Query-Efficient Gradients for a Fourier Relaxation of Junta Testing

J. R. Landers

July 3, 2026

Abstract

In tolerant junta testing, one is given query access to a Boolean function and asks whether it is well approximated by a function that depends on at most k input coordinates. A natural optimization-based approach replaces the hard search over k-subsets with a continuous search over soft coordinate weights $w \in [0,1]^n$ satisfying $\sum_i w_i \le k$. This note proves a simple but useful fact: for the Fourier-energy relaxation

$$\Phi_f(w) = \sum_{S \subseteq [n]} \widehat{f}(S)^2 \prod_{i \in S} w_i,$$

each gradient coordinate is exactly an anisotropic noisy influence, and it can be estimated from black-box queries to f. One unbiased sample uses four queries to f, and $O(\eta^{-2}\log(1/\delta))$ samples estimate one coordinate to additive error η with failure probability δ.

1 Context

Intuition

- Think of all Boolean functions as points in a huge space, and think of $\mathcal{J}_k$ as the low-dimensional surface of functions that use at most k coordinates.
- Tolerant testing asks how far f is from that surface, allowing f to be noisy but still near a good low-coordinate explanation.
- The optimizer viewpoint replaces jumping among discrete supports with moving continuously through soft coordinate weights toward that surface.

Let

$$f : \{\pm 1\}^n \to \{\pm 1\}$$

be an unknown Boolean function. A k-junta is a function that depends on at most k coordinates. The tolerant junta testing problem is to estimate

$$\operatorname{dist}(f, \mathcal{J}_k) = \min_{g \in \mathcal{J}_k} \mathbb{P}_x[f(x) \neq g(x)].$$

For Boolean f, this distance is determined by the best junta correlation

$$\operatorname{corr}(f, \mathcal{J}_k) = \max_{\substack{A \subseteq [n] \\ |A| \le k}} \max_{g \text{ depends only on } A} \mathbb{E}_x[f(x)g(x)]$$

by

$$\operatorname{dist}(f, \mathcal{J}_k) = \frac{1}{2}\big(1 - \operatorname{corr}(f, \mathcal{J}_k)\big).$$

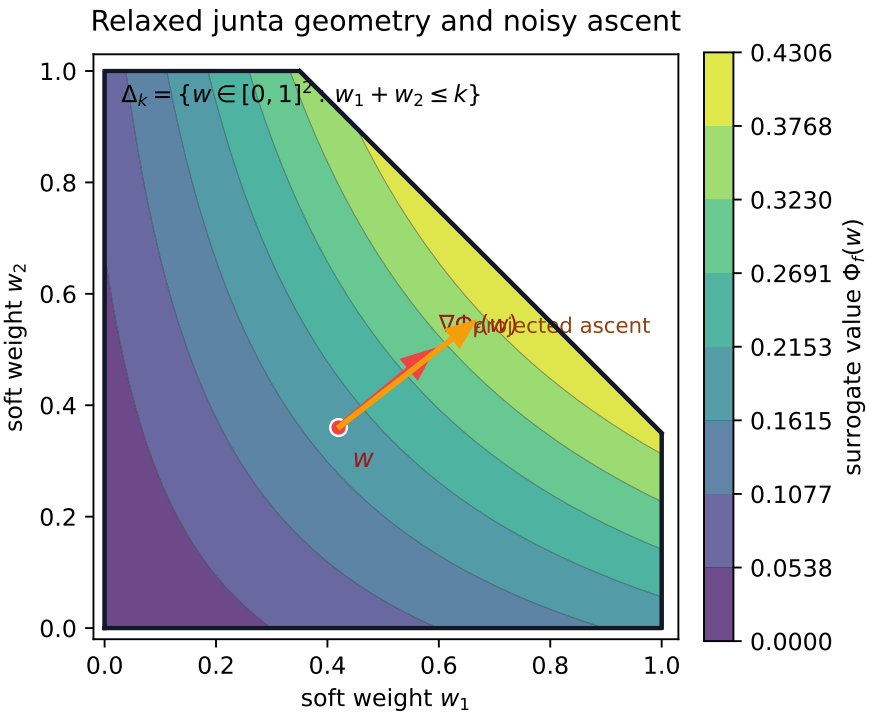

Figure 1: A two-coordinate slice of the relaxed junta polytope. The objective Φ_f extends captured Fourier energy from discrete supports to soft coordinate weights, so a gradient step must be projected back into Δ_k.

