A new method to bound rate of convergence

Arup Bose
Indian Statistical Institute, Kolkata,
abose@isical.ac.in

Sourav Chatterjee
University of California, Berkeley
Outline

• Empirical Spectral Distribution   Limiting Spectral Distribution (LSD)

• Different Types of Convergence

• Rates of Convergence

• A “new” technique of bounding the rates of convergence

• Application to the matrices: Wigner, Sample Covariance, Reverse Circulant, Toeplitz, Hankel
$A_n$ is an $n \times n$ Hermitian matrix with random entries.

Eigenvalues (characteristic roots): $\lambda_0, \lambda_1, \ldots, \lambda_{n-1}$.

The empirical spectral distribution (ESD) function of $A_n$:

$$F_n(x) = n^{-1} \sum_{i=0}^{n-1} I\{\lambda_i \leq x\}.$$ 

The empirical spectral measure: the corresponding probability measure $P_n$

The expected spectral distribution function: $E(F_n(\cdot))$.

Expected spectral measure: The corresponding probability measure
The **Limiting Spectral Distribution** (or measure) (LSD): weak limit of \( \{ P_n \} \), in some probabilistic sense, such as “almost surely” or “in probability”.

The weak convergence of \( E(F_n) \) often serves as an intermediate step in showing the weak convergence of \( F_n \).

LSD known for Wigner, Sample covariance, Toeplitz, Hankel, Reverse Circulant...

Two main tools to establish the LSD: the moment method and the Stieltjes transform method.
Our goal

When the LSD exists, how does one obtain the rate of convergence?

So far the main method to establish such rates is the use of Stieltjes transform.

Our goal:

1. Establish some general results useful in establishing the probabilistic weak convergence of $F_n$ from the convergence of $E(F_n)$ and the corresponding rates of convergence.

2. Apply these to establish some new rates of convergence.
Empirical characteristic function

\( \varphi = \text{Empirical characteristic function of } A: \) the characteristic function of the empirical spectral distribution of \( A \).

\[ \varphi(t) = \frac{1}{n} \text{tr}(e^{itA}) \]

where \( n = \) the order of \( A \) and \( e^M = \sum_{k=0}^{\infty} \frac{1}{k!} M^k \).

In general, for any distribution function \( G \) (which may be random) on \( \mathbb{R} \), its (random) characteristic function is defined as \( \varphi_G(t) = \int e^{itx} dG(x) \).
Outline of our approach

**STEP 1.** Obtain bounds for \( \text{Var}(\varphi(t)) \) and its partial derivatives.

(For a complex random variable \( X \), its variance is defined to be \( E|X - E(X)|^2 \).)

**STEP 2.** Deduce the concentration of the spectral measure near its mean, and also get the magnitude of concentration using the bounds.
Step 2: Bound on expected K-S distance

Kolmogorov distance:

\[ \| F - G \|_\infty = \sup_{x \in \mathbb{R}} |F(x) - G(x)|. \]

**Theorem 1.** Suppose \( F \) is a random distribution function on \( \mathbb{R} \) with (random) characteristic function \( \varphi \). Suppose \( \text{Var}(\varphi(t)) \leq C t^2 \) for each \( t \). If \( G \) is a nonrandom distribution function on \( \mathbb{R} \), such that \( \sup_{x \in \mathbb{R}} |G'(x)| \leq \lambda \), then

\[
E\| F - G \|_\infty \leq 2\| E(F) - G \|_\infty + \frac{8(3)^{1/2}\lambda^{1/2}}{\pi} C^{1/4}.
\]

This will be eventually used to establish rates of convergence for the ESD.

\( C \) determines the rate.
Proof: Notation

The convolution $F \ast G$ of $F$ and $G$ is defined in the usual way. That is, $F \ast G(x) = \int F(x - y)dG(y) = \int G(x - y)dF(y)$.

Define the probability density

$$h_L(x) = \frac{1 - \cos Lx}{\pi Lx^2}.$$  

Let $H_L$ be the corresponding distribution function.

The characteristic function of $H_L$ is given by

$$\psi_L(t) = (1 - \frac{|t|}{L})I(|t| \leq L).$$

Note that

$$\int_{-\infty}^{\infty} |\psi_L(t)|dt = L.$$
Let $F_0 = E(F)$, and let $\eta$ be the characteristic function of $F_0$.

