Probab. Theory Relat. Fields (2017) 169:667–727
DOI 10.1007/s00440-016-0740-2

Universality for general Wigner-type matrices

Oskari H. Ajanki[1] · **László Erdős**[1] ·
Torben Krüger[1]

Received: 9 October 2015 / Revised: 5 January 2016 / Published online: 19 September 2016

Abstract We consider the local eigenvalue distribution of large self-adjoint $N \times N$ random matrices $\mathbf{H} = \mathbf{H}^*$ with centered independent entries. In contrast to previous works the matrix of variances $s_{ij} = \mathbb{E}|h_{ij}|^2$ is not assumed to be stochastic. Hence the density of states is not the Wigner semicircle law. Its possible shapes are described in the companion paper (Ajanki et al. in Quadratic Vector Equations on the Complex Upper Half Plane. arXiv:1506.05095). We show that as N grows, the resolvent, $\mathbf{G}(z) = (\mathbf{H} - z)^{-1}$, converges to a diagonal matrix, $\mathrm{diag}(\mathbf{m}(z))$, where $\mathbf{m}(z) = (m_1(z), \ldots, m_N(z))$ solves the vector equation $-1/m_i(z) = z + \sum_j s_{ij} m_j(z)$ that has been analyzed in Ajanki et al. (Quadratic Vector Equations on the Complex Upper Half Plane. arXiv:1506.05095). We prove a local law down to the smallest spectral resolution scale, and bulk universality for both real symmetric and complex hermitian symmetry classes.

Oskari H. Ajanki: Partially supported by ERC Advanced Grant RANMAT No. 338804, and SFB-TR 12 Grant of the German Research Council.
László Erdős: Partially supported by ERC Advanced Grant RANMAT No. 338804.
Torben Krüger: Partially supported by ERC Advanced Grant RANMAT No. 338804, and SFB-TR 12 Grant of the German Research Council.

✉ László Erdős
lerdos@ist.ac.at

Oskari H. Ajanki
oskari.ajanki@iki.fi

Torben Krüger
torben.krueger@ist.ac.at

[1] IST Austria, Klosterneuburg, Austria

Keywords Eigenvector delocalization · Rigidity · Anisotropic local law · Local spectral statistics

Mathematics Subject Classification 60B20 · 15B52

Contents

1 Introduction

In the seminal paper [31] Wigner introduced random self-adjoint matrices, $\mathbf{H} = \mathbf{H}^*$, with centered, identically distributed and independent entries (subject to the symmetry constraint). He proved that the empirical density of the eigenvalues converges to the semicircle distribution. Wigner also conjectured that the distribution of the distance between consecutive eigenvalues (*gap statistics*) is universal, hence it is the same as in the Gaussian model. His revolutionary observation was that these universality phenomena hold for much larger classes of physical systems and only the basic symmetry type determines local spectral statistics. It is generally believed, but mathematically unproven, that random matrix theory (RMT), among many other examples, also describes the local statistics of random Schrödinger operators in the delocalized regime and quantization of chaotic classical Hamiltonians.

The first rigorous results on the local eigenvalue statistics in the bulk spectrum were given by Dyson, Mehta and Gaudin in the 60's. These concerned the Gaussian models and identified their local correlation functions. According to Wigner's universality hypothesis, these statistics should hold independently of the common law of the matrix elements. This conjecture was resolved recently in a series of works. The strongest result on Wigner matrices in the bulk spectrum is Theorem 7.2 in [13], see [19,30] for a summary of the history and related results. In fact, the *three-step* approach developed in [14,17,20] also applies for *generalized Wigner matrices* that allow for non-identically distributed matrix elements as long as the variance matrix $s_{ij} := \mathbb{E}|h_{ij}|^2$ is stochastic, i.e. $\sum_j s_{ij} = 1$ (in particular, independent of the index i). The stochasticity of $\mathbf{S}$ guarantees that the eigenvalue density is given by the semicircle law and the diagonal elements $G_{ii} = G_{ii}(z)$ of the resolvent

$$\mathbf{G}(z) = (\mathbf{H} - z)^{-1}, \quad \operatorname{Im} z > 0, \tag{1.1}$$

become not only deterministic but also independent of i as the the matrix size N goes to infinity. They asymptotically satisfy a system of self-consistent equations

$$-\frac{1}{G_{ii}} \approx z + \sum_j s_{ij} G_{jj}, \tag{1.2}$$

that reduces to a particularly simple scalar equation

$$-\frac{1}{m} = z + m, \tag{1.3}$$

for the common value $m \approx G_{ii}$ for all i as $N \to \infty$. The solution $m = m(z)$ of (1.3) is the Stieltjes transform of the Wigner semicircle law.

In this paper we consider a general variance matrix $\mathbf{S}$ without stochasticity condition. We show that the approximate self-consistent Eq. (1.2) still holds, but it does not simplify to a scalar equation. In fact, G_{ii} remains i-dependent even as $N \to \infty$ and it is close to the solution m_i of the *Quadratic Vector Equation* (QVE)

$$-\frac{1}{m_i} = z + \sum_j s_{ij} m_j, \tag{1.4}$$

under the additional condition that $\operatorname{Im} m_i > 0$.

In the context of random matrices importance of this equation has been realized by Girko [23], Shlyakhtenko [29], Khorunzhy and Pastur [25], see also Guionnet [24], as well as Anderson and Zeitouni [5,6], but no detailed study has been initiated. In the companion paper [1] we analyzed (1.4) in full detail. See also Section 3 of [2] for how the QVE is related to other random matrix models. We showed that $\langle m \rangle := N^{-1} \sum_i m_i$ is the Stieltjes transform of a probability density ρ that is supported on a finite number of intervals, inside of which it is a real analytic function. We also described the behavior of ρ near the edges of its support; it features only square root or cubic root (cusp) singularities and an explicit one parameter family of profiles interpolating between them as a gap in the support closes.

The main result of the current paper is the universality of the local eigenvalue statistics in the bulk for Wigner-type matrices with a general variance matrix (cf. Theorem 1.16). This extends Wigner's vision towards full universality by considering a much larger class of matrix ensembles than previously studied. In particular, we demonstrate that local statistics, as expected, are fully independent of the global density. This fact has already been established for very general β-ensembles in [10] (see also [8,28]) and for additively deformed Wigner ensembles having a density with a single interval support [27]. Our class admits a general variance matrix and allows for densities with several intervals (we do not, however, consider non-centered distributions here; an extension to matrices with non-centered entries on the diagonal may be incorporated in our analysis with additional technical effort).

Following the three-step approach, we first prove *local laws* for $\mathbf{G}$ on the scale $\eta = \operatorname{Im} z \gg N^{-1}$, i.e. down to the optimal scale just slightly above the eigenvalue spacing. This is the main and novel part of our analysis. The previous proofs (see [14] for a pedagogical presentation) heavily relied on properties of the semicircle law, especially on its square root edge singularity. With possible cubic root singularities and small gaps in the support of ρ an additional scale appears which needs to be controlled. The second step is to prove universality for Wigner-type matrices with a tiny Gaussian component via *Dyson Brownian motion* (DBM). The method of *local relaxation flow*, introduced first in [16,17], also heavily relies on the semicircle law since it requires that the global density remain unchanged along DBM. In [18] a new method was developed to localize the DBM that proves universality of the gap statistics around a fixed energy τ in the bulk, assuming that the local law holds near τ (we remark that a similar result was obtained independently in [26]). Since Wigner-type matrices were one of the main motivations for [18], it was formulated such that it could be directly applied once the local laws are available. Finally, the third step is a perturbation result to remove the tiny Gaussian component using the *Green function comparison* method that first appeared in [20] and can be applied to our case basically without any modifications.

In a separate paper [3] we apply the results of this work and [1] to treat Gaussian random matrices with correlated entries. Assuming translation invariance of the correlation structure in these Gaussian matrix ensembles we prove an optimal local law, bulk universality and non-trivial decay of off-diagonal resolvent entries.

1.1 Set-up and main results

Let $\mathbf{H}^{(N)} \in \mathbb{C}^{N\times N}$ be a sequence of self-adjoint random matrices. In particular, if the entries of $\mathbf{H}$ are real then $\mathbf{H}^{(N)}$ is symmetric. The matrix ensemble $\mathbf{H} = \mathbf{H}^{(N)}$ is said to be of **Wigner-type** if its entries h_{ij} are independent for $i \le j$ and centered, i.e.,

$$\mathbb{E}h_{ij} = 0 \quad \text{for all} \quad i, j = 1, \dots, N. \tag{1.5}$$

The dependence of $\mathbf{H}$ and other quantities on the dimension N will be suppressed in our notation. The matrix of variances, $\mathbf{S} = (s_{ij})_{i,j=1}^{N}$, is defined through

$$s_{ij} := \mathbb{E}|h_{ij}|^2. \tag{1.6}$$

It is symmetric with non-negative entries. In [1] it was shown that for every such matrix $\mathbf{S}$ the **quadratic vector equation** (QVE),

$$-\frac{1}{m_i(z)} = z + \sum_{j=1}^{N} s_{ij} m_j(z), \quad \text{for all} \quad i = 1, \dots, N \text{ and } z \in \mathbb{H}, \tag{1.7}$$

for a function $\mathbf{m} = (m_1, \ldots, m_N) : \mathbb{H} \to \mathbb{H}^N$ on the complex upper half plane, $\mathbb{H} = \{z \in \mathbb{C} : \operatorname{Im} z > 0\}$, has a unique solution. The main result of this paper is to establish the local law for Wigner-type matrices, i.e. that for large N the resolvent, $\mathbf{G}(z) = (\mathbf{H} - z)^{-1}$, with spectral parameter $z = \tau + i\eta \in \mathbb{H}$, is close to the diagonal matrix, $\operatorname{diag}(\mathbf{m}(z))$, as long as $\eta \gg N^{-1}$. As a consequence, we obtain rigidity estimates on the eigenvalues and complete delocalization for the eigenvectors. Combining this information with the Dyson–Brownian motion, we obtain universality of the eigenvalue gap statistics in the bulk.

We now list the assumptions on the variance matrices $\mathbf{S} = \mathbf{S}^{(N)}$. The restrictions on $\mathbf{S}$ are controlled by three **model parameters**, $p, P > 0$ and $L \in \mathbb{N}$, which do not depend on N. These parameters will remain fixed throughout this paper.

(A) For all N the matrix $\mathbf{S}$ is **flat**, i.e.,

$$s_{ij} \;\le\; \frac{1}{N}, \quad i, j = 1, \ldots, N. \tag{1.8}$$

(B) For all N the matrix $\mathbf{S}$ is **uniformly primitive**, i.e.,

$$(\mathbf{S}^L)_{ij} \;\ge\; \frac{p}{N}, \quad i, j = 1, \ldots, N. \tag{1.9}$$

(C) For all N the matrix $\mathbf{S}$ induces a **bounded solution** of the QVE, i.e., the unique solution $\mathbf{m}$ of (1.7) corresponding to $\mathbf{S}$ is bounded,

$$|m_i(z)| \;\le\; P, \quad i = 1, \ldots, N, \quad z \in \mathbb{H}. \tag{1.10}$$

Remark 1.1 (Boundedness and normalization) The assumption on the boundedness of $\mathbf{m}$ is an implicit condition in the sense that it can be checked only after solving (1.7). In Theorem 6.1 of [1] we list sufficient, explicitly checkable conditions on $\mathbf{S}$, which ensure (1.10). We also remark that the assumption (1.8) can be replaced by $s_{ij} \le C/N$ for some positive constant C. This will lead to a rescaling (cf. Remark 2.2 of [1]) of $\mathbf{m}$. We pick the normalization $C = 1$ just for convenience.

Remark 1.2 (Primitivity) The primitivity condition (1.9) excludes some important models, e.g. matrices of the form

$$\mathbf{H} = \begin{pmatrix} \mathbf{0} & \mathbf{X} \\ \mathbf{X}^* & \mathbf{0} \end{pmatrix},$$

whose eigenvalues yield the singular values of the Gram matrix $\mathbf{X}\mathbf{X}^*$, where $\mathbf{X}$ has independent centered entries with an arbitrary variance profile. Condition (B) is not a mere technicality; Gram matrices may have singularities in the spectrum near 0 (often referred to as the ‘hard-edge’) that require separate treatment; but even away from 0 some new ideas are needed. The complete analysis is presented in [4], where we prove local laws for Gram matrices.

In addition to the assumptions on the variances of h_{ij}, we also require uniform boundedness of higher moments. This leads to another basic model parameter, $\underline{\mu} = (\mu_1, \mu_2, \dots)$, which is a sequence of non-negative real numbers.

(D) For all N the entries h_{ij} of the random matrix $\mathbf{H}$ have **bounded moments**,

$$\mathbb{E}|h_{ij}|^k \;\le\; \mu_k s_{ij}^{k/2}, \quad k \in \mathbb{N},\; i, j = 1, \dots, N. \tag{1.11}$$

In order to state our main result, in the next corollary we collect a few facts about the solution of the QVE that are proven in [1]. Although these properties are sufficient for the formulation of our results, for their proofs we will need much more precise information about the solution of the QVE. Theorems 4.1 and 4.2 summarize everything that is needed from [1] besides the existence and uniqueness of the solution of the QVE. In particular, the statement of Corollary 1.3 follows easily from Theorem 4.1 below.

Corollary 1.3 (Solution of QVE) *Suppose* $\mathbf{S}$ *satisfies* (A)–(C). *Let* $\mathbf{m} : \mathbb{H} \to \mathbb{H}^N$ *be the solution the QVE* (1.7) *corresponding to* $\mathbf{S}$. *Then* $\mathbf{m}$ *is analytic and has a continuous extension (denoted again by* $\mathbf{m}$*) to the closed upper half plane,* $\mathbf{m} : \overline{\mathbb{H}} \to \overline{\mathbb{H}}^N$*, with* $\overline{\mathbb{H}} := \mathbb{H} \cup \mathbb{R}$. *The function* $\rho : \mathbb{R} \to [0, \infty)$*, defined by*

$$\rho(\tau) \;:=\; \lim_{\eta \downarrow 0} \frac{1}{\pi N} \sum_{i=1}^{N} \mathrm{Im}\, m_i(\tau + \mathrm{i}\eta), \tag{1.12}$$

is a probability density. Its support is contained in $[-2, 2]$ *and is a union of closed disjoint intervals*

$$\operatorname{supp}\rho \;=\; \bigcup_{k=1}^{K} [\alpha_k, \beta_k], \quad \textit{where} \quad \alpha_k < \beta_k < \alpha_{k+1}. \tag{1.13}$$

There exists a positive constant δ_**, depending only on the model parameters p, P and L, such that the sizes of these intervals are bounded from below by*

$$\beta_k - \alpha_k \;\ge\; 2\delta_*. \tag{1.14}$$

Note that (1.14) provides a lower bound on the length of the intervals that constitute $\operatorname{supp}\rho$, while the length of the gaps, $\alpha_{k+1} - \beta_k$, between neighboring intervals can be arbitrarily small. Figure 1 shows a shape that the density of states typically might have. In particular, ρ may have gaps in its support and may have additional zeros (cusps) in the interior of $\operatorname{supp}\rho$. However, the behavior of ρ on the domain $\rho \le \varepsilon$, for some sufficiently small $\varepsilon > 0$, can be completely characterized by universal shape functions. For more details see Theorem 2.6 of [1].

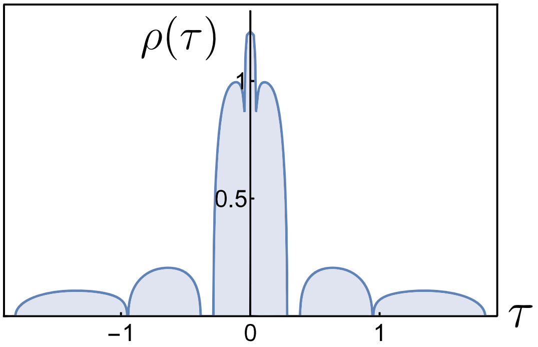

Fig. 1 The density of states may have gaps, cusps and local minima. It is always a symmetric function around zero, i.e., $\rho(\tau) = \rho(-\tau)$

Definition 1.4 (*Density of states*) The function ρ defined in (1.12) is called the **density of states**. Its harmonic extension to the upper half plane

$$\begin{aligned} \rho(\tau + \mathrm{i}\eta) &:= \int_{\mathbb{R}} \Pi_\eta(\tau - \sigma)\rho(\sigma)\mathrm{d}\sigma, \\ \Pi_\eta(\tau) &:= \frac{1}{\pi}\frac{\eta}{\tau^2 + \eta^2}; \quad \tau \in \mathbb{R}, \quad \eta > 0, \end{aligned} \tag{1.15}$$

is again denoted by ρ. With a slight abuse of notation we still write supp ρ, as in (1.13), for the support of the density of states as a function on the real line.

The density of states will be shown to be the eigenvalue distribution of $\mathbf{H}$ in the large N limit on the macroscopic scale. For any fixed values $\tau_1, \tau_2 \in \mathbb{R}$ with $\tau_1 < \tau_2$ it satisfies

$$\lim_{N\to\infty} \frac{\left|\,\mathrm{Spec}(\mathbf{H}^{(N)}) \cap [\tau_1, \tau_2]\,\right|}{N\int_{\tau_1}^{\tau_2} \rho^{(N)}(\tau)\,\mathrm{d}\tau} = 1, \tag{1.16}$$

provided the denominator does not vanish in the limit. The identity (1.16) motivates the terminology of density of states for the function ρ. The harmonic extension of ρ to $\mathbb{H}$ is a version of the density of states, that is smoothed out on the scale η. It satisfies the identity $\rho(z) = (\pi N)^{-1}\sum_k \mathrm{Im} m_k(z)$ not just for $z \in \mathbb{R}$ (cf. (1.12)) but for all $z \in \overline{\mathbb{H}}$ and it will be used in the statement of our main result, Theorem 1.7.

In fact, Theorem 1.7, implies a local version of (1.16), where the length of the interval, $[\tau_1, \tau_2]$, may converge to zero as N tends to infinity. Our estimates and thus the proven speed of convergence depend on the distance of the interval to the edges of supp ρ and even on the length of the closest gap in this case. We introduce a function $\Delta : \mathbb{R} \to [0, \infty)$, which encodes this relation.

Definition 1.5 (*Local gap size*) Let α_k and β_k be the **edges** of the support of the density of states (cf. (1.13)) and δ_* the constant introduced in Corollary 1.3. Then for any $\delta \in [0, \delta_*)$ we set

$$\Delta_\delta(\tau) := \begin{cases} \alpha_{k+1}-\beta_k & \text{if } \beta_k-\delta \le \tau \le \alpha_{k+1}+\delta \text{ for some } k=1,\dots,K-1, \\ 1 & \text{if } \tau \le \alpha_1+\delta \text{ or } \tau \ge \beta_K-\delta, \\ 0 & \text{otherwise.} \end{cases} \tag{1.17}$$

Finally, we fix an arbitrarily small **tolerance exponent** $\gamma \in (0,1)$. This number will stay fixed throughout this paper in the same fashion as the model parameters P, p, L and $\underline{\mu}$. Our main result is stated for spectral parameters $z = \tau + \mathrm{i}\eta$ whose imaginary parts satisfy

$$\eta \ge N^{\gamma-1}. \tag{1.18}$$

For a compact statement of the main theorem we define the notion of stochastic domination, introduced in [12,14]. This notion is designed to compare sequences of random variables in the large N limit up to small powers of N on high probability sets.

Definition 1.6 (*Stochastic domination*) Suppose $N_0 : (0,\infty)^2 \to \mathbb{N}$ is a given function, depending only on the model parameters p, P, L and $\underline{\mu}$, as well as on the tolerance exponent γ. For two sequences, $\varphi = (\varphi^{(N)})_N$ and $\psi = (\psi^{(N)})_N$, of non-negative random variables we say that φ is **stochastically dominated** by ψ if for all $\varepsilon > 0$ and $D > 0$,

$$\mathbb{P}\Big(\varphi^{(N)} > N^\varepsilon \psi^{(N)}\Big) \le N^{-D}, \quad N \ge N_0(\varepsilon, D). \tag{1.19}$$

In this case we write $\varphi \prec \psi$.

Basic properties of the stochastic domination that are used extensively in this paper are listed in Lemma A.1. The threshold $N_0(\varepsilon, D) = N_0(\varepsilon, D; P, p, L, \underline{\mu}, \gamma)$ will always be an explicit function whose value will be increased throughout the paper, though we will not follow its form. This will happen only finitely many times, ensuring that N_0 stays finite. The threshold is uniform in all other parameters, e.g. in the spectral parameter z, as well as in the indices $i, j, \dots$ of the matrix entries, that the sequences φ and ψ may depend on. Typically, we will not mention the existence of N_0 - it is implicit in the notation $\varphi \prec \psi$. As an example, we see that the bounded moment condition, (D), implies

$$|h_{ij}| \prec N^{-1/2}.$$

Actually, the function N_0 depends only on finitely many moment parameters $(\mu_1, \dots, \mu_M)$ instead of the entire sequence $\underline{\mu}$, where now the number of required moments $M = M(\varepsilon, D; P, p, L, \gamma)$, is an explicit function.

Now we are ready to state our main result on the local law. Suppose $\mathbf{H} = \mathbf{H}^{(N)}$ is a sequence of self-adjoint random matrices with the corresponding sequence $\mathbf{S} = \mathbf{S}^{(N)}$ of variance matrices and $\rho = \rho^{(N)}$ the induced sequence of densities of state. Recall that δ_* is the positive constant, depending only on p, P and L, introduced in Corollary 1.3 and Δ_δ is defined as in Definition 1.5.

Theorem 1.7 (Local law) *Suppose that assumptions* (A)–(D) *are satisfied and fix an arbitrary* $\gamma \in (0,1)$. *There is a deterministic function* $\kappa = \kappa^{(N)} : \mathbb{H} \to (0,\infty]$ *such that uniformly for all* $z = \tau + \mathrm{i}\eta \in \mathbb{H}$ *with* $\eta \ge N^{\gamma-1}$ *the resolvents* (1.1) *of the random matrices* $\mathbf{H} = \mathbf{H}^{(N)}$ *satisfy*

$$\max_{i,j} |G_{ij}(z) - m_i(z)\delta_{ij}| \prec \sqrt{\frac{\rho(z)}{N\eta}} + \frac{1}{N\eta} + \min\left\{\frac{1}{\sqrt{N\eta}}, \frac{\kappa(z)}{N\eta}\right\}. \tag{1.20}$$

Furthermore, for any sequence of deterministic vectors $\mathbf{w} = \mathbf{w}^{(N)} \in \mathbb{C}^N$ *with* $\max_i |w_i| \le 1$ *the averaged resolvent diagonal has an improved convergence rate,*

$$\left|\frac{1}{N}\sum_{i=1}^{N} \overline{w}_i \left(G_{ii}(z) - m_i(z)\right)\right| \prec \min\left\{\frac{1}{\sqrt{N\eta}}, \frac{\kappa(z)}{N\eta}\right\}. \tag{1.21}$$

In particular, for $w_i = 1$ *this implies*

$$\left|\frac{1}{N}\mathrm{Im}\mathrm{Tr}\,\mathbf{G}(z) - \pi\rho(z)\right| \prec \min\left\{\frac{1}{\sqrt{N\eta}}, \frac{\kappa(z)}{N\eta}\right\}. \tag{1.22}$$

The function κ *may be chosen to be*

$$\kappa(z) = \frac{1}{\Delta(\tau)^{1/3} + \rho(z)}, \tag{1.23}$$

where $\Delta = \Delta_\delta$*, with some* $\delta \in (0, \delta_*)$ *that depends only on the model parameters* p*,* P *and* L*.*

In the regime, where z *is not too close to the support of the density of states in the sense that*

$$(\Delta(\tau)^{1/3} + \rho(z))\,\mathrm{dist}(z, \mathrm{supp}\,\rho) \ge \frac{N^\gamma}{(N\eta)^2}, \tag{1.24}$$

κ *maybe improved to*

$$\kappa(z) = \frac{\eta}{\mathrm{dist}(z, \mathrm{supp}\,\rho)\,(\Delta(\tau)^{1/3} + \rho(z))} + \frac{1}{N\eta\mathrm{dist}(z, \mathrm{supp}\,\rho)^{1/2}(\Delta(\tau)^{1/3} + \rho(z))^{1/2}}. \tag{1.25}$$

The size of $\rho(z)$ is described in (4.5) below. Theorem 1.7 can be localized to a spectral interval $I \subset \mathbb{R}$, i.e., the statements hold for $\mathrm{Re}\, z \in I$ provided (1.10) applies for $z \in I + \mathrm{i}(0, \infty)$. In particular, in the bulk of the spectrum Theorem 1.7 simplifies considerably.

Corollary 1.8 (Local law in the bulk) *Assume* (A), (B) *and* (D) *with* $L = 1$. *Suppose there is a constant* $\rho_* > 0$ *and an interval* $I \subset \mathrm{supp}\,\rho$ *such that* $\rho(\tau) \ge \rho_*$ *for all* $\tau \in I$. *Then uniformly for all* $z = \tau + \mathrm{i}\eta$*, with* $\tau \in I$ *and* $\eta \ge N^{\gamma-1}$*, and non-random* $\mathbf{w} \in \mathbb{C}^N$ *satisfying* $\max_i |w_i| \le 1$*, the local laws hold*

$$\begin{aligned} \max_{i,j=1}^{N} |G_{ij}(z) - m_i(z)\delta_{ij}| &\prec \frac{1}{\sqrt{N\eta}}, \quad \textit{and} \\ \left|\frac{1}{N}\sum_{i=1}^{N} \overline{w}_i \left(G_{ii}(z) - m_i(z)\right)\right| &\prec \frac{1}{N\eta}, \end{aligned} \tag{1.26}$$

where ρ_* *is considered as an additional model parameter.*

Here the additional assumption $L = 1$ is only used to guarantee (cf. (i) of Theorem 6.1 in [1]) that the solution $\mathbf{m}(z)$ of the QVE stays bounded around $z = 0$. Indeed, if z is bounded away from zero then (i) of Lemma 5.4 in [1] implies $\|\mathbf{m}\|_\infty := \max_i |m_i|$ is bounded by a constant independent of N in the bulk of the spectrum. Therefore, if $\operatorname{dist}(I, 0) \geq \delta$ for some $\delta > 0$, or $\sup\{\, \|\mathbf{m}(z)\|_\infty : \operatorname{Re} z \in I\} \leq P$ is known for some $P < \infty$, then the assumption $L = 1$ can be removed, and (1.26) holds with δ or P, respectively, considered as model parameters.

Theorem 1.7 generalizes the previous local laws for stochastic variance matrices $\mathbf{S}$ (see [14] and references therein). It is valid for densities ρ with an edge behavior different from the square root growth that is known from Wigner's semicircular law. In particular, singularities that interpolate between a square root and a cubic root are possible. In the bulk of the support of the density of states, i.e., where ρ is bounded away from zero, the function κ is bounded. The same is true near the edges, unless the nearby gap is small. The bound deteriorates near small gaps in the support of ρ.