The recent work of Chen, Patel, and Servedio gives an adaptive tolerant tester with query complexity $2^{\widetilde{O}(k^{1/3})}$. Their algorithm is Fourier-analytic and combinatorial: it refines candidate coordinate sets using influence-like information and local estimation.

This note isolates a smaller optimization-theoretic fact. Instead of searching directly over sets A, introduce soft weights

$$\Delta_k = \left\{ w \in [0,1]^n : \sum_{i=1}^{n} w_i \le k \right\}.$$

The value $w_i = 1$ means that coordinate i is fully selected, while $w_i = 0$ means that it is ignored. The geometry is the capped cube shown in Figure 1.

2 The Relaxed Objective

Intuition

- Fourier mass tells us which coordinate directions f uses; ignoring a direction leaves some of f's structure at a distance from the chosen model.
- Choosing A is like projecting f onto the subspace spanned by Fourier terms supported inside A.
- The score $\Phi_f(w)$ softens this projection distance by letting each coordinate direction be partially present.

Write the Fourier expansion

$$f(x) = \sum_{S \subseteq [n]} \widehat{f}(S)\chi_S(x), \qquad \chi_S(x) = \prod_{i \in S} x_i.$$

For $A \subseteq [n]$, the conditional expectation of f onto the coordinates in A is

$$f_A(x) = \mathbb{E}[f(y) \mid y_A = x_A] = \sum_{S \subseteq A} \widehat{f}(S)\chi_S(x).$$

Hence

$$\|f_A\|_2^2 = \sum_{S \subseteq A} \widehat{f}(S)^2.$$

This is the Fourier energy captured by A. It is not the same as junta correlation, but it is a natural smooth surrogate for the amount of structure in f that lives on these coordinates.

Define

$$\Phi_f(w) = \sum_{S \subseteq [n]} \widehat{f}(S)^2 \prod_{i \in S} w_i, \qquad w \in \Delta_k.$$

For an integral point $w = \mathbf{1}_A$,

$$\Phi_f(\mathbf{1}_A) = \sum_{S \subseteq A} \widehat{f}(S)^2 = \|f_A\|_2^2.$$

Thus Φ_f is exactly the multilinear extension of captured Fourier energy.

3 Main Result

Intuition

- The derivative $D_i f$ asks how much distance remains in the coordinate-i direction.
- Smoothing $D_i f$ by the current weights measures whether that direction still helps after the present soft projection has been applied.
- The theorem says this remaining smoothed distance is exactly the gradient coordinate that tells the optimizer where to move.

For $i \in [n]$, define the discrete derivative

$$D_i f(x_{-i}) = \frac{f(x_{-i}, +1) - f(x_{-i}, -1)}{2}.$$

If f is Boolean-valued, then $D_i f(x_{-i}) \in \{-1, 0, 1\}$. Its Fourier expansion is

$$D_i f = \sum_{T \subseteq [n] \setminus \{i\}} \widehat{f}(T \cup \{i\}) \chi_T.$$

For $w \in [0,1]^n$, set

$$\rho_j = \sqrt{w_j}.$$

Let $T_\rho^{(-i)}$ be the anisotropic noise operator on all coordinates except i, meaning

$$T_\rho^{(-i)} \chi_T = \left(\prod_{j \in T} \rho_j \right) \chi_T, \qquad T \subseteq [n] \setminus \{i\}.$$

Theorem 1 (Gradient equals noisy derivative energy)**.** *For every $f : \{\pm 1\}^n \to \{\pm 1\}$, every $w \in [0,1]^n$, and every $i \in [n]$,*

$$\frac{\partial \Phi_f}{\partial w_i}(w) = \left\| T_{\sqrt{w}}^{(-i)} D_i f \right\|_2^2.$$

Proof. Differentiate the multilinear polynomial:

$$\frac{\partial \Phi_f}{\partial w_i}(w) = \sum_{\substack{S \subseteq [n] \\ i \in S}} \widehat{f}(S)^2 \prod_{j \in S \setminus \{i\}} w_j.$$

Writing $S = T \cup \{i\}$, this becomes

$$\sum_{T \subseteq [n] \setminus \{i\}} \widehat{f}(T \cup \{i\})^2 \prod_{j \in T} w_j.$$

On the other hand,

$$T_{\sqrt{w}}^{(-i)} D_i f = \sum_{T \subseteq [n] \setminus \{i\}} \widehat{f}(T \cup \{i\}) \left(\prod_{j \in T} \sqrt{w_j} \right) \chi_T.$$

Taking the squared L_2 norm and using Parseval gives exactly the previous display. □

Thus the gradient coordinate is not merely analogous to influence. It is an influence measurement obtained after smoothing the derivative in all other coordinates. This is precisely the quantity an optimizer needs in order to decide whether coordinate i should receive more mass.

