LOCAL LAW AND TRACY–WIDOM LIMIT FOR SPARSE SAMPLE COVARIANCE MATRICES

BY JONG YUN HWANG[*,1], JI OON LEE[*,1] AND KEVIN SCHNELLI[†,2]

KAIST * *and KTH Royal Institute of Technology*[†]

We consider spectral properties of sparse sample covariance matrices, which includes biadjacency matrices of the bipartite Erdős–Rényi graph model. We prove a local law for the eigenvalue density up to the upper spectral edge. Under a suitable condition on the sparsity, we also prove that the limiting distribution of the rescaled, shifted extremal eigenvalues is given by the GOE Tracy–Widom law with an explicit formula on the deterministic shift of the spectral edge. For the biadjacency matrix of an Erdős–Rényi graph with two vertex sets of comparable sizes M and N, this establishes Tracy–Widom fluctuations of the second largest eigenvalue when the connection probability p is much larger than $N^{-2/3}$ with a deterministic shift of order $(Np)^{-1}$.

1. Introduction. Sample covariance matrices form one fundamental class of random matrices. They are of great importance in high-dimensional data and multivariate statistics. Spectral properties of sample covariance matrices are particularly interesting as they are crucial in various procedures in real data analysis such as principal component analysis (PCA).

It is well known that the limiting spectral distribution of a sample covariance matrix is governed by the Marchenko–Pastur law when the sample size is comparable to the dimension [28]. In this case, the fluctuations of the largest eigenvalue of the sample covariance matrix follow the Tracy–Widom laws. This was first proved for the complex Wishart ensemble by Johansson [19] and for the real Wishart ensemble by Johnstone [20]. For sample covariance matrices with non-Gaussian entries but identity population covariance, the Tracy–Widom limit for the largest eigenvalue was established by Pillai and Yin [31], and a necessary and sufficient condition on the entries' distribution for the limit to hold was obtained by Ding and Yang [9]. In the general nonnull case, where the general population matrix is not a multiple of the identity matrix, the Tracy–Widom limit was identified in [10, 24] and the edge universality was proved in [3, 21].

The Tracy–Widom laws are also the answer to the question about the fluctuations of the largest eigenvalues of various random matrix models, including the adjacency matrix of random graphs. The simplest case is the Erdős–Rényi graph, where any pair of vertices are connected with probability p, independently from other pairs. For the adjacency matrices of the Erdős–Rényi graph, while the largest eigenvalue is macroscopically separated from the bulk of the spectrum, the second largest eigenvalue exhibits Tracy–Widom fluctuations as long as the connection probability p is independent of the size N. We remark that the fluctuations of the second largest eigenvalue encode key properties of the random graph in many applications, including hypothesis testing for community detection in stochastic block model [5, 26].

The study of the behavior of the second largest eigenvalue becomes significantly harder when $p \equiv p(N)$ scales as N varies. If $p \to 0$ as $N \to \infty$, we say that the graph (or the corresponding adjacency matrix) is sparse. The analysis for the sparse case was only recently done by Erdős, Knowles, Yau and Yin [11, 12] and was further extended in [24] and [17]. The main idea in a deeper analysis of the local eigenvalue density, especially for sparse matrices, is the local semicircle law. It is the further local refinement of Wigner's semicircle law that governs the global behavior of the eigenvalue density of Wigner matrices.

If the underlying random graph is directed, we lose the symmetry of the adjacency matrix, and we are led to consider the singular values of the matrices instead of the eigenvalues. Related to such a model is a bipartite random graph, where the vertices can be decomposed into two groups within which vertices are not connected to each other. In these cases, the adjacency matrices can naturally be identified with sample covariance matrices, and when the connection probability p tends to zero as the size of matrix grows, we are led to consider sparse sample covariance matrices.

Another motivation for our research on the largest eigenvalues of sample covariance matrices stems from PCA in statistics and signal processing. In practice, one of the most significant challenges in using PCA is to determine the number of components, that is, the rank of the signal matrix in a noisy matrix. For the null hypothesis testing, which tests whether a signal is present, the largest eigenvalues can be exploited for the test statistics as discussed in [4, 29]. The idea can be further extended to the estimation of the true rank by a sequential application of the strategy in [22] or by a method based on the conditional singular value test [8]. We refer to [3] for more applications of the largest eigenvalues in high-dimensional statistical inference.

In this paper, we study the spectral properties of sparse sample covariance matrices, including the local law for $p \gg N^{-1}$ and the behavior of the largest eigenvalues. We prove that the limiting distribution of the shifted, rescaled largest eigenvalue is given by GOE Tracy–Widom distribution if $p \gg N^{-2/3}$, which was also assumed in [25]. The shift is deterministic, and it can be precisely computed in

terms of N and p. As in [25], while the shift for the case $p \gg N^{-1/3}$ is negligible, it is larger than the Tracy–Widom fluctuations if $p \ll N^{-1/3}$.

Our analysis is based on a refined local Marchenko–Pastur law that provides estimates for the local eigenvalue density down to the optimal scale at the upper edge of the spectrum. Following the strategy in [25], we introduce a polynomial P that describes a deterministic correction to Marcenko–Pastur law. These corrections provide the crucial information about the precise location of the largest eigenvalue. Adapting the approach in [25], we obtain the polynomial by using resolvent expansions instead of the Schur complement formula, but due to the structure of the sample covariance matrix, we employ a linearization trick before applying resolvent expansion methods. Technically, we control high moments of $|P|$ by a recursive moment estimate, where the bound utilizes the lower moments of $|P|$. After establishing the local Marchenko–Pastur law, we use the Green function comparison method to prove the Tracy–Widom limit of the largest eigenvalue. In the Green function comparison, instead of applying the Lindberg swapping trick with matching moments, we use a continuous flow that interpolates the given sparse sample covariance matrix and a "nicely-behaving" sample covariance matrix, which is a Wishart ensemble. The continuous interpolation has been successfully used in the context of edge universality in various random matrix models; we refer to Section 2.4 for further details. Another technical difficulty stems from the lack of symmetry in Marchenko–Pastur law. Compared to the Wigner case in [25], the polynomial P has a more complicated form, and it makes the analysis for the limiting distribution more involved; see Section 3 for technical details.

The sparse sample covariance matrices we consider in this work is of the form $X^\dagger X$, which means that the population covariance matrix is the identity. A natural extension of the model is of the form $X^\dagger \Sigma X$, where Σ is a general positive semidefinite matrix. We believe that our method can be adapted to the general model to prove the improved local law at the edge and also the Tracy–Widom limit of the largest eigenvalue. It will be an interesting result as the shift of the largest eigenvalue of $X^\dagger \Sigma X$ may depend on the eigenvalues of Σ. However, the proof will be more involved for the general model, especially for nondiagonal Σ, and we will hence focus only on the null case of the form $X^\dagger X$.

For sparse matrices of Wigner-type, the result in [25] was extended by Huang, Landon and Yau [17] to cover the regime $N^{-7/9} \ll p \leq N^{-2/3}$. In that case they proved that the shifted, rescaled largest eigenvalue exhibits a phase transition from Tracy–Widom to Gaussian fluctuations at $p \sim N^{-2/3}$. We expect an analogous transition to occur for sparse sample covariance matrices, but we do not pursue this direction in the current article.

This paper is organized as follows: In Section 2, we define the model and introduce the main results with the outline of the strategy for the proof. In Section 3, we prove some properties of the deterministic refinement of the Marchenko–Pastur law. In Section 4, we prove the local law based on the key technical result on the recursive moment estimates. In Section 5, we prove the Tracy–Widom limit of the

largest eigenvalue via the Green function comparison method. Some technical results are adaptions from [25] and are therefore postponed to the Supplementary Material [18].

REMARK 1.1 (Notational remark 1). We use the symbols $O(\cdot)$ and $o(\cdot)$ for the standard big-O and little-o notation. The notation O, o, $\ll$, $\gg$ refer to the limit $N, M \to \infty$ with $N/M = d$ fixed unless otherwise stated. The notation $a \ll b$ means $a = o(b)$. We use c and C to denote positive constants that do not depend on N, usually with the convention $c \leq C$. Their values may change from line to line. We write $a \sim b$ if there is $C \geq 1$ such that $C^{-1}|b| \leq |a| \leq C|b|$.

2. The model and main results.

2.1. *Motivating examples.* Before presenting our model in detail and stating our main results, we outline a few motivating examples for the present work.

2.1.1. *Signal detection with missing values.* Consider an M-dimensional signal-plus-noise vector

$$(2.1) \qquad \mathbf{y} := A\mathbf{s} + \mathbf{z},$$

where $\mathbf{s}$ is an N-dimensional signal vector, A is an $M \times N$ deterministic matrix, and $\mathbf{z}$ is an M-dimensional random vector. To understand the case where the coordinates z_i $(1 \leq i \leq M)$ are mostly zero, hence $\mathbf{z}$ is sparse, we let

$$(2.2) \qquad z_i = B_i V_i,$$

where (B_i) and (V_i) are independent families of independent, identically distributed (i.i.d.) centered random variables. The random variables (V_i) represent the noise and satisfy $\mathbb{E}V_i^2 = 1$ and $\mathbb{E}V_i^{2k} \leq C_k$ for some constants (C_k). The random variables (B_i) are of Bernoulli type with parameter p, that is,

$$(2.3) \qquad \mathbb{P}\left(B_i = \frac{1}{\sqrt{Np}}\right) = p, \qquad \mathbb{P}(B_i = 0) = 1 - p.$$

The sparsity of the noise, which is determined by the probability p, may originate from the existence of a certain error threshold or the lack of data due to missing observations.

The first step in the analysis is to determine whether there is any signal present. Under the null hypothesis, the sample covariance matrix associated with $\mathbf{y}$ is nothing more than the one associated with $\mathbf{z}$. The limiting distribution of its largest eigenvalue is assumed to be determined by Tracy–Widom–Airy statistics if $p \sim 1$. One can also use the statistics of Onatski [30] given by $R := (\mu_1 - \mu_2)/(\mu_2 - \mu_3)$, where μ_1, μ_2, μ_3 denote the first, second, and the third largest eigenvalue, respectively. It was shown [30] in the complex setting that R is asymptotically pivotal under the null hypothesis.

We introduce the sparsity parameter q through

$$(2.4) \qquad\qquad p = \frac{q^2}{N}$$

with $0 < q = \sqrt{Np} \leq N^{1/2}$, where we allow q to depend on N. When $q = N^{1/2}$, we regain a usual sample covariance matrix.

The concept of the sparsity is crucial in the analysis for small N. For example, while the case with probability $p = 0.3$ would not be classified as a sparse case for large N, with $N = 30$ the sparsity parameter $q = 9 = N^\phi$ for $\phi \approx 0.323$, which falls into the regime we are interested in.

2.1.2. *Simple Markov switching model.* The simplest Markov switching model involves a random vector

$$(2.5) \qquad\qquad \mathbf{y} := \boldsymbol{\mu_s} + \sigma_s \mathbf{z},$$

where $\mathbf{s}$ denotes the unobserved state indicator; see for example, *e*quation (1) in [34]. If the entries of $\mathbf{s}$ are i.i.d. Bernoulli random variables, it is called the random two-regime model; see, for example, [32]. Under the null hypothesis

$$(2.6) \qquad\qquad \mathbf{H}_0 : \boldsymbol{\mu_s} = \mathbf{0},$$

this model reduces to the one in Section 2.1.1, if we simply write

$$(2.7) \qquad\qquad \sigma_s z_i = \sigma B_i V_i,$$

where B_i is a Bernoulli random variable defined in (2.3). To generalize it further, one can choose a Markov chain with states $\{0, 1\}$ instead of a Bernoulli variable. Then, one can apply the results on the largest eigenvalues of sparse sample covariance matrices by analyzing the 2×2 transition matrix.

2.1.3. *Bipartite stochastic block model.* Consider a graph with two vertex sets V_1 and V_2 of size M and N, respectively. Each V_i is further divided into two or more communities. In the simple case of two communities in each vertex set, each V_i ($i = 1, 2$) is with partition (P_i, Q_i), and edges are added independently at random between V_1 and V_2 with probabilities depending on which parts the vertices are in; edges between P_1 and P_2 or Q_1 and Q_2 are added with probability p, while the other edges are added with probability p'. This model was proposed by Feldman, Perkins, and Vempala [14] as a generalization of the classic stochastic block model to unify graph partitioning and planted constraint satisfaction problems (CSP's) into one problem. The spectral analysis on the model was studied in [15].