In applications, the sequence $\mathbf{S} = \mathbf{S}^{(N)}$ satisfying (A)–(C) may be constructed by discretizing a piecewise 1/2-Hölder continuous limit function (cf. Remark 6.2 in [1]). As a particularly simple example, suppose f is a smooth, non-negative, symmetric, $f(x, y) = f(y, x)$, function on $[0, 1]^2$ with a positive diagonal, $f(x, x) > 0$. Then the sequence of variance matrices,

$$s_{ij}^{(N)} := \frac{1}{N} f\left(\frac{i}{N}, \frac{j}{N}\right), \quad i, j = 1, \ldots, N,$$

satisfies conditions (A)–(C). The validity of (C) can be verified by using the general criteria (cf. Theorem 2.10 and Theorem 6.1 of [1]) for uniform boundedness. In this case the solution, $\mathbf{m} = (m_1, \ldots, m_N)$, of the QVE converges to a limit in the sense that

$$\sup_{z \in \mathbb{H}} \max_{i=1}^{N} \left| m_i(z) - m(i/N; z) \right| \to 0,$$

where $m : [0, 1] \times \overline{\mathbb{H}} \to \overline{\mathbb{H}}$ is the solution of the continuous QVE,

$$-\frac{1}{m(x; z)} = z + \int_0^1 f(x, y) m(y; z) \mathrm{d}y, \quad x \in [0, 1], \ z \in \overline{\mathbb{H}}.$$

The continuous QVE such as this one fall into the class of general QVEs thoroughly analyzed in the companion paper [1]. In particular, the stability analysis applies and the density of states converges to a limit

$$\rho^{(N)}(\tau) \to \frac{1}{\pi} \int_0^1 \operatorname{Im} m(x; \tau) \mathrm{d}x.$$

We introduce a notion for expressing that events hold with high probability in the limit as N tends to infinity.

Definition 1.9 (*Overwhelming probability*) Suppose $N_0 : (0, \infty) \to \mathbb{N}$ is a given function, depending only on the model parameters p, P, L and $\underline{\mu}$, as well as on the tolerance exponent γ. For a sequence $A = (A^{(N)})_N$ of random events we say that A hold **asymptotically with overwhelming probability** (a.w.o.p.), if for all $D > 0$:

$$\mathbb{P}(A^{(N)}) \geq 1 - N^{-D}, \quad N \geq N_0(D). \tag{1.27}$$

There is a simple connection between the notions of stochastic domination and asymptotically overwhelming probability. For two sequences $A = A^{(N)}$ and $B = B^{(N)}$ the statement 'A implies B a.w.o.p.' is equivalent to $\mathbb{1}_A \prec \mathbb{1}_B$, where the threshold N_0, implicit in the stochastic domination, does not depend on ε, i.e., $N_0(\varepsilon, D) = N_0(D)$.

We denote by $\lambda_1 \leq \cdots \leq \lambda_N$ the eigenvalues of the random matrix $\mathbf{H}$. The following corollary shows that the eigenvalue distribution converges to the density of states as N tends to infinity.

Corollary 1.10 (Convergence of cumulative eigenvalue distribution) *Assume* (A)–(D). *Then uniformly for all* $\tau \in \mathbb{R}$ *the cumulative empirical eigenvalue distribution approaches the integrated density of states,*

$$\left| \#\{i : \lambda_i \leq \tau\} - N \int_{-\infty}^{\tau} \rho(\omega) \mathrm{d}\omega \right| \prec \min \left\{ \frac{1}{\Delta(\tau)^{1/3} + \rho(\tau)}, \, N^{1/5} \right\}. \tag{1.28}$$

Furthermore, for an arbitrary tolerance exponent $\gamma \in (0, 1)$ *there are no eigenvalues away from the support of the density of states,*

$$\max_{k=0}^{K} \#\big\{ i : \beta_k + \delta_k < \lambda_i < \alpha_{k+1} - \delta_k \big\} = 0 \quad \textit{a.w.o.p.}, \tag{1.29}$$

where we interpret $\beta_0 := -\infty$, $\alpha_{K+1} := +\infty$ *and* δ_k *is defined as* $\delta_0 := \delta_K := N^{\gamma - 2/3}$, *as well as*

$$\delta_k := \frac{N^{\gamma}}{(\alpha_{k+1} - \beta_k)^{1/3} N^{2/3}}, \quad k = 1, \ldots, K-1. \tag{1.30}$$

Based on (1.16) we define the index, $i(\tau)$, of an eigenvalue that we expect to be located close to the spectral parameter τ by

$$i(\tau) := \left\lceil N \int_{-\infty}^{\tau} \rho(\omega) \mathrm{d}\omega \right\rceil. \tag{1.31}$$

Here, $\lceil \omega \rceil$ denotes the smallest integer that is bigger or equal to ω for any $\omega \in \mathbb{R}$.

Corollary 1.11 (Rigidity of eigenvalues) *Assume* (A)–(D), *and let* $\gamma \in (0, 1)$ *be an arbitrary tolerance exponent. Denote*

$$\varepsilon_k := N^{\gamma} \min \left\{ \frac{1}{N^{3/5}}, \frac{1}{(\alpha_{k+1} - \beta_k)^{1/9} N^{2/3}} \right\}, \quad k = 1, \ldots, K-1, \tag{1.32}$$

and $\varepsilon_0 := \varepsilon_K := N^{\gamma-2/3}$. *Then uniformly for every*

$$\tau \in \bigcup_{k=1}^{K}\big[\alpha_k + \varepsilon_{k-1}, \beta_k - \varepsilon_k\big], \tag{1.33}$$

the eigenvalues satisfy the rigidity

$$|\lambda_{i(\tau)} - \tau| \prec \min\left\{ \frac{1}{(\Delta(\tau)^{1/3} + \rho(\tau))\rho(\tau)N}, \frac{1}{N^{3/5}} \right\}. \tag{1.34}$$

Furthermore, if τ *is close to the extreme edge,* $\tau \in (\alpha_1, \alpha_1 + \varepsilon_0)$ *or* $\tau \in (\beta_K - \varepsilon_K, \beta_K]$, *then*

$$|\lambda_{i(\tau)} - \tau| \prec N^{-2/3}. \tag{1.35}$$

Finally, if $\tau \in (\beta_k - \varepsilon_k, \alpha_{k+1} + \varepsilon_k)$ *for some* $1 \le k \le K-1$, *then the corresponding eigenvalue is close to an internal edge in the sense that*

$$\lambda_{i(\tau)} \in \big[\beta_k - 2\varepsilon_k, \beta_k + \delta_k\big] \cup \big[\alpha_{k+1} - \delta_k, \alpha_{k+1} + 2\varepsilon_k\big] \quad \textit{a.w.o.p.}, \tag{1.36}$$

where δ_k *is defined in* (1.30).

Remark 1.12 (Eigenvalues outside supp ρ) The statements (1.35) and (1.36) are an immediate consequence of (1.34) and (1.29). They simply express the fact that the small number of $\mathcal{O}(N^\varepsilon)$ eigenvalues, very close to the edges, are found in the space that is left for them by the other eigenvalues for which the rigidity statement (1.34) applies. For an illustration see Fig. 2. We also note that results of this type date back to at least [7] (in the sample covariance context).

Theorem 1.13 (Anisotropic law) *Assume* (A)–(D) *and fix arbitrary* $\gamma > 0$. *Then uniformly for all* $z = \tau + \mathrm{i}\eta \in \mathbb{H}$ *with* $\eta \ge N^{\gamma-1}$, *and for any two deterministic* ℓ^2*-unit vectors* $\mathbf{w}, \mathbf{v}$ *we have*

$$\left| \sum_{i,j=1}^{N} \overline{w}_i G_{ij}(z) v_j - \sum_{i=1}^{N} m_i(z)\overline{w}_i v_i \right| \prec \sqrt{\frac{\rho(z)}{N\eta}} + \frac{1}{N\eta} + \min\left\{ \frac{1}{\sqrt{N\eta}}, \frac{\kappa(z)}{N\eta} \right\}, \tag{1.37}$$

where κ *is the function from Theorem* 1.7.

Corollary 1.14 (Delocalization of eigenvectors) *Assume* (A)–(D) *and fix arbitrary* $\gamma > 0$. *Let* $\mathbf{u}^{(i)} \in \mathbb{C}^N$ *be the* ℓ^2*-normalized eigenvector of* $\mathbf{H}$ *corresponding to the eigenvalue* λ_i. *All eigenvectors are delocalized in the sense that for any deterministic unit vector* $\mathbf{b} \in \mathbb{C}^N$ *we have*

$$\big|\mathbf{b} \cdot \mathbf{u}^{(i)}\big| \prec \frac{1}{\sqrt{N}}. \tag{1.38}$$

In particular, the eigenvectors are completely delocalized, i.e., $\|\mathbf{u}^{(i)}\|_\infty = \max_j |u_j^{(i)}| \prec N^{-1/2}$.

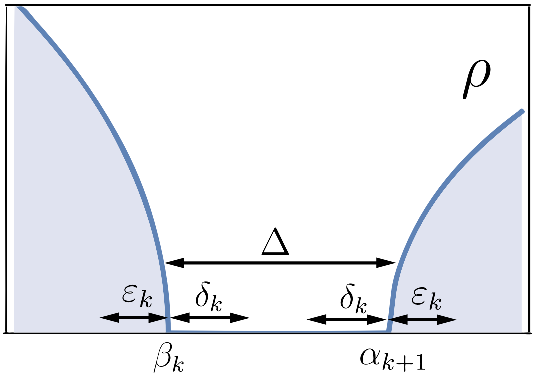

Fig. 2 Notations of Corollary 1.11: At the edges of a gap of length Δ in supp ρ the bound on the eigenvalue fluctuation is δ_k inside the gap and ε_k inside the support

Definition 1.15 (*q-full random matrix*) We say that $\mathbf{H}$ is q**-full** for some $q > 0$ (independent of N) if either of the following applies:

- $\mathbf{H}$ is real symmetric and $\mathbb{E}h_{ij}^2 \geq q/N$ for all $i, j = 1, \ldots, N$;
- $\mathbf{H}$ is complex hermitian and for all $i, j = 1, \ldots, N$ the real symmetric 2×2-matrix,

$$\boldsymbol{\sigma}_{ij} := \begin{pmatrix} \mathbb{E}(\mathrm{Re}h_{ij})^2 & \mathbb{E}(\mathrm{Re}h_{ij})(\mathrm{Im}h_{ij}) \\ \mathbb{E}(\mathrm{Re}h_{ij})(\mathrm{Im}h_{ij}) & \mathbb{E}(\mathrm{Im}h_{ij})^2 \end{pmatrix},$$

is strictly positive definite such that $\boldsymbol{\sigma}_{ij} \geq q/N$.

If $\mathbf{H}$ is real symmetric, then the q-fullness of $\mathbf{H}$ is equivalent to the property (B) with $L = 1$ and $q = p$. On the other hand, in the complex hermitian case the q-fullness condition is stronger than a lower bound on $\mathbb{E}\,|h_{ij}|^2 = s_{ij}$, and it can not be captured by the matrix $\mathbf{S}$ alone.

Theorem 1.16 (Universality) *Suppose* (A) *and* (D) *hold, and* $\mathbf{H}$ *is q-full. Then for all $\varepsilon > 0$, $n \in \mathbb{N}$ and all smooth compactly supported observables $F : \mathbb{R}^n \to \mathbb{R}$, there are two positive constants C and c, depending on ε, q and F in addition to the model parameters, such that for any $\tau \in \mathbb{R}$ with $\rho(\tau) \geq \varepsilon$ the local eigenvalue distribution is universal,*

$$\Big| \mathbb{E}F\left(\left(N\rho(\lambda_{i(\tau)})(\lambda_{i(\tau)} - \lambda_{i(\tau)+j})\right)_{j=1}^{n}\right) \\ - \mathbb{E}_{\mathrm{G}}F\left(\left(N\rho_{\mathrm{sc}}(0)(\lambda_{\lceil N/2\rceil} - \lambda_{\lceil N/2\rceil+j})\right)_{j=1}^{n}\right)\Big| \leq CN^{-c}.$$

Here, $\mathbb{E}_{\mathrm{G}}$ denotes the expectation with respect to the standard Gaussian ensemble, i.e., with respect to GUE and GOE in the cases of complex Hermitian and real symmetric $\mathbf{H}$, *respectively, and $\rho_{\mathrm{sc}}(0) = 1/(2\pi)$ is the value of Wigner's semicircle law at the origin.*

This theorem concerns the universality in the bulk. With the help of our local law one may also prove a weaker version of the universality at the edges (including the internal edges). Since our local law, Theorem 1.7, is optimal at the edges, a direct application of the Green function comparison theorem from Section 6 of [22] (with straightforward adjustments) shows edge universality in the sense that the edge statistics may depend only on the second moments encoded in the matrices $\boldsymbol{\sigma}_{ij}$. In particular, it is the same as the edge statistics of a Wigner-type matrix with centered Gaussian entries with coinciding second moments. This argument holds for the extreme edges as well as for the internal edges. However, it does not yet prove the *Tracy-Widom law*, i.e. that the edge statistics is independent even of the variances $\mathbf{S}$.

Convention 1.17 (Constants and comparison relation) *We use the convention that every positive constant with a lower star index, such as δ_*, c_* and λ_*, explicitly depends only on the model parameters P, p and L from* (B)–(D). *These dependencies can be reconstructed from the proofs, but we will not follow them. Constants $c, c_1, c_2, \ldots, C, C_1, C_2, \ldots$ also depend only on P, p and L. They will have a local meaning within a specific proof.*

For two non-negative functions φ and ψ depending on a set of parameters $u \in U$, we use the **comparison relation**

$$\varphi \gtrsim \psi, \tag{1.39}$$

if there exists a positive constant c, depending explicitly on P, p and L such that $\varphi(u) \geq c\psi(u)$ for all $u \in U$. The notation $\psi \sim \varphi$ means that both $\psi \lesssim \varphi$ and $\psi \gtrsim \varphi$ hold true. In this case we say that ψ and φ are **comparable**. *We also write $\psi = \varphi + \mathcal{O}(\vartheta)$, if $|\psi - \varphi| \lesssim \vartheta$.*

We denote the normalized scalar product between two vectors $\mathbf{u}, \mathbf{w} \in \mathbb{C}^N$ and the average of a vector by

$$\langle \mathbf{u}, \mathbf{w} \rangle := \frac{1}{N} \sum_{i=1}^{N} \overline{u}_i w_i, \quad \text{and} \quad \langle \mathbf{w} \rangle := \frac{1}{N} \sum_{i=1}^{N} w_i, \tag{1.40}$$

respectively. Note that with this convention $|\langle \mathbf{u}, \mathbf{u} \rangle| = N^{-1} \|\mathbf{u}\|_{\ell^2}^2$.

2 Bound on the random perturbation of the QVE

We will make the following standing assumptions for the rest of this paper,

- *The assumptions* (A)–(D) *hold true and an arbitrary tolerance exponent $\gamma \in (0, 1)$ is fixed;*

which are always assumed to hold unless explicitly otherwise stated.

We introduce the notation $\mathbf{G}^{(V)}$ for the resolvent of the matrix $\mathbf{H}^{(V)}$, which is identical to $\mathbf{H}$ except for the removal of the rows and columns corresponding to the indices $V \subseteq \{1, \ldots, N\}$. The enumeration of the indices is kept, even though $\mathbf{G}^{(V)}$ has a lower dimension.

The diagonal elements of the resolvent, $\mathbf{g} := (G_{11}, \ldots, G_{NN})$, satisfy the perturbed quadratic vector equation

$$-\frac{1}{g_i(z)} = z + \sum_{j=1}^{N} s_{ij} g_j(z) + d_i(z), \tag{2.1}$$

for all $z \in \mathbb{H}$ and $i = 1, \ldots, N$. The random perturbation $\mathbf{d} = (d_1, \ldots, d_N)$ is given by

$$\begin{aligned} d_k \;:=\; & \sum_{i \neq j}^{(k)} h_{ki} G_{ij}^{(k)} h_{jk} \;+\; \sum_{i}^{(k)} (|h_{ki}|^2 - s_{ki}) G_{ii}^{(k)} \\ & - \sum_{i}^{(k)} s_{ki} \frac{G_{ik} G_{ki}}{g_k} \;-\; h_{kk} \;-\; s_{kk} g_k. \end{aligned} \tag{2.2}$$

Here and in the following, the upper indices on the sums indicate which indices are not summed over. For the proof of this simple identity as well as (2.3) below via the Schur complement formula we refer to [14]. As in (2.2) we will often omit the dependence on the spectral parameter z in our notation, i.e., $G_{ij} = G_{ij}(z)$, $d_k = d_k(z)$, etc.

We will now derive an upper bound on $\|\mathbf{d}\|_\infty = \max_i |d_i|$, provided $|g_i - m_i|$ is bounded by a small constant. At the same time we will control the off-diagonal elements G_{kl} of the resolvent. These satisfy the identity

$$G_{kl} \;=\; G_{kk} G_{ll}^{(k)} \sum_{i,j}^{(kl)} h_{ki} G_{ij}^{(kl)} h_{jl} \;-\; G_{kk} G_{ll}^{(k)} h_{kl}, \tag{2.3}$$

for $k \neq l$. The strategy in what follows below is that (2.2) and (2.3) are used to improve a rough bound on the entries of the resolvent $\mathbf{G}$ to get the correct bounds on the random perturbation and the off-diagonal resolvent elements. Later, in Sect. 3, the stability of the QVE under the small perturbation, $\mathbf{d}$, will provide the improved bound on the diagonal elements, $G_{ii} - m_i = g_i - m_i$.

We introduce a short notation for the difference between $\mathbf{g}$ and the solution $\mathbf{m}$ of the unperturbed Eq. (1.7),

$$\begin{aligned} \Lambda_{\mathrm{d}}(z) \;&:=\; \max_i |G_{ii}(z) - m_i(z)|, \\ \Lambda_{\mathrm{o}}(z) \;&:=\; \max_{i \neq j} |G_{ij}(z)|, \\ \Lambda(z) \;&:=\; \max\big\{\Lambda_{\mathrm{d}}(z), \Lambda_{\mathrm{o}}(z)\big\}. \end{aligned} \tag{2.4}$$

The following lemma is analogous to Lemma 5.2 in [14] with minor modifications. For the completeness of this paper, we repeat these arguments. One small modification is that our estimates also deal with the regime where $|z|$ is large. To keep the formulas short we denote

$$[z] \;:=\; 1 + |z|.$$

The dependence of the upcoming error bounds on $[z]$ is not always optimal and this dependence is not kept in the statement of our main result Theorem 1.7, either. In fact, the regime $[z] \sim 1$ is the most interesting, since our results show that the spectrum of $\mathbf{H}$ lies a.w.o.p. inside a compact interval (cf. Corollary 1.10). For the first reading we therefore recommend to think of $[z] = 1$ in most of our proofs. The $[z]$-dependence is used mainly in order to propagate a bound from the regime of very large imaginary part of the spectral parameter ($\mathrm{Im} z \geq N^5$) to the entire domain, on which Theorem 1.7 holds.

Lemma 2.1 (Bound on perturbation) *There is a small positive constant $\lambda_* \sim 1$, such that uniformly for all spectral parameters $z = \tau + \mathrm{i}\eta \in \mathbb{H}$ with $\eta \geq N^{\gamma-1}$:*

$$|d_k(z)|\,\mathbb{1}\big(\Lambda(z) \leq \lambda_*/[z]\big) \;\prec\; [z]^{-2}\sqrt{\frac{\mathrm{Im}\langle \mathbf{g}(z)\rangle}{N\eta}} + \frac{1}{\sqrt{N}}, \tag{2.5a}$$

$$\Lambda_{\mathrm{o}}(z)\,\mathbb{1}\big(\Lambda(z) \leq \lambda_*/[z]\big) \;\prec\; [z]^{-2}\left(\sqrt{\frac{\mathrm{Im}\langle \mathbf{g}(z)\rangle}{N\eta}} + \frac{1}{\sqrt{N}}\right). \tag{2.5b}$$

For the proof of this lemma we will need an additional property of the solution of the QVE that is a corollary of Theorem 4.1, where all properties of $\mathbf{m}$ taken from [1] are summarized.

Corollary 2.2 (Bounds on solution) *The absolute value of the solution of the QVE satisfies*

$$|m_i(z)| \sim [z]^{-1}, \quad z \in \mathbb{H}, \; i = 1, \ldots, N. \tag{2.6}$$

Proof of Lemma 2.1 Here we use the three large deviation estimates,

$$\left|\sum_{i \neq j}^{(k)} h_{ki} G_{ij}^{(k)} h_{jk}\right| \prec \left(\sum_{i \neq j}^{(k)} s_{ki} s_{jk} \big|G_{ij}^{(k)}\big|^2\right)^{1/2}, \tag{2.7a}$$

$$\left|\sum_{i,j}^{(kl)} h_{ki} G_{ij}^{(kl)} h_{jl}\right| \prec \left(\sum_{i,j}^{(kl)} s_{ki} s_{jl} \big|G_{ij}^{(kl)}\big|^2\right)^{1/2}, \tag{2.7b}$$

$$\left|\sum_{i}^{(k)} \big(|h_{ki}^2| - s_{ki}\big) G_{ii}^{(k)}\right| \prec \left(\sum_{i}^{(k)} s_{ki}^2 \big|G_{ii}^{(k)}\big|^2\right)^{1/2}. \tag{2.7c}$$

Since $\mathbf{G}^{(V)}$ is independent of the rows and columns of $\mathbf{H}$ with indices in V, these estimates follow directly from the large deviation bounds in Appendix C of [14]. Furthermore, we use

$$|h_{ij}| \prec N^{-1/2}, \quad s_{ij} \leq N^{-1}, \tag{2.8}$$

where latter the inequality is just assumption (1.8) and the bound on h_{ij} follows from (1.11). We remark that the stochastic domination in (2.7) and (2.8) is uniform in k, l

and i, j, respectively, i.e., the threshold function N_0 in Definition 1.6 does not depend on i, j, k, l.

We will now show that the removal of a few rows and columns in $\mathbf{H}$ will only have a small effect on the entries of the resolvent. The general resolvent identity,

$$G_{ij} = G_{ij}^{(k)} + \frac{G_{ik}G_{kj}}{G_{kk}}, \quad k \notin \{i, j\}, \tag{2.9}$$

leads to the bound

$$\big|G_{ij}^{(k)} - G_{ij}\big|\, \mathbb{1}\big(\Lambda \le \lambda_*/[z]\big) = \frac{|G_{ik}G_{kj}|}{|g_k|}\, \mathbb{1}\big(\Lambda \le \lambda_*/[z]\big) \lesssim [z]\Lambda_{\mathrm{o}}^2. \tag{2.10}$$

In the inequality we used that $|m_k(z)| \sim [z]^{-1}$ (cf. Corollary 2.2), $|g_k| = |m_k| + \mathcal{O}(\Lambda)$ and that λ_* is chosen to be small enough. We use (2.10) in a similar calculation for $G_{ij}^{(l)}$ and find that on the event where $\Lambda \le \lambda_*/[z]$,

$$\big|G_{ij}^{(kl)} - G_{ij}^{(l)}\big| = \frac{\big|G_{ik}^{(l)}G_{kj}^{(l)}\big|}{\big|G_{kk}^{(l)}\big|} \lesssim \frac{(\,|G_{ik}| + \mathcal{O}([z]\Lambda_{\mathrm{o}}^2)\,)(\,|G_{kj}| + \mathcal{O}([z]\Lambda_{\mathrm{o}}^2)\,)}{|g_k| + \mathcal{O}([z]\Lambda_{\mathrm{o}}^2)}. \tag{2.11}$$

Again using (2.10) and that the denominator of the last expression is comparable to $[z]^{-1}$, we conclude

$$|G_{ij}^{(kl)} - G_{ij}|\, \mathbb{1}\big(\Lambda \le \lambda_*/[z]\big) \lesssim [z]\Lambda_{\mathrm{o}}^2, \tag{2.12}$$

provided λ_* is small. Therefore, we see that it is possible to remove one or two upper indices from G_{ij} for the price of a term, whose size is at most $[z]\Lambda_{\mathrm{o}}^2$.

We have now collected all necessary ingredients and use them to estimate all the terms in (2.2) one by one. We start with the first summand. By (2.7a) we find

$$\Bigg|\sum_{i\neq j}^{(k)} h_{ki}G_{ij}^{(k)}h_{jk}\Bigg|^2 \prec \sum_{i\neq j}^{(k)} s_{ki}s_{jk}\big|G_{ij}^{(k)}\big|^2 \le \frac{1}{N^2}\sum_{i\neq j}^{(k)} \big|G_{ij}^{(k)}\big|^2. \tag{2.13}$$

With the help of (2.10) we remove the upper index from $G_{ij}^{(k)}$ and get

$$\Bigg|\sum_{i\neq j}^{(k)} h_{ki}G_{ij}^{(k)}h_{jk}\Bigg|^2 \mathbb{1}\big(\Lambda \le \lambda_*/[z]\big) \prec \big(\Lambda_{\mathrm{o}}^2 + [z]^2\Lambda_{\mathrm{o}}^4\big)\mathbb{1}\big(\Lambda \le \lambda_*/[z]\big) \lesssim \Lambda_{\mathrm{o}}^2. \tag{2.14}$$

For the second summand in (2.2) we use the large deviation bound for the diagonal, (2.7c), and find that

$$\Bigg|\sum_i^{(k)}(|h_{ki}|^2-s_{ki})G_{ii}^{(k)}\Bigg|^2 \;\prec\; \sum_i^{(k)} s_{ki}^2\big|G_{ii}^{(k)}\big|^2 \;\le\; \frac{1}{N^2}\sum_i^{(k)}\big|G_{ii}^{(k)}\big|^2. \tag{2.15}$$

By removing the upper index again we estimate

$$\big|G_{ii}^{(k)}\big|\,\mathbb{1}\big(\Lambda\le\lambda_*/[z]\big) \;\lesssim\; |m_i|+\Lambda_{\mathrm{d}}+[z]\Lambda_{\mathrm{o}}^2. \tag{2.16}$$

We use this in (2.15) and for sufficiently small λ_* we arrive at

$$\Bigg|\sum_i^{(k)}(|h_{ki}|^2-s_{ki})G_{ii}^{(k)}\Bigg|^2\,\mathbb{1}\big(\Lambda\le\lambda_*/[z]\big) \;\prec\; \frac{1}{[z]^2N}. \tag{2.17}$$

The third summand in (2.2) is estimated directly by

$$\Bigg|\sum_i^{(k)} s_{ki}\frac{G_{ik}G_{ki}}{g_k}\Bigg|\,\mathbb{1}\big(\Lambda\le\lambda_*/[z]\big) \;\le\; \frac{\Lambda_{\mathrm{o}}^2}{|g_k|}\,\mathbb{1}\big(\Lambda\le\lambda_*/[z]\big) \;\lesssim\; \Lambda_{\mathrm{o}}. \tag{2.18}$$

We combine the estimates for the individual terms (2.14), (2.17), (2.18) and (2.8). Altogether we conclude that

$$|d_k|\,\mathbb{1}\big(\Lambda\le\lambda_*/[z]\big) \;\prec\; \Lambda_{\mathrm{o}}(z)+\frac{1}{\sqrt{N}}. \tag{2.19}$$

We will now derive in a similar fashion a stochastic domination bound for the off-diagonal error term Λ_{o}. Afterwards, we will combine the two bounds and infer the claim of the lemma. For the off-diagonal error term we proceed along the same lines as for $|d_k|$, using (2.3) instead of (2.2). For $k\neq l$ we find

$$|G_{kl}|^2 \;\prec\; |g_k|^2\big|G_{ll}^{(k)}\big|^2\left(\frac{1}{N^2}\sum_{i,j}^{(kl)}\big|G_{ij}^{(kl)}\big|^2+\frac{1}{N}\right). \tag{2.20}$$

Here, we applied the large deviation bound (2.7b). Using the Ward identity for the resolvent $\mathbf{G}^{(kl)}$,

$$\sum_j^{(kl)}\big|G_{ij}^{(kl)}\big|^2 = \frac{\operatorname{Im} G_{ii}^{(kl)}}{\eta}, \tag{2.21}$$

and (2.10) for removing the upper index of $G_{ll}^{(k)}$ we get

$$|G_{kl}|^2\,\mathbb{1}\big(\Lambda\le\lambda_*/[z]\big) \;\prec\; [z]^{-4}\left(\frac{1}{N^2\eta}\sum_i^{(kl)}\operatorname{Im}G_{ii}^{(kl)}+\frac{1}{N}\right). \tag{2.22}$$

We remove the upper indices from $G_{ii}^{(kl)}$ and end up with

$$\Lambda_{\mathrm{o}}\, \mathbb{1}\big(\Lambda \le \lambda_*/[z]\big) \;\prec\; [z]^{-2}\left(\sqrt{\frac{\operatorname{Im}\langle \mathbf{g}\rangle}{N\eta}} + \sqrt{\frac{[z]}{N\eta}}\,\Lambda_{\mathrm{o}} + \frac{1}{\sqrt{N}}\right). \tag{2.23}$$

The bound remains true without the summand containing Λ_{o} on the right hand side, since this term can be absorbed into the left hand side, as its coefficient is bounded by $N^{-\gamma/2}$, while on the left Λ_{o} is not multiplied by a small coefficient. Putting (2.19) and (2.23) together yields the desired result (2.5). □

3 Local law away from local minima

In this section we will use the stability of the QVE to establish the main result away from the local minima of the density of states inside its own support, i.e. away from the set

$$\mathbb{M} := \big\{\tau \in \operatorname{supp} \rho : \;\tau \text{ is the location of a local minimum of } \rho\big\}. \tag{3.1}$$

The case where z is close to $\mathbb{M}$ requires a more detailed analysis. This is given is Sect. 4. At the end of this section we will also sketch the proof of Corollary 1.8. In this section we prove the following.

