Penn State University
CSE 564 Complexity of Combinatorial Problems

Instructor: Pei Wu

Module
9

Rubinfeld-Sudan low degree test

In the PCP construction we encode an exponentially-long object (a proof) as a low-degree polynomial over a finite field, and then verify local consistency using only a few oracle queries. This lecture develops the *low-degree test*: a randomized procedure that, given oracle access to a function $f : \mathbb{F}^v \to \mathbb{F}$, distinguishes the case that f is (exactly) a low-degree polynomial from the case that f is far from *every* low-degree polynomial.
We start with the univariate situation, where interpolation gives a simple test, and then derive a more "local" characterization based on finite differences. Randomizing the step size yields a robust constraint that naturally extends to multivariate functions by checking restrictions to random lines. Finally, we sketch a BLR-style analysis showing that if the line test accepts f with high probability, then f must be close to an actual low-degree polynomial.

9.1 Parameter setting and goal

In the PCP regime, the parameters are chosen so that the domain $\mathbb{F}^v$ has size about n while the degree bounds remain polylogarithmic in n. A typical choice (as in the notes) is:

$$v \;=\; \frac{\log n}{\log\log n} \qquad (\#\text{ variables}), \qquad |\mathbb{F}| \;=\; \text{poly}(\log n),$$

together with an *individual* degree bound

$$d \;=\; \log n,$$

and a *total* degree bound

$$D \;=\; \frac{(\log n)^2}{\log\log n} \;=\; d \cdot v.$$

(For the low-degree test below, the relevant bound is D; an individual degree bound is often tracked in PCP applications as well.)

THEOREM 9.1 (Low-degree test). *Given oracle access to a function $f : \mathbb{F}^v \to \mathbb{F}$, there is a randomized tester making $q = \text{poly}(D)$ queries such that:*

1. *If f is a polynomial of total degree at most D, then the tester accepts with probability 1.*

2. *If f is δ-far from every degree-$\leq D$ polynomial (i.e. $\Delta(f, \mathcal{P}_{\leq D}) \geq \delta$), then the tester accepts with probability at most $1/3$.*

The rest of the lecture develops the main ingredient: a low-degree test based on checking that f looks low-degree when restricted to a random line.

9.2 Warm-up: the univariate case

Suppose $f : \mathbb{F} \to \mathbb{F}$ is univariate. A natural idea is to reconstruct the candidate polynomial by interpolation.

Example 9.2 (Naive interpolation tester). Fix a degree bound D.

1. Choose $D+1$ random points $x_0, \ldots, x_D \in \mathbb{F}$ and query $f(x_i)$ for all i.

2. Interpolate the unique degree-$\leq D$ polynomial $\widetilde{p}$ satisfying $\widetilde{p}(x_i) = f(x_i)$.

3. Query one additional random point $x' \in \mathbb{F}$ and accept iff $f(x') = \widetilde{p}(x')$.

If f is truly a degree-$\leq D$ polynomial, then $\widetilde{p} = f$ and the test accepts. If f is δ-far from every degree-$\leq D$ polynomial, then for any fixed $\widetilde{p}$ we have $\Pr_{x'}[f(x') \neq \widetilde{p}(x')] \geq \delta$, so the test rejects with probability at least δ. The drawback is that this uses $D+2$ queries and does not immediately generalize to multivariate domains in a "local" way.

9.3 A local characterization via finite differences

A standard tool for low-degree testing is the *finite difference* (discrete derivative) operator

$$(\nabla f)(x) \;:=\; f(x+1) - f(x).$$

For a polynomial p of degree D, each application of ∇ reduces the degree by 1, so after $D+1$ applications one obtains the zero function.

Proposition 9.3 (Local characterization of univariate low degree). *Let $D \geq 0$. Define constants*

$$C_i \;:=\; (-1)^i \binom{D+1}{i} \qquad (i = 0, 1, \ldots, D+1).$$

Then a univariate function $p : \mathbb{F} \to \mathbb{F}$ is a polynomial of degree at most D if and only if, for every $x \in \mathbb{F}$,

$$\sum_{i=0}^{D+1} C_i \, p(x+i) \;=\; 0. \tag{9.1}$$

Proof sketch. The $(D+1)$-st finite difference can be written explicitly as

$$(\nabla^{D+1} p)(x) \;=\; \sum_{i=0}^{D+1} (-1)^i \binom{D+1}{i} p(x+i),$$

which is exactly (9.1). If $\deg(p) \le D$ then $\nabla^{D+1} p \equiv 0$, so the identity holds for all x.

