Influence as soft sparsity: Estimation of monotone functions on $\{0,1\}^d$

Gérard Biau
Sorbonne Université and Institut universitaire de France
gerard.biau@sorbonne-universite.fr

Abstract

We study the problem of estimating a monotone function $f : \{0,1\}^d \to [0,1]$ from noisy observations at uniformly random vertices of the Boolean hypercube. As a measure of complexity for the target f, we use the total L^1-influence $I(f) = \sum_{i=1}^d (\mathbb{E}[f(X) \mid X_i = 1] - \mathbb{E}[f(X) \mid X_i = 0])$, a classical quantity in Boolean analysis that is nonnegative for monotone functions and controls the effective dimensionality of the estimation problem: through a spectral concentration result in the spirit of Friedgut's junta theorem, the Fourier spectrum of any f with $I(f) \leqslant K$ concentrates on low-degree subsets of the influential coordinates. We establish minimax bounds over the class $\mathcal{F}_K = \{f : \{0,1\}^d \to [0,1],\ f \text{ monotone},\ I(f) \leqslant K\}$:

$$c\,\frac{K^2}{(\log n)^{3/2}} \ \leqslant\ \inf_{\hat{f}} \sup_{f \in \mathcal{F}_K} \mathbb{E}\big[\|\hat{f} - f\|_2^2\big] \ \leqslant\ C\,\frac{K}{\sqrt{\log n}},$$

where n is the sample size. The upper bound holds for all $K \geqslant 1$ and is uniform in the ambient dimension d (under the mild condition $\log d \leqslant n^{1-\varepsilon}$). It is achieved by a Fourier thresholding estimator that adapts to the unknown K. The lower bound relies on a Varshamov–Gilbert packing on the middle layer of the hypercube combined with Fano's inequality.

1 Introduction

1.1 Context and motivation

Binary features arise naturally in many applied settings. In clinical medicine, patient profiles are described by the presence or absence of risk factors—smoking, hypertension, diabetes, family history of disease—and the probability of an adverse outcome is naturally modelled as a *monotone* function of these binary indicators, encoding the prior that each additional risk factor can only increase the risk (Feelders, 2010). In credit scoring, a borrower is characterized by binary flags (employed or not, homeowner or not, previous default or not) and the default probability is monotone in these covariates (Ben-David, 1995). In pharmacogenomics, the response to a drug may depend monotonically on the presence or absence of deleterious genetic variants (Weinreich et al., 2013).

In all these examples the number d of binary features can be large (tens to thousands), while the sample size n may be comparatively modest. The monotonicity constraint alone does not overcome the curse of dimensionality: even in the simplest case of Boolean-valued functions, the number of monotone $f : \{0,1\}^d \to \{0,1\}$ grows faster than any exponential in d (it is the Dedekind number $D(d)$, satisfying $\log_2 D(d) \sim \binom{d}{\lfloor d/2 \rfloor} \approx 2^d/\sqrt{d}$), and estimation over the full class $\mathcal{M}_d$ of monotone functions $\{0,1\}^d \to [0,1]$ is at least as hard. In practice, however, domain experts often believe that only a moderate number of features meaningfully affect the outcome—but *which* features is unknown.

This motivates the search for a complexity measure that captures the effective number of relevant features, is compatible with the monotonicity constraint, and leads to a tractable estimation problem when it is small. A natural candidate is the *total L^1-influence* of f, a quantity studied in the analysis of Boolean functions (Kahn et al., 1988; O'Donnell, 2014; Kelman et al., 2021) that, as we show, is particularly well suited to the statistical setting.

1.2 The model

Let $d \geqslant 1$. We work on the Boolean hypercube $\{0,1\}^d$, equipped with the coordinatewise partial order ($x \leqslant y$ iff $x_i \leqslant y_i$ for all i) and the uniform measure $\mu = \mathrm{Unif}(\{0,1\}^d)$. We observe n independent pairs

$$Y_j = f(X_j) + \varepsilon_j, \qquad j = 1, \dots, n,$$

where $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mu$, $\varepsilon_1, \dots, \varepsilon_n$ are independent centered random variables satisfying the sub-Gaussian tail condition $\mathbb{E}[e^{t\varepsilon_j}] \leqslant e^{\sigma^2 t^2/2}$ for all $t \in \mathbb{R}$, and the ε_j are independent of the X_j.

The function $f : \{0,1\}^d \to [0,1]$ is assumed to be *monotone* (coordinatewise nondecreasing):

$$x \leqslant y \implies f(x) \leqslant f(y).$$

We denote by $\mathcal{M}_d$ the class of all monotone functions $\{0,1\}^d \to [0,1]$, and by

$$\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$$

the observed sample. The performance of an estimator $\hat{f} = \hat{f}(\mathcal{D}_n)$ is measured by the integrated squared risk

$$R(\hat{f}, f) := \mathbb{E}_{\mathcal{D}_n}\big[\|\hat{f} - f\|_2^2\big], \qquad \|g\|_2^2 := \mathbb{E}_\mu[g(X)^2] = \frac{1}{2^d} \sum_{x \in \{0,1\}^d} g(x)^2,$$

where $\mathbb{E}_{\mathcal{D}_n}$ denotes expectation over the sample and $\mathbb{E}_\mu$ denotes expectation over $X \sim \mu$. When there is no ambiguity, we simply write $\mathbb{E}$.

For $f : \{0,1\}^d \to \mathbb{R}$, $i \in [d]$, and $x \in \{0,1\}^d$, write $x^{i \to b}$ for x with the i-th coordinate replaced by b, and $\Delta_i f(x) := f(x^{i\to 1}) - f(x^{i \to 0})$ for the discrete derivative. The *L^2-influence* of coordinate i on f is $I_i^{(2)}(f) := \mathbb{E}[(\Delta_i f(X))^2]$, and the *total L^2-influence* is $I^{(2)}(f) := \sum_{i=1}^d I_i^{(2)}(f)$. This is the classical notion of influence introduced by Kahn et al. (1988); see O'Donnell (2014, Chapter 2) for a thorough treatment.

When $f : \{0,1\}^d \to [0,1]$ is monotone, the discrete derivative $\Delta_i f$ takes values in $[0,1]$, and a second notion becomes natural: the *L^1-influence* $I_i(f) := \mathbb{E}[|\Delta_i f(X)|] = \mathbb{E}[\Delta_i f(X)]$, with total $I(f) := \sum_{i=1}^d I_i(f)$. The pointwise bound $(\Delta_i f)^2 \leqslant \Delta_i f$ (valid since $\Delta_i f \in [0,1]$) gives

$$I_i(f)^2 \;\leqslant\; I_i^{(2)}(f) \;\leqslant\; I_i(f) \qquad \text{for all } i \in [d], \tag{1}$$

where the left inequality is Jensen's. Summing, we see that $I^{(2)}(f) \leqslant I(f)$.

The reason we work with the L^1-influence rather than the L^2-influence is the identity

$$I_i(f) = \mathbb{E}[f(X) \mid X_i = 1] - \mathbb{E}[f(X) \mid X_i = 0], \tag{2}$$

which holds because $X^{i\to b} \overset{d}{=} (X \mid X_i = b)$ under the uniform measure. Each $I_i(f)$ is thus a simple difference of conditional means, directly estimable from data at parametric rate. At the same time,

the bound $I^{(2)}(f) \leqslant I(f)$ ensures that controlling the L^1-influence is enough to obtain Fourier-analytic approximation guarantees. It is this combination—statistical tractability and analytical power—that makes the L^1-influence the right complexity measure for our problem.

With this in mind, we study the class

$$\mathcal{F}_K := \big\{ f : \{0,1\}^d \to [0,1] : f \in \mathcal{M}_d,\ I(f) \leqslant K \big\}$$

for a parameter $K > 0$. Since $I_i(f) \leqslant 1$ for each i, we always have $I(f) \leqslant d$, so $\mathcal{F}_d = \mathcal{M}_d$ and the constraint $I(f) \leqslant K$ is informative when $K < d$.

The following examples, borrowed from the analysis of Boolean functions (see O'Donnell, 2014), illustrate the range of behaviors captured by the total influence.

Example 1.1. *(a) Dictator.* The output copies a single coordinate: $f(x) = x_1$. Only one variable matters, and $I(f) = 1$.

(b) Additive junta. The output averages s coordinates: $f(x) = (1/s)\sum_{i=1}^{s} x_i$ for $s \leqslant d$. The function depends on s coordinates, yet its total influence is 1, independent of s.

(c) Tribes. The coordinates are divided into ℓ blocks of size $w \approx \log_2 d$, and the output is 1 if and only if at least one block is entirely composed of 1's: $f(x) = \bigvee_{j=1}^{\ell} \bigwedge_{i \in T_j} x_i$. This function is monotone, depends on all d coordinates, and satisfies $I(f) = \Theta(\log d)$.

(d) Majority. The output is 1 if more than half the coordinates are 1: $f(x) = \mathbf{1}\{\sum_i x_i > d/2\}$. Every coordinate contributes equally, and $I(f) = \Theta(\sqrt{d})$.

These examples cover the full range $K \in \{1, \Theta(\log d), \Theta(\sqrt{d})\}$, showing that the parameter K need not be a universal constant: it may depend on d, and our results are stated for all $K \leqslant d$.

1.3 Main contributions

We establish minimax bounds on the estimation risk over $\mathcal{F}_K$ that reveal a sharp dependence on the influence budget K and the sample size n, uniform in the ambient dimension d. The upper bound $O(K/\sqrt{\log n})$ is achieved by a Fourier thresholding estimator that adapts to the unknown K and remains valid for d growing nearly exponentially in n. The lower bound $\Omega(K^2/(\log n)^{3/2})$ is proved by a Varshamov–Gilbert packing on the middle layer of the Boolean hypercube, combined with Fano's inequality. A finer analysis, stratifying $\mathcal{F}_K$ by the L^2-influence $I^{(2)}(f)$, reveals that the apparent $\log n$ gap between the two bounds is an aggregation artifact: on each sub-class $\mathcal{F}_{K,B} := \{f \in \mathcal{F}_K : I^{(2)}(f) \leqslant B\}$, the gap reduces to $\sqrt{\log n}$. Precise statements are given in Section 2.

2 Main results

2.1 Minimax bounds

Our goal is to determine how the worst-case estimation risk over $\mathcal{F}_K$ depends on the sample size n and the complexity parameter K. The following theorem provides matching upper and lower bounds, up to a gap discussed below.

Theorem 2.1 (Minimax bounds)**.** *For every $\varepsilon \in (0,1)$, there exist constants $c, C > 0$ (depending only on σ and ε) such that the following holds for all integers $n \geqslant 2$ and $d \geqslant 1$.*

(i) ***Upper bound.*** *For every $1 \leqslant K \leqslant d$, provided $\log d \leqslant n^{1-\varepsilon}$, there exists an estimator $\hat{f}_n$ (depending on $\mathcal{D}_n$ but not on K) such that*

$$\sup_{f \in \mathcal{F}_K} R(\hat{f}_n, f) \ \leqslant\ C\, \frac{K}{\sqrt{\log n}}. \tag{3}$$

(ii) ***Lower bound.*** *For every* $K \leqslant c\sqrt{\log n}$ *and* $d \geqslant (1/c)\log n$,

$$\inf_{\tilde{f}} \sup_{f \in \mathcal{F}_K} R(\tilde{f}, f) \;\geqslant\; c\,\frac{K^2}{(\log n)^{3/2}},$$

where the infimum is over all estimators $\tilde{f} = \tilde{f}(\mathcal{D}_n)$.