Then by assumption,
$$E|\varphi(t) - \eta(t)| \leq \sqrt{C}|t|.$$  

By Esseen’s lemma,
$$\|F - G\|_\infty \leq 2\|F \ast H_L - G \ast H_L\|_\infty + \frac{24\lambda}{\pi L}$$

Now
$$\|F \ast H_L - G \ast H_L\|_\infty \leq \|F_0 \ast H_L - G \ast H_L\|_\infty + \|F \ast H_L - F_0 \ast H_L\|_\infty$$

This in turn is bounded by
$$\|F_0 - G\|_\infty + \|F \ast H_L - F_0 \ast H_L\|_\infty.$$
Proof continued

So by applying the inversion formula (see Feller 1966, page 482-484) and the hypothesis about Var(ϕ(t)),

\[
E \| F * H_L - F_0 * H_L \|_\infty \leq \frac{1}{\pi} \int_{-\infty}^{\infty} |\psi(t)| \frac{E|\varphi(t) - \eta(t)|}{|t|} dt
\]
\[
\leq \frac{C^{1/2} L}{\pi}.
\]

Combining all these observations, we have

\[
E \| F - G \|_\infty \leq 2 \| F_0 - G \|_\infty + \frac{2\sqrt{C} L}{\pi} + \frac{24\lambda}{\pi L}
\]

Choosing \( L^2 = 12\lambda C^{-1/2} \) gives the desired conclusion. \( \square \)
Bound on variance: Efron-Stein inequality

Note the condition on the variance in Theorem 1.

Note that the matrix $A$ is a function of independent real random variables $x_1, x_2, \ldots, x_m$, (note that $m$ is large).

- Write $\varphi(t)$ as a function of $x_1, x_2, \ldots, x_m$.


**Theorem 2. (Efron-Stein type inequality).** Suppose $Z_1, \ldots, Z_n, Z_1^*, \ldots, Z_n^*$ are independent $m$-dimensional random vectors where $Z_i$ has the same distribution as $Z_i^*$ for all $i$. Suppose that $f : \mathbb{R}^{m \times n} \to \mathbb{R}$ is a square integrable function. Then

$$\text{Var}(f(Z_1, \ldots, Z_n)) \leq \frac{1}{2} \sum_{k=1}^{n} E|f(Z_1, \ldots, Z_n) - f(Z_1, \ldots, Z_{k-1}, Z_k^*, Z_{k+1}, \ldots, Z_n)|^2.$$
**Theorem 3.** Suppose \( f(x_1, x_2, \ldots, x_n) \) is a coordinatewise differentiable map from \( \mathbb{R}^n \) into \( \mathbb{C} \) and \( M_j, 1 \leq j \leq n \) are constants such that

\[
\sup_x \left| \frac{\partial f}{\partial x_j} \right| \leq M_j \quad \text{for } j = 1, 2, \ldots, n.
\]

Then for independent complex random variables \( X_1, X_2, \ldots, X_n \),

\[
\text{Var}(f(X_1, X_2, \ldots, X_n)) \leq \sum_{j=1}^n M_j^2 \text{Var}(X_j).
\]
There exists a constant $K > 0$ such that if $X \sim F$, and $g : \mathbb{R} \to \mathbb{R}$ is any differentiable map, then

$$\text{Var}(g(X)) \leq KE|g'(X)|^2.$$ 

Log-concave densities (i.e. densities of the form $e^{U(x)}$ where $U$ is a concave function) satisfy **POIN**:


Double exponential, and uniform distributions

See also Chen (1982).
The complete characterization of all absolutely continuous distributions which satisfy POIN is given in the following theorem, proved in a more general form in Muckenhoupt (1972).