Proposition 3.1 (Local law away from local minima) *Let δ_* be any positive constant, depending only on the model parameters p, P and L. Then, uniformly for all $z = \tau + \mathrm{i}\eta$ with $\eta \ge N^{\gamma-1}$ and* $\operatorname{dist}(z, \mathbb{M}) \ge \delta_*$, *we have*

$$[z]^2\Lambda_{\mathrm{d}}(z) + \|\mathbf{d}(z)\|_\infty \;\prec\; [z]^{-2}\sqrt{\frac{\rho(z)}{N\eta}} + \frac{[z]^{-6}}{N\eta} + \frac{1}{\sqrt{N}}\,, \tag{3.2a}$$

$$\Lambda_{\mathrm{o}}(z) \;\prec\; [z]^{-2}\sqrt{\frac{\rho(z)}{N\eta}} + \frac{[z]^{-4}}{N\eta} + \frac{[z]^{-2}}{\sqrt{N}}. \tag{3.2b}$$

Furthermore, on the same domain, for any sequence of deterministic vectors $\mathbf{w} = \mathbf{w}^{(N)} \in \mathbb{C}^N$ with the uniform bound, $\|\mathbf{w}\|_\infty \le 1$, we have

$$|\langle \mathbf{w}, \mathbf{g}(z) - \mathbf{m}(z)\rangle| \;\prec\; [z]^{-3}\frac{\rho(z)}{N\eta} + \frac{[z]^{-7}}{(N\eta)^2} + \frac{[z]^{-2}}{N}\,. \tag{3.3}$$

This proposition, combined with the properties of ρ given in Theorem 4.1 later, yields the local law (Theorem 1.7) away from the set $\mathbb{M}$. Indeed, using $\rho(z) \gtrsim [z]^{-2}\eta$ (cf. relations (4.5) below) and $\kappa(z) \ge 0$ we see that (3.2) implies (1.20).

In order to see that also the averaged local law (1.21) follows from (3.3) we split the domain $\{z \in \mathbb{H} : \operatorname{dist}(z, \mathbb{M})\} \ge \delta_*$ into three subdomains that are considered separately. To this end, let B_0 and B_1 be the upper bounds on κ from (1.23) and (1.25), respectively.

First we consider the regime $\eta \ge \delta_*/2$. Using $\Delta^{1/3} + \rho \lesssim 1$ we see that $B_0 \gtrsim 1$. Similarly, we get $B_1 \gtrsim \eta[z]^{-1}$. Since $(N\eta)^{-1}B_k$, $k = 0, 1$, are both bigger than the right hand side of (3.3), we obtain (1.21) for $\eta \ge \delta_*/2$.

Now we consider the regime $\eta \le \delta_*/2$, which is split into two cases depending on whether $\mathrm{dist}(\mathrm{Re}\, z, \mathrm{supp}\, \rho) = 0$, or not. In the former case $[z] \lesssim 1$ and $\mathrm{dist}(z, \mathrm{supp}\, \rho) = \eta$, and (4.5a) implies $\rho(\mathrm{Re}\, z) \sim 1$. Feeding these estimates into (1.23) and (1.25) yields $B_0 \sim 1$ and $B_1 \gtrsim 1$. These imply (1.21).

Finally, suppose $\mathrm{dist}(\mathrm{Re}\, z, \mathrm{supp}\, \rho) \ge \delta_*/2$ and $\eta \le \delta_*/2$. In this regime $\Delta \sim 1$ (cf. (1.17)), while (4.5f) implies $\rho \sim \eta\, [z]^{-2}$. Hence, $B_0 \sim 1$ and $B_1 \gtrsim \eta\, [z]^{-1}$, and $(N\eta)^{-1} \min\{B_0, B_1\} \ge [z]^{-1}N^{-1}$. By comparing with the right hand side of (3.3) we conclude that (1.21) applies for all $\mathrm{dist}(z, \mathbb{M}) \ge \delta_*$.

The proof of Proposition 3.1 uses a continuity argument in z. In particular, continuity of the solution of the QVE is needed. The statement of the following corollary is part of the properties of $\mathbf{m}$ listed in Theorem 4.1 below.

Corollary 3.2 (Stieltjes-transform representation) *For every $i = 1, \ldots, N$ there is a probability density $p_i : \mathbb{R} \to [0, \infty)$ with support in $[-2, 2]$ such that m_i is the Stieltjes-transform of this density, i.e.,*

$$m_i(z) = \int_{\mathbb{R}} \frac{p_i(\tau)\mathrm{d}\tau}{\tau - z}, \quad z \in \mathbb{H}. \tag{3.4}$$

The solution of the QVE is uniformly Hölder-continuous,

$$\|\mathbf{m}(z_1) - \mathbf{m}(z_2)\|_\infty \lesssim |z_1 - z_2|^{1/3}, \quad z_1, z_2 \in \overline{\mathbb{H}}. \tag{3.5}$$

Since the solution can be extended to the real line, it is the harmonic extension to the complex upper half plane of its own restriction to the real line. Therefore, $\mathrm{Im} m_i(\tau) = \pi p_i(\tau)$ for $\tau \in \mathbb{R}$. The density of states is the average of the probability densities p_i, i.e., $\rho = \langle \mathbf{p} \rangle$.

Since we will estimate the difference, $\mathbf{g} - \mathbf{m}$, we start by deriving an equation for this quantity. Using the QVE for $\mathbf{m}$ and the perturbed Eq. (2.1) for $\mathbf{g}$ we find

$$\begin{aligned} g_i - m_i &= -\frac{1}{z + (\mathbf{Sg})_i + d_i} + \frac{1}{z + (\mathbf{Sm})_i} \\ &= \frac{(\mathbf{S}(\mathbf{g} - \mathbf{m}))_i + d_i}{(z + (\mathbf{Sg})_i + d_i)(z + (\mathbf{Sm})_i)} \\ &= m_i^2 (\mathbf{S}(\mathbf{g} - \mathbf{m}))_i + m_i (g_i - m_i)(\mathbf{S}(\mathbf{g} - \mathbf{m}))_i + m_i\, g_i\, d_i. \end{aligned} \tag{3.6}$$

Rearranging the terms leads to

$$\begin{aligned} \big((\mathbf{1} - \mathrm{diag}(\mathbf{m})^2\mathbf{S})(\mathbf{g} - \mathbf{m})\big)_i &= m_i (g_i - m_i)(\mathbf{S}(\mathbf{g} - \mathbf{m}))_i \\ &\quad + m_i^2\, d_i + m_i\, (g_i - m_i)\, d_i. \end{aligned} \tag{3.7}$$

In the proof of Proposition 3.1 we will view (3.7) as a quadratic equation for $\mathbf{g} - \mathbf{m}$ and we use its stability to bound Λ_{d} in terms of $\|\mathbf{d}\|_\infty$. We will now demonstrate this effect in the case when z is far away from the support of the density of states.

Lemma 3.3 (Stability far away from support) *For $z \in \mathbb{H}$ with $|z| \geq 10$, we have*

$$\Lambda_{\mathrm{d}}(z)\mathbb{1}(\Lambda_{\mathrm{d}}(z) \leq 4|z|^{-1}) \lesssim |z|^{-2}\|\mathbf{d}(z)\|_\infty. \tag{3.8}$$

Furthermore, there is a matrix valued function $\mathbf{T} : \mathbb{H} \to \mathbb{C}^{N\times N}$, depending only on $\mathbf{S}$ and satisfying the uniform bound $\|\mathbf{T}(z)\|_{\infty\to\infty} \lesssim 1$, such that for all $\mathbf{w} \in \mathbb{C}^N$ and $|z| \geq 10$ the averaged difference between $\mathbf{g}$ and $\mathbf{m}$ satisfies the improved bound

$$\begin{aligned}&\big|\langle \mathbf{w}, \mathbf{g}(z) - \mathbf{m}(z)\rangle\big|\mathbb{1}\big(\Lambda_{\mathrm{d}}(z) \leq 4|z|^{-1}\big)\\ &\quad\lesssim |z|^{-2}\big(\|\mathbf{w}\|_\infty\|\mathbf{d}(z)\|_\infty^2 + |\langle \mathbf{T}(z)\mathbf{w}, \mathbf{d}(z)\rangle|\big).\end{aligned} \tag{3.9}$$

For a matrix $\mathbf{A}$ we denote by $\|\mathbf{A}\|_{\infty\to\infty}$ the operator norm of $\mathbf{w} \mapsto \mathbf{A}\mathbf{w}$ on ℓ^∞.

Proof Since the matrix $\mathbf{S}$ is flat (cf. (1.8)), it satisfies the norm bound $\|\mathbf{S}\|_{\infty\to\infty} \leq 1$.

We also have the trivial bound $|m_i(z)| \leq 1/\mathrm{dist}(z, \operatorname{supp}\rho) \leq 2|z|^{-1} \leq 1/5$ at our disposal. This follows directly from the Stieltjes transform representation (3.4). In particular,

$$\|(\mathbf{1} - \mathrm{diag}(\mathbf{m})^2\mathbf{S})^{-1}\|_{\infty\to\infty} \leq 2, \tag{3.10}$$

from the geometric series. By inverting the matrix $\mathbf{1} - \mathrm{diag}(\mathbf{m})^2\mathbf{S}$ and using the trivial bound on $\mathbf{m}$ in (3.7) we find

$$\Lambda_{\mathrm{d}}(z) \leq 4\Big(|z|^{-1}\Lambda_{\mathrm{d}}(z)^2 + |z|^{-1}\Lambda_{\mathrm{d}}(z)\|\mathbf{d}(z)\|_\infty + 2|z|^{-2}\|\mathbf{d}(z)\|_\infty\Big). \tag{3.11}$$

Using the bound inside the indicator function from (3.8) and $|z| \geq 10$ the assertion (3.8) of the lemma follows.

The bound for the average, (3.9), follows by taking the inverse of $\mathbf{1} - \mathrm{diag}(\mathbf{m})^2\mathbf{S}$ on both sides of (3.7) and using (3.8) and $|m_i| \sim |z|^{-1}$.

For the proof of Proposition 3.1 we use the stability of (3.7) also close to $\operatorname{supp}\rho$. This requires more care and is carried out in detail in [1]. The result of that analysis is Theorem 4.2. Here we will only need the following consequence of that theorem and (4.5a).

Corollary 3.4 (Stability away from minima) *Suppose δ_* is an arbitrary positive constant, depending only on the model parameters p, P and L. Let $\mathbf{d} : \mathbb{H} \to \mathbb{C}^N$, $\mathbf{g} : \mathbb{H} \to (\mathbb{C}\backslash\{0\})^N$ be arbitrary vector valued functions on the complex upper half plane that satisfy*

$$-\frac{1}{g_i(z)} = z + \sum_{j=1}^N s_{ij}g_j(z) + d_i(z), \quad z \in \mathbb{H}. \tag{3.12}$$

There exists a positive constant $\lambda_* \sim 1$, such that the QVE is stable away from $\mathbb{M}$,

$$\begin{aligned}&\|\mathbf{g}(z) - \mathbf{m}(z)\|_\infty\,\mathbb{1}\big(\|\mathbf{g}(z) - \mathbf{m}(z)\|_\infty \leq \lambda_*\big) \lesssim \|\mathbf{d}(z)\|_\infty,\\ &\qquad\qquad z \in \mathbb{H},\ \mathrm{dist}(z, \mathbb{M}) \geq \delta_*.\end{aligned} \tag{3.13}$$

Furthermore, there is a matrix valued function $\mathbf{T} : \mathbb{H} \to \mathbb{C}^{N \times N}$*, depending only on* $\mathbf{S}$ *and satisfying the uniform bound* $\|\mathbf{T}(z)\|_{\infty \to \infty} \lesssim 1$*, such that for all* $\mathbf{w} \in \mathbb{C}^N$*,*

$$|\langle \mathbf{w}, \mathbf{g}(z) - \mathbf{m}(z) \rangle| \mathbb{1}\big(\|\mathbf{g}(z) - \mathbf{m}(z)\|_\infty \le \lambda_*\big) \lesssim \|\mathbf{w}\|_\infty \|\mathbf{d}(z)\|_\infty^2 + |\langle \mathbf{T}(z)\mathbf{w}, \mathbf{d}(z) \rangle|, \tag{3.14}$$

for $z \in \mathbb{H}$ *with* $\operatorname{dist}(z, \mathbb{M}) \ge \delta_*$.

Furthermore, the following fluctuation averaging result is needed. It was first established for generalized Wigner matrices with Bernoulli distributed entries in [21].

Theorem 3.5 (Fluctuation averaging) *For any* $z \in \mathbb{H}$*, with* $\operatorname{Im} z \ge N^{\gamma - 1}$*, and any sequence of deterministic vectors* $\mathbf{w} = \mathbf{w}^{(N)} \in \mathbb{C}^N$ *with the uniform bound,* $\|\mathbf{w}\|_\infty \le 1$ *the following holds true: If* $\Lambda_{\mathrm{o}}(z) \prec \Phi/[z]^2$ *for some deterministic (N-dependent)* $\Phi \le N^{-\gamma/3}$ *and* $\Lambda(z) \prec N^{-\gamma/3}/(1 + |z|)$ *, then*

$$|\langle \mathbf{w}, \mathbf{d}(z) \rangle| \prec [z]^{-1} \Phi^2 + \frac{1}{N}, \tag{3.15}$$

where $\mathbf{d}(z)$ *is defined in* (2.2)*.*

Proof The proof directly follows the one given in [14]. We only mention some minor necessary modifications. Let $Q_k X := X - \mathbb{E}[X|\mathbf{H}^{(k)}]$ be the complementary projection to the conditional expectation of a random variable X given the matrix $\mathbf{H}^{(k)}$, in which the k-th row and column are removed. From the definition of $\mathbf{d}$ in (2.2) and Schur's complement formula in the form,

$$\frac{1}{G_{kk}} = h_{kk} - z - \sum_{i,j}^{(k)} h_{ki} G_{ij}^{(k)} h_{jk}, \tag{3.16}$$

we infer the identity

$$d_k = -Q_k \frac{1}{G_{kk}} - s_{kk} G_{kk} - \sum_{i}^{(k)} s_{ki} \frac{G_{ik} G_{ki}}{G_{kk}}.$$

In particular, we have that a.w.o.p.

$$\left| d_k + Q_k \frac{1}{G_{kk}} \right| \lesssim \frac{[z]^{-1}}{N} + [z] \Lambda_{\mathrm{o}}^2.$$

Thus, proving (3.15) reduces to showing

$$\left| \frac{1}{N} \sum_{k=1}^{N} \overline{w}_k Q_k \frac{1}{G_{kk}} \right| \prec [z]^{-1} \Phi^2 + \frac{1}{N}.$$

In the setting where $\mathbf{H}$ is a generalized Wigner matrix and $|z| \le 10$ this bound is precisely the content of Theorem 4.7 from [14].

The a priori bound used in the proof of that theorem is replaced by

$$\left| Q_k \frac{1}{G_{kk}^{(V)}} \right| \prec \Lambda_{\mathrm{o}} + \frac{1}{\sqrt{N}}, \tag{3.17}$$

for any $V \subseteq \{1, \ldots, N\}$ with N-independent size. This bound is proven in the same way as (2.19). Here, the N_0 hidden in the stochastic domination depends on the size $|V|$ of the index set. Following the proof of Theorem 4.7 given in [14] with (3.17) and tracking the z-dependence,

$$\frac{1}{|G_{kk}^{(V)}(z)|} \prec [z],$$

yields the fluctuation averaging, Theorem 3.5. □

Proof of Proposition 3.1 Let us show first that (3.3) follows directly from (3.2) by applying the fluctuation averaging, Theorem 3.5. Indeed, (3.2) provides a deterministic bound on the off-diagonal error, Λ_{o}, which is needed to apply the fluctuation averaging to the second terms on the right hand sides of (3.14) and (3.9). It also shows that the indicator functions on the left hand sides of (3.14) and (3.9) are a.w.o.p. nonzero. The stability bound (3.9) valid in the large $|z|$ regime is necessary to get the correct $[z]$-factors in (3.3). Thus, (3.3) is proven, provided (3.2) is true.

The proof of (3.2) is split into the consideration of two different regimes. In the first regime the absolute value of z is large, $|z| \geq N^5$. In this case we make use only of weak a priori bounds on the resolvent elements and the entries of $\mathbf{d}$. Together with Lemma 3.3 they will suffice to prove (3.2) in this case. In the second regime, $|z| \leq N^5$, we use a continuity argument. We will establish a gap in the possible values that the continuous function, $z \mapsto [z]\Lambda(z)$, might have. Here, the stability result Corollary 3.4 is used. We use this gap to propagate the bound with the help of Lemma A.2 in the appendix from $|z| = N^5$ to the whole domain where $|z| \leq N^5$, $\eta \geq N^{\gamma-1}$ and we stay away from $\mathbb{M}$.

Regime 1: Let $|z| \geq N^5$. We show that the indicator functions in the statement of Lemma 2.1 are a.w.o.p. not vanishing. We start by showing that the diagonal contribution, Λ_{d}, to Λ is sufficiently small. The reduced resolvent elements for an arbitrary $V \subseteq \{1, \ldots, N\}$ satisfy

$$|G_{ij}^{(V)}(z)| \leq \eta^{-1} \leq N^{1-\gamma}. \tag{3.18}$$

From this and the definition of $\mathbf{d}$ in (2.2) we read off the a priori bound,

$$\|\mathbf{d}(z)\|_\infty \prec N^{2-\gamma}. \tag{3.19}$$

Here, we used the general resolvent identity (2.9) in the form $G_{ik}G_{ki} = g_k(g_i - G_{ii}^{(k)})$. Since $\mathbf{g}$ satisfies the perturbed QVE (2.1) and $|\sum_{j=1}^N s_{ij} g_j(z) + d_i(z)| \prec N^{2-\gamma}$ from (3.19) and (3.18) we conclude that uniformly for $|z| \geq N^2$ we have

$$|g_k(z)| \leq 2|z|^{-1}, \quad \text{a.w.o.p.} \tag{3.20}$$

With the trivial bound $|m_i(z)| \leq 1/\mathrm{dist}(z, \operatorname{supp}\rho)$ on the solution of the QVE we infer that on this domain the indicator function in (3.8) is a.w.o.p. non-zero and therefore uniformly for $|z| \geq N^2$. Lemma 3.3 yields

$$\Lambda_{\mathrm{d}}(z) \lesssim |z|^{-2}\|\mathbf{d}(z)\|_\infty \leq N^{-\gamma/2}|z|^{-1}, \quad \text{a.w.o.p.} \tag{3.21}$$

In the last inequality we have used (3.19) in the form $\|\mathbf{d}\|_\infty \leq N^{-\gamma/2}N^2$ a.w.o.p. (cf. Definitions 1.6 and 1.9) and the extra factor $[z]^{-2}$ on the right hand side of (3.8). Thus, for $|z| \geq N^2$ the diagonal contribution to Λ does not play a role in the indicator function in the statement of Lemma 2.1.

Now we derive a similar bound for the off-diagonal contribution Λ_{o}. Using the resolvent identity (2.9) for $i = j$ again, the bound $|h_{ij}| \prec N^{-1/2}$ on the entries of the random matrix and the a priori bound on the reduced resolvent elements, (3.18), in the expansion formula (2.3) yields

$$|G_{kl}(z)| \prec \big(|g_k(z)g_l(z)| + |G_{kl}(z)G_{lk}(z)|\big)N^{2-\gamma}, \quad |G_{kl}(z)| \prec |g_k(z)|N^{3-\gamma}, \tag{3.22}$$

for $k \neq l$. With the bound (3.20) we conclude that

$$\Lambda_{\mathrm{o}}(z) \prec |z|^{-2}N^{2-\gamma} + |z|^{-1}N^{5-2\gamma}\Lambda_{\mathrm{o}}(z), \quad |z| \geq N^2. \tag{3.23}$$

Thus, $\Lambda_{\mathrm{o}} \prec N^{-3}|z|^{-1}$ on the domain where $|z| \geq N^5$. We conclude that Lemma 2.1 applies in this regime even without the indicator functions in the formulas (2.5). We use the bound from this lemma for the norm of $\mathbf{d}$ and the off-diagonal contribution, Λ_{o}, to Λ, while we use the first inequality in (3.21) for the diagonal contribution, Λ_{d}. In this way, we get

$$\begin{aligned} |z|^2\Lambda_{\mathrm{d}} + \|\mathbf{d}\|_\infty &\prec |z|^{-2}\sqrt{\frac{\rho}{N\eta}} + |z|^{-2}\sqrt{\frac{\Lambda_{\mathrm{d}}}{N\eta}} + \frac{1}{\sqrt{N}}, \\ |z|^2\Lambda_{\mathrm{o}} &\prec \sqrt{\frac{\rho}{N\eta}} + \sqrt{\frac{\Lambda_{\mathrm{d}}}{N\eta}} + \frac{1}{\sqrt{N}}, \end{aligned} \tag{3.24}$$

where we also used $g_k = m_k + \mathcal{O}(\Lambda_{\mathrm{d}})$. Applying the weighted Cauchy-Schwarz inequality, $\sqrt{\alpha\beta} \leq \theta\,\alpha + \theta^{-1}\beta$, we find for any $\varepsilon \in (0, \gamma)$ that the right hand side of the first inequality can be estimated further by

$$|z|^2\Lambda_{\mathrm{d}} + \|\mathbf{d}\|_\infty \prec |z|^{-2}\sqrt{\frac{\rho}{N\eta}} + N^{-\varepsilon}|z|^2\Lambda_{\mathrm{d}} + |z|^{-6}\frac{N^\varepsilon}{N\eta} + \frac{1}{\sqrt{N}}.$$

The term $N^{-\varepsilon}|z|^2\Lambda_{\mathrm{d}}$ can be absorbed into the left hand side and by the definition of the stochastic domination and since ε is arbitrarily small the remaining N^ε on the right hand side can be replaced by 1 without changing the correctness of this bound (cf. (i) and (ii) of Lemma A.1). In this way we arrive at

$$|z|^2\Lambda_{\mathrm{d}} + \|\mathbf{d}\|_\infty \prec |z|^{-2}\sqrt{\frac{\rho}{N\eta}} + \frac{|z|^{-6}}{N\eta} + \frac{1}{\sqrt{N}}.$$

For the bound on the off-diagonal error term we plug this result into (3.24) and get

$$\Lambda_{\mathrm{o}} \prec |z|^{-2}\sqrt{\frac{\rho}{N\eta}} + \frac{|z|^{-6}}{N\eta} + \frac{|z|^{-3}}{N^{1/4}}\sqrt{\frac{1}{N\eta}} + \frac{|z|^{-2}}{\sqrt{N}}\,.$$

Regime 2: Now let $|z| \le N^5$ and suppose that δ_* is a positive constant, depending only on the model parameters p, P and L. The diagonal contribution, Λ_{d}, satisfies

$$\Lambda_{\mathrm{d}}(z)\,\mathbb{1}\big(\Lambda_{\mathrm{d}}(z) \le \lambda_*/[z]\big) \lesssim [z]^{-2}\|\mathbf{d}(z)\|_\infty, \tag{3.25}$$

according to (3.8) in Lemma 3.3 (for $|z| \ge 10$) and (3.13) from Corollary 3.4 (for $|z| \le 10$), where λ_* is a sufficiently small positive constant.

We will now establish a gap in the possible values of $\Lambda(z)$ by showing (cf. (3.29) below) that the right hand side of (3.25) is much less than $\lambda_*/[z]$. To this end we estimate the norm of $\mathbf{d}$ in (3.25) by Lemma 2.1 and also use the bound on the off-diagonal contribution, Λ_{o}, from the same lemma,

$$\begin{aligned}\big([z]^2\Lambda_{\mathrm{d}} + \|\mathbf{d}\|_\infty\big)\,\mathbb{1}\big(\Lambda \le \lambda_*/[z]\big) &\prec [z]^{-2}\sqrt{\frac{\operatorname{Im}\langle\mathbf{g}\rangle}{N\eta}} + \frac{1}{\sqrt{N}},\\ [z]^2\Lambda_{\mathrm{o}}\,\mathbb{1}\big(\Lambda \le \lambda_*/[z]\big) &\prec \sqrt{\frac{\operatorname{Im}\langle\mathbf{g}\rangle}{N\eta}} + \frac{1}{\sqrt{N}}.\end{aligned} \tag{3.26}$$

Now we use $\operatorname{Im}\langle\mathbf{g}\rangle = \pi\rho + \operatorname{Im}\langle\mathbf{g}-\mathbf{m}\rangle \lesssim \rho + \Lambda_{\mathrm{d}}$ to estimate the first terms on the right hand side of (3.26):

$$\sqrt{\frac{\operatorname{Im}\langle\mathbf{g}\rangle}{N\eta}} \lesssim \sqrt{\frac{\pi\rho}{N\eta}} + \sqrt{\frac{1}{N\eta}\,\Lambda_{\mathrm{d}}}.$$

Using again the weighted Cauchy-Schwarz inequality in the second term yields

$$\big([z]^2\Lambda_{\mathrm{d}} + \|\mathbf{d}\|_\infty\big)\,\mathbb{1}\big(\Lambda \le \lambda_*/[z]\big) \prec [z]^{-2}\sqrt{\frac{\rho}{N\eta}} + [z]^{-6}\frac{N^{\varepsilon}}{N\eta} + \frac{1}{\sqrt{N}} + N^{-\varepsilon}[z]^2\Lambda_{\mathrm{d}}.$$

The term $N^{-\varepsilon}[z]^2\Lambda_{\mathrm{d}}$ can be absorbed (cf. (ii) of Lemma A.1) into the left hand side and we arrive at

$$\big([z]^2\Lambda_{\mathrm{d}} + \|\mathbf{d}\|_\infty\big)\,\mathbb{1}\big(\Lambda \le \lambda_*/[z]\big) \prec [z]^{-2}\sqrt{\frac{\rho}{N\eta}} + \frac{[z]^{-6}}{N\eta} + \frac{1}{\sqrt{N}}. \tag{3.27}$$

For the off-diagonal error terms we plug this into the second bound of (3.26) after using $\mathrm{Im}\langle \mathbf{g}\rangle \lesssim \rho + \Lambda_{\mathrm{d}}$ and get

$$\Lambda_{\mathrm{o}} \;\prec\; [z]^{-2}\sqrt{\frac{\rho}{N\eta}} + \frac{[z]^{-6}}{N\eta} + \frac{[z]^{-3}}{N^{1/4}}\sqrt{\frac{1}{N\eta}} + \frac{[z]^{-2}}{\sqrt{N}}. \tag{3.28}$$

In particular, we combine (3.27) and (3.28) to establish a gap in the values that Λ can take,

$$\Lambda\, \mathbb{1}\big(\Lambda \leq \lambda_*/[z]\big) \;\prec\; [z]^{-1}N^{-\gamma/2}. \tag{3.29}$$

Here we used $\eta \geq N^{\gamma-1}$. This shows that either $\Lambda \geq \lambda_*/[z]$ or $\Lambda \leq N^{-\gamma/4}/[z]$ a.w.o.p.

