

Delocalization for a class of random block band matrices

Zhigang Bao¹ · László Erdős¹

Received: 7 April 2015 / Revised: 11 November 2015

© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract We consider $N \times N$ Hermitian random matrices H consisting of blocks of size $M \geq N^{6/7}$. The matrix elements are i.i.d. within the blocks, close to a Gaussian in the four moment matching sense, but their distribution varies from block to block to form a block-band structure, with an essential band width M . We show that the entries of the Green's function $G(z) = (H - z)^{-1}$ satisfy the local semicircle law with spectral parameter $z = E + i\eta$ down to the real axis for any $\eta \gg N^{-1}$, using a combination of the supersymmetry method inspired by Shcherbina (J Stat Phys 155(3): 466–499, 2014) and the Green's function comparison strategy. Previous estimates were valid only for $\eta \gg M^{-1}$. The new estimate also implies that the eigenvectors in the middle of the spectrum are fully delocalized.

Keywords Random band matrix · Supersymmetry · Green's function comparison · Local semicircle law · Delocalization

Mathematics Subject Classification 15B52 · 81Q60 · 82B44

Z. Bao was supported by ERC Advanced Grant RANMAT No. 338804; L. Erdős was partially supported by ERC Advanced Grant RANMAT No. 338804.

✉ László Erdős
lerdos@ist.ac.at

Zhigang Bao
Zhigang.Bao@ist.ac.at

¹ IST Austria, Am Campus 1, 3400 Klosterneuberg, Austria

1 Introduction

Grassmann integration with supersymmetric (SUSY) methods is ubiquitous in the physics literature of random quantum systems, see e.g. the basic monograph of Efetov [7]. This approach is especially effective for analyzing the Green function in the middle of the bulk spectrum with spectral parameter close to the real axis, i.e. precisely in the regime where other methods often fail. The main *algebraic* strength lies in the fact that Gaussian integrals with Grassmann variables counterbalance the determinants obtained in the partition functions of complex Gaussian integrals. This greatly simplifies algebraic manipulations as it was demonstrated in several papers, see e.g. the proof of the absolutely continuous spectrum on the Bethe lattice by Klein [17] or the bounds on the Lyapunov exponents for random walks in random environment at the critical energy by Wang [29]. However, in theoretical physics Grassmann integrations are also commonly used as an *analytic* tool by performing saddle point analysis on the superspace coordinatized by complex and Grassmann variables. Since Grassmann variables lack the concept of size, the rigorous justification of this very appealing idea is notoriously difficult.

Initiated by Spencer (see [26] for a summary) and starting with the paper [4] by Disertori, Pinson and Spencer, only a handful of mathematical papers have succeeded in exploiting this powerful tool in an essentially analytic way. We still lack the mathematical framework of a full-fledged analysis on the superspace that would enable us to translate physics arguments into proofs directly, but a combination of refined algebraic identities from physics (such as the superbosonization formula) and a careful analysis have yielded results that are currently inaccessible with more standard probabilistic methods. In this paper we present such results on random band matrices that surpass a well known limitation of the recently developed probabilistic techniques to prove the local versions of the celebrated Wigner semicircle law. We start with introducing the physical motivation of our model.

The Hamiltonian of quantum systems on a graph with vertex set Γ is a self-adjoint matrix $H = (h_{ab})_{a,b \in \Gamma}$, $H = H^*$. The matrix elements h_{ab} represent the quantum transition rates from vertex a to b . Disordered quantum systems have random matrix elements. We assume they are centered, $\mathbb{E}h_{ab} = 0$, and independent subject to the basic symmetry constraint $h_{ab} = \bar{h}_{ba}$. The variance $\sigma_{ab}^2 := \mathbb{E}|h_{ab}|^2$ represents the strength of the transition from a to b and we use a scaling where the norm $\|H\|$ is typically order 1. The simplest case is the mean field model, where h_{ab} are identically distributed; this is the standard Wigner matrix ensemble [31]. The other prominent example is the Anderson model [2] or random Schrödinger operator, $H = \Delta + V$, where the kinetic energy Δ is the (deterministic) graph Laplacian and the potential $V = (V_x)_{x \in \Gamma}$ is an on-site multiplication operator with random multipliers. If Γ is a discrete d -dimensional torus then only few matrix elements h_{ab} are nonzero and they connect nearest neighbor points in the torus, $\text{dist}(a, b) \leq 1$. This is in sharp contrast to the mean field character of the Wigner matrices.

Random band matrices naturally interpolate between the mean field Wigner matrices and the short range Anderson model. They are characterized by a parameter M , called the *band width*, such that the matrix elements h_{ab} for $\text{dist}(a, b) \geq M$ are zero

or negligible. If M is comparable with the diameter L of the system then we are in the mean field regime, while $M \sim 1$ corresponds to the short range model.

The Anderson model exhibits a metal-insulator phase transition: at high disorder the system is in the localized (insulator) regime, while at small disorder it is in the delocalized (metallic) regime, at least in $d \geq 3$ dimensions and away from the spectral edges. The localized regime is characterized by exponentially decaying eigenfunctions and off diagonal decay of the Green's function, while in the complementary regime the eigenfunctions are supported in the whole physical space. In terms of the *localization length* ℓ , the characteristic length scale of the decay, the localized regime corresponds to $\ell \ll L$, while in the delocalized regime $\ell \sim L$. Starting from the basic papers [1, 15], the localized regime is well understood, but the delocalized regime is still an open mathematical problem for the d -dimensional torus.

Let $N = L^d$ be the number of vertices in the discrete torus. Since the eigenvectors of the mean field Wigner matrices are always delocalized [13, 14], while the short range models are localized, by varying the parameter M in the random band matrix, one expects a (de)localization phase transition. Indeed, for $d = 1$ it is conjectured (and supported by non rigorous supersymmetric calculations [16]) that the system is delocalized for broad bands, $M \gg N^{1/2}$ and localized for $M \ll N^{1/2}$. The optimal power $1/2$ has not yet been achieved from either sides. Localization has been shown for $M \ll N^{1/8}$ in [23], while delocalization in a certain weak sense for the most eigenvectors was proven for $M \gg N^{4/5}$ in [11]. Interestingly, for a special Gaussian model even the sine kernel behavior of the 2-point correlation function of the characteristic polynomials could be proven down to the optimal band width $M \gg N^{1/2}$, see [19, 21]. Note that the sine kernel is consistent with the delocalization but does not imply it. We remark that our discussion concerns the bulk of the spectrum; the transition at the spectral edge is much better understood. In [25] it was shown with moment method that the edge spectrum follows the Tracy–Widom distribution, characteristic to mean field model, for $M \gg N^{5/6}$, but it yields a different distribution for narrow bands, $M \ll N^{5/6}$.

Delocalization is closely related to estimates on the diagonal elements of the resolvent $G(z) = (H - z)^{-1}$ at spectral parameters with small imaginary part $\eta = \text{Im}z$. Indeed, if $G_{ii}(E + i\eta)$ is bounded for all i and all $E \in \mathbb{R}$, then each ℓ^2 -normalized eigenvector \mathbf{u} of H is delocalized on scale η^{-1} in a sense that $\max_i |u_i|^2 \lesssim \eta$, i.e. u is supported on at least η^{-1} sites. In particular, if G_{ii} can be controlled down to the scale $\eta \sim 1/N$, then the system is in the complete delocalized regime.

For band matrices with band width M , or even under the more general condition $\sigma_{ab}^2 \leq M^{-1}$, the boundedness of G_{ii} was shown down to scale $\eta \gg M^{-1}$ in [14] (see also [12]). If $M \gg N^{1/2}$, it is expected that G_{ii} remains bounded even down to $\eta \gg N^{-1}$ which is the typical eigenvalue spacing, the smallest relevant scale in the model. However, the standard approach [12, 14] via the self-consistent equations for the Green's function does not seem to work for $\eta \leq 1/M$; the fluctuation is hard to control. The more subtle approach using the self-consistent *matrix* equation in [11] could prove delocalization and the off-diagonal Green's function profile that are consistent with the conventional quantum diffusion picture, but it was valid only for relatively large

η , far from M^{-1} . Moment methods, even with a delicate renormalization scheme [24] could not break the barrier $\eta \sim M^{-1}$ either.

In this paper we attack the problem differently; with supersymmetric (SUSY) techniques. Our main result is that $G_{ii}(z)$ is bounded, and the local semicircle law holds for any $\eta \gg N^{-1}$, i.e. down to the optimal scale, if the band width is not too small, $M \gg N^{6/7}$, but under two technical assumptions. First, we consider a generalization of Wegner's n -orbital model [22, 30], namely, we assume that the band matrix has a block structure, i.e. it consists of $M \times M$ blocks and the matrix elements within each block have the same distribution. This assumption is essential to reduce the number of integration variables in the supersymmetric representation, since, roughly speaking, each $M \times M$ block will be represented by a single supermatrix with 16 complex or Grassmann variables. Second, we assume that the distribution of the matrix elements matches a Gaussian up to four moments in the spirit of [28]. Supersymmetry heavily uses Gaussian integrations, in fact all mathematically rigorous works on random band matrices with supersymmetric method assume that the matrix elements are Gaussian, see [4–6, 19–21, 26, 27]. The Green's function comparison method [14] allows one to compare Green's functions of two matrix ensembles provided that the distributions match up to four moments and provided that G_{ii} are bounded. This was an important motivation to reach the optimal scale $\eta \gg N^{-1}$.

In the next subsections we introduce the model precisely and state our main results. Our supersymmetric analysis was inspired by [20], but our observable, G_{ab} , requires a partly different formalism, in particular we use the singular version of the superbosonization formula [3]. Moreover, our analysis is considerably more involved since we consider relatively narrow bands. In Sect. 1.3, we explain our novelties compared with [20].

1.1 Matrix model

Let $H_N = (h_{ab})$ be an $N \times N$ random Hermitian matrix, in which the entries are independent (up to symmetry), centered, complex variables. In this paper, we are concerned with H_N possessing a block band structure. To define this structure explicitly, we set the additional parameters $M \equiv M(N)$ and $W \equiv W(N)$ satisfying

$$W = N/M.$$

For simplicity, we assume that both M and W are integers. Let $S = (s_{jk})$ be a $W \times W$ symmetric matrix, which will be chosen as a weighted Laplacian of a connected graph on W vertices. Now, we decompose H_N into $W \times W$ blocks of size $M \times M$ and relabel

$$h_{jk,\alpha\beta} := h_{ab}, \quad j, k = 1, \dots, W, \quad \alpha, \beta = 1, \dots, M,$$

where $j \equiv j(a)$ and $k \equiv k(b)$ are the spatial indices that describe the location of the block containing h_{ab} and $\alpha \equiv \alpha(a)$ and $\beta \equiv \beta(b)$ are the orbital indices that describe

the location of the entry in the block. More specifically, we have

$$j = \lceil a/M \rceil, \quad k = \lceil b/M \rceil, \quad \alpha = (a - (j - 1)M), \quad \beta = b - (k - 1)M.$$

We will call $(j(a), \alpha(a))$ (resp. $(k(b), \beta(b))$) as the spatial-orbital parametrization of a (resp. b). Moreover, we assume

$$\mathbb{E}h_{jk,\alpha\beta}h_{j'k',\alpha'\beta'} = \frac{1}{M}\delta_{jk'}\delta_{j'k}\delta_{\alpha\beta'}\delta_{\beta\alpha'}(\delta_{jk} + \mathfrak{s}_{jk}). \tag{1.1}$$

That means, the variance profile of the random matrix $\sqrt{M}H_N$ is given by

$$\tilde{S} = (\tilde{\mathfrak{s}}_{jk}) := I + S, \tag{1.2}$$

in which each entry represents the common variance of the entries in the corresponding block of $\sqrt{M}H_N$.

1.2 Assumptions and main results

In the sequel, for some matrix $A = (a_{ij})$ and some index sets I and J , we introduce the notation $A^{(I|J)}$ to denote the submatrix obtained by deleting the i th row and j th column of A for all $i \in I$ and $j \in J$. We will adopt the abbreviation

$$A^{(i|j)} := A^{(\{i\}|\{j\})}, \quad i \neq j, \quad A^{(i)} := A^{(\{i\}|\{i\})}. \tag{1.3}$$

In addition, we use $\|A\|_{\max} := \max_{i,j} |a_{ij}|$ to denote the max norm of A . Throughout the paper, we need some assumptions on S .

Assumption 1.1 (On S) Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a connected simple graph with $\mathcal{V} = \{1, \dots, W\}$. Assume that S is a $W \times W$ symmetric matrix satisfying the following four conditions.

- (i) S is a weighted Laplacian on \mathcal{G} , i.e. for $i \neq j$, we have $\mathfrak{s}_{ij} > 0$ if $\{i, j\} \in \mathcal{E}$ and $\mathfrak{s}_{ij} = 0$ if $\{i, j\} \notin \mathcal{E}$, and for the diagonal entries, we have

$$\mathfrak{s}_{ii} = - \sum_{j:j \neq i} \mathfrak{s}_{ij}, \quad \forall i = 1, \dots, W.$$

- (ii) \tilde{S} defined in (1.2) is strictly diagonally dominant, i.e., there exists some constant $c_0 > 0$ such that

$$1 + 2\mathfrak{s}_{ii} > c_0, \quad \forall i = 1, \dots, W.$$

- (iii) For the discrete Green's functions, we assume that there exist some positive constants C and γ such that

$$\max_{i=1,\dots,W} \|(S^{(i)})^{-1}\|_{\max} \leq CW^\gamma. \tag{1.4}$$

(iv) There exists a spanning tree $\mathcal{G}_0 = (\mathcal{V}, \mathcal{E}_0) \subset \mathcal{G}$, on which the weights are bounded below, i.e. for some constant $c > 0$, we have

$$s_{ij} \geq c, \quad \forall \{i, j\} \in \mathcal{E}_0.$$

Remark 1.2 From Assumption 1.1 (ii), we easily see that

$$\tilde{S} \geq c_0 I. \tag{1.5}$$

Later, in Lemma 7.4, we will see that $\|(S^{(i)})^{-1}\|_{\max} \leq CW^2$ always holds. Hence, we may assume $\gamma \leq 2$.

Example 1.1 Let Δ be the standard discrete Laplacian on the d -dimensional torus $[1, w]^d \cap \mathbb{Z}^d$, with periodic boundary condition, where $w = W^{1/d}$. Here by *standard* we mean the weights on the edges of the box are all 1. Now let $S = a\Delta$ for some positive constant $a < 1/4d$. It is then easy to check Assumption 1.1 (i), (ii) and (iv) are satisfied. In addition, if $d = 1$, it is well known that we can choose $\gamma = 1$ in Assumption 1.1 (iii). For $d \geq 3$, one can choose $\gamma = 0$. For $d = 2$, one can choose $\gamma = \varepsilon$ for arbitrarily small constant ε . For instance, one can refer to [8] for more details.

For simplicity, we also introduce the notation

$$\sigma_{ab}^2 := \mathbb{E}|h_{ab}|^2, \quad \mathcal{T} := (\sigma_{ab}^2) = \frac{1}{M} \tilde{S} \otimes (\mathbf{1}_M \mathbf{1}'_M), \quad a, b = 1, \dots, N, \tag{1.6}$$

where $\mathbf{1}_M$ is the M -dimensional vector whose components are all 1 and \tilde{S} is the variance matrix in (1.2). It is elementary that

$$\text{Spec}(\mathcal{T}) = \text{Spec}(\tilde{S}) \cup \{0\} \subset [0, 1]. \tag{1.7}$$

Our assumption on M depends on the constant γ in Assumption 1.1 (iii).

Assumption 1.3 (*On M*) We assume that there exists a (small) positive constant ε_1 such that

$$M \geq W^{4+2\gamma+\varepsilon_1}. \tag{1.8}$$

Remark 1.4 A direct consequence of (1.8) and $N = MW$ is

$$M \geq N^{\frac{4+2\gamma+\varepsilon_1}{5+2\gamma+\varepsilon_1}}. \tag{1.9}$$

Especially, when $\gamma = 1$, one has $M \gg N^{6/7}$. Actually, through a more involved analysis, (1.8) [or (1.9)] can be further improved. At least, for $\gamma \leq 1$, we expect that $M \gg N^{4/5}$ is enough. However, we will not pursue this direction here.

Besides Assumption 1.1 on the variance profile of H , we need to impose some additional assumption on the distribution of its entries. To this end, we temporarily employ the notation $H^g = (h_{ab}^g)$ to represent a random block band matrix with Gaussian entries, satisfying (1.1), Assumptions 1.1 and 1.3.

Assumption 1.5 (On distribution) We assume that for each $a, b \in \{1, \dots, N\}$, the moments of the entry h_{ab} match those of h_{ab}^g up to the 4th order, i.e.

$$\mathbb{E}(\operatorname{Re}h_{ab})^k (\operatorname{Im}h_{ab})^\ell = \mathbb{E}(\operatorname{Re}h_{ab}^g)^k (\operatorname{Im}h_{ab}^g)^\ell, \quad \forall k, \ell \in \mathbb{N}, \quad \text{s.t. } k + \ell \leq 4. \tag{1.10}$$

In addition, we assume the distribution of h_{ab} possesses a subexponential tail, namely, there exist positive constants c_1 and c_2 such that for any $\tilde{\gamma} > 0$,

$$\mathbb{P}\left(|h_{ab}| \geq \tilde{\gamma}^{c_1} (\mathbb{E}|h_{ab}|^2)^{\frac{1}{2}}\right) \leq c_2 e^{-\tilde{\gamma}} \tag{1.11}$$

holds uniformly for all $a, b = 1, \dots, N$.

The four moment condition (1.10) in the context of random matrices first appeared in [28].

To state our results, we will need the following notion on the comparison of two random sequences, which was introduced in [9, 12].

Definition 1.6 (Stochastic domination) For some possibly N -dependent parameter set \mathbf{U}_N , and two families of random variables $\mathbf{X} = (\mathbf{X}_N(u) : N \in \mathbb{N}, u \in \mathbf{U}_N)$ and $\mathbf{Y} = (\mathbf{Y}_N(u) : N \in \mathbb{N}, u \in \mathbf{U}_N)$, we say that \mathbf{X} is stochastically dominated by \mathbf{Y} , if for all $\varepsilon' > 0$ and $D > 0$ we have

$$\sup_{u \in \mathbf{U}_N} \mathbb{P}\left(\mathbf{X}_N(u) \geq N^{\varepsilon'} \mathbf{Y}_N(u)\right) \leq N^{-D} \tag{1.12}$$

for all sufficiently large $N \geq N_0(\varepsilon', D)$. In this case we write $\mathbf{X} \prec \mathbf{Y}$.

The set \mathbf{U}_N is omitted from the notation $\mathbf{X} \prec \mathbf{Y}$. Whenever we want to emphasize the role of \mathbf{U}_N , we say that $\mathbf{X}_N(u) \prec \mathbf{Y}_N(u)$ holds for all $u \in \mathbf{U}_N$. For example, by (1.1) and Assumption 1.5, we have

$$|h_{ab}| \prec 1/\sqrt{M}, \quad \forall a, b = 1, \dots, N. \tag{1.13}$$

Note that here $\mathbf{U}_N = \{u = (a, b) : a, b = 1, \dots, N\}$. In some applications, we also use this notation for random variables without any parameter or with a fixed parameter, i.e. the set of parameters \mathbf{U}_N plays no role.

Note that \tilde{S} is doubly stochastic. It is known that the empirical eigenvalue distribution of H_N converges to the semicircle law, whose density function is given by

$$\varrho_{sc}(x) := \frac{1}{2\pi} \sqrt{4 - x^2} \cdot \mathbf{1}(|x| \leq 2),$$

see [14] for instance. We denote the Green’s function of H_N by

$$G(z) \equiv G_N(z) := (H_N - z)^{-1}, \quad z = E + i\eta \in \mathbb{C}^+ := \{w \in \mathbb{C} : \text{Im} w > 0\}$$

and its (a, b) matrix element is $G_{ab}(z)$. Throughout the paper, we will always use E and η to denote the real and imaginary part of z without further mention. In addition, for simplicity, we suppress the subscript N from the notation of the matrices here and there. The Stieltjes transform of $\varrho_{sc}(x)$ is

$$m_{sc}(z) = \int_{-2}^2 \frac{\varrho_{sc}(x)}{x - z} dx = \frac{-z + \sqrt{z^2 - 4}}{2},$$

where we chose the branch of the square root with positive imaginary part for $z \in \mathbb{C}^+$. Note that $m_{sc}(z)$ is a solution to the following self-consistent equation

$$m_{sc}(z) = \frac{1}{-z - m_{sc}(z)}. \tag{1.14}$$

The semicircle law also holds in a local sense, see Theorem 2.3 in [12]. For simplicity, we cite this result with a slight modification adjusted to our assumption.

Proposition 1.7 (Erdős, Knowles, Yau, Yin, [12]) *Let H be a random block band matrix satisfying Assumptions 1.1, 1.3 and 1.5. Then, for any fixed small positive constants κ and ε , we have*

$$\begin{aligned} \max_{a,b} |G_{ab}(z) - \delta_{ab}m_{sc}(z)| &< (M\eta)^{-\frac{1}{2}}, \quad \text{if } E \in [-2 + \kappa, 2 - \kappa] \\ \text{and } M^{-1+\varepsilon} &\leq \eta \leq 10. \end{aligned} \tag{1.15}$$

Remark 1.8 We remark that Theorem 2.3 in [12] was established under a more general assumption $\sum_k \sigma_{jk}^2 = 1$ and $\sigma_{jk}^2 \leq C/M$. Especially, the block structure on the variance profile is not needed. In addition, Theorem 2.3 in [12] also covers the edges of the spectrum, which will not be discussed in this paper. We also refer to [14] for a previous result, see Theorem 2.1 therein.

Our aim in this paper is to extend the local semicircle law to the regime $\eta \gg N^{-1}$ and replace M with N in (1.15). More specifically, we will work in the following set, defined for arbitrarily small constant $\kappa > 0$ and any sufficiently small positive constant $\varepsilon_2 := \varepsilon_2(\varepsilon_1)$,

$$\mathbf{D}(N, \kappa, \varepsilon_2) := \left\{ z = E + i\eta \in \mathbb{C} : |E| \leq \sqrt{2} - \kappa, N^{-1+\varepsilon_2} \leq \eta \leq M^{-1}N^{\varepsilon_2} \right\}. \tag{1.16}$$

Throughout the paper, we will assume that ε_2 is much smaller than ε_1 , see (1.8) for the latter. Specifically, there exists some large enough constant C such that $\varepsilon_2 \leq \varepsilon_1/C$.

Theorem 1.9 (Local semicircle law) *Suppose that H is a random block band matrix satisfying Assumptions 1.1, 1.3 and 1.5. Let κ be an arbitrarily small positive constant and ε_2 be any sufficiently small positive constant. Then*

$$\max_{a,b} |G_{ab}(z) - \delta_{ab}m_{sc}(z)| < (N\eta)^{-\frac{1}{2}} \tag{1.17}$$

for all $z \in \mathbf{D}(N, \kappa, \varepsilon_2)$.

Remark 1.10 In fact, (1.17) together with the fact that $G_{ab}(z)$ and $m_{sc}(z)$ are Lipschitz functions of z with Lipschitz constant η^{-2} imply the uniformity of the estimate in z in the following stronger sense

$$\max_{z \in \mathbf{D}(N, \kappa, \varepsilon_2)} \max_{a,b} \left[(N\eta)^{\frac{1}{2}} |G_{ab}(z) - \delta_{ab}m_{sc}(z)| \right] < 1. \tag{1.18}$$

Remark 1.11 The restriction $|E| \leq \sqrt{2} - \kappa$ in (1.16) is technical. We believe the result can be extended to the whole bulk regime of the spectrum, i.e., $|E| \leq 2 - \kappa$. The upper bound of η in (1.16) is also technical. However, for $\eta > M^{-1}N^{\varepsilon_2}$, one can control the Green’s function by (1.15) directly.

Let $\lambda_1, \dots, \lambda_N$ be the eigenvalues of H_N . We denote by $\mathbf{u}_i := (u_{i1}, \dots, u_{iN})$ the normalized eigenvector of H_N corresponding to λ_i . From Theorem 1.9, we can also get the following delocalization property for the eigenvectors.

Theorem 1.12 (Complete delocalization) *Let H be a random block band matrix satisfying Assumptions 1.1, 1.3 and 1.5. We have*

$$\max_{i: |\lambda_i| \leq \sqrt{2} - \kappa} \|\mathbf{u}_i\|_\infty < N^{-\frac{1}{2}}. \tag{1.19}$$

Remark 1.13 We remark that delocalization in a certain weak sense was proven in [11] for an even more general class of random band matrices if $M \gg N^{4/5}$. However, Theorem 1.12 asserts delocalization for **all** eigenvectors in a very strong sense (supremum norm), while Proposition 7.1 of [11] stated that **most** eigenvectors are delocalized in a sense that their substantial support cannot be too small.

1.3 Outline of the proof strategy and novelties

In this section, we briefly outline the strategy for the proof of Theorem 1.9.

The first step, which is the main task of the whole proof, is to establish the following Theorem 1.15, namely, a prior estimate of the Green’s function in the Gaussian case. For technical reason, we need the following slight modification of Assumption 1.3, to state the result.

Assumption 1.14 (On M) Let ε_1 be the small positive constant in Assumption 1.3. We assume

$$N(\log N)^{-10} \geq M \geq W^{4+2\gamma+\varepsilon_1}. \tag{1.20}$$

In the regime $M \geq N(\log N)^{-10}$, we see that (1.17) anyway follows from (1.15) directly.

Theorem 1.15 *Assume that H is a Gaussian block band matrix, satisfying Assumptions 1.1 and 1.14. Let n be any fixed positive integer. Let κ be an arbitrarily small positive constant and ε_2 be any sufficiently small positive constant. There is $N_0 = N_0(n)$, such that for all $N \geq N_0$ and all $z \in \mathbf{D}(N, \kappa, \varepsilon_2)$, we have*

$$\mathbb{E}|G_{ab}(z)|^{2n} \leq N^{C_0} \left(\delta_{ab} + \frac{1}{(N\eta)^n} \right), \quad \forall a, b = 1, \dots, N \quad (1.21)$$

for some positive constant C_0 independent of n and z .

Remark 1.16 Much more delicate analysis can show that the prefactor N^{C_0} can be improved to some n -dependent constant C_n . We refer to Sect. 12 for further comment on this issue.

Using the definition of stochastic domination in Definition 1.6, a simple Markov inequality shows that (1.21) implies

$$|G_{ab}(z)| < \delta_{ab} + (N\eta)^{-\frac{1}{2}}, \quad \forall a, b = 1, \dots, N. \quad (1.22)$$

The proof of Theorem 1.15 is the main task of our paper. We will use the supersymmetry method. We partially rely on the arguments from Shcherbina's work [20] concerning universality of the local 2-point function and we develop new techniques to treat our observable, the high moment of the entries of $G(z)$, under a more general setting. We will comment on the novelties later in this subsection.

The second step is to generalize Theorem 1.15 from the Gaussian case to more general distribution satisfying Assumption 1.5, via a Green's function comparison strategy initiated in [14], see Lemma 2.1 below.

The last step is to use Lemma 2.1 and its Corollary 2.2 to prove our main theorems. Using (1.22) above to bound the error term in the self-consistent equation for the Green's function, we can prove Theorem 1.9 by a continuity argument in z , with the aid of the initial estimate for large η provided in Proposition 1.7. Theorem 1.12 will then easily follow from Theorem 1.9.

The second and the last steps are carried out in Sect. 2. The main body of this paper, Sects. 3–11 is devoted to the proof of Theorem 1.15.

One of the main novelty of this work is to combine the supersymmetry method and the Green's function comparison strategy to go beyond the Gaussian ensemble, which was so far the only random band matrix ensemble amenable to the supersymmetry method, as mentioned at the beginning. The comparison strategy requires an a priori control on the individual matrix elements of the Green's function with high probability [(see (1.22)], this is one of our main motivations behind Theorem 1.15.

Although we consider a different observable than [20], many technical aspects of the supersymmetric analysis overlaps with [20]. For the convenience of the reader, we now briefly introduce the strategy of [20], and highlight the main novelties of our work.

In [20], the author considers the 2-point correlation function of the trace of the resolvent of the Gaussian block band matrix H , with the variance profile $\tilde{S} = 1 + a\Delta$, under the assumption $M \sim N$ (note that we use M instead of W in [20] for the size of the blocks). The 2-point correlation function can be expressed in terms of a superintegral of a superfunction $F(\{\check{S}_i\}_{i=1}^W)$ with a collection of 4×4 supermatrices $\check{S}_i := \mathcal{Z}_i^* \mathcal{Z}_i$. Here for each i , $\mathcal{Z}_i = (\Psi_{1,i}, \Psi_{2,i}, \Phi_{1,i}, \Phi_{2,i})$ is an $M \times 4$ matrix and \mathcal{Z}_i^* is its conjugate transpose, where $\Psi_{1,i}$ and $\Psi_{2,i}$ are Grassmann M -vectors whilst $\Phi_{1,i}$ and $\Phi_{2,i}$ are complex M -vectors. Then, by using the superbosonization formula in the nonsingular case ($M \geq 4$) from [18], one can transform the superintegral of $F(\{\check{S}_i\}_{i=1}^W)$ to a superintegral of $F(\{S_i\}_{i=1}^W)$, where each S_i is a supermatrix akin to \check{S}_i , but only consists of 16 independent variables (either complex or Grassmann). We will call the integral representation of the observable after using the superbosonization formula as the *final integral representation*. Schematically it has the form

$$\int \mathbf{g}(\mathcal{S}_c) e^{Mf_c(\mathcal{S}_c) + f_g(\mathcal{S}_g, \mathcal{S}_c)} d\mathcal{S}, \tag{1.23}$$

for some functions $\mathbf{g}(\cdot)$, $f_c(\cdot)$ and $f_g(\cdot)$, where we used the abbreviation $\mathcal{S} := \{S_i\}_{i=1}^W$ and \mathcal{S}_c and \mathcal{S}_g represents the collection of all complex variables and Grassmann variables in \mathcal{S} , respectively. Here, $\mathbf{g}(\mathcal{S}_c)$ and $f_c(\mathcal{S}_c)$ are some complex functions and $f_g(\mathcal{S}_g, \mathcal{S}_c)$ will be mostly regarded as a function of the Grassmann variables with complex variables as its parameters. The number of variables (either complex or Grassmann) in the final integral representation then turns out to be of order W , which is much smaller than the original order N . In fact, in [20] it is assumed that $W = O(1)$ although the author also mentions the possibility to deal with the case $W \sim N^\varepsilon$ for some small positive ε , see the remark below Theorem 1 therein.

Performing a saddle point analysis for the complex measure $\exp\{Mf_c(\mathcal{S}_c)\}$, one can restrict the integral in a small vicinity of some saddle point, say, $\mathcal{S}_c = \mathcal{S}_{c0}$. It turns out that $f_c(\mathcal{S}_{c0}) = 0$ and $f_c(\mathcal{S}_c)$ decays quadratically away from \mathcal{S}_{c0} . Consequently, by plugging in the saddle point \mathcal{S}_{c0} , one can estimate $\mathbf{g}(\mathcal{S}_c)$ by $\mathbf{g}(\mathcal{S}_{c0})$ directly. However, for $\exp\{Mf_c(\mathcal{S}_c)\}$ and $\exp\{f_g(\mathcal{S}_g, \mathcal{S}_c)\}$, one shall expand them around the saddle point. Roughly speaking, in some vicinity of \mathcal{S}_{c0} , one will find that the expansions read

$$e^{Mf_c(\mathcal{S}_c)} = \exp\{-\mathbf{u}' \mathbb{A} \mathbf{u} + \mathbf{e}_c(\mathbf{u})\}, \quad e^{f_g(\mathcal{S}_g, \mathcal{S}_c)} = \exp\{-\boldsymbol{\rho}' \mathbb{H} \boldsymbol{\tau}\} \mathbf{p}(\boldsymbol{\rho}, \boldsymbol{\tau}, \mathbf{u}). \tag{1.24}$$

Here \mathbf{u} is a real vector of dimension $O(W)$, which is essentially a vectorization of $\sqrt{M}(\mathcal{S}_c - \mathcal{S}_{c0})$; $\mathbf{e}_c(\mathbf{u}) = o(1)$ is some error term; $\boldsymbol{\rho}$ and $\boldsymbol{\tau}$ are two Grassmann vectors of dimension $O(W)$. \mathbb{H} is a complex matrix [(c.f. (9.26)], and \mathbb{A} is a complex matrix with positive-definite Hermitian part [(the explicit form of \mathbb{A} can be read from (8.30)]. Moreover, \mathbb{A} is closely related to \mathbb{H} in the sense that determinant of a certain minor of \mathbb{H} (after two rows and two columns removed) is proportional to the square root of the determinant of \mathbb{A} , up to trivial factors. In addition, $\mathbf{p}(\boldsymbol{\rho}, \boldsymbol{\tau}, \mathbf{u})$ is the expansion of $\exp\{f_g(\mathcal{S}_g, \mathcal{S}_c) - f_g(\mathcal{S}_g, \mathcal{S}_{c0})\}$, which possesses the form

$$\mathbf{p}(\boldsymbol{\rho}, \boldsymbol{\tau}, \mathbf{u}) = \sum_{\ell=0}^{O(W)} M^{-\frac{\ell}{2}} \mathbf{p}_\ell(\boldsymbol{\rho}, \boldsymbol{\tau}, \mathbf{u}), \tag{1.25}$$

where $\mathfrak{p}_\ell(\boldsymbol{\rho}, \boldsymbol{\tau}, \mathbf{u})$ is a polynomial of the components of $\boldsymbol{\rho}$ and $\boldsymbol{\tau}$ with degree 2ℓ , regarding \mathbf{u} as fixed parameters. Now, keeping the leading order term of $\mathfrak{p}(\boldsymbol{\rho}, \boldsymbol{\tau}, \mathbf{u})$, and discarding the remainder terms, one can get the final estimate of the integral by taking the Gaussian integral over \mathbf{u} , $\boldsymbol{\rho}$ and $\boldsymbol{\tau}$. This completes the summary of [20].

Similarly to [20], we also use the superbosonization formula to reduce the number of variables and perform the saddle point analysis on the resulting integral. However, owing to the following three main aspects, our analysis is significantly different from [20].

- (Different observable) Our objective is to compute high moments of the single entry of the Green’s function. By using Wick’s formula (see Proposition 3.1), we express $\mathbb{E}|G_{jk}|^{2n}$ in terms of a superintegral of some superfunction of the form

$$\tilde{F} \left(\{\Psi_{a,j}, \Psi_{a,j}^*, \Phi_{a,j}, \Phi_{a,j}^*\}_{a=1,2; j=1,\dots,W} \right) := (\bar{\phi}_{1,q,\beta} \phi_{1,p,\alpha} \bar{\phi}_{2,p,\alpha} \phi_{2,q,\beta})^n F(\{\check{\mathcal{S}}_i\}_{i=1}^W)$$

for some $p, q \in \{1, \dots, W\}$ and $\alpha, \beta \in \{1, \dots, M\}$, where $\phi_{1,p,\alpha}$ is the α th coordinate of $\Phi_{1,p}$, and the others are defined analogously. Unlike the case in [20], \tilde{F} is not a function of $\{\check{\mathcal{S}}_i\}_{i=1}^W$ only. Hence, using the superbosonization formula to change $\check{\mathcal{S}}_i$ to \mathcal{S}_i directly is not feasible in our case. In order to handle the factor $(\bar{\phi}_{1,q,\beta} \phi_{1,p,\alpha} \bar{\phi}_{2,p,\alpha} \phi_{2,q,\beta})^n$, the main idea is to split off certain rank-one supermatrices from $\check{\mathcal{S}}_p$ and $\check{\mathcal{S}}_q$ such that this factor can be expressed in terms of the entries of these rank-one supermatrices. Then we use the superbosonization formula not only in the nonsingular case from [18] but also in the singular case from [3] to change and reduce the variables, resulting the final integral representation of $\mathbb{E}|G_{jk}|^{2n}$. Though this final integral representation, very schematically, is still of the form (1.23), due to the decomposition of the supermatrices $\check{\mathcal{S}}_p$ and $\check{\mathcal{S}}_q$, it is considerably more complicated than its counterpart in [20]. Especially, the function $\mathfrak{g}(\mathcal{S}_c)$ differs from its counterpart in [20], and its estimate at the saddle point follows from a different argument.

- (Small band width) In [20], the author considers the case that the band width M is comparable with N , i.e. the number of blocks W is finite. Though the derivation of the 2-point correlation function is highly nontrivial even with such a large band width, our objective, the local semicircle law and delocalization of the eigenvectors, however, can be proved for the case $M \sim N$ in a similar manner as for the Wigner matrix ($M = N$), see [12, 14]. In our work, we will work with much smaller band width to go beyond the results in [12, 14], see Assumption 1.3. Several main difficulties stemming from a narrow band width can be heuristically explained as follows.

At first, let us focus on the integral over the small vicinity of the saddle point, in which the exponential functions in the integrand in (1.23) approximately look like (1.24).

We regard the first term in (1.24) as a complex Gaussian measure, of dimension $O(W)$. When $W \sim 1$, one can discard the error term $\mathbf{e}_c(\mathbf{u})$ directly and perform the Gaussian integral over \mathbf{u} , due to the fact $\int d\mathbf{u} \exp\{-\mathbf{u}'\text{Re}(\mathbb{A})\mathbf{u}\} |\mathbf{e}_c(\mathbf{u})| = o(1)$.

However, such an estimate is not allowed when $W \sim N^\varepsilon$ (say), because the normalization of the measure $\exp\{-\mathbf{u}'\text{Re}(\mathbb{A})\mathbf{u}\}$ might be exponentially larger than that of $\exp\{-\mathbf{u}'\mathbb{A}\mathbf{u}\}$. In order to handle this issue, in Sect. 8.2, we will do a second deformation of the contours of the variables in \mathbf{u} , following the steepest descent paths exactly, whereby we can transform the complex Gaussian measure to a real one (c.f., (8.45)), thus the error term of the integral can be controlled.