In the spectral analysis for the graph partitioning, the $M \times N$ biadjacency matrix X can be considered, which is defined as follows:

$$(2.8) \qquad\qquad X_{\alpha i} = \begin{cases} 1 & \text{if } v_\alpha \in V_1 \text{ and } w_i \in V_2 \text{ are connected,} \\ 0 & \text{otherwise.} \end{cases}$$

Then, the singular vector of X corresponding to the second largest singular value, or equivalently, the second eigenvector $XX^\dagger$ is correlated with the partition of V_1. Such an algorithm requires that the second largest eigenvalue is well-separated from the spectral norm of the noise matrix. If the probability p and p' tend to 0 as $M, N \to \infty$, the graph is sparse and the noise matrix becomes a sparse sample covariance matrix. The null model in this case is the biadjacency matrix of the bipartite Erdős–Rényi graph where $p = p'$.

2.2. *Definitions and notation.*   We consider the following type of sample covariance matrices:

DEFINITION 2.1 (Sample covariance matrices).   Let $X = (X_{\alpha i})$ be a real $M \times N$ matrix with independent entries satisfying the moment conditions

$$(2.9) \qquad\qquad \mathbb{E}X_{\alpha i} = 0, \qquad \mathbb{E}(X_{\alpha i})^2 = \frac{1}{N}.$$

The sample covariance matrix associated with X is given by $S = X^\dagger X$. Furthermore, $M \equiv M(N)$ with

$$(2.10) \qquad\qquad d_N = \frac{N}{M} \to d \in [1, \infty),$$

as $N \to \infty$. For simplicity, we assume that d_N is constant, hence we use d instead of d_N.

The assumption $M \equiv M(N)$ with $N/M \to d$ is reasonable in high-dimensional settings. We remark that our results also hold for the case $d < 1$ after taking $(M - N)$ zero eigenvalues into consideration.

Next, we introduce some basic definitions and set notation.

DEFINITION 2.2 (High probability event).   We say that an N-dependent event $\Xi \equiv \Xi^{(N)}$ holds with high probability if, for any large $D > 0$,

$$(2.11) \qquad\qquad \mathbb{P}(\Xi^{(N)}) \geq 1 - N^{-D},$$

for sufficiently large $N \geq N_0(D)$.

DEFINITION 2.3 (Stochastic domination).   Let $Y_1 \equiv Y_1^{(N)}$, $Y_2 \equiv Y_2^{(N)}$ be N-dependent nonnegative random variables. We say that Y_1 stochastically dominates Y_2 if, for all small $\epsilon > 0$ and large $D > 0$,

$$(2.12) \qquad\qquad \mathbb{P}(Y_1^{(N)} > N^\epsilon Y_2^{(N)}) \leq N^{-D},$$

for sufficiently large $N \geq N_0(\epsilon, D)$, and we write $Y_1 \prec Y_2$.

DEFINITION 2.4 (Stieltjes transform). Given a probability measure ν on $\mathbb{R}$, its Stieltjes transform is the analytic function $m_\nu : \mathbb{C}^+ \to \mathbb{C}^+$, with $\mathbb{C}^+ := \{z = E + i\eta : E \in \mathbb{R}, \eta > 0\}$, defined by

$$(2.13) \qquad m_\nu(z) := \int_\mathbb{R} \frac{d\nu(x)}{x - z} \quad (z \in \mathbb{C}^+).$$

Note that $\lim_{\eta \to \infty} i\eta m_\nu(i\eta) = -1$ since ν is a probability measure. Conversely, if an analytic function $m : \mathbb{C}^+ \to \mathbb{C}^+$ satisfies $\lim_{\eta \to \infty} i\eta m(i\eta) = -1$, then it is the Stieltjes transform of a probability measure; see, for example, [1].

Choosing ν to be the Marchenko–Pastur law with density

$$\frac{1}{2\pi x} \sqrt{(\lambda_+ - x)(x - \lambda_-)}$$

on $[\lambda_-, \lambda_+]$, where

$$(2.14) \qquad \lambda_+ = \left(1 + \frac{1}{\sqrt{d}}\right)^2, \qquad \lambda_- = \left(1 - \frac{1}{\sqrt{d}}\right)^2,$$

one can shows that m_ν, denoted by m_{MP}, is explicitly given by

$$(2.15) \qquad m_{\mathrm{MP}}(z) = \frac{-(z + 1 - \frac{1}{d}) + \sqrt{(z + 1 - \frac{1}{d})^2 - 4z}}{2z},$$

where we choose the branch of the square root such that $m_{\mathrm{MP}}(z) \in \mathbb{C}^+$, $z \in \mathbb{C}^+$. It directly follows that

$$(2.16) \qquad 1 + \left(z + 1 - \frac{1}{d}\right) m_{\mathrm{MP}}(z) + z m_{\mathrm{MP}}(z)^2 = 0 \quad (z \in \mathbb{C}^+).$$

DEFINITION 2.5 (Green function and its normalized trace). Given a real symmetric matrix H we define its Green function by

$$(2.17) \qquad G^H(z) := \frac{1}{H - zI} \quad (z \in \mathbb{C}^+),$$

and the normalized trace of G^H by

$$(2.18) \qquad m^H(z) := \frac{1}{N} \operatorname{Tr} G^H(z) \quad (z \in \mathbb{C}^+).$$

The matrix entries of $G^H(z)$ are denoted by $G_{ij}^H(z)$. In the following we often drop the explicit z-dependence from the notation for $G^H(z)$ and $m^H(z)$.

2.3. *Main results*. The sparse sample covariance matrices we consider satisfies the following assumption.

ASSUMPTION 2.6. Fix any small $\phi > 0$. We assume that $X = (X_{\alpha i})$ is a real $M \times N$ matrix whose entries are independent, identically distributed (i.i.d.) random variables. We also assume that $(X_{\alpha i})$ satisfy the moment conditions

$$(2.19) \quad \mathbb{E}X_{\alpha i} = 0, \qquad \mathbb{E}(X_{\alpha i})^2 = \frac{1}{N}, \qquad \mathbb{E}|X_{\alpha i}|^k \leq \frac{(Ck)^{ck}}{Nq^{k-2}} \quad (k \geq 3),$$

with sparsity parameter q satisfying

$$(2.20) \qquad N^\phi \leq q \leq N^{1/2}.$$

Note that the real Wishart ensemble corresponds to the case $\phi = 1/2$. We denote by $\kappa^{(k)}$ the kth cumulant of the i.i.d. random variables $(X_{\alpha i})$. Under Assumption 2.6 we have $\kappa^{(1)} = 0$, $\kappa^{(2)} = 1/N$,

$$(2.21) \qquad |\kappa^{(k)}| \leq \frac{(2Ck)^{2(c+1)k}}{Nq^{k-2}} \quad (k \geq 3).$$

For example, if $X_{\alpha i}$ is the Bernoulli random variable defined in (2.8) with $p = p'$, then letting $q = \sqrt{Np}$,

$$(2.22) \quad \begin{aligned} \mathbb{E}[(X_{\alpha i})^k] &= \frac{(-p)^k(1-p) + (1-p)^k p}{(Np(1-p))^{k/2}} \\ &= \frac{1}{Nq^{k-2}}(1 + O(p)) = \kappa^{(k)}(1 + O(p)). \end{aligned}$$

We will also use the normalized cumulants, $s^{(k)}$, by setting

$$(2.23) \qquad s^{(1)} := 0, \qquad s^{(2)} := 1, \qquad s^{(k)} := Nq^{k-2}\kappa^{(k)} \quad (k \geq 3).$$

2.3.1. *Improved local law up to the upper edge for sparse random matrices.* Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$ be the ordered eigenvalues of $X^\dagger X$. Note that $m^{X^\dagger X}$ is the Stieltjes transform of the empirical eigenvalue distributions, $\mu^{X^\dagger X}$, of $X^\dagger X$ given by

$$(2.24) \qquad \mu^{X^\dagger X} := \frac{1}{N}\sum_{i=1}^{N}\delta_{\lambda_i}.$$

Recall d from (2.10). We introduce the following domain of the upper half-plane:

$$(2.25) \qquad \mathcal{E} = \left\{ E + i\eta : \frac{\lambda_-}{2} \leq E \leq \lambda_+ + 1, 0 < \eta < 3 \right\} \quad \text{if } d > 1,$$

respectively,

$$(2.26) \qquad \mathcal{E} = \left\{ E + \mathrm{i}\eta : \frac{1}{10} \le E \le \lambda_+ + 1, 0 < \eta < 3 \right\} \quad \text{if } d = 1.$$

Our first main result is the local law for $m^{X^\dagger X}$ up to the upper spectral edge.

THEOREM 2.7. *Let X satisfy Assumption 2.6 with $\phi > 0$. Then, there exist deterministic numbers $L_+ \ge L_- \ge 0$ and an algebraic function $\widetilde{m} : \mathbb{C}^+ \to \mathbb{C}^+$ such that the following hold*:

(1) *The function $\widetilde{m}$ is the Stieltjes transform of a deterministic probability measure $\widetilde{\rho}$, that is, $\widetilde{m}(z) = m_{\widetilde{\rho}}(z)$. The measure $\widetilde{\rho}$ is supported on the interval $[L_-, L_+]$ and is absolutely continuous with respect to Lebesgue measure with a strictly positive, continuous density on (L_-, L_+).*
(2) *The function $\widetilde{m} \equiv \widetilde{m}(z), z \in \mathbb{C}^+$, is a solution to the polynomial equation*

$$(2.27) \quad P_z(\widetilde{m}) := 1 + \left(z + 1 - \frac{1}{d} \right)\widetilde{m} + z\widetilde{m}^2 + \frac{s^{(4)}}{q^2}\widetilde{m}^2 \left(z\widetilde{m} + 1 - \frac{1}{d} \right)^2 = 0.$$

(3) *The normalized trace $m^{X^\dagger X}$ of the Green function of $X^\dagger X$ satisfies the local law*

$$(2.28) \qquad \left| m^{X^\dagger X}(z) - \widetilde{m}(z) \right| \prec \frac{1}{q^2} + \frac{1}{N\eta},$$

uniformly on the domain $\mathcal{E}, z = E + \mathrm{i}\eta$.

The right endpoint L_+ is given by

$$(2.29) \qquad L_+ = \left(1 + \frac{1}{\sqrt{d}} \right)^2 + \frac{1}{\sqrt{d}}\left(1 + \frac{1}{\sqrt{d}} \right)^2 \frac{s^{(4)}}{q^2} + O(q^{-4})$$

and the left endpoint L_- for the case $d > 1$ is given by

$$(2.30) \qquad L_- = \left(1 - \frac{1}{\sqrt{d}} \right)^2 - \frac{1}{\sqrt{d}}\left(1 - \frac{1}{\sqrt{d}} \right)^2 \frac{s^{(4)}}{q^2} + O(q^{-4}).$$

The sparsity of the entries of X makes its eigenvalues follow the deterministic law $\widetilde{\rho}$ that depends on the sparsity parameter q. While this law approaches the Marchenko–Pastur law, ρ_{MP}, when $N \to \infty$, its deterministic refinement to the standard Marchenko–Pastur law for finite N accounts for the nonoptimality at the edge of previous results obtained in [9]. (See (2.52) below for the estimate in [9].)

From the local law (2.28), we can also obtain estimates on the local density of states of $X^\dagger X$. For $E_1 < E_2$, define

$$(2.31) \quad \mathbf{n}(E_1, E_2) := \frac{1}{N}\left| \{ i : E_1 < \lambda_i \le E_2 \} \right|, \qquad n_{\widetilde{\rho}}(E_1, E_2) := \int_{E_1}^{E_2} \widetilde{\rho}(x)\,\mathrm{d}x.$$

Applying the Helffer–Sjöstrand calculus with Theorem 2.7, we get the following corollary.