Now we apply Lemma A.2 on the connected domain

$$\Big\{z \in \mathbb{H} : \mathrm{Im} z \geq N^{\gamma-1}, \mathrm{dist}(z, \mathbb{M}) \geq \delta_*, |z| \leq N^5\Big\},$$

with the choices

$$\varphi(z) \;:=\; [z]\Lambda(z), \quad \Phi(z) \;:=\; N^{-\gamma/3}, \quad z_0 \;:=\; \mathrm{i}N^5. \tag{3.30}$$

The continuity condition (A.1) of the lemma for these two functions follows from the Hölder-continuity, (3.5), of the solution of the QVE and the weak continuity of the resolvent elements,

$$|G_{ij}(z_1) - G_{ij}(z_2)| \;\leq\; \frac{|z_1 - z_2|}{(\mathrm{Im} z_1)(\mathrm{Im} z_2)} \;\leq\; N^2|z_1 - z_2|. \tag{3.31}$$

The condition (A.3) holds since by (3.2) on the first regime we have a.w.o.p. $\varphi(z_0) \leq \Phi(z_0)$. Finally, (3.29) implies a.w.o.p. $\varphi\, \mathbb{1}(\varphi \in [\Phi - N^{-1}, \Phi]) < \Phi - N^{-1}$ and thus (A.2). We infer that a.w.o.p. $\varphi \leq \Phi$. In particular, the indicator function in (3.27) and (3.28) is non-zero a.w.o.p. Thus, (3.27) and (3.28) imply (3.2) in the second regime. □

We will now sketch the proof of Corollary 1.8. The set-up in this corollary differs slightly from the one used in the rest of this paper, because the uniform bound (assumption (C)) on the solution of (1.7) is not assumed. We therefore use additional information from [1] about $\mathbf{m}$ in this more general setting.

Proof of Corollary 1.8 Since the boundedness assumption (C) on the solution of the QVE is dropped in this corollary, its proof starts by showing that nevertheless for some constant $P > 0$ we have

$$|m_i(z)| \;\leq\; P, \quad i = 1, \ldots, N, \;\; z \in I + \mathrm{i}(0, \infty). \tag{3.32}$$

In this setting the solution $\mathbf{m}(z)$ is not guaranteed to be extendable as a Hölder-continuous function with N-independent Hölder-norm to $z \in \overline{\mathbb{H}}$. The density of states, defined by (1.12), however still has a Hölder-norm with Hölder-exponent $1/13$ that is independent of N (cf. (i) of Proposition 7.1 and (i) of Theorem 6.1 in [1]). Here, we used $L = 1$ for the model parameter from assumption (B). Furthermore, (3.32) follows from the lower bound on the density of states and (i) of Lemma 5.4 and (i) of Theorem 6.1 in [1]. For the proof of Proposition 3.1 we only used the properties of the solution of (1.7), valid for z in the entire complex upper half plane, that are listed in Corollaries 2.2, 3.2 and 3.4. These properties remain true for $\mathrm{Re} z \in I$ (cf. Theorem 2.1, (i) of Theorem 2.12, (i) of Proposition 5.3 and Proposition 7.1 in [1]) if only (3.32) instead of (1.10) is satisfied. Thus, (3.2a), (3.2b) and (3.3) hold for $z \in I + \mathrm{i}[N^{\gamma-1}, \infty)$ and Corollary 1.8 is proven. □

4 Local law close to local minima

4.1 The solution of the QVE

In this subsection we state a few facts about the solution $\mathbf{m}$ of the QVE (1.7) and about the stability of this equation against perturbations. These facts are summarized in two theorems that are taken from the companion paper [1]. The first theorem contains regularity properties of $\mathbf{m}$. Furthermore, it provides lower and upper bounds on the imaginary part, $\mathrm{Im}\langle \mathbf{m} \rangle = \pi\rho$, by explicit functions. It is a combination of the statements from Theorem 2.1, Theorem 2.4, Theorem 2.6 and Corollary A.1 of [1].

Theorem 4.1 (Solution of the QVE) *Let the sequence* $\mathbf{S} = \mathbf{S}^{(N)}$ *satisfy the assumptions* (A)-(C). *Then for every component,* $m_i : \mathbb{H} \to \mathbb{H}$*, of the unique solution,* $\mathbf{m} = (m_1, \ldots, m_N)$*, of the QVE there is a probability density* $p_i : \mathbb{R} \to [0, \infty)$ *with support in the interval* $[-2, 2]$*, such that*

$$m_i(z) = \int_{\mathbb{R}} \frac{p_i(\tau)\mathrm{d}\tau}{\tau - z}, \quad z \in \mathbb{H}, \ i = 1, \ldots, N. \tag{4.1}$$

The probability densities are comparable,

$$p_i(\tau) \sim p_j(\tau), \quad \tau \in \mathbb{R}, \ i, j = 1, \ldots, N. \tag{4.2}$$

The solution $\mathbf{m}$ *has a uniformly Hölder-continuous extension (denoted again by* $\mathbf{m}$*) to the closed complex upper half plane* $\overline{\mathbb{H}} = \mathbb{H} \cup \mathbb{R}$*,*

$$\|\mathbf{m}(z_1) - \mathbf{m}(z_2)\|_\infty \lesssim |z_1 - z_2|^{1/3}, \quad z_1, z_2 \in \overline{\mathbb{H}}. \tag{4.3}$$

Its absolute value satisfies

$$|m_i(z)| \sim [z]^{-1}, \quad z \in \overline{\mathbb{H}}, \ i = 1, \ldots, N.$$

Let $\rho : \mathbb{R} \to [0, \infty)$, $\tau \mapsto \langle \mathbf{p}(\tau) \rangle$ *be the density of states, defined in* (1.12)*. Then there exists a positive constant* $\delta_* \sim 1$ *such that the following holds true. The support of the density consists of* $K \sim 1$ *disjoint intervals of lengths at least* $2\delta_*$*, i.e.,*

$$\operatorname{supp}\rho = \bigcup_{i=1}^{K} [\alpha_i, \beta_i], \quad \textit{where} \quad \beta_i - \alpha_i \geq 2\delta_*, \quad \textit{and} \quad \alpha_i < \beta_i < \alpha_{i+1}. \tag{4.4}$$

The size of the harmonic extension (1.15) *of* ρ*, up to constant factors, is given by explicit functions as follows. Let* $\eta \in [0, \delta_*]$.

- **Bulk:** *Close to the support of the density of states but away from the local minima in* $\mathbb{M}$ *(cf.* (3.1)*) the function* ρ *is comparable to* 1*, i.e.,*

$$\rho(\tau + \mathrm{i}\eta) \sim 1, \quad \tau \in \operatorname{supp}\rho, \ \operatorname{dist}(\tau, \mathbb{M}) \geq \delta_*. \tag{4.5a}$$

- **At an internal edge:** *At the edges* α_i, β_{i-1} *with* $i = 2, \ldots, K$ *in the direction where the support of the density of states continues the size of* ρ *is*

$$\rho(\alpha_i + \omega + \mathrm{i}\eta) \sim \rho(\beta_{i-1} - \omega + \mathrm{i}\eta) \sim \frac{(\omega + \eta)^{1/2}}{(\alpha_i - \beta_{i-1} + \omega + \eta)^{1/6}}, \tag{4.5b}$$

for all $\omega \in [0, \delta_*]$.

- **Inside a gap:** *Between two neighboring edges* β_{i-1} *and* α_i *with* $i = 2, \ldots, K$*, the function* ρ *satisfies*

$$\rho(\beta_{i-1} + \omega + \mathrm{i}\eta) \sim \rho(\alpha_i - \omega + \mathrm{i}\eta) \sim \frac{\eta}{(\alpha_i - \beta_{i-1} + \eta)^{1/6}(\omega + \eta)^{1/2}}, \tag{4.5c}$$

for all $\omega \in [0, (\alpha_i - \beta_{i-1})/2]$.

- **Around an extreme edge:** *At the extreme points* α_1 *and* β_K *of* $\operatorname{supp}\rho$ *the density of states grows like a square root,*

$$\rho(\alpha_1 + \omega + \mathrm{i}\eta) \sim \rho(\beta_K - \omega + \mathrm{i}\eta) \sim \begin{cases} (\omega + \eta)^{1/2}, & \omega \in [0, \delta_*], \\ \dfrac{\eta}{(|\omega| + \eta)^{1/2}}, & \omega \in [-\delta_*, 0]. \end{cases} \tag{4.5d}$$

- **Close to a local minimum:** *In a neighborhood of a local minimum in the interior of the support of the density of states, i.e., for* $\tau_0 \in \mathbb{M} \cap \operatorname{int}\operatorname{supp}\rho$*, we have*

$$\rho(\tau_0 + \omega + \mathrm{i}\eta) \sim \rho(\tau_0) + (|\omega| + \eta)^{1/3}, \quad \omega \in [-\delta_*, \delta_*]. \tag{4.5e}$$

- **Away from the support:** *Away from the interval in which* $\operatorname{supp}\rho$ *is contained*

$$\rho(z) \sim \frac{\operatorname{Im} z}{|z|^2}, \quad z \in \overline{\mathbb{H}}, \ \operatorname{dist}(z, [\alpha_1, \beta_K]) \geq \delta_*. \tag{4.5f}$$

The next theorem shows that the QVE is stable under small perturbations, $\mathbf{d}$, in the sense that once a solution of the perturbed QVE (4.6) is sufficiently close to $\mathbf{m}$, then the difference between the two can be estimated in terms of $\|\mathbf{d}\|_\infty$. In [1] it is stated as Proposition 10.1.

Theorem 4.2 (Stability) *There exists a scalar function $\sigma : \overline{\mathbb{H}} \to [0,\infty)$, three vector valued functions $\mathbf{s}, \mathbf{t}^{(1)}, \mathbf{t}^{(2)} : \overline{\mathbb{H}} \to \mathbb{C}^N$, a matrix valued function $\mathbf{T} : \overline{\mathbb{H}} \to \mathbb{C}^{N\times N}$, all depending only on $\mathbf{S}$, and a positive constant λ_*, depending only on the model parameters p, P and L, such that for two arbitrary vector valued functions $\mathbf{d} : \mathbb{H} \to \mathbb{C}^N$ and $\mathbf{g} : \mathbb{H} \to (\mathbb{C}\backslash\{0\})^N$ that satisfy*

$$-\frac{1}{g_i(z)} = z + \sum_{j=1}^{N} s_{ij} g_j(z) + d_i(z), \quad z \in \mathbb{H}, \tag{4.6}$$

the difference between $\mathbf{g} = \mathbf{g}(z)$ and $\mathbf{m} = \mathbf{m}(z)$ is bounded in terms of

$$\Theta = \Theta(z) := \big|\big\langle \mathbf{s}(z), \mathbf{g}(z) - \mathbf{m}(z)\big\rangle\big|, \quad z \in \mathbb{H}, \tag{4.7}$$

in the following two ways. On the whole complex upper half plane

$$\|\mathbf{g}-\mathbf{m}\|_\infty \mathbb{1}\big(\|\mathbf{g}-\mathbf{m}\|_\infty \le \lambda_*\big) \lesssim \Theta + \|\mathbf{d}\|_\infty, \tag{4.8}$$

$$|\langle \mathbf{w}, \mathbf{g}-\mathbf{m}\rangle| \mathbb{1}\big(\|\mathbf{g}-\mathbf{m}\|_\infty \le \lambda_*\big) \lesssim \|\mathbf{w}\|_\infty \Theta + \|\mathbf{w}\|_\infty \|\mathbf{d}\|_\infty^2 + |\langle \mathbf{T}\mathbf{w}, \mathbf{d}\rangle|, \tag{4.9}$$

for any non-random $\mathbf{w} \in \mathbb{C}^N$. The scalar function $\Theta : \overline{\mathbb{H}} \to [0,\infty)$ satisfies a cubic equation

$$\big|\Theta^3 + \pi_2 \Theta^2 + \pi_1 \Theta\big| \mathbb{1}\big(\|\mathbf{g}-\mathbf{m}\|_\infty \le \lambda_*\big) \lesssim \|\mathbf{d}\|_\infty^2 + |\langle \mathbf{t}^{(1)}, \mathbf{d}\rangle| + |\langle \mathbf{t}^{(2)}, \mathbf{d}\rangle|. \tag{4.10}$$

The coefficients $\pi_1, \pi_2 : \mathbb{H} \to \mathbb{C}$ may depend on $\mathbf{S}$ and $\mathbf{g}$. They satisfy

$$|\pi_1(z)| \sim \frac{\mathrm{Im} z}{\rho(z)} + \rho(z)(\rho(z) + \sigma(z)), \tag{4.11a}$$

$$|\pi_2(z)| \sim \rho(z) + \sigma(z), \tag{4.11b}$$

for all $z \in \mathbb{H}$. Moreover, the functions σ, $\mathbf{s}$, $\mathbf{t}^{(1)}$, $\mathbf{t}^{(2)}$ and $\mathbf{T}$ are regular in the sense that

$$|\sigma(z_1) - \sigma(z_2)| + \|\mathbf{s}(z_1) - \mathbf{s}(z_2)\| \lesssim |z_1 - z_2|^{1/3}, \quad z_1, z_2 \in \overline{\mathbb{H}}, \tag{4.12}$$

$$\sigma(z) + \|\mathbf{s}(z)\|_\infty + \|\mathbf{t}^{(1)}(z)\|_\infty + \|\mathbf{t}^{(2)}(z)\|_\infty + \|\mathbf{T}(z)\|_{\infty\to\infty} \lesssim 1, \quad z \in \overline{\mathbb{H}}. \tag{4.13}$$

Furthermore, the function σ is related to the density of states by

$$\sigma(\alpha_i) \sim \sigma(\beta_{i-1}) \sim (\alpha_i - \beta_{i-1})^{1/3}, \quad i = 2, \ldots, K, \tag{4.14a}$$

$$\sigma(\alpha_1) \sim \sigma(\beta_K) \sim 1, \tag{4.14b}$$

$$\sigma(\tau_0) \lesssim \rho(\tau_0)^2, \quad \tau_0 \in \mathbb{M}\backslash\{\alpha_i, \beta_i\}. \tag{4.14c}$$

We warn the reader that in this paper Θ and σ denote the absolute values of the quantities denoted by the same symbols in Proposition 10.1 of [1]. The function σ appears naturally in the analysis of the QVE. Analogous to the more explicitly constructed function Δ from Definition 1.5, at an edge the value of σ^3 encodes the size of the corresponding gap in supp ρ. At the local minima in $\mathbb{M}\backslash\{\alpha_i, \beta_i\}$ the value of σ^3 is small, provided the density of states has a small value at the minimum. In this sense it is again analogous to Δ, which vanishes at these internal minima.

4.2 Coefficients of the cubic equation

The stability of QVE near the points in $\mathbb{M}$ requires a careful analysis of the cubic equation (4.10) for Θ from Theorem 4.2. For this, we will provide a more explicit description of the upper and lower bounds from (4.11) on the coefficients, π_1 and π_2, of the cubic equation.

Proposition 4.3 (Behavior of the coefficients) *There exist $\delta_*, c_* \sim 1$ such that for all $\eta \in [0, \delta_*]$ the coefficients, π_1 and π_2, of the cubic Eq.* (4.10) *satisfy the following bounds.*

- **Around an internal edge:** *At the edges α_i, β_{i-1} of the gap with length $\Delta := \alpha_i - \beta_{i-1}$ for $i = 2, \ldots, K$, we have*

$$\begin{aligned} |\pi_1(\alpha_i + \omega + \mathrm{i}\eta)| &\sim |\pi_1(\beta_{i-1} - \omega + \mathrm{i}\eta)| \sim (|\omega| + \eta)^{1/2}(|\omega| + \eta + \Delta)^{1/6}, \\ |\pi_2(\alpha_i + \omega + \mathrm{i}\eta)| &\sim |\pi_2(\beta_{i-1} - \omega + \mathrm{i}\eta)| \sim (|\omega| + \eta + \Delta)^{1/3}, \\ &\qquad \omega \in [-c_*\Delta, \delta_*]. \end{aligned} \tag{4.15a}$$

- **Well inside a gap:** *Between two neighboring edges β_{i-1} and α_i of the gap with length $\Delta := \alpha_i - \beta_{i-1}$ for $i = 2, \ldots, K$, the first coefficient, π_1, satisfies*

$$|\pi_1(\alpha_i - \omega + \mathrm{i}\eta)| \sim |\pi_1(\beta_{i-1} + \omega + \mathrm{i}\eta)| \sim (\eta + \Delta)^{2/3}, \quad \omega \in \left[c_*\Delta, \frac{\Delta}{2}\right]. \tag{4.15b}$$

The second coefficient, π_2, satisfies the upper bounds,

$$\begin{aligned} |\pi_2(\alpha_i - \omega + \mathrm{i}\eta)| &\lesssim (\eta + \Delta)^{1/3}, \\ |\pi_2(\beta_{i-1} + \omega + \mathrm{i}\eta)| &\lesssim (\eta + \Delta)^{1/3}, \end{aligned} \quad \omega \in \left[c_*\Delta, \frac{\Delta}{2}\right]. \tag{4.15c}$$

- **Around an extreme edge:** *Around the extreme points α_1 and β_K of* supp ρ, *we have*

$$\begin{aligned} |\pi_1(\alpha_1 + \omega + \mathrm{i}\eta)| &\sim |\pi_1(\beta_K - \omega + \mathrm{i}\eta)| \sim (\omega + \eta)^{1/2} \\ |\pi_2(\alpha_1 + \omega + \mathrm{i}\eta)| &\sim |\pi_2(\beta_K - \omega + \mathrm{i}\eta)| \sim 1, \end{aligned} \quad \omega \in [-\delta_*, \delta_*]. \tag{4.15d}$$

- **Close to a local minimum:** *In a neighborhood of the local minimum in the interior of the support of the density of states, i.e. for $\tau_0 \in \mathbb{M} \cap$* int supp ρ, *we have*

$$\begin{aligned} |\pi_1(\tau_0 + \omega + \mathrm{i}\eta)| &\sim \rho(\tau_0)^2 + (|\omega| + \eta)^{2/3}, \\ |\pi_2(\tau_0 + \omega + \mathrm{i}\eta)| &\sim \rho(\tau_0) + (|\omega| + \eta)^{1/3}, \end{aligned} \quad \omega \in [-\delta_*, \delta_*]. \tag{4.15e}$$

Proof The proof is split according to the cases above. In each case we combine the general formulas (4.11) with the knowledge about the harmonic extension, ρ, of the density of states from Theorem 4.1 and about the behavior of the positive Hölder-continuous function, σ, at the minima in $\mathbb{M}$ from (4.14). The positive constant δ_* is chosen to have at most the same value as in Theorem 4.1. We start with the simplest case.

Around an extreme edge: By the Hölder-continuity of σ (cf. (4.12)) and because σ is comparable to 1 at the points α_1 and β_K (cf. (4.14)), this function is comparable to 1 in the whole δ_*-neighborhood of the extreme edges. Thus, using (4.11) inside this neighborhood, we find

$$|\pi_1(z)| \sim \frac{\mathrm{Im}z}{\rho(z)} + \rho(z), \quad |\pi_2(z)| \sim 1.$$

The claim now follows from the behavior of ρ, given in Theorem 4.1, inside this domain.

Close to a local minimum: In this case $\rho + \sigma$ is comparable to ρ. In fact, using the $1/3$-Hölder-continuity of σ (cf. (4.12)) and its bound at the minimum, $\tau_0 \in \mathbb{M}$, (cf. (4.14)) we find

$$\rho(z) \le \rho(z) + \sigma(z) \lesssim \rho(z) + \rho(\tau_0)^2 + |z - \tau_0|^{1/3} \sim \rho(z), \quad |z - \tau_0| \le \delta_*. \tag{4.16}$$

In the last relation we used the behavior (4.5e) of ρ from Theorem 4.1. By (4.11) we conclude that inside the δ_*-neighborhood of τ_0,

$$|\pi_1(z)| \sim \frac{\mathrm{Im}z}{\rho(z)} + \rho(z)^2, \quad |\pi_2(z)| \sim \rho(z). \tag{4.17}$$

Using the upper and lower bounds on $\rho(z)$ again, gives the desired result, (4.15e).

Around an internal edge: First we prove the bounds on $|\pi_2|$, starting from (4.11). The upper bound simply uses the $1/3$-Hölder-continuity and the behavior at the edge points of σ,

$$|\pi_2(z)| \sim \rho(z) + \sigma(z) \lesssim \rho(z) + \Delta^{1/3} + |z - \tau_0|^{1/3}, \tag{4.18}$$

where τ_0 is one of the edge points α_i or β_{i-1}. The claim follows from plugging in the size of ρ from the two corresponding domains in Theorem 4.1, i.e., the domain close to an edge, (4.5b), and the domain inside a gap, (4.5c).

For the lower bound we consider two different regimes. In the first case z is close to the edge point, $|z - \tau_0| \le c\Delta$, for some small positive constant c, depending only on the model parameters p, P and L. We find

$$|\pi_2(z)| \sim \rho(z) + \sigma(z) \gtrsim \rho(z) + \Delta^{1/3} - C|z - \tau_0|^{1/3} \sim \rho(z) + \Delta^{1/3},$$

provided c is small enough. This bound coincides with the lower bound on π_2 in (4.15a), once the size of ρ from (4.5b) is used.

In the second regime, $|z - \tau_0| \geq c\Delta$, we simply use $|\pi_2(z)| \gtrsim \rho(z)$ from (4.11). If $\mathrm{Re} z \in \operatorname{supp} \rho$, then the size of ρ from (4.5b) yields the desired lower bound. If, on the other hand, $\mathrm{Re} z$ lies inside a gap of $\operatorname{supp} \rho$, then we use the freedom of choosing the constant c_* in Proposition 4.3. Suppose $c_* \leq c/2$. Then $|z - \tau_0| \geq c\Delta$ and $|\mathrm{Re} z - \tau_0| \leq c_* \Delta$ imply $\mathrm{Im} z \gtrsim \Delta$ and

$$\rho(z) \sim (\mathrm{Im} z)^{1/3} \gtrsim \Delta^{1/3} + |z - \tau_0|^{1/3}.$$

This finishes the proof of the upper and lower bound on $|\pi_2|$ on this domain. For the claim about $|\pi_1|$ we plug the result about $|\pi_2|$ and the size of ρ into

$$|\pi_1| \sim \frac{\mathrm{Im} z}{\rho(z)} + \rho(z)|\pi_2(z)|. \tag{4.19}$$

Well inside a gap: For the upper bound on $|\pi_2|$ we simply use (4.18) again, which follows from (4.12) and (4.14). The comparison relation for $|\pi_1|$ now follows from (4.19) again. For the lower bound, $|\pi_1| \gtrsim \mathrm{Im} z/\rho$ and (4.5c) from Theorem 4.1 are sufficient. This finishes the proof of the proposition. □

4.3 Rough bound on Λ close to local minima

In the following lemma we will see that a.w.o.p. $\Lambda \leq c$ for some arbitrarily small constant $c > 0$. Since the local law away from $\mathbb{M}$ is already shown in Proposition 3.1 we may restrict to bounded z in the following. From here on until the end of Section 4 we assume $|z| \leq 10$.

Lemma 4.4 (Rough bound) *Let λ_* be a positive constant. Then, uniformly for all $z = \tau + \mathrm{i}\eta \in \mathbb{H}$ with $\eta \geq N^{\gamma - 1}$, the function Λ is uniformly small,*

$$\Lambda(z) \leq \lambda_* \quad \textit{a.w.o.p.} \tag{4.20}$$

Proof Away from the local minima in $\mathbb{M}$ the claim follows from (3.2) in Proposition 3.1. We will therefore prove that Λ is smaller than any fixed positive constant in some δ-neighborhood of $\mathbb{M}$. We will use the freedom to choose the size $\delta \sim 1$ of these neighborhoods as small as we like.

Let us sketch the upcoming argument. Close to the points in $\mathbb{M}$ we make use of Theorem 4.2. Using Lemma 2.1, we will see that the right hand side of the cubic equation in Θ, (4.10), is smaller than a small negative power, $N^{-\varepsilon}$, of N, provided Λ is bounded by a small constant, $\Lambda \leq \lambda_*$. This will imply that Θ itself is small and through (4.8) that the bound on Λ can be improved to $\Lambda \leq \lambda_*/2$. In this way we establish a gap in the possible values that the continuous function Λ can take. Lemma A.2 in the appendix is then used to propagate the bound on Λ from Proposition 3.1 into the δ-neighborhoods of the points in $\mathbb{M}$.