Now, we turn to the second term in (1.24). When $W \sim 1$, there are only finitely many Grassmann variables. Hence, the complex coefficient of each term in the polynomial $\mathfrak{p}(\boldsymbol{\rho}, \boldsymbol{\tau}, \mathbf{u})$, which is of order $M^{-\ell/2}$ for some $\ell \in \mathbb{N}$ (see (1.25)), actually controls the magnitude of the integral of this term against the Gaussian measure $\exp\{-\boldsymbol{\rho}'\mathbb{H}\boldsymbol{\tau}\}$. Consequently, in case of $W \sim 1$, it suffices to keep the leading order term (according to $M^{-\ell/2}$), one may discard the others trivially, and compute the Gaussian integral over $\boldsymbol{\rho}$ and $\boldsymbol{\tau}$ explicitly. However, when $W \sim N^\varepsilon$ (say), in light of the Wick's formula (3.2) and the fact that the coefficients are of order $M^{-\ell/2}$, the order of the integral of each term of $\mathfrak{p}(\boldsymbol{\rho}, \boldsymbol{\tau}, \mathbf{u})$ against the Gaussian measure reads $M^{-\ell/2} \det \mathbb{H}^{(|\mathbb{I}||\mathbb{J}|)}$ for some index sets \mathbb{I} and \mathbb{J} and some $\ell \in \mathbb{N}$. Due to the fact $W \sim N^\varepsilon$, $\det \mathbb{H}^{(|\mathbb{I}||\mathbb{J}|)}$ is typically exponential in W . Hence, it is much more complicated to determine and compare the orders of the integrals of all $e^{O(W)}$ terms. In Sect. 9.1, in particular using Assumption 1.1 (iii) and Lemma 9.4, we perform a unified estimate for the integrals of all the terms, rather than simply estimate them by $M^{-\ell/2}$.

In addition, the analysis for the integral away from the vicinity of the saddle point in our work is also quite different from [20]. Actually, the integral over the complement of the vicinity can be trivially ignored in [20], since each factor in the integrand of (1.23) is of order 1, thus gaining any $o(1)$ factor for the integrand outside the vicinity is enough for the estimate. However, in our case, either $\exp\{Mf_c(\mathcal{S}_c)\}$ or $\int d\mathcal{S}_g \exp\{f_g(\mathcal{S}_g, \mathcal{S}_c)\}$ is essentially exponential in W . This fact forces us to provide an a priori bound for $\int d\mathcal{S}_g \exp\{f_g(\mathcal{S}_g, \mathcal{S}_c)\}$ in the full domain of \mathcal{S}_c rather than in the vicinity of the saddle point only. This step will be done in Sect. 6. In addition, in Sect. 7, an analysis of the tail behavior of the measure $\exp\{Mf_c(\mathcal{S}_c)\}$ will also be performed, in order to control the integral away from the vicinity of the saddle point.

- (General variance profile \tilde{S}) In [20], the authors considered the special case $S = a\Delta$ with $a < 1/4d$. We generalize the discussion to more general weighted Laplacians S satisfying Assumption 1.1, which, as a special case, includes the standard Laplacian Δ for any fixed dimension d .

1.4 Notation and organization

Throughout the paper, we will need some notation. At first, we conventionally use $U(r)$ to denote the unitary group of degree r , as well, $U(1, 1)$ denotes the group of 2×2 matrices Q obeying

$$Q^* \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} Q = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Furthermore, we denote

$$\mathring{U}(r) = U(r)/U(1)^r, \quad \mathring{U}(1, 1) = U(1, 1)/U(1)^2. \tag{1.26}$$

Recalling the real part E of z , we will frequently need the following two parameters

$$a_+ = \frac{\mathbf{i}E + \sqrt{4 - E^2}}{2}, \quad a_- = \frac{\mathbf{i}E - \sqrt{4 - E^2}}{2}. \tag{1.27}$$

Correspondingly, we define the following four matrices

$$\begin{aligned} D_{\pm} &= \text{diag}(a_+, a_-), & D_{\mp} &= \text{diag}(a_-, a_+), & D_+ &= \text{diag}(a_+, a_+), \\ D_- &= \text{diag}(a_-, a_-). \end{aligned} \tag{1.28}$$

We remark here D_{\pm} does not mean “ D_+ or D_- ”. In addition, we introduce the matrix

$$\mathfrak{J} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{1.29}$$

For simplicity, we introduce the following notation for some domains used throughout the paper.

$$\begin{aligned} \mathbb{I} &:= [0, 1], & \mathbb{L} &:= [0, 2\pi), & \Sigma &: \text{unit circle}, & \mathbb{R}_+ &:= [0, \infty), \\ \mathbb{R}_- &:= -\mathbb{R}_+, & \Gamma &:= a_+\mathbb{R}_+. \end{aligned} \tag{1.30}$$

For some $\ell \times \ell$ Hermitian matrix A , we use $\lambda_1(A) \leq \dots \leq \lambda_{\ell}(A)$ to represent its ordered eigenvalues. For some possibly N -dependent parameter set \mathbf{U}_N , and two families of complex functions $\{a_N(u) : N \in \mathbb{N}, u \in \mathbf{U}_N\}$ and $\{b_N(u) : N \in \mathbb{N}, u \in \mathbf{U}_N\}$, if there exists a positive constant $C > 1$ such that $C^{-1}|b_N(u)| \leq |a_N(u)| \leq C|b_N(u)|$ holds uniformly in N and u , we write $a_N(u) \sim b_N(u)$. Conventionally, we use $\{\mathbf{e}_i : i = 1, \dots, \ell\}$ to denote the standard basis of \mathbb{R}^{ℓ} , in which the dimension ℓ has been suppressed for simplicity. For some real quantities a and b , we use $a \wedge b$ and $a \vee b$ to represent $\min\{a, b\}$ and $\max\{a, b\}$, respectively.

Throughout the paper, $c, c', c_1, c_2, C, C', C_1, C_2$ represent some generic positive constants that are possibly n -dependent and may differ from line to line. In contrast, we use C_0 to denote some generic positive constant independent of n .

The paper will be organized in the following way. In Sect. 2, we prove Theorem 1.9 and Theorem 1.12, with Theorem 1.15. The proof of Theorem 1.15 will be done in Sects. 3–11. More specifically, in Sect. 3, we use the supersymmetric formalism to represent $\mathbb{E}|G_{ij}|^{2n}$ in terms of a superintegral, in which the integrand can be factorized into several functions; Sect. 4 is devoted to a preliminary analysis on these functions; Sects. 5–10 are responsible for different steps of the saddle point analysis, whose organization will be further clarified at the end of Sect. 5; Sect. 11 is devoted to the final proof of Theorem 1.15, by summing up the discussions in Sects. 3–10. In Sect. 12, we make a comment on how to remove the prefactor N^{C_0} in (1.21). At the end of the paper, we also collect some frequently used symbols in a table, for the convenience of the reader.

2 Proofs of Theorem 1.9 and Theorem 1.12

Assuming Theorem 1.15, we prove Theorems 1.9 and 1.12 in this section. At first, (1.21) can be generalized to the generally distributed matrix with the four moment matching condition via the Green’s function comparison strategy.

Lemma 2.1 *Assume that H is a random block band matrix, satisfying Assumptions 1.1, 1.5 and 1.14. Let κ be an arbitrarily small positive constant and ε_2 be any sufficiently small positive constant. There is $N_0 = N_0(n)$, such that for all $N \geq N_0$ and all $z \in \mathbf{D}(N, \kappa, \varepsilon_2)$, we have*

$$\mathbb{E}|G_{ab}(z)|^{2n} \leq N^{C_0} \left(\delta_{ab} + \frac{1}{(N\eta)^n} \right), \quad \forall a, b = 1, \dots, N \tag{2.1}$$

for some positive constant C_0 uniform in n and z .

By the definition of stochastic domination in Definition 1.6, we can get the following corollary immediately.

Corollary 2.2 *Under the assumptions of Lemma 2.1, we have*

$$|G_{ab}(z)| \prec \delta_{ab} + (N\eta)^{-\frac{1}{2}}, \quad \forall a, b = 1, \dots, N, \quad \forall z \in \mathbf{D}(N, \kappa, \varepsilon_2). \tag{2.2}$$

In the sequel, at first, we prove Lemma 2.1 from Theorem 1.15 via the Green’s function comparison strategy. Then we prove Theorem 1.9, using Lemma 2.1. Finally, we will show that Theorem 1.12 follows from Theorem 1.9 simply.

2.1 Green’s function comparison: Proof of Lemma 2.1

To show (2.1), we use Lindeberg’s replacement strategy to compare the Green’s functions of the Gaussian case and the general case. That means, we will replace the entries of H^s by those of H one by one, and compare the Green’s functions step by step. Choose and fix a bijective ordering map

$$\varpi : \{(i, j) : 1 \leq i \leq j \leq N\} \rightarrow \{1, \dots, \zeta(N)\}, \quad \zeta(N) := N(N + 1)/2. \tag{2.3}$$

Then we use H_k to represent the $N \times N$ random Hermitian matrix whose (i, j) th entry is $h_{i,j}$ if $\varpi(i, j) \leq k$, and is $h_{i,j}^g$ otherwise. Especially, we have $H_0 = H^g$ and $H_{\zeta(N)} = H$. Correspondingly, we define the Green’s functions by

$$G_k(z) := (H_k - z)^{-1}, \quad k = 1, \dots, \zeta(N).$$

Fix k and denote

$$\varpi^{-1}(k) = (a, b). \tag{2.4}$$

Then, we write

$$\begin{aligned}
 H_{k-1} &= H_k^0 + V_{ab}, \quad V_{ab} := \left(1 - \frac{\delta_{ab}}{2}\right) (h_{ab}^g \mathbf{e}_a \mathbf{e}_b^* + h_{ba}^g \mathbf{e}_b \mathbf{e}_a^*), \\
 H_k &= H_k^0 + W_{ab}, \quad W_{ab} := \left(1 - \frac{\delta_{ab}}{2}\right) (h_{ab} \mathbf{e}_a \mathbf{e}_b^* + h_{ba} \mathbf{e}_b \mathbf{e}_a^*),
 \end{aligned}$$

where H_k^0 is obtained via replacing h_{ab} and h_{ba} by 0 in H_k (or replacing h_{ab}^g and h_{ba}^g by 0 in H_{k-1}). In addition, we denote

$$G_k^0(z) = (H_k^0 - z)^{-1}.$$

Set $\varepsilon_3 \equiv \varepsilon_3(\gamma, \varepsilon_1)$ to be a sufficiently small positive constant, satisfying (say)

$$\varepsilon_3 \leq \frac{1}{100} \cdot \frac{\varepsilon_1}{5 + 2\gamma + \varepsilon_1}, \tag{2.5}$$

where γ is from Assumption 1.1 (iii) and ε_1 is from (1.8). For simplicity, we introduce the following parameters for $\ell = 1, \dots, \zeta(N)$ and $\iota, j = 1, \dots, N$,

$$\begin{aligned}
 \widehat{\Theta}_0 &:= N^{C_0}, \\
 \widehat{\Theta}_{\ell, \iota j} &:= \widehat{\Theta}_0 \left(1 + C \left(\frac{N^{\varepsilon_3}}{\sqrt{M}}\right)^5\right)^\ell \prod_{\varpi(a,b) \leq \ell} \left(1 + C \delta_{\{\iota, j\}\{a, b\}} \left(\frac{N^{\varepsilon_3} \sqrt{N\eta}}{\sqrt{M}}\right)^5\right),
 \end{aligned} \tag{2.6}$$

where C is a positive constant. Here we used the notation $\delta_{IJ} = 1$ if two index sets I and J are the same and $\delta_{IJ} = 0$ otherwise. It is easy to see that for $\eta \leq M^{-1} N^{\varepsilon_2}$, we have

$$\widehat{\Theta}_{\ell, \iota j} \leq 2\widehat{\Theta}_0, \quad \forall \ell = 1, \dots, \zeta(N), \quad \iota, j = 1, \dots, N, \tag{2.7}$$

by using (1.9). Now, we compare $G_{k-1}(z)$ and $G_k(z)$. We will prove the following lemma.

Lemma 2.3 *Suppose that the assumptions in Lemma 2.1 hold. Additionally, we assume that for some sufficiently small positive constant ε_3 satisfying (2.5),*

$$\begin{aligned}
 |(G_\ell)_{\iota j}(z)| &< N^{\varepsilon_3}, \quad |(G_\ell^0)_{\iota j}(z)| < N^{\varepsilon_3}, \quad \forall \ell = 1, \dots, \zeta(N), \quad \forall \iota, \\
 & \quad j = 1, \dots, N, \quad \forall z \in \mathbf{D}(N, \kappa, \varepsilon_2).
 \end{aligned} \tag{2.8}$$

Let $n \in \mathbb{N}$ be any given integer. Then, if

$$\begin{aligned}
 \mathbb{E}|(G_{k-1})_{\iota j}(z)|^{2n} &\leq \widehat{\Theta}_{k-1, \iota j} \left(\delta_{\iota j} + \frac{1}{(N\eta)^n}\right), \\
 \forall \iota, j &= 1, \dots, N, \quad \forall z \in \mathbf{D}(N, \kappa, \varepsilon_2)
 \end{aligned} \tag{2.9}$$

we also have

$$\mathbb{E}|(G_k)_{lJ}(z)|^{2n} \leq \widehat{\Theta}_{k,lJ} \left(\delta_{lJ} + \frac{1}{(N\eta)^n} \right), \quad \forall l, J = 1, \dots, N, \quad \forall z \in \mathbf{D}(N, \kappa, \varepsilon_2) \tag{2.10}$$

for any $k = 1, \dots, \zeta(N)$.

Proof of Lemma 2.3 Fix k and omit the argument z from now on, but all formulas are understood to hold for all $z \in \mathbf{D}(N, \kappa, \varepsilon_2)$. At first, under the conditions (2.8) and (2.9), we show that

$$\mathbb{E}|(G_k^0)_{lJ}|^{2n} \leq 3\widehat{\Theta}_0 \left(\delta_{lJ} + \frac{1}{(N\eta)^n} \right), \quad \forall l, J = 1, \dots, N. \tag{2.11}$$

To see this, we use the expansion with (2.4)

$$(G_k^0)_{lJ} = (G_{k-1})_{lJ} + (G_{k-1} \mathbf{V}_{ab} G_k^0)_{lJ},$$

which implies that for a sufficiently small $\varepsilon' > 0$ and a sufficiently large constant $D > 0$

$$\begin{aligned} \mathbb{E}|(G_k^0)_{lJ}|^{2n} &\leq \mathbb{E} \left(|(G_{k-1})_{lJ}| + \frac{N^{2\varepsilon_3 + \varepsilon'}}{\sqrt{M}} \right)^{2n} + \eta^{-2n} N^{-D} \\ &= \sum_{\ell=0}^{2n} \binom{2n}{\ell} \left(\frac{N^{2\varepsilon_3 + \varepsilon'}}{\sqrt{M}} \right)^\ell \mathbb{E}|(G_{k-1})_{lJ}|^{2n-\ell} + \eta^{-2n} N^{-D} \end{aligned} \tag{2.12}$$

where the first step follows from (1.13), (2.8), Definition 1.6 and the trivial bound η^{-1} for the Green’s functions. Now, using (2.9), (2.7) and Hölder inequality, we have

$$\mathbb{E}|(G_{k-1})_{lJ}|^{2n-\ell} \leq 2\widehat{\Theta}_0 \left(\delta_{lJ} + \frac{1}{(N\eta)^{n-\frac{\ell}{2}}} \right), \quad 0 \leq \ell \leq 2n. \tag{2.13}$$

In addition, for sufficiently small ε' , it is easy to check that there exists a constant $c > 0$ such that

$$\frac{N^{2\varepsilon_3 + \varepsilon'}}{\sqrt{M}} \leq N^{-c} \frac{1}{(N\eta)^{\frac{1}{2}}} \tag{2.14}$$

for $z \in \mathbf{D}(N, \kappa, \varepsilon_2)$, in light of the fact $M \gg N^{\frac{4}{5}}$, c.f., (1.9). Substituting (2.13) and (2.14) into (2.12) and choosing D to be sufficiently large, we can easily get the bound (2.11).

Now, recall (2.4) again and expand $G_{k-1}(z)$ and $G_k(z)$ around $G_k^0(z)$, namely

$$\begin{aligned}
 G_{k-1} &= G_k^0 + \sum_{\ell=1}^m (-1)^\ell (G_k^0 \mathbf{V}_{ab})^\ell G_k^0 + (-1)^{m+1} (G_k^0 \mathbf{V}_{ab})^{m+1} G_{k-1}, \\
 G_k &= G_k^0 + \sum_{\ell=1}^m (-1)^\ell (G_k^0 \mathbf{W}_{ab})^\ell G_k^0 + (-1)^{m+1} (G_k^0 \mathbf{W}_{ab})^{m+1} G_k. \tag{2.15}
 \end{aligned}$$

We always choose m to be sufficiently large, depending on ε_3 but independent of N . Then, we can write

$$\begin{aligned}
 (G_{k-1})_{iJ} &= (G_k^0)_{iJ} + \sum_{\ell=1}^m \mathbf{R}_{\ell,iJ} + \tilde{\mathbf{R}}_{m+1,iJ}, \\
 (G_k)_{iJ} &= (G_k^0)_{iJ} + \sum_{\ell=1}^m \mathbf{S}_{\ell,iJ} + \tilde{\mathbf{S}}_{m+1,iJ}, \tag{2.16}
 \end{aligned}$$

where

$$\begin{aligned}
 \mathbf{R}_{\ell,iJ} &:= (-1)^\ell \left((G_k^0 \mathbf{V}_{ab})^\ell G_k^0 \right)_{iJ}, \quad \mathbf{S}_{\ell,iJ} := (-1)^\ell \left((G_k^0 \mathbf{W}_{ab})^\ell G_k^0 \right)_{iJ}, \quad \ell = 1, \dots, m, \\
 \tilde{\mathbf{R}}_{m+1,iJ} &:= (-1)^{m+1} \left((G_k^0 \mathbf{V}_{ab})^{m+1} G_{k-1} \right)_{iJ}, \quad \tilde{\mathbf{S}}_{m+1,iJ} := (-1)^{m+1} \left((G_k^0 \mathbf{W}_{ab})^{m+1} G_k \right)_{iJ}. \tag{2.17}
 \end{aligned}$$

At first, by taking m sufficiently large, from (2.8) and (1.13), we have the trivial bound

$$|\tilde{\mathbf{R}}_{m+1,iJ}|, |\tilde{\mathbf{S}}_{m+1,iJ}| < M^{-\frac{m+1}{2}} N^{(m+2)\varepsilon_3} \ll \frac{1}{M^3 \sqrt{N\eta}}. \tag{2.18}$$

For $\mathbf{R}_{\ell,iJ}$ and $\mathbf{S}_{\ell,iJ}$, we split the discussion into off-diagonal case and diagonal case. In the case of $i \neq J$, we keep the first and the last factors of the terms in the expansions of $((G_k^0 \mathbf{V}_{ab})^\ell G_k^0)_{iJ}$ and $((G_k^0 \mathbf{W}_{ab})^\ell G_k^0)_{iJ}$, namely, $(G_k^0)_{iJ'}$ and $(G_k^0)_{i'J}$ for some $i', J' = a$ or b , and bound the factors in between by using (1.13) and (2.8), resulting the bound

$$|\mathbf{R}_{\ell,iJ}|, |\mathbf{S}_{\ell,iJ}| < M^{-\frac{\ell}{2}} N^{(\ell-1)\varepsilon_3} \sum_{i', J'=a,b} |(G_k^0)_{iJ'}| |(G_k^0)_{i'J}|, \quad \ell = 1, \dots, m. \tag{2.19}$$

For $i = J$, we only keep the first factor of the terms in the expansions of $((G_k^0 \mathbf{V}_{ab})^\ell G_k^0)_{ii}$ and $((G_k^0 \mathbf{W}_{ab})^\ell G_k^0)_{ii}$, and bound the others by using (1.13) and (2.8), resulting the bound

$$|\mathbf{R}_{\ell,ii}|, |\mathbf{S}_{\ell,ii}| < M^{-\frac{\ell}{2}} N^{\ell\varepsilon_3} \left(|(G_k^0)_{ia}| + |(G_k^0)_{ib}| \right), \quad \ell = 1, \dots, m. \tag{2.20}$$

Observe that, in case $i \neq J$, if $\{i, J\} \neq \{a, b\}$, at least one of $(G_k^0)_{iJ'}$ and $(G_k^0)_{i'J}$ is an off-diagonal entry of G_k^0 for $i', J' = a$ or b .

Now we compare the $2n$ th moment of $|(G_{k-1})_{iJ}|$ and $|(G_k)_{iJ}|$. At first, we write

$$\mathbb{E}|(G_d)_{iJ}|^{2n} = \mathbb{E}((G_d)_{iJ})^n (\overline{(G_d)_{iJ}})^n, \quad d = k - 1, k. \tag{2.21}$$

By substituting the expansion (2.16) into (2.21), we can write

$$\mathbb{E}|(G_d)_{iJ}|^{2n} = \mathbf{A}(t, J) + \mathbf{R}_d(t, J), \quad d = k - 1, k, \tag{2.22}$$

where $\mathbf{A}(t, J)$ is the sum of the terms which depend only on H_k^0 and the first four moments of h_{ab} , and $\mathbf{R}_d(t, J)$ is the sum of all the other terms. We claim that $\mathbf{R}_d(t, J)$ satisfies the bound

$$|\mathbf{R}_d(t, J)| \leq C \widehat{\Theta}_0 \left(\frac{N^{\varepsilon_3}}{\sqrt{M}} \right)^5 \left(\delta_{iJ} + \frac{\delta_{\{t,J\}\{a,b\}}}{(N\eta)^{n-\frac{5}{2}}} + \frac{1}{(N\eta)^n} \right), \quad d = k - 1, k, \tag{2.23}$$

for some positive constant C independent of n . Now, we verify (2.23). According to (2.11) and the fact that the sequence $\mathbf{R}_{1,iJ}, \dots, \mathbf{R}_{m,iJ}, \widetilde{\mathbf{R}}_{m+1,iJ}$, as well as $\mathbf{S}_{1,iJ}, \dots, \mathbf{S}_{m,iJ}, \widetilde{\mathbf{S}}_{m+1,iJ}$, decreases by a factor $N^{\varepsilon_3}/\sqrt{M}$ in magnitude, it is not difficult to check the leading order terms of $\mathbf{R}_{k-1}(t, J)$ are of the form

$$\mathbb{E} \left((G_k^0)_{iJ} \right)^p \left(\overline{(G_k^0)_{iJ}} \right)^{2n-p-\sum_{\ell=1}^5 (q_\ell + q'_\ell)} \prod_{\ell=1}^5 \mathbf{R}_{\ell,iJ}^{q_\ell} \overline{\mathbf{R}}_{\ell,iJ}^{q'_\ell}, \tag{2.24}$$

with some $p, q_\ell, q'_\ell \in \mathbb{N}$ such that

$$\sum_{\ell=1}^5 \ell(q_\ell + q'_\ell) = 5, \quad 0 \leq p \leq 2n - \sum_{\ell=1}^5 (q_\ell + q'_\ell), \tag{2.25}$$

and the leading order terms of $\mathbf{R}_k(t, J)$ possess the same form (with \mathbf{R} replaced by \mathbf{S}). Every other term has at least 6 factors of h_{ab} or h_{ab}^s or their conjugates, thus their sizes are typically controlled by $M^{-3}(N\eta)^{-n}$, i.e. they are subleading. Hence, it suffices to bound (2.24).

Now, the five factors of h_{ab} or h_{ba} within the $\mathbf{R}_{\ell,iJ}$'s in (2.24) are independent of the rest and estimated by $M^{-5/2}$. For the remaining factors from G_k^0 , we use (2.11) to bound $2n$ of them and use (2.8) to bound the rest. In the case that $i \neq J$ and $\{i, J\} \neq \{a, b\}$, by the discussion above, we must have an off-diagonal entry of G_k^0 in the product $(G_k^0)_{iJ'}(G_k^0)_{i'J}$ for any choice of $i', J' = a$ or b . Then, in the bound for $\mathbf{R}_{\ell,iJ}$ in (2.19), for each $(G_k^0)_{iJ'}(G_k^0)_{i'J}$, we keep the off-diagonal entry and bound the other by N^{ε_3} from assumption (2.8). Hence, by using (2.19) and (2.25), we see that for some $i_r, J_r \in \{i, J, a, b\}$ with $i_r \neq J_r, r = 1, \dots, \sum (q_\ell + q'_\ell)$, the following

bound holds

$$\begin{aligned}
 (2.24) &\leq \left(\frac{N^{\varepsilon_3}}{\sqrt{M}}\right)^5 \mathbb{E} \left(|(G_k^0)_{i,j}|^{2n - \sum_{\ell=1}^5 (q_\ell + q'_\ell)} \prod_{r=1}^{\sum_{\ell=1}^5 (q_\ell + q'_\ell)} |(G_k^0)_{i_r, j_r}| \right) \\
 &\leq 3 \left(\frac{N^{\varepsilon_3}}{\sqrt{M}}\right)^5 \frac{\widehat{\Theta}_0}{(N\eta)^n}, \tag{2.26}
 \end{aligned}$$

where the last step follows from (2.11) and Hölder’s inequality. In case of $i \neq j$ but $\{i, j\} = \{a, b\}$, we keep an entry in the product $(G_k^0)_{i, j'}(G_k^0)_{i', j}$ and bound the other by N^{ε_3} . We remark here in this case the entry being kept can be either diagonal or off-diagonal. Consequently, for some $i_r, j_r \in \{i, j, a, b\}, r = 1, \dots, \sum(q_\ell + q'_\ell)$, we have the bound

$$\begin{aligned}
 (2.24) &\leq \left(\frac{N^{\varepsilon_3}}{\sqrt{M}}\right)^5 \mathbb{E} \left(|(G_k^0)_{i,j}|^{2n - \sum_{\ell=1}^5 (q_\ell + q'_\ell)} \prod_{r=1}^{\sum_{\ell=1}^5 (q_\ell + q'_\ell)} |(G_k^0)_{i_r, j_r}| \right) \\
 &\leq 3 \left(\frac{N^{\varepsilon_3}}{\sqrt{M}}\right)^5 \frac{\widehat{\Theta}_0}{(N\eta)^{n - \frac{5}{2}}}, \tag{2.27}
 \end{aligned}$$

by using (2.11) and Hölder’s inequality again. Hence, we have shown (2.23) in the case of $i \neq j$. For $i = j$, it is analogous to show

$$(2.24) \leq 3 \left(\frac{N^{\varepsilon_3}}{\sqrt{M}}\right)^5 \widehat{\Theta}_0 \tag{2.28}$$

by using (2.11), (2.20) and Hölder’s inequality. Hence, we verified (2.23). Consequently, by Assumption 1.5, (2.22) and (2.23) we have

$$\left| \mathbb{E}|(G_{k-1})_{i,j}|^{2n} - \mathbb{E}|(G_k)_{i,j}|^{2n} \right| \leq C \widehat{\Theta}_0 \left(\frac{N^{\varepsilon_3}}{\sqrt{M}}\right)^5 \left(\delta_{i,j} + \frac{\delta_{\{i,j\}\{a,b\}}}{(N\eta)^{n - \frac{5}{2}}} + \frac{1}{(N\eta)^n} \right),$$

which together with the assumption (2.9) for $\mathbb{E}|(G_{k-1})_{i,j}|^{2n}$ and the definition of $\widehat{\Theta}_{\ell, i, j}$ ’s in (2.6), we can get (2.10). Hence, we completed the proof of Lemma 2.3. □

To show (2.1), we also need the following lemma.

Lemma 2.4 *Suppose that the assumptions in Lemma 2.1 hold. Fix the indices $a, b \in \{1, \dots, N\}$. Let H^0 be a matrix obtained from H with its (a, b) th entry replaced by 0. Then, if for some $\eta_0 \geq 1/N$ there exists*

$$|G_{ii}(z)| < 1, \quad |(G^0)_{ii}(z)| < 1 \quad \text{for } \eta \geq \eta_0, \quad \forall i = 1, \dots, N, \tag{2.29}$$

then we also have

$$|G_{\iota J}(z)| < \frac{\eta_0}{\eta}, \quad |(G^0)_{\iota J}(z)| < \frac{\eta_0}{\eta}, \quad \text{for } \frac{1}{N} < \eta \leq \eta_0, \quad \forall \iota, J = 1, \dots, N.$$

Proof of Lemma 2.4 The proof is almost the same as the discussion on pages 2311–2312 in [10]. For the convenience of the reader, we sketch it below. At first, according to the discussion below (4.28) in [10], for any $\iota, J = 1, \dots, N$, we have

$$|G_{\iota J}(E + \mathbf{i}\eta)| \leq C \max_{\ell} \sum_{k \geq 0} \text{Im} G_{\ell\ell}(E + \mathbf{i}2^k \eta).$$

Now, we set $k_1 := \max\{k : 2^k \eta < \eta_0\}$ and $k_2 := \max\{k : 2^k \eta < 1\}$. According to our assumption, both k_1 and k_2 are of the order $\log N$. Now, we have

$$\begin{aligned} \sum_{k \geq 0} \text{Im} G_{\ell\ell}(E + \mathbf{i}2^k \eta) &= \sum_{k=0}^{k_1} \text{Im} G_{\ell\ell}(E + \mathbf{i}2^k \eta) + \sum_{k=k_1}^{k_2} \text{Im} G_{\ell\ell}(E + \mathbf{i}2^k \eta) \\ &\quad + \sum_{k=k_2+1}^{\infty} \text{Im} G_{\ell\ell}(E + \mathbf{i}2^k \eta) \\ &< \frac{\eta_0}{\eta} \sum_{k=0}^{k_1} \frac{1}{2^k} \text{Im} G_{\ell\ell}(E + \mathbf{i}\eta_0) + (k_2 - k_1) + 1 < \frac{\eta_0}{\eta} \end{aligned}$$

where in the second step, we used the fact that the function $y \mapsto y \text{Im} G_{\ell\ell}(E + \mathbf{i}y)$ is monotonically increasing, the condition (2.29) and the fact $\eta \leq \eta_0$. Hence, we conclude the proof of Lemma 2.4. \square

Now, with Theorem 1.15, Lemmas 2.3 and 2.4, we can prove Lemma 2.1.

Proof for Lemma 2.1 The proof relies on the following bootstrap argument, namely, we show that once

$$|(G_{\ell})_{\iota J}| < 1, \quad \forall \ell = 1, \dots, \zeta(N), \quad \forall \iota, J = 1, \dots, N \tag{2.30}$$

holds for $\eta \geq \eta_0$ with $\eta_0 \in [N^{-1+\varepsilon_2+\varepsilon_3}, M^{-1}N^{\varepsilon_2}]$, it also holds for $\eta \geq \eta_0 N^{-\varepsilon_3}$ for any ε_3 satisfying (2.5). Assuming (2.30) holds for $\eta \geq \eta_0$, we see that

$$\begin{aligned} \max_{\iota, J} |(G_{\ell}^0)_{\iota J}| &= \max_{\iota, J} |(G_{\ell})_{\iota J} + ((G_{\ell})W_{ab}G_{\ell}^0)_{\iota J}| \\ &< \max_{\iota, J} |(G_{\ell})_{\iota J}| \left(1 + \frac{1}{\sqrt{M}} \max_{\iota, J} |(G_{\ell}^0)_{\iota J}| \right) \\ &< 1 + \frac{1}{\sqrt{M}} \max_{\iota, J} |(G_{\ell}^0)_{\iota J}|. \end{aligned}$$

Consequently, for $\eta \geq \eta_0$, we also have

$$|(G_\ell^0)_{\iota J}| < 1, \quad \forall \ell = 1, \dots, \zeta(N), \quad \forall \iota, J = 1, \dots, N. \tag{2.31}$$

Therefore, (2.29) holds. Then, by Lemma 2.4, we see that (2.8) holds for $\eta \geq \eta_0 N^{-\varepsilon_3}$. Furthermore, by Lemma 2.3 and Theorem 1.15 for G_0 , i.e. the Gaussian case, one can get that for any given n ,

$$\begin{aligned} \mathbb{E}|(G_\ell)_{\iota J}|^{2n} &\leq \widehat{\Theta}_{\ell, \iota J} \left(\delta_{\iota J} + \frac{1}{(N\eta)^n} \right) \leq 2\widehat{\Theta}_0 \left(\delta_{\iota J} + \frac{1}{(N\eta)^n} \right), \\ &\text{for } M^{-1}N^{\varepsilon_2} \geq \eta \geq \eta_0 N^{-\varepsilon_3}, \\ &\forall \ell = 1, \dots, \zeta(N), \quad \forall \iota, J = 1, \dots, N. \end{aligned} \tag{2.32}$$

Note that since (2.32) holds for any given n , we get (2.30) for $M^{-1}N^{\varepsilon_2} \geq \eta \geq \eta_0 N^{-\varepsilon_3}$.

Now we start from $\eta_0 = M^{-1}N^{\varepsilon_2}$. By Proposition 1.7 we see that (2.30) holds for all $\eta \geq \eta_0$. Then we can use the bootstrap argument above finitely many times to show (2.30) holds for all $\eta \geq N^{-1+\varepsilon_2}$. Consequently, we have (2.8) for all $\eta \geq N^{-1+\varepsilon_2}$. Then, Lemma 2.1 follows from Lemma 2.3 and Theorem 1.15 immediately. \square

2.2 Proof of Theorem 1.9

Without loss of generality we can assume that $M \leq N(\log N)^{-10}$, otherwise, Proposition 1.7 implies (1.17) immediately. We only need to consider the diagonal entries G_{ii} below, since the bound for the off-diagonal entries of $G(z)$ is implied by (2.1) directly. For simplicity, we introduce the notation

$$\Lambda_d \equiv \Lambda_d(z) := \max_i |G_{ii}(z) - m_{sc}(z)|.$$

To bound Λ_d , a key technical input is the estimate for the quantity

$$\zeta_i \equiv \zeta_i(z) := (G_{ii})^{-1} + z + \sum_a \sigma_{ai}^2 G_{aa}, \tag{2.33}$$

which is given in the following lemma.

Lemma 2.5 *Suppose that H satisfies Assumptions 1.1, 1.5 and 1.14. We have*

$$\max_{i=1, \dots, N} |\zeta_i(z)| < (N\eta)^{-\frac{1}{2}}, \tag{2.34}$$

for all $z \in \mathbf{D}(N, \kappa, \varepsilon_2)$.

The proof of Lemma 2.5 will be postponed. Using Lemma 2.5, we see that, with high probability, (2.33) is a small perturbation of the self-consistent equation of m_{sc} , i.e. (1.14), considering $\sum_a \sigma_{ai}^2 = 1$. To control Λ_d , we use a continuity argument from [12].

We remind here that in the sequel, the parameter set of the stochastic dominance is always $\mathbf{D}(N, \kappa, \varepsilon_2)$, without further mention. We need to show that

$$\Lambda_d < (N\eta)^{-\frac{1}{2}}, \tag{2.35}$$

and first we claim that it suffices to show that

$$\mathbf{1}\left(\Lambda_d \leq N^{-\frac{\varepsilon_2}{4}}\right) \Lambda_d < (N\eta)^{-\frac{1}{2}}. \tag{2.36}$$

Indeed, if (2.36) were proven, we see that with high probability either $\Lambda_d > N^{-\frac{\varepsilon_2}{4}}$ or $\Lambda_d < (N\eta)^{-\frac{1}{2}} \leq N^{-\frac{\varepsilon_2}{2}}$ for $z \in \mathbf{D}(N, \kappa, \varepsilon_2)$. That means, there is a gap in the possible range of Λ_d . Now, choosing ε in (1.15) to be sufficiently small, we are able to get for $\eta = M^{-1}N^{\varepsilon_2}$,

$$\Lambda_d < N^{-\frac{\varepsilon_2}{2}}, \quad \forall E \in [-2 + \kappa, 2 - \kappa], \quad \forall i = 1, \dots, N. \tag{2.37}$$

By the fact that Λ_d is continuous in z , we see that with high probability, Λ_d can only stay in one side of the range, namely, (2.35) holds. The rigorous details of this argument involve considering a fine discrete grid of the z -parameter and using that $G(z)$ is Lipschitz continuous (albeit with a large Lipschitz constant $1/\eta$). The details are found in Section 5.3 of [12].

Hence, what remains is to verify (2.36). The proof of (2.36) is almost the same as that for Lemma 3.5 in [14]. For the convenience of the reader, we sketch it below without reproducing the details. We set

$$\bar{m} \equiv \bar{m}(z) := \frac{1}{N} \sum_{i=1}^N G_{ii}(z), \quad \mathbf{u}_i \equiv \mathbf{u}_i(z) := G_{ii} - \bar{m}, \quad i = 1, \dots, N.$$

We also denote $\mathbf{u} := (\mathbf{u}_1, \dots, \mathbf{u}_N)$. By the assumption $\Lambda_d \leq N^{-\frac{\varepsilon_2}{4}}$, we have

$$\mathbf{u}_i = O\left(N^{-\frac{\varepsilon_2}{4}}\right). \tag{2.38}$$

Now we rewrite (2.33) as

$$0 = G_{ii} + \frac{1}{z + \sum_a \sigma_{ai}^2 G_{aa} - \zeta_i} =: G_{ii} + \frac{1}{z + \bar{m}(z)} + \tilde{\zeta}_i. \tag{2.39}$$

By using (2.34), Lemma 5.1 in [14], and the assumption $\Lambda_d \leq N^{-\frac{\varepsilon_2}{4}}$, we can show that

$$\tilde{\zeta}_i = -\frac{\sum_a \sigma_{ai}^2 \mathbf{u}_a}{(z + \bar{m}(z))^2} + O\left(\|\mathbf{u}\|_\infty^2\right) + O\left(\max_i |\zeta_i|\right). \tag{2.40}$$

One can refer to the derivation of (5.14) in [14] for more details. Averaging over i for (2.39) and (2.40) leads to

$$\bar{m}(z) + \frac{1}{z + \bar{m}(z)} = -\tilde{\zeta}, \tag{2.41}$$

where

$$\tilde{\zeta} := \frac{1}{N} \sum_{i=1}^N \tilde{\zeta}_i = O\left(\|\mathbf{u}\|_\infty^2\right) + O\left(\max_i |\zeta_i|\right) \tag{2.42}$$

Plugging (2.38) and (2.34) into (2.42) yields

$$|\tilde{\zeta}| < N^{-\frac{\varepsilon_2}{4}}. \tag{2.43}$$

Using (2.43), the fact $|\bar{m}(z) - m_{sc}(z)| \leq \Lambda_d \leq N^{-\frac{\varepsilon_2}{4}}$, and Lemma 5.2 in [14], to (2.41), we have

$$|\bar{m}(z) - m_{sc}(z)| \leq |\tilde{\zeta}| = O\left(\|\mathbf{u}\|_\infty^2\right) + O\left(\max_i |\zeta_i|\right), \tag{2.44}$$

where in the first step we have used the fact that $z \in \mathbf{D}(N, \kappa, \varepsilon_2)$ thus away from the edges of the semicircle law. Now, we combine (2.39), (2.40) and (2.41), resulting

$$\begin{aligned} u_i &= \frac{\sum_a \sigma_{ai}^2 u_a}{(z + \bar{m}(z))^2} + \tilde{\zeta} + O\left(\|\mathbf{u}\|_\infty^2\right) + O\left(\max_i |\zeta_i|\right) \\ &= w_i + \frac{\sum_a \sigma_{ai}^2 u_a}{(z + m_{sc}(z))^2}, \quad i = 1, \dots, N. \end{aligned} \tag{2.45}$$

We just take the above identity as the definition of w_i . Analogously, we set $\mathbf{w} := (w_1, \dots, w_N)'$. Then (2.42) and (2.45) imply

$$\begin{aligned} \|\mathbf{w}\|_\infty &= O\left(\|\mathbf{u}\|_\infty^2\right) + O\left(\max_i |\zeta_i|\right) + O\left(\|\mathbf{u}\|_\infty \cdot |(z + \bar{m}(z))^{-2} - (z + m_{sc}(z))^{-2}|\right) \\ &\leq O\left(\|\mathbf{u}\|_\infty^2\right) + O\left(\max_i |\zeta_i|\right) + O\left(\|\mathbf{u}\|_\infty \cdot |\bar{m}(z) - m_{sc}(z)|\right) \\ &\leq O\left(\|\mathbf{u}\|_\infty^2\right) + O\left(\max_i |\zeta_i|\right), \end{aligned} \tag{2.46}$$

where the second step follows from the fact $|z + m_{sc}(z)| \geq 1$ in $\mathbf{D}(N, \kappa, \varepsilon_2)$ (see (5.1) in [14] for instance), (2.43) and (2.44), and in the last step we used (2.44) again.

