component of the response. The resulting rank-aware bounds are sharp up to the explicit factor $\gamma^s_{m,d}$, and hence up to constants when $m = \min(k, d-k)$ and s are fixed.

An important conclusion is that random sampling creates a difficulty distinct from observation noise. The empirical projection coefficients fluctuate with the sampled subsets even when the responses are noiseless. For a fixed cutoff D, least squares eliminates the signal-dependent design fluctuation of the component $P^\circ_{\leqslant D} f$ already contained in W_D, on the event that the corresponding empirical Gram matrix is stable. The constant leverage of the Johnson approximation spaces gives a simple sufficient condition for this stability. When the design does not identify the fitted space, however, no estimator can recover the corresponding components uniformly over the response class, even without noise. At the full interaction level, the unidentified subspace has dimension $(M_n - 1)_+$, where M_n is the number of subsets that have not been observed; the subtraction of one reflects the centering constraint. The centered completion estimator shows how repeated observations and unobserved subsets can be handled jointly. Thus, the statistical error is governed by three distinct mechanisms: approximation of the response, observation noise, and incomplete coverage of the subset domain.

Our results concern statistical rather than computational optimality. Once the feature vectors $x_D(S_j)$ are available, empirical projection requires $O(np_D)$ arithmetic operations to compute its coefficients. On $\mathcal{A}_D$, which necessarily implies $n \geqslant p_D$, a standard QR implementation of stable least squares requires $O(np_D^2)$ operations. Centered completion can be constructed in a single pass through the n observations. It can then be stored implicitly using the observed cell means together with the common value assigned to all unobserved subsets. These operation counts exclude the cost of constructing the harmonic feature vectors.

Several questions remain open. When m grows, removing the factor $\gamma^s_{m,d}$ would require an estimator that exploits the smoothness geometry when controlling unresolved components, rather than treating all of them uniformly. A related problem is to obtain more explicit expressions for the intermediate rank deficiencies $\rho_{D,n}$. Nonuniform sampling presents another challenge, because the leverage score need no longer be constant and the analysis must account for the sampling distribution. Finally, practical applications on large domains require sample-based harmonic algorithms that exploit low-order interactions without enumerating all $\binom{d}{k}$ subsets. The same general approach may extend to regression on other constrained combinatorial objects, provided that one can identify the admissible local changes, the associated harmonic decomposition, and the information retained by the sampling design.

References

Balcan, M.-F., Constantin, F., Iwata, S., and Wang, L. (2012). Learning valuation functions. In Mannor, S., Srebro, N., and Williamson, R. C., editors, *Proc. 25th COLT*, volume 23, pages 4.1–4.24. PMLR.

Balcan, M.-F., Vitercik, E., and White, C. (2016). Learning combinatorial functions from pairwise comparisons. In Feldman, V., Rakhlin, A., and Shamir, O., editors, *Proc. 29th COLT*, volume 49, pages 310–335. PMLR.

Bannai, E. and Ito, T. (1984). *Algebraic Combinatorics I: Association Schemes*. Benjamin/Cummings, Menlo Park.

Benavoli, A., Azzimonti, D., and Piga, D. (2023). Learning choice functions with Gaussian processes. In Evans, R. J. and Shpitser, I., editors, *Proc. 39th UAI*, volume 216, pages 141–151. PMLR.

Brouwer, A. E., Cohen, A. M., and Neumaier, A. (1989). *Distance-Regular Graphs*. Springer, Berlin.

Cohen, A., Davenport, M. A., and Leviatan, D. (2013). On the stability and accuracy of least squares approximations. *Found. Comput. Math.*, 13:819–834.

Cohen, A., Davenport, M. A., and Leviatan, D. (2019). Correction to: On the stability and accuracy of least squares approximations. *Found. Comput. Math.*, 19:239.

De, A. and Chakrabarti, S. (2022). Neural estimation of submodular functions with applications to differentiable subset selection. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Adv. Neural Inf. Process. Syst.*, volume 35, pages 19537–19552. Curran Associates, Inc.

Delsarte, P. (1973). *An Algebraic Approach to the Association Schemes of Coding Theory*. Number 10 in Philips Research Reports Supplements. N.V. Philips' Gloeilampenfabrieken, Eindhoven. Doctoral dissertation, Université Catholique de Louvain.

Diaconis, P. (1988). *Group Representations in Probability and Statistics*, volume 11 of *Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, Hayward.

Diaconis, P. (1989). A generalization of spectral analysis with application to ranked data. *Ann. Statist.*, 17:949–979.

Diaconis, P. and Rockmore, D. (1993). Efficient computation of isotypic projections for the symmetric group. In Finkelstein, L. and Kantor, W. M., editors, *Groups and Computation*, pages 87–104. American Mathematical Society, Providence.

Dolhansky, B. W. and Bilmes, J. A. (2016). Deep submodular functions: Definitions and learning. In Lee, D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R., editors, *Adv. Neural Inf. Process. Syst.*, volume 29, pages 3404–3412. Curran Associates, Inc.

Filmus, Y. (2016). An orthogonal basis for functions over a slice of the Boolean hypercube. *Electron. J. Combin.*, 23:P1.23.

Gottlieb, D. H. (1966). A certain class of incidence matrices. *Proc. Amer. Math. Soc.*, 17:1233–1237.

Green, A., Balakrishnan, S., and Tibshirani, R. J. (2021). Minimax optimal regression over Sobolev spaces via Laplacian regularization on neighborhood graphs. In Banerjee, A. and Fukumizu, K., editors, *Proc. 24th AISTATS*, volume 130, pages 2602–2610. PMLR.

Iglesias, R. and Natale, M. (2022). A fast Fourier transform for the Johnson graph. *J. Fourier Anal. Appl.*, 28:62.

Kirichenko, A. and van Zanten, H. (2017). Estimating a smooth function on a large graph by Bayesian Laplacian regularisation. *Electron. J. Statist.*, 11:891–915.

Kirichenko, A. and van Zanten, H. (2018). Minimax lower bounds for function estimation on graphs. *Electron. J. Statist.*, 12:651–666.

Liang, H., Wan, X., and Dong, X. (2024). Bayesian optimization of functions over node subsets in graphs. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C., editors, *Adv. Neural Inf. Process. Syst.*, volume 37, pages 38199–38234. Curran Associates, Inc.

Massart, P. (2007). *Concentration Inequalities and Model Selection*. Springer, Berlin.

O'Donnell, R. (2014). *Analysis of Boolean Functions*. Cambridge University Press, Cambridge.

Oh, C., Tomczak, J. M., Gavves, E., and Welling, M. (2019). Combinatorial Bayesian optimization using the graph Cartesian product. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Adv. Neural Inf. Process. Syst.*, volume 32, pages 2914–2924. Curran Associates, Inc.

Puy, G., Tremblay, N., Gribonval, R., and Vandergheynst, P. (2018). Random sampling of bandlimited signals on graphs. *Appl. Comput. Harmon. Anal.*, 44:446–475.

Qian, C., Shi, J.-C., Yu, Y., Tang, K., and Zhou, Z.-H. (2017). Subset selection under noise. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Adv. Neural Inf. Process. Syst.*, volume 30, pages 3563–3573. Curran Associates, Inc.

Shi, Z., Balasubramanian, K., and Polonik, W. (2024). Adaptive and non-adaptive minimax rates for weighted Laplacian-eigenmap based nonparametric regression. In Dasgupta, S., Mandt, S., and Li, Y., editors, *Proc. 27th AISTATS*, volume 238, pages 2800–2808. PMLR.

Stobbe, P. and Krause, A. (2012). Learning Fourier sparse set functions. In Lawrence, N. D. and Girolami, M., editors, *Proc. 15th AISTATS*, volume 22, pages 1125–1133. PMLR.

Tajdini, A., Jain, L., and Jamieson, K. (2024). Nearly minimax optimal submodular maximization with bandit feedback. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C., editors, *Adv. Neural Inf. Process. Syst.*, volume 37, pages 96254–96281. Curran Associates, Inc.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12:389–434.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York.

Wagstaff, E., Fuchs, F., Engelcke, M., Posner, I., and Osborne, M. A. (2019). On the limitations of representing functions on sets. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proc. 36th ICML*, volume 97, pages 6487–6494. PMLR.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Póczos, B., Salakhutdinov, R., and Smola, A. J. (2017). Deep sets. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Adv. Neural Inf. Process. Syst.*, volume 30, pages 3394–3404. Curran Associates, Inc.

Supplementary Material

This supplementary material contains the numerical experiments, the canonical frame construction, and proofs of all results stated in the main article.

A Numerical experiments

We use $(d,k) = (12,4)$, for which $N = 495$, and independent Gaussian noise with standard deviation 0.5. Prediction error is measured by the squared $L^2(\nu_{12,4})$-distance and is evaluated exactly over all 495 subsets.

For each spectral profile, we generate 120 independent centered target functions $f^{(1)}, \dots, f^{(120)}$ directly from the Johnson decomposition and normalize them so that $\|f^{(q)}\|_2 = 1$. Computationally, an $L^2(\nu_{12,4})$-orthonormal basis of each V_r is obtained by projecting the order-r inclusion indicators onto $U_{r-1}^{\perp}$ and applying a QR orthonormalization. Within each level, the coefficient vector is drawn from a standard Gaussian distribution and normalized to unit Euclidean norm before being rescaled to the prescribed squared L^2-mass. The resulting direction is uniform on the unit sphere of V_r, so its distribution does not depend on the particular orthonormal basis produced by the QR decomposition.

The *pairwise* profile allocates 20% of the squared L^2-norm to V_1 and 80% to V_2. The *diffuse* profile allocates 50%, 30%, 15%, and 5% to $V_1, V_2, V_3,$, and V_4, respectively.

Each target $f^{(q)}$ is evaluated at every sample size $n \in \{40, 80, 160, 320, 640\}$. For each pair $(f^{(q)}, n)$, we independently draw a new design and a new noise sample. Thus, the target functions are shared across sample sizes, whereas the design and noise are always resampled. At each fixed sample size, the reported Monte Carlo mean is computed from 120 losses, each corresponding to a different target and an independently generated design and noise sample.

The numerical study has two parts. We first compare predictive performance under the pairwise and diffuse target profiles. We then use separate diagnostics to examine the random-design phenomena identified by the theory: the signal-dependent variation of empirical projection, the stability of least squares, and the loss of information caused by incomplete coverage.

The prediction comparison involves five procedures. The first is empirical harmonic projection, implemented as $\overline{Y} + \widehat{f}_D^{\mathrm{proj}}$: the sample mean estimates the constant component, while $\widehat{f}_D^{\mathrm{proj}}$ estimates the nonconstant levels through D. Although the simulated targets are centered, the constant component is included to reflect the practical case of an unknown intercept. The second is Johnson spectral shrinkage, which leaves the constant level unchanged and multiplies the empirical component in V_r, $r \geqslant 1$, by

$$(1 + \lambda \gamma_{r,d})^{-1}.$$

The third is a Nadaraya–Watson estimator based on the Johnson distance, which we call the Johnson kernel estimator. Its weights are proportional to

$$\exp\left\{ -\operatorname{dist}_J(S, S')/h \right\}.$$

The remaining two procedures are additive least squares over $V_0 \oplus^{\perp} V_1 = U_1$ and the sample mean, which serve as low-dimensional baselines.

The cutoff D, shrinkage parameter λ, and kernel bandwidth h are selected using the same randomly chosen validation quarter within each repetition. The candidate sets are

$$D \in \{0, \ldots, 4\}, \qquad \lambda \in \left\{10^{-2.5+4.5j/15} : j = 0, \ldots, 15\right\},$$

and

$$h \in \{0.20, 0.35, 0.55, 0.85, 1.30, 2.00\}.$$

After selection, the corresponding estimator is refitted on the full sample. Final performance is evaluated against the known target function, using the exact $L^2(\nu_{12,4})$-error over the complete domain rather than the validation loss. Additive least squares and the sample mean have no tuning parameter and are computed directly from the full sample. This validation scheme is used only for the numerical comparison and is not analyzed theoretically in the article. Stable least squares is not included in this first comparison because the Gram-stability event need not hold at these sample sizes; it is studied separately in the third panel in a regime where the empirical Gram matrix is stable. Centered completion is likewise not used as a general prediction competitor: it is a full-level estimator designed specifically to address incomplete coverage, whose extent is displayed in the final panel.

Figure 1 summarizes the two prediction experiments and the three random-design diagnostics described below.

The pairwise experiment shows when the larger interaction space becomes estimable. At $n = 320$, harmonic projection has mean error 0.258, compared with 0.340 for the Johnson kernel, 0.554 for spectral shrinkage, 0.840 for additive least squares, and 1.004 for the sample mean. At $n = 640$, the corresponding errors are 0.126, 0.218, 0.410, 0.821, and 1.001. Since $p_2 = 65$, fitting the second harmonic level is costly at the smallest sample sizes. The validation-selected projection estimator is then highly variable and occasionally selects an excessively large cutoff. At $n = 320$ and $n = 640$, however, the validation rule selects $D = 2$ in every repetition, and estimating the pairwise component becomes beneficial.

The diffuse profile favors gradual regularization. The Johnson kernel has the smallest mean error throughout the displayed range, reaching 0.220 at $n = 640$, compared with 0.333 for validation-selected hard projection. This is consistent with the allocation of squared L^2-norm across several harmonic levels. Hard projection must select a single cutoff: it retains all components below that cutoff without shrinkage and discards all components above it. By contrast, the Johnson kernel averages observations across nearby subsets without imposing a hard interaction cutoff. The comparison is therefore mechanistic, not a claim that one estimator dominates uniformly.

The third panel compares the centered empirical projection $\widehat{f}_2^{\mathrm{proj}}$ and stable least squares over W_2 as the signal strength increases relative to the observation noise. Unlike in the preceding prediction experiments, we fix once and for all one of the pairwise targets described above, denoted by $f_0 \in W_2$, and consider $f_t = tf_0$, $t \in \{0.5, 1, 2, 4, 8\}$. Since $\|f_0\|_2 = 1$, the parameter t is the L^2-amplitude of the signal. For any fixed $s > 0$, the smallest Johnson–Sobolev budget for which $f_t \in \mathcal{E}^s_{d,k}(B)$ is

$$B_t = I^{(2,s)}_{J,d,k}(f_t) = t^2 I^{(2,s)}_{J,d,k}(f_0).$$

Thus, with the noise variance held fixed, increasing t increases the ratio B_t/σ^2.