Now we start the detailed proof from the fact that Θ satisfies the cubic equation (4.10), whose right hand side is bounded by $C\|\mathbf{d}\|_\infty$ for some constant C, depending only on the model parameters. Note that $\|\mathbf{d}\|_\infty \lesssim 1$ as long as $\Lambda \le \lambda_*$ because in this case $|m_i| \sim 1$, $|g_i| \sim 1$ and $\mathbf{g}$ satisfies the perturbed QVE with perturbation $\mathbf{d}$. From the definition of Θ in (4.7) and the uniform bound on $\mathbf{s}$ from (4.13), we get $\Theta \lesssim \Lambda$. Since the coefficient $|\pi_2|$ is uniformly bounded (cf. (4.11)), the cubic equation for Θ implies the three bounds

$$\Theta\, \mathbb{1}(\Lambda \le \varepsilon_1, |\pi_1| \ge C_1\varepsilon_1) \;\lesssim\; \frac{\|\mathbf{d}\|_\infty}{|\pi_1|}\,, \tag{4.21a}$$

$$\Theta\, \mathbb{1}(\Lambda \le \varepsilon_2, |\pi_2| \ge C_2\varepsilon_2) \;\lesssim\; \frac{|\pi_1|}{|\pi_2|} + \frac{\|\mathbf{d}\|_\infty^{1/2}}{|\pi_2|^{1/2}}\,, \tag{4.21b}$$

$$\Theta\, \mathbb{1}(\Lambda \le \lambda_*) \;\lesssim\; |\pi_2| + \sqrt{|\pi_1|} + \|\mathbf{d}\|_\infty^{1/3}. \tag{4.21c}$$

Here, $\varepsilon_1, \varepsilon_2 \in (0, \lambda_*)$, with $\lambda_* \in (0, 1)$ from Theorem 4.2, are arbitrary constants and $C_1, C_2 > 0$ depend only on the model parameters. We prove (4.21b); the other two bounds are obtained similarly. First we show that under the assumptions $\Lambda \le \varepsilon_2$ and $|\pi_2| \ge C_2\varepsilon_2$ the second order term $\pi_2\Theta^2$ is at least three times larger than Θ^3 provided $C_2 \sim 1$ is chosen to be sufficiently large. Indeed, since $\Theta \le \|\mathbf{s}\|_\infty\Lambda \le \|\mathbf{s}\|_\infty\varepsilon_2$ and $|\pi_2| \ge C_2\varepsilon_2$, it suffices to choose $C_2 \ge 3\|\mathbf{s}\|_\infty \sim 1$. Here we have also used (4.13). Next we compare the second order term to the linear term $\pi_1\Theta$. We may assume that $\Theta \ge 3|\pi_1/\pi_2|$, otherwise (4.21b) holds trivially. Together with $|\pi_2|\Theta^2 \ge 3\Theta^3$ proved above this implies that the second order term $\pi_2\Theta^2$ dominates the left hand side of (4.10). Combining this with $|\langle \mathbf{t}^{(j)}, \mathbf{d}\rangle| \lesssim \|\mathbf{d}\|_\infty$ (cf. (4.13)) on the right hand side of (4.10), hence yields

$$\frac{1}{3}|\pi_2|\Theta^2 \;\le\; |\Theta^3 + \pi_2\Theta^2 + \pi_1\Theta| \;\lesssim\; \|\mathbf{d}\|_\infty. \tag{4.22}$$

In order to satisfy the constraint of (4.10) we have also used $\varepsilon_2 \le \lambda_*$. This together with (4.22) yields (4.21b).

Let $\delta \in (0, 1)$ be another constant to be chosen later which depends only on the model parameters p, P, and L. We split $\mathbb{M}$ into four subsets, which are treated separately,

$$\begin{aligned}
\mathbb{M}_1(\delta) &:= \big\{\tau_0 \in \mathbb{M}\backslash\partial \operatorname{supp}\rho : \rho(\tau_0) > \delta^{1/3}\big\},\\
\mathbb{M}_2(\delta) &:= \big\{\tau_0 \in \partial \operatorname{supp}\rho : \Delta(\tau_0) > \delta^{1/2}\big\},\\
\mathbb{M}_3(\delta) &:= \big\{\tau_0 \in \mathbb{M}\backslash\partial \operatorname{supp}\rho : \rho(\tau_0) \le \delta^{1/3}\big\},\\
\mathbb{M}_4(\delta) &:= \big\{\tau_0 \in \partial \operatorname{supp}\rho : \Delta(\tau_0) \le \delta^{1/2}\big\}.
\end{aligned}$$

The function Δ is from Definition 1.5 and its value is simply the length of the gap at the point $\tau_0 \in \partial \operatorname{supp}\rho$ where it is evaluated. We also define the δ-neighborhoods of these subsets,

$$\mathbb{D}_k(\delta) \;:=\; \big\{z \in \mathbb{H} :\; \operatorname{dist}(z, \mathbb{M}_k(\delta)) \le \delta\big\}, \quad k = 1, 2, 3, 4.$$

As an immediate consequence of the upper and lower bounds on the coefficients, π_1 and π_2, presented in Proposition 4.3, we see that

$$|\pi_1(z)| \gtrsim \delta^{2/3}, \quad z \in \mathbb{D}_1(\delta), \tag{4.23a}$$
$$|\pi_1(z)| \lesssim \delta^{1/2}, \quad |\pi_2(z)| \gtrsim \delta^{1/6}, \quad z \in \mathbb{D}_2(\delta), \tag{4.23b}$$
$$|\pi_1(z)| \lesssim \delta^{1/2}, \quad |\pi_2(z)| \lesssim \delta^{1/6}, \quad z \in \mathbb{D}_3(\delta) \cup \mathbb{D}_4(\delta). \tag{4.23c}$$

On $\mathbb{D}_2(\delta)$ only the regimes around an internal edge, (4.15a), and around an extreme edge, (4.15d), are relevant. The case well inside the gap, (4.15b) and (4.15c), does not apply for small enough δ, since $\Delta(\tau_0) > \delta^{1/2}$ but $|z - \tau_0| \leq \delta$.

Now we make a choice for the two constants ε_1 and ε_2. We express them in terms of δ as

$$\varepsilon_1 := \delta, \quad \varepsilon_2 := \delta^{1/5}.$$

We pair the bounds on Θ from (4.21) with the corresponding bounds from (4.23) on the coefficients of the cubic equation. For small enough δ the conditions on π_1 in (4.21a) and π_2 in (4.21b) are automatically satisfied by the choice of ε_1 and ε_2, as well as the upper and lower bounds from (4.23a) and (4.23b). Thus, for small enough δ we end up with

$$\begin{gathered}
\Theta(z)\mathbb{1}(\Lambda(z) \leq \delta) \lesssim \delta^{-2/3}\|\mathbf{d}(z)\|_\infty, z \in \mathbb{D}_1(\delta), \\
\Theta(z)\mathbb{1}(\Lambda(z) \leq \delta^{1/5}) \lesssim \delta^{1/3} + \delta^{-1/12}\|\mathbf{d}(z)\|_\infty^{1/2}, \quad z \in \mathbb{D}_2(\delta), \\
\Theta(z)\mathbb{1}(\Lambda(z) \leq \lambda_*) \lesssim \delta^{1/6} + \|\mathbf{d}(z)\|_\infty^{1/3}, z \in \mathbb{D}_3(\delta) \cup \mathbb{D}_4(\delta).
\end{gathered}$$

At this stage we use Lemma 2.1 in the form of $\|\mathbf{d}\|_\infty \prec N^{-\gamma/2}$ on the set where $\Lambda \leq \lambda_*/10$, say, and (4.8) from Theorem 4.2. We may choose λ_* to be sufficiently small compared to the constants with the same name from these two statements. Furthermore, we choose δ so small that $\delta^{1/5} \leq \lambda_*$. Since $\|\mathbf{d}\|_\infty \leq N^{-\gamma/2+c}$ a.w.o.p. for an arbitrary $c > 0$ we obtain

$$\Lambda(z)\, \mathbb{1}(\Lambda(z) \leq \delta) \lesssim \delta^{-2/3} N^{-\gamma/3}, z \in \mathbb{D}_1(\delta), \tag{4.24a}$$
$$\text{a.w.o.p.} \quad \Lambda(z)\, \mathbb{1}(\Lambda(z) \leq \delta^{1/5}) \lesssim \delta^{1/3} + \delta^{-1/12} N^{-\gamma/5}, \quad z \in \mathbb{D}_2(\delta), \tag{4.24b}$$
$$\Lambda(z)\, \mathbb{1}(\Lambda(z) \leq \lambda_*) \lesssim \delta^{1/6} + N^{-\gamma/7}, z \in \mathbb{D}_3(\delta) \cup \mathbb{D}_4(\delta). \tag{4.24c}$$

The right hand sides, including the constants from the comparison relation, can be made smaller than any given constant λ_* by choosing $\delta = \delta_*$, depending only on the model parameters, small enough and N sufficiently large. Furthermore, (4.24) establish a gap in the possible values that Λ can take on the δ_*-neighborhood of any point in $\mathbb{M}$. By Proposition 3.1 we have the bound $\Lambda \prec N^{-\gamma/2}$ outside these δ_*-neighborhoods and thus also for at least one point in the boundary of each neighborhood. Now we apply Lemma A.2 to each neighborhood and in this way we propagate the bound $\Lambda \leq \lambda_*$ to every point z in the δ_*-neighborhood of $\mathbb{M}$ with $\mathrm{Im}\, z \geq N^{\gamma-1}$. □

4.4 Proof of Theorem 1.7

According to Proposition 3.1 the local law, Theorem 1.7, holds outside the δ_*-neighborhoods of the points in $\mathbb{M}$. It remains to show that it is true inside these neighborhoods as well. From here on we assume that $z \in \mathbb{H}$ satisfies $\operatorname{dist}(z, \mathbb{M}) \leq \delta_*$ and $\operatorname{Im} z \geq N^{\gamma-1}$. Let $\tau_0 \in \mathbb{M}$ be one of the closest points to z in $\mathbb{M}$, i.e.,

$$|z - \tau_0| = \operatorname{dist}(z, \mathbb{M}).$$

When $\tau_0 \in \partial \operatorname{supp} \rho$ we denote by $\theta = \theta(\tau_0) \in \{\pm 1\}$ the direction that points towards the gap in $\operatorname{supp} \rho$ at τ_0. In case $\tau_0 \notin \partial \operatorname{supp} \rho$ we make the arbitrary choice $\theta := +1$, i.e.,

$$\theta := \begin{cases} -1 & \text{if } \tau_0 \in \{\alpha_i\}, \\ +1 & \text{if } \tau_0 \in \{\beta_i\}, \\ +1 & \text{if } \tau_0 \in \mathbb{M} \backslash \partial \operatorname{supp} \rho. \end{cases}$$

The minimum τ_0 will be considered fixed in the following analysis. We parametrize z as follows in the neighborhood of $\tau_0 \in \mathbb{M}$:

$$z = \tau_0 + \theta\omega + \mathrm{i}\eta, \tag{4.25}$$

where $\eta \in (0, \delta_*]$ and $\omega \in [-\delta_*, \delta_*]$. We will then prove the local law in the form

$$\Lambda(z) \prec \sqrt{\frac{\rho(z)}{N\eta}} + \frac{1}{N\eta} + \mathcal{E}(\omega, \eta), \tag{4.26a}$$

$$\big|\big\langle \mathbf{w}, \mathbf{g}(z) - \mathbf{m}(z) \big\rangle\big| \prec \mathcal{E}(\omega, \eta), \tag{4.26b}$$

where the positive error function $\mathcal{E} : [-\delta_*, \delta_*] \times (0, \delta_*] \to (0, \infty)$ is given as the unique solution of an explicit cubic equation in (4.30) below.

To define $\mathcal{E}$ we introduce explicit auxiliary functions $\widetilde{\pi}_1$, $\widetilde{\pi}_2$ and $\widetilde{\rho}$ that are comparable in size to the corresponding functions π_1, π_2 and ρ. The reason for using these auxiliary quantities for the definition of $\mathcal{E}$ instead of the original ones is twofold. Firstly, in this way $\mathcal{E}$ will be an explicit function instead of one that is implicitly defined through the solution of the QVE. The function $\mathcal{E}$ is explicit in the sense that there is a formula for the solution of the cubic equation that defines it and the coefficients are given by the explicit functions $\widetilde{\pi}_1$, $\widetilde{\pi}_2$ and $\widetilde{\rho}$. Secondly, $\mathcal{E}$ will be monotonic of its second variable, η. This property will be used later. The definition of the three auxiliary functions will be different, depending on whether τ_0 is in the boundary of the support of the density of states or not. Recall the definition (1.17) of $\Delta_\delta(\tau)$.

- **Edge:** If $\tau_0 \in \partial \operatorname{supp} \rho$, i.e. τ_0 is an edge of a gap of size $\Delta := \Delta_0(\tau_0)$ in the support of the density of states or an extreme edge. Then we define the three explicit functions

$$\widetilde{\rho}(\omega,\eta) := \begin{cases} \dfrac{(|\omega|+\eta)^{1/2}}{(\Delta+|\omega|+\eta)^{1/6}}, & \omega \in \big[-\delta_*, 0\big], \\ \dfrac{\eta}{(\Delta+\eta)^{1/6}(\omega+\eta)^{1/2}}, & \omega \in \big[0, c_*\Delta\big], \\ \dfrac{\eta}{(\Delta+\eta)^{2/3}}, & \omega \in \Big[c_*\Delta, \dfrac{\Delta}{2}\Big]. \end{cases} \tag{4.27a}$$

$$\widetilde{\pi}_1(\omega,\eta) := \begin{cases} (|\omega|+\eta)^{1/2}(|\omega|+\eta+\Delta)^{1/6}, & \omega \in \big[-\delta_*, 0\big], \\ (\omega+\eta)^{1/2}(\Delta+\eta)^{1/6}, & \omega \in \big[0, c_*\Delta\big], \\ (\Delta+\eta)^{2/3}, & \omega \in \big[c_*\Delta, \frac{\Delta}{2}\big] \end{cases} \tag{4.27b}$$

$$\widetilde{\pi}_2(\omega,\eta) := \begin{cases} (|\omega|+\eta+\Delta)^{1/3}, & \omega \in \big[-\delta_*, 0\big], \\ (\Delta+\eta)^{1/3}, & \omega \in \big[0, c_*\Delta\big], \\ (\Delta+\eta)^{1/3}, & \omega \in \Big[c_*\Delta, \dfrac{\Delta}{2}\Big] \end{cases} \tag{4.27c}$$

Here, $c_* \sim 1$ is the constant from Proposition 4.3.

- **Internal minimum:** If $\tau_0 \in \mathbb{M}\backslash\partial\,\mathrm{supp}\,\rho$, then we define for $\omega \in [-\delta_*, \delta_*]$ the three functions

$$\widetilde{\rho}(\omega,\eta) := \rho(\tau_0) + (|\omega|+\eta)^{1/3}, \tag{4.28a}$$

$$\widetilde{\pi}_1(\omega,\eta) := \rho(\tau_0)^2 + (|\omega|+\eta)^{2/3}, \tag{4.28b}$$

$$\widetilde{\pi}_2(\omega,\eta) := \rho(\tau_0) + (|\omega|+\eta)^{1/3}, \tag{4.28c}$$

By design (cf. Proposition 4.3 and Theorem 4.1) these functions satisfy

$$\rho(\tau_0+\theta\omega+\mathrm{i}\eta) \sim \widetilde{\rho}(\omega,\eta), \quad \text{and} \quad |\pi_k(\tau_0+\theta\omega+\mathrm{i}\eta)| \sim \widetilde{\pi}_k(\omega,\eta), \tag{4.29}$$

except in one special case where the second bound does not hold, namely when $k = 2$, $\tau_0 \in \partial\,\mathrm{supp}\,\rho$ and $\omega \in [c_*\Delta, \Delta/2]$. In this case only the direction $|\pi_2| \lesssim \widetilde{\pi}_2$ is true (cf. (4.15c)).

We fix a positive constant $\widetilde{\varepsilon} \in (0, \gamma/16)$. The value of the function $\mathcal{E}$ at (ω,η) is then defined to be the unique positive solution of the cubic equation

$$\begin{aligned} &\mathcal{E}(\omega,\eta)^3 + \widetilde{\pi}_2(\omega,\eta)\mathcal{E}(\omega,\eta)^2 + \widetilde{\pi}_1(\omega,\eta)\mathcal{E}(\omega,\eta) \\ &\quad = N^{8\widetilde{\varepsilon}}\,\frac{\mathcal{E}(\omega,\eta)}{N\eta} + \frac{\widetilde{\rho}(\omega,\eta)}{N\eta} + \frac{1}{(N\eta)^2}, \end{aligned} \tag{4.30}$$

With the choices (1.23) and (1.25) for $\kappa = \kappa(z)$ we have

$$\mathcal{E} \le N^{9\widetilde{\varepsilon}} \min\left\{\frac{1}{\sqrt{N\eta}}, \frac{\kappa}{N\eta}\right\}, \tag{4.31}$$

for any $N \ge N_0$, where the threshold N_0 here depends on $\widetilde{\varepsilon}$ in addition to p, P, L, $\underline{\mu}$ and γ. The inequality (4.31) is verified by plugging its right hand side into (4.30) in place of $\mathcal{E}$ and checking that on each regime the resulting expression on the right hand side of (4.30) is smaller than the resulting expression on the left hand side of (4.30). The factor of $N^{9\widetilde{\varepsilon}}$ in (4.31) can be absorbed in the stochastic domination in (4.26). Thus (4.26) becomes (1.20) and (1.21) of Theorem 1.7.

Before we start the proof of the local law (4.26), let us motivate the definition of $\mathcal{E}$. As a consequence of Lemma 4.4 the indicator function equals one a.w.o.p. in the statement of Lemma 2.1. Thus, uniformly in the δ_*-neighborhood of τ_0 we have

$$\|\mathbf{d}\|_\infty + \Lambda_{\mathrm{o}} \prec \sqrt{\frac{\rho + |\langle \mathbf{g} - \mathbf{m}\rangle|}{N\eta}} + \frac{1}{\sqrt{N}}. \tag{4.32}$$

Here we used $\mathrm{Im}\langle \mathbf{g}\rangle \lesssim \rho + |\langle \mathbf{g} - \mathbf{m}\rangle|$. Since at the end the local law implies $|\langle \mathbf{g} - \mathbf{m}\rangle| \prec \mathcal{E}$, heuristically we may replace $|\langle \mathbf{g} - \mathbf{m}\rangle|$ in (4.32) by $\mathcal{E}$. In this case, from the fluctuation averaging, Theorem 3.5, we would be able to conclude that for any deterministic vector $\mathbf{w}$ with bounded entries,

$$\|\mathbf{d}\|_\infty^2 + |\langle \mathbf{w}, \mathbf{d}\rangle| \prec \frac{\mathcal{E}}{N\eta} + \frac{\rho}{N\eta} + \frac{1}{(N\eta)^2}. \tag{4.33}$$

Up to the technical factor of $N^{8\varepsilon}$ the right hand side coincides with the right hand side of the cubic equation defining $\mathcal{E}$. On the other hand, the right hand side of the cubic equation (4.10) for the quantity Θ from Theorem 4.2 is of the same form as the left hand side of (4.33). Therefore, we infer

$$|\Theta^3 + \pi_2\Theta^2 + \pi_1\Theta| \prec \frac{\mathcal{E}}{N\eta} + \frac{\rho}{N\eta} + \frac{1}{(N\eta)^2}. \tag{4.34}$$

We will argue that on appropriately chosen domains out of the three summands in the cubic expression in Θ always one is the biggest by far. Therefore, the error function $\mathcal{E}$, defined by (4.30), is essentially the best bound on Θ that one may hope to deduce from (4.34). Indeed, since Θ is by definition an average of $\mathbf{g} - \mathbf{m}$, we expect $\Theta \prec \mathcal{E}$.

We will now prove (4.26). To this end we gradually improve the bound on Θ. Fix some $\varepsilon \in (0, \widetilde{\varepsilon})$. The sequence of deterministic bounds on this quantity is defined as

$$\Phi_0 := 1, \quad \Phi_{k+1} := \max\big\{N^{-\varepsilon}\Phi_k,\ N^{9\varepsilon}\mathcal{E}\big\}. \tag{4.35}$$

From here on until the end of this section the threshold function N_0 from the definition of the stochastic domination (cf. Definition 1.6) as well as the definition of 'a.w.o.p.' (cf. Definition 1.9) may depend on ε in addition to p, P, L, $\underline{\mu}$ and γ. At the end of the proof we will remove this dependence. The following lemma is essential for doing one step in the upcoming iteration.

Lemma 4.5 (Improving bound through cubic) *Suppose that for all* $z \in \tau_0 + [-\delta_*, \delta_*] + \mathrm{i}[N^{\gamma-1}, \delta_*]$ *and some* $k \in \mathbb{N}$ *the quantity* $\Theta(z)$ *from* (4.7) *fulfills*

$$\left|\,\Theta(z)^3+\pi_2(z)\Theta(z)^2+\pi_1(z)\Theta(z)\,\right| \prec \frac{\rho(z)+\Phi_k(\omega,\eta)}{N\eta}+\frac{1}{(N\eta)^2}. \tag{4.36}$$

Then $\Theta(z) \prec \Phi_{k+1}(\omega,\eta)$.

We will postpone the proof of this lemma until the end of this section. First we show how to use this result to prove the local law (Theorem 1.7). Fix an integer $k \geq 0$ and assume that $\Theta + |\langle \mathbf{g}-\mathbf{m}\rangle| \prec \Phi_k$ is already proven. For $k=0$ this follows from the rough bound on Λ in Lemma 4.4, $\Lambda \prec 1 = \Phi_0$. As an induction step we show that $\Theta + |\langle \mathbf{g}-\mathbf{m}\rangle| \prec \Phi_{k+1}$.

From (4.32) we see that

$$\|\mathbf{d}\|_\infty + \Lambda_{\mathrm{o}} \prec \sqrt{\frac{\rho+\Phi_k}{N\eta}}+\frac{1}{N\eta}. \tag{4.37}$$

The right hand side is a deterministic bound on the off-diagonal error Λ_{o}. Therefore the fluctuation averaging (Theorem 3.5) is applicable to $\langle \mathbf{t}^{(1)},\mathbf{d}\rangle$ and $\langle \mathbf{t}^{(2)},\mathbf{d}\rangle$ on right hand side of the cubic equation (4.10)

$$\left|\langle \mathbf{t}^{(j)},\mathbf{d}\rangle\right| \prec \left(\sqrt{\frac{\rho+\Phi_k}{N\eta}}+\frac{1}{N\eta}\right)^2,$$

where N^{-1} from (3.15) has been neglected since $\rho \gtrsim \eta$. In this way we see that the hypothesis (4.36) of Lemma 4.5 is satisfied. Using the lemma the bound on Θ is improved to

$$\Theta(z) \prec \Phi_{k+1}(\omega,\eta). \tag{4.38}$$

In order to improve the bound on $|\langle \mathbf{g}-\mathbf{m}\rangle|$ as well, we use the bound (4.9) from Theorem 4.2 for averages of $\mathbf{g}-\mathbf{m}$ against bounded vectors. Since by Lemma 4.4 the deviation function Λ is bounded by a small constant, the indicator function in (4.9) is a.w.o.p. non-zero. Choosing $\mathbf{w}=(1,\ldots,1)$, we find that

$$|\langle \mathbf{g}-\mathbf{m}\rangle| \lesssim \Theta + \|\mathbf{d}\|_\infty^2 + |\langle \widetilde{\mathbf{w}},\mathbf{d}\rangle|, \quad \text{a.w.o.p.}, \tag{4.39}$$

where $\widetilde{\mathbf{w}} = \mathbf{T}\mathbf{w}$ is a bounded, $\|\widetilde{\mathbf{w}}\|_\infty \lesssim 1$, deterministic vector. Together with the bound (4.37) we apply the fluctuation averaging (Theorem 3.5) again,

$$|\langle \mathbf{g}-\mathbf{m}\rangle| \prec \Phi_{k+1}+\frac{\rho+\Phi_k}{N\eta}+\frac{1}{(N\eta)^2} \lesssim N^{-\varepsilon}\Phi_k+\Phi_{k+1} \lesssim \Phi_{k+1}. \tag{4.40}$$

This concludes one step in the iteration, i.e., we have shown $\Theta + |\langle \mathbf{g}-\mathbf{m}\rangle| \prec \Phi_{k+1}$.

We repeat this step finitely many times and each time improve Φ_k by a factor of $N^{-\varepsilon}$ until it reaches its target value $N^{9\varepsilon}\mathcal{E}$ and is not improved anymore. Note that all constants in our estimates, explicit and hidden, depend only on the model parameters and ε. In particular, the number of steps needed is uniform in (ω,η). At that stage we have

$$\Theta + |\langle \mathbf{g} - \mathbf{m} \rangle| \prec_\varepsilon N^{9\varepsilon}\mathcal{E},$$

where the subindex ε indicates that the threshold N_0 from the stochastic domination may depend on ε. But since $\varepsilon > 0$ was arbitrary, we infer (cf. (i) of Lemma A.1) that $\Theta + |\langle \mathbf{g} - \mathbf{m} \rangle| \prec \mathcal{E}$, where now and until the start of the proof of Lemma 4.5 below the stochastic domination is ε-independent. By (4.32) we conclude

$$\|\mathbf{d}\|_\infty + \Lambda_{\mathrm{o}} \prec \sqrt{\frac{\rho}{N\eta}} + \frac{1}{N\eta} + \mathcal{E}. \tag{4.41}$$

For the bound on the diagonal contribution, Λ_{d}, we use (4.8) to get

$$\Lambda_{\mathrm{d}} \lesssim \Theta + \|\mathbf{d}\|_\infty \prec \sqrt{\frac{\rho}{N\eta}} + \frac{1}{N\eta} + \mathcal{E}.$$

Finally, with the help of (4.9), (4.41) and the fluctuation averaging, we prove the bound on averages of $\mathbf{g} - \mathbf{m}$ against any bounded, $\|\mathbf{w}\|_\infty \leq 1$, deterministic vector,

$$|\langle \mathbf{w}, \mathbf{g} - \mathbf{m} \rangle| \prec \frac{\rho}{N\eta} + \frac{1}{(N\eta)^2} + \Theta \prec \frac{\rho}{N\eta} + \frac{1}{(N\eta)^2} + \mathcal{E}.$$

This finishes the proof of Theorem 1.7 apart from the proof of Lemma 4.5 which we will tackle now.

Proof of Lemma 4.5 The spectral parameter $z = \tau_0 + \theta\omega + \mathrm{i}\eta$ lies inside the δ_*-neighborhood of τ_0. We fix $\omega \in [-\delta_*, \delta_*]$ and show that the claim holds for any choice of $\eta \in [N^{\gamma-1}, \delta_*]$. We split the interval of possible values of η into two or three regimes, depending on the case we are treating.

- **Edge:** If $\tau_0 \in \partial \operatorname{supp} \rho$ is an edge of a gap of size $\Delta := \Delta_0(\tau_0)$, then we define

$$I_1(\omega) := \left\{ \eta \in [N^{\gamma-1}, \delta_*] : \frac{(|\omega| + \eta)^{1/2}}{(|\omega| + \eta + \Delta)^{1/6}} \geq N^{-5\varepsilon}\Phi_k(\omega, \eta) \right\},$$
$$I_2(\omega) := \left\{ \eta \in [N^{\gamma-1}, \delta_*] : N^{5\varepsilon}\frac{(|\omega| + \eta)^{1/2}}{(|\omega| + \eta + \Delta)^{1/6}} \leq \Phi_k(\omega, \eta) \right.$$
$$\left. \leq N^{2\varepsilon}(|\omega| + \eta + \Delta)^{1/3} \right\},$$
$$I_3(\omega) := \left\{ \eta \in [N^{\gamma-1}, \delta_*] : (|\omega| + \eta + \Delta)^{1/3} \leq N^{-2\varepsilon}\Phi_k(\omega, \eta) \right\}.$$

  If any of the two regimes $I_l(\omega)$ with $l = 2, 3$ consists of a single point only, then we set $I_l(\omega) := \emptyset$.
- **Internal minimum:** If $\tau_0 \in \mathbb{M}\backslash\partial \operatorname{supp} \rho$, then we set $I_2(\omega) := \emptyset$ and define

$$I_1(\omega) := \left\{ \eta \in [N^{\gamma-1}, \delta_*] : \rho(\tau_0) + (|\omega| + \eta)^{1/3} \geq N^{-2\varepsilon}\Phi_k(\omega, \eta) \right\},$$
$$I_3(\omega) := \left\{ \eta \in [N^{\gamma-1}, \delta_*] : \rho(\tau_0) + (|\omega| + \eta)^{1/3} \leq N^{-2\varepsilon}\Phi_k(\omega, \eta) \right\}.$$

If $I_3(\omega)$ consists of a single point only, then we set $I_3(\omega) := \emptyset$.