Now, using the fact $m_{sc}^2(z) = (m_{sc}(z) + z)^{-2}$ (see (1.14)), we rewrite (2.45) in terms of the matrix \mathcal{T} introduced in (1.6) as $\mathbf{u} = (1 - m_{sc}^2(z)\mathcal{T})^{-1}\mathbf{w}$. Consequently, we have

$$\|\mathbf{u}\|_\infty \leq \left\| (1 - m_{sc}^2(z)\mathcal{T})^{-1} \right\|_{\ell_\infty \rightarrow \ell_\infty} \|\mathbf{w}\|_\infty := \Gamma(z) \|\mathbf{w}\|_\infty. \tag{2.47}$$

Then for $z \in \mathbf{D}(N, \kappa, \varepsilon_2)$, using (1.7) and Proposition A.2 (ii) in [12] (with $\delta_- = 1$ and $\theta > c$), we can get

$$\Gamma(z) = O(\log N). \tag{2.48}$$

Plugging (2.48) and (2.46) into (2.47) yields

$$\|\mathbf{u}\|_\infty < O\left(\|\mathbf{u}\|_\infty^2 + O\left(\max_i |\zeta_i|\right)\right) < \|\mathbf{u}\|_\infty^2 + (N\eta)^{-\frac{1}{2}},$$

where the second step follows from (2.34). Then (2.38) further implies that $\|\mathbf{u}\|_\infty < (N\eta)^{-\frac{1}{2}}$, which together with (2.44) and (2.34) also implies

$$|\bar{m}(z) - m_{sc}(z)| < (N\eta)^{-\frac{1}{2}}.$$

Hence

$$A_d \leq \|\mathbf{u}\|_\infty + |\bar{m}(z) - m_{sc}(z)| < (N\eta)^{-\frac{1}{2}}.$$

Therefore, we completed the proof of Theorem 1.9.

Proof of Lemma 2.5 At first, recalling the notation defined in (1.3), we denote the Green’s function of $H^{(i)}$ as

$$G^{(i)}(z) := (H^{(i)} - z)^{-1},$$

with a little abuse of notation. For simplicity, we omit the variable z from the notation below. At first, we recall the elementary identity by Schur’s complement, namely,

$$G_{ii} = \left(h_{ii} - z - (\mathbf{h}_i^{(i)})^* G^{(i)} \mathbf{h}_i^{(i)}\right)^{-1}. \tag{2.49}$$

where we used the notation $\mathbf{h}_i^{(i)}$ to denote the i th column of H , with the i th component deleted. Now, we use the identity for $a, b \neq i$ (see Lemma 4.5 in [12] for instance),

$$G_{ab}^{(i)} = G_{ab} - G_{ai} G_{ib} (G_{ii})^{-1} = G_{ab} - G_{ai} G_{ib} \left(h_{ii} - z - (\mathbf{h}_i^{(i)})^* G^{(i)} \mathbf{h}_i^{(i)}\right). \tag{2.50}$$

By using (1.11) and the large deviation estimate for the quadratic form (see Theorem C.1 of [12] for instance), we have

$$\left| (\mathbf{h}_i^{(i)})^* G^{(i)} \mathbf{h}_i^{(i)} - \sum_{a \neq i} \sigma_{ai}^2 \cdot G_{aa}^{(i)} \right| < \sqrt{\frac{1}{M} \max_a |G_{aa}^{(i)}|^2 + \max_{a \neq b} |G_{ab}^{(i)}|^2}, \tag{2.51}$$

which implies that

$$\left| (\mathbf{h}_i^{(i)})^* G^{(i)} \mathbf{h}_i^{(i)} \right| < \max_{a \neq i} |G_{aa}^{(i)}| + \sqrt{\frac{1}{M} \max_{a \neq i} |G_{aa}^{(i)}|^2 + \max_{a \neq b} |G_{ab}^{(i)}|^2} \leq 3 \max_{a, b \neq i} |G_{ab}^{(i)}|, \tag{2.52}$$

where we have used the fact that $\sum_a \sigma_{ai}^2 = 1$ in the first inequality above. Plugging (1.22) and (2.52) into (2.50) and using Corollary 2.2 we obtain

$$\max_{a, b \neq i} |G_{ab}^{(i)}| < 1 + \frac{1}{N\eta} \left(1 + 3 \max_{a, b \neq i} |G_{ab}^{(i)}| \right),$$

which implies

$$\max_{a, b \neq i} |G_{ab}^{(i)}| < 1, \quad \left| (\mathbf{h}_i^{(i)})^* G^{(i)} \mathbf{h}_i^{(i)} \right| < 1. \tag{2.53}$$

In addition, (1.22), (2.50) and (2.53) lead to the fact that

$$\left| G_{ab}(z) - G_{ab}^{(i)}(z) \right| < \frac{1}{N\eta}, \quad \left| G_{ab}^{(i)}(z) \right| < \delta_{ab} + (N\eta)^{-\frac{1}{2}}, \quad \forall a, b \neq i. \tag{2.54}$$

Now, using (2.33), (2.49), (2.51) and (2.54), we can see that

$$|\zeta_i| = \left| -h_{ii} + (\mathbf{h}_i^{(i)})^* G^{(i)} \mathbf{h}_i^{(i)} - \sum_a \sigma_{ai}^2 G_{aa} \right| < (N\eta)^{-\frac{1}{2}}. \tag{2.55}$$

Therefore, we completed the proof of Lemma 2.5. □

2.3 Proof of Theorem 1.12

With Theorem 1.9, we can prove Theorem 1.12 routinely. At first, according to the uniform bound (1.18), we have

$$\max_{a, b} \sup_{z \in \mathbf{D}(N, \kappa, \varepsilon_2)} |G_{ab}(z) - \delta_{ab} m_{sc}(z)| < 1,$$

which implies that

$$\max_a \sup_{z \in \mathbf{D}(N, \kappa, \varepsilon_2)} |G_{aa}(z)| < 1 \tag{2.56}$$

due to the fact that $m_{sc}(z) \sim 1$. Recalling the normalized eigenvector $\mathbf{u}_i = (u_{i1}, \dots, u_{iN})$ corresponding to λ_i , and using the spectral decomposition, we have

$$\max_a \operatorname{Im} G_{aa}(z) = \max_a \sum_{i=1}^N \frac{|u_{ia}|^2 \eta}{|\lambda_i - E|^2 + \eta^2} = \sum_{i=1}^N \frac{\|\mathbf{u}_i\|_\infty^2 \eta}{|\lambda_i - E|^2 + \eta^2}. \tag{2.57}$$

For any $|\lambda_i| \leq \sqrt{2} - \kappa$, we set $E = \lambda_i$ on the r.h.s. of (2.57) and use (2.56) to bound the l.h.s. of it. Then we obtain $\|\mathbf{u}_i\|_\infty^2 / \eta < 1$. Choosing $\eta = N^{-1+\varepsilon_2}$ above and using the fact that ε_2 can be arbitrarily small, we can get (1.19). Hence, we completed the proof of Theorem 1.12.

3 Supersymmetric formalism and integral representation for the Green’s function

In this section, we will represent $\mathbb{E}|G_{ij}(z)|^{2n}$ for the Gaussian case by a superintegral. The final representation is stated in (3.31). We make the convention here, for any real argument in an integral below, its region of the integral is always \mathbb{R} , unless specified otherwise.

3.1 Gaussian integrals and superbosonization formulas

Let $\phi = (\phi_1, \dots, \phi_k)'$ be a vector of complex components, $\psi = (\psi_1, \dots, \psi_k)'$ be a vector of Grassmann components. In addition, let ϕ^* and ψ^* be the conjugate transposes of ϕ and ψ , respectively. We recall the following well-known formulas for Gaussian integrals.

Proposition 3.1 (Gaussian integrals or Wick’s formulas)

- (i) Let A be a $k \times k$ complex matrix with positive-definite Hermitian part, i.e. $\operatorname{Re} A > 0$. Then for any $\ell \in \mathbb{N}$, and $i_1, \dots, i_\ell, j_1, \dots, j_\ell \in \{1, \dots, k\}$, we have

$$\int \prod_{a=1}^k \frac{d\operatorname{Re} \phi_a d\operatorname{Im} \phi_a}{\pi} \exp\{-\phi^* A \phi\} \prod_{b=1}^\ell \bar{\phi}_{i_b} \phi_{j_b} = \frac{1}{\det A} \sum_{\sigma \in \mathbb{P}(\ell)} \prod_{b=1}^\ell (A^{-1})_{j_b, i_{\sigma(b)}}, \tag{3.1}$$

where $\mathbb{P}(\ell)$ is the permutation group of degree ℓ .

- (ii) Let B be any $k \times k$ matrix. Then for any $\ell \in \{0, \dots, k\}$, any ℓ distinct integers i_1, \dots, i_ℓ and another ℓ distinct integers $j_1, \dots, j_\ell \in \{1, \dots, k\}$, we have

$$\int \prod_{a=1}^k d\bar{\psi}_a d\psi_a \exp\{-\psi^* B \psi\} \prod_{b=1}^\ell \bar{\psi}_{i_b} \psi_{j_b} = (-1)^{\ell + \sum_{\alpha=1}^\ell (i_\alpha + j_\alpha)} \det B^{(\mathbf{l}|\mathbf{J})}, \tag{3.2}$$

where $\mathbf{l} = \{i_1, \dots, i_\ell\}$, and $\mathbf{J} = \{j_1, \dots, j_\ell\}$.

Now, we introduce the superbosonization formula for superintegrals. Let $\chi = (\chi_{ij})$ be an $\ell \times r$ matrix with Grassmann entries, $\mathbf{f} = (f_{ij})$ be an $\ell \times r$ matrix with complex entries. In addition, we denote their conjugate transposes by χ^* and \mathbf{f}^* respectively. Let F be a function of the entries of the matrix

$$S(\mathbf{f}, \mathbf{f}^*; \chi, \chi^*) := \begin{pmatrix} \chi^* \chi & \chi^* \mathbf{f} \\ \mathbf{f}^* \chi & \mathbf{f}^* \mathbf{f} \end{pmatrix}.$$

Let $\mathcal{A}(\chi, \chi^*)$ be the Grassmann algebra generated by χ_{ij} 's and $\bar{\chi}_{ij}$'s. Then we can regard F as a function defined on a complex vector space, taking values in $\mathcal{A}(\chi, \chi^*)$. Hence, we can and do view $F(S(\mathbf{f}, \mathbf{f}^*; \chi, \chi^*))$ as a polynomial in χ_{ij} 's and $\bar{\chi}_{ij}$'s, in which the coefficients are functions of f_{ij} 's and \bar{f}_{ij} 's. Under this viewpoint, we state the assumption on F as follows.

Assumption 3.2 Suppose that $F(S(\mathbf{f}, \mathbf{f}^*; \chi, \chi^*))$ is a holomorphic function of f_{ij} 's and \bar{f}_{ij} 's if they are regarded as independent variables, and F is a Schwarz function of $\operatorname{Re} f_{ij}$'s and $\operatorname{Im} f_{ij}$'s, by those we mean that all of the coefficients of $F(S(\mathbf{f}, \mathbf{f}^*; \chi, \chi^*))$, as functions of f_{ij} 's and \bar{f}_{ij} 's, possess the above properties.

Proposition 3.3 (Superbosonization formula for the nonsingular case, [18]) *Suppose that F satisfies Assumption 3.2. For $\ell \geq r$, we have*

$$\int F \begin{pmatrix} \chi^* \chi & \chi^* \mathbf{f} \\ \mathbf{f}^* \chi & \mathbf{f}^* \mathbf{f} \end{pmatrix} d\mathbf{f} d\chi = (\mathbf{i}\pi)^{-r(r-1)} \int d\hat{\mu}(\mathbf{x}) d\hat{\nu}(\mathbf{y}) d\omega d\xi \times F \begin{pmatrix} \mathbf{x} & \omega \\ \xi & \mathbf{y} \end{pmatrix} \frac{\det^\ell \mathbf{y}}{\det^\ell (\mathbf{x} - \omega \mathbf{y}^{-1} \xi)}, \tag{3.3}$$

where $\mathbf{x} = (x_{ij})$ is a unitary matrix; $\mathbf{y} = (y_{ij})$ is a positive-definite Hermitian matrix; ω and ξ are two Grassmann matrices, and all of them are $r \times r$. Here

$$d\mathbf{f} = \prod_{i,j} \frac{d\operatorname{Re} f_{ij} d\operatorname{Im} f_{ij}}{\pi}, \quad d\chi = \prod_{i,j} d\bar{\chi}_{ij} d\chi_{ij},$$

$$d\hat{\nu}(\mathbf{y}) = \mathbf{1}(\mathbf{y} > 0) \prod_{i=1}^r dy_{ii} \prod_{j>k} d\operatorname{Re} y_{jk} d\operatorname{Im} y_{jk}, \quad d\omega d\xi = \prod_{i,j=1}^r d\omega_{ij} d\xi_{ij},$$

and $d\hat{\mu}(\cdot)$ is defined by

$$d\hat{\mu}(\mathbf{x}) = \frac{\pi^{r(r-1)/2}}{\prod_{i=1}^r i!} \cdot \prod_{i=1}^r \frac{dx_i}{2\pi \mathbf{i}} \cdot (\Delta(x_1, \dots, x_r))^2 \cdot d\mu(V),$$

under the parametrization induced by the eigendecomposition, namely,

$$\mathbf{x} = V^* \hat{\mathbf{x}} V, \quad \hat{\mathbf{x}} = \operatorname{diag}(x_1, \dots, x_r), \quad V \in \mathring{U}(r).$$

Here $d\mu(V)$ is the Haar measure on $\mathring{U}(r)$, and $\Delta(\cdot)$ is the Vandermonde determinant. In addition, the integral w.r.t. \mathbf{x} ranges over $U(r)$, that w.r.t. \mathbf{y} ranges over the cone of positive-definite matrices.

For the singular case, i.e. $r > \ell$, we only state the formula for the case of $r = 2$ and $\ell = 1$, which is enough for our purpose. We can refer to formula (11) in [3] for the result under more general setting.

Proposition 3.4 (Superbosonization formula for the singular case, [3]) *Suppose that F satisfies Assumption 3.2. If $r = 2$ and $\ell = 1$, we have*

$$\int F \begin{pmatrix} \boldsymbol{\chi}^* \boldsymbol{\chi} & \boldsymbol{\chi}^* \mathbf{f} \\ \mathbf{f}^* \boldsymbol{\chi} & \mathbf{f}^* \mathbf{f} \end{pmatrix} d\mathbf{f} d\boldsymbol{\chi} = \frac{-1}{\pi^2} \int d\boldsymbol{\omega} d\hat{\mu}(\mathbf{x}) \cdot \mathbf{1}(y \geq 0) dy \cdot d\boldsymbol{\omega} d\xi F \begin{pmatrix} \mathbf{x} & \boldsymbol{\omega} \mathbf{w}^* \\ \mathbf{w} \xi & y \mathbf{w} \mathbf{w}^* \end{pmatrix} \frac{y(y - \xi \mathbf{x}^{-1} \boldsymbol{\omega})^2}{\det^2 \mathbf{x}}, \quad (3.4)$$

where y is a positive variable; \mathbf{x} is a 2-dimensional unitary matrix; $\boldsymbol{\omega} = (\omega_1, \omega_2)'$ and $\xi = (\xi_1, \xi_2)$ are two vectors with Grassmann components. In addition, \mathbf{w} is a unit vector, which can be parameterized by

$$\mathbf{w} = \left(v, \sqrt{1 - v^2} e^{i\theta} \right)', \quad v \in \mathbb{I}, \quad \theta \in \mathbb{L}.$$

Moreover, the differentials are defined as

$$d\boldsymbol{\omega} = v dv d\theta, \quad d\boldsymbol{\omega} d\xi = \prod_{i=1,2} d\omega_i d\xi_i.$$

In addition, the integral w.r.t. \mathbf{x} ranges over $U(2)$.

3.2 Initial representation

For $a = 1, 2$ and $j = 1, \dots, W$, we set

$$\begin{aligned} \Phi_{a,j} &= (\phi_{a,j,1}, \dots, \phi_{a,j,M})', & \Psi_{a,j} &= (\psi_{a,j,1}, \dots, \psi_{a,j,M})' \\ \Phi_a &= (\Phi'_{a,1}, \dots, \Phi'_{a,W})', & \Psi_a &= (\Psi'_{a,1}, \dots, \Psi'_{a,W})'. \end{aligned}$$

For each j and each a , $\Phi_{a,j}$ is a vector with complex components, and $\Psi_{a,j}$ is a vector with Grassmann components. In addition, we use $\Phi_{a,j}^*$ and $\Psi_{a,j}^*$ to represent the conjugate transposes of $\Phi_{a,j}$ and $\Psi_{a,j}$ respectively. Analogously, we adopt the notation Φ_a^* and Ψ_a^* to represent the conjugate transposes of Φ_a and Ψ_a , respectively. We have the following integral representation for the moments of the Green's function.

Lemma 3.5 For any $p, q = 1, \dots, W$ and $\alpha, \beta = 1, \dots, M$, we have

$$\begin{aligned}
 |G_{pq,\alpha\beta}(z)|^{2n} &= \frac{1}{(n!)^2} \int d\Phi d\Psi (\bar{\phi}_{1,q,\beta} \phi_{1,p,\alpha} \bar{\phi}_{2,p,\alpha} \phi_{2,q,\beta})^n \\
 &\quad \times \exp \left\{ \mathbf{i} \Psi_1^*(z - H) \Psi_1 + \mathbf{i} \Phi_1^*(z - H) \Phi_1 \right. \\
 &\quad \left. - \mathbf{i} \Psi_2^*(\bar{z} - H) \Psi_2 - \mathbf{i} \Phi_2^*(\bar{z} - H) \Phi_2 \right\}, \tag{3.5}
 \end{aligned}$$

where

$$d\Phi = \prod_{a=1,2} \prod_{j=1}^W \prod_{\alpha'=1}^M \frac{d\operatorname{Re} \phi_{a,j,\alpha'}}{\pi}, \quad d\Psi = \prod_{a=1,2} \prod_{j=1}^W \prod_{\alpha'=1}^M d\bar{\psi}_{a,j,\alpha'} d\psi_{a,j,\alpha'}.$$

Proof By using Proposition 3.1 (i) with $\ell = n$ and Proposition 3.1 (ii) with $\ell = 0$, we can get (3.5). □

3.3 Averaging over the Gaussian random matrix

Recall the variance profile \tilde{S} in (1.2). Now, we take expectation of the Green’s function, i.e average over the random matrix. By elementary Gaussian integral, we get

$$\begin{aligned}
 \mathbb{E}|G_{pq,\alpha\beta}(z)|^{2n} &= \frac{1}{(n!)^2} \int d\Phi d\Psi (\bar{\phi}_{1,q,\beta} \phi_{1,p,\alpha} \bar{\phi}_{2,p,\alpha} \phi_{2,q,\beta})^n \\
 &\quad \times \exp \left\{ \mathbf{i} \sum_{j=1}^W (Tr \check{X}_j J Z + Tr \check{Y}_j J Z) \right\} \\
 &\quad \times \exp \left\{ \frac{1}{2M} \sum_{j,k} \tilde{s}_{jk} Tr \check{X}_j J \check{X}_k J - \frac{1}{2M} \sum_{j,k} \tilde{s}_{jk} Tr \check{Y}_j J \check{Y}_k J \right\} \\
 &\quad \times \exp \left\{ -\frac{1}{M} \sum_{j,k} \tilde{s}_{jk} Tr \check{\Omega}_j J \check{\Xi}_k J \right\}, \tag{3.6}
 \end{aligned}$$

where $J := \operatorname{diag}(1, -1)$ and $Z := \operatorname{diag}(z, \bar{z})$, and for each $j = 1, \dots, W$, the matrices $\check{X}_j, \check{Y}_j, \check{\Omega}_j$ and $\check{\Xi}_j$ are 2×2 blocks of a supermatrix, namely,

$$\check{S}_j = \left(\begin{array}{c|c} \check{X}_j & \check{\Omega}_j \\ \hline \check{\Xi}_j & \check{Y}_j \end{array} \right) := \left(\begin{array}{cc|cc} \Psi_{1,j}^* \Psi_{1,j} & \Psi_{1,j}^* \Psi_{2,j} & \Psi_{1,j}^* \Phi_{1,j} & \Psi_{1,j}^* \Phi_{2,j} \\ \Psi_{2,j}^* \Psi_{1,j} & \Psi_{2,j}^* \Psi_{2,j} & \Psi_{2,j}^* \Phi_{1,j} & \Psi_{2,j}^* \Phi_{2,j} \\ \hline \Phi_{1,j}^* \Psi_{1,j} & \Phi_{1,j}^* \Psi_{2,j} & \Phi_{1,j}^* \Phi_{1,j} & \Phi_{1,j}^* \Phi_{2,j} \\ \Phi_{2,j}^* \Psi_{1,j} & \Phi_{2,j}^* \Psi_{2,j} & \Phi_{2,j}^* \Phi_{1,j} & \Phi_{2,j}^* \Phi_{2,j} \end{array} \right).$$

Remark 3.6 The derivation of (3.6) from (3.5) is quite standard. We refer to the proof of (2.14) in [20] for more details and will not reproduce it here.

3.4 Decomposition of the supermatrices

From now on, we split the discussion into the following three cases

- (Case 1): Entries in the off-diagonal blocks, i.e. $p \neq q$,
- (Case 2): Off-diagonal entries in the diagonal blocks, i.e. $p = q, \alpha \neq \beta$,
- (Case 3): Diagonal entries, i.e. $p = q, \alpha = \beta$.

For each case, we will perform a decomposition for the supermatrix \check{S}_j ($j = p$ or q). For a vector \mathbf{v} and some index set l , we use $\mathbf{v}^{(l)}$ to denote the subvector obtained by deleting the i th component of \mathbf{v} for all $i \in l$. Then, we adopt the notation

$$\check{S}_j^{(l)} = \left(\begin{array}{c|c} \check{X}_j^{(l)} & \check{\Omega}_j^{(l)} \\ \check{\Xi}_j^{(l)} & \check{Y}_j^{(l)} \end{array} \right), \quad \check{S}_j^{[i]} = \left(\begin{array}{c|c} \check{X}_j^{[i]} & \check{\Omega}_j^{[i]} \\ \check{\Xi}_j^{[i]} & \check{Y}_j^{[i]} \end{array} \right).$$

Here, for $\mathbf{A} = \check{X}_j, \check{Y}_j, \check{\Omega}_j$ or $\check{\Xi}_j$, the notation $\mathbf{A}^{(l)}$ is defined via replacing $\Phi_{a,j}, \Psi_{a,j}, \Phi_{a,j}^*$ and $\Psi_{a,j}^*$ by $\Phi_{a,j}^{(l)}, \Psi_{a,j}^{(l)}, (\Phi_{a,j}^*)^{(l)}$ and $(\Psi_{a,j}^*)^{(l)}$, respectively, for $a = 1, 2$, in the definition of \mathbf{A} . In addition, the notation $\mathbf{A}^{[i]}$ is defined via replacing $\Phi_{a,j}, \Psi_{a,j}, \Phi_{a,j}^*$ and $\Psi_{a,j}^*$ by $\phi_{a,j,i}, \psi_{a,j,i}, \bar{\phi}_{a,j,i}$ and $\bar{\psi}_{a,j,i}$ respectively, for $a = 1, 2$, in the definition of \mathbf{A} . Moreover, for $\mathbf{A} = \check{S}_j, \check{X}_j, \check{Y}_j, \check{\Omega}_j$ or $\check{\Xi}_j$, we will simply abbreviate $\mathbf{A}^{(a,b)}$ and $\mathbf{A}^{(\{a\})}$ by $\mathbf{A}^{(a,b)}$ and $\mathbf{A}^{(a)}$, respectively. Note that $\check{S}_j^{[i]}$ is of rank-one.

Recalling the spatial-orbital parametrization for the rows or columns of H , it is easy to see from the block structure that for any $\alpha, \alpha' \in \{1, \dots, M\}$, exchanging the (j, α) th row with the (j, α') th row and simultaneously exchanging the corresponding columns will not change the distribution of H .

For Case 1, due to the symmetry mentioned above, we can assume $\alpha = \beta = 1$. Then we extract two rank-one supermatrices from \check{S}_p and \check{S}_q such that the quantities $\bar{\phi}_{2,p,1}\phi_{1,p,1}$ and $\bar{\phi}_{1,q,1}\phi_{2,q,1}$ can be expressed in terms of the entries of these supermatrices. More specifically, we decompose the supermatrices

$$\check{S}_p = \check{S}_p^{(1)} + \check{S}_p^{[1]}, \quad \check{S}_q = \check{S}_q^{(1)} + \check{S}_q^{[1]}. \tag{3.7}$$

Consequently, we can write

$$\bar{\phi}_{1,q,1}\phi_{1,p,1}\bar{\phi}_{2,p,1}\phi_{2,q,1} = \left(\check{Y}_q^{[1]}\right)_{12} \left(\check{Y}_p^{[1]}\right)_{21}. \tag{3.8}$$

For Case 2, due to symmetry, we can assume that $\alpha = 1, \beta = 2$. Then we extract two rank-one supermatrices from \check{S}_p , namely,

$$\check{S}_p = \check{S}_p^{(1,2)} + \check{S}_p^{[1]} + \check{S}_p^{[2]}. \tag{3.9}$$

Consequently, we can write

$$\bar{\phi}_{1,p,2}\phi_{1,p,1}\bar{\phi}_{2,p,1}\phi_{2,p,2} = \left(\check{Y}_p^{[2]}\right)_{12} \left(\check{Y}_p^{[1]}\right)_{21}. \tag{3.10}$$

Finally, for Case 3, due to symmetry, we can assume that $\alpha = 1$. Then we extract only one rank-one supermatrix from $\check{\mathcal{S}}_p$, namely,

$$\check{\mathcal{S}}_p = \check{\mathcal{S}}_p^{(1)} + \check{\mathcal{S}}_p^{[1]}. \tag{3.11}$$

Consequently, we can write

$$\bar{\phi}_{1,p,1}\phi_{1,p,1}\bar{\phi}_{2,p,1}\phi_{2,p,1} = \left(\check{Y}_p^{[1]}\right)_{12} \left(\check{Y}_p^{[1]}\right)_{21} = \left(\check{Y}_p^{[1]}\right)_{11} \left(\check{Y}_p^{[1]}\right)_{22}.$$

Since the discussion for all three cases are similar, we will only present the details for Case 1. More specifically, in the remaining part of this section and Sect. 4 to Sect. 10, we will only treat Case 1. In Sect. 11, we will sum up the discussions in the previous sections and explain how to adapt them to Case 2 and Case 3, resulting a final proof of Theorem 1.15.

3.5 Variable reduction by superbosonization formulae

We will work with Case 1. Recall the decomposition (3.7). We use the superbosonization formulae to reduce the number of variables. We shall treat $\check{\mathcal{S}}_k (k \neq p, q)$ and $\check{\mathcal{S}}_j^{(1)} (j = p, q)$ on an equal footing and use the formula (3.3) with $r = 2, \ell = M$ for the former and $r = 2, \ell = M - 1$ for the latter, while we separate the terms $\check{\mathcal{S}}_j^{[1]} (j = p, q)$ and use the formula (3.4). For simplicity, we introduce the notation

$$\tilde{\mathcal{S}}_j = \begin{cases} \check{\mathcal{S}}_j, & \text{if } j \neq p, q, \\ \check{\mathcal{S}}_j^{(1)}, & \text{if } j = p, q. \end{cases} \tag{3.12}$$

Accordingly, we will use $\tilde{X}_j, \tilde{\Omega}_j, \tilde{\mathcal{E}}_j$ and \tilde{Y}_j to denote four blocks of $\tilde{\mathcal{S}}_j$. With this notation, we can rewrite (3.6) with $\alpha = \beta = 1$ as

$$\begin{aligned} \mathbb{E}|G_{pq,11}(z)|^{2n} &= \frac{1}{(n!)^2} \int d\Phi d\Psi (\bar{\phi}_{1,q,1}\phi_{1,p,1}\bar{\phi}_{2,p,1}\phi_{2,q,1})^n \\ &\times \exp \left\{ \mathbf{i} \sum_{j=1}^W (Tr \tilde{X}_j J Z + Tr \tilde{Y}_j J Z) \right\} \\ &\times \exp \left\{ \frac{1}{2M} \sum_{j,k} \tilde{s}_{jk} (Tr \tilde{X}_j J \tilde{X}_k J - Tr \tilde{Y}_j J \tilde{Y}_k J) \right\} \\ &\times \exp \left\{ -\frac{1}{M} \sum_{j,k} \tilde{s}_{jk} Tr \tilde{\Omega}_j J \tilde{\mathcal{E}}_k J \right\} \\ &\times \prod_{k=p,q} \exp \left\{ \mathbf{i} Tr \check{X}_k^{[1]} J Z + \mathbf{i} Tr \check{Y}_k^{[1]} J Z \right\} \end{aligned}$$

$$\begin{aligned}
 & \times \prod_{k=p,q} \exp \left\{ \frac{1}{M} \sum_{j=1}^W \tilde{s}_{jk} \left(Tr \tilde{X}_j J \check{X}_k^{[1]} J - Tr \tilde{Y}_j J \check{Y}_k^{[1]} J \right) \right\} \\
 & \times \prod_{k,\ell=p,q} \exp \left\{ \frac{\tilde{s}_{k\ell}}{2M} \left(Tr \check{X}_k^{[1]} J \check{X}_\ell^{[1]} J - Tr \check{Y}_k^{[1]} J \check{Y}_\ell^{[1]} J \right) \right\} \\
 & \times \prod_{k,\ell=p,q} \exp \left\{ -\frac{\tilde{s}_{k\ell}}{M} Tr \check{\Omega}_k^{[1]} J \check{\Xi}_\ell^{[1]} J \right\} \\
 & \times \prod_{k=p,q} \exp \left\{ -\frac{1}{M} \sum_j \tilde{s}_{jk} \left(Tr \tilde{\Omega}_j J \check{\Xi}_k^{[1]} J + Tr \check{\Omega}_k^{[1]} J \tilde{\Xi}_j J \right) \right\}, \tag{3.13}
 \end{aligned}$$

where the first factor $(\dots)^n$ of integrand is the observable and all other factors constitute a normalized measure, written in a somewhat complicated form according to the decomposition from Sect. 3.4.

Now, we use the superbosonization formulae (3.3) and (3.4) to the supermatrices $\check{S}_k (k \neq p, q)$, $\check{S}_j^{(1)}$ and $\check{S}_j^{[1]} (j = p, q)$ one by one, to change to the reduced variables as

$$\begin{aligned}
 \check{X}_k^{[1]} &\rightarrow X_k^{[1]}, \quad \check{\Omega}_k^{[1]} \rightarrow \omega_k^{[1]} \left(\mathbf{w}_k^{[1]} \right)^*, \quad \check{\Xi}_k^{[1]} \rightarrow \mathbf{w}_k^{[1]} \xi_k^{[1]}, \\
 \check{Y}_k^{[1]} &\rightarrow Y_k^{[1]} := y_k^{[1]} \mathbf{w}_k^{[1]} \left(\mathbf{w}_k^{[1]} \right)^*, \quad k = p, q, \\
 \tilde{X}_j &\rightarrow X_j, \quad \tilde{Y}_j \rightarrow Y_j, \quad \tilde{\Omega}_j \rightarrow \Omega_j, \quad \tilde{\Xi}_j \rightarrow \Xi_j, \quad j = 1, \dots, W. \tag{3.14}
 \end{aligned}$$

Here, for $j = 1, \dots, W$, X_j is a 2×2 unitary matrix; Y_j is a 2×2 positive-definite matrix; $\Omega_j = (\omega_{j,\alpha\beta})$ and $\Xi_j = (\xi_{j,\alpha\beta})$ are 2×2 Grassmann matrices. For $k = p$ or q , $X_k^{[1]}$ is a 2×2 unitary matrix; $y_k^{[1]}$ is a positive variable; $\omega_k^{[1]} = (\omega_{k,1}^{[1]}, \omega_{k,2}^{[1]})'$ is a column vector with Grassmann components; $\xi_k^{[1]} = (\xi_{k,1}^{[1]}, \xi_{k,2}^{[1]})$ is a row vector with Grassmann components. In addition, for $k = p, q$,

$$\mathbf{w}_k^{[1]} = \left(v_k^{[1]}, u_k^{[1]} e^{i\sigma_k^{[1]}} \right)', \quad u_k^{[1]} = \sqrt{1 - \left(v_k^{[1]} \right)^2}, \quad v_k^{[1]} \in \mathbb{I}, \quad \sigma_k^{[1]} \in \mathbb{L}. \tag{3.15}$$

Then by using superbosonization formulae, we arrive at the representation

$$\begin{aligned}
 & \mathbb{E} |G_{pq,11}(z)|^{2n} \\
 & = \frac{(-1)^W}{(n!)^2 \pi^{2W+4}} \int dX^{[1]} d\mathbf{y}^{[1]} d\mathbf{w}^{[1]} d\omega^{[1]} d\xi^{[1]} dX dY d\Omega d\Xi \\
 & \times \left(y_p^{[1]} y_q^{[1]} \left(\mathbf{w}_q^{[1]} \left(\mathbf{w}_q^{[1]} \right)^* \right)_{12} \left(\mathbf{w}_p^{[1]} \left(\mathbf{w}_p^{[1]} \right)^* \right)_{21} \right)^n \\
 & \times \exp \left\{ \mathbf{i} \sum_{j=1}^W \left(Tr X_j J Z + Tr Y_j J Z \right) \right\}
 \end{aligned}$$

$$\begin{aligned}
 & \times \exp \left\{ \frac{1}{2M} \sum_{j,k} \tilde{s}_{jk} \left(\text{Tr} X_j J X_k J - \text{Tr} Y_j J Y_k J \right) \right\} \\
 & \times \exp \left\{ -\frac{1}{M} \sum_{j,k} \tilde{s}_{jk} \text{Tr} \Omega_j J \mathcal{E}_k J \right\} \prod_j \frac{\det^M Y_j}{\det^M (X_j - \Omega_j Y_j^{-1} \mathcal{E}_j)} \\
 & \times \prod_{k=p,q} \frac{\det (X_k - \Omega_k Y_k^{-1} \mathcal{E}_k)}{\det Y_k} \prod_{k=p,q} \exp \left\{ \mathbf{i} \text{Tr} X_k^{[1]} J Z + \mathbf{i} \text{Tr} Y_k^{[1]} J Z \right\} \\
 & \times \prod_{k=p,q} \exp \left\{ \frac{1}{M} \sum_{j=1}^W \tilde{s}_{jk} \left(\text{Tr} X_j J X_k^{[1]} J - \text{Tr} Y_j J Y_k^{[1]} J \right) \right\} \\
 & \times \prod_{k,\ell=p,q} \exp \left\{ \frac{\tilde{s}_{k\ell}}{2M} \left(\text{Tr} X_k^{[1]} J X_\ell^{[1]} J - \text{Tr} Y_k^{[1]} J Y_\ell^{[1]} J \right) \right\} \\
 & \times \prod_{k,\ell=p,q} \exp \left\{ -\frac{\tilde{s}_{k\ell}}{M} \text{Tr} \boldsymbol{\omega}_k^{[1]} (\mathbf{w}_k^{[1]})^* J \mathbf{w}_\ell^{[1]} \boldsymbol{\xi}_\ell^{[1]} J \right\} \\
 & \times \prod_{k=p,q} \exp \left\{ -\frac{1}{M} \sum_{j=1}^W \tilde{s}_{jk} \text{Tr} \Omega_j J \mathbf{w}_k^{[1]} \boldsymbol{\xi}_k^{[1]} J \right\} \\
 & \times \prod_{k=p,q} \exp \left\{ -\frac{1}{M} \sum_{j=1}^W \tilde{s}_{jk} \text{Tr} \boldsymbol{\omega}_k^{[1]} (\mathbf{w}_k^{[1]})^* J \mathcal{E}_j J \right\} \\
 & \times \prod_{k=p,q} \frac{y_k^{[1]} \left(y_k^{[1]} - \boldsymbol{\xi}_k^{[1]} (X_k^{[1]})^{-1} \boldsymbol{\omega}_k^{[1]} \right)^2}{\det^2 (X_k^{[1]})}, \tag{3.16}
 \end{aligned}$$

where we used the notation $\mathbf{y}^{[1]} := (y_p^{[1]}, y_q^{[1]})$, $\mathbf{w}^{[1]} := (\mathbf{w}_p^{[1]}, \mathbf{w}_q^{[1]})$. The differentials are defined by

$$\begin{aligned}
 dX^{[1]} &= \prod_{j=p,q} d\hat{\mu} \left(X_j^{[1]} \right), & d\mathbf{y}^{[1]} &= \prod_{j=p,q} \mathbf{1} \left(y_j^{[1]} > 0 \right) dy_j^{[1]}, \\
 d\mathbf{w}^{[1]} &= \prod_{j=p,q} d\mathbf{w}_j^{[1]} = \prod_{j=p,q} v_j^{[1]} dv_j^{[1]} d\sigma_j^{[1]}, & d\boldsymbol{\omega}^{[1]} d\boldsymbol{\xi}^{[1]} &= \prod_{\alpha=1,2} \prod_{j=p,q} d\omega_{j,\alpha}^{[1]} d\xi_{j,\alpha}^{[1]}, \\
 dX &= \prod_{j=1}^W d\hat{\mu}(X_j), & dY &= \prod_{j=1}^W d\hat{\nu}(Y_j), & d\Omega d\mathcal{E} &= \prod_{\alpha,\beta=1,2} \prod_{j=1}^W d\omega_{j,\alpha\beta} d\xi_{j,\alpha\beta}.
 \end{aligned}$$