This result calls for several comments. The condition $K \geqslant 1$ is a normalization: for very small K, the rate $K/\sqrt{\log n}$ vanishes and the estimation problem becomes degenerate. The requirement $\log d \leqslant n^{1-\varepsilon}$ is extremely mild: for any fixed $\varepsilon \in (0,1)$, it allows d as large as $\exp(n^{1-\varepsilon})$, far beyond any practical scenario. The estimator $\hat{f}_n$ does not depend on K and achieves (3) simultaneously for all $1 \leqslant K \leqslant d$, without prior knowledge of the influence budget. The lower bound is stated for $K \leqslant c\sqrt{\log n}$; beyond this regime, the upper bound $CK/\sqrt{\log n}$ exceeds a constant, and the estimation problem is inherently limited by the size of the class $\mathcal{F}_K$. The condition $d \geqslant (1/c)\log n$ is a mild dimensionality requirement: the lower bound construction uses $\Theta(\log n)$ coordinates, so d must be at least of this order for the packing to be feasible.

The rate $K/\sqrt{\log n}$ depends on d only through the mild condition $\log d \leqslant n^{1-\varepsilon}$, and is otherwise dimension-free. This stands in contrast with the classical theory of multivariate isotonic regression on $[0,1]^d$, where Han et al. (2019) showed that the minimax rate is $n^{-1/d}$ (up to logarithmic factors)—a rate that becomes uninformative for $d \gtrsim \log n$. The influence constraint thus provides a mechanism for meaningful estimation even when d grows nearly exponentially in n.

The constant estimator $\hat{f} \equiv \bar{Y}$ achieves $R(\bar{Y}, f) \leqslant \mathrm{Var}(f) + O(\sigma^2/n) \leqslant K/4 + O(\sigma^2/n)$, where $\mathrm{Var}(f) := \mathbb{E}_\mu[(f(X) - \mathbb{E}_\mu f(X))^2]$ denotes the variance of f under μ, and we used $\mathrm{Var}(f) \leqslant I(f)/4$ (see, e.g., O'Donnell 2014, Chapter 2). This risk is bounded away from zero for fixed K: the dictator $f(x) = x_1 \in \mathcal{F}_1$ has $\mathrm{Var}(f) = 1/4$, so $R(\bar{Y}, f) \geqslant 1/4$ for all n. In contrast, our estimator $\hat{f}_n$ achieves $R(\hat{f}_n, f) \leqslant CK/\sqrt{\log n}$, which tends to zero as $n \to \infty$ for every fixed K—the estimation problem is genuinely consistent, uniformly over $\mathcal{F}_K$.

2.2 Adaptive minimax bounds

For fixed K, the upper and lower bounds in Theorem 2.1 differ by a factor of $\log n$. This apparent gap, however, conceals a finer structure: the upper bound is largest on *near-Boolean* functions f for which $I^{(2)}(f) \approx I(f) \approx K$, whereas the lower bound construction achieves $I^{(2)}(f_\omega) \lesssim K^2/\sqrt{\log n}$ (see Section 6). Stratifying $\mathcal{F}_K$ by the value of $I^{(2)}(f)$ reveals that the actual gap on each stratum is only $\sqrt{\log n}$, and that the apparent $\log n$ gap on $\mathcal{F}_K$ is an aggregation artifact.

Theorem 2.2 (Adaptive minimax bounds)**.** *For every* $\varepsilon \in (0,1)$*, there exist constants* $c, C > 0$ *(depending only on* σ *and* ε*) such that the following holds for all integers* $n \geqslant 2$ *and* $d \geqslant 1$*. For* $K > 0$ *and* $B \in (0, K]$*, define the sub-class*

$$\mathcal{F}_{K,B} \;:=\; \big\{ f \in \mathcal{F}_K \,:\, I^{(2)}(f) \leqslant B \big\}.$$

(i) ***Upper bound.*** *For every* $1 \leqslant K \leqslant \sqrt{\log n}$*,* $1 \leqslant B \leqslant K$*, and* $\log d \leqslant n^{1-\varepsilon}$*, the estimator* $\hat{f}_n$ *of Theorem 2.1(i) satisfies*

$$\sup_{f \in \mathcal{F}_{K,B}} R(\hat{f}_n, f) \;\leqslant\; C\,\frac{B}{\sqrt{\log n}}.$$

(ii) ***Lower bound.*** *For every* $K \leqslant c\sqrt{\log n}$*,* $B \in (0, K^2/\sqrt{\log n}]$*, and* $d \geqslant (1/c)\log n$*,*

$$\inf_{\tilde{f}} \sup_{f \in \mathcal{F}_{K,B}} R(\tilde{f}, f) \;\geqslant\; c\,\frac{B}{\log n}, \tag{4}$$

where the infimum is over all estimators $\tilde{f} = \tilde{f}(\mathcal{D}_n)$.

Part (i) follows immediately from the refined bound Theorem 5.4(ii), since every $f \in \mathcal{F}_{K,B}$ satisfies $I^{(2)}(f) \leqslant B$. Part (ii) is proved in Section 6 by a variant of the lower bound construction of Theorem 2.1(ii), where the scaling parameter is chosen to saturate the constraint $I^{(2)}(f) \leqslant B$ rather than $I(f) \leqslant K$.

The two bounds differ by a factor $\sqrt{\log n}$, uniformly over $B \in (0, K^2/\sqrt{\log n}]$. Note that $K^2/\sqrt{\log n} \leqslant K$ under the assumption $K \leqslant c\sqrt{\log n}$ with $c \leqslant 1$, so the constraint $B \leqslant K^2/\sqrt{\log n}$ in (ii) is compatible with the definition $B \in (0, K]$ of the sub-class.

The original lower bound of Theorem 2.1(ii) is recovered as the boundary case $B = K^2/\sqrt{\log n}$:

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}_K} R(\hat{f}, f) \;\geqslant\; \inf_{\hat{f}} \sup_{f \in \mathcal{F}_{K,\, K^2/\sqrt{\log n}}} R(\hat{f}, f) \;\geqslant\; c\, \frac{K^2}{(\log n)^{3/2}}.$$

Closing the remaining $\sqrt{\log n}$ gap on $\mathcal{F}_{K,B}$ is an open problem discussed in Section 7.

2.3 Notation and organization

Section 3 discusses the three roles of the L^1-influence, its analogy with soft sparsity in linear regression, and related work. Section 4 develops the Fourier analysis on the hypercube and establishes the spectral concentration result for monotone functions under an influence budget. Section 5 constructs the estimator and proves the upper bounds (Theorem 2.1(i) and Theorem 2.2(i)). Section 6 constructs the lower bound families and proves the lower bounds (Theorem 2.1(ii) and Theorem 2.2(ii)). Section 7 discusses the gap between the bounds, extensions, and open problems.

3 The L^1-influence as soft sparsity

The estimator achieving Theorem 2.1(i) exploits three distinct properties of the L^1-influence, which together make $I(\cdot)$ the right complexity measure for our problem.

Statistical tractability. By identity (2), each $I_i(f)$ is a simple difference of conditional means, directly estimable from data at parametric rate $n^{-1/2}$, uniformly in i. This stands in contrast with the L^2-influence $I_i^{(2)}(f) = \mathbb{E}[(\Delta_i f(X))^2]$, which involves f at two points simultaneously and admits no such plug-in estimator.

Analytical bridge. Every $f : \{0,1\}^d \to \mathbb{R}$ admits an expansion $f = \sum_{S \subseteq [d]} \hat{f}(S)\, \chi_S$ in the orthonormal basis $\{\chi_S\}$ of $L^2(\{0,1\}^d, \mu)$ (the Fourier basis on the hypercube, recalled in Section 4.1). For any $f \in \mathcal{F}_K$, the Fourier coefficients concentrate on low-degree subsets of the influential coordinates of f (Proposition 4.2), in the spirit of Friedgut's junta theorem (Friedgut, 1998). The bridge $I^{(2)}(f) \leqslant I(f) \leqslant K$ (see (1)) allows the L^2-tools of Boolean analysis to apply to the L^1-constrained class $\mathcal{F}_K$, and the summation over non-influential coordinates yields a factor K rather than d in the spectral bound, giving dimension-free control.

Budget control and estimation. The constraint $\sum_i I_i(f) \leqslant K$, combined with a threshold on the estimated influences, limits the number of selected coordinates to at most $O(K/\delta)$ for a threshold δ. This keeps manageable the set of Fourier coefficients to be estimated. Our estimator is then simply constructed by estimating each coefficient in this set by its empirical counterpart, and truncating to $[0,1]$. The bias-variance trade-off in the threshold δ yields the rate $K/\sqrt{\log n}$. Under an L^2-influence constraint, the analogous bound on the number of selected coordinates would be quadratically worse, leading to a much larger estimation set.

The triple role of the L^1-influence—statistical tractability, analytical bridge, and budget control—parallels the role of the ℓ_1 norm in high-dimensional linear regression, where the constraint $\|\beta\|_1 \leqslant K$

simultaneously enables support estimation, provides approximation guarantees, and controls the complexity of the estimator. The analogy is summarized in the table below.

	Linear model	**Monotone model on** $\{0,1\}^d$
Exact sparsity	$\|\beta\|_0 \leqslant s$	f is an s-junta
Soft sparsity	$\|\beta\|_1 \leqslant K$	$I(f) \leqslant K$
Key implication	$\|\beta\|_1 \leqslant K \Rightarrow \beta \approx$ sparse	$I(f) \leqslant K \Rightarrow f \approx$ junta
Structural result	ℓ_1-ball geometry	Spectral concentration (Friedgut-type)
Complexity measure	estimable at rate $n^{-1/2}$	estimable at rate $n^{-1/2}$
Minimax rate	$K\sqrt{(\log d)/n}$	$K/\sqrt{\log n}$ (this paper)

In both settings, the soft constraint does not impose exact low-dimensional structure but guarantees approximate low-dimensionality, which suffices for estimation. The analogy extends to the proof strategy: influence estimation parallels support recovery in the Lasso, and Fourier thresholding parallels soft thresholding of regression coefficients.

Related work.

Isotonic regression. Estimation of monotone functions is a classical topic in nonparametric statistics. In dimension 1, Brunk (1955) introduced the isotonic least squares estimator, with L^2 risk of order $n^{-2/3}$ (van Eeden, 1958; Brunk, 1970); see Groeneboom and Jongbloed (2014) for a modern treatment. For multivariate isotonic regression on $[0,1]^d$, Chatterjee et al. (2015) established risk bounds for the least squares estimator over partially ordered domains, and Han et al. (2019) determined the minimax rate: $n^{-1/d}$ up to logarithmic factors, which degrades rapidly with d. On the Boolean hypercube, the influence constraint provides an alternative mechanism for controlling the complexity of the estimation problem, even when $d \gg \log n$.