Conversely, if (9.1) holds for all x, then $\nabla^{D+1} p \equiv 0$. This forces p to be a polynomial of degree at most D (over a field, repeated discrete differentiation eventually reaching 0 characterizes polynomials of bounded degree). □

For intuition, note that the second difference is

$$(\nabla^2 f)(x) = \nabla(f(x+1) - f(x)) = f(x+2) - 2f(x+1) + f(x),$$

and higher differences expand according to the binomial coefficients.

A first attempt at a tester suggested by Proposition 9.3 is: pick random x, query $f(x), f(x+1), \ldots, f(x+D+1)$, and check (9.1). In the notes this is marked "does not work" as a robust tester (soundness does not follow just from the fact that the identity characterizes low degree when it holds *everywhere*).

9.4 Random step size and the RS line test

A standard fix is to randomize the step size. Instead of checking the constraint on consecutive points $x, x+1, \ldots$, we check it on a random arithmetic progression.

Corollary 9.4 (Adjusted local characterization). *A univariate function $f : \mathbb{F} \to \mathbb{F}$ has degree at most D if and only if, for all $x, y \in \mathbb{F}$,*

$$\sum_{i=0}^{D+1} C_i \, f(x + iy) \;=\; 0, \tag{9.2}$$

where $C_i = (-1)^i \binom{D+1}{i}$ as above.

Proof sketch. Fix x, y (with $y \neq 0$) and consider the univariate function $q(t) = f(ty)$. If f is a degree-$\le D$ polynomial, then so is q, and Proposition 9.3 applied to q gives (9.2). Conversely, if (9.2) holds for all x, y, then in particular every restriction $t \mapsto f(ty)$ satisfies the finite-difference characterization of degree $\le D$. □

This immediately suggests the (Rabinfield–Sudan / Reed–Solomon) *line test*.

Definition 9.5 (RS line test). Given oracle access to $f : \mathbb{F}^m \to \mathbb{F}$ and a degree bound D:

1. Pick $x, y \in \mathbb{F}^m$ uniformly at random.
2. Query $f(x), f(x+y), \ldots, f(x+(D+1)y)$.
3. Accept iff $\sum_{i=0}^{D+1} C_i f(x+iy) = 0$.

The test makes $D+2$ queries. If f is a degree-$\le D$ polynomial, then every restriction to a line $t \mapsto f(x+ty)$ is a univariate degree-$\le D$ polynomial, so the test accepts with probability 1. A key point (used later) is that when $|\mathbb{F}| > 2D$, the local "all lines have degree $\le D$" condition forces f to be a genuine multivariate polynomial of total degree at most D.

9.5 From local line conditions to multivariate low degree

We sketch why the local characterization generalizes to multivariate polynomials when $|\mathbb{F}| > 2D$. The overall structure is induction on the number of variables; it suffices to explain the bivariate case.

CLAIM 9.6 (Proof idea in the bivariate case). *Let $f : \mathbb{F}^2 \to \mathbb{F}$ satisfy the line condition*

$$\sum_{i=0}^{D+1} C_i f((x_1, x_2) + i(y_1, y_2)) = 0 \qquad \text{for all } (x_1, x_2), (y_1, y_2) \in \mathbb{F}^2.$$

Assuming $|\mathbb{F}| > 2D$, f must be a bivariate polynomial of total degree at most D.

Proof sketch. Fix $\beta \in \mathbb{F}$ and define the slice $f_\beta(x) := f(x, \beta)$. By applying the line condition to lines parallel to the x-axis, each f_β is a univariate degree-$\leq D$ polynomial.