We take $D = 2$ and $n = 2600 = 40p_2$. Since $f_t \in W_2$, no harmonic component is omitted and neither estimator incurs approximation error. For each value of t, we generate 200 independent

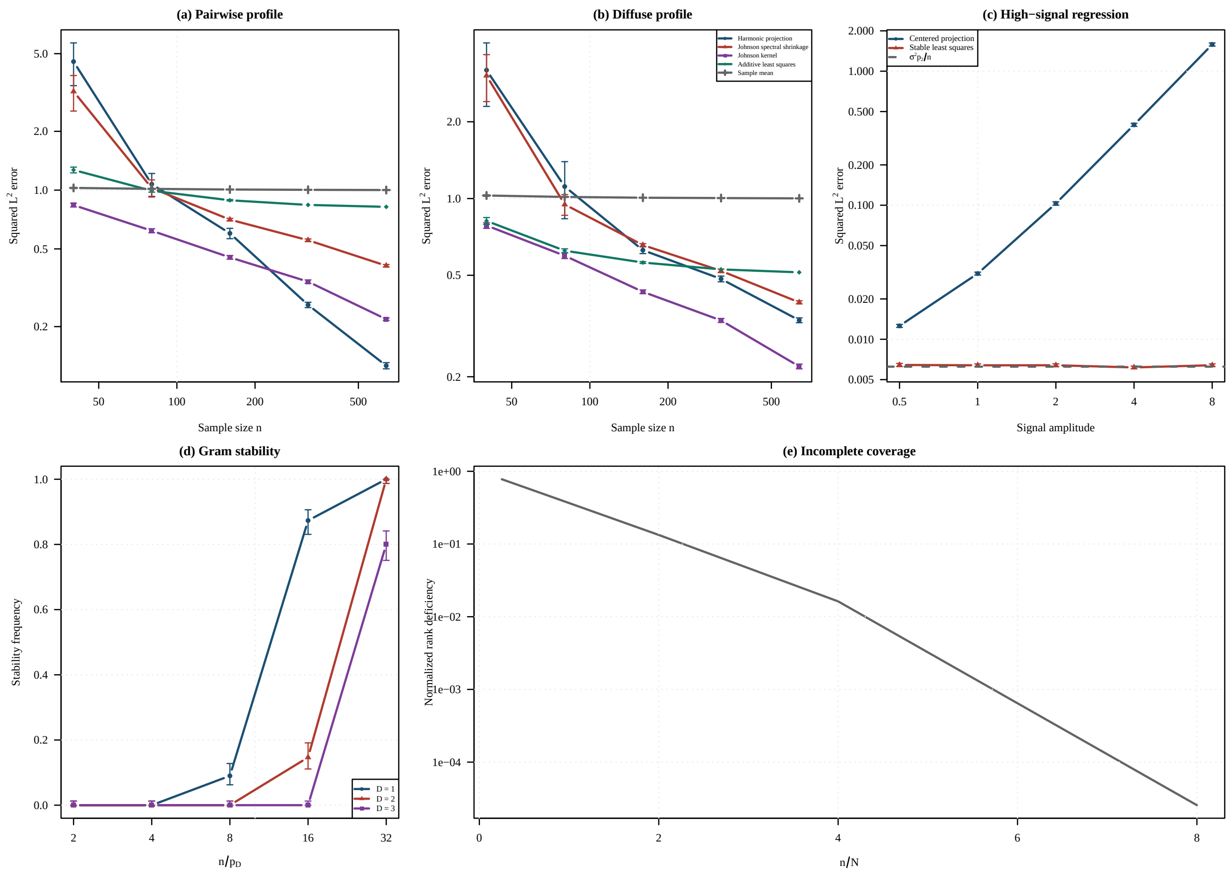

Figure 1: **Prediction and random-design mechanisms on $\mathcal{S}_{12,4}$.** The first two panels show squared $L^2(\nu_{12,4})$-error for the pairwise and diffuse profiles. Points are Monte Carlo means over 120 joint target-and-data realizations. Where visible, error bars are normal-approximation 95% confidence intervals for these means, computed as the mean plus or minus 1.96 Monte Carlo standard errors. The third panel compares centered empirical projection and stable least squares at $D = 2$ and $n = 2600$ as a fixed signal is rescaled; the dashed line is $\sigma^2 p_2/n$, and results are based on 200 independent repetitions at each amplitude. The fourth panel reports the frequency of $\|\widehat{G}_D - I_{p_D}\|_{\mathrm{op}} \leqslant 1/2$ for $D = 1, 2, 3$, based on 300 independent Gram matrices at each configuration; bars are 95% Wilson confidence intervals for the corresponding stability probabilities. The fifth panel plots the exact normalized full-level rank deficiency $\rho_{m,n}$ from (23) as a function of n/N.

design and noise samples; the target direction f_0 itself remains unchanged. Each design produces a new empirical Gram matrix $\widehat{G}_2$, and we check whether

$$\mathcal{A}_2 = \left\{ \|\widehat{G}_2 - I_{p_2}\|_{\mathrm{op}} \leqslant \frac{1}{2} \right\}$$

occurs. This stability event occurs in all 200 repetitions at every displayed value of t. Consequently, stable least squares coincides with ordinary least squares over W_2 in every repetition. On this event, empirical and population squared norms are comparable throughout W_2, so the least-squares fit is well conditioned.

As t increases from 0.5 to 8, the least-squares risk remains between 0.00617 and 0.00644, close to the observation-noise benchmark

$$\frac{\sigma^2 p_2}{n} = 0.00625.$$

By contrast, the projection risk increases from 0.0126 to 1.582. This illustrates the distinction established in Theorem 5.2: once the Gram matrix is stable, least squares removes the random-design fluctuation of the component already contained in W_2, whereas empirical projection continues to incur a variance that increases with the squared amplitude of the signal.

The fourth panel examines the stability event itself. For each $D \in \{1, 2, 3\}$ and each displayed value of n/p_D, we generate 300 independent designs and estimate

$$\mathbb{P}_{\nu_{d,k}^{\otimes n}}(\mathcal{A}_D), \qquad \mathcal{A}_D = \left\{ \|\widehat{G}_D - I_{p_D}\|_{\mathrm{op}} \leqslant \frac{1}{2} \right\}.$$

At $n/p_D = 16$, the empirical frequencies are 0.873, 0.147, and 0 for $D = 1, 2, 3$, respectively. At $n/p_D = 32$, they are 1, 1, and 0.80. The transition therefore depends not only on the number of observations per fitted coefficient but also on the dimension p_D. This is consistent with Proposition 5.1, whose sufficient stability condition is of order $p_D \log p_D$.

The final panel illustrates the incomplete-coverage mechanism using the exact expression (23). The normalized full-level rank deficiency decreases from 0.366 at $n = N$ to 0.133 at $n = 2N$, 0.0163 at $n = 4N$, and 2.56×10^{-5} at $n = 8N$. These values quantify the normalized expected dimension of the entire unresolved centered subspace.

All experiments concern the same fixed-cardinality domain $\mathcal{S}_{12,4}$. The prediction experiments use Gaussian noise and targets generated from the harmonic decomposition underlying the theory, whereas the Gram experiment and the exact rank-deficiency illustration concern the random design alone. These experiments illustrate the mechanisms identified by the theory rather than establish broad empirical superiority.

B Johnson harmonic analysis and the canonical frame

B.1 Proof of Proposition 3.1

At level $r = 0$, the space V_0 consists of the constant functions. Hence, $K\mathbf{1} = \mathbf{1}$, which proves (2) because $\theta_{0,d,k} = 1$.

Fix $1 \leqslant r \leqslant m$, let $T \subseteq [d]$ have cardinality r, and recall that

$$h_T(S) = \mathbf{1}_{\{T \subseteq S\}}.$$

We first compute $(Kh_T)(S)$ by counting the $k(d-k)$ possible ordered swaps from a fixed $S \in \mathcal{S}_{d,k}$.

If $T \subseteq S$, the inclusion remains valid after the swap exactly when the outgoing item lies in $S \setminus T$. There are $k - r$ possible outgoing items and $d - k$ possible incoming items, giving $(k-r)(d-k)$ successful swaps. If $|T \setminus S| = 1$, the incoming item must be the unique missing element of T, while the outgoing item can be any of the $k - r + 1$ elements of $S \setminus T$. If at least two elements of T are missing, one exchange cannot produce a subset containing T. Dividing these counts by $k(d-k)$ gives

$$(Kh_T)(S) = \left(1 - \frac{r}{k}\right) h_T(S) + \frac{k - r + 1}{k(d-k)} \mathbf{1}_{\{|T \setminus S| = 1\}}. \tag{29}$$

The remaining indicator satisfies

$$\mathbf{1}_{\{|T\setminus S|=1\}} = \sum_{\substack{R\subset T\\ |R|=r-1}} h_R(S) - rh_T(S). \tag{30}$$

Indeed, the sum counts the $(r-1)$-subsets of T that are contained in S. If $T \subseteq S$, all r such subsets are counted. If exactly one element of T is missing from S, only the subset obtained by removing that element is counted. If at least two elements are missing, deleting a single element from T cannot produce a subset contained in S. The sum is therefore equal to r, 1, or 0, respectively, and subtracting $rh_T(S)$ proves (30).

Substituting (30) into (29) gives the functional identity

$$(K-\theta_{r,d,k}I)h_T = \frac{k-r+1}{k(d-k)} \sum_{\substack{R\subset T\\ |R|=r-1}} h_R \in U_{r-1},$$

because

$$1-\frac{r}{k}-\frac{r(k-r+1)}{k(d-k)} = 1-\frac{r(d-r+1)}{k(d-k)} = \theta_{r,d,k}.$$

Since the functions h_T, with $|T| = r$, span U_r, it follows that

$$(K-\theta_{r,d,k}I)U_r \subseteq U_{r-1}. \tag{31}$$

We also need the invariance of U_{r-1} under K. If $r \geqslant 2$, applying the preceding calculation at order $r-1$ gives

$$(K-\theta_{r-1,d,k}I)h_C \in U_{r-2} \qquad \text{for every } C \subseteq [d] \text{ with } |C| = r-1.$$

Since $h_C \in U_{r-1}$ and $U_{r-2} \subseteq U_{r-1}$, this implies $Kh_C \in U_{r-1}$. These indicators span U_{r-1}, so K preserves U_{r-1}. When $r = 1$, the same conclusion follows from $U_0 = \operatorname{span}\{\mathbf{1}\}$ and $K\mathbf{1} = \mathbf{1}$.

Because K is self-adjoint, the invariance of U_{r-1} implies that $U_{r-1}^{\perp}$ is also invariant. Indeed, if $f \perp U_{r-1}$ and $g \in U_{r-1}$, then

$$\langle Kf, g\rangle = \langle f, Kg\rangle = 0,$$

because $Kg \in U_{r-1}$.

Now let $f \in V_r = U_r \cap U_{r-1}^{\perp}$. Relation (31) gives

$$(K-\theta_{r,d,k}I)f \in U_{r-1}.$$

On the other hand, both Kf and f belong to $U_{r-1}^{\perp}$, so the same difference belongs to $U_{r-1}^{\perp}$. Since $U_{r-1} \cap U_{r-1}^{\perp} = \{0\}$, we conclude that

$$(K-\theta_{r,d,k}I)f = 0,$$

which proves (2).

It remains to verify the Dirichlet identity. Suppose that $S_0 \sim \nu_{d,k}$ and, conditionally on $S_0 = S$, let S_1 have transition probabilities $q(S,\cdot)$. We first check that S_1 is also uniform on $\mathcal{S}_{d,k}$. For every $S' \in \mathcal{S}_{d,k}$, detailed balance gives

$$\begin{aligned}
\mathbb{P}(S_1 = S') &= \sum_{S\in\mathcal{S}_{d,k}} \nu_{d,k}(S)q(S,S') \\
&= \sum_{S\in\mathcal{S}_{d,k}} \nu_{d,k}(S')q(S',S) \\
&= \nu_{d,k}(S'),
\end{aligned}$$

because $\sum_{S\in\mathcal{S}_{d,k}} q(S',S) = 1$. Thus, $\nu_{d,k}$ is stationary for the one-exchange walk.

For a fixed $S \in \mathcal{S}_{d,k}$, the conditional distribution of S_1 gives

$$\mathbb{E}\left[f(S_1)^2 \mid S_0 = S\right] = \frac{1}{k(d-k)} \sum_{a\in S} \sum_{b\notin S} f(S^{a\to b})^2.$$

Averaging this identity with respect to $S_0 \sim \nu_{d,k}$, and using $S_1 \sim \nu_{d,k}$, gives

$$\frac{1}{k(d-k)} \mathbb{E}_{S\sim\nu_{d,k}}\left[\sum_{a\in S}\sum_{b\notin S} f(S^{a\to b})^2\right] = \mathbb{E}_{S\sim\nu_{d,k}}[f(S)^2] = \|f\|_2^2. \tag{32}$$

Similarly, conditioning on S_0 gives

$$\begin{aligned}\frac{1}{k(d-k)} \mathbb{E}_{S\sim\nu_{d,k}}\left[\sum_{a\in S}\sum_{b\notin S} f(S)f(S^{a\to b})\right] &= \mathbb{E}_{S\sim\nu_{d,k}}\left[f(S)\,\mathbb{E}[f(S_1) \mid S_0 = S]\right] \\ &= \mathbb{E}_{S\sim\nu_{d,k}}[f(S)Kf(S)] = \langle f, Kf\rangle\,.\end{aligned}$$

Expanding the squared differences, using (32) for the second quadratic term and the preceding identity for the cross term, yields

$$\begin{aligned}\frac{1}{2k(d-k)} \mathbb{E}_{S\sim\nu_{d,k}}\left[\sum_{a\in S}\sum_{b\notin S} \{f(S) - f(S^{a\to b})\}^2\right] &= \mathbb{E}_{S\sim\nu_{d,k}}[f(S)^2] - \mathbb{E}_{S\sim\nu_{d,k}}[f(S)Kf(S)] \\ &= \langle f, (I-K)f\rangle\,,\end{aligned}$$

which proves (3).

B.2 A canonical equivariant frame

The decomposition into the spaces V_r is canonical, whereas an orthonormal basis within a given level is not. Recall that, in a finite-dimensional Hilbert space $\mathcal{H}$, a frame is a finite spanning family $\{\psi_a : a \in \mathcal{I}\}$. Its frame operator is

$$\mathcal{F}g = \sum_{a\in\mathcal{I}} \langle g, \psi_a\rangle \psi_a, \qquad g \in \mathcal{H}.$$

The frame is tight if there exists $C > 0$ such that $\mathcal{F} = CI_{\mathcal{H}}$. In that case,

$$g = C^{-1} \sum_{a\in\mathcal{I}} \langle g, \psi_a\rangle \psi_a.$$

Taking the inner product of the reconstruction identity with g gives

$$\|g\|_{\mathcal{H}}^2 = C^{-1} \sum_{a\in\mathcal{I}} \langle g, \psi_a\rangle_{\mathcal{H}}^2,$$

and therefore

$$\sum_{a\in\mathcal{I}} \langle g, \psi_a\rangle_{\mathcal{H}}^2 = C\|g\|_{\mathcal{H}}^2.$$

Thus, a tight frame may be redundant while retaining reconstruction and energy identities analogous to those of an orthonormal basis. We now construct such a frame in a way that respects the permutation symmetry of the fixed-cardinality domain.

