The Polynomial Paradigm in Algorithm Design **Winter 2020**

Lecture 4: Strongly Rayleigh Distribution

Lecturer: Shayan Oveis Gharan *Jan 24th*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

In this lecture we discuss applications of real stable polynomials to probability theory.

Let $\mu : 2^{[n]} \to \mathbb{R}_{\geq 0}$ be a probability distribution. For $X \sim \mu$, the *generating polynomial* of μ is defined as follows:

$$g_\mu(z_1, \dots, z_n) = \mathbb{E}\left[z^X\right] = \sum_{S \subseteq [n]} \mathbb{P}\left[X = S\right] z^S.$$

For example, say B_1, B_2 are two independent Bernoullis with success probabilities p_1, p_2 respectively. Then, their generating polynomial is defined as follows:

$$p_1 p_2 z_1 z_2 + p_1(1-p_2)z_1 + p_2(1-p_1)z_2 + (1-p_1)(1-p_2) = (p_1 z_1 + (1-p_1))(p_2 z_2 + (1-p_2))$$

As a sanity check, observe that $g_\mu(\mathbf{1}) = 1$. We say μ is strongly Rayleigh (SR) if g_μ is a real stable polynomial.

It turns out that closure properties of real stable polynomials translate to closure properties of strongly Rayleigh distributions. Say μ is strongly Rayleigh. Then it remains so under the following operations:

Conditioning In $\mu | i$. This is nothing but $z_i \partial_{z_i} g_\mu$ (up to a normalizing constant).

Conditioning Out $\mu | \bar{i}$. This exactly $g_\mu|_{z_i = 0}$.

Projection. Given a set T, $\mu|_T$ is the distribution supported on subsets of T where for any $A \subseteq T$,

$$\mu|_T(A) = \sum_{S : S \cap T = A} \mu(S).$$

Observe that projection is exactly $g_\mu|_{z_i = 1, \forall i \notin T}$.

External Field. Given a non-negative vector $(\lambda_1, \dots, \lambda_n)$, we define $\mu * \lambda$ as the distribution where

$$\mu * \lambda(S) = \mu(S) \lambda^S.$$

Closure under external fields just follows from the closure of real stable polynomials under external fields, $g_\mu(\lambda_1 z_1, \dots, \lambda_n z_n)$.

Rank Sequence. The rank sequence of μ is the sequence $a_0, \dots, a_d$ where $a_i = \mathbb{P}_{S \sim \mu}\left[|S| = i\right]$. It follows that the rank sequence of any strongly Rayleigh distribution corresponds to a sum of independent Bernoullis. This is because $g_\mu(1, \dots, 1)$ is univariate real rooted polynomial.

In the next section we discuss several examples of Strongly Rayleigh distributions. An important example is the uniform spanning tree distribution: given a graph $G = (V, E)$, let μ be a uniform distribution over all spanning trees of G. Then, μ is strongly Rayleigh. As a consequence we prove the following lemma that we promised in the first lecture:

Lemma 4.1. *Given a set $F \subseteq E$, the univariate polynomial*

$$\sum_T t^{|F \cap T|}$$

is real rooted.

Proof. Let μ be uniform distribution over spanning trees; it is SR. Then, the projection μ_F is also SR. So, $p(z_1, \dots, z_{|F|}) = g_{\mu|F}$ is real stable. So, $p(t, \dots, t)$ is real rooted. But $p(t, \dots, t)$ is the same as $\sum_T t^{|F \cap T|}$ up to a normalizing constant. □

4.1 Determinantal Point Processes

One of the main classes of strongly Rayleigh Distributions are determinantal Point processes (DPPs). Given a PSD matrix $L \in \mathbb{R}^{n \times n}$, a.k.a., the ensemble matrix, for any $S \subseteq [n]$ we have

$$\mathbb{P}[S] \propto \det(L_{S,S})$$

Geometrically, given n vectors $v_1, \dots, v_n \in \mathbb{R}^d$, we think of the ensemble matrix as the Gram-matrix of these vectors. So, for any set $S \subseteq [n]$, say $S = \{i_1, \dots, i_k\}$, $\mathbb{P}[S]$ is proportional to the square of the k-th volume of the parallelepiped spanned by these k vectors.

Lemma 4.2. *Let $L \succeq 0$ be the ensemble matrix of a DPP. Then, the polynomial*

$$\sum_{S: S \subseteq [n]} \det(L_{S,S}) z^S$$

is real stable.