COROLLARY 2.8.   *Suppose that X satisfies Assumption 2.6 with $\phi > 0$. Let $E_1, E_2 \in \mathbb{R}$, $E_1 < E_2$. Then,*

$$(2.32) \qquad \left| \mathbf{n}(E_1, E_2) - n_{\widetilde{\rho}}(E_1, E_2) \right| \prec \frac{E_2 - E_1}{q^2} + \frac{1}{N}.$$

We can obtain the following estimates of the operator norm $\|X^\dagger X\|$ of $X^\dagger X$, by combining the local law with the deterministic refinement to the Marchenko–Pastur law. This allows us to assert a strong statement on the location of the extremal eigenvalues of $X^\dagger X$.

THEOREM 2.9.   *Suppose that X satisfies Assumption 2.6 with $\phi > 0$. Then,*

$$(2.33) \qquad \left| \|X^\dagger X\| - L_+ \right| \prec \frac{1}{q^4} + \frac{1}{N^{2/3}},$$

where L_+ is the right endpoint of the support of the measure $\widetilde{\rho}$ given in (2.29).

2.3.2. *Tracy–Widom limit of the largest eigenvalues.*   Our second main result shows that the rescaled largest eigenvalues of the sparse sample covariance matrices of sparse random matrix converge in distribution to Tracy–Widom law, if the sparsity parameter q satisfies $q \gg N^{1/6}$.

THEOREM 2.10.   *Suppose that H satisfies Assumption 2.6 with $\phi > 1/6$. Denote by $\lambda_1^{X^\dagger X}$ the largest eigenvalue of $X^\dagger X$. Then*

$$(2.34) \qquad \lim_{N \to \infty} \mathbb{P}\big(\gamma N^{2/3}\big(\lambda_1^{X^\dagger X} - L_+\big) \leq s\big) = F_1(s),$$

where L_+ is given in (2.29), *γ is a constant defined as*

$$(2.35) \qquad \gamma = d^{1/2}(1 + \sqrt{d})^{-4/3},$$

and F_1 is the cumulative distribution function of the GOE Tracy–Widom law.

In the regime $q \leq N^{1/3}$, the deterministic shift of the right endpoint L_+ is crucial in the analysis of the largest eigenvalue, since the shift is of order q^{-2}, which is larger than $N^{-2/3}$, the scale of the Tracy–Widom fluctuation. Thus, to use the Tracy–Widom fluctuation of the largest eigenvalue in this regime, the correction from the fourth cumulant should be taken into consideration.

We further remark that all results in this subsection also hold for complex sparse sample covariance matrices with the GUE Tracy–Widom law as the governing law for the limiting edge fluctuation.

2.4. *Outline of proofs.* In the proof of the local law, we apply the strategy of cumulant expansion, which was used for sparse Wigner-type matrices in [25]. For the basic ideas of the cumulant expansion method in the proof of the local law, we refer to Sections 3.1 and 3.2 in [25], where the local law for the Gaussian Orthogonal Ensemble (GOE) matrices is proved with the method. For a GOE matrix W, it is based on the observation

$$1 + zG_{ii}^W = \sum_{k=1}^N W_{ik}G_{ki}^W$$

and the expectation of the right-hand side can be obtained by the Stein lemma. Since

$$\mathbb{E}[W_{ik}G_{ki}^W] = -\mathbb{E}[G_{kk}^W G_{ii}^W] - \mathbb{E}[G_{ki}^W G_{ki}^W],$$

the method suggests to consider the polynomial $1 + zm^W + (m^W)^2$. In case the matrix is sparse, we need to add a correction term to the polynomial, which can be obtained from the following generalized Stein lemma, introduced in [33] and applied to the proof of CLT for the linear eigenvalue statistics of random matrices by Lytova and Pastur [27].

LEMMA 2.11. *Fix $\ell \in \mathbb{N}$ and let $F \in C^{\ell+1}(\mathbb{R}; \mathbb{C}^+)$. Let Y be a centered random variable with finite moments to order $\ell + 2$. Then,*

$$(2.36) \qquad \mathbb{E}[YF(Y)] = \sum_{r=1}^\ell \frac{\kappa^{(r+1)}(Y)}{r!}\mathbb{E}[F^{(r)}(Y)] + \mathbb{E}[\Omega_\ell(YF(Y))],$$

where $\mathbb{E}$ denotes the expectation with respect to Y, $\kappa^{(r+1)}(Y)$ denotes the $(r+1)$-st cumulant of Y and $F^{(r)}$ denotes the rth derivative of the function F. The error term $\Omega_\ell(YF(Y))$ satisfies

$$\mathbb{E}[\Omega_\ell(YF(Y))] \le C_\ell \mathbb{E}[|Y|^{\ell+2}] \sup_{|t| \le Q} |F^{(\ell+1)}(t)|$$

$$(2.37)$$
$$+ C_\ell \mathbb{E}[|Y|^{\ell+2}\mathbf{1}(|Y| > Q) \sup_{t \in \mathbb{R}} |F^{(\ell+1)}(t)|,$$

where $Q > 0$ is an arbitrary fixed cutoff and C_ℓ satisfies $C_\ell \le (C\ell)^\ell/\ell!$ for some numerical constant C.

While the cumulant expansion method is applicable to the sample covariance matrices we consider, we face a new problem with the direct application of the method; since the Green function of $X^\dagger X$ is $G^{X^\dagger X}(z) = (X^\dagger X - zI)^{-1}$, we are led to consider

$$1 + zG_{ii}^{X^\dagger X} = \sum_{k=1}^N (X^\dagger X)_{ik}G_{ki}^{X^\dagger X},$$

and the Stein lemma applied on the right-hand side proposes a cubic polynomial instead of the quadratic polynomial we use for the Wigner case. This makes the analysis significantly harder, especially when the correction term due to the sparsity is taken into consideration.

The difficulty is resolved by introducing the self-adjoint linearization of a sample covariance matrix $X^\dagger X$, which showed to be useful in the study of sample covariance matrices, for example, in [2, 9, 16, 21, 24]. Let H be an $(N + M) \times (N + M)$ matrix such that

$$H(X, z) \equiv H = THT + TH\overline{T} + \overline{T}HT + \overline{T}H\overline{T},$$

where

$$THT := -zI, \qquad TH\overline{T} := X^\dagger, \qquad \overline{T}HT := X, \qquad \overline{T}H\overline{T} := -I.$$

Here T is the projection on the first N coordinates in $\mathbb{R}^{N+M}$ and $\overline{T} := \mathbf{1} - T$. In block matrix form, this is written as

$$(2.38) \qquad H = \left(\begin{array}{c|c} -zI & X^\dagger \\ \hline X & -I \end{array} \right).$$

Define the inverse matrix of $H(X, z)$ by $G := H(X, z)^{-1}$. From the Schur complement formula, we find that

$$(2.39) \qquad G_{ab}(z) = (TG(z)T)_{ab} = (X^\dagger X - zI)^{-1}_{ab}, \quad 1 \le a, b \le N,$$

$$(2.40) \qquad G_{\alpha\beta}(z) = (\overline{T}G(z)\overline{T})_{\alpha\beta} = z(XX^\dagger - zI)^{-1}_{\alpha\beta}, \quad N + 1 \le \alpha, \beta \le M + N.$$

Thus,

$$(2.41) \qquad \frac{1}{N} \sum_{a=1}^{N} G_{aa} = \frac{1}{N} \operatorname{Tr}(X^\dagger X - zI)^{-1} = m^{X^\dagger X}(z),$$

and

$$(2.42) \qquad \begin{aligned} \frac{1}{N} \sum_{\alpha=N+1}^{M+N} G_{\alpha\alpha} &= \frac{1}{N} \operatorname{Tr}(z(XX^\dagger - zI)^{-1}) \\ &= zm^{X^\dagger X}(z) + \frac{1}{N}(N - M) = zm^{X^\dagger X}(z) + 1 - \frac{1}{d}. \end{aligned}$$

We get the last line from the fact that $XX^\dagger$ have M eigenvalues identical to eigenvalues of $X^\dagger X$ except $(N - M)$ zero eigenvalues.

With the linearization trick, we can identify the quadratic polynomial that naturally arises from the sample covariance matrices, and we can also estimate the correction term as in (2.27). With the quartic polynomial in (2.27), we perform the analysis and prove the local law by the recursive moment method. The details for

the recursive moment method are discussed in Appendix A of the Supplementary Material [18].

To prove the Tracy–Widom fluctuations of the largest eigenvalue, we consider the Dyson matrix flow with initial condition $X_0 = X$ defined by

$$(2.43) \qquad X_t := e^{-t/2} X_0 + \sqrt{1 - e^{-t}} W^G \quad (t \geq 0),$$

where W^G is an $M \times N$ matrix with i.i.d. Gaussian entries independent of H_0. The Dyson matrix flow is one of the key ideas in the proof of universality results in random matrix theory. For its application to the proof of Tracy–Widom limit, we refer to [6, 7, 23].

In this work, we use the Dyson matrix flow in the Green function comparison in Section 5 in conjunction with the linearization trick. Let $H_t := H(X_t, z)$ be an $(N + M) \times (N + M)$ matrix defined as in (2.38). The local law can also be established for the normalized trace of the Green function of H_t by considering

$$(2.44) \qquad G_t(z) = (H_t)^{-1}, \qquad m_t(z) = \frac{1}{N} \sum_{i=1}^{N} (G_t)_{ii}(z) \quad (z \in \mathbb{C}^+).$$

Let $\kappa_t^{(k)}$ be the kth cumulant of $(X_t)_{ij}$. Then, by the linearity of the cumulants under the addition of independent random variables, we have $\kappa_t^{(1)} = 0$, $\kappa_t^{(2)} = 1/N$ and $\kappa_t^{(k)} = e^{-kt/2} \kappa^{(k)}$ for $k \geq 3$. In particular, we have the bound

$$(2.45) \qquad |\kappa_t^{(k)}| \leq e^{-t} \frac{(Ck)^{ck}}{N q_t^{k-2}} \quad (k \geq 3),$$

where we introduced the time-dependent sparsity parameter

$$(2.46) \qquad q_t := q e^{t/2}.$$

We also let $s_t^{(k)}$ be the normalized cumulants defined by

$$(2.47) \qquad s_t^{(k)} := N q_t^{k-2} \kappa_t^{(k)}.$$

With the parameters depending on t, we generalize Theorem 2.9 as follows:

PROPOSITION 2.12. *Let X_0 satisfy Assumption 2.6 with $\phi > 0$. Then, for any $t \geq 0$, there exist deterministic numbers $L_t^+ \geq L_t^- \geq 0$ and an algebraic function $\widetilde{m}_t : \mathbb{C}^+ \to \mathbb{C}^+$ such that the following hold:*

(1) *The function $\widetilde{m}_t$ is the Stieltjes transform of a deterministic probability measure $\widetilde{\rho}_t$, that is, $\widetilde{m}_t(z) = m_{\widetilde{\rho}_t}(z)$. The measure ρ_t is supported on $[L_t^-, L_t^+]$ and $\widetilde{\rho}_t$ is absolutely continuous with respect to Lebesgue measure with a strictly positive density on (L_t^-, L_t^+).*

(2) *The function $\widetilde{m}_t \equiv \widetilde{m}_t(z)$, $z \in \mathbb{C}^+$, is a solution to the polynomial equation*

(2.48)
$$P_z(\widetilde{m}_t) := 1 + \left(z + 1 - \frac{1}{d}\right)\widetilde{m}_t + z\widetilde{m}_t^2 + \frac{e^{-t}s_t^{(4)}}{q_t^2}\widetilde{m}_t^2\left(z\widetilde{m}_t + 1 - \frac{1}{d}\right)^2$$
$$= 0.$$

(3) *The normalized trace $m_t^{X^\dagger X}$ of the Green function of $X_t^\dagger X_t$ satisfies the local law*

(2.49)
$$\left|m_t^{X^\dagger X}(z) - \widetilde{m}_t(z)\right| \prec \frac{1}{q_t^2} + \frac{1}{N\eta},$$

uniformly on the domain $\mathcal{E}$ and uniformly in $t \in [0, 6\log N]$.

We remark that the local eigenvalue statistics of X_t and W^G agree up to negligible error for $t \geq 6\log N$.