Fix $1 \leqslant r \leqslant m$ and $T \subseteq [d]$ with $|T| = r$. For $S \in \mathcal{S}_{d,k}$, let

$$L_T(S) = |S \cap T|$$

be the number of items shared by S and T. If $S \sim \nu_{d,k}$, then $L_T(S)$ has the hypergeometric distribution

$$\mathbb{P}_{S\sim\nu_{d,k}}\{L_T(S) = \ell\} = w_{r,d,k}(\ell) := \frac{\binom{r}{\ell}\binom{d-r}{k-\ell}}{\binom{d}{k}}, \qquad 0 \leqslant \ell \leqslant r. \tag{33}$$

Indeed, among the $\binom{d}{k}$ possible k-subsets, those having overlap ℓ with T are obtained by choosing ℓ items from T and $k - \ell$ items from its complement. Moreover, $r \leqslant m = \min(k, d-k)$ ensures that every overlap $\ell \in \{0, \dots, r\}$ is attainable. Hence,

$$w_{r,d,k}(\ell) > 0, \qquad 0 \leqslant \ell \leqslant r.$$

Our aim is to transform the overlap $L_T(S) = |S \cap T|$ into a normalized function that isolates the pure harmonic level V_r. The natural construction uses orthogonal polynomials for the hypergeometric distribution of $L_T(S)$.

Let $\mathcal{P}_r$ be the space of polynomials of degree at most r. On this space, define

$$\langle p, q\rangle_w = \sum_{\ell=0}^{r} w_{r,d,k}(\ell)p(\ell)q(\ell), \qquad p, q \in \mathcal{P}_r.$$

This is an inner product because all the weights are positive and no nonzero polynomial in $\mathcal{P}_r$ can vanish at all $r+1$ distinct points $0, \dots, r$. Within the $(r+1)$-dimensional space $\mathcal{P}_r$, the subspace $\mathcal{P}_{r-1}$ has dimension r and hence codimension one. Its orthogonal complement is therefore one-dimensional. Every nonzero element of this complement has degree exactly r, and any two such elements are proportional.

Subsection B.3 constructs an explicit nonzero generator of this one-dimensional space from the coefficients of the r-th finite-difference operator. It also shows that the value of this generator at r is nonzero. After normalizing with respect to $\langle\cdot,\cdot\rangle_w$ and choosing its sign, we obtain the unique polynomial $H_{r,d,k}$ characterized by

$$\langle H_{r,d,k}, p\rangle_w = 0 \quad \text{for every } p \in \mathcal{P}_{r-1}, \qquad \langle H_{r,d,k}, H_{r,d,k}\rangle_w = 1, \qquad H_{r,d,k}(r) > 0.$$

We then lift this polynomial from the overlap variable to the full domain by defining

$$\phi_{r,T}(S) = H_{r,d,k}(|S \cap T|), \qquad S \in \mathcal{S}_{d,k}. \tag{34}$$

At level $r = 0$, set $\phi_{0,\varnothing} \equiv 1$.

The proof of Theorem B.1, given in Subsection B.4, also establishes that

$$\phi_{r,T} = \frac{P_r h_T}{\|P_r h_T\|_2}.$$

Thus, $\phi_{r,T}$ is precisely the normalized order-r harmonic component of the inclusion indicator h_T.

Theorem B.1 (Canonical tight frame). *For every $0 \leqslant r \leqslant m$, the family*

$$\{\phi_{r,T} : T \subseteq [d],\ |T| = r\}$$

consists of unit-norm functions in V_r and spans V_r. With

$$A_{r,d,k} = \frac{\dim(V_r)}{\binom{d}{r}} = 1 - \frac{r}{d-r+1}, \tag{35}$$

the frame satisfies, for every $f \in L^2(\mathcal{S}_{d,k}, \nu_{d,k})$,

$$P_r f = A_{r,d,k} \sum_{\substack{T\subseteq[d] \\ |T|=r}} \langle f, \phi_{r,T}\rangle\, \phi_{r,T}, \tag{36}$$

$$\|P_r f\|_2^2 = A_{r,d,k} \sum_{\substack{T\subseteq[d] \\ |T|=r}} \langle f, \phi_{r,T}\rangle^2 . \tag{37}$$

The construction is permutation-equivariant: for every permutation π of $[d]$,

$$\phi_{r,\pi(T)}(\pi(S)) = \phi_{r,T}(S), \qquad S \in \mathcal{S}_{d,k}.$$

The frame uses $\binom{d}{r}$ vectors to represent the space V_r, whose dimension is $\binom{d}{r} - \binom{d}{r-1}$, with the convention $\binom{d}{-1} = 0$. Its exact redundancy, defined as the number of frame vectors divided by the dimension of their span, is therefore $A_{r,d,k}^{-1}$. This representation will be used to write the projection estimator in an explicit form that is invariant under relabeling of the ground set. The statistical risk bounds depend only on the harmonic spaces V_r, not on a particular frame or orthonormal basis.

B.3 Overlap polynomial

The case $r = 0$, for which $\phi_{0,\varnothing} \equiv 1$, is immediate and does not require an overlap polynomial. We therefore fix $1 \leqslant r \leqslant m$. All the hypergeometric weights $w_{r,d,k}(\ell)$, $0 \leqslant \ell \leqslant r$, are strictly positive: the inequalities $r \leqslant k$ and $r \leqslant d-k$ ensure that every overlap $\ell \in \{0, \dots, r\}$ is attainable.

We first recall the finite-difference identity underlying the construction. For a function p on the integers, define its forward difference by

$$(\Delta p)(x) = p(x+1) - p(x).$$

Iterating this operator r times gives

$$(\Delta^r p)(0) = \sum_{\ell=0}^{r} (-1)^{r-\ell} \binom{r}{\ell} p(\ell). \tag{38}$$

If p is a polynomial of degree at most $r-1$, then each application of Δ lowers its degree by one, and hence

$$\Delta^r p \equiv 0.$$

The coefficients used below are the normalized coefficients in (38). Define

$$\lambda_{r,\ell} = \frac{(-1)^{r-\ell}}{\ell!(r-\ell)!} = \frac{1}{r!}(-1)^{r-\ell}\binom{r}{\ell}, \qquad Z_{r,d,k} = \sum_{\ell=0}^{r} \frac{\lambda_{r,\ell}^2}{w_{r,d,k}(\ell)}. \tag{39}$$

Since all the weights are positive, $0 < Z_{r,d,k} < \infty$.

We now prescribe the values

$$H_{r,d,k}(\ell) := \frac{\lambda_{r,\ell}}{w_{r,d,k}(\ell)\sqrt{Z_{r,d,k}}}, \qquad \ell = 0, \ldots, r. \tag{40}$$

Because the interpolation nodes $0, \ldots, r$ are distinct, the Lagrange interpolation theorem yields a unique polynomial of degree at most r taking the values in (40). We use the notation $H_{r,d,k}$ both for this polynomial and for its restriction to the interpolation grid.

Lemma B.2 (Top hypergeometric polynomial)**.** *For every $1 \leqslant r \leqslant m$, the interpolating polynomial $H_{r,d,k}$ has degree exactly r and satisfies*

$$\sum_{\ell=0}^{r} w_{r,d,k}(\ell) H_{r,d,k}(\ell) p(\ell) = 0$$

for every polynomial p of degree at most $r-1$. Moreover,

$$\sum_{\ell=0}^{r} w_{r,d,k}(\ell) H_{r,d,k}(\ell)^2 = 1, \qquad H_{r,d,k}(r) > 0.$$

Proof. Fix $1 \leqslant r \leqslant m$, and let p be a polynomial of degree at most $r-1$. By (38),

$$\sum_{\ell=0}^{r} \lambda_{r,\ell} p(\ell) = \frac{1}{r!}(\Delta^r p)(0) = 0.$$

Using (40), we obtain

$$\sum_{\ell=0}^{r} w_{r,d,k}(\ell) H_{r,d,k}(\ell) p(\ell) = \frac{1}{\sqrt{Z_{r,d,k}}} \sum_{\ell=0}^{r} \lambda_{r,\ell} p(\ell) = 0,$$

which proves the orthogonality assertion. By (39),

$$\sum_{\ell=0}^{r} w_{r,d,k}(\ell) H_{r,d,k}(\ell)^2 = \frac{1}{Z_{r,d,k}} \sum_{\ell=0}^{r} \frac{\lambda_{r,\ell}^2}{w_{r,d,k}(\ell)} = 1.$$

If $H_{r,d,k}$ had degree at most $r-1$, the orthogonality assertion could be applied with $p = H_{r,d,k}$, giving

$$\sum_{\ell=0}^{r} w_{r,d,k}(\ell) H_{r,d,k}(\ell)^2 = 0,$$

in contradiction with its unit norm. Hence, $\deg(H_{r,d,k}) = r$. Finally, $\lambda_{r,r} = 1/r! > 0$, while $w_{r,d,k}(r) > 0$ and $Z_{r,d,k} > 0$. Formula (40) therefore gives $H_{r,d,k}(r) > 0$. □

The lemma shows that the interpolating polynomial constructed here coincides with the polynomial $H_{r,d,k}$ introduced in Subsection B.2. Indeed, within the $(r+1)$-dimensional space of polynomials of degree at most r, the orthogonal complement of the polynomials of degree at most $r-1$ is one-dimensional. Unit normalization and positivity at r select a unique representative of this one-dimensional space.

We now lift this overlap polynomial to functions on $\mathcal{S}_{d,k}$ and verify that the resulting functions belong to V_r.

Lemma B.3. *For every $1 \leqslant r \leqslant m$ and every $T \subseteq [d]$ with $|T| = r$, the function $\phi_{r,T}$ defined in (34) belongs to V_r, has unit norm, and satisfies*

$$\phi_{r,\pi(T)}(\pi(S)) = \phi_{r,T}(S), \qquad S \in \mathcal{S}_{d,k},$$

for every permutation π of $[d]$.

Proof. The equivariance property follows from

$$|\pi(S) \cap \pi(T)| = |\pi(S \cap T)| = |S \cap T|.$$

Indeed,

$$\phi_{r,\pi(T)}(\pi(S)) = H_{r,d,k}(|\pi(S) \cap \pi(T)|) = H_{r,d,k}(|S \cap T|) = \phi_{r,T}(S).$$

The norm of $\phi_{r,T}$ is determined by the distribution of the overlap $L_T(S) = |S \cap T|$. Using (33) and Lemma B.2, we obtain

$$\begin{aligned}\|\phi_{r,T}\|_2^2 &= \mathbb{E}_{S \sim \nu_{d,k}}\left[H_{r,d,k}(L_T(S))^2\right] \\ &= \sum_{\ell=0}^{r} w_{r,d,k}(\ell) H_{r,d,k}(\ell)^2 = 1.\end{aligned}$$

It remains to verify the two conditions defining $V_r = U_r \cap U_{r-1}^{\perp}$. We first prove that $\phi_{r,T} \in U_r$. For $0 \leqslant j \leqslant r$,

$$\binom{|S \cap T|}{j} = \sum_{\substack{R \subseteq T \\ |R| = j}} h_R(S). \tag{41}$$

Both sides count the j-subsets of $S \cap T$. Moreover, the polynomials

$$\ell \longmapsto \binom{\ell}{j}, \qquad j = 0, \ldots, r,$$

form a basis of the space of polynomials of degree at most r: the j-th polynomial has degree j and leading coefficient $1/j!$. Since $H_{r,d,k}$ has degree r, there exist coefficients $c_0, \ldots, c_r$ such that

$$H_{r,d,k}(\ell) = \sum_{j=0}^{r} c_j \binom{\ell}{j}.$$

Evaluating this identity at $\ell = |S \cap T|$ and applying (41) gives

$$\phi_{r,T}(S) = \sum_{j=0}^{r} c_j \binom{|S \cap T|}{j} = \sum_{j=0}^{r} c_j \sum_{\substack{R \subseteq T \\ |R| = j}} h_R(S).$$

Each function h_R in the last display belongs to U_j, and $U_j \subseteq U_r$ by nesting. Therefore, $\phi_{r,T} \in U_r$.

We next show that $\phi_{r,T} \perp U_{r-1}$. By nesting, U_{r-1} is spanned by the inclusion indicators h_R with $|R| \leqslant r-1$. Fix such a set R, and write

$$a = |R \cap T|, \qquad b = |R \setminus T|, \qquad a+b = |R|.$$

All probabilities and conditional expectations below are taken under $S \sim \nu_{d,k}$. Conditional on $L_T(S) = \ell$, the set $S \cap T$ is uniformly distributed among the ℓ-subsets of T, while $S \setminus T$ is uniformly distributed among the $(k-\ell)$-subsets of $[d] \setminus T$. Hence, the conditional probability that $R \cap T$ is contained in $S \cap T$ is, with the convention that $\binom{u}{v} = 0$ whenever $v < 0$ or $v > u$,

$$\frac{\binom{r-a}{\ell-a}}{\binom{r}{\ell}} = \frac{\binom{\ell}{a}}{\binom{r}{a}},$$

and the conditional probability that $R \setminus T$ is contained in $S \setminus T$ is

$$\frac{\binom{d-r-b}{k-\ell-b}}{\binom{d-r}{k-\ell}} = \frac{\binom{k-\ell}{b}}{\binom{d-r}{b}}.$$

The two selections are conditionally independent, and therefore

$$\mathbb{E}\left[h_R(S) \mid L_T(S) = \ell\right] = \frac{\binom{\ell}{a}}{\binom{r}{a}} \frac{\binom{k-\ell}{b}}{\binom{d-r}{b}}.$$

As a function of ℓ, the right-hand side is a polynomial of degree at most $a+b = |R| \leqslant r-1$. Indeed, $\binom{\ell}{a}$ and $\binom{k-\ell}{b}$ are polynomials in ℓ of degrees a and b, respectively. The orthogonality property of $H_{r,d,k}$ established in Lemma B.2 now gives

$$\begin{aligned}
\langle \phi_{r,T}, h_R \rangle &= \mathbb{E}_{S \sim \nu_{d,k}}\left[H_{r,d,k}(L_T(S)) h_R(S)\right] \\
&= \mathbb{E}_{S \sim \nu_{d,k}}\left[H_{r,d,k}(L_T(S)) \mathbb{E}\left[h_R(S) \mid L_T(S)\right]\right] \\
&= 0.
\end{aligned}$$

Thus, $\phi_{r,T} \perp U_{r-1}$. Together with $\phi_{r,T} \in U_r$, this proves that $\phi_{r,T} \in V_r$. $\square$

B.4 Proof of Theorem B.1

The assertion for $r = 0$ follows from $\phi_{0,\varnothing} \equiv 1$, $V_0 = \mathrm{span}\{\mathbf{1}\}$, and $A_{0,d,k} = 1$. Fix $1 \leqslant r \leqslant m$. By Lemma B.3, all the functions $\phi_{r,T}$, with $T \subseteq [d]$ and $|T| = r$, belong to V_r and have unit norm.