Sparse nonparametric regression. Exploiting low-dimensional structure in high-dimensional nonparametric problems has been extensively studied under exact sparsity, where the target depends on $s \ll d$ unknown coordinates (Yang and Barron, 1999; Kpotufe, 2011; Comminges and Dalalyan, 2012). Our work differs in that $I(f) \leqslant K$ does not require f to depend on exactly s variables—it merely implies proximity to a junta. This is a strictly weaker assumption: Tribes, for instance, depends on all d coordinates yet satisfies $I = \Theta(\log d)$, and Majority satisfies $I = \Theta(\sqrt{d})$.

Monotone classification. Learning monotone classifiers from binary data has received attention in applied machine learning, including monotone decision trees (Ben-David, 1995; Potharst and Bioch, 1999; Feelders, 2010) and ensemble methods (González et al., 2015). These works focus on algorithms and empirical performance rather than minimax rates.

Analysis of Boolean functions. The influence of a variable was introduced by Kahn et al. (1988), who proved that every Boolean function on $\{0,1\}^d$ has a coordinate with influence $\Omega(\log d/d)$ (the KKL inequality); for Boolean functions, the L^1- and L^2-influences coincide. Friedgut (1998) showed that bounded total L^2-influence implies proximity to a junta; extensions to monotone functions on general product spaces were obtained by Friedgut (2004), and to general (not necessarily monotone) functions by Hatami (2009), who showed that small total influence implies proximity to a decision tree even without monotonicity. The connection between influences and sharp thresholds is surveyed in Garban and Steif (2014). We use spectral concentration arguments originating in Friedgut's work, adapted to our statistical setting (monotone real-valued functions under an L^1-influence budget), where the distinction between L^1- and L^2-influences is essential.

Learning juntas. PAC learnability of monotone functions and juntas has been studied by Bshouty and Tamon (1996), Mossel et al. (2003), and O'Donnell and Servedio (2007), with a focus on computational complexity. Our results are information-theoretic and establish minimax rates without computational constraints.

4 Fourier analysis and spectral concentration

This section develops the analytical tools used in the proofs. Section 4.1 recalls the Fourier expansion on the Boolean hypercube, following O'Donnell (2014) (transposed from $\{-1,1\}^d$ to $\{0,1\}^d$ via $x_i \leftrightarrow 2x_i - 1$). Section 4.2 establishes a spectral concentration result for monotone functions under an influence budget.

4.1 Fourier expansion on the hypercube

For each $S \subseteq [d]$, define the character $\chi_S : \{0,1\}^d \to \{-1,+1\}$ by

$$\chi_S(x) := \prod_{i \in S} (2x_i - 1), \qquad \chi_\emptyset \equiv 1.$$

The system $\{\chi_S\}_{S \subseteq [d]}$ forms an orthonormal basis of $L^2(\{0,1\}^d, \mu)$, and every $f : \{0,1\}^d \to \mathbb{R}$ admits the Fourier expansion

$$f(x) = \sum_{S \subseteq [d]} \hat{f}(S)\, \chi_S(x), \qquad \hat{f}(S) := \mathbb{E}_\mu[f(X)\, \chi_S(X)].$$

Parseval's identity gives $\|f\|_2^2 = \sum_{S \subseteq [d]} \hat{f}(S)^2$ and $\mathrm{Var}(f) = \sum_{S \neq \emptyset} \hat{f}(S)^2$.

Fourier representation of influences. For any $f : \{0,1\}^d \to \mathbb{R}$ and $i \in [d]$, the discrete derivative $\Delta_i f(x) = f(x^{i \to 1}) - f(x^{i \to 0})$ satisfies

$$\Delta_i f(x) = 2 \sum_{S \ni i} \hat{f}(S)\, \chi_{S \setminus \{i\}}(x). \tag{5}$$

Indeed, for $i \notin S$ the character χ_S does not depend on x_i, so $\chi_S(x^{i \to 1}) = \chi_S(x^{i \to 0})$. For $i \in S$, writing $\chi_S = \chi_{\{i\}}\, \chi_{S \setminus \{i\}}$ and using $\chi_{\{i\}}(x^{i \to 1}) - \chi_{\{i\}}(x^{i \to 0}) = 1 - (-1) = 2$ gives the result.

Squaring (5) and taking expectations (using the orthonormality of the characters), we obtain

$$I_i^{(2)}(f) = 4 \sum_{S \ni i} \hat{f}(S)^2. \tag{6}$$

For monotone f, taking expectations of (5) directly and noting that only the term $S = \{i\}$ survives (since $\mathbb{E}_\mu[\chi_{S \setminus \{i\}}] = 0$ for $S \neq \{i\}$), we get

$$I_i(f) = \mathbb{E}[\Delta_i f(X)] = 2\hat{f}(\{i\}) \geqslant 0,$$

so every first-order Fourier coefficient satisfies $\hat{f}(\{i\}) \geqslant 0$. Summing (6) over i, we also note that

$$I^{(2)}(f) = 4 \sum_{S \neq \emptyset} |S|\, \hat{f}(S)^2. \tag{7}$$

Finally, we recall the noise operator and the hypercontractive inequality, which are central to the proof of spectral concentration in Section 4.2. For $\rho \in [0,1]$, the noise operator T_ρ acts on $f : \{0,1\}^d \to \mathbb{R}$ by $T_\rho f(x) = \sum_{S \subseteq [d]} \rho^{|S|} \hat{f}(S)\, \chi_S(x)$ (it damps each Fourier coefficient by the factor $\rho^{|S|}$).

Theorem 4.1 (Bonami 1970; Beckner 1975). *For $1 \leqslant p \leqslant q$ and $0 \leqslant \rho \leqslant \sqrt{\frac{p-1}{q-1}}$, every $f : \{0,1\}^d \to \mathbb{R}$ satisfies $\|T_\rho f\|_q \leqslant \|f\|_p$, where $\|g\|_r := (\mathbb{E}_\mu[|g(X)|^r])^{1/r}$.*

4.2 Spectral concentration under an influence budget

The following result is the main analytical tool behind the upper bound. It shows that the Fourier spectrum of a monotone function with bounded L^1-influence concentrates on low-degree subsets of the influential coordinates, with explicit dimension-free bounds. The proof adapts the strategy of Friedgut (1998), who established that Boolean functions with bounded L^2-influence are well approximated by juntas, to our setting of monotone real-valued functions under an L^1-influence constraint. Two features are specific to our setting. First, the bound is stated in terms of the L^2-influence $I^{(2)}(f)$ in the high-degree component (Part 1 below), which yields the K-dependent uniform bound through the bridge $I^{(2)}(f) \leqslant I(f) \leqslant K$ but is strictly stronger in the *fluid* regime $I^{(2)}(f) \ll K$. Second, in the low-degree component (Part 2), the ambient dimension d of the classical statement is replaced by the influence budget K, which is essential for dimension-free estimation.

Proposition 4.2 (Spectral concentration). *Let $f : \{0,1\}^d \to [0,1]$ be monotone with $I(f) \leqslant K$. For any integer $d_0 \geqslant 1$ and any $\delta > 0$, define*

$$J := \big\{i \in [d] : I_i(f) \geqslant \delta\big\} \qquad \text{and} \qquad \mathcal{S} := \big\{S \subseteq J : |S| \leqslant d_0\big\}.$$

Then:

(i) $|J| \leqslant K/\delta$.

(ii) $$\sum_{S \notin \mathcal{S}} \hat{f}(S)^2 \;\leqslant\; \frac{I^{(2)}(f)}{4\,d_0} \;+\; \frac{K \cdot 3^{d_0} \cdot \delta^{1/2}}{12}.$$

Proof. Part (i) is immediate: $\sum_{i \in J} I_i(f) \leqslant I(f) \leqslant K$ and each term is $\geqslant \delta$, so $|J| \leqslant K/\delta$.

For Part (ii), we split the Fourier weight outside $\mathcal{S}$ into high-degree terms and low-degree terms not supported on J.

Part 1: High-degree terms. By (7),

$$\sum_{|S|>d_0} \hat{f}(S)^2 \;\leqslant\; \frac{1}{d_0} \sum_{|S|>d_0} |S|\,\hat{f}(S)^2 \;\leqslant\; \frac{I^{(2)}(f)}{4\,d_0}. \tag{8}$$

Part 2: Low-degree terms outside J. Define $W := \sum_{|S|\leqslant d_0,\, S \not\subseteq J} \hat{f}(S)^2$. Every such S contains some $i \notin J$, so $W \leqslant \sum_{i \notin J} W_i$ where $W_i := \sum_{S \ni i,\, |S| \leqslant d_0} \hat{f}(S)^2$.

Fix $i \notin J$. Since $\Delta_i f(x) = f(x^{i\to 1}) - f(x^{i\to 0})$ does not depend on x_i, we may define $\varphi_i : \{0,1\}^{d-1} \to [0, \frac{1}{2}]$ by $\varphi_i(x_{-i}) := \Delta_i f(x)/2$, where $x_{-i} \in \{0,1\}^{d-1}$ denotes x with the i-th coordinate removed. Identity (5) then reads $\varphi_i(x_{-i}) = \sum_{S \ni i} \hat{f}(S)\,\chi_{S\setminus\{i\}}(x_{-i})$, which is the Fourier expansion of φ_i on $\{0,1\}^{d-1}$, with coefficients $\widehat{\varphi}_i(S') = \hat{f}(S' \cup \{i\})$ for $S' \subseteq [d] \setminus \{i\}$. Therefore,

$$W_i = \sum_{|S'| \leqslant d_0 - 1} \widehat{\varphi}_i(S')^2.$$

We apply Theorem 4.1 with $p = 4/3$, $q = 2$, $\rho = 1/\sqrt{3}$ (since $(p-1)/(q-1) = 1/3 = \rho^2$). This gives $\|T_{1/\sqrt{3}}\varphi_i\|_2 \leqslant \|\varphi_i\|_{4/3}$. Since

$$\|T_{1/\sqrt{3}}\varphi_i\|_2^2 = \sum_{S' \subseteq [d]\setminus\{i\}} 3^{-|S'|}\widehat{\varphi}_i(S')^2 \geqslant 3^{-(d_0-1)} \sum_{|S'|\leqslant d_0-1} \widehat{\varphi}_i(S')^2 = 3^{-(d_0-1)} W_i,$$

we obtain

$$W_i \;\leqslant\; 3^{d_0-1}\,\|\varphi_i\|_{4/3}^2. \tag{9}$$

We now bound $\|\varphi_i\|_{4/3}^2$. Since $\Delta_i f \in [0,1]$ (by monotonicity and $f \in [0,1]$), we have $(\Delta_i f)^p \leqslant \Delta_i f$ pointwise for any $p \geqslant 1$, and therefore

$$\|\varphi_i\|_{4/3}^{4/3} = \mathbb{E}\big[|\varphi_i(X_{-i})|^{4/3}\big] = 2^{-4/3}\,\mathbb{E}\big[(\Delta_i f(X))^{4/3}\big] \leqslant 2^{-4/3}\,I_i(f).$$

Raising to the power $3/2$, we obtain

$$\|\varphi_i\|_{4/3}^2 \leqslant \big(2^{-4/3}\,I_i(f)\big)^{3/2} = \frac{I_i(f)^{3/2}}{4}.$$

Substituting into (9), we are led to

$$W_i \;\leqslant\; \frac{3^{d_0-1}}{4}\,I_i(f)^{3/2}.$$

For $i \notin J$: $I_i(f) < \delta$ by definition of J. Hence $I_i(f)^{3/2} = I_i(f)\cdot I_i(f)^{1/2} < I_i(f)\cdot\delta^{1/2}$. Summing over $i \notin J$, we conclude

$$W \;\leqslant\; \frac{3^{d_0-1}\,\delta^{1/2}}{4}\sum_{i\notin J} I_i(f) \;\leqslant\; \frac{3^{d_0-1}\,\delta^{1/2}}{4}\,K \;=\; \frac{K\cdot 3^{d_0}\cdot\delta^{1/2}}{12}. \tag{10}$$

Combining (8) and (10) gives Part (ii). □

Comparison with the classical statement.