For simplicity, we let $L_t \equiv L_t^+$, the upper edge of the support of $\widetilde{\rho}_t$. In Section 3, we show that

(2.50)
$$L_t = \left(1 + \frac{1}{\sqrt{d}}\right)^2 + \frac{1}{\sqrt{d}}\left(1 + \frac{1}{\sqrt{d}}\right)^2 e^{-t}q_t^{-2}s^{(4)} + O\left(e^{-2t}q_t^{-4}\right),$$

and also satisfies

(2.51)
$$\dot{L}_t = -\frac{2}{\sqrt{d}}\left(1 + \frac{1}{\sqrt{d}}\right)^2 e^{-t}q_t^{-2}s^{(4)} + O\left(e^{-2t}q_t^{-4}\right),$$

where $\dot{L}_t$ denotes the derivative of L_t with respect to t. The actual proof of the Tracy–Widom fluctuation in Section 5 will be done by comparing $\dot{L}_t$ and the time change of a suitable functional of the Green function.

In the proof of the local law, Theorem 2.7, we use the following results of [9] as a priori estimates.

PROPOSITION 2.13 (Lemma 3.11 and Lemma 3.12 of [9]). *Suppose X satisfies Assumption 2.6 with $\phi > 0$. Then,*

(1) (*Local Marchenko–Pastur law*) *The following estimates hold uniformly for $z \in \mathcal{E}$:*

(2.52)
$$\left|m^{X^\dagger X}(z) - m_{\mathrm{MP}}(z)\right| \prec \min\left\{\frac{1}{q}, \frac{1}{q^2\sqrt{\kappa + \eta}}\right\} + \frac{1}{N\eta},$$

(2.53)
$$\max_{i,j}\left|G_{ij}(z) - \delta_{ij}\Pi_{ij}(z)\right| \prec \frac{1}{q} + \sqrt{\frac{\mathrm{Im}\, m_{\mathrm{MP}}(z)}{N\eta}} + \frac{1}{N\eta},$$

where $\kappa \equiv \kappa(z) := |E - \lambda_+|$, $z = E + i\eta$, and

$$\Pi(z) = \begin{pmatrix} m_{\mathrm{MP}}(z)I_{N\times N} & 0 \\ 0 & -(1 + m_{\mathrm{MP}}(z))^{-1}I_{M\times M} \end{pmatrix}.$$

(2) (*Delocalization*) *For the ℓ^2-normalized eigenvectors* $(\mathbf{u}_k^{X^\dagger X})$ *and* $(\mathbf{v}_\alpha^{XX^\dagger})$,

$$(2.54) \qquad \max_k \|\mathbf{u}_k^{X^\dagger X}\|_\infty + \max_\alpha \|\mathbf{v}_\alpha^{XX^\dagger}\|_\infty \prec \frac{1}{\sqrt{N}}.$$

(3) *For the* $(N+M) \times (N+M)$ *matrix H defined as* (2.38),

$$(2.55) \qquad \|H\|^2 - \lambda_+ \prec \frac{1}{q^2} + \frac{1}{N^{2/3}}.$$

Note that the estimate (2.52) is essentially optimal as long as the spectral parameter z stay away from the spectral edges, for example, for energies in the bulk $E \in [\lambda_- + \delta, \lambda_+ - \delta]$ for some (N-independent) $\delta > 0$. For the individual Green function entries $G_{i\alpha}$, we believe that the estimate (2.53) is already essentially optimal.

REMARK 2.14 (Notational remark 2). We use Latin letters for indices in $[1, N]$, Greek letters for indices in $[N + 1, M + N]$, and fraktur letters for indices ranging from 1 to $M + N$. For Latin, respectively Greek, indices, we abbreviate

$$\sum_i := \sum_{i=1}^N, \qquad \sum_\alpha := \sum_{\alpha=N+1}^{M+N}.$$

For simplicity, we also let

$$\sum_{i,\alpha} := \sum_{\substack{1 \le i \le N \\ N+1 \le \alpha \le N+M}}, \qquad \sum_{\mathfrak{i}} := \sum_{\mathfrak{i}=1}^{M+N}.$$

3. Stieltjes transform of $\widetilde{\rho}$. In this section, we prove several properties of $\widetilde{m}_t$ and its Stieltjes inversion $\widetilde{\rho}_t$.

LEMMA 3.1. *For fixed $z = E + i\eta \in \mathcal{E}$ and $t \ge 0$, the equation $P_{t,z}(w_t) = 0$ has a unique solution $w_t \equiv w_t(z)$ satisfying $\operatorname{Im} w_t > 0$ and $|w_t| \le \frac{6\lambda_+}{\lambda_-}$. Furthermore, w_t satisfies the following properties*:

(1) *There exists a probability measure $\widetilde{\rho}_t$ such that the analytic continuation of $w_t(z)$ coincides with the Stieltjes transform, $m_{\widetilde{\rho}_t}(z)$, of $\widetilde{\rho}_t$.*
(2) *The probability measure $\widetilde{\rho}_t$ is supported on $[L_t^-, L_t^+]$ for some $L_t^- \ge 0$ and $L_t^+ \equiv L_t \ge (1 + \sqrt{1/d})^2$, and it exhibits a square-root decay at the upper edge, that is,*

$$(3.1) \qquad \widetilde{\rho}_t(E) \sim \sqrt{L_t - E} \quad \left(E \in \left[L_t - \frac{1}{\sqrt{d}}, L_t \right] \right).$$

Moreover,

$$(3.2) \qquad L_t = \left(1 + \frac{1}{\sqrt{d}} \right)^2 + \frac{1}{\sqrt{d}} \left(1 + \frac{1}{\sqrt{d}} \right)^2 e^{-t} q_t^{-2} s^{(4)} + O(e^{-2t} q_t^{-4}).$$

(3) *Setting*

(3.3) $$\varkappa_t \equiv \varkappa_t(E) := \min\{|E + L_t|, |E - L_t|\},$$

the solution w_t satisfies that

(3.4) $$\left|P'_{t,z}(w_t)\right| \sim \sqrt{\varkappa_t(E) + \eta}$$

and

$$\operatorname{Im} w_t(E + i\eta) \sim \frac{\eta}{\sqrt{\varkappa_t(E) + \eta}} \quad \text{if } E > L_t,$$

(3.5) $$\operatorname{Im} w_t(E + i\eta) \sim \sqrt{\varkappa_t(E) + \eta} \quad \text{if } L_t - \frac{1}{\sqrt{d}} \le E \le L_t,$$

$$\operatorname{Im} w_t(E + i\eta) = O(1) \quad \text{if } E < L_t - \frac{1}{\sqrt{d}}.$$

PROOF. Assume first that $d > 1$. Recall the definition of $P \equiv P_{t,z}$,

(3.6) $$P(w) = 1 + \left(z + 1 - \frac{1}{d}\right)w + zw^2 + \frac{s^{(4)}e^{-t}}{q_t^2}w^2\left(zw + 1 - \frac{1}{d}\right)^2.$$

Solving the equation $P = 0$ for z by the quadratic formula, we get

$$z = Q(w) \equiv Q_{z,t}(w_t)$$

(3.7)
$$= \frac{-(2e^{-t}q_t^{-2}s^{(4)}(1 - \frac{1}{d})w^2 + w + 1) + \sqrt{-4e^{-t}q_t^{-2}s^{(4)}w^2 \cdot \frac{1}{d} + (w+1)^2}}{2e^{-t}q_t^{-2}s^{(4)}w^3}.$$

For $w \sim 1$, its derivative is given by

$$Q'(w) = \frac{1}{2e^{-t}q_t^{-2}s^{(4)}}\left(2e^{-t}q_t^{-2}s^{(4)}\left(1 - \frac{1}{d}\right)w^{-2} + 2w^{-3} + 3w^{-4}\right.$$

$$\left. + \frac{8e^{-t}q_t^{-2}s^{(4)}w^{-5} \cdot \frac{1}{d} - (w^{-3} + w^{-2})(2w^{-3} + 3w^{-4})}{\sqrt{-4e^{-t}q_t^{-2}s^{(4)}w^{-4} \cdot \frac{1}{d} + (w^{-3} + w^{-2})^2}}\right)$$

$$= \frac{1}{2e^{-t}q_t^{-2}s^{(4)}}\left(2w^{-3} + 3w^{-4}\right.$$

(3.8)
$$- (2w^{-3} + 3w^{-4})\left[1 - 4e^{-t}q_t^{-2}s^{(4)}\frac{1}{d}\frac{w^2}{(1+w)^2}\right]^{-1/2}$$

$$+ 2e^{-t}q_t^{-2}s^{(4)}\left(1 - \frac{1}{d}\right)w^{-2} + 8e^{-t}q_t^{-2}s^{(4)}\frac{1}{w^2(1+w)} \cdot \frac{1}{d}\right)$$

$$+ O(q_t^{-2})$$

$$= -\frac{2w + 3}{w^2(1+w)^2} \cdot \frac{1}{d} + \left(1 - \frac{1}{d}\right)w^{-2} + \frac{4}{w^2(1+w)} \cdot \frac{1}{d} + O(e^{-t}q_t^{-2}).$$

Let

$$
\widetilde{Q}'(w) = -\frac{2w+3}{w^2(1+w)^2} \cdot \frac{1}{d} + \left(1 - \frac{1}{d}\right)w^{-2} + \frac{4}{w^2(1+w)} \cdot \frac{1}{d}
$$

(3.9)

$$
= \frac{1}{w^2} - \frac{1}{d}\frac{1}{(1+w)^2}.
$$

Then, $\widetilde{Q}'(w)$ is independent of q_t and strictly increasing on $(-1, 0)$. Furthermore,

(3.10)
$$
\widetilde{Q}'\left(-\frac{1}{1 + \frac{1}{\sqrt{d}}}\right) = 0.
$$

Thus, there exists a unique solution $w = \tau_t$ of the equation $Q'(w) = 0$ in $(-1, 0)$, which satisfies

(3.11)
$$
\tau_t = -\frac{1}{1 + \frac{1}{\sqrt{d}}} + O\left(e^{-t}q_t^{-2}\right).
$$

Let $L_t = Q(\tau_t)$. Calculating the terms of order $e^{-t}q_t^{-2}$ precisely, we get

$$
\tau_t = \frac{1}{-1 - \frac{1}{\sqrt{d}}} + O\left(e^{-2t}q_t^{-4}\right),
$$

(3.12)

$$
L_t = \left(1 + \frac{1}{\sqrt{d}}\right)^2 + \frac{1}{\sqrt{d}}\left(1 + \frac{1}{\sqrt{d}}\right)^2 e^{-t}q_t^{-2}s^{(4)} + O\left(e^{-2t}q_t^{-4}\right).
$$

For simplicity, we let $L \equiv L_t$ and $\tau = \tau_t$. We now expand z about τ to find that

$$
z = Q(\tau) + Q'(\tau)(w - \tau) + \frac{Q''(\tau)}{2}(w - \tau)^2 + O\left(|w - \tau|^3\right)
$$

(3.13)

$$
= L + \frac{Q''(\tau)}{2}(w - \tau)^2 + O\left(|w - \tau|^3\right)
$$

in a $q_t^{-1/2}$-neighborhood of τ. Since

(3.14)
$$
Q''(\tau) = \widetilde{Q}''(\tau) + O\left(e^{-t}q_t^{-2}\right)
$$

and $\widetilde{Q}''(\tau) \sim 1$ where $\widetilde{Q}'$ is monotone increasing on $(-1, 0)$, we find that $Q''(\tau) > 0$. We hence find that

(3.15)
$$
w = \tau + \left(\frac{2}{Q''(\tau)}\right)^{1/2}\sqrt{z - L} + O\left(|z - L|\right)
$$

in this neighborhood. Choosing the branch of the square root so that $\sqrt{z - L} \in \mathbb{C}^+$, we find that $\operatorname{Im} w > 0$.