**Theorem 4.** A distribution function \( F \) with \( F(0) = 1/2 \) and density \( f \) w.r.t. Lebesgue measure satisfies POIN if and only if

\[
B = \max \{ \sup_{x \geq 0} (1 - F(x)) \int_0^x \frac{1}{f(t)} \, dt, \sup_{x < 0} F(x) \int_x^0 \frac{1}{f(t)} \, dt \} < \infty.
\]

Moreover, the smallest constant \( K \) admissible in POIN satisfies \( \frac{1}{2} B \leq K \leq 4B \).
The next theorem follows easily from the Efron-Stein inequality:

**Theorem 5.** If \( X_1, X_2, \ldots, X_n \) \( i.i.d \sim F \) where \( F \) has property POIN with constant \( K \), then for any coordinatewise differentiable map \( f : \mathbb{R}^n \rightarrow \mathbb{C} \),

\[
\text{Var}(f(X_1, \ldots, X_n)) \leq K \sigma^2 E|\nabla f(X_1, \ldots, X_n)|^2
\]

where \( |\cdot| \) denotes the Euclidean norm, and \( \sigma^2 = \text{Var}(X_1) \).
Step 1: Bound on $\text{Var}(\varphi(t))$

- Use the identity
  \[
  \frac{\partial \varphi(t)}{\partial x_j} = \frac{1}{n} \text{tr} \left( it \frac{\partial A}{\partial x_j} e^{itA} \right)
  \]

- Use the fact that $e^{itA}$ is a unitary matrix.

- Either the partial derivatives are bounded by small numbers in sup norm

Then use an Efron-Stein type inequality.

- Or the expected value of the norm-squared of $\nabla \varphi(t)$ is small.

Then use a Poincaré type inequality.
Suppose $A(u)$ is a matrix with complex entries. Derivative of product' formula:

$$
\frac{d}{du} (A(u)B(u)) = A'(u)B(u) + A(u)B'(u).
$$

**Lemma 1.** If $A(u)$ is an elementwise differentiable map from $\mathbb{R}$ or $\mathbb{C}$ into $\mathbb{C}^{n \times n}$ then

$$
\frac{d}{du} \text{tr}(e^A) = \text{tr} \left( \frac{dA}{du} e^A \right).
$$

**Lemma 2.** If $A$ is Hermitian and $t$ is real, then $e^{itA}$ is a unitary matrix. In particular, for any vector $x$, $|e^{itA}x| = |x|$, (where $| \cdot |$ denotes the Euclidean norm) and also all entries of $e^{itA}$ have modulus $\leq 1$. 
$W_n$ is a Wigner matrix with independent entries $(X_{jk}^{(n)})$ having common variance 1. We shall drop the superscript $n$.

In many situations, the LSD of $n^{-1/2}W_n$ exists and is given by

$$F(x) = \frac{1}{2\pi} \int_{-\infty}^{x} (4 - y^2)^{1/2} I_{[-2, 2]} dy.$$ 

Note that $\sup_{-2 \leq x \leq 2} F'(x) = \pi^{-1}$.

Let $F_n$ be the ESD of $n^{-1/2}W_n$. 
Wigner–rates of Bai et. al.

(W1) \( E(X_{jk}) = 0, \ E(X_{jk}^2) = 1. \)

(W2) \( \sup_{i,j,n} \ E X_{ij}^4 < \infty. \)

(W3) \( \sum \ E \left( X_{ij}^4 I\{|X_{ij}| \geq \epsilon n^{1/2} \} \right) = o(n^2) \) for any \( \epsilon > 0. \)


\[ \|F_n - F\|_\infty = O_p(n^{-2/5}). \]

(W3*) \( \sum \ E(X_{ij}^8 I\{|X_{ij}| \geq \epsilon n^{1/2} \}) = o(n^2) \) for any \( \epsilon > 0. \)

Then \( \|E(F_n) - F\|_\infty = O(n^{-1/2}). \)

(W3**) \( \sup_n \sup_{i,j} E|X_{ij}|^k < \infty \) for every \( k \geq 1. \)

Then \( \|F_n - F\|_\infty = O(n^{-2/5+\eta}) \) almost surely for every \( \eta > 0. \)
$W_n : \mathbb{R}^{n(n+1)/2} \rightarrow \mathbb{R}^{n \times n}$

Takes a tuple $(a_{jk})$ to the Wigner matrix with $(j, k)$-th entry $n^{-1/2}a_{jk}$ ($j \geq k$), and $n^{-1/2}a_{kj}$ otherwise.