Proof. First, by mother-of-all

$$p(z_1, \dots, z_n) = \det(Z - L) = \sum_{k=0}^{n} (-1)^k \sum_{S \in \binom{n}{k}} z^{\overline{S}} \det(L_{S,S})$$

is real stable, where Z is the diagonal matrix with $Z_{i,i} = z_i$. Now, to prove the lemma we use the closure of real stable polynomials under inversion:

$$z_1 \dots z_n p(-1/z_1, \dots, -1/z_n) = (-1)^n \sum_{k=0}^{n} \sum_{S \in \binom{n}{k}} z^S \det(L_{S,S})$$

is real stable. □

Truncation Given a distribution μ and an integer $k \geq 1$, the *truncation* of μ is defined as the distribution μ_k where

$$\mu_k(S) \propto \begin{cases} \mu(S) & \text{if } |S| = k, \\ 0 & \text{otherwise} \end{cases}$$

Theorem 4.3. *For any strongly Rayleigh distribution μ and any $1 \leq k \leq n$, μ_k is strongly Rayleigh.*

To prove this theorem we use the following homogenization lemma on real stable polynomials.

Given a polynomial $p \in \mathbb{R}[z_1, \ldots, z_n]$ the *homogenized* version of p, p_H is defined as follows:

$$p_H(z_1, \ldots, z_n, z_{n+1}) = z_{n+1}^{\deg p} p(z_1/z_{n+1}, \ldots, z_n/z_{n+1})$$

For example, if $p = 1 - z_1 z_2$, $p_H = z_3^2 - z_1 z_2$.

Lemma 4.4. *For any real stable polynomial $p \in \mathbb{R}[z_1, \ldots, z_n]$ with non-negative coefficients, p_H is real stable.*

Note that p having non-negative coefficients is a necessary condition as $1 - z_1 z_2$ is real stable but the homogenized version $z_3^2 - z_1 z_2$ is not.

We do not prove this lemma here, we just use to prove that strongly Rayleigh distributions are closed under truncation. To prove Theorem 4.3 we use that $g_{\mu H}(z_1, \ldots, z_{n+1})$ is real stable. Say g_μ has degree d, then, we note that

$$g_{\mu_k} \propto \partial_{z_{n+1}}^{d-k} g_{\mu H} \mid_{z_{n+1}=0} .$$

Finally, the statement follows from the closure of real stable polynomials under differentiation and specialization. This completes the proof of Theorem 4.3.

For example, given n independent Bernoullis, $B_1, \ldots, B_n$ with corresponding SR distribution μ. It follows that μ_k is SR and inherits all properties of SR distributions even though μ_k is no longer a distribution over independent Bernoullis. In particular, we will see in the the next section that μ_k is negatively correlated.

Given a DPP μ, a k-DPP is μ_k. It is well-known that μ_k is no longer a DPP. Therefore, for years researchers had difficulties studying properties of μ_k. However, it follows that μ_k is also SR so we can use properties of SR distributions to study k-DPPs.

Random Spanning Trees A special case of DPPs are random spanning tree distributions. Given a graph $G = (V, E)$, choose an arbitrary direction for any edge; so for edge $e = (i, j)$ let $v_e = (\mathbf{1}_i - \mathbf{1}_j)$. Let $L \in \mathbb{R}^{E \times E}$ be the Gram-matrix of these vectors, i.e., for any two edges e, f, $L_{e,f} = \langle v_e, v_f \rangle$. Let μ be the corresponding determinantal point process (which is SR). We claim that μ_{n-1} is the uniform distribution on spanning trees; so random spanning trees are SR. To see this, it is enough to show that for any set $F \subseteq E$ of size $|F| = n-1$,

$$\det(L_{F,F}) = \begin{cases} n & \text{if } F \text{ is a tree,} \\ 0 & \text{otherwise.} \end{cases}$$

Let $F = \{e_1, \ldots, e_{n-1}\}$ and let

$$B = \begin{bmatrix} v_{e_1} \\ \vdots \\ v_{e_{n-1}} \end{bmatrix}$$

Then, $L_{F,F} = BB^T$. First, observe that if F has a cycle, say $e_1, \ldots, e_k$ in F form a cycle, then

$$(\pm) v_{e_1} (\pm) \ldots (\pm) v_{e_k} = 0,$$

where we choose the signs to negate the directions such that $e_1, \ldots, e_k$ form a directed cycle. It follows that the vectors in F are linearly dependent, so $\det(L_{F,F}) = 0$.