4 A Four-Query Unbiased Estimator

Intuition

- A gradient coordinate is a squared distance, so the estimator measures a norm rather than a signed slope directly.
- Two independent noisy copies give two views of the same local distance; their product is unbiased for the squared smoothed derivative.
- Repeating this bounded distance measurement makes the estimate concentrate around the true gradient coordinate.

The identity above is useful only if the right-hand side can be estimated from queries to f. The next theorem gives such an estimator. Figure 2 shows one sample.

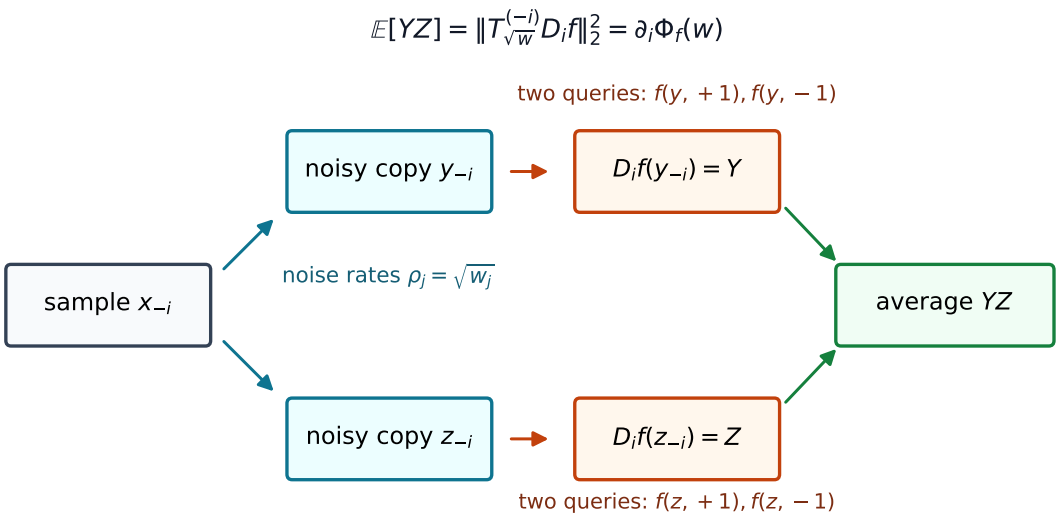

Figure 2: One unbiased sample for a gradient coordinate. Draw a base point x_{-i}, draw two independent noisy copies y_{-i}, z_{-i}, estimate the two discrete derivatives using four total queries, and average YZ.

Theorem 2 (Coordinate estimator)**.** *Fix $w \in [0,1]^n$, $i \in [n]$, and $\eta, \delta \in (0,1)$. There is a randomized query algorithm that outputs $\widehat{g}_i$ satisfying*

$$\left| \widehat{g}_i - \frac{\partial \Phi_f}{\partial w_i}(w) \right| \leq \eta$$

with probability at least $1 - \delta$, using

$$O\left(\eta^{-2} \log(1/\delta)\right)$$

queries to f.

Proof. One sample is generated as follows.

1. Draw $x_{-i} \in \{\pm 1\}^{n-1}$ uniformly.
2. Independently draw y_{-i} and z_{-i} as anisotropic noisy copies of x_{-i}: for each $j \neq i$, keep x_j with probability $\rho_j = \sqrt{w_j}$, and otherwise resample uniformly from $\{\pm 1\}$.
3. Query f four times to compute

$$Y = D_i f(y_{-i}) = \frac{f(y_{-i}, +1) - f(y_{-i}, -1)}{2},$$

and

$$Z = D_i f(z_{-i}) = \frac{f(z_{-i}, +1) - f(z_{-i}, -1)}{2}.$$

4. Return YZ.

Conditioned on x_{-i}, the two noisy copies are independent. Also,

$$\mathbb{E}[Y \mid x_{-i}] = T^{(-i)}_{\sqrt{w}} D_i f(x_{-i}), \qquad \mathbb{E}[Z \mid x_{-i}] = T^{(-i)}_{\sqrt{w}} D_i f(x_{-i}).$$