For $g \in V_r$, define the frame operator

$$\mathcal{F}_r g = \sum_{\substack{T \subseteq [d] \\ |T| = r}} \langle g, \phi_{r,T} \rangle \, \phi_{r,T}.$$

A finite family is a tight frame for V_r precisely when its frame operator is a positive scalar multiple of the identity on V_r. We now prove that this is the case and determine the scalar.

The kernel of $\mathcal{F}_r$, relative to the uniform inner product, is

$$\mathcal{Q}_r(S, S') = \sum_{\substack{T \subseteq [d] \\ |T| = r}} \phi_{r,T}(S)\phi_{r,T}(S'),$$

in the sense that

$$(\mathcal{F}_r g)(S) = \frac{1}{N} \sum_{S' \in \mathcal{S}_{d,k}} \mathcal{Q}_r(S, S') g(S').$$

The permutation equivariance established in Lemma B.3 gives

$$\phi_{r,T}(\pi(S)) = \phi_{r,\pi^{-1}(T)}(S).$$

Reindexing the sum defining $\mathcal{Q}_r$ therefore yields

$$\mathcal{Q}_r(\pi(S), \pi(S')) = \mathcal{Q}_r(S, S')$$

for every permutation π of $[d]$.

All the sets considered below belong to $\mathcal{S}_{d,k}$ and therefore have the same cardinality k. For two pairs (S, S') and (R, R'), there exists a permutation π of $[d]$ satisfying

$$\pi(S) = R \qquad \text{and} \qquad \pi(S') = R'$$

if and only if

$$|S \cap S'| = |R \cap R'|.$$

Indeed, because all four sets have cardinality k, equality of the intersection sizes implies equality of the cardinalities of the four corresponding parts: the intersection, the two set differences, and the complement of the union. Bijections between these four pairs of parts combine to form the required permutation of $[d]$. Consequently, $\mathcal{Q}_r(S, S')$ depends only on $\mathrm{dist}_J(S, S')$. Thus, there exist constants $q_{r,0}, \ldots, q_{r,m}$ such that

$$\mathcal{Q}_r(S, S') = \sum_{\ell=0}^{m} q_{r,\ell} \mathbf{1}_{\{\mathrm{dist}_J(S,S')=\ell\}}.$$

For $0 \leqslant \ell \leqslant m$, let A_ℓ denote the Johnson distance matrix indexed by $\mathcal{S}_{d,k}$, with entries

$$(A_\ell)_{S,S'} = \mathbf{1}_{\{\mathrm{dist}_J(S,S')=\ell\}}.$$

We regard A_ℓ as an operator on functions $g : \mathcal{S}_{d,k} \to \mathbb{R}$, acting by

$$\begin{aligned}(A_\ell g)(S) &= \sum_{S' \in \mathcal{S}_{d,k}} (A_\ell)_{S,S'} g(S') \\ &= \sum_{\substack{S' \in \mathcal{S}_{d,k} \\ \mathrm{dist}_J(S,S')=\ell}} g(S').\end{aligned}$$

Thus, $A_\ell g$ sums the values of g over all subsets at Johnson distance ℓ from S. Using the preceding expression for $\mathcal{Q}_r$, we obtain, for every $g \in V_r$,

$$\begin{aligned}(\mathcal{F}_r g)(S) &= \frac{1}{N} \sum_{S' \in \mathcal{S}_{d,k}} \sum_{\ell=0}^{m} q_{r,\ell} (A_\ell)_{S,S'} g(S') \\ &= \frac{1}{N} \sum_{\ell=0}^{m} q_{r,\ell} (A_\ell g)(S).\end{aligned}$$

Equivalently, as an operator on V_r,

$$\mathcal{F}_r = \frac{1}{N} \sum_{\ell=0}^{m} q_{r,\ell} A_\ell \bigg|_{V_r}.$$

The Johnson graph is distance-regular, so each distance matrix A_ℓ is a polynomial in its adjacency matrix A_1 (Brouwer et al., 1989). Since $A_1 = k(d-k)K$, Proposition 3.1 then implies that A_ℓ acts as multiplication by a scalar on every harmonic space V_j. Since $\mathcal{F}_r$ is a linear combination of these matrices, its restriction to V_r also acts as multiplication by a scalar. Thus, there exists $c_r \in \mathbb{R}$ such that

$$\mathcal{F}_r g = c_r g, \qquad g \in V_r.$$

We determine c_r by taking the trace on V_r. The rank-one operator

$$g \longmapsto \langle g, \phi_{r,T} \rangle \, \phi_{r,T}$$

has trace $\|\phi_{r,T}\|_2^2 = 1$. Since there are $\binom{d}{r}$ subsets $T \subseteq [d]$ of cardinality r,

$$\begin{aligned} c_r \dim(V_r) &= \operatorname{tr}(\mathcal{F}_r) \\ &= \sum_{\substack{T \subseteq [d] \\ |T| = r}} \|\phi_{r,T}\|_2^2 = \binom{d}{r}. \end{aligned}$$

Hence,

$$c_r = \frac{\binom{d}{r}}{\dim(V_r)} = A_{r,d,k}^{-1} > 0.$$

The second equality uses (35). It follows that

$$\mathcal{F}_r = A_{r,d,k}^{-1} I_{V_r},$$

which proves tightness.

The positivity of the scalar also implies that the frame functions span V_r. For every $v \in V_r$,

$$v = \mathcal{F}_r(A_{r,d,k} v).$$

By definition, every value of $\mathcal{F}_r$ is a linear combination of the functions $\phi_{r,T}$. Hence, V_r is contained in their span. The reverse inclusion follows from Lemma B.3, since every $\phi_{r,T}$ belongs to V_r. Therefore, the frame functions span V_r.

For an arbitrary $f \in L^2(\mathcal{S}_{d,k}, \nu_{d,k})$, the fact that $\phi_{r,T} \in V_r$ gives

$$\langle f, \phi_{r,T} \rangle = \langle P_r f, \phi_{r,T} \rangle.$$

Applying

$$A_{r,d,k} \mathcal{F}_r = I_{V_r}$$

to $P_r f$ proves (36). Taking the inner product of (36) with $P_r f$, and using again $\langle P_r f, \phi_{r,T} \rangle = \langle f, \phi_{r,T} \rangle$, gives (37).

It remains to establish the projection identity

$$\phi_{r,T} = \frac{P_r h_T}{\|P_r h_T\|_2},$$

announced in Subsection B.2. Let G_T be the group of permutations of $[d]$ that preserve T setwise, and define

$$(\mathcal{U}_\pi f)(S) = f(\pi^{-1}(S)).$$

The action of a permutation on an inclusion indicator simply relabels its indexing subset. Indeed, for every $R \subseteq [d]$,

$$\begin{aligned}(\mathcal{U}_\pi h_R)(S) &= h_R(\pi^{-1}(S)) \\ &= \mathbf{1}_{\{R \subseteq \pi^{-1}(S)\}} = \mathbf{1}_{\{\pi(R) \subseteq S\}} = h_{\pi(R)}(S).\end{aligned}$$

Since π maps the collection of j-subsets of $[d]$ bijectively onto itself, the action $\mathcal{U}_\pi$ permutes the generators

$$\{h_R : R \subseteq [d],\ |R| = j\}$$

of U_j. Consequently,

$$\mathcal{U}_\pi(U_j) = U_j, \qquad 0 \leqslant j \leqslant m.$$

Now let $\pi \in G_T$. By definition of G_T, the permutation π preserves T setwise, that is, $\pi(T) = T$. Applying the preceding identity with $R = T$ gives

$$\mathcal{U}_\pi h_T = h_{\pi(T)} = h_T.$$

Thus, h_T is invariant under every permutation in G_T.

The action also preserves the uniform inner product. Indeed, the map $S \mapsto \pi^{-1}(S)$ is a bijection of $\mathcal{S}_{d,k}$, and therefore

$$\begin{aligned}\langle \mathcal{U}_\pi f, \mathcal{U}_\pi g\rangle &= \frac{1}{N} \sum_{S \in \mathcal{S}_{d,k}} f(\pi^{-1}(S)) g(\pi^{-1}(S)) \\ &= \frac{1}{N} \sum_{R \in \mathcal{S}_{d,k}} f(R) g(R) = \langle f, g\rangle .\end{aligned}$$

The isometry $\mathcal{U}_\pi$ also preserves $U_{r-1}^\perp$. Indeed, if $f \in U_{r-1}^\perp$ and $u \in U_{r-1}$, then

$$\langle \mathcal{U}_\pi f, u\rangle = \langle f, \mathcal{U}_{\pi^{-1}} u\rangle = 0,$$

because $\mathcal{U}_{\pi^{-1}} u \in U_{r-1}$. Since $\mathcal{U}_\pi$ also preserves U_r, it preserves

$$V_r = U_r \cap U_{r-1}^\perp.$$

We next show that P_r commutes with this action. Since $P_r f \in V_r$ and V_r is invariant under $\mathcal{U}_\pi$,

$$\mathcal{U}_\pi P_r f \in V_r.$$

Moreover, for every $v \in V_r$,

$$\langle \mathcal{U}_\pi (f - P_r f), v\rangle = \langle f - P_r f, \mathcal{U}_{\pi^{-1}} v\rangle = 0,$$

because $\mathcal{U}_{\pi^{-1}}v \in V_r$, whereas $f - P_r f \perp V_r$. Thus,

$$\mathcal{U}_\pi f = \mathcal{U}_\pi P_r f + \mathcal{U}_\pi (f - P_r f)$$

is the orthogonal decomposition of $\mathcal{U}_\pi f$ into a component in V_r and a component in $V_r^\perp$. Therefore,

$$P_r(\mathcal{U}_\pi f) = \mathcal{U}_\pi (P_r f),$$

or equivalently,

$$P_r \mathcal{U}_\pi = \mathcal{U}_\pi P_r.$$

Applying this identity to h_T gives, for every $\pi \in G_T$,

$$\mathcal{U}_\pi (P_r h_T) = P_r(\mathcal{U}_\pi h_T) = P_r h_T.$$

Hence, $P_r h_T$ is invariant under G_T.

The G_T-orbits in $\mathcal{S}_{d,k}$ are the level sets of $L_T(S) = |S \cap T|$. Consequently, if $g \in V_r$ is G_T-invariant, there exist values $y_0, \dots, y_r$ such that $g(S) = y_\ell$ whenever $L_T(S) = \ell$.

Because $r \leqslant m$, all overlap values $0, \dots, r$ are attainable. Lagrange interpolation therefore gives a unique polynomial q of degree at most r satisfying $q(\ell) = y_\ell$ for $0 \leqslant \ell \leqslant r$. Hence,

$$g(S) = q(L_T(S)).$$

For $0 \leqslant j \leqslant r - 1$, (41) gives

$$S \longmapsto \binom{L_T(S)}{j} \in U_j \subseteq U_{r-1}.$$

Since the polynomials $\ell \mapsto \binom{\ell}{j}$, $0 \leqslant j \leqslant r-1$, form a basis of the polynomials of degree at most $r-1$, it follows that

$$S \longmapsto p(L_T(S)) \in U_{r-1}$$

for every polynomial p of degree at most $r-1$. Since $g \in V_r \subseteq U_{r-1}^\perp$, every such polynomial p satisfies

$$0 = \langle g, p \circ L_T \rangle = \sum_{\ell=0}^{r} w_{r,d,k}(\ell) q(\ell) p(\ell).$$

Thus q belongs to the orthogonal complement of the polynomials of degree at most $r-1$ within the polynomials of degree at most r. This orthogonal complement is one-dimensional, and Lemma B.2 shows that it is spanned by $H_{r,d,k}$. Therefore q is a scalar multiple of $H_{r,d,k}$. It follows that every G_T-invariant vector in V_r is a scalar multiple of $\phi_{r,T}$. Conversely, Lemma B.3 shows that $\phi_{r,T}$ is a G_T-invariant unit vector in V_r. Hence the G_T-invariant subspace of V_r is exactly $\operatorname{span}\{\phi_{r,T}\}$. Since $P_r h_T$ is G_T-invariant, there exists $\beta_{r,T} \in \mathbb{R}$ such that

$$P_r h_T = \beta_{r,T} \phi_{r,T}.$$

It remains to determine the proportionality constant and its sign. Using self-adjointness of P_r, together with $P_r \phi_{r,T} = \phi_{r,T}$, we obtain

$$\begin{aligned} \beta_{r,T} &= \langle P_r h_T, \phi_{r,T} \rangle = \langle h_T, P_r \phi_{r,T} \rangle = \langle h_T, \phi_{r,T} \rangle \\ &= \mathbb{P}_{S \sim \nu_{d,k}}(T \subseteq S) H_{r,d,k}(r) > 0. \end{aligned}$$

Because $\|\phi_{r,T}\|_2 = 1$, it follows that

$$\|P_r h_T\|_2 = \beta_{r,T}.$$

Dividing the proportionality identity by this norm gives

$$\phi_{r,T} = \frac{P_r h_T}{\|P_r h_T\|_2},$$

as claimed.

B.5 Frame representation of the projection estimator

Applying the reconstruction identity in Theorem B.1 separately on the harmonic levels $V_1, \ldots, V_D$ gives the alternative representation

$$\widehat{f}_D^{\mathrm{proj}} = \sum_{r=1}^{D} A_{r,d,k} \sum_{\substack{T \subseteq [d] \\ |T|=r}} \widehat{\beta}_{r,T} \phi_{r,T}, \qquad \widehat{\beta}_{r,T} = \frac{1}{n} \sum_{j=1}^{n} Y_j \phi_{r,T}(S_j).$$

This formula is explicit and equivariant under relabeling of the ground set, but it uses the redundant canonical frame and is not intended as a computationally optimal implementation.

C Projection and minimax proofs

C.1 Proof of Proposition 4.3

When $D = 0$, one has $W_0 = \{0\}$, $\widehat{f}_0^{\mathrm{proj}} = 0$, and $p_0 = 0$. Since $\gamma_{1,d} = 1$,

$$\mathbb{E}_{f,P}\left[\left\|\widehat{f}_0^{\mathrm{proj}} - f\right\|_2^2\right] = \|f\|_2^2 \leqslant B = b_0(s, B).$$

Thus, the result holds for $D = 0$. We henceforth fix $1 \leqslant D \leqslant m$, $f \in \mathcal{E}_{d,k}^{s}(B)$, and $P \in \mathcal{P}_{\sigma^2}$.