Let $B_0 := \{w \in \mathbb{C} : |w| < \frac{6\lambda_+}{\lambda_-}\}$. For $z \in \mathcal{E}$ and $|w| = \frac{6\lambda_+}{\lambda_-}$,

(3.16) $\quad |zw^2| = \frac{|w|}{2}|zw| + \frac{|zw|}{2}|w| \geq 3|zw| + \lambda_+|w| > 2 + |zw| + \left|1 - \frac{1}{d}\right||w|$

hence

$$
\left| 1 + \left(z + 1 - \frac{1}{d} \right) w + z w^2 \right| \geq |z w^2| - \left| \left(z + 1 - \frac{1}{d} \right) w \right| - 1
$$

(3.17)

$$
> 1 > \left| \frac{s^{(4)} e^{-t}}{q_t^2} w^2 \left(z w + 1 - \frac{1}{d} \right)^2 \right|.
$$

Now, Rouché's theorem implies that the polynomial $P(w)$ has the same number of roots as the quadratic polynomial $1 + (z + 1 - \frac{1}{d}) w + z w^2 = 0$ in B_0. Hence, we conclude that $P(w) = 0$ has two solutions on B_0.

Let us extend $w \equiv w(z)$ to cover $z \in \partial \mathcal{E} \cap \mathbb{R}$. Then, w forms a curve $w : \partial \mathcal{E} \cap \mathbb{R} \to \mathbb{C}$, which we will denote by Γ. We already know that Γ intersects the real axis at τ. Let $\widetilde{\tau}$ be the largest real number such that $\widetilde{\tau} < \tau$ and Γ intersects the real axis at $\widetilde{\tau}$. Since

$$
(3.18) \qquad 1 + \left(\widetilde{\tau} + 1 - \frac{1}{d} \right) w(\widetilde{\tau}) + \widetilde{\tau} w(\widetilde{\tau})^2 = O(e^{-t} q_t^{-2}),
$$

it can be easily checked from the quadratic formula

$$
\begin{aligned}
w = \frac{1}{2z} \Bigg[&-\left(z + 1 - \frac{1}{d} \right) \\
&+ \sqrt{ \left(z + 1 - \frac{1}{d} \right)^2 - 4z \left(1 + e^{-t} q_t^{-2} s^{(4)} w^2 \left(z w + 1 - \frac{1}{d} \right)^2 \right) } \Bigg]
\end{aligned}
$$

(3.19)

that $m_{\mathrm{MP}}(\widetilde{\tau}) - w(\widetilde{\tau}) = O(e^{-t/2} q_t^{-1})$ and also $|\widetilde{\tau} - \lambda_-| = O(e^{-t} q_t^{-2})$, where m_{MP} is the Stieltjes transform of the Marchenko–Pastur law; see (2.15). Since we chose the branch of the square root in (3.15) so that $\sqrt{z - L} \in \mathbb{C}^+$, we find that the curve $\Gamma \in \mathbb{C}^+ \cup \mathbb{R}$, joining τ and $\widetilde{\tau}$. This shows that one solution of $P(w) = 0$ is in $\mathbb{C}^+$. Choosing the other branch for the square root in (3.15), we can identify another solution of $P(w) = 0$ in $\mathbb{C}^-$. Since there are only two solutions of $P(w) = 0$ in B_0, this proves the uniqueness statement of the lemma. Furthermore, by the analytic inverse function theorem, we also find that $w(z)$ is analytic for $z \in (w^{-1}(\widetilde{\tau}), L)$ since $Q'(w) \neq 0$ for such z.

For a large but N-independent z, we can find from the quadratic formula (3.19) that $w(z) = -\frac{1}{z} + o(\frac{1}{z})$. By continuity, this shows that the analytic continuation of $w(z)$ for $z \in \mathbb{C}^+$ is in the domain D_Γ enclosed by Γ and the real axis. In particular, $|w(z)| < \frac{6\lambda_+}{\lambda_-}$ for all $z \in \mathbb{C}^+$.

We then prove the analyticity of $w(z)$ in $\mathbb{C}^+$. It suffices to show that $Q'(w) \neq 0$ for $w \in D_\Gamma$. If $Q'(w) = 0$ for some $w \in D_\Gamma$, we have

$$
(3.20) \qquad 0 = w^2 Q'(w) = 1 - \frac{1}{d} \frac{w^2}{(1 + w)^2} + O(e^{-t} q_t^{-2}).
$$

We again use Rouché's theorem. Since $d \geq 1$, we have $1 - \frac{1}{d}\frac{w^2}{(1+w)^2} > c \gg e^{-t}q_t^{-2}$ for $|w| = \frac{6\lambda_+}{\lambda_-}$. Hence, $w^2 Q'(w) = 0$ has two solutions in the disk B_0. We already know that those solutions are τ and $\tilde{\tau}$. Thus, $Q'(w) \neq 0$ for $w \in D_\Gamma$ and $w(z)$ is analytic.

Let $\tilde{\rho} \equiv \tilde{\rho}$ be the Stieltjes inversion of $w \equiv w(z)$. In order to show that $\tilde{\rho}$ is a probability measure, it suffices to show that $\lim_{y \to \infty} iyw(iy) = -1$. By considering $z = iy$ in (3.19), it can be easily checked. This proves the first part of the lemma.

The second part of the lemma is already proved in the previous computation in the proof. To prove the last part of the lemma, with a slight abuse of notation, let $P_t(w, z) = P_{t,z}(w)$. We notice that

$$(3.21) \qquad 0 = \frac{d}{dw} P_t(w, z) = \frac{\partial z}{\partial w} \cdot \frac{\partial}{\partial z} P(w, z) + \frac{\partial}{\partial w} P(w, z).$$

From (3.13) and (3.15), we find that

$$(3.22) \qquad \frac{\partial z}{\partial w} \sim \sqrt{z - L} \sim \sqrt{\varkappa + \eta}.$$

We claim that $\frac{\partial}{\partial z} P(w, z) \sim 1$, which would prove the first relation in the last part of the lemma. Since

$$(3.23) \qquad \frac{\partial}{\partial z} P(w, z) = w + w^2 + \frac{2s^{(4)}e^{-t}w}{q_t^2} w^2 \left(zw + 1 - \frac{1}{d}\right),$$

it suffices to prove that

$$(3.24) \qquad |w|, |1 + w| > c > 0,$$

for some constant c independent of N. If we assume $|w| \leq c$, then

$$(3.25) \qquad |P(w, z)| \geq 1 - c\left|z + 1 - \frac{1}{d}\right| - c^2|z| - Cq_t^{-2} > c,$$

for some (small) $c > 0$, which contradicts that $P(w, z) = 0$. Similarly, if we assume $|1 + w| \leq c$, then

$$(3.26) \qquad |P(w, z)| \geq \frac{1}{d} - |1 + w| - |zw| \cdot |1 + w| - \frac{|1 + w|}{d} - Cq_t^{-2} > c,$$

for some (small) $c > 0$. This proves that $\frac{\partial}{\partial z} P(w, z) \sim 1$, and we find that

$$(3.27) \qquad P'_{t,z}(w) = \frac{\partial}{\partial w} P(w, z) = -\frac{\partial z}{\partial w} \cdot \frac{\partial}{\partial z} P(w, z) \sim \sqrt{\varkappa + \eta}.$$

This proves the first relation in the last part of the lemma. Other relations in the last part of the lemma can be easily proved from the first property and (3.15).

If $d = 1$, the polynomial $P(w)$ reduces to

$$(3.28) \qquad P(w) = 1 + zw + zw^2 + e^{-t}q_t^{-2}s^{(4)}z^2w^4$$

and

$$(3.29) \qquad Q(w) = \frac{-(w+1) + \sqrt{-4e^{-t}q_t^{-2}s^{(4)}w^2 + (w+1)^2}}{2e^{-t}q_t^{-2}s^{(4)}w^3}.$$

The square root behavior does not change in this case, with $\tau_t = -\frac{1}{2} + O(e^{-t}q_t^{-2})$. For uniqueness, we consider the disk $B_{10} = \{w \in \mathbb{C} : |w| < 10\}$. For $z \in \mathcal{E}$ and $w \in \partial B_{10}$, $|zw^2| \geq |zw| + 2$. Hence, we can use Rouché's theorem and the uniqueness statement follows.

For the analyticity, if $Q'(w) = 0$ then

$$(3.30) \qquad 0 = w^2 Q'(w) = 1 - \frac{w^2}{(1+w)^2} + O(e^{-t}q_t^{-2}).$$

In this case, the equation $1 - \frac{w^2}{(1+w)^2} = 0$ has only one solution $w = -\frac{1}{2}$. Thus, $w^2 Q'(w) = 0$ has only one solution in the disk B_{10}, and it proves the analyticity. The remaining parts can be proved with suitable changes. $\square$

4. Proof of local laws.

4.1. *Proof of Proposition* 2.12. In this subsection, we prove Proposition 2.12. The following lemma provides the main tool of the proof, which is the recursive moment estimate for $P(m_t)$. It is analogous to Lemma 5.1 in [25] for sparse Wigner matrices.

LEMMA 4.1 (Recursive moment estimate). *Fix $\phi > 0$ and $t \geq 0$. Let X_0 satisfies Assumption* 2.6. *Then, for any $D > 10$ and small $\epsilon > 0$, the normalized trace of the Green function, $m_t \equiv m_t(z)$, of the matrix H_t satisfies*

$$\mathbb{E}|P(m_t)|^{2D}$$

$$\leq N^\epsilon \mathbb{E}\left[\left(\frac{1}{q_t^4} + \frac{\operatorname{Im} m_t}{N\eta} + \frac{N-M}{N^2}\right)|P(m_t)|^{2D-1}\right]$$

$$+ N^{-\epsilon/4}q_t^{-1}\mathbb{E}[|m_t - \tilde{m}_t|^2|P(m_t)|^{2D-1}] + N^\epsilon q_t^{-8D}$$

$$(4.1) \quad + N^\epsilon q_t^{-1} \sum_{s=2}^{2D} \sum_{u'=0}^{s-2} \mathbb{E}\left[\left(\frac{\operatorname{Im} m_t}{N\eta} + \frac{N-M}{N^2}\right)^{2s-u'-2}|P'(m_t)|^{u'}|P(m_t)|^{2D-s}\right]$$

$$+ N^\epsilon \sum_{s=2}^{2D} \mathbb{E}\left[\left(\frac{1}{N\eta} + \frac{1}{q_t}\left(\frac{\operatorname{Im} m_t}{N\eta} + \frac{N-M}{N^2}\right)^{1/2} + \frac{1}{q_t^2}\right)\right.$$

$$\left. \times \left(\frac{\operatorname{Im} m_t}{N\eta} + \frac{N-M}{N^2}\right)^{s-1}|P'(m_t)|^{s-1}|P(m_t)|^{2D-s}\right],$$

uniformly on the domain $\mathcal{E}$, for any sufficiently large N.

We prove Lemma 4.1 in Appendix A of the Supplementary Material [18]. In the following, we sketch the proof of the local law in Proposition 2.12; the details are found in Appendix B of the Supplementary Material [18]. The remaining parts of this section are partly adapted from Section 5 in [25], and we reproduce the argument here in a more structured and clear way.

Let $\widetilde{m}_t$ be the solution w_t in Lemma 3.1. To simplify the notation, we introduce the following z- and t-dependent deterministic parameters:

$$(4.2) \quad \alpha_1(z) := \operatorname{Im}\widetilde{m}_t(z), \qquad \alpha_2(z) := P'(\widetilde{m}_t(z)), \qquad \beta := \frac{1}{N\eta} + \frac{1}{q_t^2},$$

with $z = E + i\eta$. From Lemma 3.1, we check that $\alpha_1 \le C|\alpha_2|$. Further let

$$(4.3) \quad \Lambda_t(z) := |m_t(z) - \widetilde{m}_t(z)| \quad (z \in \mathbb{C}^+).$$

Note that from Proposition 2.13 and (3.19), we have that $\Lambda_t(z) \prec 1$ uniformly on $\mathcal{E}$.