Now we change the variables as $X_j J \rightarrow X_j$, $Y_j J \rightarrow B_j$, $\Omega_j J \rightarrow \Omega_j$, $\mathcal{E}_j J \rightarrow \mathcal{E}_j$ and perform the scaling $X_j \rightarrow -M X_j$, $B_j \rightarrow M B_j$, $\Omega_j \rightarrow \sqrt{M} \Omega_j$ and $\mathcal{E}_j \rightarrow$

$\sqrt{M}\mathcal{E}_j$. Consequently, we can write

$$\begin{aligned} \mathbb{E}|G_{pq,11}(z)|^{2n} &= \frac{(-1)^W M^{4W}}{(n!)^2 \pi^{2W+4}} \int dX^{[1]} d\mathbf{y}^{[1]} d\mathbf{w}^{[1]} d\boldsymbol{\omega}^{[1]} d\xi^{[1]} dX dB d\Omega d\mathcal{E} \\ &\quad \times \exp \left\{ -M(K(X) + L(B)) \right\} \\ &\quad \times \mathcal{P}(\Omega, \mathcal{E}, X, B) \cdot \mathcal{Q}(\Omega, \mathcal{E}, \boldsymbol{\omega}^{[1]}, \xi^{[1]}, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}) \\ &\quad \times \mathcal{F}(X, B, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}), \end{aligned} \tag{3.17}$$

where the functions in the integrand are defined as

$$\begin{aligned} K(X) &:= -\frac{1}{2} \sum_{j,k} \tilde{s}_{jk} Tr X_j X_k + \mathbf{i}E \sum_j Tr X_j + \sum_j \log \det X_j, \\ L(B) &:= \frac{1}{2} \sum_{j,k} \tilde{s}_{jk} Tr B_j B_k - \mathbf{i}E \sum_j Tr B_j - \sum_j \log \det B_j, \\ \mathcal{P}(\Omega, \mathcal{E}, X, B) &:= \exp \left\{ -\sum_{j,k} \tilde{s}_{jk} Tr \Omega_j \mathcal{E}_k \right\} \times \prod_j \frac{1}{\det^M \left(1 + \frac{1}{M} X_j^{-1} \Omega_j B_j^{-1} \mathcal{E}_j \right)} \\ &\quad \times \prod_{k=p,q} \frac{\det \left(X_k + \frac{1}{M} \Omega_k B_k^{-1} \mathcal{E}_k \right)}{\det B_k}, \end{aligned} \tag{3.18}$$

$$\begin{aligned} \mathcal{Q}(\Omega, \mathcal{E}, \boldsymbol{\omega}^{[1]}, \xi^{[1]}, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}) &:= \prod_{k=p,q} \left(1 - (y_k^{[1]})^{-1} \xi_k^{[1]} (X_k^{[1]})^{-1} \boldsymbol{\omega}_k^{[1]} \right)^2 \\ &\quad \times \prod_{k=p,q} \exp \left\{ -\frac{1}{\sqrt{M}} \sum_j \tilde{s}_{jk} \left(Tr \Omega_j \mathbf{w}_k^{[1]} \xi_k^{[1]} J + Tr \boldsymbol{\omega}_k^{[1]} (\mathbf{w}_k^{[1]})^* J \mathcal{E}_j \right) \right\} \\ &\quad \times \prod_{k,\ell=p,q} \exp \left\{ -\frac{1}{M} \tilde{s}_{k\ell} Tr \boldsymbol{\omega}_k^{[1]} (\mathbf{w}_k^{[1]})^* J \mathbf{w}_\ell^{[1]} \xi_\ell^{[1]} J \right\} \\ \mathcal{F}(X, B, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}) &:= f(X, X^{[1]}) g(B, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}), \end{aligned} \tag{3.19}$$

with

$$\begin{aligned} f(X, X^{[1]}) &:= \exp \left\{ M\eta \sum_{j=1}^W Tr X_j J \right\} \prod_{k,\ell=p,q} \exp \left\{ \frac{\tilde{s}_{k\ell}}{2M} Tr X_k^{[1]} J X_\ell^{[1]} J \right\} \\ &\quad \times \prod_{k=p,q} \frac{1}{\det^2 \left(X_k^{[1]} \right)} \exp \left\{ \mathbf{i} Tr X_k^{[1]} J Z - \sum_j \tilde{s}_{jk} Tr X_j X_k^{[1]} J \right\}, \end{aligned} \tag{3.20}$$

$$\begin{aligned}
 g(B, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}) &:= \exp \left\{ -M\eta \sum_{j=1}^W \text{Tr} B_j J \right\} \prod_{k, \ell=p, q} \exp \left\{ -\frac{\tilde{\mathfrak{s}}_{k\ell}}{2M} \text{Tr} Y_k^{[1]} J Y_\ell^{[1]} J \right\} \\
 &\times \left(\left(\mathbf{w}_q^{[1]} \left(\mathbf{w}_q^{[1]} \right)^* \right)_{12} \left(\mathbf{w}_p^{[1]} \left(\mathbf{w}_p^{[1]} \right)^* \right)_{21} \right)^n \\
 &\times \prod_{k=p, q} \left(y_k^{[1]} \right)^{n+3} \exp \left\{ i \text{Tr} Y_k^{[1]} J Z - \sum_j \tilde{\mathfrak{s}}_{jk} \text{Tr} B_j Y_k^{[1]} J \right\}.
 \end{aligned} \tag{3.21}$$

In (3.17), the regions of X_j 's and $X_k^{[1]}$'s are all $U(2)$, and those of B_j 's are the set of the matrices A satisfying $AJ > 0$. We remind the reader here that if we parametrize the unitary matrix X_j according to its eigendecomposition, the scaling $X_j \rightarrow -MX_j$ is equivalent to changing the contour of the eigenvalues of X_j from the unit circle Σ to $\frac{1}{M}\Sigma$, up to the orientation. Afterwards, we deformed the contour back to Σ in (3.17). This is possible since the only singularity of the integrand in (3.17) in the variables of the eigenvalues of X_j is at 0, c.f., the matrix X_j^{-1} in the factor $\mathcal{P}(\Omega, \mathcal{E}, X, B)$.

3.6 Parametrization for X, B

Similarly to the discussion in [20], we start with some preliminary parameterization. At first, we do the eigendecomposition

$$X_j = P_j^* \hat{X}_j P_j, \quad B_j = Q_j^{-1} \hat{B}_j Q_j, \quad P_j \in \mathring{U}(2), \quad Q_j \in \mathring{U}(1, 1), \tag{3.22}$$

where

$$\hat{X}_j = \text{diag}(x_{j,1}, x_{j,2}), \quad \hat{B}_j = \text{diag}(b_{j,1}, -b_{j,2}), \quad x_{j,1}, x_{j,2} \in \Sigma, \quad b_{j,1}, b_{j,2} \in \mathbb{R}_+. \tag{3.23}$$

Further, we introduce

$$V_j = P_j P_1^* \in \mathring{U}(2), \quad T_j = Q_j Q_1^{-1} \in \mathring{U}(1, 1), \quad j = 1, \dots, W. \tag{3.24}$$

Especially, we have $V_1 = T_1 = I$. Now, we parameterize P_1, Q_1, V_j and T_j for all $j = 2, \dots, W$ as follows

$$\begin{aligned}
 P_1 &= \begin{pmatrix} u & v e^{i\theta} \\ -v e^{-i\theta} & u \end{pmatrix}, \quad V_j = \begin{pmatrix} u_j & v_j e^{i\theta_j} \\ -v_j e^{-i\theta_j} & u_j \end{pmatrix}, \\
 u &= \sqrt{1 - v^2}, \quad u_j = \sqrt{1 - v_j^2}, \quad v, v_j \in \mathbb{I}, \quad \theta, \theta_j \in \mathbb{L}, \\
 Q_1 &= \begin{pmatrix} s & t e^{i\sigma} \\ t e^{-i\sigma} & s \end{pmatrix}, \quad T_j = \begin{pmatrix} s_j & t_j e^{i\sigma_j} \\ t_j e^{-i\sigma_j} & s_j \end{pmatrix}, \\
 s &= \sqrt{1 + t^2}, \quad s_j = \sqrt{1 + t_j^2}, \quad t, t_j \in \mathbb{R}_+, \quad \sigma, \sigma_j \in \mathbb{L}.
 \end{aligned} \tag{3.25}$$

Under the parametrization above, we can express the corresponding differentials as follows.

$$\begin{aligned}
 dXdB &= d\mu(P_1)d\nu(Q_1) \cdot \prod_{j=2}^W d\mu(V_j)d\nu(T_j) \cdot \prod_{j=1}^W db_{j,1}db_{j,2} \cdot \frac{dx_{j,1}}{2\pi i} \frac{dx_{j,2}}{2\pi i} \\
 &\times 2^W (\pi/2)^{2W} \prod_{j=1}^W (x_{j,1} - x_{j,2})^2 (b_{j,1} + b_{j,2})^2, \tag{3.26}
 \end{aligned}$$

where

$$\begin{aligned}
 d\mu(P_1) &= 2vdv \cdot \frac{d\theta}{2\pi}, & d\mu(V_j) &= 2v_jdv_j \cdot \frac{d\theta_j}{2\pi}, & d\nu(Q_1) &= 2tdt \cdot \frac{d\sigma}{2\pi}, \\
 d\nu(T_j) &= 2t_jdt_j \cdot \frac{d\sigma_j}{2\pi}.
 \end{aligned}$$

In addition, for simplicity, we do the change of variables

$$\Omega_j \rightarrow P_1^* \Omega_j Q_1, \quad \mathcal{E}_j \rightarrow Q_1^{-1} \mathcal{E}_j P_1. \tag{3.27}$$

Note that the Berezinian of such a change is 1. After this change, $\mathcal{P}(\Omega, \mathcal{E}, X, B, \mathbf{y}^{[1]}, \mathbf{w}^{[1]})$ turns out to be independent of P_1 and Q_1 .

To adapt to the new parametrization, we change the notation

$$\begin{aligned}
 K(X) &\rightarrow K(\hat{X}, V), & L(B) &\rightarrow L(\hat{B}, T), & \mathcal{P}(\Omega, \mathcal{E}, X, B) &\rightarrow \mathcal{P}(\Omega, \mathcal{E}, \hat{X}, \hat{B}, V, T), \\
 \mathcal{Q}(\Omega, \mathcal{E}, \omega^{[1]}, \xi^{[1]}, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}) &\rightarrow \mathcal{Q}(\Omega, \mathcal{E}, \omega^{[1]}, \xi^{[1]}, P_1, Q_1, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}), \\
 \mathcal{F}(X, B, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}) &\rightarrow \mathcal{F}(\hat{X}, \hat{B}, V, T, P_1, Q_1, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}), \\
 f(X, X^{[1]}) &\rightarrow f(P_1, V, \hat{X}, X^{[1]}), & g(B, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}) &\rightarrow g(Q_1, T, \hat{B}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}).
 \end{aligned} \tag{3.28}$$

We recall here that $K(X)$ does not depend on P_1 , as well, $L(B)$ does not depend on Q_1 . Moreover, according to the change (3.27), we have

$$\begin{aligned}
 \mathcal{P}(\Omega, \mathcal{E}, \hat{X}, \hat{B}, V, T) &= \exp \left\{ - \sum_{j,k} \tilde{s}_{jk} Tr \Omega_j \mathcal{E}_k \right\} \\
 &\times \prod_j \frac{1}{\det^M (1 + M^{-1} V_j^* \hat{X}_j^{-1} V_j \Omega_j T_j^{-1} \hat{B}_j^{-1} T_j \mathcal{E}_j)} \\
 &\times \prod_{k=p,q} \frac{\det (V_k^* \hat{X}_k V_k + M^{-1} \Omega_k T_k^{-1} \hat{B}_k^{-1} T_k \mathcal{E}_k)}{\det \hat{B}_k} \tag{3.29}
 \end{aligned}$$

and

$$\mathcal{Q}(\Omega, \mathcal{E}, \omega^{[1]}, \xi^{[1]}, P_1, Q_1, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]})$$

$$\begin{aligned}
 &= \prod_{k=p,q} \exp \left\{ -\frac{1}{\sqrt{M}} \sum_j \tilde{s}_{jk} \left(Tr P_1^* \Omega_j Q_1 \mathbf{w}_k^{[1]} \xi_k^{[1]} J + Tr \omega_k^{[1]} (\mathbf{w}_k^{[1]})^* J Q_1^{-1} \Xi_j P_1 \right) \right\} \\
 &\times \prod_{k,\ell=p,q} \exp \left\{ -\frac{1}{M} \tilde{s}_{k\ell} Tr \omega_k^{[1]} (\mathbf{w}_k^{[1]})^* J (\mathbf{w}_\ell^{[1]}) \xi_\ell^{[1]} J \right\} \\
 &\times \prod_{k=p,q} \left(1 - (y_k^{[1]})^{-1} \xi_k^{[1]} (X_k^{[1]})^{-1} \omega_k^{[1]} \right)^2. \tag{3.30}
 \end{aligned}$$

Consequently, using (3.26), from (3.17) we can write

$$\begin{aligned}
 \mathbb{E}|G_{pq,11}(z)|^{2n} &= \frac{M^{4W}}{(n!)^2 8^W \pi^{2W+4}} \int \prod_{j=2}^W d\mu(V_j) dv(T_j) \\
 &\times \int_{\mathbb{R}_+^{2W}} \prod_{j=1}^W db_{j,1} db_{j,2} \oint_{\Sigma^{2W}} \prod_{j=1}^W dx_{j,1} dx_{j,2} \\
 &\times \exp \left\{ -M(K(\hat{X}, V) + L(\hat{B}, T)) \right\} \\
 &\times \prod_{j=1}^W (x_{j,1} - x_{j,2})^2 (b_{j,1} + b_{j,2})^2 \cdot \mathbf{A}(\hat{X}, \hat{B}, V, T). \tag{3.31}
 \end{aligned}$$

where we introduced the notation

$$\begin{aligned}
 \mathbf{A}(\hat{X}, \hat{B}, V, T) &:= \int dX^{[1]} dy^{[1]} d\mathbf{w}^{[1]} d\omega^{[1]} d\xi^{[1]} d\Omega d\Xi d\mu(P_1) dv(Q_1) \\
 &\mathcal{P}(\Omega, \Xi, \hat{X}, \hat{B}, V, T) \mathcal{Q}(\Omega, \Xi, \omega^{[1]}, \xi^{[1]}, P_1, Q_1, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}) \\
 &\times \mathcal{F}(\hat{X}, \hat{B}, V, T, P_1, Q_1, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}). \tag{3.32}
 \end{aligned}$$

In (3.31), the regions of V_j 's are all $\mathring{U}(2)$, and those of T_j 's are all $\mathring{U}(1, 1)$. Observe that all Grassmann variables are inside the integrand of the integral $\mathbf{A}(\hat{X}, \hat{B}, V, T)$. Hence, (3.31) separates the saddle point calculation from the observable $\mathbf{A}(\hat{X}, \hat{B}, V, T)$.

To facilitate the discussions in the remaining part, we introduce some additional terms and notation here. Henceforth, we will employ the notation $(X^{[1]})^{-1} = \{(X_p^{[1]})^{-1}, (X_q^{[1]})^{-1}\}$ and $(\mathbf{y}^{[1]})^{-1} = \{(y_p^{[1]})^{-1}, (y_q^{[1]})^{-1}\}$ for the collection of inverse matrices and reciprocals, respectively. For a matrix or a vector A under discussion, we will use the term *A-variables* to refer to all the variables parametrizing it. For example, \hat{X}_j -variables means $x_{j,1}$ and $x_{j,2}$, and \hat{X} -variables refer to the collection of all \hat{X}_j -variables. Analogously, we can define the terms *T-variables*, $\mathbf{y}^{[1]}$ -variables, Ω -variables and so on. We use another term *A-entries* to refer to the non-zero entries of A . Note that \hat{X}_j -variables are just \hat{X}_j -entries. However, for T_j , they are different, namely,

$$T_j\text{-variables : } t_j, \sigma_j, \quad \text{vs. } T_j\text{-entries : } s_j, t_j e^{i\sigma_j}, t_j e^{-i\sigma_j}.$$

Analogously, we will also use the term T -entries to refer to the collection of all T_j -entries. Then V -entries, $\mathbf{w}^{[1]}$ -entries, etc. are defined in the same manner. It is easy to check that Q_1^{-1} -entries are the same as Q_1 -entries, up to a sign, as well, T_j^{-1} -entries are the same as T_j -entries, for all $j = 2, \dots, W$.

Moreover, to simplify the notation, we make the convention here that we will frequently use a *dot* to represent all the arguments of a function. That means, for instance, we will write $\mathcal{P}(\Omega, \mathcal{E}, \hat{X}, \hat{B}, V, T)$ as $\mathcal{P}(\cdot)$ if there is no confusion. Analogously, we will also use the abbreviation $\mathcal{Q}(\cdot), \mathcal{F}(\cdot), \mathbf{A}(\cdot)$, and so on.

Let $\mathbf{a} := \{a_1, \dots, a_\ell\}$ be a set of variables, we will adopt the notation

$$\Omega(\mathbf{a}; \kappa_1, \kappa_2, \kappa_3)$$

to denote the class of all multivariate polynomials $\mathfrak{p}(\mathbf{a})$ in the arguments a_1, \dots, a_ℓ such that the following three conditions are satisfied: (i) The total number of the monomials in $\mathfrak{p}(\mathbf{a})$ is bounded by κ_1 ; (ii) the coefficients of all monomials in $\mathfrak{p}(\mathbf{a})$ are bounded by κ_2 in magnitude; (iii) the power of each a_i in each monomial is bounded by κ_3 , for all $i = 1, \dots, \ell$. For example, $5b_{j,1}^{-1} + 3b_{j,1}t_j^2 + 1 \in \Omega(\{b_{j,1}^{-1}, b_{j,1}, t_j\}; 3, 5, 2)$. In addition, we define the subset of $\Omega(\mathbf{a}; \kappa_1, \kappa_2, \kappa_3)$, namely,

$$\Omega_{\text{deg}}(\mathbf{a}; \kappa_1, \kappa_2, \kappa_3)$$

consisting of those polynomials in $\Omega(\mathbf{a}; \kappa_1, \kappa_2, \kappa_3)$ such that the degree is bounded by κ_3 , i.e. the total degree of each monomial is bounded by κ_3 . For example $5b_{j,1}^{-1} + 3b_{j,1}t_j^2 + 1 \in \Omega_{\text{deg}}(\{b_{j,1}^{-1}, b_{j,1}, t_j\}; 3, 5, 3)$.

4 Preliminary discussion on the integrand

In this section, we perform a preliminary analysis on the factors of the integrand in (3.17). Recall the matrix \mathcal{J} defined in (1.29).

4.1 Factor $\exp\{-M(K(\hat{X}, V) + L(\hat{B}, T))\}$

Recall the parametrization of $\hat{B}_j, \hat{X}_j, T_j$ and V_j in (3.23) and (3.25), as well as the matrices defined in (1.28). According to the discussion in [20], there are three types of saddle points of this function, namely,

- Type I: For each j , $(\hat{B}_j, T_j, \hat{X}_j) = (D_\pm, I, D_\pm)$ or (D_\pm, I, D_\mp) ,
 $\theta_j \in \mathbb{L}$, $v_j = 0$ if $\hat{X}_j = \hat{X}_1$, and $v_j = 1$ if $\hat{X}_j \neq \hat{X}_1$.
- Type II: For each j , $(\hat{B}_j, T_j, \hat{X}_j) = (D_\pm, I, D_+)$ and $V_j \in \mathring{U}(2)$.
- Type III: For each j , $(\hat{B}_j, T_j, \hat{X}_j) = (D_\pm, I, D_-)$ and $V_j \in \mathring{U}(2)$.

(Actually, since θ_j and v_j vary on continuous sets, it would be more appropriate to use the term *saddle manifolds*.) We will see that the main contribution to the integral (3.17) comes from some small vicinities of the Type I saddle points. At first, by the definition in (3.24), we have $V_1 = I$. If we regard θ_j 's in the parametrization of V_j 's

as fixed parameters, it is easy to see that there are totally 2^W choices of Type I saddle points. Furthermore, the contributions from all the Type I saddle points are the same, since one can always do the transform $V_j \rightarrow \mathfrak{I}V_j$ or $(\hat{X}_j, P_j) \rightarrow (\mathfrak{I}\hat{X}_j, \mathfrak{I}P_j)$ for several j to change one saddle to another. That means, for Type I saddle points, it suffices to consider

- Type I' : For each j , $(\hat{B}_j, T_j, \hat{X}_j, V_j) = (D_{\pm}, I, D_{\pm}, I)$.

Therefore, the total contribution to the integral (3.17) from all Type I saddle points is 2^W times that from the Type I' saddle point.

Following the discussion in [20], we will show in Sect. 5 that both $K(\hat{X}, V) - K(D_{\pm}, I)$ and $L(\hat{B}, T) - L(D_{\pm}, I)$ have positive real parts, bounded by some positive quadratic forms from below, which allows us to perform the saddle point analysis. In addition, it will be seen that in a vicinity of the Type I' saddle point, $\exp\{-M(K(\hat{X}, V) + L(\hat{B}, T))\}$ is approximately Gaussian.

4.2 Factor $\mathcal{Q}(\Omega, \mathcal{E}, \omega^{[1]}, \xi^{[1]}, P_1, Q_1, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]})$

The function $\mathcal{Q}(\cdot)$ contains both the Ω, \mathcal{E} -variables from $\mathcal{P}(\cdot)$, and the $P_1, Q_1, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}$ -variables from $\mathcal{F}(\cdot)$. In addition, note that in the integrand in (3.17), $\mathcal{Q}(\cdot)$ is the only factor containing the $\omega^{[1]}$ and $\xi^{[1]}$ - variables. Hence, we can compute the integral

$$\begin{aligned} &\mathcal{Q}(\Omega, \mathcal{E}, P_1, Q_1, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}) \\ &:= \int d\omega^{[1]} d\xi^{[1]} \mathcal{Q}(\Omega, \mathcal{E}, \omega^{[1]}, \xi^{[1]}, P_1, Q_1, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}) \end{aligned} \tag{4.1}$$

At first, the explicit formula for $\mathcal{Q}(\cdot)$ is complicated and irrelevant for us. From (3.30) and the definition of the Grassmann integral, it is not difficult to see that $\mathcal{Q}(\cdot)$ is a polynomial of the $(X^{[1]})^{-1}, (\mathbf{y}^{[1]})^{-1}, \mathbf{w}^{[1]}, P_1, Q_1, \Omega$ and \mathcal{E} -entries. In principle, for each monomial in the polynomial $\mathcal{Q}(\cdot)$, we can combine the Grassmann variables with $\mathcal{P}(\cdot)$, then perform the integral over Ω and \mathcal{E} -variables, whilst we combine the complex variables with $\mathcal{F}(\cdot)$, and perform the integral over $X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}, P_1$ and Q_1 -variables. A formal discussion on $\mathcal{Q}(\cdot)$ will be given in Sect. 6.1. However, the terms from $\mathcal{Q}(\cdot)$ turn out to be irrelevant in our proof. Therefore, in the arguments with $\mathcal{Q}(\cdot)$ involved, a typical strategy that we will adopt is as follows: we usually neglect $\mathcal{Q}(\cdot)$ at first, and perform the discussion on $\mathcal{P}(\cdot)$ and $\mathcal{F}(\cdot)$ separately, at the end, we make necessary comments on how to slightly modify the discussions to take $\mathcal{Q}(\cdot)$ into account.

4.3 Factor $\mathcal{P}(\Omega, \mathcal{E}, \hat{X}, \hat{B}, V, T)$

We will mainly regard $\mathcal{P}(\cdot)$ as a function of the Ω and \mathcal{E} -variables. As mentioned above, we also have some Ω and \mathcal{E} -variables from the irrelevant term $\mathcal{Q}(\cdot)$. But we temporarily ignore them and regard as if the integral over Ω and \mathcal{E} -variables reads

$$\mathcal{P}(\hat{X}, \hat{B}, V, T) := \int d\Omega d\mathcal{E} \mathcal{P}(\Omega, \mathcal{E}, \hat{X}, \hat{B}, V, T). \tag{4.2}$$

We shall estimate $P(\cdot)$ in three different regions: (1) the complement of the vicinities of the saddle points; (2) the vicinity of Type I saddle point; (3) the vicinities of Type II and III saddle points, which will be done in Sects. 6.2, 9.1 and 10.1, respectively. (Definition 5.5 gives the precise definition of the vicinities.) In each case we will decompose the function $\mathcal{P}(\cdot)$ as a product of a Gaussian measure and a multivariate polynomial of Grassmann variables. Consequently, we can employ (3.2) to perform the integral of this polynomial against the Gaussian measure, whereby $P(\cdot)$ can be estimated.

4.4 Factor $\mathcal{F}(\hat{X}, \hat{B}, V, T, P_1, Q_1, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]})$

Observe that \mathcal{F} is the only term containing the energy scale η . As in the previous discussion of $\mathcal{P}(\cdot)$, here we also ignore the $P_1, Q_1, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}$ -variables from the irrelevant term $Q(\cdot)$ temporarily, and investigate the integral

$$\begin{aligned} F(\hat{X}, \hat{B}, V, T) &= \int dX^{[1]} d\mathbf{y}^{[1]} d\mathbf{w}^{[1]} d\mu(P_1) d\nu(Q_1) \\ &\quad \times \mathcal{F}(\hat{X}, \hat{B}, V, T, P_1, Q_1, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}) \\ &= \int dX^{[1]} d\mu(P_1) f(\hat{X}, V, P_1) \\ &\quad \times \int d\mathbf{y}^{[1]} d\mathbf{w}^{[1]} d\nu(Q_1) g(\hat{B}, T, Q_1, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}). \end{aligned} \tag{4.3}$$

We shall also estimate $F(\cdot)$ in three different regions: (1) the complement of the vicinities of the saddle points; (2) the vicinity of Type I saddle point; (3) the vicinities of Type II and III saddle points, which will be done in Sects. 6.3, 9.2 and 10.2, respectively.

Especially, when we restrict the \hat{X}, \hat{B}, V and T -variables to the vicinity of the Type I saddle points, the above integral can be performed approximately, resulting our main term, a factor of order $1/(N\eta)^{n+2}$. This step will be done in Sect. 9. It is instructive to give a heuristic sketch of this calculation. At first, we plug the Type I saddle points into (4.3). We will show that the integral of $f(\cdot)$ approximately reads

$$e^{-(a_+ - a_-)N\eta} \int dX^{[1]} d\mu(P_1) f(D_{\pm}, I, P_1) \sim \frac{1}{N\eta},$$

which is the easy part. Then, recalling the definition of $g(\cdot)$ in (3.21) and the parameterization (3.15), we will show that the integral of $g(\cdot)$ approximately reads

$$\begin{aligned} &e^{(a_+ - a_-)N\eta} \int d\mathbf{y}^{[1]} d\mathbf{w}^{[1]} d\nu(Q_1) g(D_{\pm}, I, Q_1, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}) \\ &\sim \int_0^\infty 2t dt \int_{\mathbb{L}^2} d\sigma_p^{[1]} d\sigma_q^{[1]} e^{i\text{ns}_p^{[1]}} e^{-i\text{ns}_q^{[1]}} \cdot e^{-cN\eta t^2 + c_1 e^{-i\sigma_p^{[1]}} t + c_2 e^{i\sigma_q^{[1]}} t} \\ &\sim \int_0^\infty 2t dt \cdot t^{2n} \cdot e^{-cN\eta t^2} \sim \frac{1}{(N\eta)^{n+1}}, \end{aligned} \tag{4.4}$$

where in the second step above we used the fact that

$$\int_{\mathbb{L}} d\sigma \cdot e^{i\sigma} e^{ce^{-i\sigma}t} \sim t^n.$$

We notice that the factor $e^{i\sigma_p^{[1]}} e^{-i\sigma_q^{[1]}}$ in (4.4) actually comes from the term $\left((\mathbf{w}_q^{[1]}(\mathbf{w}_q^{[1]})^*)_{12} (\mathbf{w}_p^{[1]}(\mathbf{w}_p^{[1]})^*)_{21} \right)^n$ in (3.21). This factor brings a strong oscillation to the integrand in the integral (4.4). In Case 2, an analogous factor will appear, resulting the same estimate as (4.4). However, in Case 3, such an oscillating factor is absent, then the estimate for the counterpart of the integral in (4.4) is of order $1/N\eta$ instead of $1/(N\eta)^{n+1}$. The detailed analysis will be presented in Sects. 10 and 11.

5 Saddle points and vicinities

In this section, we study the saddle points of $K(\hat{X}, V)$ and $L(\hat{B}, T)$ and deform the contours of the \hat{B} -variables to pass through the saddle points. Then we introduce and classify some small vicinities of these saddle points. The derivation of the saddle points of $K(\hat{X}, V)$ and $L(\hat{B}, T)$ in Sects. 5.1 and 5.2 below is essentially the same as the counterpart in [20], the only difference is that we are working under a more general setting on S . Hence, in Sects. 5.1 and 5.2, we just sketch the discussion, list the results, and make necessary modifications to adapt to our setting. In the sequel, we employ the notation

$$\begin{aligned} \mathbf{b}_a &:= (b_{1,a}, \dots, b_{W,a}), & \mathbf{x}_a &:= (x_{1,a}, \dots, x_{W,a}), & a &= 1, 2, \\ \mathbf{t} &:= (t_2, \dots, t_W), & \mathbf{v} &:= (v_2, \dots, v_W), \\ \boldsymbol{\sigma} &:= (\sigma_2, \dots, \sigma_W), & \boldsymbol{\theta} &:= (\theta_2, \dots, \theta_W). \end{aligned} \tag{5.1}$$

As mentioned above, later we also need to deform the contours, and discuss the integral over some vicinities of the saddle points, thus it is convenient to introduce a notation for the integral over specific domains. To this end, for $a = 1, 2$, we use \mathbf{I}_a^b and \mathbf{I}_a^x to denote generic domains of \mathbf{b}_a and \mathbf{x}_a respectively. Analogously, we use \mathbf{I}^t and \mathbf{I}^v to represent generic domains of \mathbf{t} and \mathbf{v} , respectively. These domains will be specified later. Now, for some collection of domains, we introduce the notation

$$\begin{aligned} \mathcal{I}(\mathbf{I}_1^b, \mathbf{I}_2^b, \mathbf{I}_1^x, \mathbf{I}_2^x, \mathbf{I}^t, \mathbf{I}^v) &:= \frac{M^{4W}}{(n!)^2 8^W \pi^{2W+4}} \int_{\mathbb{L}^{2W-2}} \prod_{j=2}^W \frac{d\theta_j}{2\pi} \prod_{j=2}^W \frac{d\sigma_j}{2\pi} \\ &\times \int_{\mathbf{I}_1^b} \prod_{j=1}^W db_{j,1} \int_{\mathbf{I}_2^b} \prod_{j=1}^W db_{j,2} \int_{\mathbf{I}_1^x} \prod_{j=1}^W dx_{j,1} \\ &\times \int_{\mathbf{I}_2^x} \prod_{j=1}^W dx_{j,2} \int_{\mathbf{I}^t} \prod_{j=2}^W 2t_j dt_j \int_{\mathbf{I}^v} \prod_{j=2}^W 2v_j dv_j \end{aligned}$$

$$\begin{aligned} & \times \exp \left\{ -M(K(\hat{X}, V) + L(\hat{B}, T)) \right\} \\ & \times \prod_{j=1}^W (x_{j,1} - x_{j,2})^2 (b_{j,1} + b_{j,2})^2 \cdot \mathbf{A}(\hat{X}, \hat{B}, V, T). \end{aligned} \tag{5.2}$$

For example, we can write (3.31) as

$$\mathbb{E}|G_{pq,11}(z)|^{2n} = \mathcal{I}(\mathbb{R}_+^W, \mathbb{R}_+^W, \Sigma^W, \Sigma^W, \mathbb{R}_+^{W-1}, \mathbb{I}^{W-1}), \tag{5.3}$$

which is the integral over the full domain.

5.1 Saddle points of $L(\hat{B}, T)$

We introduce the function

$$\mathbb{k}(a) := \frac{a^2}{2} - \mathbf{i}Ea - \log a, \quad a \in \mathbb{C}. \tag{5.4}$$

Recalling the definition of $L(\cdot)$ in (3.18), the decomposition of B_j 's in (3.22) and the definition of T_j 's in (3.24), we can write

$$L(\hat{B}, T) =: \ell(\mathbf{b}_1) + \ell(-\mathbf{b}_2) + \ell_S(\hat{B}, T), \tag{5.5}$$

where we used the notation introduced in (5.1), and the functions $\ell(\cdot)$ and $\ell_S(\cdot)$ are defined as

$$\begin{aligned} \ell(\mathbf{a}) & := -\frac{1}{4} \sum_{j,k} \mathfrak{s}_{jk} (a_j - a_k)^2 + \sum_j \mathbb{k}(a_j), \quad \mathbf{a} = (a_1, \dots, a_W) \in \mathbb{C}^W, \\ \ell_S(\hat{B}, T) & := \frac{1}{2} \sum_{j,k} \mathfrak{s}_{jk} |(T_k T_j^{-1})_{12}|^2 (b_{j,1} + b_{j,2})(b_{k,1} + b_{k,2}). \end{aligned} \tag{5.6}$$

Following the discussion in [20] with slight modification (see Section 3 therein), we see that for $|E| \leq \sqrt{2} - \kappa$, the saddle point of $L(\hat{B}, T)$ is

$$(\hat{B}_j, T_j) = (D_{\pm}, I), \quad \forall j = 1, \dots, W, \tag{5.7}$$

where D_{\pm} is defined in (1.28). For simplicity, we will write (5.7) as $(\hat{B}, T) = (D_{\pm}, I)$ in the sequel. Observe that

$$L(D_{\pm}, I) = \ell(a_+) + \ell(a_-), \quad \ell_S(D_{\pm}, I) = 0. \tag{5.8}$$

We introduce the notation

$$\mathring{\ell}_{++}(\mathbf{a}) := \ell(\mathbf{a}) - \ell(a_+), \quad \mathring{\ell}_{+-}(\mathbf{a}) := \ell(\mathbf{a}) - \ell(a_-), \quad \mathring{\ell}_{--}(\mathbf{a}) := \ell(-\mathbf{a}) - \ell(a_-),$$

$$(5.9)$$

where $\ell(a_+)$ represents the value of $\ell(\mathbf{a})$ at the point $\mathbf{a} = (a_+, \dots, a_+)$, and $\ell(a_-)$ is defined analogously. Correspondingly, we adopt the notation

$$\mathring{L}(\hat{B}, T) := L(\hat{B}, T) - L(D_{\pm}, I) = \mathring{\ell}_{++}(\mathbf{b}_1) + \mathring{\ell}_{--}(\mathbf{b}_2) + \ell_S(\hat{B}, T), \quad (5.10)$$

where the second step is implied by (5.5), (5.8) and (5.9). Now, for each $j = 1, \dots, W$, we deform the contours of $b_{j,1}$ and $b_{j,2}$ to

$$b_{j,1} \in \Gamma := \{ra_+ | r \in \mathbb{R}_+\}, \quad b_{j,2} \in \bar{\Gamma} = \{-ra_- | r \in \mathbb{R}_+\} \quad (5.11)$$

to pass through the saddle points of \hat{B} -variables, based on the following lemma which will be proved in Sect. 7.

Lemma 5.1 *With the notation introduced in (5.2), we have*

$$\begin{aligned} \mathcal{I}(\Gamma^W, \bar{\Gamma}^W, \Sigma^W, \Sigma^W, \mathbb{R}_+^{W-1}, \mathbb{I}^{W-1}) &= \mathcal{I}(\mathbb{R}_+^W, \mathbb{R}_+^W, \Sigma^W, \Sigma^W, \mathbb{R}_+^{W-1}, \mathbb{I}^{W-1}) \\ &= \mathbb{E}|G_{pq,11}(z)|^{2n}. \end{aligned}$$

We introduce the notation

$$r_{j,1} = |b_{j,1}|, \quad r_{j,2} = |b_{j,2}|, \quad j = 1, \dots, W. \quad (5.12)$$

Along the new contours, we have the following lemma.

Lemma 5.2 *Suppose that $|E| \leq \sqrt{2} - \kappa$. Let $\mathbf{b}_1 \in \Gamma^W$, $\mathbf{b}_2 \in \bar{\Gamma}^W$. We have*

$$\begin{aligned} \operatorname{Re} \mathring{L}(\hat{B}, T) &\geq c \sum_{a=1,2} \sum_{j=1}^W ((r_{j,a} - 1)^2 + (r_{j,a} - \log r_{j,a} - 1)) + \operatorname{Re} \ell_S(\hat{B}, T) \\ &\geq c \sum_{a=1,2} \sum_{j=1}^W (r_{j,a} - 1)^2 \end{aligned} \quad (5.13)$$

for some positive constant c .