Let (S, Y) denote a generic observation having the same distribution as each pair (S_j, Y_j). Thus,

$$S \sim \nu_{d,k}, \qquad Y = f(S) + \varepsilon, \qquad \varepsilon \sim P,$$

where ε is independent of S.

Let $e_1, \ldots, e_{p_D}$ be an orthonormal basis of W_D, and write

$$\theta_\ell = \langle f, e_\ell \rangle, \qquad 1 \leqslant \ell \leqslant p_D.$$

Since the noise is centered and independent of the design,

$$\mathbb{E}_{f,P}[\widehat{\theta}_\ell] = \mathbb{E}_{f,P}[Y e_\ell(S)] = \mathbb{E}_{S \sim \nu_{d,k}}[f(S) e_\ell(S)] = \theta_\ell.$$

Hence,

$$P_{\leqslant D}^{\circ} f = \sum_{\ell=1}^{p_D} \theta_\ell e_\ell$$

is the expectation of the empirical projection estimator.

We next establish the constant-diagonal identity (11). For a permutation π of $[d]$, recall the action

$$(\mathcal{U}_\pi g)(S) = g(\pi^{-1}(S)).$$

As shown in Subsection B.4, $\mathcal{U}_\pi$ preserves every harmonic space V_r and is an isometry for the uniform inner product. It therefore preserves $W_D = V_1 \oplus^\perp \cdots \oplus^\perp V_D$. Thus, if $e_1, \ldots, e_{p_D}$ is an orthonormal basis of W_D, then so is

$$\mathcal{U}_{\pi^{-1}} e_1, \ldots, \mathcal{U}_{\pi^{-1}} e_{p_D}.$$

Since the projection kernel is independent of the orthonormal basis used to represent W_D, it follows that

$$\begin{aligned}
\mathcal{K}_D^\circ(\pi(S), \pi(S')) &= \sum_{\ell=1}^{p_D} e_\ell(\pi(S)) e_\ell(\pi(S')) \\
&= \sum_{\ell=1}^{p_D} (\mathcal{U}_{\pi^{-1}} e_\ell)(S)(\mathcal{U}_{\pi^{-1}} e_\ell)(S') \\
&= \mathcal{K}_D^\circ(S, S').
\end{aligned}$$

The permutation action on $\mathcal{S}_{d,k}$ is transitive. Indeed, for any $S, R \in \mathcal{S}_{d,k}$, choose a bijection from S onto R and a bijection from $[d] \setminus S$ onto $[d] \setminus R$. Together, these two bijections define a permutation π of $[d]$ such that $\pi(S) = R$. The preceding invariance identity then gives

$$\mathcal{K}_D^\circ(R, R) = \mathcal{K}_D^\circ(S, S).$$

Thus, the diagonal $S \mapsto \mathcal{K}_D^\circ(S, S)$ is constant over $\mathcal{S}_{d,k}$.

Its value is determined by averaging under the uniform measure:

$$\begin{aligned}
\mathbb{E}_{S \sim \nu_{d,k}}[\mathcal{K}_D^\circ(S, S)] &= \sum_{\ell=1}^{p_D} \mathbb{E}_{S \sim \nu_{d,k}}[e_\ell(S)^2] \\
&= \sum_{\ell=1}^{p_D} \|e_\ell\|_2^2 = p_D.
\end{aligned}$$

Consequently,

$$\mathcal{K}_D^\circ(S, S) = \sum_{\ell=1}^{p_D} e_\ell(S)^2 = p_D, \qquad S \in \mathcal{S}_{d,k},$$

which proves (11).

We now bound the coefficient-estimation term. By the definitions of the estimator and of the population projection,

$$\widehat{f}_D^{\text{proj}} - P_{\leqslant D}^\circ f = \sum_{\ell=1}^{p_D} (\widehat{\theta}_\ell - \theta_\ell) e_\ell.$$

Since the basis is orthonormal, the squared norm is the sum of the squared coefficient differences. Using also the unbiasedness of $\widehat{\theta}_\ell$, we obtain

$$\begin{aligned}
\mathbb{E}_{f,P}\left[\left\|\widehat{f}_D^{\text{proj}} - P_{\leqslant D}^\circ f\right\|_2^2\right] &= \sum_{\ell=1}^{p_D} \mathbb{E}_{f,P}\left[(\widehat{\theta}_\ell - \theta_\ell)^2\right] \\
&= \sum_{\ell=1}^{p_D} \mathrm{Var}_{f,P}(\widehat{\theta}_\ell).
\end{aligned}$$

Since the observations are i.i.d.,

$$\mathrm{Var}_{f,P}(\widehat{\theta}_\ell) = \frac{1}{n}\mathrm{Var}_{f,P}(Ye_\ell(S)) \leqslant \frac{1}{n}\mathbb{E}_{f,P}\left[Y^2 e_\ell(S)^2\right].$$

Summing over ℓ and using the constant-diagonal identity gives

$$\begin{aligned}\mathbb{E}_{f,P}\left[\left\|\widehat{f}_D^{\mathrm{proj}} - P^\circ_{\leqslant D} f\right\|_2^2\right] &\leqslant \frac{1}{n}\mathbb{E}_{f,P}\left[Y^2\sum_{\ell=1}^{p_D} e_\ell(S)^2\right] \\ &= \frac{p_D}{n}\mathbb{E}_{f,P}[Y^2].\end{aligned}$$

The last equality is precisely where $\mathcal{K}^\circ_D(S,S) = p_D$ is used.

Because $Y = f(S) + \varepsilon$, and because ε is centered and independent of S,

$$\mathbb{E}_{f,P}[f(S)\varepsilon] = 0.$$

Therefore,

$$\mathbb{E}_{f,P}[Y^2] = \|f\|_2^2 + \mathbb{E}_P[\varepsilon^2] \leqslant B + \sigma^2.$$

We conclude that

$$\mathbb{E}_{f,P}\left[\left\|\widehat{f}_D^{\mathrm{proj}} - P^\circ_{\leqslant D} f\right\|_2^2\right] \leqslant (\sigma^2 + B)\frac{p_D}{n}.$$

Finally, for every realization of the sample,

$$\widehat{f}_D^{\mathrm{proj}} - P^\circ_{\leqslant D} f \in W_D,$$

whereas

$$f - P^\circ_{\leqslant D} f \in W_D^\perp.$$

Pythagoras therefore gives

$$\left\|\widehat{f}_D^{\mathrm{proj}} - f\right\|_2^2 = \left\|\widehat{f}_D^{\mathrm{proj}} - P^\circ_{\leqslant D} f\right\|_2^2 + \|f - P^\circ_{\leqslant D} f\|_2^2.$$

If $D < m$, the harmonic tail bound (6) gives

$$\|f - P^\circ_{\leqslant D} f\|_2^2 = \sum_{r>D}\|P_r f\|_2^2 \leqslant b_D(s,B).$$

If $D = m$, then $W_m = V_0^\perp$; since f is centered, $f \in W_m$ and $P^\circ_{\leqslant m} f = f$. Hence, the approximation term is zero.

Combining the estimation and approximation bounds, and then taking the supremum over $f \in \mathcal{E}^s_{d,k}(B)$ and $P \in \mathcal{P}_{\sigma^2}$, proves (10).

C.2 Proof of Theorem 4.4

Write

$$a_D = a_D(s,B), \qquad v_D = v_D(n,\sigma^2), \qquad \mathfrak{R} = \mathfrak{R}_{n,d,k}(s,B,\sigma^2).$$

Lower bound. Fix $1 \leqslant D \leqslant m$, and put $p = p_D$. For every $g \in W_D$, the spectral decomposition of the Johnson–Sobolev energy and the monotonicity of $r \mapsto \gamma_{r,d}$ give

$$I^{(2,s)}_{J,d,k}(g) = \sum_{r=1}^{D} \gamma^s_{r,d} \|P_r g\|_2^2 \leqslant \gamma^s_{D,d} \|g\|_2^2.$$

Consequently,

$$\mathcal{B}_D := \left\{ g \in W_D : \|g\|_2^2 \leqslant a_D \right\} \subseteq \mathcal{E}^s_{d,k}(B).$$

We restrict the noise distribution to the Gaussian member $G = N(0, \sigma^2)$ of $\mathcal{P}_{\sigma^2}$. Since the design law does not depend on the response function, the chain rule for Kullback–Leibler divergence gives, for $g, g' \in W_D$,

$$\begin{aligned} \mathrm{KL}(\mathbb{P}_{g,G}, \mathbb{P}_{g',G}) &= \frac{1}{2\sigma^2} \sum_{j=1}^{n} \mathbb{E}_{S_j \sim \nu_{d,k}} \left[\{ g(S_j) - g'(S_j) \}^2 \right] \\ &= \frac{n}{2\sigma^2} \|g - g'\|_2^2. \end{aligned} \tag{42}$$

Let $e_1, \ldots, e_p$ be an orthonormal basis of W_D. For $\omega, \omega' \in \{0,1\}^p$, define their Hamming distance by

$$\mathrm{Ham}(\omega, \omega') = \sum_{\ell=1}^{p} \mathbf{1}_{\{\omega_\ell \neq \omega'_\ell\}}.$$

Massart's version of the Varshamov–Gilbert lemma (Massart, 2007, Lemma 4.7), applied with $\alpha = 1/2$, provides a set $\Omega \subseteq \{0,1\}^p$ such that

$$\mathrm{Ham}(\omega, \omega') > \frac{p}{4} \quad \text{for all distinct } \omega, \omega' \in \Omega, \qquad \log |\Omega| > \frac{p}{8}. \tag{43}$$

Translating Ω by any one of its elements, using coordinatewise addition modulo two, preserves these properties. We may therefore assume that $0 \in \Omega$.

For $\omega \in \Omega$, set

$$g_\omega = \delta \sum_{\ell=1}^{p} \omega_\ell e_\ell, \qquad \delta^2 = c_1 \min\left\{ \frac{a_D}{p}, \frac{\sigma^2}{n} \right\},$$

where $c_1 \in (0,1]$ is a sufficiently small universal constant. Orthonormality gives

$$\|g_\omega\|_2^2 \leqslant p\delta^2 \leqslant a_D,$$

so every g_ω belongs to $\mathcal{B}_D$. Moreover,

$$\|g_\omega - g_{\omega'}\|_2^2 = \delta^2 \,\mathrm{Ham}(\omega, \omega') > \frac{p\delta^2}{4}, \qquad \omega \neq \omega'. \tag{44}$$

Suppose that $p \geqslant 8$, and let $M = |\Omega| - 1$. Since $0 \in \Omega$, the zero response may be used as the reference distribution. By (42),

$$\mathrm{KL}(\mathbb{P}_{g_\omega,G}, \mathbb{P}_{0,G}) \leqslant \frac{c_1 p}{2}, \qquad \omega \in \Omega \setminus \{0\}.$$

Moreover, $\log|\Omega| > p/8$ and $p \geqslant 8$ imply $|\Omega| \geqslant 3$, and hence $M = |\Omega| - 1 \geqslant |\Omega|/2$. Therefore,

$$\log M \geqslant \log|\Omega| - \log 2 > \frac{p}{8} - \log 2 \geqslant \frac{1-\log 2}{8}\, p.$$

Thus, by choosing $c_1 > 0$ sufficiently small, the preceding Kullback–Leibler divergences are bounded by a sufficiently small fixed fraction of $\log M$, as required by Fano's inequality.

By (44), the signals are separated by more than $2s_0$ in L^2, where

$$s_0 = \frac{\sqrt{p}\,\delta}{4}.$$

Fano's inequality in the form stated by Tsybakov (2009, Theorem 2.5) gives a universal constant $c_{\mathrm{F}} > 0$ such that

$$\inf_{\widehat{f}} \sup_{\omega\in\Omega} \mathbb{P}_{g_\omega,G}\left(\|\widehat{f} - g_\omega\|_2 \geqslant s_0\right) \geqslant c_{\mathrm{F}}.$$

Since

$$\mathbb{E}_{g_\omega,G}\left[\|\widehat{f} - g_\omega\|_2^2\right] \geqslant s_0^2 \mathbb{P}_{g_\omega,G}\left(\|\widehat{f} - g_\omega\|_2 \geqslant s_0\right),$$

it follows that

$$R_n^\star(\mathcal{E}^s_{d,k}(B)) \geqslant c_{\mathrm{F}} s_0^2 = \frac{c_{\mathrm{F}} p \delta^2}{16} = \frac{c_{\mathrm{F}} c_1}{16} \min\left\{a_D, \frac{\sigma^2 p}{n}\right\} = c_{\geqslant 8} \min\{a_D, v_D\},$$

where

$$c_{\geqslant 8} = \frac{c_{\mathrm{F}} c_1}{16} > 0$$

is universal. The condition $p \geqslant 8$ is needed only for this uniform application of Fano's inequality; the packing (43) exists for every $p \geqslant 1$.

It remains to cover the finitely many dimensions $1 \leqslant p < 8$. Let e_1 be any unit vector in W_D and consider

$$g_+ = \delta_0 e_1, \qquad g_- = -\delta_0 e_1, \qquad \delta_0^2 = \min\left\{a_D, \frac{\sigma^2}{n}\right\}.$$

Both signals belong to $\mathcal{B}_D$, and

$$\|g_+ - g_-\|_2 = 2\delta_0.$$

Moreover, (42) gives

$$\mathrm{KL}(\mathbb{P}_{g_+,G}, \mathbb{P}_{g_-,G}) = \frac{n}{2\sigma^2}\|g_+ - g_-\|_2^2 = \frac{2n\delta_0^2}{\sigma^2} \leqslant 2.$$

To connect testing and estimation, associate with any estimator $\widehat{f}$ the minimum-distance test

$$\psi_{\widehat{f}} \in \operatorname*{arg\,min}_{\eta\in\{+,-\}} \|\widehat{f} - g_\eta\|_2,$$

with an arbitrary rule for breaking ties. Since the two signals are separated by $2\delta_0$,

$$\{\psi_{\widehat{f}} \neq \eta\} \subseteq \{\|\widehat{f} - g_\eta\|_2 \geqslant \delta_0\}, \qquad \eta \in \{+,-\}.$$

Consequently,

$$\inf_{\widehat{f}} \max_{\eta \in \{+,-\}} \mathbb{P}_{g_\eta, G} \left(\|\widehat{f} - g_\eta\|_2 \geqslant \delta_0 \right) \geqslant p_{e,1},$$

where

$$p_{e,1} := \inf_{\psi} \max_{\eta \in \{+,-\}} \mathbb{P}_{g_\eta, G}(\psi \neq \eta),$$

the infimum being taken over all measurable tests ψ with values in $\{+,-\}$. By the Kullback–Leibler version of the two-hypothesis bound (Tsybakov, 2009, Theorem 2.2(iii)),

$$p_{e,1} \geqslant \frac{1}{4} \exp(-2).$$

For every estimator $\widehat{f}$ and every $\eta \in \{+,-\}$,

$$\mathbb{E}_{g_\eta, G} \left[\|\widehat{f} - g_\eta\|_2^2 \right] \geqslant \delta_0^2 \mathbb{P}_{g_\eta, G} \left(\|\widehat{f} - g_\eta\|_2 \geqslant \delta_0 \right).$$

Combining this inequality with the preceding reduction from estimation to testing gives

$$R_n^\star(\mathcal{E}_{d,k}^s(B)) \geqslant \delta_0^2 p_{e,1} \geqslant \frac{\delta_0^2}{4} \exp(-2) = c_3 \min\left\{ a_D, \frac{\sigma^2}{n} \right\},$$

where $c_3 := \frac{1}{4}\exp(-2) > 0$ is universal. Since $1 \leqslant p < 8$,

$$\min\left\{ a_D, \frac{\sigma^2}{n} \right\} \geqslant \frac{1}{8} \min\left\{ a_D, \frac{\sigma^2 p}{n} \right\} = \frac{1}{8} \min\{a_D, v_D\}.$$

Consequently, with

$$c = \min\left\{ c_{\geqslant 8}, \frac{c_3}{8} \right\} > 0,$$

we have proved, for every $1 \leqslant D \leqslant m$,

$$R_n^\star(\mathcal{E}_{d,k}^s(B)) \geqslant c \min\{a_D, v_D\}.$$

Maximizing over D gives

$$R_n^\star(\mathcal{E}_{d,k}^s(B)) \geqslant c\,\mathfrak{R},$$

which proves the lower bound.

