II. Perron-Frobenius theory for $\mathbb{N} \times \mathbb{N}$ nonnegative matrices

Mike Boyle

University of Maryland

1. Introduction

The Perron-Frobenius theory for $\mathbb{N}\times\mathbb{N}$ nonnegative matrices comes from Vere-Jones (1962, 1967).

In this lecture, A is an $\mathbb{N} \times \mathbb{N}$ or $n \times n$ matrix with nonnegative real entries. As in the finite ($n \times n$) case, the analysis goes by cases:
(1) A is irreducible with period 1
("irreducible and aperiodic")
(2) A is irreducible with period p
(3) general A.

When A is finite ($n \times n$),
(1) holds iff A is primitive.
For A infinite ($\mathbb{N}\times\mathbb{N}$), it is too restrictive (and it is unnecessary) to assume some power of A is positive. (1) holds if for each (i, j), $A^n(i, j)$ is positive for all large n.

The reductions from general to irreducible, and from irreducible to aperiodic, are similar to the finite case. We concentrate on case (1). Until further notice, A is assumed to be irreducible aperiodic $\mathbb{N} \times \mathbb{N}$ or $n \times n$.

In the finite case, the Perron Theorem holds for ALL aperiodic irreducible A.

In the infinite case, the aperiodic irreducible matrices split into four classes. The split reflects how conclusions and corollaries of the Perron Theorem for finite matrices break down.

The goal of this lecture is to come to some understanding of the statements and some key ideas in the infinite matrix case. It would take another lecture or two to go through a proof. Anyway, the proof is intricate enough that it is better done in private.

2. The Perron Theorem for finite matrices

PERRON THEOREM
For A an $n \times n$ primitive matrix,
there is a number $\lambda > 0$ such that
• A has a nonnegative (hence positive) eigenvector with eigenvalue λ
• λ is a simple root of χ_A (the characteristic polynomial of A) and $\lambda > |\alpha|$ for every other root α of χ_A.

The critical (and easy) first step of the proof was to get the nonnegative eigenvector. That step used the action of A as an operator on $\mathbb{R}^n$. But we have no suitable space V on which all the $\mathbb{N} \times \mathbb{N}$ irreducible aperiodic matrices act in a way allowing an analogous argument.

So Vere-Jones gave a very different kind of argument.

3. Matrices as labeled graphs

For visual support and language: associate to A a labeled graph G_A ("graph" means "directed graph").

EXAMPLE

$$A = \begin{pmatrix} 3 & \pi \\ .43 & 0 \end{pmatrix}$$

G_A has vertices 1 and 2,
with
an edge from 1 to 1 labeled 3;
an edge from 1 to 2 labeled π;
an edge from 2 to 1 labeled .43 .

A path p of length n in G_A from i to j is a finite sequence of n head-to-tail edges,
$p = (i_0 i_1)(i_1 i_2) \cdots (i_{n-1} i_n)$,
such that $i_0 = i$ and $i_n = j$.

p is a *loop* if $i_0 = i_n$.
p is a *first return loop* if no vertex i_j with $0 < j < n$ equals i_0. For example:
The length 2 path (11)(11) is a loop but not a first return loop.
An n-*path*(n-*loop*) is a path(loop) of length n.

We restrict from here to A such that all entries of all A^n are finite.
The *weight* of p is $A(i_0, i_1)A(i_1, i_2) \cdots A(i_{n-1}, i_n)$.
The weight of an edge from i to j is $A(i, j)$.
$A^n(i, j)$ is the sum of the weights of the length n paths from i to j.

4. The two series $f(z), t(z)$

We pick a base index in $\mathbb{N}$ – the choice is unimportant; for definiteness we choose 1 – and with it define two formal power series critical for the theory. Define for $n \in \mathbb{N}$:
$t_n = A^n(1,1)$, the sum of the weights of the n-loops from 1 to 1
$f_n =$ the sum of the weights of the first return n-loops from 1 to 1 .