Let $S = \{0, 1, \ldots, D\} \subseteq \mathbb{F}$ and for each $\beta \in S$ let $L_\beta(y)$ be the Lagrange basis polynomial of degree $\leq D$ that equals 1 at $y = \beta$ and 0 at the other points of S. Define

$$g(x, y) \;:=\; \sum_{\beta \in S} L_\beta(y)\, f_\beta(x).$$

By construction, $g(x, \beta) = f(x, \beta)$ for all $\beta \in S$ and all $x \in \mathbb{F}$. On the other hand, for each fixed x, the function $y \mapsto f(x, y)$ is degree-$\leq D$ (line condition on vertical lines), so it is uniquely determined by its values on the $D + 1$ points in S. Hence $g(x, y) = f(x, y)$ for all x, y.

Now g is a polynomial: it has degree $\leq D$ in y (by construction) and for every fixed y it has degree $\leq D$ in x (since it is a linear combination of the f_β's). It remains to show that the *total* degree is $\leq D$. Suppose for contradiction that g has total degree $D' > D$, and let $g_{D'}(x, y) = \sum_{i+j=D'} a_{ij} x^i y^j$ be its homogeneous degree-D' part. Consider the restriction to a line through the origin:

$$t \;\mapsto\; g(\alpha t, \beta t).$$

The coefficient of $t^{D'}$ in this univariate polynomial is $\sum_{i+j=D'} a_{ij} \alpha^i \beta^j$, which is a nonzero bivariate polynomial in (α, β) of degree D'. By Schwartz–Zippel, it is nonzero for some $(\alpha, \beta) \in \mathbb{F}^2$ as long as $|\mathbb{F}| > D'$; in the notes one uses $D' \leq 2D < |\mathbb{F}|$. For such (α, β), the restriction has degree D', contradicting the assumption that *every* line restriction has degree at most D. Therefore $D' \leq D$ and $f = g$ has total degree at most D. □

9.6 Soundness: a BLR-style analysis

We now state the main soundness statement proved in the notes for the RS line test. For a function $f : \mathbb{F}^m \to \mathbb{F}$, define its (relative) Hamming distance to another function g by

$$\Delta(f, g) \;:=\; \Pr_{x \in \mathbb{F}^m}[f(x) \neq g(x)], \qquad \Delta(f, \mathcal{P}_{\leq D}) \;:=\; \min_{p \in \mathcal{P}_{\leq D}} \Delta(f, p).$$

THEOREM 9.7 (Soundness of the RS line test). *If the RS line test accepts f with probability $1 - \varepsilon$, then either*

$$\varepsilon = \Omega\left(\frac{1}{D^2}\right) \qquad \text{or} \qquad \Delta(f, \mathcal{P}_{\leq D}) \leq 2\varepsilon.$$

Equivalently, the RS line test rejects with probability at least

$$\min\Big\{\, \Omega(1/D^2),\ \tfrac{1}{2}\,\Delta(f,\mathcal{P}_{\le D}) \,\Big\}.$$

COROLLARY 9.8. *There is an $O(D^3)$-query tester that rejects with probability at least $1/2$ whenever $\Delta(f,\mathcal{P}_{\le D}) = \Omega(1/D^2)$.*

Proof sketch. The analysis is "very much like the combinatorial proof of the BLR linearity test" in the notes. The key is to define a *decoded* function g_f that aggregates the local constraints by majority.

Step 1: majority decoding. For a fixed point $x \in \mathbb{F}^m$ and direction $y \in \mathbb{F}^m$, the RS constraint $\sum_{i=0}^{D+1} C_i f(x+iy) = 0$ can be solved for $f(x)$ because $C_0 = 1$:

$$f(x) \stackrel{?}{=} -\sum_{i=1}^{D+1} C_i f(x+iy).$$

Define $g_f(x)$ to be the most popular value of the right-hand side over random y:

$$g_f(x) \;:=\; \mathrm{Maj}_{y\in\mathbb{F}^m}\left[-\sum_{i=1}^{D+1} C_i\, f(x+iy)\right].$$

CLAIM 9.9 (Claim 1). $\varepsilon \ge \frac{1}{2}\,\Delta(f,g_f)$.