Upper bound. The sequence $(v_D)_{D=1}^m$ is increasing because

$$p_D = \binom{d}{D} - 1$$

is increasing for $D \leqslant m \leqslant d/2$. Similarly, $(a_D)_{D=1}^m$ is decreasing because $D \mapsto \gamma_{D,d}$ is increasing on this range. Adopt the conventions

$$v_0 = 0, \qquad a_0 = +\infty,$$

and let

$$D_\star = \max\left\{ 0 \leqslant D \leqslant m : v_D \leqslant a_D \right\}.$$

The set in this definition is nonempty because it contains $D = 0$.

We first prove the lower comparison asserted after (12). Fix $0 \leqslant D \leqslant m$. If $D \geqslant 1$ and $1 \leqslant q \leqslant D$, then

$$\min\{v_q, a_q\} \leqslant v_q \leqslant v_D.$$

If $D < m$ and $D < q \leqslant m$, then

$$\min\{v_q, a_q\} \leqslant a_q \leqslant a_{D+1} = b_D(s, B).$$

The endpoint cutoffs are covered as well: when $D = 0$, only the second comparison is needed, whereas when $D = m$, only the first is needed and $b_m(s, B) = 0$. Taking the maximum over q shows that

$$\mathfrak{R} \leqslant v_D + b_D(s, B).$$

Since this holds for every cutoff D,

$$\mathfrak{R} \leqslant \min_{0 \leqslant D \leqslant m} \{v_D + b_D(s, B)\}.$$

We now prove the reverse comparison up to the factor two.

If $D_\star = m$, then $v_m \leqslant a_m$, and hence

$$\min_{0 \leqslant D \leqslant m} \left\{v_D + \mathbf{1}_{\{D<m\}} a_{D+1}\right\} \leqslant v_m = \min\{v_m, a_m\} \leqslant \mathfrak{R}.$$

Suppose now that $D_\star < m$. By maximality,

$$v_{D_\star+1} > a_{D_\star+1},$$

so

$$\mathfrak{R} \geqslant \min\{v_{D_\star+1}, a_{D_\star+1}\} = a_{D_\star+1}.$$

If $D_\star \geqslant 1$, then $v_{D_\star} \leqslant a_{D_\star}$, and therefore

$$\mathfrak{R} \geqslant \min\{v_{D_\star}, a_{D_\star}\} = v_{D_\star}.$$

Using $D = D_\star$ in the minimum consequently gives

$$\min_{0 \leqslant D \leqslant m} \left\{v_D + \mathbf{1}_{\{D<m\}} a_{D+1}\right\} \leqslant v_{D_\star} + a_{D_\star+1} \leqslant 2\mathfrak{R}.$$

If $D_\star = 0$, maximality gives $v_1 > a_1$, and the candidate $D = 0$ instead yields

$$v_0 + a_1 = a_1 = \min\{v_1, a_1\} \leqslant \mathfrak{R}.$$

Thus, in every case,

$$\min_{0 \leqslant D \leqslant m} \left\{v_D + \mathbf{1}_{\{D<m\}} a_{D+1}\right\} \leqslant 2\mathfrak{R}. \tag{45}$$

Finally, set

$$\tau = 1 + \frac{B}{\sigma^2} \geqslant 1.$$

For every $0 \leqslant D \leqslant m$,

$$(\sigma^2 + B)\frac{p_D}{n} = \tau v_D, \qquad b_D(s, B) = \mathbf{1}_{\{D<m\}} a_{D+1}.$$

Moreover, since $\tau \geqslant 1$,

$$v_D + b_D(s,B) \leqslant \tau v_D + b_D(s,B) \leqslant \tau\{v_D + b_D(s,B)\}.$$

Taking minima over D, and using the two comparisons proved above together with (45), gives

$$\begin{aligned}\mathfrak{R} \leqslant \min_{0\leqslant D\leqslant m}\{v_D + b_D(s,B)\} &\leqslant \min_{0\leqslant D\leqslant m}\{\tau v_D + b_D(s,B)\}\\ &\leqslant \tau \min_{0\leqslant D\leqslant m}\{v_D + b_D(s,B)\}\\ &\leqslant 2\tau\mathfrak{R}.\end{aligned}$$

This is precisely the deterministic comparison stated after (12).

Proposition 4.3, optimized over $0 \leqslant D \leqslant m$, now gives

$$\begin{aligned}R_n^\star(\mathcal{E}^s_{d,k}(B)) &\leqslant \min_{0\leqslant D\leqslant m}\{\tau v_D + b_D(s,B)\}\\ &\leqslant 2\tau\mathfrak{R}\\ &= 2\left(1+\frac{B}{\sigma^2}\right)\mathfrak{R},\end{aligned}$$

which proves the upper bound in (13).

If $B/\sigma^2 \leqslant C_{\mathrm{snr}}$, then the two bounds just proved yield

$$c\,\mathfrak{R}_{n,d,k}(s,B,\sigma^2) \leqslant R_n^\star(\mathcal{E}^s_{d,k}(B)) \leqslant 2(1+C_{\mathrm{snr}})\mathfrak{R}_{n,d,k}(s,B,\sigma^2).$$

The constant $c > 0$ is universal, and the upper constant depends only on C_{snr}. The comparison is therefore uniform over all d, k, n, s, B, σ^2 satisfying the assumptions of the theorem and $B/\sigma^2 \leqslant C_{\mathrm{snr}}$.

D Random-design proofs

D.1 Proof of Proposition 5.1

Fix $1 \leqslant D \leqslant m$. By (11), the leverage score is constant, i.e., $\|x_D(S)\|_2^2 = p_D$, $S \in \mathcal{S}_{d,k}$. Write

$$\widehat{G}_D = \sum_{j=1}^n Z_j, \qquad Z_j = \frac{1}{n}x_D(S_j)x_D(S_j)^\top.$$

The matrices $Z_1, \ldots, Z_n$ are independent and positive semidefinite. Moreover, orthonormality gives

$$\mathbb{E}_{S_j\sim\nu_{d,k}}[Z_j] = \frac{1}{n}I_{p_D}, \qquad \sum_{j=1}^n \mathbb{E}_{S_j\sim\nu_{d,k}}[Z_j] = I_{p_D}.$$

Since a rank-one matrix $xx^\top$ has unique nonzero eigenvalue $\|x\|_2^2$, the constant-leverage identity also gives

$$\lambda_{\max}(Z_j) = \frac{1}{n}\|x_D(S_j)\|_2^2 = \frac{p_D}{n} \qquad \text{almost surely.}$$

Thus, in the notation of the matrix Chernoff inequalities (Tropp, 2012, Corollary 5.2),

$$\mu_{\min} = \lambda_{\min}\left(\sum_{j=1}^{n} \mathbb{E}_{S_j \sim \nu_{d,k}}[Z_j]\right) = 1, \qquad \mu_{\max} = \lambda_{\max}\left(\sum_{j=1}^{n} \mathbb{E}_{S_j \sim \nu_{d,k}}[Z_j]\right) = 1,$$

and the uniform eigenvalue bound is $R = p_D/n$.

At relative deviation $1/2$, the upper-tail inequality gives

$$\mathbb{P}_{\nu_{d,k}^{\otimes n}}\left\{\lambda_{\max}(\widehat{G}_D) > \frac{3}{2}\right\} \leqslant p_D \exp\left(-c_0 \frac{n}{p_D}\right),$$

where

$$c_0 = \frac{3}{2}\log\left(\frac{3}{2}\right) - \frac{1}{2}.$$

The corresponding lower-tail exponent is

$$c_- = \frac{1}{2} + \frac{1}{2}\log\left(\frac{1}{2}\right) \approx 0.1534,$$

which is larger than $c_0 \approx 0.1082$. Therefore

$$\mathbb{P}_{\nu_{d,k}^{\otimes n}}\left\{\lambda_{\min}(\widehat{G}_D) < \frac{1}{2}\right\} \leqslant p_D \exp\left(-c_0 \frac{n}{p_D}\right).$$

Since $\widehat{G}_D$ is symmetric,

$$\|\widehat{G}_D - I_{p_D}\|_{\mathrm{op}} = \max\left\{\lambda_{\max}(\widehat{G}_D) - 1,\, 1 - \lambda_{\min}(\widehat{G}_D)\right\}.$$

Consequently,

$$\mathcal{A}_D^c = \left\{\lambda_{\max}(\widehat{G}_D) > \frac{3}{2}\right\} \cup \left\{\lambda_{\min}(\widehat{G}_D) < \frac{1}{2}\right\}.$$

A union bound consequently yields

$$\mathbb{P}_{\nu_{d,k}^{\otimes n}}(\mathcal{A}_D^c) \leqslant 2p_D \exp\left(-c_0 \frac{n}{p_D}\right),$$

which is (14).

D.2 Proof of Theorem 5.2

Fix $1 \leqslant D \leqslant m$, $f \in \mathcal{E}_{d,k}^s(B)$, and $P \in \mathcal{P}_{\sigma^2}$. As in the main text, write

$$g_D = P_{\leqslant D}^{\circ} f, \qquad h_D = f - g_D,$$

and let θ_D be the coefficient vector of g_D in the chosen orthonormal basis. Then $g_D(S) = x_D(S)^\top \theta_D$, and population orthogonality gives

$$\mathbb{E}_{S \sim \nu_{d,k}}[x_D(S) h_D(S)] = 0. \tag{46}$$

Set

$$z_h = \frac{1}{n}\sum_{j=1}^{n} x_D(S_j)h_D(S_j), \qquad z_\varepsilon = \frac{1}{n}\sum_{j=1}^{n} x_D(S_j)\varepsilon_j.$$

Then (17) reads, on $\mathcal{A}_D$,

$$\widehat{\theta}_D^{\mathrm{LS}} - \theta_D = \widehat{G}_D^{-1}(z_h + z_\varepsilon).$$

Conditional on the design, z_h, $\widehat{G}_D$, and $\mathbf{1}_{\mathcal{A}_D}$ are fixed, whereas $\mathbb{E}_{f,P}[z_\varepsilon \mid S_1, \dots, S_n] = 0$. Consequently, the cross term between $\widehat{G}_D^{-1}z_h$ and $\widehat{G}_D^{-1}z_\varepsilon$ has conditional expectation zero. Since $\left\|\widehat{G}_D^{-1}\right\|_{\mathrm{op}} \leqslant 2$ on $\mathcal{A}_D$,

$$\mathbb{E}_{f,P}\left[\mathbf{1}_{\mathcal{A}_D}\left\|\widehat{\theta}_D^{\mathrm{LS}} - \theta_D\right\|_2^2\right] \leqslant 4\mathbb{E}_{f,P}[\|z_h\|_2^2] + 4\mathbb{E}_{f,P}[\|z_\varepsilon\|_2^2]. \tag{47}$$

The summands defining z_h are centered by (46). Independence across observations and constant leverage give

$$\mathbb{E}_{f,P}[\|z_h\|_2^2] = \frac{1}{n}\mathbb{E}_{S\sim\nu_{d,k}}[h_D(S)^2\,\|x_D(S)\|_2^2] = \frac{p_D}{n}\,\|h_D\|_2^2. \tag{48}$$

Similarly, independence and centering of the noise eliminate all cross-observation terms, and

$$\mathbb{E}_{f,P}[\|z_\varepsilon\|_2^2] = \frac{1}{n}\mathbb{E}_P[\varepsilon_1^2]\mathbb{E}_{S\sim\nu_{d,k}}[\|x_D(S)\|_2^2] \leqslant \sigma^2\frac{p_D}{n}. \tag{49}$$

On $\mathcal{A}_D$, orthonormality of the chosen basis and orthogonality of h_D to W_D give

$$\left\|\widehat{f}_D^{\mathrm{LS}} - f\right\|_2^2 = \left\|\widehat{\theta}_D^{\mathrm{LS}} - \theta_D\right\|_2^2 + \|h_D\|_2^2.$$

On $\mathcal{A}_D^c$, the estimator is zero and the loss is $\|f\|_2^2 \leqslant B$ by (8). Combining this observation with (47)–(49) gives

$$\mathbb{E}_{f,P}\left[\left\|\widehat{f}_D^{\mathrm{LS}} - f\right\|_2^2\right] \leqslant \left(1 + 4\frac{p_D}{n}\right)\|h_D\|_2^2 + 4\sigma^2\frac{p_D}{n} + B\mathbb{P}_{\nu_{d,k}^{\otimes n}}(\mathcal{A}_D^c).$$

The harmonic tail bound gives $\|h_D\|_2^2 \leqslant b_D(s,B)$, and Proposition 5.1, together with the trivial bound $\mathbb{P}_{\nu_{d,k}^{\otimes n}}(\mathcal{A}_D^c) \leqslant 1$, gives

$$\mathbb{P}_{\nu_{d,k}^{\otimes n}}(\mathcal{A}_D^c) \leqslant \min\left\{1, 2p_D\exp\left(-c_0\frac{n}{p_D}\right)\right\}.$$

Taking the stated suprema proves (18).

D.3 Proof of Proposition 5.3

Fix $1 \leqslant D \leqslant m$, and abbreviate

$$a_D = \frac{B}{\gamma_{D,d}^{s}}.$$

As in the proof of Theorem 4.4, the Euclidean ball

$$\mathcal{B}_D = \{g \in W_D : \|g\|_2^2 \leqslant a_D\}$$

is contained in $\mathcal{E}^s_{d,k}(B)$. We restrict the noise law to the degenerate distribution δ_0, under which $\varepsilon = 0$ almost surely. This noiseless law belongs to $\mathcal{P}_{\sigma^2}$ for every $\sigma^2 > 0$.