Easily checked:

- $t_{m+n} \geq t_m t_n$ for all m, n
- therefore $\sup(t_n)^{1/n} = \lim(t_n)^{1/n} := \lambda$
- λ doesn't depend on the choice of base index

Define the Perron value λ_A of A to be this λ (possibly ∞). Then Rad(t), the radius of convergence of the series t, is $1/\lambda$.

EXAMPLE:

For positive numbers a, b, c, d and $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$,

$f(z) = az + (bc)z^2 + (bdc)z^3 + (bd^2c)z^4 + \cdots$
$f(z) = az + bc \sum_{n=0}^{\infty} dz^{k+2}$
$t(z) = az + (a^2 + bc)z^2 + \cdots$

Easily checked:

• If A is finite, then λ_A is the usual spectral radius.

• λ_A equals the sup of the λ_B such that B is a finite principal submatrix of A.

An essential functional relation

There is an essential relation:

$$\frac{1}{1-f(z)} \;=\; 1+t(z)\ .$$

This is an equation of formal power series: by definition, $\frac{1}{1-f(z)}$ is the series

$$1+f(z)+[f(z)]^2+[f(z)]^3+\cdots\ .$$

In this infinite sum of series, for each n only finitely many terms contribute to the coefficient of z^n. Therefore the infinite sum is well defined as a formal power series. As the right hand side is a series with positive coefficients, for a positive real α greater than than Rad(t) it makes sense to say $\frac{1}{1-f(z)} = \infty$ when $z=\alpha$.

RADII OF CONVERGENCE:
Note, Rad$(f) \le$ Rad(t).

EXAMPLE: $A = (2)$
G_A is a single vertex with a self loop labeled 2.
$1 - f(z) = 1 - 2z$
$1 + t(z) = 1 + 2z + 2^2z^2 + 2^3z^3 + \cdots$
$\frac{1}{1-f(z)} = \frac{1}{1-2z} = 1 + t(z)$.

PROOF OF THE RELATION:
Let $g(z) = \sum g_n z^n = (1 - f(z))\,(1 + t(z))$.
We must show $g(z) = 1$. We have

$$g_n = t_n - f_1 t_{n-1} - f_2 t_{n-2} \cdots - f_{n-1} t_1$$

so we must show
$t_n = f_1 t_{n-1} + \cdots + f_{n-1} t_1$. Well,
$t_n =$ sum of weight of n-loops from 1 to 1
$f_n =$ sum of weight of simple n-loops from 1

An n-loop from 1 which is not simple must be (for some $k < n$) a simple k-loop from 1 followed by an $(n-k)$-loop. For a given k, the weights of these loops total to $f_k t_{n-k}$. QED

Dividing irreducible period 1 into classes

Let $t(1/\lambda)$ denote $\sum_n t_n(1/\lambda)^n$.
DEFN A is *transient* (T) if $t(\lambda) :=< \infty$.
DEFN A is *recurrent* (R) if $t(\lambda) = \infty$.

For PF theory, the transient matrices are the bad ones. Almost everything goes wrong.

We divide the recurrent matrices into 3 classes.
Below, $f'(z)$ is the formal power series
$f'(z) = \sum_{n=1}^{\infty} n f_n z^{n-1}$.

DEFN A is *null recurrent* (NR) if it is recurrent and $f'(1/\lambda) = \infty$.

DEFN A is *positive recurrent* (PR) if it is recurrent and $f'(1/\lambda) < \infty$.

The great Vere-Jones work (1967) gave the theory of T, NR, PR.
Let us also single out another class (however it's named).

DEFN A is *exponentially recurrent* (ER) if Rad$(f) <$ Rad(t) (i.e., $\limsup (f_n)^{1/n} < \lambda$) .

(Caveat: terminology variations are discussed later.)

We have $R \supset PR \supset ER$. The PF properties will improve along this progression. ER is the class very close to finite primitive matrices for PF.