The junta theorem of Friedgut (1998), stated for Boolean functions $f : \{0,1\}^d \to \{0,1\}$ with bounded total L^2-influence $I^{(2)}(f) \leqslant K$, asserts that f is ε-close in L^2 to a function depending on at most $2^{O(K/\varepsilon)}$ coordinates. Proposition 4.2 extends this to monotone real-valued functions under an L^1-influence budget, via the bridge $I^{(2)}(f) \leqslant I(f) \leqslant K$.

Specifically, fix $\varepsilon \in (0,1)$ and set $d_0 := \lceil K/(2\varepsilon)\rceil$ and $\delta := e^{-\gamma d_0}$ with $\gamma > 2\log 3$. Define $J := \{i : I_i(f) \geqslant \delta\}$ and $g := \sum_{S\in\mathcal{S}} \hat{f}(S)\,\chi_S$, where $\mathcal{S} := \{S \subseteq J : |S| \leqslant d_0\}$. By Proposition 4.2(ii) and $I^{(2)}(f) \leqslant K$,

$$\|f-g\|_2^2 \;\leqslant\; \frac{K}{4d_0} + \frac{K}{12}\,(3e^{-\gamma/2})^{d_0} \;\leqslant\; \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \;=\; \varepsilon,$$

provided K/ε is larger than a constant depending only on γ (so that the second term is also at most $\varepsilon/2$). Since g depends only on the coordinates in J, and Proposition 4.2(i) gives $|J| \leqslant K/\delta = Ke^{\gamma d_0}$, the function g is a junta with at most

$$|J| \leqslant Ke^{\gamma\lceil K/(2\varepsilon)\rceil} \;=\; 2^{O(K/\varepsilon)}$$

coordinates, where the last equality uses $\gamma d_0 \leqslant \gamma(K/(2\varepsilon)+1) = O(K/\varepsilon)$ and $\log K \leqslant K/(2\varepsilon)$ for $K \geqslant 1$ and $\varepsilon \leqslant 1/2$. This recovers the classical junta theorem with explicit, dimension-free bounds.

Our result provides two additional pieces of information exploited in Section 5. First, the approximation is through the Fourier projection onto $\mathcal{S}$, whose cardinality is much smaller than that of the full junta. Indeed, by Proposition 4.2(i), $|J| \leqslant Ke^{\gamma d_0}$, and the standard binomial estimate $|\mathcal{S}| \leqslant \sum_{j=0}^{d_0}\binom{|J|}{j} \leqslant (e|J|/d_0)^{d_0}$ gives

$$\log|\mathcal{S}| \;\leqslant\; d_0\log\frac{e|J|}{d_0} \;\leqslant\; d_0\big(\gamma d_0 + \log(eK/d_0)\big) \;=\; \gamma d_0^2 + d_0\log(eK/d_0),$$

which grows at most quadratically in d_0, whereas the full junta $\{0,1\}^J$ has $2^{|J|} = 2^{2^{O(K/\varepsilon)}}$ elements—a double exponential in K/ε. This gap is the key reason the Fourier estimator achieves a better rate than a junta-based estimator; the precise variance calculation exploiting this bound is carried out in Section 5. Second, the bound in Proposition 4.2(ii) involves the budget K rather than the ambient dimension d, which is essential for dimension-free estimation: in the original argument of Friedgut (1998), the analogous bound carries a prefactor d instead of K.

Finally, we note that the exponential dependence $|J| = 2^{O(K/\varepsilon)}$ cannot in general be improved. The Tribes function (Example 1.1(c)) satisfies $I(\text{Tribes}) = \Theta(\log d)$, so $K = \Theta(\log d)$ and $d = 2^{\Theta(K)}$. Yet any g with $\|\text{Tribes} - g\|_2^2 \leqslant 1/10$ must depend on at least $\Omega(d/\log d)$ coordinates (O'Donnell, 2014, Section 4.2), which is $2^{\Theta(K)}/\Theta(K)$—super-polynomial in K, confirming that the exponential dependence on K in the junta size is unavoidable.

5 Proof of the upper bound

This section constructs the estimator and proves Theorem 2.1(i) and Theorem 2.2(i).

The procedure combines influence estimation with Fourier coefficient estimation on the spectral concentration set identified in Proposition 4.2. A sample-splitting step ensures independence between variable selection and coefficient estimation.

5.1 The estimator

The estimator proceeds as follows.

Procedure 5.1 (Fourier thresholding estimator)**.**

Input: data $\mathcal{D}_n = \{(X_j, Y_j)\}_{j=1}^n$; universal constants $\gamma > 2\log 3$ and $c_0 > 0$.

Step 0 (Sample splitting). Split $\mathcal{D}_n$ into $\mathcal{D}_1 = \{(X_j, Y_j)\}_{j=1}^{n_1}$ and $\mathcal{D}_2 = \{(X_j, Y_j)\}_{j=n_1+1}^n$, where $n_1 = \lfloor n/2 \rfloor$ and $n_2 = n - n_1$.

Step 1 (Variable selection). Set

$$d_0 := \max\left(\left\lceil \sqrt{\frac{(\log n - c_0\sqrt{\log n})_+}{\gamma}} \right\rceil, 1\right) \qquad \text{and} \qquad \delta := e^{-\gamma d_0}.$$

Using $\mathcal{D}_1$, compute for each $i \in [d]$:

$$\hat{I}_i := \bar{Y}_{i,1} - \bar{Y}_{i,0}, \qquad \bar{Y}_{i,b} := \frac{\sum_{j \leqslant n_1} Y_j \mathbf{1}_{\{X_{j,i}=b\}}}{\sum_{j \leqslant n_1} \mathbf{1}_{\{X_{j,i}=b\}}}.$$

Select coordinates $\hat{J} := \{i \in [d] : \hat{I}_i \geqslant \delta/2\}$.

Step 2 (Fourier estimation). Using $\mathcal{D}_2$, compute the empirical Fourier coefficients

$$\tilde{f}(S) := \frac{1}{n_2} \sum_{j=n_1+1}^{n} Y_j \chi_S(X_j)$$

for every S in the estimated spectral set $\hat{\mathcal{S}} := \{S \subseteq \hat{J} : |S| \leqslant d_0\}$.

Step 3 (Reconstruction and truncation). Output

$$\hat{f}_n(x) := \max\Big(0,\ \min\Big(\sum_{S \in \hat{\mathcal{S}}} \tilde{f}(S)\,\chi_S(x),\ 1\Big)\Big).$$

We emphasize that d_0 depends only on n, γ, and c_0 (universal constants), not on K or d. The estimator $\hat{f}_n$ therefore does not require knowledge of K and achieves the bound of Theorem 2.1 simultaneously for all $K \leqslant d$ and all $f \in \mathcal{F}_K$.

Proposition 5.2. *There exists $C_\sigma > 0$ (depending only on σ) such that for every $t > 0$,*

$$\mathbb{P}\left(\max_{i\in[d]} |\hat{I}_i - I_i(f)| > t\right) \;\leqslant\; 2d \exp\left(-\frac{n_1\, t^2}{C_\sigma}\right) + 2d\, e^{-n_1/8}.$$

Proof. Fix $i \in [d]$ and let $N_{i,b} := \sum_{j\leqslant n_1} \mathbf{1}_{\{X_{j,i}=b\}}$. Since $N_{i,1} \sim \mathrm{Bin}(n_1, 1/2)$, Hoeffding's inequality (Vershynin, 2018, Theorem 2.2.6) gives

$$\mathbb{P}(N_{i,1} \leqslant n_1/4) \;\leqslant\; e^{-n_1/8},$$

and the same bound holds for $N_{i,0} = n_1 - N_{i,1}$. Define the event $\mathcal{E}_i := \{N_{i,0} \geqslant n_1/4,\ N_{i,1} \geqslant n_1/4\}$, which satisfies $\mathbb{P}(\mathcal{E}_i) \geqslant 1 - 2e^{-n_1/8}$. On $\mathcal{E}_i$, all sample means $\bar{Y}_{i,b}$ are well defined since $N_{i,b} \geqslant 1$.

Conditionally on $N_{i,b} = n_b$, the observations $(X_j, Y_j)_{j:X_{j,i}=b}$ are n_b i.i.d. copies of (X, Y) drawn from the law of $(X, f(X) + \varepsilon)$ conditioned on $X_i = b$. Each centered variable $Y_j - \mathbb{E}[f(X) \mid X_i = b] = (f(X_j) - \mathbb{E}[f(X) \mid X_i = b]) + \varepsilon_j$ is a sum of two independent terms: ε_j, which is sub-Gaussian with parameter σ^2 by assumption, and $f(X_j) - \mathbb{E}[f(X) \mid X_i = b]$, which takes values in an interval of length at most 1 (since $f \in [0,1]$) and is therefore sub-Gaussian with parameter $1/4$ by Hoeffding's lemma. Hence $Y_j - \mathbb{E}[f(X) \mid X_i = b]$ is sub-Gaussian with parameter $\sigma^2 + 1/4$, and $\bar{Y}_{i,b} - \mathbb{E}[f(X) \mid X_i = b]$ is sub-Gaussian with parameter $(\sigma^2 + 1/4)/n_b \leqslant 4(\sigma^2 + 1/4)/n_1$ on $\mathcal{E}_i$.

Since $\hat{I}_i - I_i = (\bar{Y}_{i,1} - \mathbb{E}[f(X) \mid X_i{=}1]) - (\bar{Y}_{i,0} - \mathbb{E}[f(X) \mid X_i{=}0])$ is a difference of two independent sub-Gaussian variables, it is sub-Gaussian with parameter at most $8(\sigma^2 + 1/4)/n_1$ on $\mathcal{E}_i$. The standard tail bound then gives

$$\mathbb{P}\big(|\hat{I}_i - I_i| > t \;\big|\; \mathcal{E}_i\big) \;\leqslant\; 2\exp\left(-\frac{n_1\, t^2}{C_\sigma}\right)$$

with $C_\sigma := 16(\sigma^2 + 1)$, using $\sigma^2 + 1/4 \leqslant \sigma^2 + 1$. Since $\mathbb{P}(|\hat{I}_i - I_i| > t) \leqslant \mathbb{P}(|\hat{I}_i - I_i| > t \mid \mathcal{E}_i) + \mathbb{P}(\mathcal{E}_i^c)$, a union bound over $i \in [d]$ gives

$$\mathbb{P}\left(\max_{i\in[d]} |\hat{I}_i - I_i(f)| > t\right) \;\leqslant\; 2d \exp\left(-\frac{n_1\, t^2}{C_\sigma}\right) + 2d\, e^{-n_1/8}.$$

□

Define the *good event*

$$\Omega_n := \left\{\max_{i\in[d]} |\hat{I}_i - I_i(f)| \leqslant \frac{\delta}{4}\right\}.$$

Setting $t = \delta/4$ in Proposition 5.2, we have

$$\mathbb{P}(\Omega_n^c) \;\leqslant\; 2d \exp\left(-\frac{n_1\, \delta^2}{16\, C_\sigma}\right) + 2d\, e^{-n_1/8}.$$

Proposition 5.3 (Variable selection)**.** *On the event Ω_n, the following properties hold.*

(i) The set $J = \{i \in [d] : I_i(f) \geqslant \delta\}$ from Proposition 4.2 satisfies $J \subseteq \hat{J}$.