Otherwise, suppose F does not have a cycle. Then, F is a tree. We need to show $|\det(L_{F,F})| = n$. The main observation is that B is a totally unimodular matrix, i.e., every square submatrix of B has determinant $0/1/-1$. This simply because given any square submatrix of B say $B_{X,Y}$ where $X \subseteq F$ and $Y \subseteq V$; either there is a vertex of degree zero in subgraph (X, Y) in which case $\det(B_{X,Y}) = 0$, or there is a vertex of degree 1. In the latter case, say a vertex y has only an edge x, we re-order the rows and columns of $B_{X,Y}$ so that x is the last row and y is the last column. This implies that the last column of $B_{X,Y}$ is all zeros except the entry on the diagonal. Following this idea recursively we can make $B_{X,Y}$ upper-diagonal with $1, -1$ on the diagonal. That implies the claim.

Finally, we use the Cauchy-Binet identity to argue that

$$\det(L_{F,F}) = \sum_{S \in \binom{n}{n-1}} \det(B_{F,S})^2 = n.$$

In the above we use that when $|S| = n-1$ then $\det(B_{F,S})$ is non-zero; so it is either $-1/+1$. We leave this as an exercise.

Lemma 4.5 (Couchy-Binet Identity)**.** *Let* $v, \dots, v_n \in \mathbb{R}^d$, *then*

$$\det\left(\sum_{i=1}^{n} v_i v_i^T\right) = \sum_{S \in \binom{n}{d}} \det\left(\sum_{i \in S} v_i v_i^T\right).$$

Proof. Let

$$A = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$$

be the matrix with v_i's as the rows. Since A is a rank d matrix, by SVD we can write $A = \sum_{i=1}^{d} s_i a_i b_i^T$ where a_i's are orthonormal and b_i's are orthonormal. Then,

$$\sum_{i=1}^{n} v_i v_i^T = A^T A = \sum_{i=1}^{d} s_i b_i b_i^T$$

So,

$$\det(\sum_{i=1}^{n} v_i v_i^T) = \prod_{i=1}^{d} s_i.$$

On the other hand, the RHS is the sum of all $d \times d$ principal minors of $AA^T = \sum_{i=1}^{d} s_i a_i a_i^T$. But this is exactly the $n-d$-th coefficient of the characteristic polynomial of AA^T which is the d-th elementary symmetric polynomial of eigenvalues of AA^T, i.e.,

$$e_d(s_1, \dots, s_d, 0, \dots, 0) = s_1 \dots s_d$$

as desired. □

4.2 Negative Correlation

A probability distribution $\mu : 2^{[n]} \to \mathbb{R}_{\geq 0}$ is negatively correlated if for any i, j,

$$\mathbb{P}[i|j] \leq \mathbb{P}[i].$$

In this section we prove that any strongly Rayleigh distribution is negatively correlated.

Theorem 4.6. *A multilinear polynomial $p \in \mathbb{R}[z_1, \dots, z_n]$ is real stable iff for any i, j*

$$\partial_{z_i} p(x) \cdot \partial_{z_j} p(x) \geq p(x) \cdot \partial_{z_i} \partial_{z_j} p(x), \tag{4.1}$$

where $x \in \mathbb{R}^n$.

Before proving this theorem, let us discuss its implications.

Lemma 4.7. *Strongly Rayleigh distributions are pairwise negatively correlated, i.e., for any $1 \leq i < j \leq n$*

$$\mathbb{P}[i] \, \mathbb{P}[j] \geq \mathbb{P}[i, j].$$

Proof. To see this just plug in $x = \mathbf{1}$ in the above equation. Observe that $\partial_{z_i} g_\mu(\mathbf{1}) = \mathbb{P}[i]$. Similarly, $\partial_{z_i z_j} g_\mu(\mathbf{1}) = \mathbb{P}[i, j]$ and $g_\mu(\mathbf{1}) = 1$. □

The fact that (4.1) holds for all $x > 0$ means that μ and all external fields of μ are negatively correlated. An immediate consequence is that k-DPPs are negatively correlated. This was unknown to researchers in machine learning and they were expecting that truncation to k implies certain positive correlations.

A probability distribution is called *Rayleigh* if it satisfies (4.1) for all $x \geq 0$. The notion of Rayleigh distributions was first introduced by Wagner [Wag06] with the purpose of generalizing the Rayleigh monotonicity law for effective resistances in graphs.