Recurrent matrices

Here is a list of properties which as conclusions or corollaries of the Perron Theorem hold for every finite primitive matrix, and which also hold for every recurrent matrix A.

1. There are nonnegative vectors ℓ, r such that $\ell A = \lambda \ell$ and $Ar = \lambda r$.
(Then ℓ and r must be positive.)
2. The nonnegative eigenvectors for λ are unique up to scalar multiples.
3. $u \geq 0$ and $uA \leq \lambda u \implies uA = \lambda u$.

Transient matrices

None of the previous properties are guaranteed for transient matrices. A transient matrix might have no nonnegative right eigenvector for λ; linearly independent right eigenvectors; a positive vector v with $Av = \beta v$ and $\beta > \lambda$. Likewise (obviously), for left eigenvectors.

Systematically, for A transient:
There is always a positive vector v such that $Av \leq \lambda v$ and $Av \neq \lambda v$.

Let's look at a transient example.
For numbers $\epsilon_2, \epsilon 3, \ldots$ define $M(1,2) = 1$
$M(k, k+1) = 1 - \epsilon_k$ if $k > 1$
$M(k, 1) = \epsilon_k$ if $k > 1$
$M = 0$ otherwise.

If every ϵ_k is zero, then $M^n \to 0$ (entrywise). If the ϵ_k are chosen to decrease sufficiently rapidly then we can easily force $\lambda_M < 1$. But M will be stochastic and we can also solve for a positive left eigenvector. This matrix M must be transient, by our definition.

Positive recurrent matrices

Below (and elsewhere), convergence means entrywise convergence.

Suppose ℓ, r are positive left, right eigenvectors for A. By abuse of notation, let $\ell \cdot r$ denote $\sum_n \ell_n r_n$.

If A is positive recurrent, then $\ell \cdot r < \infty$. Then we can choose them so that $\ell \cdot r = 1$. In this case, as for finite A,

$$((1/\lambda)A)^n \longrightarrow r\ell \ .$$

In contrast, if A is null recurrent or transient then $((1/\lambda)A)^n \longrightarrow 0$.

Exponentially recurrent matrices

This is the class which enjoys essentially all good properties Perron-Frobenius. In addition to the PR properties:

1. The convergence

$$((1/\lambda)A)^n \longrightarrow r\ell \ .$$

is entrywise exponentially fast. This means that for each (i,j) there are constants $C > 0$ and $0 < \kappa < 1$ such that for all n

$$|((1/\lambda)A)^n(i,j) - r(i)\ell(j)| < C\kappa^n \ .$$

(The convergence need not be uniform over all entries: given the choice of a base vertex, the same κ can be chosen for all entries (i,j) (any number from the interval $(\text{Rad}(t)/\text{Rad}(\text{f}), 1)$), but in general C depends on (i,j).)

That number κ, reflecting the exponential rate, in general cannot be chosen independent of the choice of base vertex.

2. If $0 \leq B \leq A$ and $B \neq A$, then $\lambda_B < \lambda_A$.

3. For a finite matrix A with nonzero spectrum $(\lambda_1, \ldots, \lambda_k)$, the function $\det(I - zA) = (1 - \lambda_1 z) \cdots (1 - \lambda_k z)$. So, the poles of the function $\zeta_A : z \mapsto 1/\det(I - zA)$ are the reciprocals of the numbers $\lambda_1, \ldots, \lambda_k$. One way of expressing the spectral gap provided by the Perron Theorem is to say that the function

$$z \mapsto \frac{(1 - \lambda z)}{\det(I - zA)}$$

(where $\lambda = \lambda_A$) has radius of convergence strictly greater than the radius of convergence of ζ_A. For A in ER, an analogous fact is that the radius of convergence of $1/\det(I - f(z))$ is $1/\lambda$ and the radius of convergence of
$(1 - \lambda z)/\det(I - f(z))$ is strictly larger.
(This becomes more meaningful when ζ_A can be interpreted in terms of periodic points.)

All three of these properties must fail if A is not ER.