(ii) Every $i \in \hat{J}$ satisfies $I_i(f) \geqslant \delta/4$.
(iii) $|\hat{J}| \leqslant 4K/\delta$.

Proof. On Ω_n, we have $|\hat{I}_i - I_i(f)| \leqslant \delta/4$ for all $i \in [d]$.

For (i), let $i \in J$. Then $I_i(f) \geqslant \delta$, so $\hat{I}_i \geqslant I_i(f) - \delta/4 \geqslant 3\delta/4 > \delta/2$, hence $i \in \hat{J}$.

For (ii), let $i \in \hat{J}$. Then $\hat{I}_i \geqslant \delta/2$, hence $I_i(f) \geqslant \hat{I}_i - \delta/4 \geqslant \delta/4$.

For (iii), summing the bound from (ii) over $i \in \hat{J}$ gives $\sum_{i\in\hat{J}} I_i(f) \geqslant |\hat{J}|\cdot\delta/4$. Since $\sum_{i=1}^{d} I_i(f) = I(f) \leqslant K$, we conclude that $|\hat{J}| \leqslant 4K/\delta$. □

5.2 Risk analysis

The following result is the technical core of the upper bound. It implies Theorem 2.1(i) via Part (i) and Theorem 2.2(i) via Part (ii).

Theorem 5.4 (Upper bound, detailed version)**.** *For every $\varepsilon \in (0,1)$, there exist constants $c, C, C' > 0$ (depending only on σ and ε) such that the following holds for all integers $n \geqslant 2$, $d \geqslant 1$ with $\log d \leqslant n^{1-\varepsilon}$, and every $K \leqslant d$.*

(i) ***Uniform bound.***

$$R(\hat{f}_n, f) \;\leqslant\; C\,\frac{K}{\sqrt{\log n}} \;+\; C'\,e^{-c\sqrt{\log n}}.$$

(ii) ***Refined bound.*** *Under the additional assumption $K \leqslant \sqrt{\log n}$,*

$$R(\hat{f}_n, f) \;\leqslant\; C\,\frac{I^{(2)}(f)}{\sqrt{\log n}} \;+\; C'\,e^{-c\sqrt{\log n}}.$$

The two bounds are consistent: Part (i) follows from Part (ii) via $I^{(2)}(f) \leqslant I(f) \leqslant K$ (see (1)). Part (ii) is strictly stronger in the *fluid* regime $I^{(2)}(f) \ll K$, and will be used to derive Theorem 2.2(i) in Section 5.3.

Proof. Throughout, $C_1, C_2, \ldots$ denote positive constants depending only on σ, γ, ε, and c_0; their values may change from line to line but never depend on n, d, or K. All conditions of the form "for n large enough" involve thresholds depending only on these parameters and can be absorbed by enlarging the constants C and C' in the statement; in particular, we assume henceforth that n is large enough that $\log n \geqslant c_0\sqrt{\log n}$, so that $d_0 = \lceil \sqrt{(\log n - c_0\sqrt{\log n})/\gamma} \rceil \geqslant 1$. The constant c_0 depends only on γ and is determined at the end of Step 1.

For $K > \sqrt{\log n}$, Part (i) holds since $CK/\sqrt{\log n} \geqslant C \geqslant 1 \geqslant R(\hat{f}_n, f)$ for any $C \geqslant 1$. It therefore suffices to prove Part (ii) under the assumption

$$K \;\leqslant\; \sqrt{\log n}, \tag{11}$$

from which Part (i) follows via $I^{(2)}(f) \leqslant K$.

Since $\hat{f}_n$ is truncated to $[0,1]$ and $f \in [0,1]$, the pointwise error satisfies $|\hat{f}_n(x) - f(x)| \leqslant 1$ for all x, hence

$$\|\hat{f}_n - f\|_2^2 \;\leqslant\; 1 \qquad \text{always.} \tag{12}$$

We decompose the risk as

$$R(\hat{f}_n, f) \;=\; \mathbb{E}\big[\|\hat{f}_n - f\|_2^2\,\mathbf{1}_{\Omega_n}\big] + \mathbb{E}\big[\|\hat{f}_n - f\|_2^2\,\mathbf{1}_{\Omega_n^c}\big] \;\leqslant\; \mathbb{E}\big[\|\hat{f}_n - f\|_2^2\,\mathbf{1}_{\Omega_n}\big] + \mathbb{P}(\Omega_n^c), \tag{13}$$

where the inequality uses (12).

Step 1: Risk on the good event (bias + variance). On Ω_n, the truncation can only reduce the L^2 error (projecting onto $[0,1]$ is a contraction when the target lies in $[0,1]$), so it suffices to bound the risk of the untruncated estimator $\tilde{f}_n = \sum_{S\in\hat{\mathcal{S}}} \tilde{f}(S)\chi_S$. By Parseval's identity, since $\tilde{f}_n$ sets to zero all Fourier coefficients outside $\hat{\mathcal{S}}$,

$$\|\tilde{f}_n - f\|_2^2 \;=\; \underbrace{\sum_{S\in\hat{\mathcal{S}}} \big(\tilde{f}(S) - \hat{f}(S)\big)^2}_{\text{variance}} \;+\; \underbrace{\sum_{S\notin\hat{\mathcal{S}}} \hat{f}(S)^2}_{\text{bias}}.$$

Bias. On Ω_n, Proposition 5.3(i) gives $J \subseteq \hat{J}$, hence $\mathcal{S} = \{S \subseteq J : |S| \leqslant d_0\} \subseteq \hat{\mathcal{S}}$. By Proposition 4.2(ii) with $\delta = e^{-\gamma d_0}$,

$$\sum_{S\notin\hat{\mathcal{S}}} \hat{f}(S)^2 \;\leqslant\; \sum_{S\notin\mathcal{S}} \hat{f}(S)^2 \;\leqslant\; \frac{I^{(2)}(f)}{4\,d_0} + \frac{K}{12}\,\eta^{d_0},$$

where $\eta := 3e^{-\gamma/2} < 1$ (since $\gamma > 2\log 3$). Since $d_0 \geqslant C_1\sqrt{\log n}$, the first term satisfies

$$\frac{I^{(2)}(f)}{4d_0} \;\leqslant\; \frac{C_2\, I^{(2)}(f)}{\sqrt{\log n}}.$$

The second term decays exponentially: since $K \leqslant \sqrt{\log n}$ by (11), $(K/12)\eta^{d_0} \leqslant C_3\, e^{-C_4\sqrt{\log n}}$. Therefore,

$$\text{Bias} \;\leqslant\; C_2\,\frac{I^{(2)}(f)}{\sqrt{\log n}} + C_3\, e^{-C_4\sqrt{\log n}} \;\leqslant\; C_2\,\frac{K}{\sqrt{\log n}} + C_3\, e^{-C_4\sqrt{\log n}}, \tag{14}$$

using $I^{(2)}(f) \leqslant K$ in the last step.

Variance. By sample splitting, $\hat{\mathcal{S}}$ is determined by $\mathcal{D}_1$ and the coefficients $\tilde{f}(S)$ are computed from the independent sample $\mathcal{D}_2$. Conditionally on $\mathcal{D}_1$, the errors $\tilde{f}(S) - \hat{f}(S)$ are centered, so by linearity of expectation,

$$\mathbb{E}\left[\sum_{S\in\hat{\mathcal{S}}} (\tilde{f}(S) - \hat{f}(S))^2 \,\middle|\, \mathcal{D}_1\right] = \sum_{S\in\hat{\mathcal{S}}} \mathrm{Var}(\tilde{f}(S)) \leqslant \frac{2(\sigma^2+1)}{n}\,|\hat{\mathcal{S}}|,$$

since $\mathrm{Var}(\tilde{f}(S)) = \mathrm{Var}(Y\chi_S(X))/n_2 \leqslant \mathbb{E}[Y^2]/n_2 \leqslant 2(\sigma^2+1)/n$ (using $n_2 \geqslant n/2$ and $\mathbb{E}[Y^2] \leqslant 1+\sigma^2$).

It remains to bound $|\hat{\mathcal{S}}|$ on Ω_n. By Proposition 5.3(iii), $|\hat{J}| \leqslant 4Ke^{\gamma d_0}$. We distinguish two cases. If $d_0 \leqslant |\hat{J}|$, the standard binomial inequality $\sum_{j=0}^{k}\binom{N}{j} \leqslant (eN/k)^k$ (valid for $k \leqslant N$) with $k = d_0$ and $N = |\hat{J}|$ gives

$$\log|\hat{\mathcal{S}}| \;\leqslant\; d_0\big(\gamma d_0 + \log(4eK/d_0)\big) \;=\; \gamma d_0^2 + d_0\log(4eK/d_0).$$

If $d_0 > |\hat{J}|$, then $\hat{\mathcal{S}}$ is the power set of $\hat{J}$ and $\log|\hat{\mathcal{S}}| = |\hat{J}|\log 2 \leqslant d_0\log 2 \leqslant \gamma d_0^2$ (since $d_0 \geqslant 1$ and $\gamma > \log 2$). In both cases,

$$\log|\hat{\mathcal{S}}| \;\leqslant\; \gamma d_0^2 + d_0\big(\log(4eK/d_0)\big)^+. \tag{15}$$

Write $u := \sqrt{(\log n - c_0\sqrt{\log n})/\gamma}$, so that $d_0 = \lceil u\rceil \leqslant u+1$. Then

$$\gamma d_0^2 \;\leqslant\; \gamma(u+1)^2 \;=\; \log n - c_0\sqrt{\log n} + 2\gamma u + \gamma \;\leqslant\; \log n - (c_0 - 2\sqrt{\gamma})\sqrt{\log n} + \gamma, \tag{16}$$

where we used $u \leqslant \sqrt{(\log n)/\gamma}$. By (11), $K/d_0 \leqslant \sqrt{\log n}/u \leqslant \sqrt{2\gamma}$ for n large enough that $u \geqslant \sqrt{(\log n)/(2\gamma)}$, so $4eK/d_0 \leqslant 4e\sqrt{2\gamma} =: C_5$, and

$$d_0\big(\log(4eK/d_0)\big)^+ \;\leqslant\; (\log C_5)\, d_0 \;\leqslant\; C_6\sqrt{\log n}. \tag{17}$$

Substituting (16) and (17) into (15),

$$\log \frac{|\hat{\mathcal{S}}|}{n} \;\leqslant\; -(c_0 - 2\sqrt{\gamma} - C_6)\sqrt{\log n} + \gamma. \tag{18}$$

We now choose $c_0 := 2\sqrt{\gamma} + C_6 + 2$, a constant depending only on γ. Then the right-hand side of (18) is at most $-2\sqrt{\log n} + \gamma \leqslant -\sqrt{\log n}$ for n large enough that $\sqrt{\log n} \geqslant \gamma$, and the variance satisfies

$$\text{Variance} \;\leqslant\; \frac{2(\sigma^2+1)}{n}\,|\hat{\mathcal{S}}| \;\leqslant\; C_7\, e^{-\sqrt{\log n}}. \tag{19}$$

Combining (14) and (19),

$$\mathbb{E}\big[\|\hat{f}_n - f\|_2^2\, \mathbf{1}_{\Omega_n}\big] \;\leqslant\; C_2\,\frac{I^{(2)}(f)}{\sqrt{\log n}} + (C_3 + C_7)\, e^{-c'\sqrt{\log n}}, \tag{20}$$

where $c' := \min(C_4, 1)$.