The strategy is now as follows. We apply Young's inequality to split up all the terms on the right side of (4.1) and absorb resulting factors of $\mathbb{E}|P(m_t)|^{2D}$ into the left hand side. For example, for the first term on the right of (4.1), we get, upon using the notation in (4.3), that

$$N^\epsilon \left(\frac{\operatorname{Im}m_t}{N\eta} + \frac{N-M}{N^2} + q_t^{-4}\right)|P(m_t)|^{2D-1}$$

$$\le N^\epsilon \frac{\alpha_1 + \Lambda_t}{N\eta}|P(m_t)|^{2D-1} + N^\epsilon q_t^{-4}|P(m_t)|^{2D-1}$$

(4.4)

$$\le \frac{N^{(2D+1)\epsilon}}{2D}C^{2D}\beta^{2D}(\alpha_1 + \Lambda_t)^{2D} + \frac{N^{(2D+1)\epsilon}}{2D}q_t^{-8D}$$

$$+ \frac{2(2D-1)}{2D}N^{-\frac{\epsilon}{2D-1}}|P(m_t)|^{2D},$$

and note that the last term can be absorbed into the left side of (4.1). The same idea can be applied to the second term on the right side of (4.1). To hand the other terms, we Taylor expand $P'(m_t)$ around $\widetilde{m}_t$ as

$$(4.5) \quad |P'(m_t) - \alpha_2 - P''(\widetilde{m}_t)(m_t - \widetilde{m}_t)| \le Cq_t^{-2}\Lambda_t^2,$$

where we used (4.2). Noticing that $P''(\widetilde{m}_t) = 2z + O(q_t^{-2})$, we proceed in a similar way as above using Young's inequality. Skipping over some details, we eventually find

$$\mathbb{E}[|P(m_t)|^{2D}]$$

$$\le CN^{(2D+1)\epsilon}\mathbb{E}[\beta^{2D}(\alpha_1 + \Lambda_t)^D(|\alpha_2| + 15\Lambda_t)^D] + C\frac{N^{(2D+1)\epsilon}}{2D}q_t^{-8D}$$

$$(4.6) \qquad + C \frac{N^{-(D/4-1)\epsilon}}{2D} q_t^{-2D} \mathbb{E}[\Lambda_t^{4D}]$$

$$\leq N^{3D\epsilon} \beta^{2D} |\alpha_2|^{2D} + N^{3D\epsilon} \beta^{2D} \mathbb{E}[\Lambda_t^{2D}]$$

$$+ N^{3D\epsilon} q_t^{-8D} + N^{-D\epsilon/8} q_t^{-2D} \mathbb{E}[\Lambda_t^{4D}],$$

uniformly on $\mathcal{E}$, where we used $\alpha_1 \leq C\alpha_2$ to get the second line.

Next, we aim to control Λ_t in terms of $|P(m_t)|$. For that we Taylor expand $P(m_t)$ around $\widetilde{m}_t$ to get

$$(4.7) \qquad \left| P(m_t) - \alpha_2(m_t - \widetilde{m}_t) - \frac{1}{2} P''(\widetilde{m}_t)(m_t - \widetilde{m}_t)^2 \right| \leq C q_t^{-2} \Lambda_t^3,$$

since $P(\widetilde{m}_t) = 0$ and $P'''(\widetilde{m}_t) = 8e^{-t} q_t^{-2} s^{(4)}(z^2 \widetilde{m}_t + z(z\widetilde{m}_t + 1 - \frac{1}{d}))$. Then using $\Lambda_t \prec 1$ and $P''(\widetilde{m}_t) = 2z + O(q_t^{-2})$ we obtain

$$(4.8) \qquad \Lambda_t^2 \prec 2|\alpha_2|\Lambda_t + 2|P(m_t)| \quad (z \in \mathcal{E}).$$

This estimate can upon applying a Schwarz inequality be fed back into (4.6), to get the bound

$$(4.9) \qquad \mathbb{E}[|P(m_t)|^{2D}] \leq N^{5D\epsilon} \beta^{2D} |\alpha_2|^{2D} + N^{5D\epsilon} \beta^{4D} + q_t^{-2D} |\alpha_2|^{4D},$$

uniformly on $\mathcal{E}$. For any fixed $z \in \mathcal{E}$, Markov's inequality then yields $|P(m_t)| \prec |\alpha_2|\beta + \beta^2 + q_t^{-1}|\alpha_2|^2$. Uniformity in z is easily achieved using a lattice argument and the Lipschitz continuity of $m_t(z)$ and $\widetilde{m}_t(z)$ on $\mathcal{E}$. Finally, a Taylor expansion of $P(m_t)$ around $\widetilde{m}_t$ will give the following self-consistent equation for $m_t(z) - \widetilde{m}_t(z)$:

LEMMA 4.2. *Under the assumptions of Lemma* 4.1, *we have*

$$(4.10) \quad \left| \alpha_2(m_t - \widetilde{m}_t) + z(m_t - \widetilde{m}_t)^2 \right| \prec \beta \Lambda_t^2 + |\alpha_2|\beta + \beta^2 + q_t^{-1}|\alpha_2|^2,$$

uniformly on $\mathcal{E}$.

The detailed proof of Lemma 4.2 is given in Appendix B of the Supplementary Material [18].

We remark that the last term on the right-hand side of (4.10) is not yet optimal. To get a better estimate, we use the behavior of the $\alpha_2(z)$ to refine the analysis and obtain the following corollary.

COROLLARY 4.3. *Under the assumptions of Lemma* 4.1, *we have*

$$(4.11) \qquad \left| \alpha_2(m_t - \widetilde{m}_t) + z(m_t - \widetilde{m}_t)^2 \right| \prec \beta \Lambda_t^2 + |\alpha_2|\beta + \beta^2,$$

uniformly on $\mathcal{E}$.

PROOF. Recall from Lemma 3.1 that there is a constant $C_0 > 1$ such that $C_0^{-1}\sqrt{\varkappa_t(E) + \eta} \le |\alpha_2| \le C_0\sqrt{\varkappa_t(E) + \eta}$, where we can choose C_0 uniform in $z \in \mathcal{E}$. Note that, for a fixed E, $\beta = \beta(E + i\eta)$ is a decreasing function of η whereas $\sqrt{\varkappa_t(E) + \eta}$ is increasing. Hence there is $\widetilde{\eta_0} \equiv \widetilde{\eta_0}(E)$ such that $\sqrt{\varkappa(E) + \widetilde{\eta_0}} = C_0 q_t \beta(E + i\widetilde{\eta_0})$. We consider the subdomain $\widetilde{\mathcal{E}} \subset \mathcal{E}$ defined by

$$(4.12) \qquad \widetilde{\mathcal{E}} := \{z = E + i\eta \in \mathcal{E} : \eta > \widetilde{\eta}_0(E)\}.$$

On this subdomain $\widetilde{\mathcal{E}}$, $\beta \le q_t^{-1}|\alpha_2|$, hence we get from (4.10) that there is a high probability event $\widetilde{\Xi}$ such that

$$\big|\alpha_2(m_t - \widetilde{m}_t) + z(m_t - \widetilde{m}_t)^2\big| \le N^\epsilon \beta \Lambda_t^2 + N^\epsilon q_t^{-1}|\alpha_2|^2$$

and thus

$$|\alpha_2|\Lambda_t \le \big(|z| + N^\epsilon \beta\big)\Lambda_t^2 + N^\epsilon q_t^{-1}|\alpha_2|^2$$

uniformly on $\widetilde{\mathcal{E}}$ on $\widetilde{\Xi}$. Hence, on $\widetilde{\Xi}$,

$$(4.13) \quad |\alpha_2| \le 2\big(|z| + 1\big)\Lambda_t \le 12\Lambda_t \quad \text{or} \quad \Lambda_t \le 2N^\epsilon q_t^{-1}|\alpha_2| \quad (z \in \widetilde{\mathcal{E}}).$$

When $\eta = N^{-\epsilon}$, it is easy to see that

$$(4.14) \qquad \begin{aligned} |\alpha_2| &\ge \left|z + 1 - \frac{1}{d} + 2z\widetilde{m}_t\right| - Cq_t^{-2} \ge 2E\,\mathrm{Im}\,\widetilde{m}_t \\ &\ge c\sqrt{\eta} \gg 2N^\epsilon q_t^{-1}|\alpha_2|, \end{aligned}$$

for some constant c and sufficiently large N. We have that either $N^{-\epsilon/2}/12 \le \Lambda_t$ or $\Lambda_t \le 2N^\epsilon q_t^{-1}|\alpha_2|$ on $\widetilde{\Xi}$. From the a priori estimate (2.52), we know that $|\Lambda_t| \prec \frac{1}{q_t} + \frac{1}{N\eta}$, we hence find that

$$(4.15) \qquad \Lambda_t \le 2N^\epsilon q_t^{-1}|\alpha_2|,$$

holds on the event $\widetilde{\Xi}$. Putting (4.15) back into (4.6), we obtain that

$$(4.16) \qquad \begin{aligned} \mathbb{E}\big[|P(m_t)|^{2D}\big] &\le N^{4D\epsilon}\beta^{2D}|\alpha_2|^{2D} + N^{3D\epsilon}q_t^{-8D} + q_t^{-6D}|\alpha_2|^{4D} \\ &\le N^{6D\epsilon}\beta^{2D}|\alpha_2|^{2D} + N^{6D\epsilon}\beta^{4D}, \end{aligned}$$

for any small $\epsilon > 0$, and large D, uniformly on $\widetilde{\mathcal{E}}$. Note that, for $z \in \mathcal{E}\backslash\widetilde{\mathcal{E}}$, it is direct to check the estimate $\mathbb{E}[|P(m_t)|^{2D}] \le N^{6D\epsilon}\beta^{2D}|\alpha_2|^{2D} + N^{6D\epsilon}\beta^{4D}$. Applying a lattice argument and the Lipschitz continuity (see, e.g., the proof of Lemma 4.2), we find from a union bound that for any small $\epsilon > 0$ and large D there exists an event Ξ with $\mathbb{P}(\Xi) \ge 1 - N^{-D}$ such that

$$(4.17) \qquad \big|\alpha_2(m_t - \widetilde{m}_t) + z(m_t - \widetilde{m}_t)^2\big| \le N^\epsilon \beta \Lambda_t^2 + N^\epsilon |\alpha_2|\beta + N^\epsilon \beta^2,$$

on Ξ, uniformly on $\mathcal{E}$ for any sufficiently large N. $\square$

We now prove Proposition 2.12, which will also imply Theorem 2.9.

PROOF OF PROPOSITION 2.12 AND THEOREM 2.9.   Fix $t \in [0, 6 \log N]$. Let $\tilde{m}_t$ be the solution w_t in Lemma 3.1. We proved statements (1) and (2) in Lemma 3.1, it hence remains to prove statement (3) of Proposition 2.12.

Recall that for fixed E $\beta = \beta(E + i\eta)$ is a decreasing function of η, $\sqrt{\varkappa_t(E) + \eta}$ is an increasing function of η, and $\eta_0 \equiv \eta_0(E)$ satisfies that $\sqrt{\varkappa(E) + \eta_0} = 10 C_0 N^\epsilon \beta(E + i\eta_0)$. Further notice that $\eta_0(E)$ is a continuous function. We consider the subdomains of $\mathcal{E}$ defined by

$$\mathcal{E}_1 := \{z = E + i\eta \in \mathcal{E} : \eta \le \eta_0(E), 10 N^\epsilon \le N\eta\},$$
$$\mathcal{E}_2 := \{z = E + i\eta \in \mathcal{E} : \eta > \eta_0(E), 10 N^\epsilon \le N\eta\}.$$

We consider the cases $z \in \mathcal{E}_1$, $z \in \mathcal{E}_2$ and $z \in \mathcal{E} \setminus (\mathcal{E}_1 \cup \mathcal{E}_2)$, and split the stability analysis accordingly. Let Ξ be a high probability event such that (4.17) holds.

Case 1: If $z \in \mathcal{E}_1$, we note that $|\alpha_2| \le C_0 \sqrt{\varkappa(E) + \eta} \le 10 C_0^2 N^\epsilon \beta(E + i\eta)$. Then, we find that

$$|z| \Lambda_t^2 \le |\alpha_2| \Lambda_t + N^\epsilon \beta \Lambda_t^2 + N^\epsilon |\alpha_2| \beta + N^\epsilon \beta^2$$
$$\le 10 C_0^2 N^\epsilon \beta \Lambda_t + N^\epsilon \beta \Lambda_t^2 + (10 C_0^2 N^\epsilon + 1) N^\epsilon \beta^2,$$

on Ξ. Hence, there is some finite constant C such that on Ξ, we have $\Lambda_t \le C N^\epsilon \beta$, $z \in \mathcal{E}_1$.