More about those classes.

Here is a table of possibilities for the four classes

	$f(1/\lambda)$	$t(1/\lambda)$	$f'(1/\lambda)$
T	< 1	finite	$\leq \infty$
NR	1	∞	∞
PR, not NR	1	∞	$< \infty$
ER	1	∞	$< \infty$

To justify the table (and know ER $\subset$ PR) we need the definitions and three facts:
(1) $0 < f(1/\lambda) \leq 1$.
(2) $f(1/\lambda) = 1$ iff $t(1/\lambda) = \infty$.
(3) ER $\subset$ PR .

Proofs (easy) for these facts are in an appendix.

Loop graphs
Loop graphs are a useful tool, e.g. to write examples.

Given $0 \neq f = \sum_{n=1}^{\infty}$ with $f_n \geq 0$ for all n, define the loop graph $G_{(f)}$ as the labeled graph which is the union of the first return loops from the base vertex 1, such that
• there is one n-loop for each n such that $f_n > 0$, and the loop is labeled such that its weight is f_n
(to be definite: let the first edge of the loop be labeled f_n and label other edges 1)
• each vertex not equal to 1 has a single incoming and a single outgoing edge.

If (after naming vertices as integers) A is the adjacency matrix of $G_{(f)}$, then the function f_A is the given f. A is irreducible with period equal to gcd$\{n : f_n \neq 0\}$.
Abuse of notation: write λ_f for λ_A.

EXAMPLE: $f(z) = 3z + z^2$. Then $A = \begin{pmatrix} 3 & 1 \\ 1 & 0 \end{pmatrix}$.

Examples of the classes.

It will often be more transparent to consider the case $\lambda = 1$. There is an easy reduction which generally lets one assume $\lambda = 1$ WLOG. Below, A is a matrix such that $G_A = G_{(f)}$.

EXAMPLE.

$f(z) = \sum c/n^2 z^n$, with c a positive constant. Here $\lambda = 1/\lambda = 1 = \limsup (f_n)^{1/n}$.

Whether $f(1) = 1$ depends on c.

- If $f(1) < 1$, then A is transient.
- If $f(1) = 1$, then A is null recurrent, because $f'(1/\lambda) = \sum c/n = \infty$.

EXAMPLE.
$f(z) = \sum(c/n^3)z^n$, with c a positive constant such that $f(1) = 1$.
Then $\lambda = 1 = 1/\lambda$, so $f(1/\lambda) = 1$ and
$f'(1/\lambda) = \sum c/n^2 < \infty$.

Therefore A is PR.
Because $\limsup(c/n^3)^{1/n} = 1 = 1/\lambda$,
the matrix A is not ER.

EXAMPLE

$f(z) = z$

$A = (1)$

$1 + t(z) = 1 + z + z^2 + z^3 + \cdots$

$\lambda = 1$

$\limsup (f_n)^{1/n} = 0 < 1/\lambda$.

This A is ER.

“Robustness” of the classes.

Here are a few remarks to suggest how prominent the different classes may be among the $\mathbb{N} \times \mathbb{N}$ irreducible aperiodic matrices.

First of all, we restrict to matrices A such that A^n is well defined as a matrix with real entries, for all n. That exludes “most” matrices. Then we also exclude the remaining A for which $\lambda_A = +\infty$. In the small remaining set where we live:

1. If A is ER, then there is $\epsilon > 0$ such that any matrix B which has the same zero entries and satisfies $(1-\epsilon)A < B < (1+\epsilon)A$ must also be ER.
(One can define a separable metric on the aperiodic irredicble matrices with a given sign pattern, in which T and ER are each open sets and their union is dense.)