In the cubic equation (4.30), used to define the error function $\mathcal{E}$, the coefficients $\widetilde{\pi}_1$ and $\widetilde{\pi}_2$ on the left hand side are monotonously increasing functions of η. The linear and the constant coefficient of $\mathcal{E}$ on the right hand side are monotonously decreasing in η. Thus, $\mathcal{E}$ itself is a monotonously decreasing function of η. From this fact and the definition of the regimes I_1, I_2 and I_3 we see that $I_1 = [\eta_1, \delta_*]$, $I_2 = [\eta_2, \eta_1]$ and $I_3 = [N^{\gamma-1}, \eta_2]$ for some $\eta_1, \eta_2 \in [N^{\gamma-1}, \delta_*]$. Here, we interpret $I_2 = \emptyset$ if $\eta_1 \leq \eta_2$ and $I_3 = \emptyset$ if $\eta_2 \leq N^{\gamma-1}$.

Now we define a z-dependent indicator function

$$\chi(\omega, \eta) := \begin{cases} \mathbb{1}\Big(N^{-7\varepsilon}\Phi_k(\omega,\eta) \leq \Theta(\tau_0 + \theta\omega + \mathrm{i}\eta) \leq N^{-6\varepsilon}\Phi_k(\omega,\eta)\Big) & \text{if } \eta \in I_1(\omega) \\ \mathbb{1}\Big(N^{-4\varepsilon}\Phi_k(\omega,\eta) \leq \Theta(\tau_0 + \theta\omega + \mathrm{i}\eta) \leq N^{-3\varepsilon}\Phi_k(\omega,\eta)\Big) & \text{if } \eta \in I_2(\omega) \ . \\ \mathbb{1}\Big(N^{-\varepsilon}\Phi_k(\omega,\eta) \leq \Theta(\tau_0 + \theta\omega + \mathrm{i}\eta) \leq \Phi_k(\omega,\eta)\Big) & \text{if } \eta \in I_3(\omega) \end{cases} \tag{4.42}$$

This function fixes the values of Θ to a small interval just below the deterministic control parameter Φ_k. We will prove that Θ cannot take these values, i.e. $\chi = 0$ a.w.o.p. Figure 3 illustrates this argument. Compared to Figure 6.1 in [14] we see that instead of two there are now three domains, $I_1(\omega)$, $I_2(\omega)$ and $I_3(\omega)$, to be distinguished. The reason for this extra complication is that (4.10) is cubic in Θ, compared to the quadratic equation for $[v]$ that appeared in the proof of Lemma 6.2 in [14]. To see that $\chi = 0$, first note that the choice of the domains, I_l, ensures that there is always one summand on the left hand side of the cubic equation (4.10) for Θ which dominates the two others by a factor N^{ε}, whenever χ does not vanish. In fact, by construction we have:

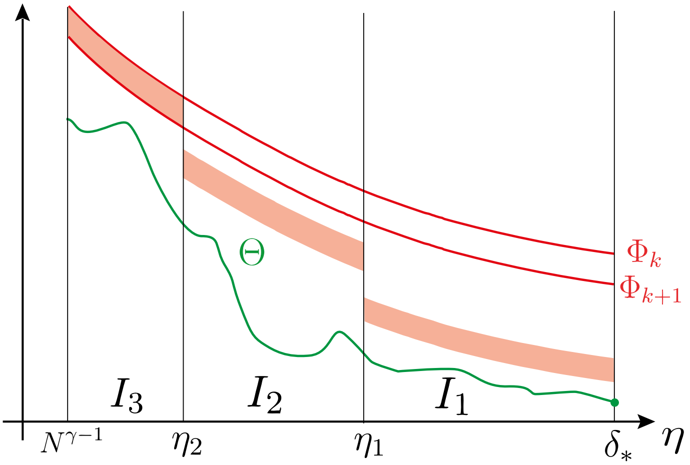

Fig. 3 The *shaded area* is forbidden for Θ. Since the continuous function Θ lies below this region at $\eta = \delta_*$ it stays below it for any $\eta \geq N^{\gamma-1}$, hence proving $\Theta \leq \Phi_{k+1}$

The random functions Θ and χ satisfy a.w.o.p.

$$\begin{aligned}&\left(\Theta(z)^3+\widetilde{\pi}_2(\omega,\eta)\Theta(z)^2+\widetilde{\pi}_1(\omega,\eta)\Theta(z)\right)\chi(\omega,\eta)\\&\lesssim\left|\Theta(z)^3+\pi_2(z)\Theta(z)^2+\pi_1(z)\Theta(z)\right|.\end{aligned}\tag{4.43}$$

We will verify this fact at the end of the proof of this lemma. Now we will simply use it. First we combine the assumption (4.36) of the lemma and (4.43) to obtain

$$N^{-\varepsilon}(\Theta^3+\widetilde{\pi}_2\Theta^2+\widetilde{\pi}_1\Theta)\,\chi\;\le\;\frac{\widetilde{\rho}+\Phi_k}{N\eta}+\frac{1}{(N\eta)^2}\quad\text{a.w.o.p.}$$

Here we also gave up a factor of N^{ε} to get an inequality instead of the stochastic domination, and replaced ρ by the comparable quantity $\widetilde{\rho}$. By the definition of the indicator function χ we have $\Theta\chi\ge N^{-7\varepsilon}\Phi_k$. Using this to bound the left hand side, and that $\varepsilon\le\widetilde{\varepsilon}$, we obtain

$$\begin{gathered}\big(\mathcal{R}^3+\widetilde{\pi}_2\mathcal{R}^2+\widetilde{\pi}_1\mathcal{R}\big)\chi\;\le\;N^{8\widetilde{\varepsilon}}\frac{\mathcal{R}}{N\eta}+\frac{\widetilde{\rho}}{N\eta}+\frac{1}{(N\eta)^2},\quad\text{a.w.o.p.,}\\ \mathcal{R}:=N^{-8\varepsilon}\Phi_k.\end{gathered}$$

Comparing this with the defining Eq. (4.30) for $\mathcal{E}$ we conclude that a.w.o.p. $N^{-8\varepsilon}\Phi_k\chi\le\mathcal{E}$.

On the other hand, by the definition of Φ_k in (4.35) we know that $\Phi_k>N^{8\varepsilon}\mathcal{E}$. These two inequalities yield

$$\chi(\omega,\eta)\;=\;0,\quad\eta\in[N^{\gamma-1},\delta_*],\quad\text{a.w.o.p.}\tag{4.44}$$

Now we successively, for $l=1,2,3$, apply Lemma A.2 on the connected domains $\tau_0+\theta\omega+\mathrm{i}I_l(\omega)$ with the choices $\varphi:=\Theta$ and

$$\Phi(\tau_0+\theta\omega+\mathrm{i}\eta):=\begin{cases}N^{-6\varepsilon}\Phi_k(\omega,\eta) & \text{if }\;l=1,\\ N^{-3\varepsilon}\Phi_k(\omega,\eta) & \text{if }\;l=2,\\ \Phi_k(\omega,\eta) & \text{if }\;l=3,\end{cases}$$

$$z_0:=\begin{cases}\tau_0+\theta\omega+\mathrm{i}\delta_* & \text{if }\;l=1,\\ \tau_0+\theta\omega+\mathrm{i}\eta_1 & \text{if }\;l=2,\\ \tau_0+\theta\omega+\mathrm{i}\eta_2 & \text{if }\;l=3,\end{cases}$$

where as explained after the definition of I_1, I_2 and I_3 above we have $I_1=[\eta_1,\delta_*]$, $I_2=[\eta_2,\eta_1]$ and $I_3=[N^{\gamma-1},\eta_2]$. The condition (A.1) of the lemma is satisfied by the definition of Θ in (4.7), the Hölder-continuity of the solution of the QVE, the weak Lipschitz-continuity of $\mathbf{g}$ with Lipschitz-constant N^2 and the Hölder-continuity of $\mathbf{s}$ from (4.12). The gap condition, (A.2), holds because of (4.44) and the definition of χ and Φ for an appropriate choice of the exponent D_3.

The condition, $\varphi(z_0) \leq \Phi(z_0)$ a.w.o.p., necessary for the application of Lemma A.2 on the first domain, $\tau_0+\theta\omega+\mathrm{i}I_1(\omega)$, is obtained form Proposition 3.1. With Lemma A.2 we propagate the bound to all $z \in \tau_0+\theta\omega+\mathrm{i}I_1(\omega)$. Now we apply Lemma A.2 on the second domain $\tau_0+\theta\omega+\mathrm{i}I_2(\omega)$, provided $I_2(\omega)$ is not empty. The bound (A.3) for the new $z_0 = \tau_0+\theta\omega+\mathrm{i}\eta_1$ is obtained from the previous step. Finally, we apply Lemma A.2 to $\tau_0+\theta\omega+\mathrm{i}I_3(\omega)$, in case it is not empty, with the new choice $z_0 = \tau_0+\theta\omega+\mathrm{i}\eta_2$. Altogether, we applied the lemma at most three times. Through this procedure we prove that a.w.o.p. $\Theta(z) \leq \Phi(z)$ for all $z \in \tau_0+\theta\omega+\mathrm{i}[N^{\gamma-1}, \delta_*]$. On the third domain, $\tau_0+\theta\omega+\mathrm{i}I_3(\omega)$, we use that a.w.o.p. $\chi = 0$ (cf. (4.44)) and thus a.w.o.p. $\Theta(z) \leq N^{-\varepsilon}\Phi_k$. Altogether we showed that in the δ_*-neighborhood of τ_0,

$$\text{a.w.o.p.}\quad \Theta(z) \;\leq\; N^{-\varepsilon}\Phi_k \;\leq\; \Phi_{k+1}.$$

This finishes the proof of Lemma 4.5 up to verifying the claim (4.43).

For the proof of (4.43) one verifies case by case that on I_1 the term $\widetilde{\pi}_1\Theta \sim |\pi_1|\Theta$ is bigger than the two other terms, $\widetilde{\pi}_2\Theta^2$ and Θ^3 by a factor of N^{ε}. If I_3 is not empty then the term Θ^3 is the biggest. If I_2 is not empty, then $|\pi_2| \sim \widetilde{\pi}_2$ and $\widetilde{\pi}_2\Theta^2$ is the biggest term by a factor of N^{ε}. More specifically, when $\eta \in I_j$ and $\chi = \chi(\omega,\eta) = 1$ we show

$$\left|\Theta^3+\pi_2\Theta^2+\pi_1\Theta\right| \;\sim\; |\pi_j|\Theta^j \;\sim\; \widetilde{\pi}_j\Theta^j \;\sim\; \Theta^3+\widetilde{\pi}_2\Theta^2+\widetilde{\pi}_1\Theta,$$

where $\pi_3 = \widetilde{\pi}_3 := 1$. As an example we demonstrate these relations in a few cases:

- **Well inside a gap:** If $\tau_0 \in \partial\,\mathrm{supp}\,\rho$ and $\omega \in [c_*\Delta, \Delta/2]$ then $I_2(\omega) = \emptyset$. We now check that on $I_1(\omega)$ the linear term in Θ is the biggest while on $I_3(\omega)$ the cubic term dominates. First, let $\eta \in I_1(\omega)$. Then the following chain of inequalities hold,

  $$\begin{aligned}\widetilde{\pi}_1\Theta \;\sim\; |\pi_1|\Theta \;\sim\; (\Delta+\eta)^{2/3}\Theta \gtrsim N^{-5\varepsilon}(\Delta+\eta)^{1/3}\Phi_k\Theta \;\sim\; N^{-5\varepsilon}\widetilde{\pi}_2\Phi_k\Theta\\ \gtrsim N^{-10\varepsilon}\Phi_k^2\Theta.\end{aligned}$$

  Here, we used (4.29), (4.15b), the definition of $I_1(\omega)$ and (4.27c) in the form $\widetilde{\pi}_2 \sim (\Delta+\eta)^{1/3}$. Now we can use χ to replace Φ_k by Θ. By definition of χ and since $\widetilde{\pi}_k \gtrsim |\pi_k|$ for $k = 1, 2$ we also get

  $$N^{-5\varepsilon}\widetilde{\pi}_2\Phi_k\Theta\chi \geq N^{\varepsilon}\widetilde{\pi}_2\Theta^2\chi \;\gtrsim\; N^{\varepsilon}|\pi_2|\Theta^2\chi, \quad N^{-10\varepsilon}\Phi_k^2\Theta\chi \geq N^{2\varepsilon}\Theta^3\chi.$$

  We conclude that on $I_1(\omega)$ the linear term in Θ dominates the others,

  $$\widetilde{\pi}_1\Theta\chi \;\gtrsim\; N^{\varepsilon}(\Theta^3+\widetilde{\pi}_2\Theta^2)\chi.$$

  Suppose now that $\eta \in I_3(\omega)$. In this case, using the choice of the indicator function χ,

  $$\Theta^3\chi \;\geq\; N^{-\varepsilon}\Phi_k\Theta^2\chi \;\geq\; N^{-2\varepsilon}\Phi_k^2\Theta\chi.$$

By definition of $I_3(\omega)$ and (4.27c) we find that

$$\begin{aligned} N^{-\varepsilon}\Phi_k\Theta^2 &\gtrsim N^{\varepsilon}(\Delta+\eta)^{1/3}\Theta^2 \sim N^{\varepsilon}\widetilde{\pi}_2\Theta^2,\\ N^{-2\varepsilon}\Phi_k^2\Theta &\gtrsim N^{2\varepsilon}(\Delta+\eta)^{2/3}\Theta \sim N^{2\varepsilon}\widetilde{\pi}_1\Theta. \end{aligned}$$

Altogether we find that the cubic term dominates the two others,

$$\Theta^3\chi \gtrsim N^{\varepsilon}(\widetilde{\pi}_2\Theta^2+\widetilde{\pi}_1\Theta)\chi.$$

- **Inside a gap close to an edge on** I_2: If $\tau_0\in\partial\operatorname{supp}\rho$, $\omega\in[0,c_*\Delta]$ and $\eta\in I_2(\omega)$, then we will show the quadratic term in Θ dominates the two other terms. We have

$$|\pi_2|\Theta^2 \sim \widetilde{\pi}_2\Theta^2 \sim (\Delta+\eta)^{1/3}\Theta^2 \gtrsim N^{-2\varepsilon}\Phi_k\Theta^2,$$

where in the inequality we used the definition of $I_2(\omega)$. The choice of χ guarantees that $\Phi_k\chi \ge N^{3\varepsilon}\Theta\chi$. Thus, the quadratic term is larger than the cubic term by a factor of N^{ε}. On the other hand

$$\begin{aligned} (\Delta+\eta)^{1/3}\Theta^2\chi &\gtrsim N^{-4\varepsilon}(\Delta+\eta)^{1/3}\Phi_k\Theta\\ &\gtrsim N^{\varepsilon}(\omega+\eta)^{1/2}(\Delta+\eta)^{1/6}\Theta \sim N^{\varepsilon}\widetilde{\pi}_1\Theta\\ &\sim N^{\varepsilon}|\pi_1|\Theta. \end{aligned}$$

Here, in the first inequality we used the indicator function χ and in the second inequality the definition of $I_2(\omega)$. Altogether, we arrive at

$$\widetilde{\pi}_2\Theta^2\chi \gtrsim N^{\varepsilon}(\Theta^3+\widetilde{\pi}_1\Theta)\chi.$$

- **Internal minimum on** I_1: If $\tau_0\in\mathbb{M}\backslash\partial\operatorname{supp}\rho$ and $\eta\in I_1(\omega)$, then the linear term is the biggest,

$$\begin{aligned} |\pi_1|\Theta &\sim \widetilde{\pi}_1\Theta \sim \big(\rho(\tau_0)^2+(|\omega|+\eta)^{2/3}\big)\Theta\\ &\gtrsim N^{-2\varepsilon}\big(\rho(\tau_0)+(|\omega|+\eta)^{1/3}\big)\Phi_k\Theta. \end{aligned}$$

Here, we used (4.29) and the definitions of $\widetilde{\pi}_1$ and $I_1(\omega)$, respectively. Since $\Phi_k\chi \ge N^{6\varepsilon}\Theta\chi$ and by the definition of $\widetilde{\pi}_2$ this shows that the linear term is larger than the quadratic term by a factor of $N^{4\varepsilon}$. In order to compare the linear with the cubic term we estimate further. By definition of $I_1(\omega)$,

$$N^{-2\varepsilon}\big(\rho(\tau_0)+(|\omega|+\eta)^{1/3}\big)\Phi_k\Theta \ge N^{-4\varepsilon}\Phi_k^2\Theta.$$

Again we use the lower bound on $\Phi_k\chi$ and get

$$N^{-4\varepsilon}\Phi_k^2\Theta\chi \ge N^{8\varepsilon}\Theta^3\chi.$$

Thus we showed that on the domain $I_1(\omega)$

$$\widetilde{\pi}_1 \Theta\, \chi \;\gtrsim\; N^{\varepsilon}\, (\Theta^3 + \widetilde{\pi}_2 \Theta^2) \chi \,.$$

The other cases are proven similarly. This completes the proof of (4.43). □

5 Rigidity and delocalization of eigenvectors

5.1 Proof of Corollary 1.10

Here we explain how the local law, Theorem 1.7, is used to estimate the difference between the cumulative density of states and the eigenvalue distribution function of the random matrix **H**. The following auxiliary result shows that the difference between two probability measures can be estimated in terms of the difference of their respective Stieltjes transforms. For completeness the proof is given in the appendix. It uses a Cauchy-integral formula that was also applied in the construction of the Helffer-Sjöstrand functional calculus (cf. [11]) and it appeared in different variants in [15,20,21].

Lemma 5.1 (Bounding measures by Stieltjes transforms) *There is a universal constant $C > 0$, such that for any two probability measures, ν_1 and ν_2, on the real line and any three numbers $\eta_1, \eta_2, \varepsilon \in (0, 1]$ with $\varepsilon \geq \max\{\eta_1, \eta_2\}$, the difference between the two measures evaluated on the interval $[\tau_1, \tau_2] \subseteq \mathbb{R}$, with $\tau_1 < \tau_2$, satisfies*

$$\begin{aligned} &\big|\nu_1([\tau_1, \tau_2]) \;-\; \nu_2([\tau_1, \tau_2])\big| \\ &\quad\leq\; C \Big(\nu_1([\tau_1 - \eta_1, \tau_1]) \;+\; \nu_1([\tau_2, \tau_2 + \eta_2]) \;+\; J_1 + J_2 + J_3 \Big). \end{aligned} \tag{5.1}$$

Here, the three contributions to the error, J_1, J_2 and J_3, are defined as

$$\begin{aligned} J_1 \;&:=\; \int_{\tau_1 - \eta_1}^{\tau_1} \mathrm{d}\omega \left(\mathrm{Im}\, m_{\nu_1}(\omega + \mathrm{i}\eta_1) + |m_{\nu_1 - \nu_2}(\omega + \mathrm{i}\eta_1)| \right. \\ &\qquad \left. + \frac{1}{\eta_1} \int_{\eta_1}^{2\varepsilon} \mathrm{d}\eta\, |m_{\nu_1 - \nu_2}(\omega + \mathrm{i}\eta)| \right), \\ J_2 \;&:=\; \int_{\tau_2}^{\tau_2 + \eta_2} \mathrm{d}\omega \left(\mathrm{Im}\, m_{\nu_1}(\omega + \mathrm{i}\eta_2) + |m_{\nu_1 - \nu_2}(\omega + \mathrm{i}\eta_2)| \right. \\ &\qquad \left. + \frac{1}{\eta_2} \int_{\eta_2}^{2\varepsilon} \mathrm{d}\eta\, |m_{\nu_1 - \nu_2}(\omega + \mathrm{i}\eta)| \right), \\ J_3 \;&:=\; \frac{1}{\varepsilon} \int_{\tau_1 - \eta_1}^{\tau_2 + \eta_2} \mathrm{d}\omega \int_{\varepsilon}^{2\varepsilon} \mathrm{d}\eta\, |m_{\nu_1 - \nu_2}(\omega + \mathrm{i}\eta)| \,, \end{aligned} \tag{5.2}$$

where m_ν denotes the Stieltjes transform of ν for any signed measure ν.

We will now apply this lemma to prove Corollary 1.10 with the choices of the measures

$$\nu_1(\mathrm{d}\omega) := \rho(\omega)\mathrm{d}\omega, \quad \text{and} \quad \nu_2(\mathrm{d}\omega) := \frac{1}{N}\sum_{i=1}^{N}\delta_{\lambda_i}(\mathrm{d}\omega). \tag{5.3}$$

As a first step we show that a.w.o.p. there are no eigenvalues with an absolute value larger or equal than 10, i.e.,

$$\#\{i : |\lambda_i| \geq 10\} = 0 \quad \text{a.w.o.p.} \tag{5.4}$$

We focus on the eigenvalues $\lambda_i \geq 10$. The ones with $\lambda_i \leq -10$ are treated in the same way. We will show first that there are no eigenvalues in a small interval around τ with $\tau \geq 10$. In fact, we prove that for $\gamma \in (0, 1/3)$,

$$\#\big\{i : \tau \leq \lambda_i \leq \tau + N^{-1}\big\} \prec N^{-\gamma}. \tag{5.5}$$

For this we apply Lemma 5.1 with the same choices of the measures ν_1 and ν_2 as in (5.3) and with

$$\eta_1 := \eta_2 := \varepsilon := N^{\gamma-1}, \quad \tau_1 := \tau, \quad \tau_2 := \tau + N^{-1}. \tag{5.6}$$

Theorem 1.7 takes the form

$$\big|\langle \mathbf{g}(\omega + \mathrm{i}\eta)\rangle - \langle \mathbf{m}(\omega + \mathrm{i}\eta)\rangle\big| \prec \frac{1}{N} + N^{-2\gamma}, \quad (\omega, \eta) \in \Gamma, \tag{5.7}$$

where $\Gamma := [\tau - N^{\gamma-1}, \tau + 2N^{\gamma-1}] \times [N^{\gamma-1}, 2N^{\gamma-1}]$. Here we used $\kappa(\omega + \mathrm{i}\eta) \lesssim \eta_1 + (N\eta)^{-1}$, that follows from the facts that we are well outside $\operatorname{supp}\rho \subset [-2, 2]$, and hence $\Delta(\omega) = 1$ by (1.17) so the condition (1.24) holds, and thus (1.25) is applicable.

Using the definition of stochastic domination (Definition 1.6), the basic union bound, and the part (iii) of Lemma A.1 we see that the estimate (5.7) holds even with supremum over $(\omega, \eta) \in \widehat{\Gamma}$, where $\widehat{\Gamma} := (N^{-10}\mathbb{Z})^2 \cap \Gamma$ is a fine grid of spacing N^{-10} with $|\widehat{\Gamma}| \leq N^{20}$. Using the Lipschitz-continuity of $z \mapsto \langle \mathbf{g}(z)\rangle$ with Lipschitz-constant bounded by N^2, as well as the uniform $1/3$-Hölder-continuity of $z \mapsto \langle \mathbf{m}(z)\rangle$, we can extend the supremum over $\widehat{\Gamma}$ to the entire domain Γ, i.e.,

$$\sup_{(\omega,\eta)\in\Gamma} \big|\langle \mathbf{g}(\omega + \mathrm{i}\eta)\rangle - \langle \mathbf{m}(\omega + \mathrm{i}\eta)\rangle\big| \prec \frac{1}{N} + N^{-2\gamma}.$$

Plugging this bound into the definitions of J_1, J_2 and J_3 from (5.2) and using (5.1) and the fact that $\rho = 0$ in this regime shows the validity of (5.5).

We conclude that a.w.o.p. there are no eigenvalues in an interval of length N^{-1} to the right of τ. By using a union bound this implies that

$$\#\{i : 10 \leq \lambda_i \leq N\} = 0 \quad \text{a.w.o.p.}$$

The eigenvalues larger than N are treated by the following simple argument,

$$\max_{i=1}^{N} \lambda_i^2 \leq \sum_{i=1}^{N} \lambda_i^2 = \sum_{i,j=1}^{N} |h_{ij}|^2 \prec N.$$

Thus (5.4) holds true.

Now we apply Lemma 5.1 to prove (1.28). In case $|\tau| \geq 10$ the bound (1.28) follows because a.w.o.p. there are no eigenvalues of $\mathbf{H}$ with absolute value larger or equal than 10. Thus, we fix $\tau \in (-10, 10)$ and make the choices

$$\eta_1 := \eta_2 := N^{\gamma-1}, \quad \tau_1 := -10, \quad \tau_2 := \tau, \quad \varepsilon := 1. \tag{5.8}$$

Again we use (1.21) from Theorem 1.7, the Lipschitz-continuity of $\langle \mathbf{g} \rangle$ and the Hölder-continuity of $\langle \mathbf{m} \rangle$ to see that uniformly for all $\eta \geq N^{\gamma-1}$,

$$\sup_{\omega \in [0,\eta_1]} \big| \langle \mathbf{g}(\tau_1 - \omega + \mathrm{i}\eta) \rangle - \langle \mathbf{m}(\tau_1 - \omega + \mathrm{i}\eta) \rangle \big| \prec \frac{1}{N} + \frac{1}{(N\eta)^2}.$$

Here we evaluated $\Delta(\tau_1) = 1$ and thus $\kappa \lesssim \eta + (N\eta)^{-1}$. With J_1 defined as in (5.2) we infer $J_1 \prec N^{-1}$. Theorem 1.7 also implies the bound

$$\sup_{\omega \in [-20,20]} \sup_{\eta \in [1,2]} \big| \langle \mathbf{g}(\omega + \mathrm{i}\eta) \rangle - \langle \mathbf{m}(\omega + \mathrm{i}\eta) \rangle \big| \prec \frac{1}{N},$$

since in this regime $\kappa \lesssim 1$, thus showing that $J_3 \prec N^{-1}$. We are left with estimating the three terms constituting J_2. The first and second of these terms are estimated trivially by using the boundedness of their integrands. Therefore, we conclude that

$$\left| \int_{-10}^{\tau} \rho(\omega) \mathrm{d}\omega - \frac{\#\{\, i : -10 \leq \lambda_i \leq \tau \}}{N} \right| \prec N^{\gamma-1} + R(\tau), \tag{5.9}$$

where the error term, R, is defined as

$$\begin{aligned} R(\tau) := N^{1-\gamma} \int_0^{N^{\gamma-1}} \mathrm{d}\omega \int_{N^{\gamma-1}}^{2} \mathrm{d}\eta \\ \min \left\{ \frac{1}{N\eta(\Delta(\tau+\omega)^{1/3} + \rho(\tau+\omega+\mathrm{i}\eta))}, \frac{1}{(N\eta)^{1/2}} \right\}. \end{aligned} \tag{5.10}$$

This expression is derived by using the bound (1.23) on κ for the integrand of the third contribution to J_2.

To estimate R further we distinguish three cases, depending on whether τ is away from $\mathbb{M}$, close to an edge or close to a local minimum in the interior of supp ρ. In each of these cases we prove

$$R(\tau) \prec \min\left\{ \frac{1}{N(\Delta(\tau)^{1/3}+\rho(\tau))}, \frac{1}{N^{4/5}} \right\}. \tag{5.11}$$

Away from $\mathbb{M}$: In case $\operatorname{dist}(\tau, \mathbb{M}) \geq \delta_*$, with δ_* the size of the neighborhood around the local minima from Theorem 4.1, we have $\Delta^{1/3}+\rho \sim 1$ and thus the η-integral in (5.10) yields a factor comparable to $N^{-1}\log N$. Thus, $R(\tau) \prec N^{-1}$, and hence (5.11) holds.