Proof. $\Rightarrow$: Assume p is real stable. First, note that for any $a > 0$ and $b \in \mathbb{R}$, by closure properties of real stable polynomials, $p(at + b, z_2, \dots, z_n)$ is real stable. Then the bivariate restriction

$$g(s, t) = p(x_1, \dots, x_{i-1}, s + x_i, \dots, x_{j-1}, t + x_j, x_{j+1}, \dots, x_n)$$

is real stable. Since p is multilinear, g is multilinear and it can be writen as

$$g(s, t) = a + bs + ct + dst,$$

where $a = p(x)$, $b = \partial_{z_i} p(x)$, $c = \partial_{z_j} p(x)$ and $d = \partial_{z_i} \partial_{z_j} p(x)$. Now, in HW1 we will see that this implies $bc \geq ad$. Therefore, (4.1)

$\Leftarrow$: Conversely, assume that (4.1) holds for all i, j. We prove by induction; suppose

$$p(z_1, \dots, z_{n+1}) = q(z_1, \dots, z_n) + z_{n+1} r(z_1, \dots, z_n).$$

By IH, for any $\alpha \in \mathbb{R}$, the polynomial $p(z_1, \dots, z_n, \alpha)$ is real stable or is identically zero. Similarly we have this for q, r. First, suppose that for some $\alpha \in \mathbb{R}$, $q + \alpha r$ is identically zero. Then, we can write

$$p = (z_{n+1} - \alpha) r(z_1, \dots, z_n)$$

and thus real stable and we are done. Otherwise, we assume $p(z_1, \dots, z_n, \alpha)$ is real stable and non-zero for any $\alpha \in \mathbb{R}$. This implies that for any $\alpha \in R$, and $(z_1, \dots, z_n) \in \mathcal{H}^n$, $p(z_1, \dots, z_n, \alpha) \neq 0$. Writing, $p = r(q/r + z_{n+1})$ and using that q/r is continuous, it follows that either

$$\Im(q/r) > 0 \quad \forall (z_1, \dots, z_n) \in \mathcal{H}^n, \text{ or}$$
$$\Im(q/r) < 0 \quad \forall (z_1, \dots, z_n) \in \mathcal{H}^n.$$

In the former case we get that $p(z_1, \dots, z_{n+1})$ is real stable and we are done. In the latter case we get that $\tilde{p}(z_1, \dots, z_{n+1}) = p(z_1, \dots, z_n, -z_{n+1})$ is real stable. Using the forward direction of the theorem this implies the reverse of (4.1), for $i, n+1$ we get

$$\partial_{z_i} \tilde{p}(x) \cdot \partial_{z_{n+1}} \tilde{p}(x) \geq \tilde{p}(x) \partial_{z_i} \partial_{z_{n+1}} \tilde{p}(x)$$

for any $x \in \mathbb{R}^{n+1}$. This implies

$$\partial_{z_i} p(x_1, \ldots, x_n, -x_{n+1}) \cdot -\partial_{z_{n+1}} p(x_1, \ldots, -x_{n+1}) \geq p(x_1, \ldots, x_{n+1}) \cdot -\partial_{z_i} \partial_{z_{n+1}} p(x_1, \ldots, -x_{n+1})$$

so

$$\partial_{z_i} p(x) \cdot \partial_{z_{n+1}} p(x) \leq p(x) \cdot \partial_{z_i} \partial_{z_{n+1}} p(x)$$

for any $x \in \mathbb{R}^{n+1}$. Since this the inverse of (4.1) indeed we have equality above. This implies that for any $1 \leq i \leq n$, $r\partial_{z_i} q = q\partial_{z_i} r$. But the latter implies that q is a multiple of r and therefore p is real stable. □

SR distributions satisfy several stronger notions of negative dependence such as *Negative Association*, *Stochastic Dominance*, and Concentration of Lipschitz functions.

Theorem 4.8. *Let $\mu : 2^{[n]} \to \mathbb{R}_{\geq 0}$ be a k-homogeneous SR distribution and $f : 2^{[n]} \to \mathbb{R}$ be a 1-Lipschitz function. Then, for any $a \geq 0$,*

$$\mathbb{P}\left[|f - \mathbb{E}\left[f\right]| > a\right] \leq \exp(-a^2/8k).$$

Here f is 1-Lipschitz if for any two sets A, B such that $|A \Delta B| = 1$ we have $|f(A) - f(B)| \leq 1$. As an immediate consequence, we can use above to show that the number of even degree vertices in a random spanning tree is tightly concentrated around its expectation.

Note that if μ is not homogeneous, then we can homogenize it first by means of Lemma 4.4 and then replace k in the above with the degree of g_μ.

References

[Wag06] D. Wagner. "NEGATIVELY CORRELATED RANDOM VARIABLES AND MASON'S CONJECTURE". 2006. URL: https://arxiv.org/pdf/math/0602648.pdf (cit. on p. 4-5).