Let Π_D be the uniform distribution, with respect to the $L^2(\nu_{d,k})$ Euclidean structure of W_D, on the sphere of squared radius a_D, and let $\Theta \sim \Pi_D$. Since this prior is supported on $\mathcal{E}^s_{d,k}(B)$, the minimax risk is at least its Bayes risk in the noiseless submodel:

$$R_n^\star(\mathcal{E}^s_{d,k}(B)) \geqslant \inf_{\widehat{f}} \int \mathbb{E}_{g,\delta_0}\left[\|\widehat{f} - g\|_2^2\right] \Pi_D(\mathrm{d}g).$$

Because the prior is supported on W_D, Pythagoras gives, for every $g \in W_D$ and every function-valued estimator $\widehat{f}$,

$$\|\widehat{f} - g\|_2^2 = \|P^\circ_{\leqslant D}\widehat{f} - g\|_2^2 + \|(I - P^\circ_{\leqslant D})\widehat{f}\|_2^2.$$

Projecting an estimator onto W_D therefore cannot increase its loss under this prior. It is consequently enough to consider estimators taking values in W_D.

As in the main text, the chosen orthonormal basis identifies W_D isometrically with $\mathbb{R}^{p_D}$. Conditional on the design, regard Θ as its coefficient vector under this identification. Write

$$\mathbb{R}^{p_D} = \mathrm{row}(\mathbf{X}_D) \oplus^\perp \ker(\mathbf{X}_D).$$

Let

$$\Theta_{\mathrm{row}} = P_{\mathrm{row}(\mathbf{X}_D)}\Theta, \qquad \Theta_{\mathrm{ker}} = P_{\ker(\mathbf{X}_D)}\Theta.$$

Then

$$\Theta = \Theta_{\mathrm{row}} + \Theta_{\mathrm{ker}}.$$

In the noiseless submodel,

$$\mathbf{Y} = \mathbf{X}_D\Theta = \mathbf{X}_D\Theta_{\mathrm{row}},$$

because $\mathbf{X}_D\Theta_{\mathrm{ker}} = 0$. The observations determine Θ_{row} uniquely. Indeed, if $u, v \in \mathrm{row}(\mathbf{X}_D)$ satisfy $\mathbf{X}_D u = \mathbf{X}_D v$, then

$$u - v \in \mathrm{row}(\mathbf{X}_D) \cap \ker(\mathbf{X}_D) = \{0\}.$$

Thus, for a fixed design, the data determine Θ_{row} but leave Θ_{ker} unobserved. Conditional on the identified value of Θ_{row}, the spherical prior distributes Θ_{ker} uniformly on the sphere in $\ker(\mathbf{X}_D)$ centered at zero and having squared radius

$$a_D - \|\Theta_{\mathrm{row}}\|_2^2.$$

Here, Θ_{row} is data-dependent: the design determines $\mathrm{row}(\mathbf{X}_D)$, and $\mathbf{Y}$ determines the unique vector $u \in \mathrm{row}(\mathbf{X}_D)$ satisfying $\mathbf{X}_D u = \mathbf{Y}$. In particular, Θ_{ker} and $-\Theta_{\mathrm{ker}}$ have the same conditional distribution, so

$$\mathbb{E}_\Theta\left[\Theta_{\mathrm{ker}} \mid S_1, \ldots, S_n, \mathbf{Y}\right] = 0.$$

Since $\Theta = \Theta_{\mathrm{row}} + \Theta_{\mathrm{ker}}$ and Θ_{row} is determined by the data,

$$\begin{aligned}\mathbb{E}_\Theta\left[\Theta \mid S_1, \ldots, S_n, \mathbf{Y}\right] &= \Theta_{\mathrm{row}} + \mathbb{E}_\Theta\left[\Theta_{\mathrm{ker}} \mid S_1, \ldots, S_n, \mathbf{Y}\right] \\ &= \Theta_{\mathrm{row}}.\end{aligned}$$

Thus, Θ_{row} is the posterior mean and hence the Bayes estimator under squared-error loss. Accordingly, for a fixed design and the noiseless observations $\mathbf{Y} = \mathbf{X}_D\Theta$,

$$\widehat{\Theta}^B(\mathbf{X}_D, \mathbf{X}_D\Theta) = P_{\text{row}(\mathbf{X}_D)}\Theta.$$

Consequently, its Bayes risk conditional on the design is

$$\mathbb{E}_\Theta\left[\left\|\widehat{\Theta}^B(\mathbf{X}_D, \mathbf{X}_D\Theta) - \Theta\right\|_2^2 \,\middle|\, S_1, \dots, S_n\right] = \mathbb{E}_\Theta\left[\|P_{\ker(\mathbf{X}_D)}\Theta\|_2^2 \,\middle|\, S_1, \dots, S_n\right].$$

It remains to compute this expectation. The uniform distribution on the sphere of squared radius a_D in $\mathbb{R}^{p_D}$ is isotropic: all coordinate directions have the same second moment, and their squared coordinates sum to a_D. Consequently,

$$\mathbb{E}_\Theta[\Theta\Theta^\top] = \frac{a_D}{p_D} I_{p_D}.$$

For the orthogonal projector $P_{\ker(\mathbf{X}_D)}$, we have

$$\|P_{\ker(\mathbf{X}_D)}\Theta\|_2^2 = \Theta^\top P_{\ker(\mathbf{X}_D)}\Theta.$$

Since Θ is independent of the design and the projector is fixed conditionally on $S_1, \dots, S_n$,

$$\begin{aligned}
\mathbb{E}_\Theta\left[\|P_{\ker(\mathbf{X}_D)}\Theta\|_2^2 \,\middle|\, S_1, \dots, S_n\right] &= \operatorname{tr}\left\{P_{\ker(\mathbf{X}_D)}\mathbb{E}_\Theta[\Theta\Theta^\top]\right\} \\
&= \frac{a_D}{p_D}\operatorname{tr}\left\{P_{\ker(\mathbf{X}_D)}\right\} \\
&= a_D\frac{p_D - \operatorname{rank}(\mathbf{X}_D)}{p_D}.
\end{aligned}$$

The last equality follows because the trace of an orthogonal projector equals the dimension of its range, here

$$\dim\ker(\mathbf{X}_D) = p_D - \operatorname{rank}(\mathbf{X}_D).$$

Averaging over the design gives

$$R_n^\star(\mathcal{E}_{d,k}^s(B)) \geqslant a_D\rho_{D,n}.$$

Since this holds for every $1 \leqslant D \leqslant m$, maximizing over D proves (20).

Moreover,

$$\operatorname{rank}(\mathbf{X}_D) \leqslant \min(n, p_D),$$

and hence

$$p_D - \operatorname{rank}(\mathbf{X}_D) \geqslant (p_D - n)_+.$$

Dividing by p_D and taking expectations proves (21).

We finally verify the full-level identities. Recall that $\mathcal{O}_n$ is the set of observed subsets and that

$$M_n = N - |\mathcal{O}_n|.$$

Set $q = |\mathcal{O}_n| = N - M_n$. Repeated observations merely duplicate rows of $\mathbf{X}_m$, so its rank depends only on the q distinct observed subsets. Recall also that W_m is the $(N-1)$-dimensional space of all centered functions on $\mathcal{S}_{d,k}$.

If $q < N$, evaluation at the q distinct observed subsets has rank q. Indeed, let $T_1, \ldots, T_q$ denote these subsets and prescribe arbitrary values $y_1, \ldots, y_q$. Choose one unobserved subset T_0, assign it the value $-\sum_{j=1}^{q} y_j$, and assign zero to every other unobserved subset. The resulting function is centered and takes the prescribed values at $T_1, \ldots, T_q$. Hence, the evaluation map from W_m onto $\mathbb{R}^q$ is surjective and has rank q.

If $q = N$, every subset in $\mathcal{S}_{d,k}$ is observed. The complete evaluation map sends $f \in W_m$ to its vector of N values. Since f is centered, this vector satisfies the single linear constraint

$$\sum_{S \in \mathcal{S}_{d,k}} f(S) = 0.$$

Conversely, every vector satisfying this constraint defines a centered function on $\mathcal{S}_{d,k}$. The image of the evaluation map is therefore an $(N-1)$-dimensional hyperplane, and its rank is $N-1$.

Combining the two cases gives, for every realization of the design,

$$\operatorname{rank}(\mathbf{X}_m) = \min(q, N-1).$$

Since $p_m = N - 1$ and $M_n = N - q$, it follows that

$$\begin{aligned} p_m - \operatorname{rank}(\mathbf{X}_m) &= (N-1) - \min(q, N-1) \\ &= (M_n - 1)_+. \end{aligned}$$

Taking expectations and dividing by $p_m = N - 1$ in (19) proves (22).

It remains to derive the explicit expression. Each fixed subset is absent from all n independent draws with probability $(1 - 1/N)^n$. Summing these N absence indicators gives

$$\mathbb{E}_{\nu_{d,k}^{\otimes n}}[M_n] = N \left(1 - \frac{1}{N}\right)^n.$$

Moreover,

$$(M_n - 1)_+ = M_n - 1 + \mathbf{1}_{\{M_n = 0\}}.$$

Therefore,

$$\mathbb{E}_{\nu_{d,k}^{\otimes n}}[(M_n - 1)_+] = N \left(1 - \frac{1}{N}\right)^n - 1 + \mathbb{P}_{\nu_{d,k}^{\otimes n}}(M_n = 0),$$

which proves (23).

To conclude, we evaluate the coverage probability appearing in this expression. For each $S \in \mathcal{S}_{d,k}$, let A_S be the event that S is absent from the sample. Since

$$\{M_n = 0\} = \bigcap_{S \in \mathcal{S}_{d,k}} A_S^c,$$

the inclusion–exclusion formula gives

$$\mathbb{P}_{\nu_{d,k}^{\otimes n}}(M_n = 0) = \sum_{\mathcal{J} \subseteq \mathcal{S}_{d,k}} (-1)^{|\mathcal{J}|} \mathbb{P}_{\nu_{d,k}^{\otimes n}}\left(\bigcap_{S \in \mathcal{J}} A_S\right).$$

If $|\mathcal{J}| = j$, avoiding every subset in $\mathcal{J}$ in all n draws has probability $(1 - j/N)^n$. Since there are $\binom{N}{j}$ such collections $\mathcal{J}$, it follows that

$$\mathbb{P}_{\nu_{d,k}^{\otimes n}}(M_n = 0) = \sum_{j=0}^{N} (-1)^j \binom{N}{j} \left(1 - \frac{j}{N}\right)^n .$$

When $n < N$, this probability is zero because n draws cannot cover N distinct subsets. This completes the proof.

D.4 Proof of Proposition 5.4

For every deterministic set $\mathcal{U} \subseteq \mathcal{S}_{d,k}$, write

$$\mathcal{H}(\mathcal{U}) = \{g \in W_m : g(S) = 0 \text{ for every } S \notin \mathcal{U}\} .$$

Recall that

$$\mathcal{U}_n = \mathcal{S}_{d,k} \setminus \mathcal{O}_n$$

is the random set of unobserved subsets. Then, consistently with the notation used in the main text,

$$\mathcal{H}_n = \mathcal{H}(\mathcal{U}_n).$$

Conditional on the design, $\mathcal{H}_n$ is the space of centered functions supported on the fixed set $\mathcal{U}_n$. If $M_n \geqslant 1$, its values on the M_n unobserved subsets satisfy one linear constraint, so its dimension is $M_n - 1$. If $M_n = 0$, then $\mathcal{H}_n = \{0\}$. Hence, in both cases,

$$\dim(\mathcal{H}_n) = (M_n - 1)_+.$$

Let $\widehat{f}^{\,\mathrm{comp},0}$ be the noiseless completion estimator introduced in the main text. It agrees with f on $\mathcal{O}_n$ and is constant on $\mathcal{U}_n$. As shown there,

$$f - \widehat{f}^{\,\mathrm{comp},0} = P_{\mathcal{H}_n} f.$$

Equivalently,

$$\widehat{f}^{\,\mathrm{comp},0} - f = -P_{\mathcal{H}_n} f.$$

This identity also holds when $M_n = 0$, since then $\mathcal{H}_n = \{0\}$ and $\widehat{f}^{\,\mathrm{comp},0} = f$.

Let $\tau : \mathcal{S}_{d,k} \to \mathcal{S}_{d,k}$ be any permutation of the N subsets, and let R_τ be its action on functions:

$$(R_\tau g)(S) = g(\tau^{-1}(S)).$$

Because the design points are sampled uniformly from $\mathcal{S}_{d,k}$, the random set of unobserved subsets $\mathcal{U}_n$ has the same distribution as $\tau(\mathcal{U}_n)$. Moreover, relabeling the unobserved subsets relabels the corresponding subspace:

$$\mathcal{H}(\tau(\mathcal{U}_n)) = R_\tau \mathcal{H}_n.$$

Since R_τ preserves the uniform inner product, the associated orthogonal projectors satisfy

$$P_{\mathcal{H}(\tau(\mathcal{U}_n))} = R_\tau P_{\mathcal{H}_n} R_\tau^{-1}.$$

Averaging over the design and using the equality in distribution of $\mathcal{U}_n$ and $\tau(\mathcal{U}_n)$ gives

$$R_\tau \mathcal{T}_n R_\tau^{-1} = \mathcal{T}_n, \qquad \mathcal{T}_n := \mathbb{E}_{\nu_{d,k}^{\otimes n}}[P_{\mathcal{H}_n}].$$

Thus, $\mathcal{T}_n$ is the deterministic operator obtained by averaging the random projector $P_{\mathcal{H}_n}$ over the design; in particular,

$$\mathcal{T}_n f = \mathbb{E}_{\nu_{d,k}^{\otimes n}}[P_{\mathcal{H}_n} f],$$

and $\mathcal{T}_n$ commutes with every permutation of the N subsets.