Then $W_n^{jk} = \frac{\partial W_n}{\partial a_{jk}}$ is a constant matrix whose only nonzero entries are $(j, k)$-th and $(k, j)$-th entries ($= n^{-1/2}$).

$\varphi_t^n = \text{the empirical characteristic function of } W_n$. Then

$$\frac{\partial \varphi_t^n}{\partial a_{jk}} = n^{-1} \text{tr} \left( it \frac{\partial W_n}{\partial a_{jk}} e^{itW_n} \right) = n^{-1} \text{tr} \left( itW_n^{jk} e^{itW_n} \right).$$

Let $B = e^{itW_n}$. Note that $B$ is unitary. Then

$$\frac{\partial \varphi_t^n}{\partial a_{jk}} = \frac{it(b_{jk} + b_{kj})}{n\sqrt{n}} = \frac{2itb_{jk}}{n\sqrt{n}} \quad \text{and} \quad |\nabla \varphi_t^n|^2 = \sum_{j \geq k} \left| \frac{\partial \varphi_t^n}{\partial a_{jk}} \right|^2 \leq \sum_{j,k} \frac{4t^2|b_{jk}|^2}{n^3} = \frac{4t^2}{n^2}.$$
Wigner–rate assuming POIN

Stronger conditions—new and stronger results.

Note that \( \sup_{-2 \leq x \leq 2} F'(x) = \pi^{-1} \).

**Theorem 6.** If \( W_n \) is a random real Wigner matrix whose entries on and above the diagonal are i.i.d. with variance 1, drawn from a distribution satisfying POIN with constant \( K \), then \( \text{Var}(\varphi_n(t)) \leq 4Kt^2/n^2 \). Consequently, by Theorem [1], if \( F \) denotes the semicircular law, then

\[
E\|F_n - F\| \leq 2\|EF_n - F\| + \frac{8(6)^{1/2}K^{1/4}}{\pi^{3/2}}n^{-1/2}
\]

where \( F_n \) denotes the empirical c.d.f. of \( n^{-1/2}W_n \).
Wigner–rate without POIN

Minimal conditions–weaker rate results.

Lemma 2: the elements of $e^{itW_n}$ are bounded in modulus by 1. Hence

$$\left\| \frac{\partial \varphi_n^t}{\partial a_{jk}} \right\|_{\infty} \leq 2 |t| n^{-3/2}.$$ 

So we can invoke Theorems 3 and 1:

**Theorem 7.** If $\{W_n\}$ is a sequence of random Wigner matrices, whose entries on and above the diagonal are independent with variance uniformly bounded by 1, then

$$\text{Var}(\varphi_n(t)) \leq \frac{4t^2}{n^3} \sum_{j \geq k} \text{Var}(a_{jk}) \leq \frac{4t^2}{n}.$$ 

Hence if $F_n$ denotes the empirical c.d.f. of $n^{-1/2}W_n$ and $F$ denotes the semicircular law, then

$$E\|F_n - F\| \leq 2\|EF_n - F\| + \frac{8(6)^{1/2}K^{1/4}}{\pi^{3/2}} n^{-1/4}.$$
Suppose $X$ is a real $p \times n$ matrix with entries $x_{jk}$, which are i.i.d. real random variables with mean zero and unit variance. Let $S = \frac{1}{n} XX^T$. If $y_n = p/n \to y \in (0, 1]$ then the ESD of $S_n$ converges almost surely to the law $F_y(\cdot)$ with the Marčenko-Pastur density

$$f_y(x) = \begin{cases} \frac{1}{2\pi y} \sqrt{(b-x)(x-a)} & \text{if } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

where $a = a(y) = (1 - \sqrt{y})^2$ and $b = b(y) = (1 + \sqrt{y})^2$. It can be easily shown that the density is bounded by $\lambda = \left[\pi \sqrt{y(1-y)}\right]^{-1}$.

In cases where $y > 1$, the LSD exists but has a point mass at the origin. If $y = 0$, then a scaling and a centering are required for the LSD of $S_n$ to exist. See Bai (1999) or Bose et al. (2003) for the precise results. We do not consider these cases.
Stieltjes transform method: Bai (1993b) proved that $E(F_n)$ converges to $F_y$ at a rate $O(n^{-1/4})$ or $O(n^{-5/48})$ depending on how close $y_n$ is to 1.