2. If A and B are recurrent with $B \leq A$ and $B \neq A$, then $\lambda_B < \lambda_A$.
(Otherwise, $f_A(1/\lambda_A) > f_B(1/\lambda_A) = f_B(1/\lambda_B) = 1$, which is impossible.)
3. Suppose A is not ER. Then there are many B such that $\lambda_B = \lambda_A$ and $A \neq B \leq A$.
For example, suppose $0 < c < 1$ and C is a matrix formed by multiplying finitely many rows and finitely many columns of A by c. Then any B such that $C \leq B \leq A$ and $B \neq A$ is transient.

There are other remarks about the size of sets. Suppose we use the box topology (the neighborhoods of a matrix A are the sets of the form $V_\epsilon = \{B : |B_(i,j)_A(i,j)| < \epsilon((i,j))$, where ϵ is a function from $\mathbb{N} \times \mathbb{N}$ to $(0, +\infty)$. (This topology is neither separable nor metrizable.)
In this topology, T with finite λ and ER are open subsets of the $\mathbb{N} \times \mathbb{N}$ matrices, and the complement in the set of A with finite λ of the union of their interiors has empty interior. In this sense, $R \backslash ER$ is the small and unstable set.
(I haven't thought this through completely; there should be sharper remarks.)

Stochasticization.

Suppose that A is a nonnegative matrix (finite or countable) and r is a strictly positive right eigenvector, $A = \beta r$ (so, $\beta > 0$).
Define a diagonal matrix R with $R(i,i) = r(i)$.
Define $P = R^{-1}(1/\beta)AR$.
Let v be the column vector with every entry 1.

By direct computation, $Pv = v$.
Therefore P is stochastic (nonnegative with every row sum 1). Let us say a matrix obtained from A by a positive diagonal similarity and a scalar multiple is a *stochasticization* of A. If A has only one positive eigenvector up to scalar multiples (for example, whenever A is recurrent), then there is only one stochasticization P of A.

If A is irreducible aperiodic and recurrent, then it has the unique stochasticization. If A is transient we can find R such that $R^{-1}(1/\lambda)AR$ is less than or equal to a stochastic matrix and is not equal to it.

Probabilistic heuristic.

Suppose we consider irreducible aperiodic matrices A with Perron value 1. (This a natural reduction. For example, multiplication of a matrix by a positive scalar does not change membership in any of the classes T, NR, PR, ER.) Now think of entries of A as being like transition probabilities. (This heuristic can be made precise for recurrent matrices by stochasticizing them. Note, for a diagonal matrix R with positive diagonal, A and $R^{-1}AR$ have exactly the same series f and t.)

Then $t(1)$ can be thought of as the expected number of returns to the base state (1, for us), and “recurrent” and “transient” match the standard probabilistic terminology. Similarly $f'(1)$ corresponds to expected return time in the recurrent case: infinite for null recurrent and finite for positive recurrent.

However, for transient matrices this doesn't work without the assumed normalization to $\lambda = 1$. A transient matrix A can be stochastic but have $\lambda < 1$ and $t(1) = \infty$.

Notation and references.

I follow Kitchens in using the terms transient, null recurrent and positive recurrent. Vere-Jones used in their place R-transient, R-null recurrent and R-positive recurrent, where R is the radius of convergence of t (this R is our $1/\lambda$.) The repetition of R emphasizes the distinction in the transient case.

"Exponentially recurrent" is a term picked for this lecture. It seems self-explanatory with regard to the meaning of the class. Gurevich and Savchenko called these positive recurrent matrices stable. What in the terms of this paper would be an ER stochastic matrix was called geometrically ergodic by Vere-Jones (1962).

One basic reference for this topic is the exposition in Chapter 7 of Kitchens' book Symbolic Dynamics, which has further references and discussion of them (and which I found very helpful as a reference). A standard reference for many years has been Seneta's book.

Defining the Perron eigenvectors.
Given the countable irreducible and aperiodic matrix A, and our choice of base vertex 1, for each state j and $n \in \mathbb{N}$ define $\ell_{1j}(n)$ to be the weight of the n-paths $(i_0)(i_1 i_2) \cdots (i_{s-1} i_s)$ from 1 to j such that $i_s \neq 1$ if $0 < s < n$.