Case 2: If $z \in \mathcal{E}_2$, we obtain that

$$(4.18) \qquad |\alpha_2| \Lambda_t \le (|z| + N^\epsilon \beta) \Lambda_t^2 + |\alpha_2| N^\epsilon \beta + N^\epsilon \beta^2,$$

on Ξ. We then notice that $C_0 |\alpha_2| \ge \sqrt{\varkappa_t(E) + \eta} \ge 10 C_0 N^\epsilon \beta$, that is, $N^\epsilon \beta \le |\alpha_2|/10$, so that

$$(4.19) \qquad |\alpha_2| \Lambda_t \le (|z| + 1) \Lambda_t^2 + (1 + N^\epsilon) |\alpha_2| \beta,$$

on Ξ, where we used that $N^\epsilon \beta \le 1$. Hence, on Ξ, either

$$(4.20) \qquad |\alpha_2| \le 2(1 + |z|) \Lambda_t \quad \text{or} \quad \Lambda_t \le 3 N^\epsilon \beta.$$

We now follow the dichotomy argument and the continuity argument used to obtain (4.15). Since $3 N^\epsilon \beta \le |\alpha_2|/8$ on $\mathcal{E}_2$, by continuity, we find that on the event Ξ, $\Lambda_t \le 3 N^\epsilon \beta$ for $z \in \mathcal{E}_2$.

Case 3: For $z \in \mathcal{E} \setminus (\mathcal{E}_1 \cup \mathcal{E}_2)$ we use that $|m_t'(z)| \le \frac{\operatorname{Im} m_t(z)}{\operatorname{Im} z}$, $z \in \mathbb{C}^+$, since m_t is a Stieltjes transform of a probability measure. Set $\tilde{\eta} := 10 N^{-1+\epsilon}$ and observe that

$$(4.21) \quad |m_t(E + i\eta)| \le \int_\eta^{\tilde{\eta}} \frac{s \operatorname{Im} m_t(E + is)}{s^2} \, ds + \Lambda_t(E + i\tilde{\eta}) + |\tilde{m}_t(E + i\tilde{\eta})|.$$

From the definition of the Stieltjes transform, it is easy to check that $s \to s \operatorname{Im} m_t(E + is)$ is monotone increasing. Thus, we find that

$$|m_t(E + i\eta)| \le \frac{2\tilde{\eta}}{\eta} \operatorname{Im} m_t(E + i\tilde{\eta}) + \Lambda_t(E + i\tilde{\eta}) + |\tilde{m}_t(E + i\tilde{\eta})|$$

$$(4.22)$$

$$\le C \frac{N^\epsilon}{N\eta} (\operatorname{Im} \tilde{m}_t(E + i\tilde{\eta}) + \Lambda_t(E + i\tilde{\eta})) + |\tilde{m}_t(E + i\tilde{\eta})|,$$

for some C where we used $\widetilde{\eta} = 10N^{-1+\epsilon}$ to obtain the second inequality. Since $z = E + i\widetilde{\eta} \in \mathcal{E}_1 \cup \mathcal{E}_2$, we have $\Lambda_t(E + i\widetilde{\eta}) \leq CN^\epsilon \beta(E + i\widetilde{\eta}) \leq C$ on Ξ. Using that $\widetilde{m}_t$ is uniformly bounded on $\mathcal{E}$, we get that, on Ξ, $\Lambda_t \leq CN^\epsilon \beta$, for all $z \in \mathcal{E} \backslash (\mathcal{E}_1 \cup \mathcal{E}_2)$.

In sum, we get $\Lambda_t \prec \beta$ uniformly on $\mathcal{E}$ for fixed $t \in [0, 6\log N]$. Choosing $t = 0$, we have proved Theorem 2.9. To prove that this bound holds for all $t \in [0, 6\log N]$, we use the continuity of the Dyson matrix flow. Choosing a lattice $\mathcal{L} \subset [0, 6\log N]$ with spacings of order N^{-3}, we find that $\Lambda_t \prec \beta$, uniformly on $\mathcal{E}$ and on $\mathcal{L}$, by a union bound. Thus, by continuity, we can extend the conclusion to all $t \in [0, 6\log N]$ and conclude the proof of Proposition 2.12. $\quad\square$

4.2. *Proof of Theorem* 2.9. Theorem 2.9 is a direct consequence of the following result.

LEMMA 4.4. *Let X_0 satisfy Assumption* 2.6 *with $\phi > 0$. Then,*

$$(4.23) \qquad \left| \|X_t^\dagger X_t\| - L_t \right| \prec \frac{1}{q_t^4} + \frac{1}{N^{2/3}},$$

uniformly in $t \in [0, 6\log N]$.

The proof of Lemma 4.4 is split into a lower and an upper bound. The lower bound is a direct consequence of the local law in Proposition 2.12. The upper bound requires an additional stability analysis starting from the first inequality in (4.6). This time we capitalize on the fact that $\alpha_1(z) = \operatorname{Im} \widetilde{m}_t(z)$ behaves as $\eta/\sqrt{\varkappa_t(E) + \eta}$, for $E \geq L_+$, to get sharper estimates outside of the spectrum. Since the arguments for the lower and upper bounds are similar to the ones in [25], we postpone their proofs to the Appendix B of the Supplementary Material [18].

5. Proof of Tracy–Widom limit for the largest eigenvalue. In this section, we prove the Theorem 2.10, the Tracy–Widom limiting distribution of the largest eigenvalue. Following the idea from [13], we consider the imaginary part of the normalized trace of the Green function $m \equiv m^{X^\dagger X}$ of $X^\dagger X$. For $\eta > 0$, let

$$(5.1) \qquad \theta_\eta(y) = \frac{\eta}{\pi(y^2 + \eta^2)} \quad (y \in \mathbb{R}).$$

It can be easily checked from the definition of the Green function that

$$(5.2) \qquad \operatorname{Im} m(E + i\eta) = \frac{\pi}{N} \operatorname{Tr} \theta_\eta(X^\dagger X - E).$$

The first proposition in this section shows how we can approximate the distribution of the largest eigenvalue by using the Green function. Recall that L_+ is the right endpoint of the deterministic probability measure in Theorem 2.7.

PROPOSITION 5.1.   *Let X satisfy Assumption 2.6, with $\phi > 1/6$. Denote by $\lambda_1^{X^\dagger X}$ the largest eigenvalue of $X^\dagger X$. Fix $\epsilon > 0$ and let $E \in \mathbb{R}$ be such that $|E - L_+| \leq N^{-2/3+\epsilon}$. Set $E_+ := L_+ + 2N^{-2/3+\epsilon}$ and define $\chi_E := \mathbf{1}_{[E,E_+]}$. Let $\eta_1 := N^{-2/3-3\epsilon}$ and $\eta_2 := N^{-2/3-9\epsilon}$. Let $K : \mathbb{R} \to [0, \infty)$ be a smooth function satisfying*

$$(5.3) \qquad K(x) = \begin{cases} 1 & \text{if } |x| < 1/3, \\ 0 & \text{if } |x| > 2/3, \end{cases}$$

which is a monotone decreasing on $[0, \infty)$. Then, for any $D > 0$,

$$(5.4) \qquad \mathbb{E}\big[K\big(\mathrm{Tr}(\chi_E * \theta_{\eta_2})(X^\dagger X)\big)\big] > \mathbb{P}(\lambda_1^{X^\dagger X} \leq E - \eta_1) - N^{-D}$$

and

$$(5.5) \qquad \mathbb{E}\big[K\big(\mathrm{Tr}(\chi_E * \theta_{\eta_2})(X^\dagger X)\big)\big] < \mathbb{P}(\lambda_1^{X^\dagger X} \leq E + \eta_1) + N^{-D}$$

for N sufficiently large, with θ_{η_2}.

For the proof, we refer to Proposition 7.1 of [25]. We remark that the lack of the improved local law near the lower edge does not alter the proof of Proposition 5.1.

Next, we state the Green function comparison result for our model. We let W^G be a $M \times N$ Gaussian matrix independent of X and denote by $m^G \equiv m^{W^G}$ the normalized trace of its Green function.

PROPOSITION 5.2.   *Under the assumptions of Proposition 5.1 the following holds. Let $\epsilon > 0$ and set $\eta_0 = N^{-2/3-\epsilon}$. Let $E_1, E_2 \in \mathbb{R}$ satisfy $|E_1|, |E_2| \leq N^{-2/3+\epsilon}$. Consider a smooth function $F : \mathbb{R} \to \mathbb{R}$ such that*

$$(5.6) \qquad \max_{x \in \mathbb{R}} |F^{(l)}(x)|(|x| + 1)^{-C} \leq C \quad (l \in [1, 11]).$$

Then, for any sufficiently small $\epsilon > 0$, there exists $\delta > 0$ such that

$$(5.7) \qquad \begin{aligned} \bigg| &\mathbb{E}F\bigg(N \int_{E_1}^{E_2} \mathrm{Im}\, m(x + L_+ + i\eta_0)\, dx\bigg) \\ &- \mathbb{E}F\bigg(N \int_{E_1}^{E_2} \mathrm{Im}\, m^G(x + \lambda_+ + i\eta_0)\, dx\bigg)\bigg| \\ &\leq N^{-\delta}, \end{aligned}$$

for large enough N.

Proposition 5.2 directly implies Theorem 2.10, the Tracy–Widom limit for the largest eigenvalue. A detailed proof is found, for example, with the same notation in [25], Section 7.

In the remainder of the section, we prove Proposition 5.2. We begin by the following application of the generalized Stein lemma.

LEMMA 5.3. *Fix $\ell \in \mathbb{N}$ and let $F \in C^{\ell+1}(\mathbb{R}; \mathbb{C}^+)$. Let $Y \equiv Y_0$ be a random variable with finite moments to order $\ell + 2$ and let W be a Gaussian random variable independent of Y. Assume that $\mathbb{E}[Y] = \mathbb{E}[W] = 0$ and $\mathbb{E}[Y^2] = \mathbb{E}[W^2]$. Introduce*

$$(5.8) \qquad\qquad Y_t := e^{-t/2}Y_0 + \sqrt{1 - e^{-t}}\, W,$$

and let $\dot{Y}_t \equiv dY_t/dt$. Then,

$$(5.9) \quad \mathbb{E}[\dot{Y}_t F(Y_t)] = -\frac{1}{2}\sum_{r=2}^{\ell} \frac{\kappa^{(r+1)}(Y_0)}{r!} e^{-\frac{(r+1)t}{2}} \mathbb{E}[F^{(r)}(Y_t)] + \mathbb{E}[\Omega_\ell(\dot{Y}_t F(Y_t))],$$

where $\mathbb{E}$ denotes the expectation with respect to Y and W, $\kappa^{(r+1)}(Y)$ denotes the $(r+1)$th cumulant of Y and $F^{(r)}$ denotes the rth derivative of the function F. The error term Ω_ℓ in (5.9) satisfies

$$
\begin{aligned}
\big|\mathbb{E}[\Omega_\ell(\dot{Y}_t F(Y_t))]\big| &\le C_\ell \mathbb{E}[|Y_t|^{|\ell+2}] \sup_{|x| \le Q} \big|F^{(\ell+1)}(x)\big| \\
(5.10) \qquad &\quad + C_\ell \mathbb{E}[|Y_t|^{\ell+2}\mathbf{1}(|Y_t| > Q)] \sup_{x \in \mathbb{R}} \big|F^{(\ell+1)}(x)\big|,
\end{aligned}
$$

where $Q > 0$ is an arbitrary fixed cutoff and C_ℓ satisfies $C_\ell \le \frac{(C\ell)^\ell}{\ell!}$ for some numerical constant C.