Proof Since $|E| \leq \sqrt{2} - \kappa$, we have $\operatorname{Re}(b_{j,1} + b_{j,2})(b_{k,1} + b_{k,2}) \geq 0$ for $\mathbf{b}_1 \in \Gamma^W$ and $\mathbf{b}_2 \in \bar{\Gamma}^W$, thus $\operatorname{Re} \ell_S(\hat{B}, T) \geq 0$, in light of the definition in (5.6). Consequently, according to (5.10), it suffices to prove

$$\operatorname{Re} \mathring{\ell}_{++}(\mathbf{b}_1) + \operatorname{Re} \mathring{\ell}_{--}(\mathbf{b}_2) \geq c \sum_{a=1,2} \sum_{j=1}^W ((r_{j,a} - 1)^2 + (r_{j,a} - \log r_{j,a} - 1)) \quad (5.14)$$

for some positive constant c . To see this, we observe the following identities obtained via elementary calculation,

$$\begin{aligned} \operatorname{Re} \mathring{\ell}_{++}(\mathbf{b}_1) &= \frac{E^2 - 2}{4} \left(\frac{1}{2} \sum_{j,k} \mathfrak{s}_{jk} (r_{j,1} - r_{k,1})^2 - \sum_j (r_{j,1} - 1)^2 \right) \\ &\quad + \sum_{j=1}^W (r_{j,1} - \log r_{j,1} - 1) \end{aligned} \tag{5.15}$$

which together with $|E| \leq \sqrt{2} - \kappa$ and (1.5) implies (5.14) immediately. The same identity holds if we replace $\mathring{\ell}_{++}(\mathbf{b}_1)$ and $r_{j,1}$'s by $\mathring{\ell}_{--}(\mathbf{b}_2)$ and $r_{j,2}$'s in (5.15), respectively. Hence, we completed the proof of Lemma 5.2. \square

5.2 Saddle points of $K(\hat{X}, V)$

Analogously, recalling the definition in (5.6), we can write

$$K(\hat{X}, V) =: -\ell(\mathbf{x}_1) - \ell(\mathbf{x}_2) + \ell_S(\hat{X}, V), \tag{5.16}$$

where $\ell(\cdot)$ is defined in the first line of (5.6) and $\ell_S(\hat{X}, V)$ is defined as

$$\ell_S(\hat{X}, V) = \frac{1}{2} \sum_{j,k} \mathfrak{s}_{jk} |(V_k V_j^*)_{12}|^2 (x_{j,1} - x_{j,2})(x_{k,1} - x_{k,2}). \tag{5.17}$$

Analogously to the notation $L(D_{\pm}, I)$, we will use $K(D_{\pm}, I)$ to represent the value of $K(\hat{X}, V)$ at $(\hat{X}_j, V_j) = (D_{\pm}, I)$ for all $j = 1, \dots, W$. In addition, $K(D_+, I)$ and $K(D_-, I)$ are defined in the same manner. Observing that

$$\ell_S(D_{\pm}, I) = \ell_S(D_+, I) = \ell_S(D_-, I) = 0, \tag{5.18}$$

we have

$$K(D_{\pm}, I) = -\ell(a_+) - \ell(a_-), \quad K(D_+, I) = -2\ell(a_+), \quad K(D_-, I) = -2\ell(a_-). \tag{5.19}$$

Moreover, we employ the notation

$$\mathring{K}(\hat{X}, V) = K(\hat{X}, V) - K(D_{\pm}, I) = -\mathring{\ell}_{++}(\mathbf{x}_1) - \mathring{\ell}_{+-}(\mathbf{x}_2) + \ell_S(\hat{X}, V). \tag{5.20}$$

We will need the following elementary observations that are easy to check from (5.19) and (5.6)

$$K(D_{\pm}, I) + L(D_{\pm}, I) = 0, \quad \operatorname{Re} K(D_+, I) = \operatorname{Re} K(D_-, I) = \operatorname{Re} K(D_{\pm}, I). \tag{5.21}$$

In addition, we introduce the $W \times W$ matrix

$$S^v = (s_{jk}^v), \quad s_{jk}^v := s_{jk} |(V_k V_j^*)_{12}|^2, \tag{5.22}$$

and the $2W \times 2W$ matrices

$$\mathbb{S} = S \oplus S, \quad \mathbb{S}^v := \mathbb{S} + \begin{pmatrix} -S^v & S^v \\ S^v & -S^v \end{pmatrix}, \tag{5.23}$$

where \mathbb{S}^v depends on the V -variables according to (5.22). Here we regard V -variables as fixed parameters. Due to the fact $|(V_k V_j^*)_{12}| \in \mathbb{I}$, it is easy to see that \mathbb{S}^v is a weighted Laplacian of a graph with $2W$ vertices. In particular, $\mathbb{S}^v \leq 0$. By the definition (5.22), one can see that $S_{ii}^v = 0$ for all $i = 1, \dots, W$. Consequently, we can obtain

$$\sum_{k \neq j} \mathbb{S}_{jk}^v = \sum_{k \neq j} \mathbb{S}_{kj}^v = -\mathbb{S}_{jj}^v = \begin{cases} -s_{jj}, & \text{if } j = 1, \dots, W \\ -s_{j-W, j-W}, & \text{if } j = W + 1, \dots, 2W. \end{cases}$$

Similarly to (1.5), we get

$$I + \mathbb{S}^v \geq c_0 I, \tag{5.24}$$

where c_0 is the constant in Assumption 1.1 (ii). Moreover, it is not difficult to see from the definitions in (5.17), (5.22) and (5.23) that

$$\frac{1}{4} \sum_{j,k} s_{jk} \text{Tr}(\hat{X}_j - \hat{X}_k)^2 + \ell_S(\hat{X}, V) = -\frac{1}{2} \mathbf{x}' \mathbb{S}^v \mathbf{x}, \tag{5.25}$$

where we used the notation $\mathbf{x} := (\mathbf{x}'_1, \mathbf{x}'_2)'$. Now let

$$\vartheta_j = \arg x_{j,1}, \quad \vartheta_{W+j} = \arg x_{j,2}, \quad \forall j = 1, \dots, W. \tag{5.26}$$

Then, recalling the parametrization of V_j 's in (3.25), we have the following lemma.

Lemma 5.3 *Assume that $x_{j,1}, x_{j,2} \in \Sigma$ for all $j = 1, \dots, W$. We have*

$$\text{Re} \hat{K}(\hat{X}, V) \geq \frac{1}{4} \sum_{j,k=1}^{2W} (\mathbb{S}^v)_{jk} (\cos \vartheta_j - \cos \vartheta_k)^2 + c \sum_{j=1}^{2W} \left(\sin \vartheta_j - \frac{E}{2} \right)^2 \tag{5.27}$$

for some positive constant c . In addition, $\text{Re} \hat{K}(\hat{X}, V)$ attains its minimum 0 at the following three types of saddle points

- Type I: For each j , $\hat{X}_j = D_{\pm}$ or D_{\mp} , $\theta_j \in \mathbb{L}$ $v_j = 0$ if $\hat{X}_j = \hat{X}_1$, and $v_j = 1$ if $\hat{X}_j \neq \hat{X}_1$,
- Type II: For each j , $\hat{X}_j = D_+$, $V_j \in \dot{U}(2)$,
- Type III: For each j , $\hat{X}_j = D_-$, $V_j \in \dot{U}(2)$,

which are the restrictions of three types of saddle points in Sect. 4.1, on \hat{X} and V -variables.

Remark 5.4 The Type I saddle points of (\hat{X}, V) are exactly those points satisfying

$$V_j^* \hat{X}_j V_j = D_{\pm}, \quad \forall j = 1, \dots, W, \quad \text{or} \quad V_j^* \hat{X}_j V_j = D_{\mp}, \quad \forall j = 1, \dots, W.$$

In Lemma 5.3, we wrote them in terms of \hat{X}_j, v_j and θ_j in order to evoke the parameterization in (3.23) and (3.25).

Proof By (5.16), (5.25), the definitions of the functions $\ell(\cdot)$ in (5.7) and $\mathbb{k}(\cdot)$ in (5.4), we can write

$$\mathring{K}(\hat{X}, V) = -\frac{1}{2} \mathbf{x}^* \mathbb{S}^v \mathbf{x} - \sum_{j=1}^W \left(\mathbb{k}(x_{j,1}) + \mathbb{k}(x_{j,2}) - \mathbb{k}(a_+) - \mathbb{k}(a_-) \right).$$

By using (5.26) and the fact $|x_{j,a}| = 1$ for all $j = 1, \dots, W$ and $a = 1, 2$, we can obtain via elementary calculation

$$\begin{aligned} \operatorname{Re} \mathring{K}(\hat{X}, V) &= \frac{1}{4} \sum_{j,k=1}^{2W} (\mathbb{S}^v)_{jk} \left((\cos \vartheta_j - \cos \vartheta_k)^2 - (\sin \vartheta_j - \sin \vartheta_k)^2 \right) \\ &+ \sum_{j=1}^{2W} \left(\sin \vartheta_j - \frac{E}{2} \right)^2. \end{aligned} \tag{5.28}$$

In light of the fact $\mathbb{S}^v \leq 0$ and (5.24), we have

$$I + \frac{1}{2} \mathbb{S}^v \geq I + \mathbb{S}^v \geq c_0 I. \tag{5.29}$$

Applying (5.29) to the r.h.s. of (5.28) yields (5.27).

Then what remains is to show that $\operatorname{Re} \mathring{K}(\hat{X}, V)$ attains its minimum 0 at three types of points listed in Lemma 5.3. This step can be verified in the same manner as the counterpart in [20]. Hence, we just omit it here and refer to the proof of Lemma 2 in [20] for more details. Therefore, we completed the proof of Lemma 5.3. \square

5.3 Vicinities of the saddle points

Having studied the saddle points of $L(\hat{B}, T)$ and $K(\hat{X}, V)$, we then introduce some small vicinities of them. To this end, we introduce the quantity

$$\Theta \equiv \Theta(N, \varepsilon_0) := WN^{\varepsilon_0} \tag{5.30}$$

for small positive constant ε_0 which will be chosen later. Let $\mathbf{a} = (a_1, \dots, a_W) \in \mathbb{C}^W$ be any complex vector. In the sequel, we adopt the notation for any $d \in \mathbb{C}$,

$$\begin{aligned} \mathbf{a} + d &:= (a_1 + d, \dots, a_W + d), \quad d\mathbf{a} := (da_1, \dots, da_W), \\ \arg(\mathbf{a}) &:= (\arg(a_1), \dots, \arg(a_W)). \end{aligned} \tag{5.31}$$

Now, we define the domains $\Upsilon_+^b, \Upsilon_-^b, \Upsilon_+^x, \Upsilon_-^x$ and Υ_S as follows

$$\begin{aligned} \Upsilon_{\pm}^b &\equiv \Upsilon_{\pm}^b(N, \varepsilon_0) := \left\{ \mathbf{a} \in \Gamma^W : \|\mathbf{a} - a_{\pm}\|_2^2 \leq \frac{\Theta}{M} \right\}, \\ \Upsilon_{\pm}^x &\equiv \Upsilon_{\pm}^x(N, \varepsilon_0) := \left\{ \mathbf{a} \in \Sigma^W : \|\arg(a_{\pm}^{-1}\mathbf{a})\|_2^2 \leq \frac{\Theta}{M} \right\}, \\ \Upsilon_S &\equiv \Upsilon_S(N, \varepsilon_0) := \left\{ \mathbf{a} \in \mathbb{R}_+^{W-1} : -\mathbf{a}'S^{(1)}\mathbf{a} \leq \frac{\Theta}{M} \right\}, \end{aligned} \tag{5.32}$$

where the superscripts b and x indicate that these will be domains of the corresponding variables. In order to define the vicinities of the Type I saddle points properly, we introduce the permutation ϵ_j of $\{1, 2\}$, for each triple $(x_{j,1}, x_{j,2}, v_j)$. Specifically, recalling the fact of $u_j = \sqrt{1 - v_j^2}$ from (3.25), we define

$$v_{j,\epsilon_j} \equiv v_{j,\epsilon_j}(\epsilon_1) := v_j \mathbf{1}(\epsilon_j = \epsilon_1) + u_j \mathbf{1}(\epsilon_j \neq \epsilon_1).$$

Denoting by $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_W)$ and $\boldsymbol{\epsilon}(a) = (\epsilon_1(a), \dots, \epsilon_W(a))$ for $a = 1, 2$, we set

$$\mathbf{x}_{\boldsymbol{\epsilon}(a)} = (x_{1,\epsilon_1(a)}, \dots, x_{W,\epsilon_W(a)}), \quad a = 1, 2, \quad \mathbf{v}_{\boldsymbol{\epsilon}} = (v_{2,\epsilon_2}, \dots, v_{W,\epsilon_W}). \tag{5.33}$$

With this notation, we now define the Type I, II, and III vicinities, parameterized by $(\mathbf{b}_1, \mathbf{b}_2, \mathbf{x}_1, \mathbf{x}_2, \mathbf{t}, \mathbf{v})$ of the corresponding saddle point types. We also define the special case of the Type I vicinity, namely, Type I' vicinity, corresponding to the Type I' saddle point defined in Sect. 4.1.

Definition 5.5 – Type I vicinity : $(\mathbf{b}_1, \mathbf{b}_2, \mathbf{x}_{\boldsymbol{\epsilon}(1)}, \mathbf{x}_{\boldsymbol{\epsilon}(2)}, \mathbf{t}, \mathbf{v}_{\boldsymbol{\epsilon}}) \in \Upsilon_+^b \times \Upsilon_-^b \times \Upsilon_+^x \times \Upsilon_-^x \times \Upsilon_S \times \Upsilon_S$ for some $\boldsymbol{\epsilon}$.

- Type I' vicinity : $(\mathbf{b}_1, \mathbf{b}_2, \mathbf{x}_1, \mathbf{x}_2, \mathbf{t}, \mathbf{v}) \in \Upsilon_+^b \times \Upsilon_-^b \times \Upsilon_+^x \times \Upsilon_-^x \times \Upsilon_S \times \Upsilon_S$.
- Type II vicinity : $(\mathbf{b}_1, \mathbf{b}_2, \mathbf{x}_1, \mathbf{x}_2, \mathbf{t}, \mathbf{v}) \in \Upsilon_+^b \times \Upsilon_-^b \times \Upsilon_+^x \times \Upsilon_-^x \times \Upsilon_S \times \mathbb{I}^{W-1}$.
- Type III vicinity : $(\mathbf{b}_1, \mathbf{b}_2, \mathbf{x}_1, \mathbf{x}_2, \mathbf{t}, \mathbf{v}) \in \Upsilon_+^b \times \Upsilon_-^b \times \Upsilon_-^x \times \Upsilon_-^x \times \Upsilon_S \times \mathbb{I}^{W-1}$.

In the following discussion, the parameter ε_0 in Θ is allowed to be different from line to line. However, given ε_1 in (1.8), we shall always choose ε_2 in (1.16) and ε_0 in (5.30) according to the rule

$$C\varepsilon_2 \leq \varepsilon_0 \leq \varepsilon_1/C \tag{5.34}$$

for some sufficiently large $C > 0$. Consequently, by Assumption 1.14 we have

$$N(\log N)^{-10} \geq M = M^{1-4\varepsilon_0} M^{4\varepsilon_0} \geq W^{(4+2\gamma+\varepsilon_1)(1-4\varepsilon_0)} M^{4\varepsilon_0}$$

$$\gg W^{(4+2\gamma+4\epsilon_0)} M^{4\epsilon_0} = W^{2\gamma} \Theta^4. \tag{5.35}$$

To prove Theorem 1.15, we split the task into three steps. The first step is to exclude the integral outside the vicinities. Specifically, we will show the following lemma.

Lemma 5.6 *Under Assumptions 1.1 and 1.14, we have,*

$$\begin{aligned} & \mathcal{I}(\Gamma^W, \bar{F}^W, \Sigma^W, \Sigma^W, \mathbb{R}_+^{W-1}, \mathbb{I}^{W-1}) \\ &= 2^W \mathcal{I}(\gamma_+^b, \gamma_-^b, \gamma_+^x, \gamma_-^x, \gamma_S, \gamma_S) + \mathcal{I}(\gamma_+^b, \gamma_-^b, \gamma_+^x, \gamma_-^x, \gamma_S, \mathbb{I}^{W-1}) \\ &+ \mathcal{I}(\gamma_+^b, \gamma_-^b, \gamma_-^x, \gamma_-^x, \gamma_S, \mathbb{I}^{W-1}) + O(e^{-\Theta}). \end{aligned} \tag{5.36}$$

Remark 5.7 The first three terms on the r.h.s. of (5.36) correspond to the integrals over vicinities of the Type I, II, and III saddle points, respectively. Note that for the first term, we have used the fact that the total contribution of the integral over the Type I vicinity is 2^W times that over the Type I' vicinity.

The second step, is to estimate the integral over the Type I vicinity. We have the following lemma.

Lemma 5.8 *Under Assumptions 1.1 and 1.14, there exists some positive constant C_0 uniform in n and some positive number $N_0 = N_0(n)$ such that for all $N \geq N_0$,*

$$2^W |\mathcal{I}(\gamma_+^b, \gamma_-^b, \gamma_+^x, \gamma_-^x, \gamma_S, \gamma_S)| \leq \frac{N^{C_0}}{(N\eta)^n}. \tag{5.37}$$

The last step is to show that the integral over the Type II and III vicinities are also negligible.

Lemma 5.9 *Under Assumptions 1.1 and 1.14, there exists some positive constant c such that,*

$$\begin{aligned} & \mathcal{I}(\gamma_+^b, \gamma_-^b, \gamma_+^x, \gamma_-^x, \gamma_S, \mathbb{I}^{W-1}) = O(e^{-cW}), \\ & \mathcal{I}(\gamma_+^b, \gamma_-^b, \gamma_-^x, \gamma_-^x, \gamma_S, \mathbb{I}^{W-1}) = O(e^{-cW}). \end{aligned}$$

Therefore, the remaining task is to prove Lemmas 5.1, 5.6, 5.8 and 5.9. For the convenience of the reader, we outline the organization of the subsequent part as follows.

At first, the proofs of Lemmas 5.1 and 5.6 require a discussion on the bound of the integrand, especially on the term $A(\cdot)$, which contains the integral over all the Grassmann variables. To this end, we perform a crude analysis for the function $A(\cdot)$ in Sect. 6 in advance, with which we are able to prove Lemmas 5.1 and 5.6 in Sect. 7. Then, we can restrict ourselves to the integral over the vicinities, i.e., prove Lemmas 5.8 and 5.9. In Sect. 8, we will analyze the factor $\exp\{-M(\hat{K}(\hat{X}, V) + \hat{L}(\hat{B}, T))\}$, which is approximately Gaussian in the Type I vicinity. The key step is to deform the contours of \hat{X} and \hat{B} -variables again to the steepest paths exactly, whereby we can control the error terms in the integrand and prove Lemma 5.8 in Sect. 9. In the

Type II and III vicinities, we only deform the contours of \hat{B} -variables again and bound $\exp\{-M\hat{K}(\hat{X}, V)\}$ by its absolute value directly. It turns out to be enough for our proof of Lemma 5.9, which is given in Sect. 10.

6 Crude bound on $\mathbf{A}(\hat{X}, \hat{B}, V, T)$

In this section, we provide a bound on the function $\mathbf{A}(\cdot)$ in terms of the \hat{B}, T -variables, which holds on all the domains under discussion in the sequel. Here, by *crude bound* we mean a bound of order $\exp\{O(WN^{\varepsilon_2})\}$, which will be specified in Lemma 6.1 below. By the definition in (3.32), we see that $\mathbf{A}(\cdot)$ is an integral of the product of $\mathcal{Q}(\cdot), \mathcal{P}(\cdot)$ and $\mathcal{F}(\cdot)$. As mentioned in Sect. 4.2, a typical procedure we will adopt is to ignore $\mathcal{Q}(\cdot)$ at first, then estimate the integrals of $\mathcal{P}(\cdot)$ and $\mathcal{F}(\cdot)$, which are denoted by $\mathbf{P}(\cdot)$ and $\mathbf{F}(\cdot)$, respectively [see (4.2) and (4.3)], finally, we make necessary comment on how to modify the bounding scheme to take $\mathcal{Q}(\cdot)$ into account, whereby we can get the desired bound for $\mathbf{A}(\cdot)$.

For the sake of simplicity, from now on, we will use the notation

$$\begin{aligned} \omega_{j,1} &:= \omega_{j,11}, & \omega_{j,2} &:= \omega_{j,12}, & \omega_{j,3} &:= \omega_{j,21}, & \omega_{j,4} &:= \omega_{j,22}, \\ \xi_{j,1} &:= \xi_{j,11}, & \xi_{j,2} &:= \xi_{j,21}, & \xi_{j,3} &:= \xi_{j,12}, & \xi_{j,4} &:= \xi_{j,22}. \end{aligned} \tag{6.1}$$

Moreover, we introduce the domains

$$\begin{aligned} \hat{\Sigma} &:= \left\{ r e^{i\vartheta} : |r - 1| \leq \frac{1}{10}, \vartheta \in \mathbb{L} \right\}, \\ \mathbb{K} \equiv \mathbb{K}(E) &:= \begin{cases} \left\{ \omega \in \mathbb{C} : 0 \leq \arg \omega \leq \frac{\arg a_+}{2} + \frac{\pi}{8} \right\}, & \text{if } E \geq 0, \\ \left\{ \omega \in \mathbb{C} : \frac{\arg a_+}{2} - \frac{\pi}{8} \leq \arg \omega \leq 0 \right\}, & \text{if } E < 0. \end{cases} \end{aligned} \tag{6.2}$$

By the assumption that $|E| \leq \sqrt{2} - \kappa$ in (1.16), it is easy to see that $|\arg \omega| \leq \pi/4 - c$ for all $\omega \in \mathbb{K} \cup \bar{\mathbb{K}}$, where c is some positive constant depending on κ . Our aim is to show the following lemma.

Lemma 6.1 *Suppose that $\mathbf{b}_1, \mathbf{b}_2, \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{K}^W \times \bar{\mathbb{K}}^W \times \hat{\Sigma}^W \times \hat{\Sigma}^W$. Under the assumption of Theorem 1.15, we have*

$$|\mathbf{A}(\hat{X}, \hat{B}, V, T)| \leq e^{O(WN^{\varepsilon_2})} \prod_{j=1}^W \left(r_{j,1}^{-1} + r_{j,2}^{-1} + t_j + 1 \right)^C := e^{O(WN^{\varepsilon_2})} \mathbf{p}(\mathbf{r}^{-1}, \mathbf{t}).$$

Remark 6.2 Obviously, using the terminology introduced at the end of Sect. 3.6, we have

$$\mathbf{p}(\mathbf{r}^{-1}, \mathbf{t}) \in \Omega \left(\left\{ r_{j,1}^{-1}, r_{j,2}^{-1}, t_j \right\}_{j=1}^W ; \kappa_1, \kappa_2, \kappa_3 \right), \quad \kappa_1 = e^{O(W)}, \quad \kappa_2, \kappa_3 = O(1). \tag{6.3}$$

6.1 Integral of \mathcal{Q}

In this section, we investigate the function

$$\begin{aligned} & \mathbf{Q}(\Omega, \mathcal{E}, P_1, Q_1, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}) \\ & := \int d\omega^{[1]} d\xi^{[1]} \mathcal{Q}(\Omega, \mathcal{E}, \omega^{[1]}, \xi^{[1]}, P_1, Q_1, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}). \end{aligned} \tag{6.4}$$

Recall $\Omega_{\text{deg}}(\mathbf{a}; \kappa_1, \kappa_2, \kappa_3)$ defined at the end of Sect. 3.6, the parameterization in (3.15) and (3.25) and the notation introduced in (6.1). We have the following lemma.

Lemma 6.3 *If we regard $\sigma, v_p^{[1]}, v_q^{[1]}, P_1$ and $(X^{[1]})^{-1}$ -entries as fixed parameters, we have*

$$\mathbf{Q}(\cdot) \in \Omega_{\text{deg}}(\mathfrak{S}; \kappa_1, \kappa_2, \kappa_3), \quad \kappa_1 = W^{O(1)}, \quad \kappa_2, \kappa_3 = O(1), \tag{6.5}$$

where \mathfrak{S} is the set of variables defined by

$$\mathfrak{S} := \left\{ t, s, (y_k^{[1]})^{-1}, e^{i\sigma_k^{[1]}}, e^{-i\sigma_k^{[1]}}, \frac{\omega_{i,a}\xi_{j,b}}{M} \right\}_{\substack{i,j=1,\dots,W; \\ k=p,q;a,b=1,\dots,4}}.$$

Proof Note that $\mathcal{Q}(\cdot)$ can be regarded as a function of the Grassmann variables in $\omega^{[1]}$ and $\xi^{[1]}$. Hence, by the definition in (3.30), it is a polynomial of these variables with bounded degree. In addition, we can always combine the factor $1/\sqrt{M}$ with Ω_j and \mathcal{E}_j -variables in the first factor of the r.h.s. of (3.30). Then it is not difficult to check that

$$\mathcal{Q}(\cdot) \in \Omega_{\text{deg}}(\mathfrak{S}'; \kappa_1, \kappa_2, \kappa_3), \quad \kappa_1 = W^{O(1)}, \quad \kappa_2, \kappa_3 = O(1), \tag{6.6}$$

where

$$\mathfrak{S}' := \left\{ t, s, (y_k^{[1]})^{-1}, e^{i\sigma_k^{[1]}}, e^{-i\sigma_k^{[1]}}, \frac{\omega_{j,r}\xi_{k,b}^{[1]}}{\sqrt{M}}, \frac{\xi_{j,r}\omega_{k,b}^{[1]}}{\sqrt{M}}, \omega_{k,a}^{[1]}\xi_{\ell,b}^{[1]} \right\}_{\substack{j=1,\dots,W;k=p,q; \\ r=1,\dots,4;a,b=1,2}}.$$

By the definition in (6.4), $\mathbf{Q}(\cdot)$ is the integral of $\mathcal{Q}(\cdot)$ over the $\omega^{[1]}$ and $\xi^{[1]}$ -variables. Now, we regard all the other variables in \mathfrak{S}' , except $\omega^{[1]}$ and $\xi^{[1]}$ -variables, as parameters. By the definition of Grassmann integral, we know that only the coefficient of the highest order term $\prod_{k=p,q} \prod_{a=1,2} \omega_{k,a}^{[1]}\xi_{k,a}^{[1]}$ in $\mathcal{Q}(\cdot)$ survives after integrating $\omega^{[1]}$ and $\xi^{[1]}$ -variables out. Then, it is easy to see (6.5) from (6.6), completing the proof. □

6.2 Integral of \mathcal{P}

In this subsection, we temporarily ignore the Ω and \mathcal{E} -variables from $\mathbf{Q}(\cdot)$, and estimate $\mathbf{P}(\hat{X}, \hat{B}, V, T)$ defined in (4.2). Recalling $r_{j,1}$ and $r_{j,2}$ defined in (5.12), we can formulate our estimate as follows.

Lemma 6.4 *Suppose that the assumptions in Lemma 6.1 hold. We have*

$$|\mathbf{P}(\hat{X}, \hat{B}, V, T)| \leq e^{O(W)} \prod_{j=1}^W \left(r_{j,1}^{-1} + r_{j,2}^{-1} + t_j + 1 \right)^{O(1)}. \tag{6.7}$$

Proof We start with one factor from $\mathcal{P}(\cdot)$ (see (3.29)), namely

$$\begin{aligned} \varpi_j &:= \frac{1}{\det^M (1 + M^{-1} V_j^* \hat{X}_j^{-1} V_j \Omega_j T_j^{-1} \hat{B}_j^{-1} T_j \mathcal{E}_j)} \\ &= \exp \left\{ -M \log \det (1 + M^{-1} V_j^* \hat{X}_j^{-1} V_j \Omega_j T_j^{-1} \hat{B}_j^{-1} T_j \mathcal{E}_j) \right\} \\ &= 1 + \sum_{\ell=1}^4 \frac{1}{M^{\ell-1}} \mathfrak{p}_\ell(\hat{X}_j, \hat{B}_j, V_j, T_j, \Omega_j, \mathcal{E}_j). \end{aligned} \tag{6.8}$$

Here $\mathfrak{p}_\ell(\cdot)$ is a polynomial in $\hat{X}_j^{-1}, \hat{B}_j^{-1}, V_j, T_j, \Omega_j$ and \mathcal{E}_j -entries with bounded degree and bounded coefficients. Moreover, if we regard $\mathfrak{p}_\ell(\cdot)$ as a polynomial of Ω_j and \mathcal{E}_j -entries, it is homogeneous, with degree 2ℓ , and the total degree for Ω_j -variables is ℓ , thus that for \mathcal{E}_j -entries is also ℓ . More specifically, we can write

$$\mathfrak{p}_\ell(\hat{X}_j, \hat{B}_j, V_j, T_j, \Omega_j, \mathcal{E}_j) = \sum_{\substack{\alpha_1, \dots, \alpha_\ell, \\ \beta_1, \dots, \beta_\ell=1}}^4 \mathfrak{p}_{\ell, \alpha, \beta}(\hat{X}_j, \hat{B}_j, V_j, T_j) \prod_{i=1}^{\ell} \omega_{j, \alpha_i} \xi_{j, \beta_i},$$

where we used the notation in (6.1) and denoted $\alpha = (\alpha_1, \dots, \alpha_\ell)$ and $\beta = (\beta_1, \dots, \beta_\ell)$. It is easy to verify that ϖ_j is of the form (6.8) by taking Taylor expansion w.r.t. the Grassmann variables. The expansion in (6.8) terminates at $\ell = 4$, owing to the fact that there are totally 8 Grassmann variables from Ω_j and \mathcal{E}_j . In addition, it is also easy to check that $\mathfrak{p}_{\ell, \alpha, \beta}(\cdot)$ is a polynomial of $\hat{X}_j^{-1}, \hat{B}_j^{-1}, V_j, T_j$ -entries with bounded degree and bounded coefficients, which implies that there exist two positive constants C_1 and C_2 , such that

$$|\mathfrak{p}_{\ell, \alpha, \beta}(\cdot)| \leq C_1 \left(r_{j,1}^{-1} + r_{j,2}^{-1} + t_j + 1 \right)^{C_2} \tag{6.9}$$

uniformly in ℓ, α and β . Here we used the fact that \hat{X}_j^{-1} and V_j -entries are all bounded and T_j -entries are bounded by $1 + t_j$.

Now, we go back to the definition of $\mathcal{P}(\cdot)$ in (3.29) and study the last factor. Similarly to the discussion above, it is easy to see that for $k = p$ or q ,

$$\hat{\omega}_k := \frac{\det (V_k^* \hat{X}_k V_k + M^{-1} \Omega_k T_k^{-1} \hat{B}_k^{-1} T_k \mathcal{E}_k)}{\det \hat{B}_k} = \hat{\mathfrak{p}}_0(\hat{X}_k, \hat{B}_k)$$

$$+ \sum_{\ell=1}^4 \sum_{\alpha, \beta} \hat{p}_{\ell, \alpha, \beta}(\hat{X}_k, \hat{B}_k, V_k, T_k) \prod_{i=1}^{\ell} \omega_{k, \alpha, i} \xi_{k, \beta, i}, \tag{6.10}$$

where $\hat{p}_0(\cdot) = \det \hat{X}_k / \det \hat{B}_k$ and $\hat{p}_{\ell, \alpha, \beta}(\cdot)$'s are some polynomials of $\hat{X}_k, \hat{B}_k^{-1}, V_k, T_k$ -entries with bounded degree and bounded coefficients. Similarly, we have

$$|\hat{p}_0(\cdot)|, |\hat{p}_{\ell, \alpha, \beta}(\cdot)| \leq C_1(r_{k,1}^{-1} + r_{k,2}^{-1} + t_k + 1)^{C_2} \tag{6.11}$$

for some positive constants C_1 and C_2 .

According to the definitions in (6.8) and (6.10), we can rewrite (3.29) as

$$\mathcal{P}(\Omega, \Xi, \hat{X}, \hat{B}, V, T) = \exp \left\{ - \sum_{j,k} \tilde{s}_{jk} Tr \Omega_j \Xi_k \right\} \prod_{j=1}^W \boldsymbol{\omega}_j \prod_{k=p,q} \hat{\boldsymbol{\omega}}_k. \tag{6.12}$$

In light of the discussion above, $\prod_{j=1}^W \boldsymbol{\omega}_j \prod_{k=p,q} \hat{\boldsymbol{\omega}}_k$ is a polynomial of $\hat{X}^{-1}, \hat{B}^{-1}, V, T, \Omega$ and Ξ -entries, in which each monomial is of the form

$$q_{\ell, \alpha, \beta}(\hat{X}^{-1}, \hat{B}^{-1}, V, T) \prod_{j=1}^W \prod_{i=1}^{\ell_j} \omega_{j, \alpha, i} \xi_{j, \beta, i}, \tag{6.13}$$

where we used the notation

$$\begin{aligned} \ell &= (\ell_1, \dots, \ell_W), \quad \alpha \equiv \boldsymbol{\alpha}(\ell) = (\alpha_1, \dots, \alpha_W), \quad \beta \equiv \boldsymbol{\beta}(\ell) = (\beta_1, \dots, \beta_W), \\ \alpha_j &= (\alpha_{j,1}, \dots, \alpha_{j,\ell_j}), \quad \beta_j = (\beta_{j,1}, \dots, \beta_{j,\ell_j}), \end{aligned} \tag{6.14}$$

and $q_{\ell, \alpha, \beta}(\cdot)$ is a polynomial of $\hat{X}, \hat{X}^{-1}, \hat{B}^{-1}, V$ and T -entries. Moreover, all the entries of ℓ, α and β are bounded by 4. By (6.9) and (6.11), we have

$$|q_{\ell, \alpha, \beta}(\hat{X}^{-1}, \hat{B}^{-1}, V, T)| \leq e^{O(W)} \prod_{j=1}^W (r_{j,1}^{-1} + r_{j,2}^{-1} + t_j + 1)^C. \tag{6.15}$$

In addition, it is easy to see that the number of the summands of the form (6.13) in $\prod_{j=1}^W \boldsymbol{\omega}_j \prod_{k=p,q} \hat{\boldsymbol{\omega}}_k$ is bounded by $e^{O(W)}$.

Now, we define the vectors

$$\boldsymbol{\Omega} := (\omega_1, \omega_2, \omega_3, \omega_4), \quad \boldsymbol{\Xi} := (\xi_1, \xi_2, \xi_3, \xi_4), \tag{6.16}$$

where $\omega_\alpha = (\omega_{1,\alpha}, \dots, \omega_{W,\alpha})$ and $\xi_\alpha = (\xi_{1,\alpha}, \dots, \xi_{W,\alpha})$ for $\alpha = 1, 2, 3, 4$. Here we used the notation (6.1). In addition, we introduce the matrix

$$\tilde{\mathbb{H}} = \tilde{S} \oplus \tilde{S} \oplus \tilde{S} \oplus \tilde{S}.$$

It is easy to check $\sum_{j,k} \tilde{s}_{jk} Tr \Omega_j \Xi_k = \mathbf{\Omega} \tilde{\mathbb{H}} \mathbf{\Xi}'$. By using the Gaussian integral formula for the Grassmann variables (3.2), we see that for each ℓ, α and β , we have

$$\left| \int d\Omega d\Xi \cdot \exp \left\{ - \sum_{j,k} \tilde{s}_{jk} Tr \Omega_j \Xi_k \right\} \cdot \prod_{j=1}^W \prod_{i=1}^{\ell_j} \omega_{j,\alpha_j,i} \xi_{j,\beta_j,i} \right| \leq |\det \tilde{\mathbb{H}}^{(\ell|\mathbf{J})}|, \tag{6.17}$$

for some index sets ℓ and \mathbf{J} with $|\ell| = |\mathbf{J}|$. By Assumption 1.1 (i) and (ii), we see that the 2-norm of each row of \tilde{S} is $O(1)$. Consequently, by using Hadamard’s inequality, we have

$$|\det \tilde{\mathbb{H}}^{(\ell|\mathbf{J})}| = e^{O(W)}. \tag{6.18}$$

Therefore, (6.12)–(6.18) and the bound $e^{O(W)}$ for the total number of summands of the form (6.13) in $\prod_{j=1}^W \hat{\omega}_j \prod_{k=p,q} \hat{\omega}_k$ imply (6.7). Thus we completed the proof. \square

6.3 Integral of \mathcal{F}

In this subsection, we also temporarily ignore the $X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}, P_1, Q_1$ -variables from $\mathbf{Q}(\cdot)$, and estimate $F(\hat{X}, \hat{B}, V, T)$ defined in (4.3). We have the following lemma.

Lemma 6.5 *Suppose that the assumptions in Lemma 6.1 hold. We have*

$$|F(\hat{X}, \hat{B}, V, T)| \leq e^{O(WN^{\varepsilon_2})} \prod_{k=p,q} (r_{k,1}^{-1} + r_{k,2}^{-1} + t_k + 1)^{O(1)}. \tag{6.19}$$

Proof Recalling the decomposition of $\mathcal{F}(\cdot)$ in (3.19) together with the parameterization in (3.29), we will study the integrals

$$\mathbb{G}(\hat{B}, T) := \int d\nu(Q_1) d\mathbf{y}^{[1]} d\mathbf{w}^{[1]} g(Q_1, T, \hat{B}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}), \tag{6.20}$$

$$\mathbb{F}(\hat{X}, V) := \int d\mu(P_1) dX^{[1]} f(P_1, V, \hat{X}, X^{[1]}) \tag{6.21}$$

separately. Recalling the convention at the end of Sect. 3, we use $f(\cdot)$ and $g(\cdot)$ to represent the integrands above. One can refer to (3.20) and (3.21) for the definition. From the assumption $\eta \leq M^{-1}N^{\varepsilon_2}$, we see

$$|\mathbb{F}(\hat{X}, V)| \leq e^{O(WN^{\varepsilon_2})}, \tag{6.22}$$

since $P_1, V, \hat{X}, X^{[1]}$ -variables are all bounded and $|\det X_p^{[1]}|, |\det X_q^{[1]}| \sim 1$ when $\mathbf{x}_1, \mathbf{x}_2 \in \hat{\Sigma}$.