Therefore

$$\mathbb{E}[YZ \mid x_{-i}] = \left(T^{(-i)}_{\sqrt{w}} D_i f(x_{-i})\right)^2.$$

Averaging over x_{-i},

$$\mathbb{E}[YZ] = \left\| T^{(-i)}_{\sqrt{w}} D_i f \right\|_2^2 = \frac{\partial \Phi_f}{\partial w_i}(w).$$

Since $Y, Z \in [-1, 1]$, also $YZ \in [-1, 1]$. Averaging

$$M = O\left(\eta^{-2} \log(1/\delta)\right)$$

independent samples and applying Hoeffding's inequality proves the claim. □

Corollary 1 (Full gradient on a reduced universe)**.** *Suppose the current coordinate universe has size m. There is a randomized query algorithm using*

$$O\left(m\eta^{-2} \log(m/\delta)\right)$$

queries to f that returns $\widehat{g} \in \mathbb{R}^m$ with

$$\|\widehat{g} - \nabla \Phi_f(w)\|_\infty \le \eta$$

with probability at least $1 - \delta$.

Proof. Run the coordinate estimator for each of the m coordinates with failure probability δ/m, then take a union bound. □

5 Precision for Projected Ascent

Intuition

- The feasible region Δ_k limits how far the optimizer may move in coordinate-weight space.
- An ℓ_∞ gradient error creates only a controlled error in the distance comparison between any two feasible moves.
- Choosing η on the order of ε/k makes the noisy arrow point close enough to a valid distance-reducing direction.

The ℓ_∞ guarantee is the right norm for optimization over Δ_k. If $u, v \in \Delta_k$, then

$$\|u - v\|_1 \le 2k.$$

Therefore

$$|\langle \widehat{g} - \nabla \Phi_f(w), u - v \rangle| \le 2k\eta.$$

Choosing

$$\eta \le \frac{\varepsilon}{8k}$$

makes every linearized ascent comparison over Δ_k accurate up to $\varepsilon/4$. In particular, if

$$\widehat{s} \in \arg\max_{s \in \Delta_k} \langle \widehat{g}, s \rangle,$$

then

$$\langle \nabla \Phi_f(w), \widehat{s} \rangle \geq \max_{s \in \Delta_k} \langle \nabla \Phi_f(w), s \rangle - 2k\eta.$$

Since a linear function over Δ_k is maximized by putting mass on the coordinates with the largest positive coefficients, this gives the query-estimated direction needed for Frank-Wolfe or projected ascent.

6 What This Does and Does Not Prove

Intuition

- This note proves that local distance-reducing directions for the relaxation can be estimated efficiently from queries.
- It does not prove that following those local directions always reaches the nearest k-junta surface.
- The result is one rigorous module in a larger program: turn distance to $\mathcal{J}_k$ into an optimizer geometry with provable query access.

This note proves that the optimizer relaxation is operational: its stochastic gradients can be estimated from black-box queries to f, and the estimator has clean concentration.

It does not prove that optimizing Φ_f solves tolerant junta testing. The missing landscape question is whether high values of Φ_f, or stationary points reached by noisy ascent, reliably yield coordinate sets with near-optimal junta correlation. That is a harder structural statement.

The result should therefore be read as a modular ingredient:

the Fourier-energy relaxation has query-efficient stochastic gradients.

This is the piece one needs before an optimizer-based approach to tolerant junta testing can be analyzed seriously.

7 A Proven Energy-to-Distance Bridge

Intuition

- The relaxation does not estimate distance directly; it estimates how much Fourier energy is captured by a coordinate subspace.
- For Boolean functions, captured energy and junta correlation are trapped between two elementary distance bounds.
- This gives a rigorous bridge in the near-junta regime: if f is close to a k-junta, then nearly maximizing captured energy gives a nearly good junta approximation.

The missing global issue is optimization landscape, not the relationship between captured energy and distance. The latter admits a simple deterministic bridge.