Close to an edge: Let $\operatorname{dist}(\tau, \{\alpha_k, \beta_k\}) \leq \delta_*$. Then from the size of ρ at an internal edge, at the extreme edges and inside the gap (cf. (4.5b), (4.5d) and (4.5c) from Theorem 4.1) we see that

$$\Delta(\tau+\omega)^{1/3} + \rho(\tau+\omega+\mathrm{i}\eta) \sim \big(\Delta(\tau)+\operatorname{dist}(\tau,\{\alpha_k,\beta_k\})+\eta\big)^{1/3}.$$

for any $\omega \in [0, N^{\gamma-1}]$ and $\eta \in [N^{\gamma-1}, 2]$. With this the size of R is given by

$$R(\tau) \sim \int_{N^{\gamma-1}}^{2} \mathrm{d}\eta \min\left\{ \frac{1}{N\eta(\Delta(\tau)+\operatorname{dist}(\tau,\{\alpha_k,\beta_k\})+\eta)^{1/3}}, \frac{1}{(N\eta)^{1/2}} \right\}.$$

Integrating over η yields that

$$R(\tau) \lesssim \min\left\{ \frac{\log N}{N(\Delta(\tau)+\operatorname{dist}(\tau,\{\alpha_k,\beta_k\}))^{1/3}}, \frac{1}{N^{4/5}} \right\}.$$

Now (5.11) follows by using the size of ρ from Theorem 4.1 again.

Close to an internal local minimum: Suppose $|\tau-\tau_0| \leq \delta_*$ for some $\tau_0 \in \mathbb{M}\backslash\partial \operatorname{supp}\rho$. Then using the size of ρ from (4.5e) of Theorem 4.1 we see that

$$R(\tau) \sim \int_{N^{\gamma-1}}^{2} \mathrm{d}\eta \min\left\{ \frac{1}{N\eta(\rho(\tau_0)+|\tau-\tau_0|^{1/3}+\eta^{1/3})}, \frac{1}{(N\eta)^{1/2}} \right\}.$$

The bound (5.11) follows by performing the integration over η.

This finishes the proof of (5.11). We insert this bound into (5.9) and use that γ was arbitrary. Thus, we find

$$\left| \int_{-10}^{\tau} \rho(\omega)\mathrm{d}\omega - \frac{\#\{i : -10 \leq \lambda_i \leq \tau\}}{N} \right| \prec \min\left\{ \frac{1}{N(\Delta(\tau)^{1/3}+\rho(\tau))}, \frac{1}{N^{4/5}} \right\}.$$

This finishes the proof of (1.28) since there are no eigenvalues below -10.

Now we prove (1.29). Let $\tau \in \mathbb{R}\backslash\operatorname{supp}\rho$. Suppose that for some $k = 1, \ldots, K$ we have $|\tau-\beta_k| = \operatorname{dist}(\tau, \partial \operatorname{supp}\rho)$. The case when τ is closer to the set $\{\alpha_k\}$ than to $\{\beta_k\}$ is treated similarly. Suppose further that

$$\tau \geq \alpha_k + \delta_k,$$

where δ_k are defined as in (1.30) and $\delta_0 = N^{\gamma-2/3}$. Note that there is nothing to show if $k > 1$ and the size of the gap, $\alpha_k - \beta_{k-1}$, is smaller than $2\delta_k$, i.e., if such a τ does not exist. In particular, we have $\alpha_k - \beta_{k-1} = \Delta(\tau) \gtrsim N^{-1/2}$. We will show that a.w.o.p. there are no eigenvalues in an interval of length $N^{-2/3}$ to the right of τ, i.e.

$$\#\big\{ i : \tau \le \lambda_i \le \tau + N^{-2/3} \big\} = 0 \quad \text{a.w.o.p.} \tag{5.12}$$

We apply Lemma 5.1 with the same choices of the measures ν_1 and ν_2 as in (5.3). Additionally, we set

$$\eta_1 := \eta_2 := \varepsilon := N^{-2/3}, \quad \tau_1 := \tau, \quad \tau_2 := \tau + N^{-2/3}. \tag{5.13}$$

We use the local law, Theorem 1.7, to estimate the differences between the Stieltjes transforms of the two measures for the integrands in the definition of the three error terms, J_1, J_2 and J_3 from (5.2). By the definition of δ_k the condition (1.24) is satisfied inside the integrals and we use the improved bound, (1.25), on κ. Indeed, we find

$$\sup \big| \langle \mathbf{g}(\omega + \mathrm{i}\eta) \rangle - \langle \mathbf{m}(\omega + \mathrm{i}\eta) \rangle \big| \prec \frac{1}{N\delta_k \Delta(\tau)^{1/3}} + \frac{1}{N^{2/3}\delta_k^{1/2}\Delta(\tau)^{1/6}},$$

where the supremum is taken over $\omega \in [\tau - N^{-2/3}, \tau + 2N^{-2/3}]$ and $\eta \in [N^{-2/3}, 2N^{-2/3}]$. With this, the definition of δ_k and the size of ρ from (4.5c) and (4.5d) we infer

$$J_1 + J_2 + J_3 \prec N^{-1-\gamma/2}.$$

From this (5.12) follows. The claim, (1.29), is now a consequence of a simple union bound taken over the events in (5.12) with different choices of τ. This finishes the proof of Corollary 1.10.

5.2 Proof of Corollary 1.11

Here we show how we get the rigidity, Corollary 1.11, from Corollary 1.10. Fix a $\tau \in [\alpha_1, \beta_K]$. We define the random fluctuation to the left, δ_-, and to the right, δ_+, of the eigenvalue $\lambda_{i(\tau)}$ as

$$\begin{aligned} \delta_+(\tau) &:= \inf\bigg\{ \delta \ge 0 : 2 + \Big| \#\big\{ i : \lambda_i \le \tau + \delta \big\} - N\int_{-\infty}^{\tau+\delta} \rho(\omega)\mathrm{d}\omega \Big| \\ &\qquad \le N \int_{\tau}^{\tau+\delta} \rho(\omega)\mathrm{d}\omega \bigg\} \end{aligned} \tag{5.14}$$

$$\begin{aligned} \delta_-(\tau) &:= \inf\bigg\{ \delta \ge 0 : 1 + \Big| \#\big\{ i : \lambda_i \le \tau - \delta \big\} - N\int_{-\infty}^{\tau-\delta} \rho(\omega)\mathrm{d}\omega \Big| \\ &\qquad \le N \int_{\tau-\delta}^{\tau} \rho(\omega)\mathrm{d}\omega \bigg\}. \end{aligned} \tag{5.15}$$

We show now that with this definition,

$$\lambda_{i(\tau)} \in \big[\tau - \delta_-(\tau), \tau + \delta_+(\tau)\big]. \tag{5.16}$$

We start with the upper bound on $\lambda_{i(\tau)}$. By the definition of $i(\tau)$ we find the inequality

$$\begin{aligned}\#\big\{i : \lambda_i \le \lambda_{i(\tau)}\big\} = i(\tau) \;\le\; & 1 + N \int_{-\infty}^{\tau} \rho(\omega)\mathrm{d}\omega \\ = & 1 + N \int_{-\infty}^{\tau+\delta_+} \rho(\omega)\mathrm{d}\omega - N \int_{\tau}^{\tau+\delta_+} \rho(\omega)\mathrm{d}\omega.\end{aligned}$$

The definition of $\delta_+ = \delta_+(\tau)$ implies that

$$\#\big\{i : \lambda_i \le \lambda_{i(\tau)}\big\} < \#\{i : \lambda_i \le \tau + \delta_+\}.$$

By monotonicity of the cumulative eigenvalue distribution, we conclude that $\lambda_{i(\tau)} \le \tau + \delta_+$. Thus, the upper bound is proven.

Now we show the lower bound. We start similarly,

$$\begin{aligned}\#\big\{i : \lambda_i \le \lambda_{i(\tau)}\big\} = i(\tau) \;\ge\; & N \int_{-\infty}^{\tau} \rho(\omega)\mathrm{d}\omega \\ = & N \int_{-\infty}^{\tau-\delta_-} \rho(\omega)\mathrm{d}\omega \,+\, N \int_{\tau-\delta_-}^{\tau} \rho(\omega)\mathrm{d}\omega.\end{aligned}$$

By definition of δ_- we get

$$\#\big\{i : \lambda_i \le \lambda_{i(\tau)}\big\} \;\ge\; 1 + \liminf_{\varepsilon \downarrow 0} \#\big\{i : \lambda_i \le \tau - \delta_- - \varepsilon\big\}.$$

Here the lim inf is necessary, since the cumulative eigenvalue distribution is not continuous from the left. We conclude that $\lambda_{i(\tau)} \ge \tau - \delta_- - \varepsilon$ for all $\varepsilon > 0$ and therefore the lower bound is proven.

Now we start with the proof of (1.34). For this we show that for any τ that is well inside the support of the density of states, i.e., that satisfies (1.33), we have

$$\delta_-(\tau) + \delta_+(\tau) \;\prec\; \delta, \quad \delta := \min\left\{ \frac{1}{\rho(\tau)(\Delta(\tau)^{1/3} + \rho(\tau))N}, \frac{1}{N^{3/5}} \right\}. \tag{5.17}$$

If τ is in the bulk, i.e., $\operatorname{dist}(\tau, \mathbb{M}) \ge \delta_*$, then $\delta \sim N^{-1}$ and thus (5.17) follows from (1.28). We distinguish the two remaining cases, namely whether τ is close to an edge or to a local minimum inside the interior of supp ρ.

Close to an edge: Suppose that $\tau \in [\beta_k - \delta_*, \beta_k - \varepsilon_k]$. The case when τ is closer to $\{\alpha_k\}$ than to $\{\beta_k\}$ is treated similarly. By the definition of ε_k in (1.32) and by the size of

ρ from (4.5d) and (4.5b) in Theorem 4.1 we see that $\varepsilon_k \gtrsim N^\gamma \delta$. Using Corollary 1.10 we find for any $\varepsilon \in (0, \gamma/2)$ that

$$\biggl| \#\bigl\{ i : \lambda_i \le \tau + N^\varepsilon \delta \bigr\} - N \int_{-\infty}^{\tau + N^\varepsilon \delta} \rho(\omega) \mathrm{d}\omega \biggr| \prec \min\Bigl\{ \bigl(\Delta(\tau) + \beta_k - \tau\bigr)^{-1/3}, N^{1/5} \Bigr\}.$$

On the other hand

$$\begin{aligned} N \int_{\tau}^{\tau + N^\varepsilon \delta} \rho(\omega) \mathrm{d}\omega &\sim \frac{N^{1+\varepsilon} \delta (\beta_k - \tau)^{1/2}}{(\Delta(\tau) + \beta_k - \tau)^{1/6}} \\ &\gtrsim N^\varepsilon \min\Bigl\{ \bigl(\Delta(\tau) + \beta_k - \tau\bigr)^{-1/3}, N^{1/5} \Bigr\}. \end{aligned}$$

Here we used the size of ρ from Theorem 4.1, the definition of δ and $\beta_k - \tau \ge \varepsilon_k$. Since ε was arbitrary we conclude that $\delta_+(\tau) \prec \delta$. The bound, $\delta_-(\tau) \prec \delta$, is shown in the same way.

Close to internal local minima: Suppose $|\tau - \tau_0| \le \delta_*$ for some $\tau_0 \in \mathbb{M} \backslash \partial \operatorname{supp} \rho$. Then by (4.5e) with $\Delta(\tau_0) = 0$ and the definition of δ in (5.17) we have

$$\delta \sim \min\left\{ \frac{1}{(\rho(\tau_0)^3 + |\tau - \tau_0|)^{2/3} N}, \frac{1}{N^{3/5}} \right\}.$$

We apply (1.28) from Corollary 1.10 and, using (4.5e) again, we get

$$\begin{aligned} &\biggl| \#\bigl\{ i : \lambda_i \le \tau + N^\varepsilon \delta \bigr\} - N \int_{-\infty}^{\tau + N^\varepsilon \delta} \rho(\omega) \mathrm{d}\omega \biggr| \\ &\quad \prec \min\Bigl\{ \bigl(\rho(\tau_0)^3 + |\tau + N^\varepsilon \delta - \tau_0|\bigr)^{-1/3}, N^{1/5} \Bigr\}. \end{aligned} \tag{5.18}$$

On the other hand, we find

$$N \int_{\tau}^{\tau + N^\varepsilon \delta} \rho(\omega) \mathrm{d}\omega \sim N^{1+\varepsilon} \delta \Bigl(\rho(\tau_0)^3 + |\tau - \tau_0| + N^\varepsilon \delta \Bigr)^{1/3}. \tag{5.19}$$

We will now verify that for large enough N,

$$\begin{aligned} &N^{\varepsilon/2} \min\Bigl\{ \bigl(\rho(\tau_0)^3 + |\tau + N^\varepsilon \delta - \tau_0| \bigr)^{-1/3}, N^{1/5} \Bigr\} \\ &\lesssim N^{1+\varepsilon} \delta \Bigl(\rho(\tau_0)^3 + |\tau - \tau_0| + N^\varepsilon \delta \Bigr)^{1/3}. \end{aligned} \tag{5.20}$$

We distinguish three cases. First let us consider the regime where $\rho(\tau_0)^3 + |\tau - \tau_0| \le N^{-3/5}$. Then we have $\delta = N^{-3/5}$ and

$$N^{1+\varepsilon} \delta \bigl(\rho(\tau_0)^3 + |\tau - \tau_0| + N^\varepsilon \delta \bigr)^{1/3} \sim N^{4\varepsilon/3} N^{1/5}.$$

Now we treat the situation where, $N^{-3/5} < \rho(\tau_0)^3 + |\tau - \tau_0| \le N^{3\varepsilon/2-3/5}$. In this case

$$N^{1+\varepsilon}\delta\left(\rho(\tau_0)^3 + |\tau - \tau_0| + N^\varepsilon\delta\right)^{1/3} \gtrsim \frac{N^\varepsilon}{(\rho(\tau_0)^3 + |\tau - \tau_0|)^{1/3}} \ge N^{\varepsilon/2}N^{1/5}.$$

Finally, we consider $\rho(\tau_0)^3 + |\tau - \tau_0| > N^{3\varepsilon/2-3/5}$. Then for large enough N we find on the one hand

$$\min\left\{\left(\rho(\tau_0)^3 + |\tau + N^\varepsilon\delta - \tau_0|\right)^{-1/3}, N^{1/5}\right\} \sim \frac{1}{(\rho(\tau_0)^3 + |\tau - \tau_0|)^{1/3}},$$

and on the other hand

$$N^{1+\varepsilon}\,\delta\left(\rho(\tau_0)^3 + |\tau - \tau_0| + N^\varepsilon\delta\right)^{1/3} \gtrsim \frac{N^\varepsilon}{(\rho(\tau_0)^3 + |\tau - \tau_0|)^{1/3}}.$$

Thus, (5.20) holds true and since ε was arbitrary, we infer from (5.18) and (5.19) that $\delta_+(\tau) \prec \delta$. Along the same lines we prove $\delta_-(\tau) \prec \delta$. Thus (5.17) and with it (1.34) are proven.

The statement about the fluctuation of the eigenvalues at the leftmost edge, (1.35) follows directly from (1.34) and (1.29) in Corollary 1.10. Indeed, for $\tau \in [\alpha_1, \alpha_1 + \varepsilon_0)$ we have $\lambda_{i(\tau)} \le \lambda_{i(\alpha_1+\varepsilon_0)}$ and from (1.34) with $\Delta(\tau) = 1$, as well as $\rho(\alpha_1+\varepsilon_0) \sim \varepsilon_0^{1/2}$, and from the definition of ε_0 we see that

$$\lambda_{i(\alpha_1+\varepsilon_0)} \le \alpha_1 + \varepsilon_0 + N^{\gamma-2/3} \le \tau + 2N^{\gamma-2/3} \quad \text{a.w.o.p.}$$

On the other hand, (1.29) shows that a.w.o.p. $\lambda_{i(\tau)} \ge \alpha_1 - N^{\gamma-2/3}$. Since γ was arbitrary, (1.35) follows. The rigidity at the rightmost edge is proven along the same lines.

The claim, (1.36), about the remaining eigenvalues follows from a similar argument. For $\tau \in (\beta_k - \varepsilon_k, \alpha_{k+1} + \varepsilon_k)$, as a consequence of (1.29), we have

$$\lambda_{i(\tau)} \in \left[\lambda_{i(\beta_k-\varepsilon_k)}, \beta_k + \delta_k\right] \cup \left[\alpha_{k+1} - \delta_k, \lambda_{i(\alpha_{k+1}+\varepsilon_k)}\right] \quad \text{a.w.o.p.}$$

From (1.34) and the definition of ε_k we infer $\lambda_{i(\beta_k-\varepsilon_k)} \ge \beta_k - 2\varepsilon_k$ a.w.o.p., as well as $\lambda_{i(\alpha_{k+1}+\varepsilon_k)} \le \alpha_{k+1} + 2\varepsilon_k$ a.w.o.p., which finishes the proof of (1.36).

5.3 Proof of Corollary 1.14

The delocalization of eigenvectors is a simple consequence of the anisotropic local law Theorem 1.13 using the argument from [14]. Expressing the resolvent in the eigenbasis, we have

$$\mathbf{b} \cdot \mathbf{G}(z)\mathbf{b} = \sum_{i=1}^N \frac{|\mathbf{b} \cdot \mathbf{u}^{(i)}|^2}{\lambda_i - z}, \tag{5.21}$$

where $\mathbf{u}^{(i)}$ is the ℓ^2-normalised eigenvector corresponding to the eigenvalue λ_i. We evaluate this at $z := \lambda_k + \mathrm{i}N^{\gamma-1}$ with $\gamma > 0$ as in the statement of Theorem 1.13. The anisotropic local law implies that also $\mathbf{b} \cdot \mathbf{G}(z)\mathbf{b}$ is uniformly bounded. Hence we get

$$1 \gtrsim \operatorname{Im} \mathbf{b} \cdot \mathbf{G}(z)\mathbf{b} \ge N^{1-\gamma} |\mathbf{b} \cdot \mathbf{u}^{(k)}|^2,$$

by keeping only a single summand $i = k$ from (5.21). As $\gamma > 0$ was arbitrary we conclude that

$$|\mathbf{b} \cdot \mathbf{u}^{(k)}| \prec N^{-1/2},$$

uniformly in k.

6 Anisotropic law and universality

6.1 Proof of Theorem 1.13

Given the entrywise local law, Theorem 1.7, the proof of the anisotropic law follows exactly as in Section 7 of [9], where the same argument was presented for generalized Wigner matrices (this argument itself mimicked the detailed proof of the isotropic law for sample covariance matrices in Section 5 of [9]). The only difference is that in our case $G_{ii}(z)$ is close to $m_i(z)$, the i-th component of the solution to the QVE, which now genuinely depends on i, while in [9] we had $G_{ii}(z) \approx m_{\mathrm{sc}}(z)$ for every i, where $m_{\mathrm{sc}}(z)$ is the solution to (1.3). However, the diagonal resolvent elements played no essential role in [9]. We now explain the small modifications.

Recall from Section 5.2 of [9] that by polarization it is sufficient to prove (1.37) for ℓ^2-normalized vectors $\mathbf{w} = \mathbf{v}$. We can then write

$$\sum_{i,j=1}^{N} \overline{v}_i\, G_{ij} v_j - \sum_{i=1}^{N} m_i |v_i|^2 = \sum_i (G_{ii} - m_i)|v_i|^2 + \mathcal{Z}, \quad \mathcal{Z} := \sum_{i \neq j}^{N} \overline{v_i} G_{ij} v_j.$$

The first term containing the diagonal elements G_{ii} is clearly bounded by the right hand side of (1.37) by Theorem 1.7. This is the first instant where the nontrivial i-dependence of m_i is used.

The main technical part of the proof in [9] is then to control $\mathcal{Z}$, the contribution of the off diagonal terms. We can follow this proof in our case to the letter; the nontrivial i-dependence of m_i requires a slight modification only at one point. To see this, we recall the main structure of the proof. For any even p, the moment

$$\mathbb{E}|\mathcal{Z}|^p = \mathbb{E} \sum_{b_{11} \neq b_{12}} \cdots \sum_{b_{p1} \neq b_{p2}} \left(\prod_{k=1}^{p/2} \overline{v}_{b_{k1}} G_{b_{k1} b_{k2}} v_{b_{k2}} \right) \left(\prod_{k=p/2+1}^{p} \overline{v}_{b_{k1}} G^*_{b_{k1} b_{k2}} v_{b_{k2}} \right), \tag{6.1}$$

is computed. Let us concentrate on a fixed summand in (6.1) and let $B = \{b_{k_1}\} \cup \{b_{k_2}\}$ be the set of $\mathbf{v}$-indices appearing in that term. Using the resolvent identity (2.9) we successively expand the resolvents until each of them appears in a *maximally expanded* form, where every resolvent entry is of the form $G_{ab}^{(B\backslash ab)}$, for some $a, b \in B$ (cf. Definition 5.4 of [9]). Each time a maximally expanded off-diagonal element is produced we use (2.3). Finally, unless we end up with an expression that contains a very large numbers of off-diagonal resolvent entries (such *trivial leaves* are treated separately in Subsection 5.11 of [9]) we apply (3.16) to expand the remaining maximally expanded diagonal resolvent entries. This way we end up with an expression where only the resolvent entries of the type $G_{ij}^{(B)}$, with $i, j \notin B$, appear. In other words, the $\mathbf{v}$-indices and the indices of the resolvent entries are completely decoupled; only explicit products of entries of $\mathbf{H}$ represent the connections between them. We can now take partial expectation w.r.t. the rows and columns of these h-terms. In this way we guarantee that each index in B appears at least twice as a value of b_{k1} or b_{k2} in (6.1), i.e., the entries of $\mathbf{v}$ must be at least paired, and therefore the $2p$-fold summation in (6.1) effectively becomes at most a p-fold summation. This renders the uncontrolled ℓ^1-norm of $\mathbf{v}$ to ℓ^q-norms of $\mathbf{v}$, with $q \geq 2$, which are bounded by one by normalization.

Along this procedure it is only at the treatment of the maximally expanded diagonal resolvent elements appearing in the non-trivial leaves (cf. Subsection 5.12 of [9]) where we need to slightly adjust the proof to the setting where $\mathbf{S}$ is not stochastic. Using the QVE (1.7) and Schur's formula, similarly as in (3.16), we obtain a representation, where all the dependence of the B-columns and -rows of $\mathbf{H}$ is explicit

$$\frac{1}{G_{bb}^{(B\backslash b)}} = \frac{1}{m_b} - \sum_{i,j}^{(B)} \Big(h_{bi} G_{ij}^{(B)} h_{jb} - s_{bi} m_i \delta_{ij} \Big) + \sum_{a \in B} s_{ba} m_a + h_{bb}, \quad b \in B. \tag{6.2}$$

This formula replaces (5.41) from [9]. Taking the inverse of this formula and expanding around the leading term m_b, we get a geometric series representation for $G_{bb}^{(B\backslash b)}$ in terms of powers of the last three term in (6.2). The resulting formula is analogous to (5.42) in [9]. The geometric series converges because the last three term on the right hand side of (6.2) are much smaller than $|1/m_b| \sim 1$ a.w.o.p. Indeed, the last two terms in (6.2) are of size N^{-1} and $N^{-1/2+c}$ a.w.o.p., respectively. The double sum in (6.2) is small by using the large deviation estimates (2.7a)–(2.7c), similarly as in the proof of Lemma 2.1. When estimating the diagonal sum $i = j$, we note that $|G_{ii}^{(B)} - m_i|$ is small by first estimating $|G_{ii}^{(B)} - G_{ii}|$ similarly to (2.12), and then we use the local law Theorem 1.7 to see that also $|G_{ii} - m_i|$ is small.

The proof in [9] did not use the specific form of the subtracted term $s_{bi} m_i \delta_{ij}$ in (6.2), just the fact that the subtraction made (2.7c) applicable for the double summation in (6.2). After this slight modification, the rest of the proof in [9] goes through without any further changes.

6.2 Proof of Theorem 1.16

For the proof of Theorem 1.16 we follow the method developed in [14,17,20]. Theorem 2.1 from [18] was designed for proving universality for a random matrix with a small independent Gaussian component and densities of state that may differ from Wigner's semicircle law. The main theorem in [18] asserts that if local laws hold in a sufficiently strong sense then bulk universality holds locally for matrices with a small Gaussian component. We remark that a similar approach was independently developed in [26] that can also be easily used to conclude bulk universality from Theorem 1.7, but here we follow [18]. In Section 2.5 of [18] a recipe was given how to use this theorem to establish universality for a quite general class of random matrix models even without the Gaussian component, as long as uniform local laws on the optimal scale are known and the matrix satisfies the appropriate q-fullness condition (cf. Definition 1.15) that allows for an application of the moment matching (Lemma 6.5 in [20]) and the Green's function comparison theorem (Theorem 2.3 in [20]).

Let $\mathbf{H}$ be the Wigner-type matrix satisfying the hypotheses of Theorem 1.16, and for which the universality is to be proven. Let τ be a bulk point of ρ, so that $\rho(\tau) \geq \varepsilon$, for some $\varepsilon > 0$, and let $I := [\tau - \delta, \tau + \delta]$ be some environment of size $\delta \sim 1$ around τ. Following the above recipe, it remains to show that the local law holds for the random matrices

$$\mathbf{H}_t = \mathrm{e}^{-t/2}\mathbf{H}_0 + (1-\mathrm{e}^{-t})^{1/2}\mathbf{U},$$

uniformly in both $t \in [0, T]$ and the spectral parameters $z \in I + \mathrm{i}[N^{\gamma-1}, \infty)$. Here T is a small negative power of N, i.e., $T = N^{-\xi}$ for some $\xi > 0$, such that $\mathbf{H}$ and $\mathbf{H}_T$ are close in the four moment comparison sense (cf. Theorem 2.3 of [20]), and $\mathbf{U}$ is a standard GUE/GOE random matrix. The random matrix $\mathbf{H}_0$ has independent entries, is independent of $\mathbf{U}$, and has a variance matrix

$$\mathbf{S}_0 := \mathrm{e}^T\mathbf{S} - (\mathrm{e}^T - 1)\mathbf{S}_{\mathrm{G}},$$

with $\mathbf{S}$ and $\mathbf{S}_{\mathrm{G}}$ denoting the variance matrices of $\mathbf{H}$ and the standard GUE/GOE-matrix, respectively. It follows that the variance matrix of $\mathbf{H}_t$ is

$$\mathbf{S}_t = \mathrm{e}^{-t}\mathbf{S}_0 + (1-\mathrm{e}^{-t})\mathbf{S}_{\mathrm{G}},$$

and hence $\mathbf{S}_T = \mathbf{S}$ as required by the moment matching.

We will now show that $\mathbf{H}_t$ satisfy the hypotheses of Corollary 1.8 uniformly in t. Since $T = N^{-\xi}$ is small, the variance matrices $\mathbf{S}_t$ are all small perturbations of $\mathbf{S}$. In particular, $\mathbf{S}_t$, $t \in [0, T]$, are hence $q/2$-full.