Proof. Fix $x \in \mathbb{F}^m$. If $f(x) \neq g_f(x)$, then by the definition of majority at least half of the directions y produce a value $-\sum_{i=1}^{D+1} C_i f(x+iy)$ that is *different* from $f(x)$. On such a choice of y, the RS identity

$$\sum_{i=0}^{D+1} C_i f(x+iy) = 0 \qquad (\text{with } C_0 = 1)$$

fails, so the RS line test rejects on the pair (x,y). Therefore

$$\Pr_y\big[\text{RS rejects on } (x,y)\big] \;\ge\; \tfrac{1}{2}\cdot \mathbf{1}[f(x)\neq g_f(x)].$$

Averaging over random x gives

$$\varepsilon \;=\; \Pr_{x,y}[\text{RS rejects on } (x,y)] \;\ge\; \tfrac{1}{2}\,\Pr_x[f(x)\neq g_f(x)] \;=\; \tfrac{1}{2}\,\Delta(f,g_f).$$

∎

CLAIM 9.10 (Claim 3). *For every fixed $x \in \mathbb{F}^m$,*

$$\Pr_{y\in\mathbb{F}^m}\left[g_f(x) = -\sum_{i=1}^{D+1} C_i f(x+iy)\right] \;\ge\; 1 - 2(D+1)\varepsilon.$$

Proof. Fix x and define the random variable

$$A_y(x) \;:=\; -\sum_{i=1}^{D+1} C_i\, f(x+iy) \qquad \text{for } y \in \mathbb{F}^m.$$

By definition, $g_f(x)$ is a value of $A_y(x)$ that occurs with maximum probability, so

$$\Pr_y[g_f(x) = A_y(x)] \;=\; \max_{v\in\mathbb{F}} \Pr_y[A_y(x) = v].$$

For any distribution $(p_v)_{v\in\mathbb{F}}$ we have $\max_v p_v \geq \sum_v p_v^2$ (since $\sum_v p_v^2 \leq (\max_v p_v)\sum_v p_v = \max_v p_v$), hence

$$\begin{aligned}\Pr_y[g_f(x) = A_y(x)] &\geq \sum_{v\in\mathbb{F}} \Pr_y[A_y(x) = v]^2 \\ &= \Pr_{y,z}[A_y(x) = A_z(x)],\end{aligned}$$

where y, z are independent and uniform.

We now lower bound $\Pr_{y,z}[A_y(x) = A_z(x)]$. Consider the $(D+2)\times(D+2)$ grid of points

$$x + iy + jz \qquad (i, j \in \{0, 1, \ldots, D+1\}).$$

Suppose the RS constraint holds on all the following $2(D+1)$ lines in this grid:

1. *Columns*: for every $i \in \{1, \ldots, D+1\}$, the RS constraint holds for basepoint $x + iy$ and direction z, i.e.
$$f(x+iy) \;=\; -\sum_{j=1}^{D+1} C_j\, f(x+iy+jz).$$
2. *Rows*: for every $j \in \{1, \ldots, D+1\}$, the RS constraint holds for basepoint $x + jz$ and direction y, i.e.
$$f(x+jz) \;=\; -\sum_{i=1}^{D+1} C_i\, f(x+iy+jz).$$

Under these assumptions we can compute:

$$\begin{aligned}A_y(x) &= -\sum_{i=1}^{D+1} C_i\, f(x+iy) = \sum_{i=1}^{D+1} C_i \left(\sum_{j=1}^{D+1} C_j\, f(x+iy+jz)\right) \\ &= \sum_{i=1}^{D+1}\sum_{j=1}^{D+1} C_i C_j\, f(x+iy+jz) \\ &= \sum_{j=1}^{D+1} C_j \left(\sum_{i=1}^{D+1} C_i\, f(x+iy+jz)\right) = -\sum_{j=1}^{D+1} C_j\, f(x+jz) = A_z(x),\end{aligned}$$

where the second and last equalities used the RS constraints on the columns and rows, respectively. Thus, if all these $2(D+1)$ RS constraints hold, then $A_y(x) = A_z(x)$.