Define the power series $L_{1j}(z) = \sum_n \ell_{1j}(n)(z)$.

Similarly, define $r_{1i}(n)$ to be the weight of the n-paths $(i_0)(i_1 i_2) \cdots (i_{s-1} i_s)$ from i to 1 such that $i_s \neq 1$ if $0 < s < n$.

Define the power series $R_{j1}(z) = \sum_n r_{1j}(n)(z)$.

Now we can define the vectors ℓ, r which for recurrent A will be the eigenvectors for the Perron value λ. Set $\ell(j) = L_{1j}(1/\lambda)$ and $r(i) = R_{i1}(1/\lambda)$.

Note $L_{11}(z) = R_{11}(z) = f(z)$, so $\ell(1) = r(1)$ is at most 1. Also for any j, the coefficient of z^n in the series $L_{1j}(z)L_{jj}(z)R_{j1}(z)$ is the sum of weights of n loops from 1 to 1 which pass through j. Because $L_jj(1) \leq 1$, we have $L_{1j}(1) \leq R_{j1}(1) \leq f(1) \leq 1$.

Proving the convergence in the recurrent case involves more recursion relations and estimates.

Appendix: Some proofs about $f(1/\lambda), f'(1/\lambda)$.
(1) $0 < f(1/\lambda) \leq 1$.
PROOF
$0 < f(1/\lambda)$ because $f_n \geq 0$ for all n and $f \neq 0$
.

To see $f(1/\lambda \leq 1$, suppose $0 < \beta < \infty$ and $1 < f(\beta) \leq \infty$. It suffices to show $\beta > 1/\lambda$.

Pick α such that $0 < \alpha < \beta$ and $f(\alpha) > 1$. Then

$$1 + t(\alpha) = 1 + f(\alpha) + [f(\alpha)]^2 + \cdots = \infty$$

and therefore $\alpha \geq$ Rad$(t) = 1/\lambda$, and $\beta > 1/\lambda$.

(2) $f(1/\lambda) = 1$ iff $t(1/\lambda) = \infty$. PROOF This holds because $0 < f(1/\lambda) \leq 1$ and $1 + t(z) = 1 + f(z) + [f(z)]^2 + \cdots$.

(3) ER $\Longrightarrow$ PR .

PROOF

By definition of ER, the function f defined from the matrix A satisfies

$$\limsup (f_n)^{1/n} = \text{Rad}(f) > 1/\lambda \ .$$

First we prove A is recurrent. If $f(1/\lambda < 1$, then (because $\text{Rad}(f) > 1/\lambda$) we may take $\alpha > 1/\lambda$ such that $f(\alpha) < 1$. Then $1+t(\alpha) < \infty$, so $\text{Rad}(t) \geq \alpha > 1/\lambda$, a contradiction. Therefore $f(1/lambda) = 1$.

Next we show $f'(1/\lambda) < \infty$. This holds because $\text{Rad}(f') = \text{Rad}(f) > 1/\lambda$.

Appendix. Finite approximation for eigenvectors. Exercise: Prove the following regularity result.

Suppose A is $\mathbb{N}\times\mathbb{N}$ recurrent and $(\mathcal{S}_k)$ is an increasing sequence of subsets of $\mathbb{N}$ whose union is $\mathbb{N}$ and such that for each k the principal submatrix of A on index set $\mathcal{S}_k$ is primitive.
Let A_k be the $\mathbb{N}\times\mathbb{N}$ matrix such that $A_k(i,j) = A(i,j)$ if $\{i,j\}\subset\mathcal{S}_k$ and $A_k = 0$ otherwise.
Let r_k be a nonnegative right eigenvector of A_k, $A_k r_k = \lambda_k r_k$, with the sequence (r_k) normalized such that for some element I of $\mathcal{S}_1$, $r_k(I) = 1$ for all k.

Then $\lim \lambda_k = \lambda$ and $\lim r_k = r$ such that $Ar = \lambda r$.