Fix an arbitrary ordering of the N elements of $\mathcal{S}_{d,k}$, and identify each function $f \in L^2(\mathcal{S}_{d,k}, \nu_{d,k})$ with its vector of values

$$(f(S))_{S \in \mathcal{S}_{d,k}} \in \mathbb{R}^N.$$

Under this identification, $\mathcal{T}_n$ is a linear operator on $\mathbb{R}^N$. Let $t_n(S, T)$ denote the entry of its matrix indexed by $S, T \in \mathcal{S}_{d,k}$. The identity $R_\tau \mathcal{T}_n R_\tau^{-1} = \mathcal{T}_n$ implies

$$t_n(\tau(S), \tau(T)) = t_n(S, T)$$

for every permutation τ of $\mathcal{S}_{d,k}$. For any $S, R \in \mathcal{S}_{d,k}$, there exists a permutation τ such that $\tau(S) = R$. Likewise, for any $S \neq T$ and $R \neq U$, there exists a permutation τ such that $\tau(S) = R$ and $\tau(T) = U$. Hence all diagonal entries of the matrix are equal, and all off-diagonal entries are equal. Therefore, for some scalars a_n and b_n,

$$\mathcal{T}_n = (a_n - b_n) I_N + b_n \mathbf{1}\mathbf{1}^\top.$$

Equivalently, for every function f and every $S \in \mathcal{S}_{d,k}$,

$$(\mathcal{T}_n f)(S) = (a_n - b_n) f(S) + b_n \sum_{T \in \mathcal{S}_{d,k}} f(T).$$

Every space $\mathcal{H}_n$ consists of centered functions and is therefore orthogonal to the constant function $\mathbf{1}$. Consequently,

$$P_{\mathcal{H}_n} \mathbf{1} = 0 \qquad \text{and hence} \qquad \mathcal{T}_n \mathbf{1} = 0.$$

Finally, if $f \in W_m$, then f is centered and $\mathbf{1}^\top f = 0$. Therefore,

$$\mathcal{T}_n f = (a_n - b_n) f.$$

Thus, the restriction of $\mathcal{T}_n$ to W_m is a scalar multiple of the identity.

Write $c_n = a_n - b_n$, so that

$$\mathcal{T}_n f = c_n f, \qquad f \in W_m.$$

Moreover, $\mathcal{T}_n \mathbf{1} = 0$. Since $L^2(\mathcal{S}_{d,k}, \nu_{d,k})$ is the orthogonal sum of the constant functions and the $(N-1)$-dimensional space W_m, it follows that

$$\operatorname{tr}(\mathcal{T}_n) = c_n (N - 1).$$

On the other hand, linearity of the trace and the fact that the trace of an orthogonal projector equals the dimension of its range give

$$\begin{aligned} \operatorname{tr}(\mathcal{T}_n) = \mathbb{E}_{\nu_{d,k}^{\otimes n}}[\operatorname{tr}(P_{\mathcal{H}_n})] &= \mathbb{E}_{\nu_{d,k}^{\otimes n}}[\dim(\mathcal{H}_n)] \\ &= \mathbb{E}_{\nu_{d,k}^{\otimes n}}[(M_n - 1)_+]. \end{aligned}$$

Therefore,

$$c_n = \frac{\mathbb{E}_{\nu_{d,k}^{\otimes n}}[(M_n - 1)_+]}{N-1} = \rho_{m,n}.$$

Consequently, for every centered f,

$$\begin{aligned}\mathbb{E}_{\nu_{d,k}^{\otimes n}}\left[\|P_{\mathcal{H}_n} f\|_2^2\right] &= \mathbb{E}_{\nu_{d,k}^{\otimes n}}\left[\langle f, P_{\mathcal{H}_n} f\rangle\right] \\ &= \langle f, \mathcal{T}_n f\rangle = \rho_{m,n}\|f\|_2^2. \end{aligned} \tag{50}$$

Here we used that an orthogonal projector is self-adjoint and idempotent, so $\|P_{\mathcal{H}_n} f\|_2^2 = \langle f, P_{\mathcal{H}_n} f\rangle$.

We next control the contribution of observation noise. Conditional on the design, define

$$\bar{\varepsilon}(S) = \frac{1}{C_n(S)} \sum_{j: S_j = S} \varepsilon_j, \qquad S \in \mathcal{O}_n.$$

These variables are conditionally independent and centered, and

$$\mathbb{E}_P\left[\bar{\varepsilon}(S)^2 \,\middle|\, S_1, \ldots, S_n\right] \leqslant \frac{\sigma^2}{C_n(S)}.$$

Since the completion rule is linear in the cell means and the noise is centered, for every $S \in \mathcal{S}_{d,k}$,

$$\mathbb{E}_{f,P}\left[\widehat{f}^{\,\mathrm{comp}}(S) \,\middle|\, S_1, \ldots, S_n\right] = \widehat{f}^{\,\mathrm{comp},0}(S).$$

If $M_n \geqslant 1$, the noise component $\widehat{f}^{\,\mathrm{comp}} - \widehat{f}^{\,\mathrm{comp},0}$ is equal to $\bar{\varepsilon}(S)$ on $\mathcal{O}_n$ and to

$$-\frac{1}{M_n} \sum_{T \in \mathcal{O}_n} \bar{\varepsilon}(T)$$

on $\mathcal{U}_n$. The squared $L^2(\nu_{d,k})$-norm of this noise component is

$$\frac{1}{N}\left\{\sum_{S \in \mathcal{O}_n} \bar{\varepsilon}(S)^2 + \frac{1}{M_n}\left(\sum_{S \in \mathcal{O}_n} \bar{\varepsilon}(S)\right)^2\right\}.$$

Since the cell averages are conditionally independent and centered,

$$\mathbb{E}_P\left[\left(\sum_{S \in \mathcal{O}_n} \bar{\varepsilon}(S)\right)^2 \,\middle|\, S_1, \ldots, S_n\right] = \sum_{S \in \mathcal{O}_n} \mathbb{E}_P\left[\bar{\varepsilon}(S)^2 \,\middle|\, S_1, \ldots, S_n\right].$$

Consequently,

$$\begin{aligned}&\mathbb{E}_{f,P}\left[\left\|\widehat{f}^{\,\mathrm{comp}} - \widehat{f}^{\,\mathrm{comp},0}\right\|_2^2 \,\middle|\, S_1, \ldots, S_n\right] \\ &\qquad \leqslant \frac{\sigma^2}{N}\left(1 + \frac{1}{M_n}\right) \sum_{S \in \mathcal{O}_n} \frac{1}{C_n(S)} \leqslant \frac{2\sigma^2}{N} \sum_{S \in \mathcal{O}_n} \frac{1}{C_n(S)}. \end{aligned} \tag{51}$$

If $M_n = 0$, every subset is observed and $\widehat{f}^{\,\mathrm{comp},0} = f$. Moreover,

$$\overline{Y}(S) = f(S) + \bar{\varepsilon}(S), \qquad S \in \mathcal{S}_{d,k}.$$

Since f is centered,

$$\frac{1}{N}\sum_{T\in\mathcal{S}_{d,k}} f(T)=0,$$

and hence

$$\begin{aligned}\widehat{f}^{\,\mathrm{comp}}(S)-\widehat{f}^{\,\mathrm{comp},0}(S)&=\overline{Y}(S)-\frac{1}{N}\sum_{T\in\mathcal{S}_{d,k}}\overline{Y}(T)-f(S)\\&=\overline{\varepsilon}(S)-\frac{1}{N}\sum_{T\in\mathcal{S}_{d,k}}\overline{\varepsilon}(T).\end{aligned}$$

Subtracting the uniform mean cannot increase the average squared norm. Taking conditional expectation therefore gives

$$\mathbb{E}_{f,P}\left[\left\|\widehat{f}^{\,\mathrm{comp}}-\widehat{f}^{\,\mathrm{comp},0}\right\|_2^2\,\middle|\,S_1,\ldots,S_n\right]\leqslant\frac{\sigma^2}{N}\sum_{S\in\mathcal{S}_{d,k}}\frac{1}{C_n(S)}.$$

Since $\mathcal{O}_n=\mathcal{S}_{d,k}$ when $M_n=0$, the preceding bound is smaller than the right-hand side of (51). Thus, (51) holds in both cases.

The decomposition

$$\widehat{f}^{\,\mathrm{comp}}-f=\big(\widehat{f}^{\,\mathrm{comp},0}-f\big)+\big(\widehat{f}^{\,\mathrm{comp}}-\widehat{f}^{\,\mathrm{comp},0}\big)$$

has conditionally vanishing cross term, because the first component is fixed by the design and the second has conditional mean zero. Let $C\sim\mathrm{Bin}(n,1/N)$, and define $r(0)=0$ and $r(c)=1/c$ for every integer $c\geqslant 1$. Then

$$\eta_{n,N}=\mathbb{E}[r(C)].$$

For each fixed $S\in\mathcal{S}_{d,k}$, the count $C_n(S)$ has the same distribution as C. Moreover, since $r(0)=0$,

$$\sum_{S\in\mathcal{O}_n}\frac{1}{C_n(S)}=\sum_{S\in\mathcal{S}_{d,k}}r(C_n(S)).$$

Therefore,

$$\begin{aligned}\frac{1}{N}\mathbb{E}_{\nu_{d,k}^{\otimes n}}\left[\sum_{S\in\mathcal{O}_n}\frac{1}{C_n(S)}\right]&=\frac{1}{N}\sum_{S\in\mathcal{S}_{d,k}}\mathbb{E}_{\nu_{d,k}^{\otimes n}}[r(C_n(S))]\\&=\mathbb{E}[r(C)]=\eta_{n,N}.\end{aligned}$$

Combining this identity with (50) and (51) gives

$$\mathbb{E}_{f,P}\left[\|\widehat{f}^{\,\mathrm{comp}}-f\|_2^2\right]\leqslant\rho_{m,n}\|f\|_2^2+2\sigma^2\eta_{n,N}.$$

Since $\|f\|_2^2\leqslant B$, taking the stated suprema proves (26).

It remains to bound $\eta_{n,N}$. Since

$$0\leqslant r(c)\leqslant 1,\qquad c\geqslant 0,$$

we have $\eta_{n,N}\leqslant 1$. Moreover,

$$r(c)\leqslant\frac{2}{c+1},\qquad c\geqslant 0,$$

and therefore

$$\eta_{n,N} \leqslant 2\mathbb{E}\left[\frac{1}{C+1}\right].$$

The identity

$$\frac{1}{c+1}\binom{n}{c} = \frac{1}{n+1}\binom{n+1}{c+1}$$

gives

$$\begin{aligned}\mathbb{E}\left[\frac{1}{C+1}\right] &= \sum_{c=0}^{n} \frac{1}{c+1}\binom{n}{c}\left(\frac{1}{N}\right)^{c}\left(1-\frac{1}{N}\right)^{n-c} \\ &= \frac{N}{n+1}\left\{1-\left(1-\frac{1}{N}\right)^{n+1}\right\} \\ &\leqslant \frac{N}{n+1}.\end{aligned}$$

Combining the two bounds proves (27).

D.5 Proof of Theorem 5.5

For brevity, write

$$\mathfrak{R} = \mathfrak{R}_{n,d,k}(s,B,\sigma^2), \qquad \mathfrak{D} = \mathfrak{D}_{n,d,k}(s,B), \qquad a = \frac{B}{\gamma_{m,d}^{s}}, \qquad \Gamma = \gamma_{m,d}^{s}.$$

Since $m \geqslant 1$, the monotonicity of the spectral weights and $\gamma_{1,d} = 1$ imply $\Gamma \geqslant 1$.

Theorem 4.4 and Proposition 5.3 give

$$R_n^{\star}(\mathcal{E}_{d,k}^{s}(B)) \geqslant c_1\mathfrak{R}, \qquad R_n^{\star}(\mathcal{E}_{d,k}^{s}(B)) \geqslant \mathfrak{D}$$

for a universal constant $c_1 > 0$. Therefore,

$$\begin{aligned}R_n^{\star}(\mathcal{E}_{d,k}^{s}(B)) &\geqslant \max\{c_1\mathfrak{R}, \mathfrak{D}\} \\ &\geqslant \frac{\min(c_1,1)}{2}(\mathfrak{R}+\mathfrak{D}),\end{aligned}$$

which proves the lower inequality in (28) with a universal constant.

For the upper bound, Theorem 4.4, together with $B = \Gamma a$, gives

$$R_n^{\star}(\mathcal{E}_{d,k}^{s}(B)) \leqslant 2\left(1+\frac{\Gamma a}{\sigma^2}\right)\mathfrak{R}. \tag{52}$$

On the other hand, Proposition 5.4 gives

$$R_n^{\star}(\mathcal{E}_{d,k}^{s}(B)) \leqslant B\rho_{m,n} + 2\sigma^2\eta_{n,N}.$$

The term $D = m$ in the definition of $\mathfrak{D}$ satisfies

$$\mathfrak{D} \geqslant \frac{B}{\Gamma}\rho_{m,n} = a\rho_{m,n}.$$

Consequently,

$$B\rho_{m,n} = \Gamma a \rho_{m,n} \leqslant \Gamma \mathfrak{D},$$

and hence

$$R_n^\star(\mathcal{E}_{d,k}^s(B)) \leqslant \Gamma \mathfrak{D} + 2\sigma^2 \eta_{n,N}. \tag{53}$$

Put

$$v_m = \frac{\sigma^2(N-1)}{n}.$$

This is precisely the quantity $v_D(n, \sigma^2)$ at $D = m$, because $p_m = N - 1$. The bounds in (27), together with $N \geqslant 2$, imply

$$\sigma^2 \eta_{n,N} \leqslant 4 \min\{\sigma^2, v_m\}. \tag{54}$$

Indeed, if $n \leqslant N - 1$, then $v_m \geqslant \sigma^2$, and $\eta_{n,N} \leqslant 1$ gives

$$\sigma^2 \eta_{n,N} \leqslant \sigma^2 = \min\{\sigma^2, v_m\}.$$

If $n \geqslant N$, then $v_m < \sigma^2$, and

$$\eta_{n,N} \leqslant \frac{2N}{n+1} \leqslant \frac{4(N-1)}{n},$$

so

$$\sigma^2 \eta_{n,N} \leqslant 4 v_m = 4 \min\{\sigma^2, v_m\}.$$

We now distinguish two cases. If $a \leqslant \sigma^2$, then

$$1 + \frac{\Gamma a}{\sigma^2} \leqslant 1 + \Gamma \leqslant 2\Gamma.$$

Equation (52) therefore gives

$$R_n^\star(\mathcal{E}_{d,k}^s(B)) \leqslant 4\Gamma \mathfrak{R} \leqslant 4\Gamma(\mathfrak{R} + \mathfrak{D}).$$

Suppose instead that $a > \sigma^2$. The term $D = m$ in the maximum defining $\mathfrak{R}$ gives

$$\mathfrak{R} \geqslant \min\{a, v_m\}.$$

Since $a > \sigma^2$,

$$\min\{\sigma^2, v_m\} \leqslant \min\{a, v_m\} \leqslant \mathfrak{R}.$$

Combining (53) and (54) yields

$$R_n^\star(\mathcal{E}_{d,k}^s(B)) \leqslant \Gamma \mathfrak{D} + 8\mathfrak{R} \leqslant 8\Gamma(\mathfrak{R} + \mathfrak{D}),$$

where the last inequality uses $\Gamma \geqslant 1$.

The two cases prove the upper inequality in (28), with the universal constant $C = 8$.