For $A \subseteq [n]$, write

$$h_A(x) = \mathbb{E}[f(y) \mid y_A = x_A].$$

Then

$$\Phi_f(\mathbf{1}_A) = \|h_A\|_2^2$$

and

$$\mathrm{corr}(f, \mathcal{J}_A) = \mathbb{E}|h_A(x)|.$$

Lemma 1 (Energy-correlation sandwich)**.** *For every Boolean $f : \{\pm 1\}^n \to \{\pm 1\}$ and every $A \subseteq [n]$,*

$$\Phi_f(\mathbf{1}_A) \le \mathrm{corr}(f, \mathcal{J}_A) \le \sqrt{\Phi_f(\mathbf{1}_A)}.$$

Proof. Since f is Boolean-valued, the conditional expectation h_A satisfies

$$|h_A(x)| \le 1$$

for every x. Hence

$$h_A(x)^2 \le |h_A(x)|.$$

Averaging gives

$$\Phi_f(\mathbf{1}_A) = \mathbb{E}[h_A(x)^2] \le \mathbb{E}|h_A(x)| = \mathrm{corr}(f, \mathcal{J}_A).$$

The other inequality is Cauchy–Schwarz:

$$\mathrm{corr}(f, \mathcal{J}_A) = \mathbb{E}|h_A(x)| \le \sqrt{\mathbb{E}[h_A(x)^2]} = \sqrt{\Phi_f(\mathbf{1}_A)}.$$

□

Define the optimal captured energy

$$P_k(f) = \max_{\substack{A \subseteq [n] \\ |A| \le k}} \Phi_f(\mathbf{1}_A),$$

and the optimal junta correlation

$$C_k(f) = \max_{\substack{A \subseteq [n] \\ |A| \le k}} \mathrm{corr}(f, \mathcal{J}_A).$$

Theorem 3 (Energy brackets junta distance)**.** *For every Boolean $f : \{\pm 1\}^n \to \{\pm 1\}$,*

$$P_k(f) \le C_k(f) \le \sqrt{P_k(f)}.$$

Equivalently,

$$\frac{1 - \sqrt{P_k(f)}}{2} \le \mathrm{dist}(f, \mathcal{J}_k) \le \frac{1 - P_k(f)}{2}.$$

Proof. The lower bound $P_k(f) \le C_k(f)$ follows by applying the previous lemma to a set A maximizing $\Phi_f(\mathbf{1}_A)$. For the upper bound, let $A^\star$ maximize $\mathrm{corr}(f, \mathcal{J}_A)$. The lemma gives

$$\mathrm{corr}(f, \mathcal{J}_{A^\star}) \le \sqrt{\Phi_f(\mathbf{1}_{A^\star})} \le \sqrt{P_k(f)}.$$

Thus $C_k(f) \le \sqrt{P_k(f)}$. The distance statement follows from

$$\mathrm{dist}(f, \mathcal{J}_k) = \frac{1}{2}(1 - C_k(f)).$$

□

This theorem already explains why the surrogate is meaningful: when $P_k(f)$ is near 1, the true distance to $\mathcal{J}_k$ is also small. Moreover, a near-maximizer of captured energy gives a concrete near-junta certificate.

Corollary 2 (Near-maximal energy gives a near-junta)**.** *Let $A \subseteq [n]$, $|A| \le k$, satisfy*

$$\Phi_f(\mathbf{1}_A) \ge P_k(f) - \alpha.$$

Then

$$\operatorname{corr}(f, \mathcal{J}_A) \ge C_k(f)^2 - \alpha,$$

and therefore

$$\operatorname{dist}(f, \mathcal{J}_A) \le 2 \operatorname{dist}(f, \mathcal{J}_k) + \frac{\alpha}{2}.$$

In particular, if f is τ-close to a k-junta, then A is

$$\left(2\tau + \frac{\alpha}{2}\right)\text{-close}$$

to a k-junta over the coordinates in A.