The same rates were obtained for convergence in probability of $F_n$ to $F_y$ in Bai, Miao and Tsay (1997).

Bai, Miao and Yao (2003) prove several results under the conditions given in Wigner case: If $y_n$ remains bounded away from 1 and suitable combinations of the above conditions hold then

$$
\|E(F_n) - F_y\|_\infty = O(n^{-1/2}), \|F_n - F_y\|_\infty = O_P(n^{-2/5}) \text{ and } \|F_n - F_y\|_\infty = O_{a.s.}(n^{-2/5+\eta}).
$$
Let \( Y = \frac{\partial X}{\partial x_{jk}} = e_j e_k^T \)

\[ x_k = k \text{th column of } X \]

\( \varphi_n^t = \text{empirical characteristic function evaluated at } t. \)

\[ S_{jk} := \frac{\partial S}{\partial x_{jk}} = \frac{1}{n} (YX^T + XY^T) \]

\[ \frac{\partial \varphi_n^t}{\partial x_{jk}} = p^{-1} \text{tr}(itS_{jk}e^{itS}) \]

\[ = it(np)^{-1} \text{tr}(YX^T e^{itS} + XY^T e^{itS}) \]

\[ = it(np)^{-1} \text{tr}(e_j x_k e^{itS} + x_k e_j^T e^{itS}) \]

\[ = it(np)^{-1} (x_k e^{itS} e_j + e_j^T e^{itS} x_k) = \frac{2itz_{kj}}{np} \]

where we have written \( z_{kj} \) for the \( j \)th component of the vector \( z_k := e^{itS} x_k \). Note that since \( e^{itS} \) is unitary, \( \|z_k\| = \|x_k\| \).
\( S \) matrix– rate with POIN

\[
\sum_{j,k} |\frac{\partial \varphi_n^t}{\partial x_{jk}}|^2 \leq \frac{4t^2}{n^2p^2} \sum_{k=1}^{n} \|z_k\|^2 = \frac{4t^2}{n^2p^2} \sum_{k=1}^{n} \|x.k\|^2
\]

\[
= \frac{4t^2}{n^2p^2} \sum_{j,k} |x_{jk}|^2 \text{ a.s.}.
\]

So, if \( \forall j, k, E|x_{jk}|^2 \leq M^2 \), then \( E|\nabla \varphi_n^t|^2 \leq \frac{4M^2t^2}{np} \).

Applying Theorems 5 and 1,

**Theorem 8.** If \( \{x_{jk}\} \) are i.i.d. from a distribution satisfying \text{POIN} with constant \( K \), mean 0 and variance 1, then \( \text{Var}(\varphi_n(t)) \leq \frac{4Kt^2}{np} \).

Consequently, if \( y_n = p/n \in (0,1) \) and \( F_y \) denotes the Marčenko-Pastur distribution with parameter \( y \), then

\[
E\|F_{n,p} - F_{y_n}\| \leq 2\|EF_{n,p} - F_{y_n}\| + \frac{8(6)^{1/2}K^{1/4}}{\pi^{3/2}[y_n(1-y_n)]^{1/2}} n^{-1/2}
\]

where \( F_{n,p} \) denotes the ESD of \( S \).
If we don’t assume **POIN** but instead impose $\forall j, k, |x_{jk}| \leq M$ a.s., then

$$|\frac{\partial \varphi^t_n}{\partial x_{jk}}| \leq 2|t|(np)^{-1}\|z_k\| = 2|t|(np)^{-1}\|x_k\| \leq \frac{2M|t|}{n\sqrt{p}} \text{ a.s.}$$

Thus, if the variance of $x_{jk}$ is bounded by 1, then

$$\text{Var}(\varphi_n(t)) \leq \frac{4M^2t^2}{n^2p}np = \frac{4M^2t^2}{n}.$$}

Hence we get

**Theorem 9.** Suppose $y = p/n \in (0, 1)$ and $x_{jk}$ are independent with mean zero and variance bounded by 1. Suppose $M$ is such that $P(|x_{jk}| \leq M) = 1$. Then

$$E\|F_n - F_y\| \leq 2\|EF_n - F_y\| + \frac{8(6)^{1/2}M^{1/2}}{\pi^{3/2}y^{1/4}(1-y)^{1/2}}n^{-1/4}.$$ 

where $F_n$ is the ESD of $S_n$. 