Step 2: Probability of the bad event. By Proposition 5.2 with $t = \delta/4$,

$$\mathbb{P}(\Omega_n^c) \;\leqslant\; \underbrace{2d\exp\!\left(-\frac{n_1\delta^2}{16C_\sigma}\right)}_{=:T_1} + \underbrace{2d\, e^{-n_1/8}}_{=:T_2},$$

where $C_\sigma = 16(\sigma^2+1)$, and we use $n_1 = \lfloor n/2 \rfloor \geqslant n/4$ throughout.

Bound on T_2.

$$\log T_2 \;=\; \log 2 + \log d - \frac{n_1}{8} \;\leqslant\; 1 + n^{1-\varepsilon} - \frac{n}{32}.$$

Since $n^{1-\varepsilon} = o(n)$, the right-hand side is at most $-n/64$ for n large enough, giving $T_2 \leqslant e^{-n/64} \leqslant e^{-\sqrt{\log n}}$.

Bound on T_1. Since $d_0 \leqslant \sqrt{(\log n)/\gamma} + 1$,

$$\delta^2 = e^{-2\gamma d_0} \;\geqslant\; e^{-2\gamma}\cdot e^{-2\sqrt{\gamma\log n}}.$$

Setting $\alpha(n) := C_8 \exp(\log n - 2\sqrt{\gamma\log n})$ with $C_8 := e^{-2\gamma}/(64C_\sigma)$,

$$\frac{n_1\delta^2}{16C_\sigma} \geqslant \alpha(n), \qquad \log T_1 \leqslant 1 + n^{1-\varepsilon} - \alpha(n).$$

Since $n^{1-\varepsilon}/\alpha(n) = C_8^{-1}\exp(-\varepsilon\log n + 2\sqrt{\gamma\log n}) \to 0$, we have $n^{1-\varepsilon} \leqslant \alpha(n)/2$ for n large enough, so $\log T_1 \leqslant 1 - \alpha(n)/2$. Finally, $\log n - 2\sqrt{\gamma\log n} \geqslant (\log n)/2$ for n large enough, so $\alpha(n) \geqslant C_8\sqrt{n} \geqslant 2(1+\sqrt{\log n})$, giving $T_1 \leqslant e^{-\sqrt{\log n}}$.

Combining,

$$\mathbb{P}(\Omega_n^c) \;\leqslant\; C_9\, e^{-\sqrt{\log n}}. \tag{21}$$

Step 3: Conclusion. Substituting (20) and (21) into (13),

$$R(\hat{f}_n, f) \;\leqslant\; C_2\,\frac{I^{(2)}(f)}{\sqrt{\log n}} + (C_3 + C_7 + C_9)\, e^{-c'\sqrt{\log n}},$$

which is Part (ii) with $C = C_2$ and $C' = C_3 + C_7 + C_9$. Part (i) follows by $I^{(2)}(f) \leqslant K$ for $K \leqslant \sqrt{\log n}$, and by $CK/\sqrt{\log n} \geqslant 1 \geqslant R(\hat{f}_n, f)$ for $K > \sqrt{\log n}$. □

5.3 Proofs of Theorem 2.1(i) and Theorem 2.2(i)

Both results follow from Theorem 5.4 by the same argument: the exponential term $C'e^{-c\sqrt{\log n}}$ is absorbed into the leading term after enlarging C, using $K/\sqrt{\log n} \gg e^{-c\sqrt{\log n}}$ for $K \geqslant 1$ (Theorem 2.1(i), via Part (i)) and $B/\sqrt{\log n} \gg e^{-c\sqrt{\log n}}$ for $B \geqslant 1$ (Theorem 2.2(i), via Part (ii), since $I^{(2)}(f) \leqslant B$ for every $f \in \mathcal{F}_{K,B}$). □

6 Proof of the lower bound

This section proves Theorem 2.1(ii) and Theorem 2.2(ii). The argument constructs a large family of well-separated monotone functions in $\mathcal{F}_K$ and applies Fano's inequality. The construction exploits the middle layer of the Boolean hypercube and the Varshamov–Gilbert bound.

We use the following standard form of Fano's inequality; see Tsybakov (2009, Theorem 2.7).

Theorem 6.1 (Fano). *Let $M \geq 2$, let $\{P_0, P_1, \dots, P_M\}$ be probability measures on a measurable space, and let $\theta_0, \dots, \theta_M$ be elements of a pseudometric space (T, d) satisfying $d(\theta_i, \theta_j) \geqslant 2\delta > 0$ for all $i \neq j$. Then*

$$\inf_{\hat{\theta}} \max_{0 \leqslant j \leqslant M} \mathbb{P}_j\big(d(\hat{\theta}, \theta_j) \geqslant \delta\big) \;\geqslant\; 1 - \frac{\bar{\mathrm{KL}} + \log 2}{\log M}, \tag{22}$$

where $\bar{\mathrm{KL}} := \frac{1}{M}\sum_{j=1}^{M} \mathrm{KL}(P_j \| P_0)$.

6.1 The construction

Throughout, all inequalities hold for n large enough (depending only on σ); the constant c in the statement of Theorem 2.1(ii) is adjusted accordingly.

Fix $K \geqslant 1$ and $d \geqslant s$, where

$$s := \lfloor 2 \log_2 n \rfloor.$$

For $z \in \{0,1\}^s$, write $|z| := \sum_{i=1}^{s} z_i$ for its Hamming weight. Let $m := \lfloor s/2 \rfloor$ and $L_m := \{z \in \{0,1\}^s : |z| = m\}$ the middle layer, with $N := |L_m| = \binom{s}{m}$.

Step 1: Packing on the middle layer. By the Varshamov–Gilbert bound (Massart, 2007, Lemma 4.7), there exists a subset $\Omega \subseteq \{0,1\}^N$ with

$$\log|\Omega| \;\geqslant\; \frac{N}{8} \;=\; \frac{1}{8}\binom{s}{m}, \qquad d_H(\omega, \omega') \;\geqslant\; \frac{N}{4} \quad \text{for all } \omega \neq \omega' \in \Omega, \tag{23}$$

where d_H denotes the Hamming distance on $\{0,1\}^N$. We index the coordinates of $\omega \in \{0,1\}^N$ by the elements of L_m, writing ω_a for $a \in L_m$.

Step 2: From binary vectors to monotone functions. Fix a subset $S_0 \subseteq [d]$ with $|S_0| = s$, and set

$$\beta := \frac{K}{A\sqrt{s}},$$

where $A > 0$ is an absolute constant determined by Stirling's approximation, chosen at the end of the influence calculation below to ensure $I(f_\omega) \leqslant K$.

For each $\omega \in \Omega$, define $f_\omega : \{0,1\}^d \to [0,1]$ by

$$f_\omega(x) := \begin{cases} 0 & \text{if } |x_{S_0}| < m, \\ \beta\,\omega_{x_{S_0}} & \text{if } |x_{S_0}| = m, \\ \beta & \text{if } |x_{S_0}| > m, \end{cases}$$

where $x_{S_0} \in \{0,1\}^s$ denotes the projection of x onto the coordinates in S_0; see Figure 1 for an illustration.

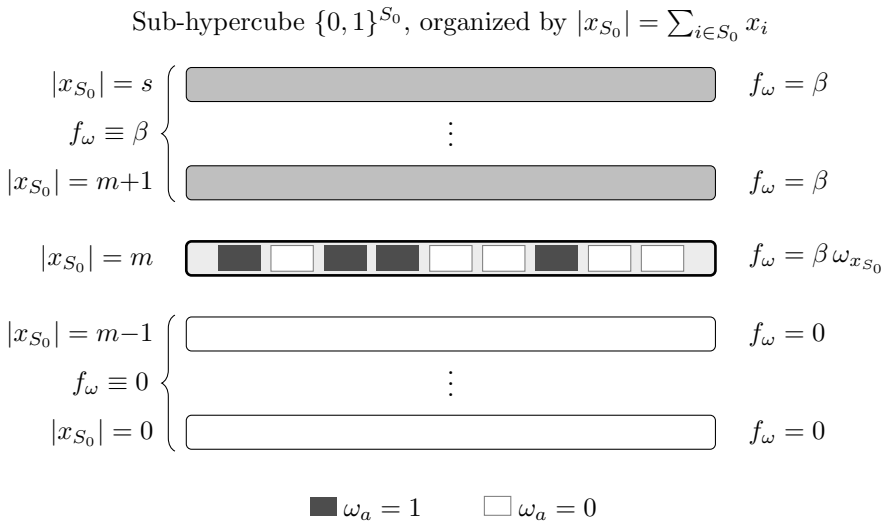

Figure 1: Structure of the functions f_ω in the lower-bound construction. Only the s coordinates in S_0 matter; the remaining $d-s$ coordinates do not affect f_ω. The sub-hypercube $\{0,1\}^{S_0}$ is organized by Hamming weight $|x_{S_0}|$. Below the middle layer ($|x_{S_0}| < m$), $f_\omega \equiv 0$; above ($|x_{S_0}| > m$), $f_\omega \equiv \beta$. On the middle layer L_m, $f_\omega(a) = \beta$ if $\omega_a = 1$ and $f_\omega(a) = 0$ if $\omega_a = 0$. Different choices of $\omega \in \Omega$ (the Varshamov–Gilbert packing) yield functions that differ only on L_m and are well separated in L^2.

We now check that each f_ω is monotone, takes values in $[0,1]$, and belongs to $\mathcal{F}_K$.

(i) *Monotonicity.* Let $x \leqslant y$ in $\{0,1\}^d$. Then $|x_{S_0}| \leqslant |y_{S_0}|$. If $|x_{S_0}| = |y_{S_0}|$, then $x_i \leqslant y_i$ for all i together with $\sum_{i \in S_0} x_i = \sum_{i \in S_0} y_i$ forces $x_i = y_i$ for every $i \in S_0$, so $f_\omega(x) = f_\omega(y)$. If $|x_{S_0}| < |y_{S_0}|$, one of the following holds:

- $|x_{S_0}| < m$ and $|y_{S_0}| < m$: $f_\omega(x) = 0 = f_\omega(y)$.
- $|x_{S_0}| < m$ and $|y_{S_0}| = m$: $f_\omega(x) = 0 \leqslant \beta\omega_{y_{S_0}} = f_\omega(y)$.
- $|x_{S_0}| < m$ and $|y_{S_0}| > m$: $f_\omega(x) = 0 \leqslant \beta = f_\omega(y)$.
- $|x_{S_0}| = m$ and $|y_{S_0}| > m$: $f_\omega(x) = \beta\omega_{x_{S_0}} \leqslant \beta = f_\omega(y)$.
- $|x_{S_0}| > m$ and $|y_{S_0}| > m$: $f_\omega(x) = \beta = f_\omega(y)$.