Proof. The previous theorem gives

$$P_k(f) \ge C_k(f)^2.$$

Thus

$$\Phi_f(\mathbf{1}_A) \ge C_k(f)^2 - \alpha.$$

By the energy-correlation sandwich,

$$\operatorname{corr}(f, \mathcal{J}_A) \ge \Phi_f(\mathbf{1}_A) \ge C_k(f)^2 - \alpha.$$

Therefore

$$\operatorname{dist}(f, \mathcal{J}_A) = \frac{1}{2}\big(1 - \operatorname{corr}(f, \mathcal{J}_A)\big) \le \frac{1}{2}\big(1 - C_k(f)^2 + \alpha\big).$$

Since

$$1 - C_k(f)^2 = (1 - C_k(f))(1 + C_k(f)) \le 2(1 - C_k(f)),$$

we get

$$\operatorname{dist}(f, \mathcal{J}_A) \le 1 - C_k(f) + \frac{\alpha}{2} = 2 \operatorname{dist}(f, \mathcal{J}_k) + \frac{\alpha}{2}.$$

□

The corollary is not yet a full tolerant tester. It proves that global or near-global optimization of the integral captured-energy objective is a valid route to a near-junta certificate when f is close to $\mathcal{J}_k$. What remains open is whether stochastic ascent on the continuous relaxation reliably finds such a near-maximizer, and whether the resulting estimator has enough resolution in the far case.

8 Continued Research

Intuition

- The present note gives a cheap local compass: it estimates which coordinate directions appear to reduce distance to the junta surface.

- The missing theorem is global: following this compass must be shown to reach a support whose junta correlation is nearly optimal.
- If such a landscape theorem exists, the optimizer picture could explain, or in special regimes compete with, the adaptive tester of Chen–Patel–Servedio.

The gradient estimator above is much cheaper than the full tolerant junta tester of Chen, Patel, and Servedio, but it solves a strictly smaller problem. Their theorem estimates

$$\mathrm{dist}(f, \mathcal{J}_k)$$

using

$$2^{\widetilde{O}(k^{1/3})}$$

queries. In contrast, the result here estimates one gradient coordinate of Φ_f using

$$O\left(\eta^{-2}\log(1/\delta)\right)$$

queries, and estimates the full gradient over a reduced coordinate universe of size m using

$$O\left(m\eta^{-2}\log(m/\delta)\right)$$

queries.

With the projected-ascent precision choice

$$\eta \le \frac{\varepsilon}{8k},$$

one full gradient estimate costs

$$O\left(mk^2\varepsilon^{-2}\log(m/\delta)\right)$$

queries. If a coordinate-oracle reduction or preprocessing step reduces the ambient dimension to

$$m = \mathrm{poly}(k, 1/\varepsilon),$$

then each gradient step is polynomial in k and $1/\varepsilon$. This is far below $2^{\widetilde{O}(k^{1/3})}$ as a local primitive. The difficulty is that a cheap local primitive is not yet a distance estimator.

The speculative target is therefore the following.

Theorem 4 (Speculative optimizer-to-tester theorem)**.** *Suppose $f : \{\pm1\}^n \to \{\pm1\}$ is accessed by queries, and suppose the optimization landscape of Φ_f over Δ_k satisfies a suitable no-bad-attractors condition: from a standard initialization, noisy projected ascent using ℓ_∞-accurate gradient estimates reaches, after $T = \mathrm{poly}(k, 1/\varepsilon)$ iterations, a point $w \in \Delta_k$ that can be rounded to a set $A \subseteq [n]$, $|A| \le k$, with*

$$\mathrm{corr}(f, \mathcal{J}_A) \ge \mathrm{corr}(f, \mathcal{J}_k) - \varepsilon.$$

Then the resulting optimizer-based tester estimates

$$\mathrm{dist}(f, \mathcal{J}_k) = \frac{1}{2}\left(1 - \mathrm{corr}(f, \mathcal{J}_k)\right)$$

to additive error $O(\varepsilon)$ using

$$\mathrm{poly}(k, 1/\varepsilon, \log(1/\delta))$$

queries after reduction to a polynomial-size coordinate universe.

This theorem is not proved here. Its main unproved hypothesis is the landscape condition. In geometric terms, the open question is whether the relaxed objective Φ_f has reliable paths down toward the nearest k-junta surface, or whether it contains local attractors that point toward coordinate sets with high captured L_2 Fourier energy but poor junta correlation.

The comparison should therefore be read carefully:

the gradient bound is stronger locally, but the Servedio-paper bound is stronger globally.

The continued research problem is to close this gap. One possible route is to prove the speculative theorem under structural assumptions, such as separation between relevant and irrelevant coordinates, bounded high-degree Fourier tail, or closeness to a unique k-junta. A second route is to show that these assumptions fail in the worst case, which would explain why the Chen–Patel–Servedio algorithm needs its sharper combinatorial refinement machinery.