Finally we bound the probability that all these constraints hold. For each fixed $i \in \{1, \dots, D+1\}$, the pair $(x+iy, z)$ is uniform over $\mathbb{F}^m \times \mathbb{F}^m$ (because y is uniform and $i \neq 0$), so the probability that the RS constraint fails on $(x+iy, z)$ is exactly ε. Similarly for each fixed $j \in \{1, \dots, D+1\}$, the pair $(x+jz, y)$ is uniform, so the failure probability is again ε. By a union bound over the $2(D+1)$ constraints,

$$\Pr_{y,z}[\text{all these constraints hold}] \;\geq\; 1 - 2(D+1)\varepsilon.$$

Therefore $\Pr_{y,z}[A_y(x) = A_z(x)] \geq 1 - 2(D+1)\varepsilon$, and plugging back into the previous inequality proves the claim. ∎

Claim 9.11 (Claim 2). *If $\varepsilon = O(1/D^2)$, then $g_f \in \mathcal{P}_{\leq D}$. In particular, for every $x, y \in \mathbb{F}^m$,*

$$\sum_{i=0}^{D+1} C_i \, g_f(x+iy) = 0.$$

Proof. Fix $x, y \in \mathbb{F}^m$, and choose $z \in \mathbb{F}^m$ uniformly at random. We will show that for at least one choice of z (hence, for *every* choice),

$$\sum_{i=0}^{D+1} C_i \, g_f(x+iy) = 0. \tag{9.3}$$

Step 1: express $g_f(x+iy)$ using a common direction z. By Claim 9.10, for each fixed point $u \in \mathbb{F}^m$ we have

$$\Pr_z\left[g_f(u) = -\sum_{j=1}^{D+1} C_j \, f(u+jz) \right] \;\geq\; 1 - 2(D+1)\varepsilon.$$

Apply this with $u = x+iy$ for every $i \in \{0, 1, \dots, D+1\}$ and take a union bound to get

$$\Pr_z\left[\forall i \in \{0, \dots, D+1\},\; g_f(x+iy) = -\sum_{j=1}^{D+1} C_j \, f(x+iy+jz) \right] \;\geq\; 1 - 2(D+2)(D+1)\varepsilon. \tag{9.4}$$

Step 2: enforce RS constraints on the shifted lines. For a fixed direction y, let

$$\varepsilon(y) \;:=\; \Pr_{u \in \mathbb{F}^m}[\text{RS rejects on } (u, y)].$$

(So $\mathbb{E}_y[\varepsilon(y)] = \varepsilon$; the handwritten notes suppress this dependence and write ε throughout.) For each $j \in \{1, \dots, D+1\}$, since $x + jz$ is uniform over $\mathbb{F}^m$ when z is uniform,

$$\Pr_z\left[\sum_{i=0}^{D+1} C_i \, f(x+jz+iy) \neq 0 \right] \;=\; \varepsilon(y).$$

By a union bound over $j \in \{1, \dots, D+1\}$,

$$\Pr_z\left[\forall j \in \{1, \dots, D+1\},\; \sum_{i=0}^{D+1} C_i \, f(x+jz+iy) = 0 \right] \;\geq\; 1 - (D+1)\varepsilon(y). \tag{9.5}$$

Step 3: choose a good z and compute. If $\varepsilon = O(1/D^2)$ (and in particular $\varepsilon(y) = O(1/D^2)$ for the fixed y in question), then the right-hand sides of (9.4) and (9.5) are both positive, hence there exists some z for which *both* events occur. Fix such a z.

Using (9.4) we can expand the left-hand side of (9.3):

$$\begin{aligned}\sum_{i=0}^{D+1} C_i\, g_f(x+iy) &= -\sum_{i=0}^{D+1} C_i \left(\sum_{j=1}^{D+1} C_j\, f(x+iy+jz) \right) \\ &= -\sum_{j=1}^{D+1} C_j \left(\sum_{i=0}^{D+1} C_i\, f(x+jz+iy) \right),\end{aligned}$$

where we swapped the order of summation in the last step. Now (9.5) says that each inner sum is 0, so the whole expression is 0. This proves (9.3).

Since x, y were arbitrary, g_f satisfies the RS constraint for *all* x, y. By the local-to-global implication from Section 5 (when $|\mathbb{F}| > 2D$), this implies $g_f \in \mathcal{P}_{\leq D}$. □

Step 2: finish. If $\varepsilon = O(1/D^2)$ then $g_f \in \mathcal{P}_{\leq D}$, and Claim 1 gives $\Delta(f, \mathcal{P}_{\leq D}) \leq \Delta(f, g_f) \leq 2\varepsilon$. If not, then necessarily $\varepsilon = \Omega(1/D^2)$, which is the other outcome in the theorem. □