For $\mathbb{G}(\hat{B}, T)$, we use the facts

$$\begin{aligned} \operatorname{Re}(Tr B_j Y_k^{[1]} J) &\geq 0, \quad \operatorname{Re}(i Tr Y_k^{[1]} J Z) = -\eta Tr Y_k^{[1]} \leq 0, \\ Tr Y_k^{[1]} J Y_\ell^{[1]} J &\geq 0, \quad k, \ell = p, q, \\ |(\mathbf{w}_q^{[1]}(\mathbf{w}_q^{[1]})^*)_{12}| &\leq 1, \quad |(\mathbf{w}_p^{[1]}(\mathbf{w}_p^{[1]})^*)_{21}| \leq 1, \end{aligned} \tag{6.23}$$

to estimate trivially several terms, whereby we can get the bound

$$|g(\cdot)| \leq \exp \left\{ -M\eta \sum_{j=1}^W Tr \operatorname{Re}(B_j) J \right\} \prod_{k=p,q} \left(y_k^{[1]} \right)^{n+3} \exp \left\{ -\tilde{s}_{kk} Tr \operatorname{Re}(B_k) Y_k^{[1]} J \right\}. \tag{6.24}$$

Here $\operatorname{Re}(B_j) = Q_1^{-1} T_j^{-1} \operatorname{Re}(\hat{B}_j) T_j Q_1$. Hence, integrating $y_p^{[1]}$ and $y_q^{[1]}$ out yields

$$\int_{\mathbb{R}_+^2} dy_p^{[1]} dy_q^{[1]} |g(Q_1, T, \hat{B}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]})| \leq C \frac{\exp \left\{ -M\eta \sum_{j=1}^W Tr \operatorname{Re}(B_j) J \right\}}{\prod_{k=p,q} \left((\mathbf{w}_k^{[1]})^* J \operatorname{Re}(B_k) \mathbf{w}_k^{[1]} \right)^{C_1}}, \tag{6.25}$$

for some positive constants C and C_1 depending on n , where we used the elementary facts that $\tilde{s}_{kk} \geq c$ for some positive constant c and

$$Tr \operatorname{Re}(B_j) Y_k^{[1]} J = y_k^{[1]} (\mathbf{w}_k^{[1]})^* J \operatorname{Re}(B_j) \mathbf{w}_k^{[1]}, \quad k = p, q, \quad j = 1, \dots, W. \tag{6.26}$$

Now, note that

$$(\mathbf{w}_k^{[1]})^* J \operatorname{Re} B_j \mathbf{w}_k^{[1]} \geq \lambda_1(J \operatorname{Re} B_j), \quad k = p, q, \quad j = 1, \dots, W. \tag{6.27}$$

In addition, it is also easy to see $\lambda_1(T_j) = s_j - t_j$ and $\lambda_1(Q_1) = s - t$, according to the definitions in (3.25). Now, by the fact $JA^{-1} = AJ$ for any $A \in \mathring{U}(1, 1)$, we have

$$J \operatorname{Re} B_j = Q_1 T_j \operatorname{diag}(\operatorname{Re} b_{j,1}, \operatorname{Re} b_{j,2}) T_j Q_1. \tag{6.28}$$

Consequently, we can get

$$\lambda_1(J \operatorname{Re} B_j) \geq (s_j - t_j)^2 (s - t)^2 \min\{\operatorname{Re} b_{j,1}, \operatorname{Re} b_{j,2}\} = \frac{\min\{\operatorname{Re} b_{j,1}, \operatorname{Re} b_{j,2}\}}{(s_j + t_j)^2 (s + t)^2}, \tag{6.29}$$

by recalling the facts $s^2 - t^2 = 1$ and $s_j^2 - t_j^2 = 1$. Therefore, combining (6.25), (6.27) and (6.29), we have

$$\int_{\mathbb{R}_+^2} dy_p^{[1]} dy_q^{[1]} |g(\cdot)| \leq C(s+t)^{4C_1} \exp \left\{ -M\eta \sum_{j=1}^W Tr(\text{Re}B_j)J \right\} \times \prod_{k=p,q} \frac{(s_k+t_k)^{2C_1}}{(\min\{\text{Re}b_{k,1}, \text{Re}b_{k,2}\})^{C_1}}. \tag{6.30}$$

Now, what remains is to estimate the exponential function in (6.30). By elementary calculation from (6.28) we obtain

$$Tr(\text{Re}B_j)J \geq (\text{Re}b_{j,1} + \text{Re}b_{j,2})(s_j^2 + t_j^2)(s^2 + t^2) - 4sts_jt_j.$$

Observe that

$$\begin{aligned} (s_j^2 + t_j^2)(s^2 + t^2) - 4sts_jt_j &= \frac{(s_j^2 + t_j^2)^2 (s^2 + t^2)^2 - 16(sts_jt_j)^2}{(s_j^2 + t_j^2)(s^2 + t^2) + 4sts_jt_j} \\ &\geq \frac{4t^4 + 4t^2 + 4t_j^4 + 4t_j^2 + 1}{2(1 + 2t_j^2)(1 + 2t^2)} \geq \frac{1 + 2t^2}{2(1 + 2t_j^2)}. \end{aligned}$$

It implies that

$$\exp \left\{ -M\eta \sum_{j=1}^W Tr(\text{Re}B_j)J \right\} \leq \exp \left\{ -cM\eta \sum_{j=1}^W \frac{r_{j,1} + r_{j,2}}{1 + 2t_j^2} (1 + 2t^2) \right\}, \tag{6.31}$$

for some positive constant c , where we have used the fact $\text{Re}b_{j,\alpha} \geq cr_{j,\alpha}$ for all $j = 1, \dots, W$ and $\alpha = 1, 2$, in light of the assumption $|E| \leq \sqrt{2} - \kappa$ and the definition of \mathbb{K} in (6.2). Plugging (6.31) into (6.30), estimating $(s+t)^2 \leq 2(1+2t^2)$, and integrating t out, we can crudely bound

$$\int_{\mathbb{R}_+^2} dy_p^{[1]} dy_q^{[1]} \int_{\mathbb{R}_+} 2tdt \cdot |g(\cdot)| \leq C \left(\frac{1}{M\eta} \right)^{C_2} \left(\sum_{j=1}^W \frac{r_{j,1} + r_{j,2}}{1 + 2t_j^2} \right)^{-C_2} \times \prod_{k=p,q} \left(\frac{1 + 2t_k^2}{\min\{r_{k,1}, r_{k,2}\}} \right)^{C_1}. \tag{6.32}$$

Now, we use the trivial bounds

$$\left(\sum_{j=1}^W \frac{r_{j,1} + r_{j,2}}{1 + 2t_j^2} \right)^{-C_2} \leq \left(\frac{1 + 2t_p^2}{r_{p,1} + r_{p,2}} \right)^{C_2} \leq \left((1 + 2t_p^2) (r_{p,1}^{-1} + r_{p,2}^{-1}) \right)^{C_2}, \tag{6.33}$$

and

$$\frac{1 + 2t_k^2}{\min\{r_{k,1}, r_{k,2}\}} \leq C(1 + 2t_k^2) (r_{k,1}^{-1} + r_{k,2}^{-1}). \tag{6.34}$$

Inserting (6.33) and (6.34) into (6.32) and integrating out the remaining variables yields

$$|\mathbb{G}(\hat{B}, T)| \leq C \left(\frac{1}{M\eta} \right)^{C_1} \prod_{k=p,q} (r_{k,1}^{-1} + r_{k,2}^{-1} + t_k + 1)^{C_3}. \tag{6.35}$$

Combining (6.22) and (6.35) we can get the bound (6.19). Thus we completed the proof of Lemma 6.5. □

6.4 Summing up: Proof of Lemma 6.1

In the discussions in Sects. 6.2 and 6.3, we ignored the irrelevant factor $\mathbf{Q}(\cdot)$. However, it is easy to modify the discussion slightly to take this factor into account, whereby we can prove Lemma 6.1.

Proof At first, by the definition in (6.4), we can rewrite (3.32) as

$$\mathbf{A}(\cdot) = \int dX^{[1]} dy^{[1]} d\mathbf{w}^{[1]} d\Omega d\mathcal{E} d\mu(P_1) d\nu(Q_1) \mathcal{P}(\cdot) \mathbf{Q}(\cdot) \mathcal{F}(\cdot).$$

Now, by the conclusion $\kappa_1 = W^{O(1)}$ in Lemma 6.3, it suffices to consider one term in $\mathbf{Q}(\cdot)$, which is a monomial of the form $\mathfrak{p}(t, s, (y_p^{[1]})^{-1}, (y_q^{[1]})^{-1})\mathfrak{q}(\Omega, \mathcal{E})$, regarding $\sigma, v_p^{[1]}, v_q^{[1]}, P_1$ -variables, $X^{[1]}$ -variables and $\mathbf{w}^{[1]}$ -variables as bounded parameters. Here $\mathfrak{p}(\cdot)$ is a monomial of $t, s, (y_p^{[1]})^{-1}, (y_q^{[1]})^{-1}$ and $\mathfrak{q}(\cdot)$ is a monomial of Ω, \mathcal{E} -variables, both with bounded coefficients and bounded degrees, according to the fact $\kappa_2, \kappa_3 = O(1)$ in Lemma 6.3. Now we define

$$\begin{aligned} \mathbf{P}_q(\hat{X}, \hat{B}, V, T) &:= \int d\Omega d\mathcal{E} \mathcal{P}(\cdot) \cdot \mathfrak{q}(\Omega, \mathcal{E}), \\ \mathbf{F}_p(\hat{X}, \hat{B}, V, T) &:= \int dX^{[1]} dy^{[1]} d\mathbf{w}^{[1]} d\mu(P_1) d\nu(Q_1) \mathcal{F}(\cdot) \\ &\cdot \mathfrak{p}(t, s, (y_p^{[1]})^{-1}, (y_q^{[1]})^{-1}). \end{aligned} \tag{6.36}$$

By repeating the discussions in Sects. 6.2 and 6.3 with slight modification, we can easily see that (6.7) and (6.19) hold as well, if we replace $P(\cdot)$ and $F(\cdot)$ by $P_q(\cdot)$ and $F_p(\cdot)$, respectively. Therefore, we completed the proof of Lemma 6.1. \square

7 Proofs of Lemmas 5.1 and 5.6

In this section, with the aid of Lemma 6.1, we prove Lemmas 5.1 and 5.6. A key problem is that the domain of \mathbf{t} -variables is not compact. This forces us to analyze the exponential function

$$\mathbb{M}(\mathbf{t}) := \exp \left\{ -M \operatorname{Re}(\ell_S(\hat{B}, T)) \right\} \tag{7.1}$$

carefully for any fixed \hat{B} -variables. Recall the definition of the sector \mathbb{K} in (6.2). For $\mathbf{b}_1 \in \mathbb{K}^W$ and $\mathbf{b}_2 \in \bar{\mathbb{K}}^W$, we have

$$\min_{j,k} \operatorname{Re}(b_{j,1} + b_{j,2})(b_{k,1} + b_{k,2}) \geq c \min_{j,k} \sum_{a,b=1,2} r_{j,a} r_{k,b} \geq c \min_{j,a} r_{j,a}^2 := \mathfrak{A}(\hat{B}), \tag{7.2}$$

From now on, we regard $\mathbb{M}(\mathbf{t})$ as a measure of the \mathbf{t} -variables and study it in the following two regions separately:

$$(i) : \mathbf{t} \in \mathbb{I}^{W-1}, \quad (ii) : \mathbf{t} \in \mathbb{R}_+^{W-1} \setminus \mathbb{I}^{W-1}.$$

Roughly speaking, when $\mathbf{t} \in \mathbb{I}^{W-1}$, we will see that $\mathbb{M}(\mathbf{t})$ can be bounded pointwisely by a Gaussian measure. More specifically, we have the following lemma.

Lemma 7.1 *With the notation above, we have*

$$\mathbb{M}(\mathbf{t}) \leq \exp \left\{ -\frac{M}{12} \mathfrak{A}(\hat{B}) \sum_{j,k} \mathfrak{s}_{jk} (t_k - t_j)^2 \right\}, \quad \forall \mathbf{t} \in \mathbb{I}^{W-1}.$$

Proof Using the definition of $\ell_S(\hat{B}, T)$ in (5.6) and $\mathfrak{A}(\hat{B})$ in (7.2) and the fact $|(T_k T_j^{-1})_{12}| = |s_j t_k e^{i\sigma_k} - s_k t_j e^{i\sigma_j}|$, we have

$$\operatorname{Re} \ell_S(\hat{B}, T) \geq \frac{1}{2} \mathfrak{A}(\hat{B}) \sum_{j,k} \mathfrak{s}_{jk} \left| s_j t_k e^{i\sigma_k} - s_k t_j e^{i\sigma_j} \right|^2. \tag{7.3}$$

Simple estimate using $s_j^2 = 1 + t_j^2$ shows that

$$\left| s_j t_k e^{i\sigma_k} - s_k t_j e^{i\sigma_j} \right|^2 \geq \frac{1}{4} (t_k - t_j)^2 \left(\frac{1}{1 + 2t_j^2} + \frac{1}{1 + 2t_k^2} \right) \geq \frac{1}{6} (t_k - t_j)^2.$$

$$(7.4)$$

Notice that the assumption $\mathbf{t} \in \mathbb{I}^{W-1}$ was used only in the last inequality. By (7.3), (7.4) and the definition (7.1), Lemma 7.1 follows immediately. \square

However, the behavior of $\mathbb{M}(\mathbf{t})$ for $\mathbf{t} \in \mathbb{R}_+^{W-1} \setminus \mathbb{I}^{W-1}$ is much more sophisticated. We will not try to provide a pointwise control of $\mathbb{M}(\mathbf{t})$ in this region. Instead, we will bound the integral of $q(\mathbf{t})$ against $\mathbb{M}(\mathbf{t})$ over this region, for any given monomial $q(\cdot)$ of interest. More specifically, recalling the definition of Θ in (5.30) and the spanning tree $\mathcal{G}_0 = (\mathcal{V}, \mathcal{E}_0)$ in Assumption 1.1, and additionally setting

$$\mathfrak{L} := \frac{M}{4} \mathfrak{A}(\hat{B}) \min_{i,j \in \mathcal{E}_0} s_{ij}, \tag{7.5}$$

we have the following lemma.

Lemma 7.2 *Let $q(\mathbf{t}) = \prod_{j=2}^W t_j^{n_j}$ be a monomial of \mathbf{t} -variables, with powers $n_j = O(1)$ for all $j = 2, \dots, W$. We have*

$$\int_{\mathbb{R}_+^{W-1} \setminus \mathbb{I}^{W-1}} \prod_{j=2}^W dt_j \mathbb{M}(\mathbf{t}) q(\mathbf{t}) \leq \left(1 + \mathfrak{L}^{-\frac{1}{2}}\right)^{O(W^2)} \exp\left\{-\Theta^2 \mathfrak{A}(\hat{B}) + O(W^2 \log N)\right\} \tag{7.6}$$

Remark 7.3 Roughly speaking, by Lemma 7.2 we see that the integral of $q(\mathbf{t})$ -variables against the measure $\mathbb{M}(\mathbf{t})$ over the region $\mathbb{R}_+^{W-1} \setminus \mathbb{I}^{W-1}$ is exponentially small, owing to the fact $\Theta^2 \gg W^2 \log N$.

We will postpone the proof of Lemma 7.2 to the end of this section. In the sequel, at first, we prove Lemmas 5.1 and 5.6 with the aid of Lemmas 6.1, 7.1 and 7.2. Before commencing the formal proofs, we mention two basic facts which are formulated as the following lemma.

Lemma 7.4 *Under Assumption 1.1, we have the following two facts.*

- For the smallest eigenvalue of $-S^{(1)}$, there exists some positive constant c such that

$$\lambda_1(-S^{(1)}) \geq c/W^2. \tag{7.7}$$

- Let $\mathbf{q} = (\rho_2, \dots, \rho_W)'$ be a real vector and $\rho_1 = 0$. If there is at least one $\alpha \in \{2, \dots, W\}$ such that $q_\alpha \geq \Theta/\sqrt{M}$, then we have

$$\sum_{j,k} s_{jk} (q_j - q_k)^2 \geq \Theta/M. \tag{7.8}$$

Proof Let $\boldsymbol{\rho} = (\rho_2, \dots, \rho_W)'$ be a real vector and $\rho_1 = 0$. Now, we assume $|\rho_\alpha| = \max_{\beta=2, \dots, W} |\rho_\beta|$. Then

$$\frac{-\boldsymbol{\rho}' S^{(1)} \boldsymbol{\rho}}{\|\boldsymbol{\rho}\|_2^2} = \frac{\frac{1}{2} \sum_{j,k} \mathfrak{s}_{jk} (\rho_j - \rho_k)^2}{\sum_j \rho_j^2} \geq \frac{c(\rho_\alpha - \rho_1)^2}{W^2 \rho_\alpha^2} = \frac{c}{W^2},$$

where the second step follows from Assumption 1.1 (iv) and Cauchy-Schwarz inequality. Analogously, we have $\sum_{j,k} \mathfrak{s}_{jk} (\varrho_j - \varrho_k)^2 \geq c\varrho_\alpha^2/W$, which implies (7.8) by the definition of Θ in (5.30). Hence, we completed the proof. \square

Recalling the notation defined in (5.2) and the facts $|x_{j,a}| = 1$ and $|b_{j,a}| = r_{j,a}$ for all $j = 1, \dots, W$ and $a = 1, 2$, for any sequence of domains, we have

$$\begin{aligned} & \left| \mathcal{I}(\mathbf{I}_1^b, \mathbf{I}_2^b, \mathbf{I}_1^x, \mathbf{I}_2^x, \mathbf{I}^t, \mathbf{I}^v) \right| \\ & \leq e^{O(W \log N)} \int_{\mathbf{I}_1^b} \prod_{j=1}^W db_{j,1} \int_{\mathbf{I}_2^b} \prod_{j=1}^W db_{j,2} \int_{\mathbf{I}_1^x} \prod_{j=1}^W dx_{j,1} \int_{\mathbf{I}_2^x} \prod_{j=1}^W dx_{j,2} \\ & \times \int_{\mathbf{I}^t} \prod_{j=2}^W 2t_j dt_j \int_{\mathbf{I}^v} \prod_{j=2}^W 2v_j dv_j \times \exp \left\{ -M(\operatorname{Re}K(\hat{X}, V) + \operatorname{Re}L(\hat{B}, T)) \right\} \\ & \cdot |\mathbf{A}(\hat{X}, \hat{B}, V, T)| \cdot \prod_{j=1}^W (r_{j,1} + r_{j,2})^2. \end{aligned} \tag{7.9}$$

In addition, according to Lemma 6.1, we have

$$|\mathbf{A}(\hat{X}, \hat{B}, V, T)| \cdot \prod_{j=2}^W 2t_j \cdot \prod_{j=2}^W 2v_j \cdot \prod_{j=1}^W (r_{j,1} + r_{j,2})^2 \leq e^{O(WN^{\varepsilon_2})} \tilde{\mathfrak{p}}(\mathbf{r}, \mathbf{r}^{-1}, \mathbf{t}), \tag{7.10}$$

for some polynomial $\tilde{\mathfrak{p}}(\mathbf{r}, \mathbf{r}^{-1}, \mathbf{t})$ with positive coefficients, and

$$\begin{aligned} \tilde{\mathfrak{p}}(\mathbf{r}, \mathbf{r}^{-1}, \mathbf{t}) & \in \mathcal{Q} \left(\{r_{j,1}, r_{j,2}, r_{j,1}^{-1}, r_{j,2}^{-1}, t_j\}_{j=1}^W; \kappa_1, \kappa_2, \kappa_3 \right), \quad \kappa_1 = e^{O(W)}, \\ \kappa_2, \kappa_3 & = O(1). \end{aligned} \tag{7.11}$$

7.1 Proof of Lemma 5.1

At first, since throughout the whole proof, the domains of $\mathbf{x}_1, \mathbf{x}_2$, and \mathbf{v} -variables, namely, Σ^W, Σ^W and \mathbb{I}^{W-1} , will not be involved, we just use $*$'s to represent them, in order to simplify the notation.

Now, we introduce the following contours with the parameter $\mathfrak{D} \in \mathbb{R}_+$,

$$\Gamma_{\mathfrak{D}} := \{ra_+ | r \in [0, \mathfrak{D}]\} \subset \Gamma, \quad \mathbb{R}_{\mathfrak{D}} = [0, (\operatorname{Re}a_+)\mathfrak{D}] \subset \mathbb{R}_+,$$

$$\mathcal{L}_{\mathfrak{D}} := \{(\operatorname{Re}a_+)\mathfrak{D} + \mathbf{i}(\operatorname{Im}a_+)r \mid r \in [0, \mathfrak{D}]\}.$$

In addition, we recall the sector \mathbb{K} defined in (6.2). Then, trivially, we have $\mathbb{R}_+, \Gamma, \mathcal{L}_{\mathfrak{D}} \subset \mathbb{K}$ and $\mathbb{R}_+, \bar{\Gamma}, \bar{\mathcal{L}}_{\mathfrak{D}} \in \bar{\mathbb{K}}$ for all $\mathfrak{D} \in \mathbb{R}_+$.

Observe that the integrand in (5.2) is an analytic function of the \hat{B} -variables. To see this, we can go back to the integral representation (3.17) and the definitions of $L(B)$ and $\mathcal{P}(\Omega, \mathcal{E}, X, B)$ in (3.18). Since $\exp\{M \log \det B_j\} = (\det B_j)^M$, actually the logarithmic terms in $L(B)$ do not produce any singularity in the integrand in (3.17); it can also compensate the factors $b_{j,a}^{-\ell}$ from $\mathcal{P}(\Omega, \mathcal{E}, X, B)$ since $\ell \ll M$. Consequently, we have

$$\mathcal{I}\left((\Gamma_{\mathfrak{D}} \cup \mathcal{L}_{\mathfrak{D}})^W, (\bar{\Gamma}_{\mathfrak{D}} \cup \bar{\mathcal{L}}_{\mathfrak{D}})^W, *, *, \mathbb{R}_+^{W-1}, *\right) = \mathcal{I}\left((\mathbb{R}_{\mathfrak{D}})^W, (\mathbb{R}_{\mathfrak{D}})^W, *, *, \mathbb{R}_+^{W-1}, *\right).$$

Hence, to prove Lemma 5.1, it suffices to prove the following lemma.

Lemma 7.5 *Suppose that $|E| \leq \sqrt{2} - \kappa$. As $\mathfrak{D} \rightarrow \infty$, the following convergence hold,*

- (i) : $\mathcal{I}\left((\Gamma_{\mathfrak{D}} \cup \mathcal{L}_{\mathfrak{D}})^W, (\bar{\Gamma}_{\mathfrak{D}} \cup \bar{\mathcal{L}}_{\mathfrak{D}})^W, *, *, \mathbb{R}_+^{W-1}, *\right) - \mathcal{I}\left((\Gamma_{\mathfrak{D}})^W, (\bar{\Gamma}_{\mathfrak{D}})^W, *, *, \mathbb{R}_+^{W-1}, *\right) \rightarrow 0,$
- (ii) : $\mathcal{I}\left((\Gamma)^W, (\bar{\Gamma})^W, *, *, \mathbb{R}_+^{W-1}, *\right) - \mathcal{I}\left((\Gamma_{\mathfrak{D}})^W, (\bar{\Gamma}_{\mathfrak{D}})^W, *, *, \mathbb{R}_+^{W-1}, *\right) \rightarrow 0,$
- (iii) : $\mathcal{I}\left(\mathbb{R}_+^W, \mathbb{R}_+^W, *, *, \mathbb{R}_+^{W-1}, *\right) - \mathcal{I}\left(\mathbb{R}_{\mathfrak{D}}^W, \mathbb{R}_{\mathfrak{D}}^W, *, *, \mathbb{R}_+^{W-1}, *\right) \rightarrow 0.$

Proof For simplicity, we use the notation

$$\begin{aligned} \mathbf{I}_{\mathfrak{D}}^{b,1} &:= (\Gamma_{\mathfrak{D}} \cup \mathcal{L}_{\mathfrak{D}})^W \times (\bar{\Gamma}_{\mathfrak{D}} \cup \bar{\mathcal{L}}_{\mathfrak{D}})^W \setminus (\Gamma_{\mathfrak{D}})^W \times (\bar{\Gamma}_{\mathfrak{D}})^W, \\ \mathbf{I}_{\mathfrak{D}}^{b,2} &:= \Gamma^W \times \bar{\Gamma}^W \setminus (\Gamma_{\mathfrak{D}})^W \times (\bar{\Gamma}_{\mathfrak{D}})^W, \quad \mathbf{I}_{\mathfrak{D}}^{b,3} := \mathbb{R}_+^W \times \mathbb{R}_+^W \setminus \mathbb{R}_{\mathfrak{D}}^W \times \mathbb{R}_{\mathfrak{D}}^W. \end{aligned}$$

Now, recall the definition of the function $\ell(\mathbf{a})$ in (5.6) and the representation of $L(\hat{B}, T)$ in (5.5). Hence, in light of the definition of $\mathbb{M}(\mathbf{t})$ in (7.1), we have

$$\exp\{-M \operatorname{Re}L(\hat{B}, T)\} = \exp\{-M(\operatorname{Re}\ell(\mathbf{b}_1) + \operatorname{Re}\ell(-\mathbf{b}_2))\} \mathbb{M}(\mathbf{t}). \tag{7.12}$$

By the assumption $|E| \leq \sqrt{2} - \kappa$, we see that $\operatorname{Re}b_{j,a}b_{k,a} > 0$ for all $b_{j,a}, b_{k,a} \in \mathbb{K} \cup \bar{\mathbb{K}}$. Consequently, when $b_{j,1} \in \mathbb{K}$ and $b_{j,2} \in \bar{\mathbb{K}}$ for all $j = 1, \dots, W$, we have

$$\begin{aligned} \operatorname{Re}\ell(\mathbf{b}_1) + \operatorname{Re}\ell(-\mathbf{b}_2) &\geq \sum_{a=1,2} \sum_j \left(\frac{1}{2}(1 + \mathfrak{s}_{jj}) \operatorname{Re}b_{j,a}^2 + (-1)^{a+1} \operatorname{Elm}b_{j,a} - \log r_{j,a} \right) \\ &\geq c \sum_j (r_{j,1}^2 + r_{j,2}^2) - \sum_j (\log r_{j,1} + \log r_{j,2}), \end{aligned} \tag{7.13}$$

where we used Assumption 1.1 (ii) and the fact that $(-1)^{a+1} \operatorname{Elm}b_{j,a} \geq 0$.

Now, when $(\mathbf{b}_1, \mathbf{b}_2) \in \mathbf{I}_{\mathfrak{D}}^{b,i}$ for $i = 1, 2, 3$, we have $\sum_{a=1,2} \sum_j r_{j,a}^2 \geq c\mathfrak{D}^2$ for some positive constant c , which implies the trivial fact

$$\sum_j (r_{j,1}^2 + r_{j,2}^2) \geq \frac{1}{2} \sum_j (r_{j,1}^2 + r_{j,2}^2) + \frac{c}{2} \mathfrak{D}^2. \tag{7.14}$$

Consequently, we can get from (7.12), (7.13) and (7.14) that for some positive constant c ,

$$\exp \{ - M \text{Re} L(\hat{B}, T) \} \leq e^{-cM\mathfrak{D}^2} \prod_{a=1,2} \prod_{j=1}^W e^{-cMr_{j,a}^2} r_{j,a}^M \cdot \mathbb{M}(\mathbf{t})$$

holds in $\mathbf{I}_{\mathfrak{D}}^{b,i}$ for $i = 1, 2, 3$. In addition, by the boundedness of V and \hat{X} -variables, we can get the trivial bound $MK(\hat{X}, V) = O(N)$. Hence, from (7.9) and (7.10) we see that the quantities in Lemma 7.5 (i), (ii) and (iii) can be bounded by the following integral with $i = 1, 2, 3$, respectively,

$$e^{-cM\mathfrak{D}^2 + O(N)} \int_{\mathbf{I}_{\mathfrak{D}}^{b,i}} \prod_{j=1}^W db_{j,1} db_{j,2} \int_{\mathbb{R}_+^{W-1}} \prod_{j=2}^W dt_j \prod_{a=1,2} \prod_{j=1}^W e^{-cMr_{j,a}^2} r_{j,a}^M \mathbb{M}(\mathbf{t}) \tilde{\mathbf{p}}(\mathbf{r}, \mathbf{r}^{-1}, \mathbf{t}). \tag{7.15}$$

According to the facts $\kappa_1 = e^{O(W)}$ and $\kappa_2 = O(1)$ in (7.11), it suffices to consider one monomial, say $\tilde{\mathbf{q}}(\mathbf{r}, \mathbf{r}^{-1}, \mathbf{t})$, instead $\tilde{\mathbf{p}}(\mathbf{r}, \mathbf{r}^{-1}, \mathbf{t})$. Then, bounding t_j 's by 1 trivially in the region $\mathbf{t} \in \mathbb{I}^{W-1}$ and using Lemma 7.2 in the region $\mathbf{t} \in \mathbb{R}_+^{W-1} \setminus \mathbb{I}^{W-1}$, we can first integrate \mathbf{t} -variables out; then performing the Gaussian integral on \hat{B} -variables, it is not difficult to get the estimate

$$(7.15) \leq e^{-cM\mathfrak{D}^2 + O(N \log N)} \rightarrow 0, \text{ as } \mathfrak{D} \rightarrow \infty.$$

for $i = 1, 2, 3$. Here we remark that after integrating \mathbf{t} -variables out, we get a singularity $(\min_{j,a} r_{j,a})^{-cW^2}$ from the factor $(1 + \mathfrak{L}^{-\frac{1}{2}})^{O(W^2)}$ according to (7.6), which can be compensated by the factor $\prod_a \prod_j r_{j,a}^M$ by the fact $M \gg W^2$. Thus we completed the proof. □

7.2 Proof of Lemma 5.6

Plugging the first identity of (5.21) and (7.10) into (7.9), we can write

$$\left| \mathcal{I}(\mathbf{I}_1^b, \mathbf{I}_2^b, \mathbf{I}_1^x, \mathbf{I}_2^x, \mathbf{I}^t, \mathbf{I}^v) \right| \leq e^{O(WN^{\varepsilon_2})} \int_{\mathbf{I}_1^b} \prod_{j=1}^W db_{j,1} \int_{\mathbf{I}_2^b} \prod_{j=1}^W db_{j,2}$$

$$\begin{aligned} & \times \int_{\mathbf{I}_1^x} \prod_{j=1}^W dx_{j,1} \int_{\mathbf{I}_2^x} \prod_{j=1}^W dx_{j,2} \times \int_{\mathbf{I}^t} \prod_{j=2}^W dt_j \int_{\mathbf{I}^v} \prod_{j=2}^W dv_j \\ & \times \exp \left\{ -M \left(\operatorname{Re} \mathring{K}(\hat{X}, V) + \operatorname{Re} \mathring{L}(\hat{B}, T) \right) \right\} \tilde{\mathfrak{p}}(\mathbf{r}, \mathbf{r}^{-1}, \mathbf{t}), \end{aligned} \tag{7.16}$$

where $\tilde{\mathfrak{p}}(\mathbf{r}, \mathbf{r}^{-1}, \mathbf{t})$ is specified in (7.11).

Lemma 5.6 immediately follows from the following two lemmas.

Lemma 7.6 *Under Assumptions 1.1 and 1.14, we have*

$$\mathcal{I}(\Gamma^W, \bar{\Gamma}^W, \Sigma^W, \Sigma^W, \mathbb{R}_+^{W-1} \setminus \mathbb{I}^{W-1}, \mathbb{I}^{W-1}) \leq e^{-\Theta^2}. \tag{7.17}$$

Lemma 7.7 *Under Assumptions 1.1 and 1.14, we have*

$$\begin{aligned} \mathcal{I}(\Gamma^W, \bar{\Gamma}^W, \Sigma^W, \Sigma^W, \mathbb{I}^{W-1}, \mathbb{I}^{W-1}) &= 2^W \mathcal{I}(\gamma_+^b, \gamma_-^b, \gamma_+^x, \gamma_-^x, \gamma_S, \gamma_S) \\ &+ \mathcal{I}(\gamma_+^b, \gamma_-^b, \gamma_+^x, \gamma_-^x, \gamma_S, \mathbb{I}^{W-1}) \\ &+ \mathcal{I}(\gamma_+^b, \gamma_-^b, \gamma_-^x, \gamma_-^x, \gamma_S, \mathbb{I}^{W-1}) + O(e^{-\Theta}). \end{aligned} \tag{7.18}$$

In the sequel, we prove Lemmas 7.6 and 7.7.

Proof of Lemma 7.6. Recall (7.16) with the choice of the integration domains

$$(\mathbf{I}_1^b, \mathbf{I}_2^b, \mathbf{I}_1^x, \mathbf{I}_2^x, \mathbf{I}^t, \mathbf{I}^v) = (\Gamma^W, \bar{\Gamma}^W, \Sigma^W, \Sigma^W, \mathbb{R}_+^{W-1} \setminus \mathbb{I}^{W-1}, \mathbb{I}^{W-1}).$$

To simplify the integral on the r.h.s. of (7.16), we use the fact $\operatorname{Re} \mathring{K}(\hat{X}, V) \geq 0$ implied by (5.27), together with the facts that the \mathbf{x} and \mathbf{v} -variables are bounded by 1. Consequently, we can eliminate the integral over \mathbf{x} and \mathbf{v} -variables from the integral on the r.h.s. of (7.16). Moreover, according to (7.11), it suffices to prove

$$\begin{aligned} & \int_{\Gamma^W} \prod_{j=1}^W db_{j,1} \int_{\bar{\Gamma}^W} \prod_{j=1}^W db_{j,2} \int_{\mathbb{R}_+^{W-1} \setminus \mathbb{I}^{W-1}} \prod_{j=2}^W dt_j \exp \left\{ -M \operatorname{Re} \mathring{L}(\hat{B}, T) \right\} \cdot \tilde{\mathfrak{q}}(\mathbf{r}, \mathbf{r}^{-1}, \mathbf{t}) \\ & \leq e^{-\Theta^2} \end{aligned} \tag{7.19}$$

instead, for some monomial $\tilde{\mathfrak{q}}(\cdot)$ in $\tilde{\mathfrak{p}}(\cdot)$. The proof of (7.19) then is just an application of the first inequality of (5.13), Lemma 7.2 and elementary Gaussian integral. We omit the details here. \square

To prove Lemma 7.7, we split the exponential function into two parts. We use one part to control the integral, and the other will be estimated by its magnitude. More specifically, we shall prove the following two lemmas.

Lemma 7.8 Under Assumptions 1.1 and 1.14, we have

$$\int_{\Gamma^W} \prod_{j=1}^W db_{j,1} \int_{\bar{\Gamma}^W} \prod_{j=1}^W db_{j,2} \int_{\Sigma^W} \prod_{j=1}^W dx_{j,1} \int_{\Sigma^W} \prod_{j=1}^W dx_{j,2} \int_{\mathbb{I}^{W-1}} \prod_{j=2}^W dt_j \int_{\mathbb{I}^{W-1}} \prod_{j=2}^W dv_j \times \exp \left\{ -\frac{1}{2} M(\text{Re}\hat{K}(\hat{X}, V) + \text{Re}\hat{L}(\hat{B}, T)) \right\} \cdot \tilde{p}(\mathbf{r}, \mathbf{r}^{-1}, \mathbf{t}) \leq e^{O(W)}. \tag{7.20}$$

Lemma 7.9 If $(\mathbf{b}_1, \mathbf{b}_2, \mathbf{x}_1, \mathbf{x}_2, \mathbf{t}, \mathbf{v}) \in \Gamma^W \times \bar{\Gamma}^W \times \Sigma^W \times \Sigma^W \times \mathbb{I}^{W-1} \times \mathbb{I}^{W-1}$, but not in any of the Types I, II, III vicinities in Definition 5.5, we have

$$\exp \left\{ -\frac{1}{2} M(\text{Re}\hat{K}(\hat{X}, V) + \text{Re}\hat{L}(\hat{B}, T)) \right\} \leq e^{-\Theta}. \tag{7.21}$$

With Lemmas 7.8 and 7.9, we can prove Lemma 7.7.

Proof of Lemma 7.7. For the sake of simplicity, in this proof, we temporarily use $\mathcal{I}_{\text{full}}$ to represent the l.h.s. of (7.18), i.e. the integral over the full domain, and use $\mathcal{I}_I, \mathcal{I}_{II}$ and \mathcal{I}_{III} to represent the first three terms on the r.h.s. of (7.18). Now, combining (7.16), (7.21) and (7.20), we see that,

$$|\mathcal{I}_{\text{full}} - \mathcal{I}_I - \mathcal{I}_{II} - \mathcal{I}_{III}| \leq e^{O(WN^{\epsilon_2})} \cdot e^{-\Theta} \cdot e^{O(W)} \leq e^{-\Theta},$$

in light of the definition of Θ in (5.30) and the assumption (5.34). Hence, we completed the proof of Lemma 7.7. □

Proof of Lemma 7.8. At first, again, the polynomial $\tilde{p}(\cdot)$ in the integrand can be replaced by some monomial $\tilde{q}(\cdot)$ in the discussion, owing to the fact that $\kappa_1 = \exp\{O(W)\}$ in (7.11). Then, the proof is similar to that of Lemma 7.6, but much simpler, since \mathbf{t} -variables are bounded by 1 now. Consequently, we can eliminate \hat{X} , \mathbf{t} and \mathbf{v} -variables from the integral directly and use the trivial bounds

$$\tilde{q}(\mathbf{r}, \mathbf{r}^{-1}, \mathbf{t}) \leq \prod_{a=1,2} \prod_{j=1}^W r_{j,a}^{\ell_{j,a}}, \quad \text{Re}\hat{L}(\hat{B}, T) \geq c \sum_{a=1,2} \sum_{j=1}^W (r_{j,a} - 1)^2, \tag{7.22}$$

where the latter is from (5.13). Then, an elementary Gaussian integral leads to the conclusion immediately. □

Proof of Lemma 7.9. According to (5.13) and (5.27), we see both $M\text{Re}\hat{L}(\hat{B}, T)$ and $M\text{Re}\hat{K}(\hat{X}, V)$ are nonnegative on the full domain. Hence, it suffices to show one of them is larger than Θ outside the Type I, II, III vicinities. For the former, using the first inequality of (5.13), (7.3) and (7.4), we have

$$\text{Re}\hat{L}(\hat{B}, T) \geq c \sum_{a=1,2} \sum_{j=1}^W (r_{j,a} - 1)^2 + \text{Re}\ell_S(\hat{B}, T)$$

$$\geq c \left(\|\mathbf{b}_1 - a_+\|_2^2 + \|\mathbf{b}_2 + a_-\|_2^2 \right) + \frac{\mathfrak{A}(\hat{B})}{12} (-\mathbf{t}' S^{(1)} \mathbf{t}).$$

Then it is easy to see that $M\text{Re}\hat{L}(\hat{B}, T) \geq \Theta$ if $(\mathbf{b}_1, \mathbf{b}_2, \mathbf{t}) \notin \Upsilon_+^b \times \Upsilon_-^b \times \Upsilon_S$, by the definitions in (5.32).

Now we turn to $M\text{Re}\hat{K}(\hat{X}, V)$. Recall the definition of ϑ_j 's in (5.26). If there is some $j \in \{1, 2W\}$ such that $(\sin \vartheta_j - E/2)^2 > \Theta/M$, we can get $M\text{Re}\hat{K}(\hat{X}, V) > \Theta$ immediately, by using (5.27). Hence, in the sequel, it suffices to consider the case $(\sin \vartheta_j - E/2)^2 \leq \Theta/M$ for all $j = 1, \dots, 2W$, which implies

$$|\arg(a_+^{-1} x_{j,a})|^2 \wedge |\arg(a_-^{-1} x_{j,a})|^2 \leq \frac{\Theta}{M}, \quad \forall j = 1, \dots, W; a = 1, 2. \tag{7.23}$$

Now, recall the notation defined in (5.33). We claim that it suffices to focus on the following three subcases of (7.23), corresponding to the Type I, II and III saddle points.