Next we show that the interval I is inside the bulk of $\mathbf{H}_t$. To this end, we consider the QVE associated to the variance matrix $\mathbf{S}_t$,

$$-\frac{1}{m_{t;i}} = z + (\mathbf{S}\mathbf{m}_t)_i + d_i, \quad \mathbf{d} = (\mathbf{S}_t - \mathbf{S})\mathbf{m}_t,$$

as a perturbation of the original QVE with $\mathbf{S} = \mathbf{S}_T$. In order to use our stability results we show $\|\mathbf{d}\|_\infty \lesssim T$. Since $\mathbf{H}_t$ is $q/2$-full we have $s_{t;ij} \geq q/2$ and hence using (i) of Theorem 6.1 of [1] we see that there is a constant $\delta' \sim 1$ such that $\|\mathbf{m}_t(z)\|_\infty \sim 1$ uniformly for $|\mathrm{Re}\, z| \leq \delta'$. Moreover, the structural L^2-bound from Theorem 2.1 of [1] implies

$$\frac{\|\mathbf{m}_t(z)\|_{\ell^2}^2}{N} = \frac{1}{N}\sum_{i=1}^{N}|m_{t;i}(z)|^2 \le \frac{4}{|z|^2}, \quad z \in \mathbb{H},\ t \in [0,T].$$

Combining these estimates we see that $\sup_{t,z}\|\mathbf{m}_t(z)\|_{\ell^2}^2 \lesssim N$, and consequently the perturbation is small in the uniform norm: $\|\mathbf{d}\|_\infty \lesssim N \sup_{i,j}|s_{t;ij} - s_{ij}| \lesssim N^{-\xi}$. Applying the stability (Theorem 4.2 or Theorem 2.12 from [1]) of the QVE associated to $\mathbf{S}$ we conclude that $\|\mathbf{m}_t(z) - \mathbf{m}(z)\|_\infty \lesssim N^{-\xi}\varepsilon^{-2}$, and hence $\rho_t(\omega) \ge \varepsilon/2$ for $\omega \in I$ and all t, provided N is sufficiently large.

The moment condition (D) is automatically satisfied uniformly for every $\mathbf{H}_t$ by construction. Since the condition (A) is merely a matter of normalization we have now shown that $\mathbf{H}_t$ satisfy the hypotheses of Corollary 1.8 uniformly in t. Thus $\mathbf{H}_t$ satisfy local law uniformly in $t \in [0,T]$ and $z \in I + \mathrm{i}[N^{\gamma-1},\infty)$. This finishes the proof of universality.

Acknowledgements Open access funding provided by Institute of Science and Technology (IST Austria). We thank the anonymous referee for several useful comments and suggestions. We are grateful to Johannes Alt for pointing out several typos.

A Appendix

The relation $\prec$ is transitive and it satisfies the following arithmetic rules:

Lemma A.1 (Basic facts about stochastic domination) *We have:*

(i) *If $\phi \prec N^\delta \psi$, for every $\delta > 0$, then $\phi \prec \psi$;*
(ii) *If $\phi \prec N^{-\delta}\phi + \psi$, for some $\delta > 0$, then $\phi \prec \psi$.*

Let ϕ_u and ψ_u be some non-negative random variables parametrized by elements u of some set $\mathbb{U}$, such that $\phi_u \prec \psi_u$, uniformly in u. If $\mathbb{U}' \subset \mathbb{U}$, then

(iii) *$\sum_{u\in\mathbb{U}'}\phi_u \prec \sum_{u\in\mathbb{U}'}\psi_u$, provided $|\mathbb{U}'| \le N^C$ for some $C < \infty$;*
(iv) *$\prod_{u\in\mathbb{U}'}\phi_u \prec \prod_{u\in\mathbb{U}'}\psi_u$, provided $|\mathbb{U}'| \le C$, for some $C < \infty$.*

These properties follow directly from the definition (Definition 1.6) of stochastic domination. For further details see [14].

Lemma A.2 (Bound propagation) *Suppose C_1, D_1, D_2, D_3 and ε_1 are positive constants, depending explicitly on p, P, L, $\underline{\mu}$, γ and possible on additional parameters in some set V. Suppose further that the threshold function N_0 from Definition 1.9 depends on the same parameters. Let $\mathbb{D}^{(N)} \subseteq \mathbb{H}$ be a sequence of connected subsets of the complex upper half plane with only polynomially growing diameter, $\sup\{|z_1 - z_2| : z_1, z_2 \in \mathbb{D}^{(N)}\} \le N^{D_1}$. Let $\varphi = (\varphi^{(N)}(z) : z \in \mathbb{D}^{(N)})_{N\in\mathbb{N}}$ be a sequence of non-negative random functions and $\Phi^{(N)} : \mathbb{D}^{(N)} \to (N^{-D_3},\infty)$ a sequence of deterministic functions on these sets. Suppose they satisfy the following conditions:*

- *Uniformly for all* $z_1, z_2 \in \mathbb{D}^{(N)}$

$$|\varphi^{(N)}(z_1) - \varphi^{(N)}(z_2)| + |\Phi^{(N)}(z_1) - \Phi^{(N)}(z_2)| \le C_1 N^{D_2} |z_1 - z_2|^{\varepsilon_1}. \quad \text{(A.1)}$$

- *Uniformly for all* $z \in \mathbb{D}^{(N)}$

$$\textit{a.w.o.p.} \quad \varphi^{(N)} \notin \big[\Phi^{(N)}(z) - N^{-D_3}, \Phi^{(N)}(z)\big]. \quad \text{(A.2)}$$

- *There is a sequence* $z_0^{(N)} \in \mathbb{D}^{(N)}$ *such that*

$$\textit{a.w.o.p.} \quad \varphi^{(N)}(z_0^{(N)}) \le \Phi^{(N)}(z_0^{(N)}). \quad \text{(A.3)}$$

Then the sequence φ *satisfies the bound*

$$\textit{a.w.o.p.} \quad \textit{for all} \quad z \in \mathbb{D}^{(N)} \ : \ \varphi^{(N)}(z) \le \Phi^{(N)}(z). \quad \text{(A.4)}$$

Proof We will not carry the upper index N in this proof. First we choose a grid $\mathbb{G} \subseteq \mathbb{D}$ with the following properties

- The number of points in $\mathbb{G}$ is polynomially large, i.e., $|\mathbb{G}| \le C_2 N^{D_4}$.
- The grid is connected and sufficiently dense in $\mathbb{D}$, i.e., for any two points $z_1, z \in \mathbb{G}$ there is a path $(z_i)_{i=2}^K \subseteq \mathbb{G}$, such that $\max\{|z_K - z|, |z_{i+1} - z_i|\} \le N^{-D_5}$ for all $i = 1, \dots, K-1$.

Here, the positive exponent D_5 is chosen sufficiently large such that $C_1 N^{D_2 - \varepsilon_1 D_5} \le N^{-D_3}/2$. Then an upper bound on the positive constants D_4 and C_2 is determined by the choice of D_5 and the diameter of $\mathbb{D}$, i.e., by D_1.

Now let $z \in \mathbb{G}$. Then we find a path $(z_i)_{i=1}^K$ in $\mathbb{G}$ that connects z_0 with $z_{K+1} := z$ in the sense of the second property of $\mathbb{G}$. We may assume the length of the path, K, to be bounded by $|\mathbb{G}|$. Inductively we show that for all $i = 0, \dots, K+1$

$$\text{a.w.o.p.} \quad \varphi(z_i) \le \Phi(z_i) - N^{-D_3}.$$

For $i = 0$ this follows from (A.3) and (A.2). For all other i it follows by induction using the continuity condition (A.1), which implies $|\varphi(z_{i+1}) - \varphi(z_i)| + |\Phi(z_{i+1}) - \Phi(z_i)| \le N^{-D_3}/2$. This shows that if $\varphi(z_i) \le \Phi(z_i) - N^{-D_3}$, then $\varphi(z_{i+1}) \le \Phi(z_{i+1})$ and with (A.2) even that $\varphi(z_{i+1}) \le \Phi(z_{i+1}) - N^{-D_3}$. In particular, $\varphi(z) \le \Phi(z) - N^{-D_3}$ a.w.o.p.

Using a union bound we infer that

$$\text{a.w.o.p.} \quad \text{for all } z \in \mathbb{G} \quad \varphi(z) \le \Phi(z) - N^{-D_3}.$$

By (A.1) and since $\mathbb{G}$ is sufficiently dense in $\mathbb{D}$ this bound extends to all $z \in \mathbb{D}$ and the lemma is proven. □

Proof of Lemma 5.1 For f, χ compactly supported on $\mathbb{R}$ the Cauchy integral formula holds true,

$$
\begin{aligned}
f(\tau) &= \frac{1}{\pi}\int_{\mathbb{R}^2} \frac{\partial_{\bar z}\widetilde{f}(\sigma+\mathrm{i}\eta)}{\tau-\sigma-\mathrm{i}\eta}\mathrm{d}\sigma\mathrm{d}\eta\\
&= \frac{1}{2\pi}\int_{\mathbb{R}^2} \frac{\mathrm{i}\eta f''(\sigma)\chi(\eta)+\mathrm{i}(f(\sigma)+\mathrm{i}\eta f'(\sigma))\chi'(\eta)}{\tau-\sigma-\mathrm{i}\eta}\mathrm{d}\sigma\mathrm{d}\eta,\\
\widetilde{f}(\sigma+\mathrm{i}\eta) &:= (f(\sigma)+\mathrm{i}\eta f'(\sigma))\chi(\eta).
\end{aligned}
$$

For a signed measure ν on $\mathbb{R}$ this implies the formula

$$
\int_{\mathbb{R}} f(\tau)\nu(\mathrm{d}\tau) = \mathrm{Re}\int_{\mathbb{R}} f(\tau)\nu(\mathrm{d}\tau) = -\frac{1}{2\pi}\big(L_1(\nu)+L_2(\nu)+L_3(\nu)\big),
$$

where the three integrals L_1, L_2 and L_3 are given as

$$
\begin{aligned}
L_1(\nu) &:= \int_{\mathbb{R}^2} \eta f''(\sigma)\chi(\eta)\mathrm{Im}m_\nu(\sigma+\mathrm{i}\eta)\mathrm{d}\sigma\mathrm{d}\eta,\\
L_2(\nu) &:= \int_{\mathbb{R}^2} f(\sigma)\chi'(\eta)\mathrm{Im}m_\nu(\sigma+\mathrm{i}\eta)\mathrm{d}\sigma\mathrm{d}\eta,\\
L_3(\nu) &:= \int_{\mathbb{R}^2} \eta f'(\sigma)\chi'(\eta)\mathrm{Re}m_\nu(\sigma+\mathrm{i}\eta)\mathrm{d}\sigma\mathrm{d}\eta,
\end{aligned}
$$

and m_ν is the Stieltjes transform of ν.

Now we choose $f \geq 0$, such that $f|_{[\tau_1,\tau_2]} = 1$ and $f|_{\mathbb{R}\setminus[\tau_1-\eta_1,\tau_2+\eta_2]} = 0$. Furthermore, we assume that the derivatives of f satisfy

$$
\begin{aligned}
&\|f'|_{[\tau_1-\eta_1,\tau_1]}\|_\infty \lesssim \eta_1^{-1}, \quad \|f''|_{[\tau_1-\eta_1,\tau_1]}\|_\infty \lesssim \eta_1^{-2},\\
&\|f'|_{[\tau_2,\tau_2+\eta_2]}\|_\infty \lesssim \eta_2^{-1}, \quad \|f''|_{[\tau_2,\tau_2+\eta_2]}\|_\infty \lesssim \eta_2^{-2}.
\end{aligned}
$$

The function $\chi \geq 0$ is chosen to be symmetric and such that $\chi|_{[-\varepsilon,\varepsilon]} = 1$, $\chi|_{\mathbb{R}\setminus[-2\varepsilon,2\varepsilon]} = 0$, as well as $\|\chi'\|_\infty \lesssim \varepsilon^{-1}$. Here the constant ε is chosen to satisfy $\varepsilon \geq \max\{\eta_1,\eta_2\}$. We now derive bounds on $L_k(\nu_1-\nu_2)$ for $k=1,2,3$.

We split the integral, L_1, into the contributions,

$$
\begin{aligned}
L_1(\nu) &= 2\big(L_{1,1,<}(\nu)+L_{1,1,>}(\nu)+L_{1,2,<}(\nu)+L_{1,2,>}(\nu)\big),\\
L_{1,1,<}(\nu) &:= \int_{\tau_1-\eta_1}^{\tau_1}\mathrm{d}\sigma\int_0^{\eta_1}\mathrm{d}\eta\,\eta f''(\sigma)\mathrm{Im}m_\nu(\sigma+\mathrm{i}\eta),\\
L_{1,1,>}(\nu) &:= \int_{\tau_1-\eta_1}^{\tau_1}\mathrm{d}\sigma\int_{\eta_1}^{2\varepsilon}\mathrm{d}\eta\,\eta f''(\sigma)\chi(\eta)\mathrm{Im}m_\nu(\sigma+\mathrm{i}\eta),\\
L_{1,2,<}(\nu) &:= \int_{\tau_2}^{\tau_2+\eta_2}\mathrm{d}\sigma\int_0^{\eta_2}\mathrm{d}\eta\,\eta f''(\sigma)\mathrm{Im}m_\nu(\sigma+\mathrm{i}\eta),\\
L_{1,2,>}(\nu) &:= \int_{\tau_2}^{\tau_2+\eta_2}\mathrm{d}\sigma\int_{\eta_2}^{2\varepsilon}\mathrm{d}\eta\,\eta f''(\sigma)\chi(\eta)\mathrm{Im}m_\nu(\sigma+\mathrm{i}\eta).
\end{aligned}
$$

For a positive measure ν the function $\eta \mapsto \eta\,\mathrm{Im} m_\nu(\sigma + \mathrm{i}\eta)$ is monotonously increasing. Thus, we estimate

$$|L_{1,1,<}(\nu)| \le \max_{\sigma\in[0,\eta_1]} |f''(\tau_1-\sigma)| \int_{\tau_1-\eta_1}^{\tau_1} \mathrm{d}\sigma \int_0^{\eta_1} \mathrm{d}\eta\, \eta_1 \mathrm{Im} m_\nu(\sigma+\mathrm{i}\eta_1)$$
$$\le \int_{\tau_1-\eta_1}^{\tau_1} \mathrm{d}\sigma\, \mathrm{Im} m_\nu(\sigma+\mathrm{i}\eta_1),\ \nu\ge 0.$$

We conclude that

$$|L_{1,1,<}(\nu_1-\nu_2)| \le \int_{\tau_1-\eta_1}^{\tau_1} \mathrm{d}\sigma \left(2\mathrm{Im} m_{\nu_1}(\sigma+\mathrm{i}\eta_1) + \left|m_{\nu_1-\nu_2}(\sigma+\mathrm{i}\eta_1)\right|\right). \tag{A.5}$$

In the same way we find

$$|L_{1,2,<}(\nu_1-\nu_2)| \le \int_{\tau_2}^{\tau_2+\eta_2} \mathrm{d}\sigma \left(2\mathrm{Im} m_{\nu_1}(\sigma+\mathrm{i}\eta_2) + \left|m_{\nu_1-\nu_2}(\sigma+\mathrm{i}\eta_2)\right|\right). \tag{A.6}$$

For the treatment of $L_{1,1,>}$ we integrate by parts, first in σ and then in η,

$$L_{1,1,>}(\nu) = -\eta_1 \int_{\tau_1-\eta_1}^{\tau_1} \mathrm{d}\sigma\, f'(\sigma) \mathrm{Re} m_\nu(\sigma+\mathrm{i}\eta_1)$$
$$- \int_{\eta_1}^{2\varepsilon} \mathrm{d}\eta \int_{\tau_1-\eta_1}^{\tau_1} \mathrm{d}\sigma\, \partial_\eta(\eta\chi(\eta)) f'(\sigma) \mathrm{Re} m_\nu(\sigma+\mathrm{i}\eta).$$

We use $\max_\eta |\chi(\eta)+\eta\chi'(\eta)| \lesssim 1$ and $\max_{\sigma\in[0,\eta_1]} |f'(\tau_1-\sigma)| \lesssim \eta_1^{-1}$. In this way we estimate for $\nu = \nu_1 - \nu_2$,

$$\begin{aligned} L_{1,1,>}(\nu_1-\nu_2) &\lesssim \int_{\tau_1-\eta_1}^{\tau_1} \mathrm{d}\sigma |m_{\nu_1-\nu_2}(\sigma+\mathrm{i}\eta_1)| \\ &\quad + \frac{1}{\eta_1} \int_{\eta_1}^{2\varepsilon} \mathrm{d}\eta \int_{\tau_1-\eta_1}^{\tau_1} \mathrm{d}\sigma |m_{\nu_1-\nu_2}(\sigma+\mathrm{i}\eta)|. \end{aligned} \tag{A.7}$$

Going through the same steps we also arrive at

$$\begin{aligned} L_{1,2,>}(\nu_1-\nu_2) &\lesssim \int_{\tau_2}^{\tau_2+\eta_2} \mathrm{d}\sigma |m_{\nu_1-\nu_2}(\sigma+\mathrm{i}\eta_2)| \\ &\quad + \frac{1}{\eta_2} \int_{\eta_2}^{2\varepsilon} \mathrm{d}\eta \int_{\tau_2}^{\tau_2+\eta_2} \mathrm{d}\sigma |m_{\nu_1-\nu_2}(\sigma+\mathrm{i}\eta)|. \end{aligned} \tag{A.8}$$

We continue by estimating L_2 from above.

$$|L_2(\nu_1-\nu_2)| \lesssim \frac{1}{\varepsilon} \int_{\tau_1-\eta_1}^{\tau_2+\eta_2} \mathrm{d}\sigma \int_\varepsilon^{2\varepsilon} \mathrm{d}\eta |m_{\nu_1-\nu_2}(\sigma+\mathrm{i}\eta)|. \tag{A.9}$$

Finally we derive a bound for L_3. We split the integral into two components,

$$\begin{aligned} L_3(\nu) &= 2\big(L_{3,1}(\nu)+L_{3,2}(\nu)\big),\\ L_{3,1}(\nu) &:= \int_{\tau_1-\eta_1}^{\tau_1}\mathrm{d}\sigma\int_\varepsilon^{2\varepsilon}\mathrm{d}\eta\,\eta f'(\sigma)\chi'(\eta)\mathrm{Re}\, m_\nu(\sigma+\mathrm{i}\eta),\\ L_{3,2}(\nu) &:= \int_{\tau_2}^{\tau_2+\eta_2}\mathrm{d}\sigma\int_\varepsilon^{2\varepsilon}\mathrm{d}\eta\,\eta f'(\sigma)\chi'(\eta)\mathrm{Re}\, m_\nu(\sigma+\mathrm{i}\eta). \end{aligned}$$

We arrive at the bound

$$\begin{aligned} L_3(\nu_1-\nu_2) &\lesssim \frac{1}{\eta_1}\int_{\tau_1-\eta_1}^{\tau_1}\mathrm{d}\sigma\int_\varepsilon^{2\varepsilon}\mathrm{d}\eta|m_{\nu_1-\nu_2}(\sigma+\mathrm{i}\eta)|\\ &\quad+\frac{1}{\eta_2}\int_{\tau_2}^{\tau_2+\eta_2}\mathrm{d}\sigma\int_\varepsilon^{2\varepsilon}\mathrm{d}\eta|m_{\nu_1-\nu_2}(\sigma+\mathrm{i}\eta)|. \end{aligned}$$

We combine this with the estimates from (A.5), (A.6), (A.7), (A.8) and (A.9). Altogether we have

$$\Big|\int f\,\mathrm{d}(\nu_1-\nu_2)\Big| \lesssim J_1+J_2+J_3,$$

where the three terms on the right hand side are given by

$$\begin{aligned} J_1 &:= \int_{\tau_1-\eta_1}^{\tau_1}\mathrm{d}\sigma\,\Big(\mathrm{Im}\, m_{\nu_1}(\sigma+\mathrm{i}\eta_1)+|m_{\nu_1-\nu_2}(\sigma+\mathrm{i}\eta_1)|\\ &\quad+\frac{1}{\eta_1}\int_{\eta_1}^{2\varepsilon}\mathrm{d}\eta|m_{\nu_1-\nu_2}(\sigma+\mathrm{i}\eta)|\Big),\\ J_2 &:= \int_{\tau_2}^{\tau_2+\eta_2}\mathrm{d}\sigma\,\Big(\mathrm{Im}\, m_{\nu_1}(\sigma+\mathrm{i}\eta_2)+|m_{\nu_1-\nu_2}(\sigma+\mathrm{i}\eta_2)|\\ &\quad+\frac{1}{\eta_2}\int_{\eta_2}^{2\varepsilon}\mathrm{d}\eta|m_{\nu_1-\nu_2}(\sigma+\mathrm{i}\eta)|\Big),\\ J_3 &:= \frac{1}{\varepsilon}\int_{\tau_1-\eta_1}^{\tau_2+\eta_2}\mathrm{d}\sigma\int_\varepsilon^{2\varepsilon}\mathrm{d}\eta|m_{\nu_1-\nu_2}(\sigma+\mathrm{i}\eta)|. \end{aligned}$$

Now we use this bound for the smoothed out indicator function to derive a bound on the difference of number of eigenvalues in the interval $[\tau_1,\tau_2]$ and the predicted number, given by the integral over the density of states. We use

$$\nu_2([\tau_1,\tau_2]) \le \int f\mathrm{d}\nu_1+\int f\mathrm{d}(\nu_1-\nu_2), \tag{A.10}$$

for f defined as above. Then we get

$$\nu_2([\tau_1,\tau_2]) \le \nu_1([\tau_1,\tau_2]) + \nu_1([\tau_1-\eta_1,\tau_1]\cup[\tau_2,\tau_2+\eta_2]) + \left|\int f\,\mathrm{d}(\nu_1-\nu_2)\right|.$$

Similarly we use

$$\nu_1([\tau_1,\tau_2]) \ge \int f\,\mathrm{d}\nu_2 - \nu_1([\tau_1-\eta_1,\tau_1]\cup[\tau_2,\tau_2+\eta_2]),$$

to get the bound

$$\nu_1([\tau_1,\tau_2]) \ge \nu_2([\tau_1,\tau_2]) - \left|\int f\,\mathrm{d}(\nu_2-\nu_1)\right| - \nu_1([\tau_1-\eta_1,\tau_1]\cup[\tau_2,\tau_2+\eta_2]).$$

Together, the two bounds imply

$$|\nu_1([\tau_1,\tau_2]) - \nu_2([\tau_1,\tau_2])| \lesssim \nu_1([\tau_1-\eta_1,\tau_1]\cup[\tau_2,\tau_2+\eta_2]) + J_1 + J_2 + J_3.$$

□

References

1. Ajanki, O., Erdős, L., Krüger, T.: Quadratic Vector Equations on the Complex Upper Half Plane. arXiv:1506.05095
2. Ajanki, O., Erdős, L., Krüger, T.: Singularities of solutions to quadratic vector equations on the complex upper half-plane. Commun. Pure Appl. Math. doi:10.1002/cpa.21639
3. Ajanki, O., Erdős, L., Krüger, T.: Local spectral statistics of Gaussian matrices with correlated entries. J. Stat. Phys. **163**(2), 280–302 (2016). doi:10.1007/s10955-016-1479-y
4. Alt, J., Erdős, L., Krüger, T.: Local laws for Gram matrices. arXiv:1606.07353
5. Anderson, G., Zeitouni, O.: A CLT for a band matrix model. Probab. Theory Relat. Fields **134**(2), 283–338 (2005)
6. Anderson, G., Zeitouni, O.: A law of large numbers for finite-range dependent random matrices. Comm. Pure Appl. Math. **61**(8), 1118–1154 (2008)
7. Bai, Z .D., Silverstein, Jack W.: No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. Ann. Probab. **26**(1), 316–345, 01 (1998)
8. Bekerman, F., Figalli, A., Guionnet, A.: Transport maps for Beta-matrix models and universality. Commun. Math. Phys. **338**, 589–619 (2015)
9. Bloemendal, A., Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: Isotropic local laws for sample covariance and generalized wigner matrices. Electr. J. Probab. **19**(33), 1–53 (2014)
10. Bourgade, P., Erdős, L., Yau, H.-T.: Universality of general β-ensembles. Duke Math. J. **163**(6), 1127–1190 (2014)
11. Davies, E .B.: The functional calculus. J. Lond. Math. Soc. (2) **52**(1), 166–176 (1995)
12. Erdős, L., Knowles, A., Yau, H.-T.: Averaging fluctuations in resolvents of random band matrices. Ann. Henri Poincaré **14**(8), 1837–1926 (2013)
13. Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: Spectral statistics of Erdős–Rényi s II: eigenvalue spacing and the extreme eigenvalues. Commun. Math. Phys. **314**(3), 587–640 (2012)
14. Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: The local semicircle law for a general class of random matrices. Electron. J. Probab. **18**, 1–58 (2013)
15. Erdős, L., Ramírez, J.A., Schlein, B., Yau, H.-T.: Universality of sine-kernel for Wigner matrices with a small Gaussian perturbation. Electron. J. Probab. **15**(18), 526–603 (2010)
16. Erdős, L., Schlein, B., Yau, H.-T.: Universality of random matrices and local relaxation flow. Invent. Math. **185**, 75–119 (2011)

17. Erdős, L., Schlein, B., Yin, J.: The local relaxation flow approach to universality of the local statistics for random matrices. Ann. Inst. Henri Poincaré Probab. Stat. **48**(1), 1–46 (2012)
18. Erdős, L., Schnelli, K.: Universality for Random Matrix Flows with Time-dependent Density. arXiv:1504.00650
19. Erdős, L., Yau, H.-T.: Universality of local spectral statistics of random matrices. Bull. Am. Math. Soc **49**, 377–414 (2012)
20. Erdős, L., Yau, H.-T., Yin, J.: Bulk universality for generalized Wigner matrices. Probab. Theory Relat. Fields **154**(1–2), 341–407 (2011)
21. Erdős, L., Yau, H.-T., Yin, J.: Universality for generalized Wigner matrices with Bernoulli distribution. J. Comb. **2**(1), 15–82 (2011)
22. Erdős, L., Yau, H.-T., Yin, J.: Rigidity of eigenvalues of generalized Wigner matrices. Adv. Math. **229**(3), 1435–1515 (2012)
23. Girko, V.L.: Theory of Stochastic Canonical Equations. Vol. I, volume 535 of Mathematics and its Applications. Kluwer Academic Publishers, Dordrecht (2001)
24. Guionnet, A.: Large deviations upper bounds and central limit theorems for non-commutative functionals of Gaussian large random matrices. Annales de l'IHP Probabilités et statistiques **38**, 341–384 (2002)
25. Khorunzhy, A.M., Pastur, L.A.: On the eigenvalue distribution of the deformed Wigner ensemble of random matrices. In: Spectral Operator Theory and Related Topics, Adv. Soviet Math., 19, pp. 97–127. Am. Math. Soc., Providence, RI (1994)
26. Landon, B., Yau, H.-T.: Convergence of local statistics of Dyson Brownian motion. arXiv:1504.03605
27. Lee, J.O., Schnelli, K., Stetler, B., Yau, H.-T.: Bulk Universality for Deformed Wigner Matrices. arXiv:1405.6634
28. Shcherbina, M.: On Fluctuations of Eigenvalues of Random Band Matrices. arXiv:1504.05762
29. Shlyakhtenko, D.: Random Gaussian band matrices and freeness with amalgamation. Int. Math. Res. Note **20**, 1013–1015 (1996)
30. Tao, T., Vu, V.: Random matrices: the universality phenomenon for Wigner ensembles. In: Modern Aspects of Random Matrix Theory, pp. 121–172. Am. Math. Soc., Providence, RI (2014)
31. Wigner, E.P.: Characteristic vectors of bordered matrices with infinite dimensions. Ann. Math. **2**(62), 548–564 (1955)