**Outline**

- Statistical
- Unity in diversity
- Spectral Distribution/Measure
- Limiting Spectral Distribution/Measure
- Our goal
- Empirical characteristic function
- Outline of our approach
- Step 1: Bound on expected K-S distance
- Step 2: Bound on variance: Efron-Stein inequality
- Consequence
- "Poincaré Condition"
- **POIN**: Absolutely continuous
- **POIN + Efron-Stein**
- Sample covariance matrix LSD
- Sample covariance matrix – derivatives
- Sample covariance matrix – with **POIN**
- $S$ matrix – without **POIN**
- Reverse circulant
- Hankel
- Toeplitz
- Remarks

Random matrices - p. 28/33
Reverse circulant: LSD

**Reverse Circulant matrix** $A_n = \left( (x_{(i+j-2) \mod n}) \right)$. Visually,

$$A_n = \begin{bmatrix}
  x_0 & x_1 & x_2 & \cdots & x_{n-1} \\
  x_1 & x_2 & \cdots & x_{n-1} & x_0 \\
  x_2 & \cdots & x_{n-1} & x_0 & x_1 \\
  & \vdots & & & \\
  x_{n-1} & x_0 & x_1 & \cdots & x_{n-2}
\end{bmatrix}$$

Bose and Mitra (2002), Bose, Chatterjee and Gangopadhyay (2003):

If $\{x_i\}$ are i.i.d. with mean zero and variance 1 then ESD of $X_n = n^{-1/2} A_n$ converges (in probability) to the LSD with density $f$ given by

$$f(x) = |x| \exp(-x^2), \quad -\infty < x < \infty.$$
Let $B = e^{itA_n}$. Then $\sum \sum |b_{ij}|^2 = n$. Also for each $k$, there are at most $2n$ pairs of $(i, j)$ such that $i + j - 2 = k \mod n$.

Let $\varphi^t_n = \text{empirical characteristic function of } A_n \text{ evaluated at } t$.

\[
\frac{\partial \varphi^t_n}{\partial x_k} = \frac{it}{n\sqrt{n}} \sum_{i+j-2=k \mod n} b_{ji}
\]

\[
\sum_k \left| \frac{\partial \varphi^t_n}{\partial x_k} \right|^2 \leq \frac{t^2}{n^3} \sum_k \left[ 2n \sum |b_{ji}|^2 \right] = \frac{2t^2}{n}
\]

If $x_k$ are i.i.d. from a density satisfying POIN, then

\[
E\|F_n - F\| \leq 2\|EF_n - F\| + O(n^{-1/4}).
\]

The eigenvalues can be explicitly obtained and using them, $\|EF_n - F\|$ is of a much smaller order than $n^{-1/4}$. 
$H_n = ((x_{i+j-2}))$ is called a Hankel matrix.

Bryc Jiang and Dembo: LSD of $n^{-1/2}H_n$. Not much known about the LSD.

**IF the limit has a bounded density**, then we can use our technique. The computations for our method are very similar to the previous example. In fact, here

$$\frac{\partial \varphi^t_n}{\partial x_k} = \frac{it}{n\sqrt{n}} \sum_{i+j-2=k} b_{ji}$$

So, similar computations show that under **POIN**

$$E\|F_n - F\| \leq 2\|EF_n - F\| + O(n^{-1/4}).$$
$T_n = \left( (x|i-j|) \right)$ is called a Toeplitz matrix of order $n$.

Bryc Dembo and Jiang: LSD exists and is unimodal.

IF the limiting distribution has a bounded density:

Exactly the same kind of computations as in the preceding examples show that in this case, too,

Under POIN,

$$E\|F_n - F\| \leq 2\|EF_n - F\| + O(n^{-1/4}).$$
Do the Toeplitz and the Hankel limits have bounded densities?

How optimal are the rates?

Sharpen the method to obtain better results?

Develop other characteristic function based methods?

Higher order estimates?