In every case, $f_\omega(x) \leqslant f_\omega(y)$. ✓

(ii) *Range.* $f_\omega \in [0,\beta] \subseteq [0,1]$, provided $\beta \leqslant 1$. Since $\beta = K/(A\sqrt{s})$, the condition $\beta \leqslant 1$ is equivalent to $K \leqslant A\sqrt{s}$. Under the assumption $K \leqslant c\sqrt{\log n}$ of Theorem 2.1(ii), this holds by choosing c small enough (depending on A). ✓

(iii) *Influence.* For $i \notin S_0$, $\Delta_i f_\omega = 0$. For $i \in S_0$, $\Delta_i f_\omega(x) = f_\omega(x^{i\to 1}) - f_\omega(x^{i\to 0})$ is nonzero only when $|x^{i\to 0}_{S_0}|$ and $|x^{i\to 1}_{S_0}| = |x^{i\to 0}_{S_0}| + 1$ straddle the middle layer. Writing $k := |x_{S_0\setminus\{i\}}|$ (which determines both sides since $|x^{i\to 0}_{S_0}| = k$ and $|x^{i\to 1}_{S_0}| = k+1$), the two cases where $\Delta_i f_\omega \neq 0$ are:

- $k = m-1$: $f_\omega(x^{i\to 0}) = 0$, $f_\omega(x^{i\to 1}) = \beta\,\omega_{x^{i\to 1}_{S_0}}$, giving $\Delta = \beta\,\omega_{x^{i\to 1}_{S_0}}$.
- $k = m$: $f_\omega(x^{i\to 0}) = \beta\,\omega_{x^{i\to 0}_{S_0}}$, $f_\omega(x^{i\to 1}) = \beta$, giving $\Delta = \beta(1 - \omega_{x^{i\to 0}_{S_0}})$.

Since $k = |X_{S_0\setminus\{i\}}|$ has the $\mathrm{Bin}(s{-}1, 1/2)$ distribution and is independent of X_i,

$$I_i(f_\omega) = \beta\Big(\mathbb{P}(k{=}m{-}1)\,\mathbb{E}[\omega_{X^{i\to 1}_{S_0}} \mid k{=}m{-}1] + \mathbb{P}(k{=}m)\,\mathbb{E}[1{-}\omega_{X^{i\to 0}_{S_0}} \mid k{=}m]\Big),$$

where $\mathbb{P}(k = j) = \binom{s-1}{j} 2^{-(s-1)}$. Each conditional expectation is at most 1. Choosing $A > 0$

large enough that $\binom{s}{m}/2^s \leqslant A/(2\sqrt{s})$ (which is possible by Stirling's approximation), Pascal's identity $\binom{s-1}{m-1} + \binom{s-1}{m} = \binom{s}{m}$ gives

$$I_i(f_\omega) \;\leqslant\; \frac{A\beta}{\sqrt{s}}.$$

Summing over $i \in S_0$ and using $\beta = K/(A\sqrt{s})$,

$$I(f_\omega) \;\leqslant\; A\beta\sqrt{s} \;=\; K. \checkmark \tag{24}$$

Step 3: Separation. For $\omega \neq \omega' \in \Omega$, the functions f_ω and $f_{\omega'}$ differ only on the middle layer. Since $f_\omega(x)$ depends only on x_{S_0}, each term $(f_\omega(z) - f_{\omega'}(z))^2$ is repeated 2^{d-s} times in the sum over $\{0,1\}^d$, and since f_ω and $f_{\omega'}$ agree outside L_m with $(\omega_a - \omega'_a)^2 = |\omega_a - \omega'_a|$ for $\omega_a \in \{0,1\}$,

$$\|f_\omega - f_{\omega'}\|_2^2 \;=\; \frac{1}{2^s} \sum_{z \in \{0,1\}^s} (f_\omega(z) - f_{\omega'}(z))^2 \;=\; \frac{\beta^2\, d_H(\omega,\omega')}{2^s}.$$

By (23), $d_H(\omega,\omega') \geqslant \binom{s}{m}/4$. Stirling's approximation gives $\binom{s}{m}/2^s \sim \sqrt{2/(\pi s)}$ as $s \to \infty$, so in particular $\binom{s}{m}/2^s \geqslant 4c'/\sqrt{s}$ for an absolute constant $c' > 0$ and all $s \geqslant 1$. Therefore,

$$\|f_\omega - f_{\omega'}\|_2^2 \;\geqslant\; \frac{c'\,\beta^2}{\sqrt{s}} \;=\; \frac{c'\,K^2}{A^2\, s^{3/2}}. \tag{25}$$

6.2 Proof of Theorem 2.1(ii)

For the lower bound, it suffices to consider Gaussian noise $\varepsilon_j \sim N(0,\sigma^2)$, which satisfies the sub-Gaussian assumption of Section 1.1 and is therefore a legitimate choice within our model. The KL divergence between two Gaussian location models gives

$$\mathrm{KL}(P_\omega \| P_{\omega'}) \;=\; \frac{n}{2\sigma^2}\,\|f_\omega - f_{\omega'}\|_2^2 \;\leqslant\; \frac{n\beta^2}{2\sigma^2} \;=\; \frac{nK^2}{2\sigma^2 A^2 s}, \tag{26}$$

where we used $\|f_\omega - f_{\omega'}\|_\infty \leqslant \beta$ (since both functions take values in $[0,\beta]$).

We apply Theorem 6.1 with $\delta := \sqrt{c'}K/(2As^{3/4})$, so that $\|f_\omega - f_{\omega'}\|_2 \geqslant 2\delta$ by (25). The Fano bound (22) gives a positive probability of error provided $\bar{\mathrm{KL}} \leqslant \frac{1}{2}\log|\Omega|$ (indeed, the right-hand side of (22) is then at least $1 - (\frac{1}{2} + \log 2/\log|\Omega|)$, which is bounded below by a positive constant p_0 since $|\Omega| \to \infty$ with n).

We now verify the condition $\bar{\mathrm{KL}} \leqslant \frac{1}{2}\log|\Omega|$. By (26),

$$\bar{\mathrm{KL}} \;\leqslant\; \frac{nK^2}{2\sigma^2 A^2 s}.$$

By (23) and Stirling's approximation,

$$\frac{1}{2}\log|\Omega| \;\geqslant\; \frac{1}{16}\binom{s}{m} \;\geqslant\; \frac{c''\,2^s}{\sqrt{s}},$$

where $c'' > 0$ is an absolute constant. The condition is therefore satisfied whenever

$$\frac{nK^2}{2\sigma^2 A^2 s} \;\leqslant\; \frac{c''\,2^s}{\sqrt{s}}. \tag{27}$$

Recall that $s = \lfloor 2\log_2 n \rfloor$, so $2^s \geqslant n^2/2$. Under the assumption $K \leqslant c\sqrt{\log n}$, the left-hand side of (27) is at most $nc^2 \log n/(2\sigma^2 A^2 s) = O(n)$. The right-hand side is at least $c'' n^2/(2\sqrt{s}) = \Omega(n^2/\sqrt{\log n})$. Since $n^2/\sqrt{\log n} \gg n$, condition (27) is satisfied for n large enough (depending only on σ).

Conclusion. With $s = \lfloor 2\log_2 n \rfloor$, the Fano condition is satisfied for n large enough, and Theorem 6.1 gives, for some $p_0 > 0$ (depending only on σ),

$$\inf_{\hat{f}} \max_{\omega} \mathbb{P}_\omega\big(\|\hat{f} - f_\omega\|_2 \geqslant \delta\big) \;\geqslant\; p_0.$$

The construction requires $d \geqslant s$ (so that the subset $S_0 \subseteq [d]$ with $|S_0| = s$ exists); the choice of S_0 is arbitrary since the uniform measure treats all coordinates symmetrically. Since every f_ω belongs to $\mathcal{F}_K$ (by (24)), for any estimator $\hat{f}$,

$$\sup_{f \in \mathcal{F}_K} R(\hat{f}, f) \;\geqslant\; \max_{\omega} \mathbb{E}_\omega\big[\|\hat{f} - f_\omega\|_2^2\big] \;\geqslant\; \max_{\omega} \delta^2\, \mathbb{P}_\omega\big(\|\hat{f} - f_\omega\|_2 \geqslant \delta\big) \;\geqslant\; p_0\, \delta^2,$$

where the second inequality holds since $Z^2 \geqslant a^2 \mathbf{1}_{Z \geqslant a}$ for any $Z \geqslant 0$ and $a > 0$, and the third uses the Fano bound. Taking the infimum over $\hat{f}$ and using $s \leqslant (2/\log 2)\log n$,

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}_K} R(\hat{f}, f) \;\geqslant\; p_0\, \delta^2 \;=\; \frac{p_0\, c'}{4A^2} \cdot \frac{K^2}{s^{3/2}} \;\geqslant\; \frac{c\, K^2}{(\log n)^{3/2}},$$

where $c > 0$ depends only on σ. The construction requires $d \geqslant s = \lfloor 2\log_2 n \rfloor$, which is guaranteed by $d \geqslant (1/c)\log n$ upon choosing $c \leqslant \log 2/2$. Taking c smaller if necessary, the condition $K \leqslant c\sqrt{\log n}$ of Theorem 2.1(ii) is also satisfied, completing the proof. □

6.3 Proof of Theorem 2.2(ii)

We adapt the construction of Section 6.1 by choosing β to saturate $I^{(2)} \leqslant B$ rather than $I \leqslant K$. Let $K \leqslant c\sqrt{\log n}$, $B \in (0, K^2/\sqrt{\log n}]$, and define

$$\beta_B := \sqrt{\frac{B}{A_1\sqrt{s}}},$$

where $A_1 > 0$ is an absolute constant chosen large enough (depending only on A) and $s = \lfloor 2\log_2 n \rfloor$ as before. The functions f_ω are defined as in Section 6.1 with β replaced by β_B; the constants c and A_1 are chosen small and large enough, respectively, for all conditions below to hold.

Range and monotonicity. Since $s \geqslant \log n/\log 2$,

$$\beta_B^2 = \frac{B}{A_1\sqrt{s}} \leqslant \frac{K^2\sqrt{\log 2}}{A_1 \log n} \leqslant \frac{c^2\sqrt{\log 2}}{A_1} \leqslant 1$$

for c small enough. Monotonicity follows by the same argument as in Section 6.1.

L^1-influence bound. By the same calculation as (24), $I_i(f_\omega) \leqslant A\beta_B/\sqrt{s}$ for each $i \in S_0$, so

$$I(f_\omega) \;\leqslant\; A\beta_B\sqrt{s} \;=\; \sqrt{\frac{A^2 B\sqrt{s}}{A_1}} \;\leqslant\; K \cdot \frac{A(2/\log 2)^{1/4}}{\sqrt{A_1}} \;\leqslant\; K,$$

where we used $\sqrt{s} \leqslant \sqrt{2/\log 2}\sqrt{\log n}$ and $B \leqslant K^2/\sqrt{\log n}$, and the last inequality holds by choosing $A_1 \geqslant A^2\sqrt{2/\log 2}$.