- (i) There is a sequence of permutations of $\{1, 2\}$, $\epsilon = (\epsilon_1, \dots, \epsilon_W)$, such that $\|\arg(a_+^{-1} \mathbf{x}_{\epsilon(1)})\|_\infty^2 \leq \Theta/M$ and $\|\arg(a_-^{-1} \mathbf{x}_{\epsilon(2)})\|_\infty^2 \leq \Theta/M$,
- (ii) For $a = 1, 2$, there is $\|\arg(a_+^{-1} \mathbf{x}_a)\|_\infty^2 \leq \Theta/M$.
- (iii) For $a = 1, 2$, there is $\|\arg(a_-^{-1} \mathbf{x}_a)\|_\infty^2 \leq \Theta/M$.

For those \hat{X} -variables which satisfy (7.23) but do not belong to any of the case (i), (ii) or (iii) listed above, one can actually get $M\text{Re}\hat{K}(\hat{X}, V) > cM \gg \Theta$. We explain this estimate for the following case, namely, there is some pair $\{i, j\} \in \mathcal{E}$ such that

$$|\arg(a_+^{-1} x_{i,1})|^2, |\arg(a_-^{-1} x_{i,2})|^2 \leq \frac{\Theta}{M}, \quad |\arg(a_-^{-1} x_{j,1})|^2, |\arg(a_-^{-1} x_{j,2})|^2 \leq \frac{\Theta}{M}, \tag{7.24}$$

all the other cases can be handled analogously. Using (5.27), we have

$$\begin{aligned} \text{Re}\hat{K}(\hat{X}, V) &\geq \frac{1}{4} (\mathbb{S}^v)_{ij} (\cos \vartheta_i - \cos \vartheta_j)^2 + \frac{1}{4} (\mathbb{S}^v)_{i+W, j} (\cos \vartheta_{i+W} - \cos \vartheta_j)^2 \\ &\quad + \frac{1}{4} (\mathbb{S}^v)_{i, j+W} (\cos \vartheta_i - \cos \vartheta_{j+W})^2 \\ &\quad + \frac{1}{4} (\mathbb{S}^v)_{i+W, j+W} (\cos \vartheta_{i+W} - \cos \vartheta_{j+W})^2. \end{aligned} \tag{7.25}$$

Now, by the assumption (7.24), we have

$$\cos \vartheta_i = \text{Re}(a_+) + o(1), \quad \cos \vartheta_{i+W}, \cos \vartheta_j, \cos \vartheta_{j+W} = \text{Re}(a_-) + o(1). \tag{7.26}$$

Consequently, from (7.25) and the definition (5.23) we have

$$\text{Re}\hat{K}(\hat{X}, V) \geq c \left((\mathbb{S}^v)_{ij} + (\mathbb{S}^v)_{i, j+W} \right) = c\mathfrak{s}_{ij} > c',$$

where the last step follows from Assumption 1.1 (iv). Therefore, we have $M\text{Re}\hat{K}(\hat{X}, V) > cM$.

Hence, we can focus on the cases (i), (ii) and (iii). Note that in case (i), we actually defined a vicinity of $(\mathbf{x}_{\epsilon(1)}, \mathbf{x}_{\epsilon(2)})$ in terms of the ℓ^∞ -norm rather than ℓ^2 -norm of $\mathbf{x}_{\epsilon(1)}$ and $\mathbf{x}_{\epsilon(2)}$, thus the vicinity is larger than $\mathcal{Y}_+^x \times \mathcal{Y}_-^x$. Without loss of generality, we assume that $\epsilon_1 = \text{id}$ which is the identity. Then by doing the transform $(\hat{X}_j, V_j) \rightarrow (\mathfrak{J}\hat{X}_j\mathfrak{J}, \mathfrak{J}V_j)$ for those j with $\epsilon_j \neq \text{id}$, we can transform $(\mathbf{x}_{\epsilon(1)}, \mathbf{x}_{\epsilon(1)}, \mathbf{v}_\epsilon)$ to $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{v})$ for all permutation sequence ϵ . Hence, it suffices to assume $\epsilon_j = \text{id}$ for all $j = 1, \dots, W$ and consider the case

$$\|a_+^{-1}\mathbf{x}_1\|_\infty < \frac{\Theta}{M}, \quad \|a_-^{-1}\mathbf{x}_2\|_\infty < \frac{\Theta}{M} \quad \text{but} \quad (\mathbf{x}_1, \mathbf{x}_2, \mathbf{v}) \notin \mathcal{Y}_+^x \times \mathcal{Y}_-^x \times \mathcal{Y}_S.$$

Now, if $(\mathbf{x}_1, \mathbf{x}_2) \notin \mathcal{Y}_+^x \times \mathcal{Y}_-^x$, we can use (5.27) again to see that

$$M\text{Re}\hat{K}(\hat{X}, V) \geq cM \sum_{j=1}^{2W} \left(\sin \vartheta_j - \frac{E}{2} \right)^2 \geq cM (\|a_+^{-1}\mathbf{x}_1\|_2^2 + \|a_-^{-1}\mathbf{x}_2\|_2^2) \geq \Theta,$$

where in the second step we used the fact $\|a_+^{-1}\mathbf{x}_1\|_\infty < \Theta/M$ and $\|a_-^{-1}\mathbf{x}_2\|_\infty < \Theta/M$, and in the last step we used the definition in (5.32). Now, if $(\mathbf{x}_1, \mathbf{x}_2) \in \mathcal{Y}_+^x \times \mathcal{Y}_-^x$ but $\mathbf{v} \notin \mathcal{Y}_S$, we go back to (5.20). It is easy to check $-\hat{\ell}_{++}(\mathbf{x}_1) - \hat{\ell}_{--}(\mathbf{x}_2) \geq 0$ for $(\mathbf{x}_1, \mathbf{x}_2) \in \mathcal{Y}_+^x \times \mathcal{Y}_-^x$. Hence it suffices to estimate $\ell_S(\hat{X}, V)$. By the definition in (5.17) and the fact $x_{j,1} - x_{j,2} = a_+ - a_- + o(1)$ for all $j = 1, \dots, W$, we have

$$\begin{aligned} \ell_S(\hat{X}, V) &\geq c \sum_{j,k} \mathfrak{s}_{jk} |(V_k V_j^*)_{12}|^2 = c \sum_{j,k} \mathfrak{s}_{jk} |u_j v_k e^{i\theta_k} - u_k v_j e^{i\theta_j}|^2 \\ &\geq c \sum_{j,k} \mathfrak{s}_{jk} (v_j - v_k)^2, \end{aligned} \tag{7.27}$$

where the last step follows from the same argument as (7.4). Consequently, $M\ell_S(\hat{X}, V) \geq \Theta$ if $\mathbf{v} \notin \mathcal{Y}_S$. Hence, we finished the discussion in case (i). For cases (ii) and (iii), the proofs are much easier since it suffices to discuss the \hat{X} -variables, we just omit the details here. Hence, we completed the proof of Lemma 7.9. \square

7.3 Proof of Lemma 7.2

Let $\mathbb{I}^c = \mathbb{R}_+ \setminus \mathbb{I}$. Now we consider the domain sequence $\mathbb{J} = (\mathbb{J}_2, \dots, \mathbb{J}_W) \in \{\mathbb{I}, \mathbb{I}^c\}^{W-1}$. We decompose the integral in Lemma 7.2 as follows

$$\int_{\mathbb{R}^{W-1} \setminus \mathbb{I}^{W-1}} \prod_{j=2}^W dt_j \mathbb{M}(\mathbf{t})\mathbf{q}(\mathbf{t}) = \sum_{\substack{\mathbb{J} \in \{\mathbb{I}, \mathbb{I}^c\}^{W-1} \\ \mathbb{J} \neq \mathbb{I}^{W-1}}} \int_{\prod_{j=2}^W \mathbb{J}_j} \prod_{j=2}^W dt_j \mathbb{M}(\mathbf{t})\mathbf{q}(\mathbf{t}). \tag{7.28}$$

Note the total number of the choices of such \mathbb{J} in the sum above is $2^{W-1} - 1$. It suffices to consider one of these sequences $\mathbb{J} \in \{\mathbb{I}, \mathbb{I}^c\}^{W-1}$ in which there is at least one i such that $\mathbb{J}_i = \mathbb{I}^c$.

Recall the spanning tree $\mathcal{G}_0 = (\mathcal{V}, \mathcal{E}_0)$ in Assumption 1.1. The simplest case is that there exists a linear spanning tree (a path) \mathcal{G}_0 with

$$\mathcal{E}_0 = \{(i, i + 1)\}_{i=1}^{W-1} \subset \mathcal{E}. \tag{7.29}$$

We first present the proof in this simplest case.

Now, we only keep the edges in the path \mathcal{E}_0 , i.e. the terms with $k = j - 1$ in (7.3), we also trivially discard the term $1/(1 + 2t_j^2)$ from the sum $1/(1 + 2t_{j-1}^2) + 1/(1 + 2t_j^2)$ in the estimate (7.4) (the first inequality), and finally we bound all $M\mathfrak{A}(\hat{B})\mathfrak{s}_{j-1,j}/4$ by \mathfrak{L} defined in (7.5) from below. That means, we use the bound

$$\mathbb{M}(\mathbf{t}) \leq \prod_{j=2}^W \exp \left\{ -\mathfrak{L} \frac{(t_j - t_{j-1})^2}{1 + 2t_{j-1}^2} \right\} := \prod_{j=2}^W \check{\mathbb{M}}_j(\mathbf{t}). \tag{7.30}$$

Consequently, we have

$$\int_{\prod_{j=2}^W \mathbb{J}_j} \prod_{j=2}^W dt_j \mathbb{M}(\mathbf{t})\mathfrak{q}(\mathbf{t}) \leq \int_{\prod_{j=2}^W \mathbb{J}_j} \prod_{j=2}^W dt_j \prod_{j=2}^W t_j^{n_j} \check{\mathbb{M}}_j(\mathbf{t}). \tag{7.31}$$

Note that, as a function of \mathbf{t} , $\check{\mathbb{M}}_j(\mathbf{t})$ only depends on t_{j-1} and t_j .

Having fixed \mathbb{J} , assume that k is the largest index such that $\mathbb{J}_k = \mathbb{I}^c$, i.e. $t_{k+1}, \dots, t_W \in \mathbb{I}$. Then, using the fact $t_1 = 0$ and $t_k \geq 1$, it is not difficult to check

$$\sum_{j=2}^W \frac{(t_j - t_{j-1})^2}{1 + 2t_{j-1}^2} \geq \sum_{j=2}^k \frac{(t_j - t_{j-1})^2}{1 + 2t_{j-1}^2} \geq \frac{1}{300k^2}, \quad \text{if } t_k \in \mathbb{I}^c \tag{7.32}$$

by employing the elementary facts

$$\frac{(t_j - t_{j-1})^2}{1 + 2t_{j-1}^2} \geq \begin{cases} \frac{1}{3}(t_j/t_{j-1} - 1)^2, & \text{if } t_{j-1} \in \mathbb{I}^c \\ \frac{1}{3}(t_j - t_{j-1})^2, & \text{if } t_{j-1} \in \mathbb{I} \end{cases} \quad j = 2, \dots, W. \tag{7.33}$$

Now, we split $\prod_{j=2}^W \check{\mathbb{M}}_j(\mathbf{t})$ into two parts. We use one to control the integral, and the other will be estimated by (7.32). Specifically, substituting (7.32) into (7.31) we have

$$\int_{\prod_{j=2}^W \mathbb{J}_j} \prod_{j=2}^W dt_j \mathbb{M}(\mathbf{t})\mathfrak{q}(\mathbf{t}) \leq e^{-\frac{\mathfrak{L}}{600k^2}} \int_{\mathbb{R}_+^{W-1}} \prod_{j=2}^W dt_j \prod_{j=2}^W t_j^{n_j} (\check{\mathbb{M}}_j(\mathbf{t}))^{\frac{1}{2}}. \tag{7.34}$$

Therefore, what remains is to estimate the integral in (7.34), which can be done by elementary Gaussian integral step by step. More specifically, using (7.33) and the change of variable $t_j/t_{j-1} - 1 \rightarrow t_j$ in case of $t_{j-1} \in \mathbb{I}^c$ and $t_j - t_{j-1} \rightarrow t_j$ in case of $t_{j-1} \in \mathbb{I}$, it is elementary to see that for any $\ell = O(W)$,

$$\int_{\mathbb{R}_+} dt_j t_j^\ell (\check{\mathbb{M}}_j(\mathbf{t}))^{\frac{1}{2}} \leq \ell!! \left(1 + c\mathfrak{L}^{-\frac{1}{2}}\right)^{O(\ell)} \left(t_{j-1}^{\ell+1} + 1\right) \leq e^{O(W \log N)} \left(1 + \mathfrak{L}^{-\frac{1}{2}}\right)^{O(\ell)} \left(t_{j-1}^{\ell+1} + 1\right). \tag{7.35}$$

Starting from $j = W$, using (7.35) to integrate (7.34) successively, the exponent of t_j increases linearly ($n_j = O(1)$), thus we can get

$$\int_{\prod_{j=2}^W \mathbb{J}_j} \prod_{j=2}^W dt_j \mathbb{M}(\mathbf{t})\mathfrak{q}(\mathbf{t}) \leq e^{-\frac{\mathfrak{L}}{600W^2}} \cdot e^{O(W^2 \log N)} \cdot \left(1 + \mathfrak{L}^{-\frac{1}{2}}\right)^{O(W^2)}.$$

Then (7.6) follows from the definition of \mathfrak{L} in (7.5) and (5.35). Hence, we completed the proof for (7.6) when the spanning tree is given by (7.29).

Now, we consider more general spanning tree \mathcal{G}_0 and regard 1 as its root. We start from the generalization of (7.30), namely,

$$\mathbb{M}(\mathbf{t}) \leq \prod_{\{i,j\} \in \mathcal{E}_0} \exp \left\{ -\mathfrak{L} \frac{(t_j - t_i)^2}{1 + 2t_i^2} \right\} := \prod_{\{i,j\} \in \mathcal{E}_0} \check{\mathbb{M}}_{i,j}(\mathbf{t}). \tag{7.36}$$

Here we make the convention that $\text{dist}(1, i) = \text{dist}(1, j) - 1$ for all $\{i, j\} \in \mathcal{E}_0$, where $\text{dist}(a, b)$ represents the distance between a and b . Now, if there is k' such that $\mathbb{J}_{k'} \in \mathbb{I}^c$, we can prove the following analogue of (7.32), namely,

$$\sum_{\{i,j\} \in \mathcal{E}_0} \frac{(t_j - t_i)^2}{1 + 2t_i^2} \geq \frac{1}{300k^2}$$

Consequently, we can get the analogue of (7.34) via replacing $\check{\mathbb{M}}_j(\mathbf{t})$'s by $\check{\mathbb{M}}_{i,j}(\mathbf{t})$'s. Finally, integrating t_j 's out successively, from the leaves to the root 1, yields the same conclusion, i.e. (7.6), for general \mathcal{G}_0 . Therefore, we completed the proof of Lemma 7.2.

8 Gaussian measure in the vicinities

From now on, we can restrict ourselves to the Type I, II and III vicinities. As a preparation of the proofs of Lemmas 5.8 and 5.9, we will show in this section that the exponential function

$$\exp \left\{ -M(\hat{K}(\hat{X}, V) + \hat{L}(\hat{B}, T)) \right\} \tag{8.1}$$

is approximately a Gaussian measure (unnormalized).

8.1 Parametrization and initial approximation in the vicinities

We change the \mathbf{x} , \mathbf{b} , \mathbf{t} , \mathbf{v} -variables to a new set of variables, namely, $\hat{\mathbf{x}}$, $\hat{\mathbf{b}}$, $\hat{\mathbf{t}}$ and $\hat{\mathbf{v}}$. The precise definition of $\hat{\mathbf{x}}$ differs in the different vicinities. To distinguish the parameterization, we set $\varkappa = \pm, +, \text{ or } -$, corresponding to Type I, II or III vicinity, respectively. Recalling D_\varkappa from (1.28). For each j and each \varkappa , we then set

$$\begin{aligned} \hat{X}_j &= D_\varkappa \text{diag} \left(\exp \{i\hat{x}_{j,1}/\sqrt{M}\}, \exp \{i\hat{x}_{j,2}/\sqrt{M}\} \right), \quad \hat{x}_{j,a}/\sqrt{M} \in [-\pi, \pi], \\ \hat{B}_j &= D_\pm + D_\pm \text{diag} \left(\hat{b}_{j,1}/\sqrt{M}, \hat{b}_{j,2}/\sqrt{M} \right), \quad t_j = \hat{t}_j/\sqrt{M}. \end{aligned} \tag{8.2}$$

If $\varkappa = \pm$, we also need to parameterize v_j by

$$v_j = \hat{v}_j/\sqrt{M}. \tag{8.3}$$

Correspondingly, we set the vectors

$$\begin{aligned} \hat{\mathbf{b}}_a &:= (\hat{b}_{1,a}, \dots, \hat{b}_{W,a}), \quad \hat{\mathbf{x}}_a := (\hat{x}_{1,a}, \dots, \hat{x}_{W,a}), \quad a = 1, 2, \\ \hat{\mathbf{t}} &:= (\hat{t}_2, \dots, \hat{t}_W), \quad \hat{\mathbf{v}} := (\hat{v}_2, \dots, \hat{v}_W). \end{aligned}$$

Accordingly, recalling the quantity Θ from (5.30), we introduce the domains

$$\begin{aligned} \hat{\mathcal{Y}} &\equiv \hat{\mathcal{Y}}(N, \varepsilon_0) := \{\mathbf{a} \in \mathbb{R}^W : \|\mathbf{a}\|_2^2 \leq \Theta\}, \\ \hat{\mathcal{Y}}_S &\equiv \hat{\mathcal{Y}}_S(N, \varepsilon_0) := \{\mathbf{a} \in \mathbb{R}_+^{W-1} : -\mathbf{a}'S^{(1)}\mathbf{a} \leq \Theta\}. \end{aligned} \tag{8.4}$$

We remind here, as mentioned above, in the sequel, the small constant ε_0 in $\hat{\mathcal{Y}}$ and $\hat{\mathcal{Y}}_S$ may be different from line to line, subject to (5.34). Now, by the definition of the Type I, II and III vicinities in Definition 5.5 and the parametrization in (8.2) and (8.3), we can redefine the vicinities as follows.

Definition 8.1 We can redefine three types of vicinities as follows.

- Type I' vicinity : $(\hat{\mathbf{b}}_1, \hat{\mathbf{b}}_2, \hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \hat{\mathbf{t}}, \hat{\mathbf{v}}) \in \hat{\mathcal{Y}} \times \hat{\mathcal{Y}} \times \hat{\mathcal{Y}} \times \hat{\mathcal{Y}} \times \hat{\mathcal{Y}}_S \times \hat{\mathcal{Y}}_S$, with $\varkappa = \pm$.
- Type II vicinity : $(\hat{\mathbf{b}}_1, \hat{\mathbf{b}}_2, \hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \hat{\mathbf{t}}, \hat{\mathbf{v}}) \in \hat{\mathcal{Y}} \times \hat{\mathcal{Y}} \times \hat{\mathcal{Y}} \times \hat{\mathcal{Y}} \times \hat{\mathcal{Y}}_S \times \mathbb{I}^{W-1}$, with $\varkappa = +$.
- Type III vicinity : $(\hat{\mathbf{b}}_1, \hat{\mathbf{b}}_2, \hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \hat{\mathbf{t}}, \hat{\mathbf{v}}) \in \hat{\mathcal{Y}} \times \hat{\mathcal{Y}} \times \hat{\mathcal{Y}} \times \hat{\mathcal{Y}} \times \hat{\mathcal{Y}}_S \times \mathbb{I}^{W-1}$, with $\varkappa = -$.

We recall from (7.8) the fact

$$\hat{\mathbf{t}} \in \hat{\mathcal{Y}}_S \implies \|\hat{\mathbf{t}}\|_\infty = O(\Theta). \tag{8.5}$$

Now, we use the representation (5.2). Then, for the Type I vicinity, we change $\mathbf{x}, \mathbf{b}, \mathbf{t}, \mathbf{v}$ -variables to $\hat{\mathbf{x}}, \hat{\mathbf{b}}, \hat{\mathbf{t}}, \hat{\mathbf{v}}$ -variables according to (8.2) with $\kappa = \pm$, thus

$$\begin{aligned}
 2^W \mathcal{I}(\gamma_+^b, \gamma_-^b, \gamma_+^x, \gamma_-^x, \gamma_S, \gamma_S) &= \frac{M^2}{(n!)^2 4^W \pi^{2W+4}} \int_{\mathbb{L}^{2W-2}} \prod_{j=2}^W \frac{d\theta_j}{2\pi} \prod_{j=2}^W \frac{d\sigma_j}{2\pi} \\
 &\times \int_{\hat{\gamma}} \prod_{j=1}^W d\hat{b}_{j,1} \int_{\hat{\gamma}} \prod_{j=1}^W d\hat{b}_{j,2} \int_{\hat{\gamma}} \prod_{j=1}^W d\hat{x}_{j,1} \int_{\hat{\gamma}} \prod_{j=1}^W d\hat{x}_{j,2} \int_{\hat{\gamma}_S} \prod_{j=2}^W 2\hat{t}_j d\hat{t}_j \\
 &\times \hat{t}_j \int_{\hat{\gamma}_S} \prod_{j=2}^W 2\hat{v}_j d\hat{v}_j \prod_{j=1}^W \exp \left\{ \mathbf{i} \frac{\hat{x}_{j,1} + \hat{x}_{j,2}}{\sqrt{M}} \right\} \times \exp \left\{ -M(\hat{K}(\hat{X}, V) + \hat{L}(\hat{B}, T)) \right\} \\
 &\times \prod_{j=1}^W (x_{j,1} - x_{j,2})^2 (b_{j,1} + b_{j,2})^2 \cdot \mathbf{A}(\hat{X}, \hat{B}, V, T). \tag{8.6}
 \end{aligned}$$

For the Type II or III vicinities, i.e. $\kappa = +$ or $-$, we change $\mathbf{x}, \mathbf{b}, \mathbf{t}$ -variables to $\hat{\mathbf{x}}, \hat{\mathbf{b}}, \hat{\mathbf{t}}$ -variables. Consequently, we have

$$\begin{aligned}
 \mathcal{I}(\gamma_+^b, \gamma_-^b, \gamma_\kappa^x, \gamma_\kappa^x, \gamma_S, \mathbb{I}^{W-1}) &= \frac{(-a_\kappa^2)^W}{(n!)^2} \cdot \frac{M^{W+1}}{8^W \pi^{2W+4}} \cdot \int_{\mathbb{L}^{2W-2}} \prod_{j=2}^W \frac{d\theta_j}{2\pi} \prod_{j=2}^W \frac{d\sigma_j}{2\pi} \\
 &\times \int_{\hat{\gamma}} \prod_{j=1}^W d\hat{b}_{j,1} \int_{\hat{\gamma}} \prod_{j=1}^W d\hat{b}_{j,2} \int_{\hat{\gamma}} \prod_{j=1}^W d\hat{x}_{j,1} \int_{\hat{\gamma}} \prod_{j=1}^W d\hat{x}_{j,2} \int_{\hat{\gamma}_S} \prod_{j=2}^W 2\hat{t}_j d\hat{t}_j \\
 &\times \int_{\mathbb{I}^{W-1}} \prod_{j=2}^W 2v_j dv_j \prod_{j=1}^W \exp \left\{ \mathbf{i} \frac{\hat{x}_{j,1} + \hat{x}_{j,2}}{\sqrt{M}} \right\} \exp \left\{ -M(\hat{K}(\hat{X}, V) + \hat{L}(\hat{B}, T)) \right\} \\
 &\times \prod_{j=1}^W (x_{j,1} - x_{j,2})^2 (b_{j,1} + b_{j,2})^2 \cdot \mathbf{A}(\hat{X}, \hat{B}, V, T). \tag{8.7}
 \end{aligned}$$

We will also need the following facts

$$\prod_{j=1}^W |(x_{j,1} - x_{j,2})^2 (b_{j,1} + b_{j,2})^2| = e^{O(W)}, \quad |\mathbf{A}(\hat{X}, \hat{B}, V, T)| \leq e^{O(WN^{\epsilon_2})} \tag{8.8}$$

if $\mathbf{x}_1, \mathbf{x}_2 \in \hat{\Sigma}^W, b_{j,1} = a_+ + o(1), b_{j,2} = -a_- + o(1)$ and $t_j = o(1)$, for all $j = 1, \dots, N$, which always hold in these types of vicinities. The first estimate in (8.8) is trivial, and the second follows from Lemma 6.1.

Now, we approximate (8.1) in the vicinities. We introduce the matrices $\mathcal{E}_+(\vartheta) = \text{diag}(e^{\mathbf{i}\vartheta}, e^{-\mathbf{i}\vartheta})\mathcal{J}$ and $\mathcal{E}_-(\vartheta) = \text{diag}(e^{\mathbf{i}\vartheta}, -e^{-\mathbf{i}\vartheta})\mathcal{J}$ for any $\vartheta \in \mathbb{L}$. Then, with the parameterization above, expanding \hat{X}_j in (3.23) and T_j in (3.25) up to the second

order, we can write

$$T_j = I + \frac{\mathring{i}_j}{\sqrt{M}} \mathcal{E}_+(\sigma_j) + \frac{1}{M} R_j^t, \quad \hat{X}_j = D_\kappa + \frac{\mathbf{i}}{\sqrt{M}} D_\kappa \text{diag}(\mathring{x}_{j,1}, \mathring{x}_{j,2}) + \frac{1}{M} R_j^\kappa, \\ \kappa = \pm, +, -. \tag{8.9}$$

For $\kappa = \pm$, we also expand V_j in (3.25) up to the second order, namely,

$$V_j = I + \frac{\mathring{v}_j}{\sqrt{M}} \mathcal{E}_-(\theta_j) + \frac{1}{M} R_j^v. \tag{8.10}$$

We just take (8.9) and (8.10) as the definition of R_j^κ , R_j^t and R_j^v . Note that R_j^κ is actually κ -dependent. However, this dependence is irrelevant for our analysis thus is suppressed from the notation. It is elementary that

$$\|R_j^\kappa\|_{\max} = O(\mathring{x}_{j,1}^2 + \mathring{x}_{j,2}^2), \quad \|R_j^t\|_{\max} = O(\mathring{i}_j^2), \quad \|R_j^v\|_{\max} = O(\mathring{v}_j^2). \tag{8.11}$$

Recall the facts (5.10) and (5.20). In addition, in light of (5.18)–(5.20), we can also represent $\mathring{K}(\hat{X}, V)$ in the following two alternative ways

$$\mathring{K}(\hat{X}, V) = \left(-\mathring{\ell}_{++}(\mathbf{x}_1) - \mathring{\ell}_{++}(\mathbf{x}_2)\right) + \ell_S(\hat{X}, V) + (K(D_+, I) - K(D_\pm, I)), \tag{8.12}$$

$$\mathring{K}(\hat{X}, V) = \left(-\mathring{\ell}_{+-}(\mathbf{x}_1) - \mathring{\ell}_{+-}(\mathbf{x}_2)\right) + \ell_S(\hat{X}, V) + (K(D_-, I) - K(D_\pm, I)). \tag{8.13}$$

We will use the representations of $\mathring{K}(\hat{X}, V)$ in (8.12) and (8.13) for Type II and III vicinities respectively. In addition, we introduce the matrices

$$\mathbb{A}_+ := (1 + a_+^2)I + a_+^2 S, \quad \mathbb{A}_- := (1 + a_-^2)I + a_-^2 S. \tag{8.14}$$

Then, we have the following lemma.

Lemma 8.2 *With the parametrization in (8.9), we have the following approximations.*

– Let $\mathring{\mathbf{b}}_1, \mathring{\mathbf{b}}_2 \in \mathbb{C}^W$ and $\|\mathring{\mathbf{b}}_1\|_\infty, \|\mathring{\mathbf{b}}_2\|_\infty = o(\sqrt{M})$, we have

$$M \left(\mathring{\ell}_{++}(\mathbf{b}_1) + \mathring{\ell}_{--}(\mathbf{b}_2)\right) = \frac{1}{2} \mathring{\mathbf{b}}_1' \mathbb{A}_+ \mathring{\mathbf{b}}_1 + \frac{1}{2} \mathring{\mathbf{b}}_2' \mathbb{A}_- \mathring{\mathbf{b}}_2 + R^b, \\ R^b = O\left(\frac{\sum_{a=1,2} \|\mathring{\mathbf{b}}_a\|_3^3}{\sqrt{M}}\right). \tag{8.15}$$

– Let $x = \pm$ and $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2 \in \mathbb{C}^W$ and $\|\hat{\mathbf{x}}_1\|_\infty, \|\hat{\mathbf{x}}_2\|_\infty = o(\sqrt{M})$, we have

$$M \left(-\hat{\ell}_{++}(\mathbf{x}_1) - \hat{\ell}_{+-}(\mathbf{x}_2) \right) = \frac{1}{2} \hat{\mathbf{x}}_1' \mathbb{A}_+ \hat{\mathbf{x}}_1 + \frac{1}{2} \hat{\mathbf{x}}_2' \mathbb{A}_- \hat{\mathbf{x}}_2 + R_{\pm}^x, \\ R_{\pm}^x = O \left(\frac{\sum_{a=1,2} \|\hat{\mathbf{x}}_a\|_3^3}{\sqrt{M}} \right). \tag{8.16}$$

– In the Type II vicinity, we have

$$M \left(-\hat{\ell}_{++}(\mathbf{x}_1) - \hat{\ell}_{++}(\mathbf{x}_2) \right) = \frac{1}{2} \hat{\mathbf{x}}_1' \mathbb{A}_+ \hat{\mathbf{x}}_1 + \frac{1}{2} \hat{\mathbf{x}}_2' \mathbb{A}_+ \hat{\mathbf{x}}_2 + R_+^x, \quad R_+^x = O \left(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}} \right). \tag{8.17}$$

– In the Type III vicinity, we have

$$M \left(-\hat{\ell}_{+-}(\mathbf{x}_1) - \hat{\ell}_{+-}(\mathbf{x}_2) \right) = \frac{1}{2} \hat{\mathbf{x}}_1' \mathbb{A}_- \hat{\mathbf{x}}_1 + \frac{1}{2} \hat{\mathbf{x}}_2' \mathbb{A}_- \hat{\mathbf{x}}_2 + R_-^x, \quad R_-^x = O \left(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}} \right). \tag{8.18}$$

Here R_{\pm}^x, R_+^x and R_-^x are remainder terms of the Taylor expansion of the function $\ell(\mathbf{a})$ defined in (5.6).

Proof It follows from the Taylor expansion of the function $\ell(\mathbf{a})$ easily. □

Then, according to (5.10), (5.20), (8.12) and (8.13), what remains is to approximate $M\ell_S(\hat{B}, T)$ and $M\ell_S(\hat{X}, V)$ in the vicinities. Recalling the definition in (5.6) and the parameterization in (8.2), we can rewrite

$$M\ell_S(\hat{B}, T) = \frac{1}{2} \sum_{j,k} \mathfrak{s}_{jk} |s_j \hat{t}_k e^{i\sigma_k} - s_k \hat{t}_j e^{i\sigma_j}|^2 \prod_{\ell=j,k} \left(a_+ - a_- + \frac{a_+ \hat{b}_{\ell,1} - a_- \hat{b}_{\ell,2}}{\sqrt{M}} \right) \\ =: \frac{(a_+ - a_-)^2}{2} \sum_{j,k} \mathfrak{s}_{jk} |\hat{t}_k e^{i\sigma_k} - \hat{t}_j e^{i\sigma_j}|^2 + R^{t,b}. \tag{8.19}$$

We take the above equation as the definition of $R^{t,b}$. Now, we set

$$\tau_{j,1} := \hat{t}_j \cos \sigma_j, \quad \tau_{j,2} := \hat{t}_j \sin \sigma_j, \quad \forall j = 2, \dots, W \tag{8.20}$$

and change the variables and the measure as

$$(\hat{t}_j, \sigma_j) \rightarrow (\tau_{j,1}, \tau_{j,2}), \quad 2\hat{t}_j d\hat{t}_j \frac{d\sigma_j}{2\pi} \rightarrow \frac{1}{\pi} d\tau_{j,1} d\tau_{j,2}. \tag{8.21}$$

In the Type I' vicinity, we can do the same thing for $M\ell_S(\hat{X}, V)$, namely,

$$M\ell_S(\hat{X}, V) =: \frac{(a_+ - a_-)^2}{2} \sum_{j,k} \mathfrak{s}_{jk} |\hat{v}_k e^{i\theta_k} - \hat{v}_j e^{i\theta_j}|^2 + R_{\pm}^{v,x}, \tag{8.22}$$

where $R_{\pm}^{v,x}$ is the remainder term. Then we set

$$v_{j,1} := \hat{v}_j \cos \theta_j, \quad v_{j,2} := \hat{v}_j \sin \theta_j, \quad \forall j = 2, \dots, W \tag{8.23}$$

and change the variables and measure as

$$(\hat{v}_j, \theta_j) \rightarrow (v_{j,1}, v_{j,2}), \quad 2\hat{v}_j d\hat{v}_j \frac{d\theta_j}{2\pi} \rightarrow \frac{1}{\pi} dv_{j,1} dv_{j,2}. \tag{8.24}$$

Now, we introduce the vectors

$$\boldsymbol{\tau}_a = (\tau_{2,a}, \dots, \tau_{W,a}), \quad \boldsymbol{v}_a = (v_{2,a}, \dots, v_{W,a}), \quad a = 1, 2.$$

With this notation, we can rewrite (8.19) and (8.22) as

$$\begin{aligned} M\ell_S(\hat{B}, T) &= -(a_+ - a_-)^2 \sum_{a=1,2} \boldsymbol{\tau}'_a S^{(1)} \boldsymbol{\tau}_a + R^{t,b}, \\ M\ell_S(\hat{X}, V) &= -(a_+ - a_-)^2 \sum_{a=1,2} \boldsymbol{v}'_a S^{(1)} \boldsymbol{v}_a + R_{\pm}^{v,x}. \end{aligned} \tag{8.25}$$

According to (8.21) and (8.24), we can express (8.6) as an integral over $\hat{\mathbf{b}}, \hat{\mathbf{x}}, \hat{\boldsymbol{\tau}}$ and $\hat{\mathbf{v}}$ -variables. However, we need to specify the domains of $\hat{\boldsymbol{\tau}}$ and $\hat{\mathbf{v}}$ -variables in advance. Our aim is to restrict the integral in the domains

$$\boldsymbol{\tau}_a \in \hat{\mathcal{Y}}_S, \quad \boldsymbol{v}_a \in \hat{\mathcal{Y}}_S, \quad a = 1, 2. \tag{8.26}$$

To see the truncation from the domain $(\hat{\mathbf{t}}, \hat{\mathbf{v}}, \boldsymbol{\sigma}, \boldsymbol{\theta}) \in \hat{\mathcal{Y}}_S \times \hat{\mathcal{Y}}_S \times \mathbb{L}^{W-1} \times \mathbb{L}^{W-1}$ to (8.26) is harmless in the integral (8.6), we need to bound $R^{t,b}$ and $R_{\pm}^{v,x}$ in terms of $\boldsymbol{\tau}'_a S^{(1)} \boldsymbol{\tau}_a$ and $\boldsymbol{v}'_a S^{(1)} \boldsymbol{v}_a$, respectively. By (8.5), we have $s_j = 1 + O(\Theta^2/M)$ for all $j = 2, \dots, W$, which also implies

$$\begin{aligned} |s_j \hat{t}_k e^{i\sigma_k} - s_k \hat{t}_j e^{i\sigma_j}|^2 &= \sum_{a=1,2} (\tau_{j,a} - \tau_{k,a})^2 + O\left(\frac{\Theta^3}{M}\right) \sum_{a=1,2} |\tau_{j,a} - \tau_{k,a}| \\ &\quad + O\left(\frac{\Theta^6}{M^2}\right). \end{aligned} \tag{8.27}$$

Then, by the definition of $R^{t,b}$ in (8.19), it is not difficult to check that in the Type I vicinity, we have

$$\begin{aligned}
 |R^{t,b}| &\leq O\left(\frac{\Theta^{\frac{1}{2}}}{\sqrt{M}}\right) \sum_{a=1,2} (-\tau'_a S^{(1)} \tau_a) \\
 &\quad + O\left(\frac{\Theta^{\frac{7}{2}}}{M}\right) \sum_{a=1,2} (-\tau'_a S^{(1)} \tau_a)^{1/2} + O\left(\frac{\Theta^7}{M^2}\right). \tag{8.28}
 \end{aligned}$$

The same estimate holds if we replace $R^{t,b}$ by $R_{\pm}^{v,x}$, τ_a by \mathbf{v}_a . Then it is obvious that if one of $\tau_1, \tau_2, \mathbf{v}_1$ and \mathbf{v}_2 is not in $\hat{\gamma}_S$, we will get (7.21). Hence, using (8.8), we can discard the integral outside the vicinity, analogously to the proof of Lemma 7.7. More specifically, in the sequel, we can and do assume (8.26). Now, plugging (8.26) into (8.28) in turn yields the bound

$$|R^{t,b}| = O\left(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}} \vee \frac{\Theta^4}{M}\right), \quad |R_{\pm}^{v,x}| = O\left(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}} \vee \frac{\Theta^4}{M}\right). \tag{8.29}$$

By the discussion above, for the Type I vicinity, we can write (8.6) as

$$\begin{aligned}
 2^W \mathcal{I}(\gamma_+^b, \gamma_-^b, \gamma_+^x, \gamma_-^x, \gamma_S, \gamma_S) &= \frac{M^2}{(n!)^2 4^W \pi^{4W+2}} \cdot \int_{\hat{\gamma}} \prod_{j=1}^W d\hat{b}_{j,1} \int_{\hat{\gamma}} \prod_{j=1}^W d\hat{b}_{j,2} \\
 &\times \int_{\hat{\gamma}} \prod_{j=1}^W d\hat{x}_{j,1} \int_{\hat{\gamma}} \prod_{j=1}^W d\hat{x}_{j,2} \int_{\hat{\gamma}_S} \prod_{j=2}^W d\tau_{j,1} \int_{\hat{\gamma}_S} \prod_{j=2}^W d\tau_{j,2} \int_{\hat{\gamma}_S} \prod_{j=2}^W d\mathbf{v}_{j,1} \int_{\hat{\gamma}_S} \prod_{j=2}^W d\mathbf{v}_{j,2} \\
 &\times \exp\left\{-\frac{1}{2} \hat{\mathbf{b}}'_1 \mathbb{A}_+ \hat{\mathbf{b}}_1 - \frac{1}{2} \hat{\mathbf{b}}'_2 \mathbb{A}_- \hat{\mathbf{b}}_2 - R^b\right\} \exp\left\{-\frac{1}{2} \hat{\mathbf{x}}'_1 \mathbb{A}_+ \hat{\mathbf{x}}_1 - \frac{1}{2} \hat{\mathbf{x}}'_2 \mathbb{A}_- \hat{\mathbf{x}}_2 - R_{\pm}^x\right\} \\
 &\times \exp\left\{(a_+ - a_-)^2 \sum_{a=1,2} \tau'_a S^{(1)} \tau_a - R^{t,b}\right\} \exp\left\{(a_+ - a_-)^2 \sum_{a=1,2} \mathbf{v}'_a S^{(1)} \mathbf{v}_a - R_{\pm}^{v,x}\right\} \\
 &\times \prod_{j=1}^W \exp\left\{\mathbf{i} \frac{\hat{x}_{j,1} + \hat{x}_{j,2}}{\sqrt{M}}\right\} \prod_{j,k} (x_{j,1} - x_{j,2})^2 (b_{j,1} + b_{j,2})^2 \cdot \mathbf{A}(\hat{X}, \hat{B}, V, T) \\
 &+ O(e^{-\Theta}), \tag{8.30}
 \end{aligned}$$

where the error term stems from the truncation of the vicinity $(\hat{\mathbf{t}}, \hat{\mathbf{v}}, \sigma, \theta) \in \hat{\gamma}_S \times \hat{\gamma}_S \times \mathbb{L}^{W-1} \times \mathbb{L}^{W-1}$ to $(\tau_1, \tau_2, \mathbf{v}_1, \mathbf{v}_2) \in \hat{\gamma}_S \times \hat{\gamma}_S \times \hat{\gamma}_S \times \hat{\gamma}_S$.