PROOF OF PROPOSITION 5.2. Fix a (small) $\epsilon > 0$. Consider $x \in [E_1, E_2]$. For simplicity, let

$$(5.11) \qquad G \equiv G_t(x + L_t + i\eta_0), \qquad m \equiv m_t(x + L_t + i\eta_0),$$

with $\eta_0 = N^{-2/3-\epsilon}$, and define

$$(5.12) \qquad\qquad Y \equiv Y_t := N \int_{E_1}^{E_2} \mathrm{Im}\, m(x + L_t + i\eta_0)\, dx.$$

Note that $Y \prec N^\epsilon$ and $|F^{(l)}(Y)| \prec N^{C\epsilon}$ for $l \in [1, 11]$. Recall from (2.50) and (2.51) that

$$
L_t = \lambda_+ + \frac{1}{\sqrt{d}}\left(1 + \frac{1}{\sqrt{d}}\right)^2 e^{-t}s^{(4)}q_t^{-2} + O(e^{-2t}q_t^{-4}),
$$

$$
\dot{L}_t = -2\frac{1}{\sqrt{d}}\left(1 + \frac{1}{\sqrt{d}}\right)^2 e^{-t}s^{(4)}q_t^{-2} + O(e^{-2t}q_t^{-4}),
$$

with $q_t = e^{t/2} q_0$. Let $z = x + L_t + i\eta_0$ and $G \equiv G(z)$. Differentiating $F(Y)$ with respect to t, we get

$$
\frac{d}{dt} \mathbb{E} F(Y) = \mathbb{E}\left[F'(Y) \frac{dY}{dt} \right] = \mathbb{E}\left[F'(Y) \operatorname{Im} \int_{E_1}^{E_2} \sum_{i=1}^{N} \frac{dG_{ii}}{dt} \, dx \right]
$$

(5.13)

$$
= \mathbb{E}\left[F'(Y) \operatorname{Im} \int_{E_1}^{E_2} \left(\sum_{i,j,\alpha} \dot{X}_{\alpha j} \frac{\partial G_{ii}}{\partial H_{\alpha j}} + \dot{L}_t \sum_{1 \le i, j \le N} G_{ij} G_{ji} \right) dx \right],
$$

where by definition

(5.14) $$\dot{X}_{\alpha j} \equiv (\dot{X}_t)_{\alpha j} = -\frac{1}{2} e^{-t/2} (X_0)_{\alpha j} + \frac{e^{-t}}{2\sqrt{1 - e^{-t}}} W_{\alpha j}^G.$$

Thus, we find that

$$
\sum_{i,j,\alpha} \mathbb{E}\left[\dot{X}_{\alpha j} F'(Y) \frac{\partial G_{ii}}{\partial X_{\alpha j}} \right]
$$

(5.15)

$$
= -2 \sum_{i,j,\alpha} \mathbb{E}[\dot{X}_{\alpha j} F'(Y) G_{i\alpha} G_{ji}]
$$

$$
= \frac{e^{-t}}{N} \sum_{r=2}^{\ell} \frac{q_t^{-(r-1)} s^{(r+1)}}{r!} \sum_{1 \le i \le j} \sum_{j,\alpha} \mathbb{E}[\partial_{\alpha j}^r (F'(Y) G_{i\alpha} G_{ji})] + O(N^{1/3 + C\epsilon}),
$$

for $\ell = 10$, where we use the short hand $\partial_{j\alpha} = \partial / \partial H_{j\alpha}$. Here, the error term $O(N^{1/3 + C\epsilon})$ in (5.15) corresponds to Ω_ℓ in (5.9), which is $O(N^{C\epsilon} N^2 q_t^{-10})$ for $Y = H_{j\alpha}$.

We claim the following lemma which is proved in Appendix C of Supplementary Material [18].

LEMMA 5.4. *For an integer $r \ge 2$, let*

(5.16) $$J_r := \frac{e^{-t}}{N} \frac{q_t^{-(r-1)} s^{(r+1)}}{r!} \sum_{i,j,\alpha} \mathbb{E}[\partial_{j\alpha}^r (F'(Y) G_{ij} G_{\alpha i})].$$

Then, for any $r \ne 3$,

(5.17) $$J_r = O(N^{2/3 - \epsilon'})$$

and

(5.18) $$J_3 = \frac{2}{\sqrt{d}} \left(1 + \frac{1}{\sqrt{d}} \right)^2 e^{-t} s^{(4)} q_t^{-2} \sum_{i,j} \mathbb{E}[F'(Y) G_{ij} G_{ji}] + O(N^{2/3 - \epsilon'}).$$

Assuming Lemma 5.4, we find that there exists $\epsilon' > 2\epsilon$ such that, for all $t \in [0, 6\log N]$,

$$(5.19) \quad \sum_{i,j,\alpha} \mathbb{E}\left[\dot{X}_{j\alpha} F'(Y) \frac{\partial G_{ii}}{\partial X_{j\alpha}} \right] = -\dot{L}_t \sum_{i,j} \mathbb{E}[G_{ij} G_{ji} F'(Y)] + O(N^{2/3-\epsilon'}),$$

which implies that the right-hand side of (5.13) is $O(N^{-\epsilon'/2})$. Integrating equation (5.13) from $t = 0$ to $t = 6\log N$, we get

$$\left| \mathbb{E}F\left(N \int_{E_1}^{E_2} \operatorname{Im} m(x + L_t + i\eta_0)\, \mathrm{d}x \right)_{t=0} \right.$$
$$\left. - \mathbb{E}F\left(N \int_{E_1}^{E_2} \operatorname{Im} m(x + L_t + i\eta_0)\, \mathrm{d}x \right)_{t=6\log N} \right| \le N^{-\epsilon'/4}.$$

We remark that the comparison between $\operatorname{Im} m|_{t=6\log N}$ and $\operatorname{Im} m^G$ is trivial as in the Wigner-type case. Comparing ith largest eigenvalues of $X^\dagger X$ and λ_i^G, we find that

$$(5.20) \qquad \qquad \left| \operatorname{Im} m|_{t=6\log N} - \operatorname{Im} m^G \right| \prec N^{-5/3}.$$

This completes the proof of desired proposition. $\square$

Acknowledgements. We thank Xiucai Ding for useful comments and suggestions. We are also grateful to the anonymous referee for carefully reading our manuscript and suggesting several improvements. Jong Yun Hwang and Ji Oon Lee thank Wooseok Yang for correcting several typos in the manuscript.

SUPPLEMENTARY MATERIAL

Supplement: Proofs of some lemmas (DOI: 10.1214/19-AAP1472SUPP; .pdf). In the Supplementary Material [18], we will provide the proofs of Lemmas 4.1, 4.2, 4.4 and 5.4.

REFERENCES

[1] AKHIEZER, N. I. (1965). *The Classical Moment Problem and Some Related Questions in Analysis*. Hafner, New York. Translated by N. Kemmer. MR0184042

[2] ANDERSON, G. W. (2013). Convergence of the largest singular value of a polynomial in independent Wigner matrices. *Ann. Probab.* **41** 2103–2181. MR3098069

[3] BAO, Z., PAN, G. and ZHOU, W. (2015). Universality for the largest eigenvalue of sample covariance matrices with general population. *Ann. Statist.* **43** 382–421. MR3311864

[4] BIANCHI, P., DEBBAH, M., MAIDA, M. and NAJIM, J. (2011). Performance of statistical tests for single-source detection using random matrix theory. *IEEE Trans. Inform. Theory* **57** 2400–2419. MR2809098

[5] BICKEL, P. J. and SARKAR, P. (2016). Hypothesis testing for automated community detection in networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 253–273. MR3453655

[6] BOURGADE, P., ERDÖS, L. and YAU, H.-T. (2014). Edge universality of beta ensembles. *Comm. Math. Phys.* **332** 261–353. MR3253704

[7] BOURGADE, P., HUANG, J. and YAU, H.-T. (2017). Eigenvector statistics of sparse random matrices. *Electron. J. Probab.* **22** Paper No. 64. MR3690289

[8] CHOI, Y., TAYLOR, J. and TIBSHIRANI, R. (2017). Selecting the number of principal components: Estimation of the true rank of a noisy matrix. *Ann. Statist.* **45** 2590–2617. MR3737903

[9] DING, X. and YANG, F. (2018). A necessary and sufficient condition for edge universality at the largest singular values of covariance matrices. *Ann. Appl. Probab.* **28** 1679–1738. MR3809475

[10] EL KAROUI, N. (2007). Tracy–Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *Ann. Probab.* **35** 663–714. MR2308592

[11] ERDŐS, L., KNOWLES, A., YAU, H.-T. and YIN, J. (2012). Spectral statistics of Erdős–Rényi graphs II: Eigenvalue spacing and the extreme eigenvalues. *Comm. Math. Phys.* **314** 587–640. MR2964770

[12] ERDŐS, L., KNOWLES, A., YAU, H.-T. and YIN, J. (2013). Spectral statistics of Erdős–Rényi graphs I: Local semicircle law. *Ann. Probab.* **41** 2279–2375. MR3098073

[13] ERDŐS, L., YAU, H.-T. and YIN, J. (2012). Rigidity of eigenvalues of generalized Wigner matrices. *Adv. Math.* **229** 1435–1515. MR2871147

[14] FELDMAN, V., PERKINS, W. and VEMPALA, S. (2015). Subsampled power iteration: A unified algorithm for block models and planted CSP's. In *Advances in Neural Information Processing Systems* 2836–2844.

[15] FLORESCU, L. and PERKINS, W. (2016). Spectral thresholds in the bipartite stochastic block model. In *Conference on Learning Theory* 943–959.

[16] GIRKO, V. L. (1984). The circular law. *Teor. Veroyatn. Primen.* **29** 669–679. MR0773436

[17] HUANG, J., LANDON, B. and YAU, H.-T. Transition from Tracy–Widom to Gaussian fluctuations of extremal eigenvalues of sparse Erdős–Rényi graphs. Available at arXiv:1712.03936.

[18] HWANG, J. Y., LEE, J. O. and SCHNELLI, K. (2019). Supplement to "Local law and Tracy–Widom limit for sparse sample covariance matrices." DOI:10.1214/19-AAP1472SUPP.

[19] JOHANSSON, K. (2000). Shape fluctuations and random matrices. *Comm. Math. Phys.* **209** 437–476. MR1737991

[20] JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961

[21] KNOWLES, A. and YIN, J. (2017). Anisotropic local laws for random matrices. *Probab. Theory Related Fields* **169** 257–352. MR3704770

[22] KRITCHMAN, S. and NADLER, B. (2008). Determining the number of components in a factor model from limited noisy data. *Chemom. Intell. Lab. Syst.* **94** 19–32.

[23] LEE, J. O. and SCHNELLI, K. (2015). Edge universality for deformed Wigner matrices. *Rev. Math. Phys.* **27** 1550018. MR3405746

[24] LEE, J. O. and SCHNELLI, K. (2016). Tracy–Widom distribution for the largest eigenvalue of real sample covariance matrices with general population. *Ann. Appl. Probab.* **26** 3786–3839. MR3582818

[25] LEE, J. O. and SCHNELLI, K. (2018). Local law and Tracy–Widom limit for sparse random matrices. *Probab. Theory Related Fields* **171** 543–616. MR3800840

[26] LEI, J. (2016). A goodness-of-fit test for stochastic block models. *Ann. Statist.* **44** 401–424. MR3449773

[27] LYTOVA, A. and PASTUR, L. (2009). Central limit theorem for linear eigenvalue statistics of random matrices with independent entries. *Ann. Probab.* **37** 1778–1840. MR2561434

[28] MARČENKO, V. A. and PASTUR, L. A. (1967). Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb.* **72** 507–536. MR0208649

[29] NADAKUDITI, R. R. and EDELMAN, A. (2008). Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples. *IEEE Trans*. *Signal Process*. **56** 2625–2638. MR1500236

[30] ONATSKI, A. (2009). Testing hypotheses about the numbers of factors in large factor models. *Econometrica* **77** 1447–1479. MR2561070

[31] PILLAI, N. S. and YIN, J. (2014). Universality of covariance matrices. *Ann*. *Appl*. *Probab.* **24** 935–1001. MR3199978

[32] QUANDT, R. E. (1972). A new approach to estimating switching regressions. *J. Amer. Statist. Assoc*. **67** 306–310.

[33] STEIN, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann*. *Statist*. **9** 1135–1151. MR0630098

[34] TIMMERMANN, A. (2000). Moments of Markov switching models. *J. Econometrics* **96** 75–111. MR1768756

J. Y. HWANG
J. O. LEE
DEPARTMENT OF MATHEMATICAL
    SCIENCES
KAIST
DAEJEON 34141
KOREA
E-MAIL: jyh9006@kaist.ac.kr
        jioon.lee@kaist.edu

K. SCHNELLI
DEPARTMENT OF MATHEMATICS
KTH ROYAL INSTITUTE OF TECHNOLOGY
STOCKHOLM SE-100 44
SWEDEN
E-MAIL: schnelli@kth.se