L^2-influence bound. Since $\Delta_i f_\omega \in \{0, \beta_B\}$, we have $(\Delta_i f_\omega)^2 = \beta_B \, \Delta_i f_\omega$ pointwise, hence $I_i^{(2)}(f_\omega) = \beta_B \, I_i(f_\omega)$ for each $i \in S_0$. Summing over $i \in S_0$,

$$I^{(2)}(f_\omega) \;=\; \beta_B \, I(f_\omega) \;\leqslant\; A\beta_B^2\sqrt{s} \;=\; \frac{AB}{A_1} \;\leqslant\; B,$$

using $I(f_\omega) \leqslant A\beta_B\sqrt{s}$ from the L^1-bound above and $A_1 \geqslant A$. Hence $f_\omega \in \mathcal{F}_{K,B}$.

Separation. By the same calculation as (25),

$$\|f_\omega - f_{\omega'}\|_2^2 \;\geqslant\; \frac{c'\beta_B^2}{\sqrt{s}} \;=\; \frac{c'B}{A_1 s},$$

where $c' > 0$ is an absolute constant.

Fano condition. $\bar{\mathrm{KL}} \leqslant n\beta_B^2/(2\sigma^2) = nB/(2\sigma^2 A_1\sqrt{s})$. Using $B \leqslant K^2/\sqrt{\log n} \leqslant c^2\sqrt{\log n}$ and $\sqrt{s} \geqslant \sqrt{\log n/\log 2}$,

$$\bar{\mathrm{KL}} \;\leqslant\; \frac{c^2 n\sqrt{\log 2}}{2\sigma^2 A_1} \;=\; O(n) \;\ll\; \tfrac{1}{2}\log|\Omega| \;=\; \Omega\!\left(\frac{n^2}{\sqrt{\log n}}\right).$$

The Fano condition is therefore satisfied for n large enough.

Conclusion. Applying Theorem 6.1 with $\delta := \sqrt{c'B/(4A_1 s)}$ and using $s \leqslant (2/\log 2)\log n$,

$$\inf_{\hat f} \sup_{f \in \mathcal{F}_{K,B}} R(\hat f, f) \;\geqslant\; p_0\,\delta^2 \;=\; \frac{p_0 c' B}{4A_1 s} \;\geqslant\; \frac{cB}{\log n},$$

which is (4), adjusting c. □

7 Discussion

Theorem 2.1 leaves a gap between the upper and lower bounds:

$$\frac{c\,K^2}{(\log n)^{3/2}} \;\leqslant\; \inf_{\hat f} \sup_{f \in \mathcal{F}_K} R(\hat f, f) \;\leqslant\; \frac{C\,K}{\sqrt{\log n}}.$$

However, Theorem 2.2 shows that the picture is sharper than it appears: stratifying $\mathcal{F}_K$ by $I^{(2)}(f)$, the gap on each sub-class $\mathcal{F}_{K,B}$ reduces to $\sqrt{\log n}$, with matching bounds $CB/\sqrt{\log n}$ (upper) and $cB/\log n$ (lower). The residual $\log n$ gap on the full class $\mathcal{F}_K$ is an aggregation artifact: the upper bound is largest when $I^{(2)}(f) \approx K$ (near-Boolean functions, for which $\Delta_i f \in \{0, 1\}$ and $I^{(2)} = I$), whereas the lower bound construction achieves $I^{(2)}(f_\omega) \asymp K^2/\sqrt{\log n}$. We discuss the sources of this remaining gap and several directions for further work.

The upper bound bottleneck. The rate $K/\sqrt{\log n}$ arises from a bias-variance trade-off in the degree parameter d_0: the bias decreases as d_0 grows, but the estimated spectral set $\hat{\mathcal{S}}$ has cardinality at most $e^{Cd_0^2}$ for a constant C depending on γ (see (15)), forcing $d_0 \lesssim \sqrt{\log n}$ to keep the variance bounded.

The root cause is that our estimator includes all subsets of $\hat J$ of size $\leqslant d_0$ in $\hat{\mathcal{S}}$, many of which carry negligible Fourier weight. Identifying from data the relevant Fourier subsets $\mathcal{S} \subseteq 2^J$ rather than all subsets of $\hat J$ would allow a larger d_0 and a faster rate, but this higher-order selection problem remains open.

Open problem 7.1. Is there an estimator that identifies the relevant Fourier subsets $\mathcal{S}$ (not just the influential coordinates J) from noisy observations, with estimation cost proportional to $|\mathcal{S}|$ rather than $|\hat{\mathcal{S}}|$?

The lower bound bottleneck. The rate $K^2/(\log n)^{3/2}$ reflects the geometry of the middle layer of $\{0,1\}^s$ ($s = \Theta(\log n)$): its μ-measure $\Theta(1/\sqrt{s})$ forces the scaling $\beta = \Theta(K/\sqrt{s})$ to fit the influence budget, yielding a squared separation of order $K^2/s^{3/2}$. Improving the lower bound would require either packing functions that vary on a larger fraction of the hypercube, or a testing argument that goes beyond Fano's inequality.

Open problem 7.2. What is the exact minimax rate of estimation over $\mathcal{F}_{K,B}$? Can the $\sqrt{\log n}$ gap between $CB/\sqrt{\log n}$ and $cB/\log n$ in Theorem 2.2 be closed?

Our conjecture is that the upper bound $CB/\sqrt{\log n}$ is the correct rate, and that the lower bound construction does not capture the full difficulty of $\mathcal{F}_{K,B}$.

Comparison with isotonic regression on $[0,1]^d$. For isotonic regression on $[0,1]^d$, Han et al. (2019) showed that the minimax rate is $n^{-1/d}$ (up to logarithmic factors), which becomes uninformative for $d \gtrsim \log n$. On the Boolean hypercube, the influence constraint rescues estimation even for d nearly exponential in n: the rate $K/\sqrt{\log n}$ remains finite under the mild condition $\log d \leqslant n^{1-\varepsilon}$. The two settings are not directly comparable (different domains, different structural assumptions), but the qualitative message is the same: meaningful high-dimensional estimation requires constraints beyond monotonicity. Whether an analogue of the influence budget can be defined on $[0,1]^d$ is an interesting open question.

Computational cost. Step 1 of the estimator (influence estimation) requires $O(nd)$ operations. Step 2 (Fourier estimation) computes one average per element of $\hat{\mathcal{S}}$, at cost $O(n|\hat{\mathcal{S}}|)$. Since $\log|\hat{\mathcal{S}}| = O(\log n)$, the total cost is $O(n^c)$ for a constant c depending on γ—polynomial in n, but with a potentially large exponent. In practice, one may choose a smaller d_0 (e.g., $d_0 = 2$ or 3), accepting a larger bias in exchange for a much smaller computational cost.

Extensions. Our results are stated for the uniform measure μ on $\{0,1\}^d$. The Fourier expansion and Bonami–Beckner inequality extend to general product measures $\mu_1 \otimes \cdots \otimes \mu_d$ on $\{0,1\}^d$ (O'Donnell, 2014), and the spectral concentration argument carries over with modified constants when the marginal probabilities $\mu_i(\{1\})$ are bounded away from 0 and 1. The degenerate case (some marginals close to 0 or 1) introduces additional difficulties and is left for future work.

One might also consider the class $\{f \in \mathcal{M}_d : I^{(2)}(f) \leqslant K\}$, constraining the L^2-influence directly. The spectral concentration still applies (since $I^{(2)} \leqslant I$, this is a larger class), but the estimation of $I_i^{(2)}(f) = \mathbb{E}[(\Delta_i f(X))^2]$ from data is harder: unlike the L^1-influence, it is not a simple difference of means. Whether the minimax rate changes under this alternative constraint is an open question.

References

Beckner, W. (1975). Inequalities in Fourier analysis. *Ann. Math.*, 102:159–182.

Ben-David, A. (1995). Monotonicity maintenance in information-theoretic machine learning algorithms. *Mach. Learn.*, 19:29–43.

Bonami, A. (1970). Étude des coefficients de Fourier des fonctions de $L^p(G)$. *Ann. Inst. Fourier*, 20:335–402.

Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.*, 26:607–616.

Brunk, H. D. (1970). Estimation of isotonic regression. In Puri, M. L., editor, *Nonparametric Techniques in Statistical Inference*, pages 177–197. Cambridge University Press, Cambridge.

Bshouty, N. H. and Tamon, C. (1996). On the Fourier spectrum of monotone functions. *J. ACM*, 43:747–770.

Chatterjee, S., Guntuboyina, A., and Sen, B. (2015). On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.*, 43:1774–1800.

Comminges, L. and Dalalyan, A. S. (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *Ann. Statist.*, 40:2667–2696.

Feelders, A. (2010). Monotone relabeling in ordinal classification. In *Proc. IEEE ICDM*, pages 803–808. IEEE Computer Society.

Friedgut, E. (1998). Boolean functions with low average sensitivity depend on few coordinates. *Combinatorica*, 18:27–35.

Friedgut, E. (2004). Influences in product spaces: KKL and BKKKL revisited. *Combin. Probab. Comput.*, 13:17–29.

Garban, C. and Steif, J. E. (2014). *Noise Sensitivity of Boolean Functions and Percolation*. Cambridge University Press, Cambridge.

González, S., Herrera, F., and García, S. (2015). Monotonic random forest with an ensemble pruning mechanism based on the degree of monotonicity. *New Gener. Comput.*, 33:367–388.

Groeneboom, P. and Jongbloed, G. (2014). *Nonparametric Estimation under Shape Constraints*. Cambridge University Press, Cambridge.

Han, Q., Wang, T., Chatterjee, S., and Samworth, R. J. (2019). Isotonic regression in general dimensions. *Ann. Statist.*, 47:2440–2471.

Hatami, H. (2009). Decision trees and influences of variables over product probability spaces. *Combin. Probab. Comput.*, 18:357–369.

Kahn, J., Kalai, G., and Linial, N. (1988). The influence of variables on Boolean functions. In *Proc. 29th IEEE FOCS*, pages 68–80.

Kelman, E., Khot, S., Kindler, G., Minzer, D., and Safra, M. (2021). Theorems of KKL, Friedgut, and Talagrand via random restrictions and log-Sobolev inequality. In Lee, J. R., editor, *Proc. 12th ITCS*, volume 185, pages 26:1–26:17. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Kpotufe, S. (2011). k-NN regression adapts to local intrinsic dimension. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Adv. Neural Inf. Process. Syst.*, volume 24, pages 729–737. Curran Associates, Inc.

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

Mossel, E., O'Donnell, R., and Servedio, R. A. (2003). Learning juntas. In *Proc. 35th ACM STOC*, pages 206–212.

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

O'Donnell, R. and Servedio, R. A. (2007). Learning monotone decision trees in polynomial time. *SIAM J. Comput.*, 37:827–844.

Potharst, R. and Bioch, J. C. (1999). A decision tree algorithm for ordinal classification. In Hand, D. J., Kok, J. N., and Berthold, M. R., editors, *Proc. IDA*, pages 187–198. Springer.

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

van Eeden, C. (1958). *Testing and Estimating Ordered Parameters of Probability Distributions*. PhD thesis, Univ. Amsterdam.

Vershynin, R. (2018). *High-Dimensional Probability*. Cambridge University Press, Cambridge.

Weinreich, D. M., Lan, Y., Wylie, C. S., and Heckendorn, R. B. (2013). Should evolutionary geneticists worry about higher-order epistasis? *Curr. Opin. Genet. Dev.*, 23:700–707.

Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27:1564–1599.