Now, for the Type II and III vicinities, the discussion on $\ell_S(\hat{B}, T)$ is of course the same. For $\ell_S(\hat{X}, V)$, we make the following approximation. For the Type II vicinity, using the notation in (5.22), we can write

$$M \ell_S(\hat{X}, V) = \frac{-Ma_+^2}{2} \sum_{j,k} \mathfrak{s}_{jk}^v \prod_{\ell=j,k} \left(\frac{\hat{x}_{\ell,1}}{\sqrt{M}} - \frac{\hat{x}_{\ell,2}}{\sqrt{M}} + O\left(\frac{\hat{x}_{\ell,1}^2 + \hat{x}_{\ell,2}^2}{M}\right) \right)$$

$$=: -\frac{a_+^2}{2} \sum_{j,k} s_{jk}^v (\hat{x}_{j,1} - \hat{x}_{j,2})(\hat{x}_{k,1} - \hat{x}_{k,2}) + R_+^{v,x}. \tag{8.31}$$

It is easy to see that

$$R_+^{v,x} = O\left(\frac{\|\hat{\mathbf{x}}_1\|_3^3 + \|\hat{\mathbf{x}}_2\|_3^3}{\sqrt{M}}\right) = O\left(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}}\right), \text{ for } \hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2 \in \hat{\mathcal{Y}}. \tag{8.32}$$

Recall \mathbb{S}^v defined in (5.23). Set

$$\mathbb{A}_+^v := (1 + a_+^2)I_{2W} + a_+^2\mathbb{S}^v, \quad \mathbb{A}_-^v := (1 + a_-^2)I_{2W} + a_-^2\mathbb{S}^v, \tag{8.33}$$

Combining (5.16), (5.19), (8.17) and (8.31) we obtain

$$\begin{aligned} M(K(\hat{X}, V) - K(D_+, I)) &= \frac{1}{2}\hat{\mathbf{x}}_1' \mathbb{A}_+ \hat{\mathbf{x}}_1 + \frac{1}{2}\hat{\mathbf{x}}_2' \mathbb{A}_+ \hat{\mathbf{x}}_2 \\ &- \frac{a_+^2}{2} \sum_{j,k} s_{jk}^v (\hat{x}_{j,1} - \hat{x}_{j,2})(\hat{x}_{k,1} - \hat{x}_{k,2}) + R_+^x + R_+^{v,x} = \frac{1}{2}\hat{\mathbf{x}}' \mathbb{A}_+^v \hat{\mathbf{x}} + R_+^x + R_+^{v,x}, \end{aligned}$$

where

$$\hat{\mathbf{x}} := (\hat{\mathbf{x}}_1', \hat{\mathbf{x}}_2')'.$$

Analogously, for the Type III vicinity, we can write

$$M(K(\hat{X}, V) - K(D_-, I)) = \frac{1}{2}\hat{\mathbf{x}}' \mathbb{A}_-^v \hat{\mathbf{x}} + R_-^x + R_-^{v,x}, \tag{8.34}$$

where

$$R_-^{v,x} = O\left(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}}\right). \tag{8.35}$$

Consequently, by (8.12) and (8.13) we can write (8.7) for $\kappa = +, -$ as

$$\begin{aligned} \mathcal{I}(\gamma_+^b, \gamma_-^b, \gamma_\times^x, \gamma_\times^x, \gamma_S, \mathbb{I}^{W-1}) &= \exp\left\{M(K(D_\pm, I) - K(D_\times, I))\right\} \\ &\times \frac{(-a_\times^2)^W}{(n!)^2} \cdot \frac{M^{W+1}}{8^W \pi^{3W+3}} \times \int_{\mathbb{L}^{W-1}} \prod_{j=2}^W \frac{d\theta_j}{2\pi} \int_{\mathbb{I}^{W-1}} \prod_{j=2}^W 2v_j dv_j \int_{\hat{\mathcal{Y}}} \prod_{j=1}^W d\hat{b}_{j,1} \\ &\times \int_{\hat{\mathcal{Y}}} \prod_{j=1}^W d\hat{b}_{j,2} \int_{\hat{\mathcal{Y}}} \prod_{j=1}^W d\hat{x}_{j,1} \int_{\hat{\mathcal{Y}}} \prod_{j=1}^W d\hat{x}_{j,2} \times \int_{\hat{\mathcal{Y}}_S} \prod_{j=2}^W d\tau_{j,1} \int_{\hat{\mathcal{Y}}_S} \prod_{j=2}^W d\tau_{j,2} \\ &\times \exp\left\{-\frac{1}{2}\hat{\mathbf{b}}_1' \mathbb{A}_+ \hat{\mathbf{b}}_1 - \frac{1}{2}\hat{\mathbf{b}}_2' \mathbb{A}_- \hat{\mathbf{b}}_2 - R^b\right\} \times \exp\left\{-\frac{1}{2}\hat{\mathbf{x}}' \mathbb{A}_\times^v \hat{\mathbf{x}} - R_\times^x - R_\times^{v,x}\right\} \end{aligned}$$

$$\begin{aligned} & \times \exp \left\{ (a_+ - a_-)^2 \sum_{a=1,2} \tau'_a S^{(1)} \tau_a - R^{t,b} \right\} \times \prod_{j=1}^W \exp \left\{ i \frac{\hat{x}_{j,1} + \hat{x}_{j,2}}{\sqrt{M}} \right\} \\ & \times \prod_{j=1}^W (x_{j,1} - x_{j,2})^2 \times (b_{j,1} + b_{j,2})^2 \mathbf{A}(\hat{X}, \hat{B}, V, T) + O(e^{-\theta}). \end{aligned} \tag{8.36}$$

8.2 Steepest descent paths in the vicinities

In order to estimate the integrals (8.30) and (8.36) properly, we need to control various remainder terms in (8.30) and (8.36) to reduce these integrals to Gaussian ones. The final result is collected in Proposition 8.4 at the end of this section. As a preparation, we shall further deform the contours of $\mathring{\mathbf{b}}$ -variables and $\mathring{\mathbf{x}}$ -variables to the steepest descent paths. We mainly provide the discussion for the $\mathring{\mathbf{b}}$ -variables, that for the $\mathring{\mathbf{x}}$ -variables is analogous.

For simplicity, in this section, we assume $0 \leq E \leq \sqrt{2} - \kappa$, the case $-\sqrt{2} + \kappa \leq E \leq 0$ can be discussed similarly. We introduce the eigendecomposition of S as

$$S = \mathbf{U} \hat{S} \mathbf{U}'.$$

Note that \mathbf{U} is an orthogonal matrix thus the entries are all real. Now, we perform the change of coordinate

$$\mathbf{c}_a = (c_{1,a}, \dots, c_{W,a})' := \mathbf{U}' \mathring{\mathbf{b}}_a, \quad a = 1, 2.$$

Obviously, for the differentials, we have $\prod_{j=1}^W d\mathring{b}_{j,a} = \prod_{j=1}^W dc_{j,a}$ for $a = 1, 2$. In addition, for the domains, it is elementary to see

$$\mathring{\mathbf{b}}_a \in \mathring{\Upsilon} \iff \mathbf{c}_a \in \Upsilon, \quad a = 1, 2. \tag{8.37}$$

Now, we introduce the notation

$$\gamma_j^+ := 1/\sqrt{1 + a_+^2 + a_+^2 \lambda_j(S)}, \quad \gamma_j^- := 1/\sqrt{1 + a_-^2 + a_-^2 \lambda_j(S)},$$

and set the diagonal matrices

$$\mathbb{D}_+ := \text{diag}(\gamma_1^+, \dots, \gamma_W^+), \quad \mathbb{D}_- := \text{diag}(\gamma_1^-, \dots, \gamma_W^-).$$

By the assumption $0 \leq E \leq \sqrt{2} - \kappa$ and (1.5), it is not difficult to check

$$\begin{aligned} |\gamma_j^+| \sim 1, \quad |\gamma_j^-| \sim 1, \quad \arg \gamma_j^+ \in \left(-\frac{\pi}{8}, 0\right], \quad \arg \gamma_j^- \in \left[0, \frac{\pi}{8}\right), \\ \forall j = 1, \dots, W. \end{aligned} \tag{8.38}$$

With the notation introduced above, we have

$$\mathbf{b}'_1 \mathbb{A}_+ \mathbf{b}_1 = \mathbf{c}'_1 \mathbb{D}_+^{-2} \mathbf{c}_1, \quad \mathbf{b}'_2 \mathbb{A}_- \mathbf{b}_2 = \mathbf{c}'_2 \mathbb{D}_-^{-2} \mathbf{c}_2.$$

To simplify the following discussion, we enlarge the domain of the \mathbf{c} -variables to

$$\mathbf{c}_a \in \mathcal{Y}_\infty \equiv \mathcal{Y}_\infty(\varepsilon) := [-\theta^{\frac{1}{2}}, \theta^{\frac{1}{2}}]^W, \quad a = 1, 2. \tag{8.39}$$

Obviously, $\hat{\mathcal{Y}} \subset \mathcal{Y}_\infty$. It is easy to check that (7.21) also holds when $\mathbf{c}_a \in \mathcal{Y}_\infty \setminus \hat{\mathcal{Y}}$ for either $a = 1$ or 2 , according to (8.37), thus such a modification of the domain will only produce an error term of order $O(\exp\{-\theta\})$ in the integral (8.30), by using (8.8).

Now we do the scaling $\mathbf{c}_1 \rightarrow \mathbb{D}_+ \mathbf{c}_1$ and $\mathbf{c}_2 \rightarrow \mathbb{D}_- \mathbf{c}_2$. Consequently, we have

$$\mathbf{b}_1 = \mathbb{U} \mathbb{D}_+ \mathbf{c}_1, \quad \mathbf{b}_2 = \mathbb{U} \mathbb{D}_- \mathbf{c}_2, \tag{8.40}$$

thus

$$\mathbf{b}'_1 \mathbb{A}_+ \mathbf{b}_1 = \sum c_{j,1}^2, \quad \mathbf{b}'_2 \mathbb{A}_- \mathbf{b}_2 = \sum c_{j,2}^2. \tag{8.41}$$

Accordingly, we should adjust the change of differentials as

$$\prod_{j=1}^W db_{j,1} \rightarrow \det \mathbb{D}_+ \cdot \prod_{j=1}^W dc_{j,1}, \quad \prod_{j=1}^W db_{j,2} \rightarrow \det \mathbb{D}_- \cdot \prod_{j=1}^W dc_{j,2}.$$

In addition, the domain of \mathbf{c}_1 should be changed from \mathcal{Y}_∞ to $\prod_{j=1}^W \mathbb{J}_j^+$ and that of \mathbf{c}_2 should be changed from \mathcal{Y}_∞ to $\prod_{j=1}^W \mathbb{J}_j^-$, where

$$\mathbb{J}_j^+ := (\gamma_j^+)^{-1}[-\theta^{\frac{1}{2}}, \theta^{\frac{1}{2}}], \quad \mathbb{J}_j^- := (\gamma_j^-)^{-1}[-\theta^{\frac{1}{2}}, \theta^{\frac{1}{2}}].$$

Now, we consider the integrand in (8.30) as a function of \mathbf{c} -variables on the disks, namely,

$$c_{j,1} \in \mathbb{O}_j^+ := \{z \in \mathbb{C} : |z| \leq \theta^{\frac{1}{2}} |\gamma_j^+|^{-1}\}, \quad c_{j,2} \in \mathbb{O}_j^- := \{z \in \mathbb{C} : |z| \leq \theta^{\frac{1}{2}} |\gamma_j^-|^{-1}\}.$$

For $\mathbf{c}_1 \in \prod_{j=1}^W \mathbb{O}_j^+$ and $\mathbf{c}_2 \in \prod_{j=1}^W \mathbb{O}_j^-$, by (8.38) and (8.40) we have

$$\|\mathbf{b}_1\|_\infty, \|\mathbf{b}_2\|_\infty \leq O(\theta). \tag{8.42}$$

Here we used the elementary fact $\|U\mathbf{a}\|_\infty \leq \sqrt{W} \|\mathbf{a}\|_\infty$ for any $\mathbf{a} \in \mathbb{C}^W$ and unitary matrix U . Then, we deform the contour of $c_{j,1}$ from \mathbb{J}_j^+ to

$$(-\Sigma_j^+) \cup \mathbb{L}_j^+ \cup \Sigma_j^+$$

for each $j = 1, \dots, W$, where

$$\mathbb{L}_j^+ := \mathbb{R} \cap \mathbb{O}_j^+, \quad \Sigma_j^+ = \left\{ z \in \partial\mathbb{O}_j^+ : 0 \leq \arg z \leq -\arg \gamma_j^+ \right\}.$$

It is not difficult to see $\operatorname{Re} c_{j,1}^2 \geq \Theta$ for $c_{j,1} \in (-\Sigma_j^+) \cup \Sigma_j^+$, by (8.38). Consequently, by (8.41), we have

$$\left| \exp \left\{ -\frac{1}{2} \mathring{\mathbf{b}}_1' \mathbb{A}_+ \mathring{\mathbf{b}}_1 \right\} \right| = \left| \exp \left\{ -\frac{1}{2} \sum_{j=1}^W c_{j,1}^2 \right\} \right| \leq O(e^{-\Theta}).$$

Then using (8.8), we can get rid of the integral over Σ_j^+ and $-\Sigma_j^+$, analogously to the discussion in Sect. 7. Similarly, we can perform the same argument for \mathbf{c}_2 . Consequently, we can restrict the \mathbf{c}_1 and \mathbf{c}_2 integral from $\mathbf{c}_1 \in \prod_{j=1}^W \mathbb{J}_j^+$ and $\mathbf{c}_2 \in \prod_{j=1}^W \mathbb{J}_j^-$ to the domain $\mathbf{c}_1 \in \prod_{j=1}^W \mathbb{L}_j^+$ and $\mathbf{c}_2 \in \prod_{j=1}^W \mathbb{L}_j^-$.

By (8.15), (8.42) and the fact $\|\mathbf{a}\|_3^3 \leq \|\mathbf{a}\|_\infty \|\mathbf{a}\|_2^2$ for any vector \mathbf{a} , we see that

$$|R^b| \leq C \frac{\|\mathring{\mathbf{b}}_1\|_\infty + \|\mathring{\mathbf{b}}_2\|_\infty}{\sqrt{M}} (\|\mathring{\mathbf{b}}_1\|_2^2 + \|\mathring{\mathbf{b}}_2\|_2^2) \leq \frac{\Theta}{\sqrt{M}} (\|\mathbf{c}_1\|_2^2 + \|\mathbf{c}_2\|_2^2) \quad (8.43)$$

for some positive constant C , where in the last step we also used the fact that $\|\mathbf{b}_a\|_2 = O(\|\mathbf{c}_a\|_2)$ for $a = 1, 2$, which is implied by (8.40) and (8.38). Consequently, we have

$$\begin{aligned} & \exp \left\{ -\frac{1}{2} \|\mathbf{c}_1\|_2^2 - \frac{1}{2} \|\mathbf{c}_2\|_2^2 - R^b \right\} \\ &= \exp \left\{ -\left(\frac{1}{2} + o(1) \right) \|\mathbf{c}_1\|_2^2 - \left(\frac{1}{2} + o(1) \right) \|\mathbf{c}_2\|_2^2 \right\}. \end{aligned}$$

This allows us to take a step further to truncate \mathbf{c}_1 and \mathbf{c}_2 according to their 2-norm, namely

$$\mathbf{c}_1, \mathbf{c}_2 \in \hat{\mathcal{Y}}. \quad (8.44)$$

Similarly to the discussion in the proof of Lemma 7.7, such a truncation will only produce an error of order $\exp\{-\Theta\}$ to the integral, by using (8.8).

Now, analogously to (8.40), we can change $\hat{\mathbf{x}}$ -variables to \mathbf{d} -variables, defined by

$$\mathbf{d}_1 = (d_{1,1}, \dots, d_{W,1}) := \mathbb{D}_+^{-1} \mathbf{U}' \hat{\mathbf{x}}_1, \quad \mathbf{d}_2 = (d_{1,2}, \dots, d_{W,2}) := \mathbb{D}_-^{-1} \mathbf{U}' \hat{\mathbf{x}}_2.$$

Thus accordingly, we change the differentials

$$\prod_{j=1}^W d\hat{x}_{j,1} \rightarrow \det \mathbb{D}_+ \cdot \prod_{j=1}^W dd_{j,1}, \quad \prod_{j=1}^W d\hat{x}_{j,2} \rightarrow \det \mathbb{D}_- \cdot \prod_{j=1}^W dd_{j,2}.$$

In addition, like (8.44), we deform the domain to $\mathbf{d}_1, \mathbf{d}_2 \in \hat{\mathcal{Y}}$. Finally, using the fact $\det \mathbb{D}_+ \mathbb{D}_- = 1/\sqrt{\det \mathbb{A}_+ \mathbb{A}_-}$, from (8.30) we arrive at the representation

$$\begin{aligned}
 & 2^W \mathcal{I}(\mathcal{Y}_+^b, \mathcal{Y}_-^b, \mathcal{Y}_+^x, \mathcal{Y}_-^x, \mathcal{Y}_S, \mathcal{Y}_S) \\
 &= \frac{M^2}{(n!)^2 4^W \pi^{4W+2}} \cdot \frac{1}{\det \mathbb{A}_+ \mathbb{A}_-} \cdot \int_{\hat{\mathcal{Y}}} \prod_{j=1}^W dc_{j,1} \int_{\hat{\mathcal{Y}}} \prod_{j=1}^W dc_{j,2} \int_{\hat{\mathcal{Y}}} \prod_{j=1}^W dd_{j,1} \int_{\hat{\mathcal{Y}}} \prod_{j=1}^W dd_{j,2} \\
 &\times \int_{\hat{\mathcal{Y}}_S} \prod_{j=2}^W d\tau_{j,1} \int_{\hat{\mathcal{Y}}_S} \prod_{j=2}^W d\tau_{j,2} \int_{\hat{\mathcal{Y}}_S} \prod_{j=2}^W dv_{j,1} \int_{\hat{\mathcal{Y}}_S} \prod_{j=2}^W dv_{j,2} \prod_{j=1}^W \exp \left\{ \mathbf{i} \frac{\hat{x}_{j,1} + \hat{x}_{j,2}}{\sqrt{M}} \right\} \\
 &\times \exp \left\{ -\frac{1}{2} \|\mathbf{c}_1\|_2^2 - \frac{1}{2} \|\mathbf{c}_2\|_2^2 - R^b \right\} \times \exp \left\{ -\frac{1}{2} \|\mathbf{d}_1\|_2^2 - \frac{1}{2} \|\mathbf{d}_2\|_2^2 - R_{\pm}^x \right\} \\
 &\times \exp \left\{ (a_+ - a_-)^2 \sum_{a=1,2} \tau'_a S^{(1)} \tau_a - R^{t,b} \right\} \exp \left\{ (a_+ - a_-)^2 \sum_{a=1,2} v'_a S^{(1)} v_a - R_{\pm}^{v,x} \right\} \\
 &\times \prod_{j=1}^W (x_{j,1} - x_{j,2})^2 (b_{j,1} + b_{j,2})^2 \cdot \mathbf{A}(\hat{X}, \hat{B}, V, T) + O(e^{-\Theta}), \tag{8.45}
 \end{aligned}$$

in which \mathbf{x} and $\hat{\mathbf{x}}$ -variables should be regarded as functions of the \mathbf{d} -variables, as well, \mathbf{b} and $\hat{\mathbf{b}}$ -variables should be regarded as functions of the \mathbf{c} -variables.

Now, in the Type II and III vicinities, we only do the change of coordinates for the $\hat{\mathbf{b}}$ -variables, which is enough for our purpose. Consequently, we have

$$\begin{aligned}
 & \mathcal{I}(\mathcal{Y}_+^b, \mathcal{Y}_-^b, \mathcal{Y}_x^x, \mathcal{Y}_x^x, \mathcal{Y}_S, \mathbb{I}^{W-1}) \\
 &= \exp \left\{ M(K(D_{\pm}, I) - K(D_x, I)) \right\} \times \frac{(-a_x^2)^W}{(n!)^2} \cdot \frac{M^{W+1}}{8^W \pi^{3W+3}} \times \frac{1}{\sqrt{\det \mathbb{A}_+ \mathbb{A}_-}} \\
 &\cdot \int_{\mathbb{I}^{W-1}} \prod_{j=2}^W \frac{d\theta_j}{2\pi} \int_{\mathbb{I}^{W-1}} \prod_{j=2}^W 2v_j dv_j \int_{\hat{\mathcal{Y}}} \prod_{j=1}^W dc_{j,1} \int_{\hat{\mathcal{Y}}} \prod_{j=1}^W dc_{j,2} \int_{\hat{\mathcal{Y}}} \prod_{j=1}^W d\hat{x}_{j,1} \\
 &\times \int_{\hat{\mathcal{Y}}} \prod_{j=1}^W d\hat{x}_{j,2} \int_{\hat{\mathcal{Y}}_S} \prod_{j=2}^W d\tau_{j,1} \int_{\hat{\mathcal{Y}}_S} \prod_{j=2}^W d\tau_{j,2} \prod_{j=1}^W \exp \left\{ \mathbf{i} \frac{\hat{x}_{j,1} + \hat{x}_{j,2}}{\sqrt{M}} \right\} \\
 &\times \exp \left\{ -\frac{1}{2} \|\mathbf{d}_1\|_2^2 - \frac{1}{2} \|\mathbf{d}_2\|_2^2 - R^b \right\} \times \exp \left\{ (a_+ - a_-)^2 \sum_{a=1,2} \tau'_a S^{(1)} \tau_a - R^{t,b} \right\} \\
 &\times \exp \left\{ -\frac{1}{2} \hat{\mathbf{x}}' \mathbb{A}_x^v \hat{\mathbf{x}} - R_x^x - R_x^{v,x} \right\} \times \prod_{j=1}^W (x_{j,1} - x_{j,2})^2 (b_{j,1} + b_{j,2})^2 \\
 &\cdot \mathbf{A}(\hat{X}, \hat{B}, V, T) + O(e^{-\Theta}). \tag{8.46}
 \end{aligned}$$

By (8.38), it is then easy to see

$$\mathbf{c}_1, \mathbf{c}_2 \in \hat{\mathcal{Y}} \implies \hat{\mathbf{b}}_1, \hat{\mathbf{b}}_2 \in \hat{\mathcal{Y}}, \quad \mathbf{d}_1, \mathbf{d}_2 \in \hat{\mathcal{Y}} \implies \hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2 \in \hat{\mathcal{Y}}. \tag{8.47}$$

We keep the terminology “Type I’, II and III vicinities” for the slightly modified domains defined in terms of $\mathbf{c}, \mathbf{d}, \boldsymbol{\tau}$ and \mathbf{v} -variables. More specifically, we redefine the vicinities as follows.

Definition 8.3 We slightly modify Definition 8.1 as follows.

- Type I’ vicinity: $\mathbf{c}_1, \mathbf{c}_2, \mathbf{d}_1, \mathbf{d}_2 \in \hat{\mathcal{Y}}, \boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \mathbf{v}_1, \mathbf{v}_2 \in \hat{\mathcal{Y}}_S$.
- Type II vicinity: $\mathbf{c}_1, \mathbf{c}_2, \hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2 \in \hat{\mathcal{Y}}, \boldsymbol{\tau}_1, \boldsymbol{\tau}_2 \in \hat{\mathcal{Y}}_S, V_j \in \hat{U}(2)$ for all $j = 2, \dots, W$,
 where $\hat{\mathbf{x}}$ – variables are defined in (8.9) with $\kappa = +$.
- Type III vicinity: $\mathbf{c}_1, \mathbf{c}_2, \hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2 \in \hat{\mathcal{Y}}, \boldsymbol{\tau}_1, \boldsymbol{\tau}_2 \in \hat{\mathcal{Y}}_S, V_j \in \hat{U}(2)$ for all $j = 2, \dots, W$,
 where $\hat{\mathbf{x}}$ – variables are defined in (8.9) with $\kappa = -$.

Now, recall the remainder terms R^b, R_{\pm}^x, R_+^x and R_-^x in Lemma 8.2, $R^{t,b}$ and $R_{\pm}^{v,x}$ in (8.25), $R_+^{v,x}$ in (8.31) and $R_-^{v,x}$ in (8.34). In light of (8.47), the bounds on these remainder terms are the same as those obtained in Sect. 8.1. For the convenience of the reader, we collect them as the following proposition.

Proposition 8.4 Under Assumptions 1.1 and 1.14, we have the following estimate, in the vicinities.

$$(i) : R^{t,b} = O\left(\frac{\Theta^4}{M}\right), \quad R_{\pm}^{v,x} = O\left(\frac{\Theta^4}{M}\right), \quad (ii) : R_+^{v,x} = O\left(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}}\right),$$

$$R_-^{v,x} = O\left(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}}\right).$$

$$(iii) : R^b = O\left(\frac{\Theta^2}{\sqrt{M}}\right), \quad R^x = O\left(\frac{\Theta^2}{\sqrt{M}}\right), \quad R_+^x = O\left(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}}\right), \quad R_-^x = O\left(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}}\right).$$

Proof Note that, (i) can be obtained from (8.28), and (ii) is implied by (8.32) and (8.35), and (iii) follows from Lemma 8.2. Hence, we completed the proof. \square

Analogously, in the vicinities, $\|\hat{\mathbf{b}}_1\|_2^2, \|\hat{\mathbf{b}}_2\|_2^2, \|\hat{\mathbf{x}}_1\|_2^2, \|\hat{\mathbf{x}}_2\|_2^2, \|\hat{\mathbf{t}}\|_{\infty}$ and $\|\hat{\mathbf{v}}\|_{\infty}$ are still bounded by Θ .

9 Integral over the Type I vicinities

With (8.45), we estimate the integral over the Type I vicinity in this section. At first, in the Type I’ vicinity, we have $\|\hat{\mathbf{x}}_a\|_{\infty} = O(\Theta^{\frac{1}{2}})$ and $\|\hat{\mathbf{b}}_a\|_{\infty} = O(\Theta^{\frac{1}{2}})$ for $a = 1, 2$. Consequently, according to the parametrization in (8.2), we can get

$$\prod_{j=1}^W (x_{j,1} - x_{j,2})^2 (b_{j,1} + b_{j,2})^2 = (a_+ - a_-)^{4W} \left(1 + O\left(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}}\right) \right). \quad (9.1)$$

Hence, what remains is to estimate the function $\mathbf{A}(\hat{X}, \hat{B}, V, T)$. We have the following lemma.

Lemma 9.1 *Suppose that the assumptions in Theorem 1.15 hold. In the Type I' vicinity, for any given positive integer n , there is $N_0 = N_0(n)$, such that for all $N \geq N_0$ we have*

$$|\mathbf{A}(\hat{X}, \hat{B}, V, T)| \leq \frac{\Theta^2 W^{C_0}}{M(N\eta)^{n+\ell}} \cdot |\det \mathbb{A}_+|^2 \cdot \det(S^{(1)})^2.$$

for some positive constant C_0 and some integer $\ell = O(1)$, both of which are independent of n .

With (8.45), (9.1) and Lemma 9.1, we can prove Lemma 5.8.

Proof of Lemma 5.8. Using (8.45), (9.1), Lemma 9.1, Proposition 8.4 with (5.35), the fact $\det \mathbb{A}_+ = \overline{\det \mathbb{A}_-}$ and the trivial estimate $M\Theta^2 W^{C_0}/(N\eta)^\ell \leq N^{C_0}$ for sufficiently large constant C_0 , we have

$$\begin{aligned} & 2^W |\mathcal{I}(\gamma_+^b, \gamma_-^b, \gamma_+^x, \gamma_-^x, \gamma_S, \gamma_S)| \\ & \leq \frac{N^{C_0}}{(N\eta)^n} \cdot \frac{1}{(2\pi^2)^{2W}} \cdot \det(S^{(1)})^2 \cdot (a_+ - a_-)^{4W} \\ & \quad \times \int_{\hat{\gamma}_+} \prod_{j=1}^W dc_{j,1} \int_{\hat{\gamma}_-} \prod_{j=1}^W dc_{j,2} \int_{\hat{\gamma}_+} \prod_{j=1}^W dd_{j,1} \int_{\hat{\gamma}_-} \prod_{j=1}^W dd_{j,2} \int_{\hat{\gamma}_S} \prod_{j=2}^W d\tau_{j,1} \int_{\hat{\gamma}_S} \prod_{j=2}^W d\tau_{j,2} \\ & \quad \times \int_{\hat{\gamma}_S} \prod_{j=2}^W dv_{j,1} \int_{\hat{\gamma}_S} \prod_{j=2}^W dv_{j,2} \exp \left\{ -\frac{1}{2} (\|\mathbf{c}_1\|_2^2 + \|\mathbf{c}_2\|_2^2 + \|\mathbf{d}_1\|_2^2 + \|\mathbf{d}_2\|_2^2) \right\} \\ & \quad \times \exp \left\{ (a_+ - a_-)^2 \left(\boldsymbol{\tau}'_1 S^{(1)} \boldsymbol{\tau}_1 + \boldsymbol{\tau}'_1 S^{(1)} \boldsymbol{\tau}_2 + \mathbf{v}'_1 S^{(1)} \mathbf{v}_1 + \mathbf{v}'_1 S^{(1)} \mathbf{v}_2 \right) \right\} + O(e^{-\Theta}). \end{aligned}$$

Then, by elementary Gaussian integral we obtain (5.37). Hence, we completed the proof of Lemma 5.8. □

The remaining part of this section will be dedicated to the proof of Lemma 9.1. Recall the definitions of the functions $\mathbf{A}(\cdot)$, $\mathbf{Q}(\cdot)$, $\mathbf{P}(\cdot)$ and $\mathbf{F}(\cdot)$ in (3.32), (4.1), (4.2) and (4.3). Using the strategy in Sect. 6 again, we ignore the irrelevant factor $\mathbf{Q}(\cdot)$ at the beginning. Hence, we bound $\mathbf{P}(\cdot)$ and $\mathbf{F}(\cdot)$ at first, and modify the bounding procedure slightly to take $\mathbf{Q}(\cdot)$ into account in the end, resulting a proof of Lemma 9.1.

9.1 $\mathbf{P}(\hat{X}, \hat{B}, V, T)$ in the Type I' vicinity

Our aim, in this section, is to prove the following lemma.

Lemma 9.2 *Suppose that the assumptions in Theorem 1.15 hold. In the Type I' vicinity, we have*

$$|\mathbf{P}(\hat{X}, \hat{B}, V, T)| \leq \frac{W^{2+\gamma} \Theta^2}{M} |\det \mathbb{A}_+|^2 \det(S^{(1)})^2. \tag{9.2}$$

Before commencing the formal proof, we introduce more notation below. At first, we introduce

$$\hat{\kappa}_j \equiv \hat{\kappa}_j(\hat{X}, \hat{B}, V, T) := |\hat{x}_{j,1}| + |\hat{x}_{j,2}| + |\hat{b}_{j,1}| + |\hat{b}_{j,2}| + |\hat{v}_j| + |\hat{t}_j| = O(\Theta), \tag{9.3}$$

where the bound holds in the Type I' vicinity.

Recalling (6.12) with ϖ_j defined in (6.8) and $\hat{\varpi}_j$ in (6.10). Now, we write

$$\begin{aligned} \varpi_j &= \exp \left\{ -M \log \det \left(1 + M^{-1} V_j^* \hat{X}_j^{-1} V_j \Omega_j T_j^{-1} \hat{B}_j^{-1} T_j \mathcal{E}_j \right) \right\} \\ &=: \exp \left\{ -Tr V_j^* \hat{X}_j^{-1} V_j \Omega_j T_j^{-1} \hat{B}_j^{-1} T_j \mathcal{E}_j \right\} \exp \left\{ \sum_{\ell=2}^4 \frac{(-1)^{\ell-1}}{\ell M^{\ell-1}} \Delta_{\ell,j} \right\}, \end{aligned} \tag{9.4}$$

where

$$\Delta_{\ell,j} := Tr \left(V_j^* \hat{X}_j^{-1} V_j \Omega_j T_j^{-1} \hat{B}_j^{-1} T_j \mathcal{E}_j \right)^\ell. \tag{9.5}$$

The second step of (9.4) follows from the Taylor expansion of the logarithmic function. Now, we expand the first factor of (9.4) around the Type I' saddle point, namely

$$\begin{aligned} &\exp \left\{ -Tr V_j^* \hat{X}_j^{-1} V_j \Omega_j T_j^{-1} \hat{B}_j^{-1} T_j \mathcal{E}_j \right\} \\ &=: \exp \left\{ -Tr D_\pm^{-1} \Omega_j D_\pm^{-1} \mathcal{E}_j \right\} \exp \left\{ -\frac{1}{\sqrt{M}} \Delta_j \right\}. \end{aligned} \tag{9.6}$$

We take (9.6) as the definition of Δ_j , which is of the form

$$\Delta_j = \sum_{\alpha, \beta=1}^4 \hat{\mathfrak{p}}_{j, \alpha, \beta} \cdot \omega_{j, \alpha} \xi_{j, \beta}$$

for some function $\hat{\mathfrak{p}}_{j, \alpha, \beta}$ of $\hat{\mathbf{x}}, \hat{\mathbf{b}}, \hat{\mathbf{v}}$ and $\hat{\mathbf{t}}$ -variables, satisfying

$$\hat{\mathfrak{p}}_{j, \alpha, \beta} = O(\hat{\kappa}_j), \quad \forall \alpha, \beta = 1, \dots, 4. \tag{9.7}$$

One can check (9.7) easily by using (8.9)–(8.11). Analogously, we can also write

$$\Delta_{\ell,j} = \sum_{\substack{\alpha_1, \dots, \alpha_\ell, \\ \beta_1, \dots, \beta_\ell=1}}^4 \hat{\mathfrak{p}}_{\ell,j, \alpha, \beta} \prod_{i=1}^{\ell} \omega_{j, \alpha_i} \xi_{j, \beta_i}, \quad \boldsymbol{\alpha} := (\alpha_1, \dots, \alpha_\ell), \quad \boldsymbol{\beta} := (\beta_1, \dots, \beta_\ell), \tag{9.8}$$

where

$$\hat{\mathfrak{p}}_{\ell,j, \alpha, \beta} = O(1), \quad \forall \ell = 2, \dots, 4; \alpha_1, \dots, \alpha_\ell, \beta_1, \dots, \beta_\ell = 1, \dots, 4. \tag{9.9}$$

The bound on $\hat{p}_{\ell,j,\alpha,\beta}$ in (9.9) follows from the fact that all the $V_j, \hat{X}_j^{-1}, T_j, T_j^{-1}$ and \hat{B}_j^{-1} -entries are bounded in the Type I' vicinity, uniformly in j . Consequently, we can write for $j \neq p, q$

$$\exp \left\{ -\frac{1}{\sqrt{M}} \Delta_j + \sum_{\ell=2}^4 \frac{(-1)^{\ell-1}}{\ell M^{\ell-1}} \Delta_{\ell,j} \right\} = 1 + \sum_{\ell=1}^4 M^{-\frac{\ell}{2}} \sum_{\substack{\alpha_1, \dots, \alpha_\ell, \\ \beta_1, \dots, \beta_\ell=1}}^4 \hat{q}_{\ell,j,\alpha,\beta} \prod_{i=1}^{\ell} \omega_{j,\alpha_i} \xi_{j,\beta_i}. \tag{9.10}$$

In a similar manner, we can also write for $k = p, q$,

$$\begin{aligned} & \exp \left\{ -\frac{1}{\sqrt{M}} \Delta_k + \sum_{\ell=2}^4 \frac{(-1)^{\ell-1}}{\ell M^{\ell-1}} \Delta_{\ell,k} \right\} \hat{\omega}_k \\ &= \hat{p}_0(\cdot) \left(1 + \sum_{\ell=1}^4 M^{-\frac{\ell}{2}} \sum_{\substack{\alpha_1, \dots, \alpha_\ell, \\ \beta_1, \dots, \beta_\ell=1}}^4 \hat{q}_{\ell,k,\alpha,\beta} \prod_{i=1}^{\ell} \omega_{k,\alpha_i} \xi_{k,\beta_i} \right), \end{aligned} \tag{9.11}$$

where $\hat{p}_0(\cdot) = \det \hat{X}_k / \det \hat{B}_k$, which is introduced in (6.10), and $\hat{q}_{\ell,j,\alpha,\beta}$ is some function of \hat{X}, \hat{B}, V and T -variables, satisfying the bound

$$\hat{q}_{\ell,j,\alpha,\beta} = O((1 + \hat{\kappa}_j)^\ell), \quad \forall \ell = 1, \dots, 4, \quad j = 1, \dots, W. \tag{9.12}$$

Obviously, we have $\hat{p}_0(\cdot) = O(1)$ in Type I' vicinity.

Now, in order to distinguish ℓ, α and β for different j , we index them as ℓ_j, α_j and β_j , where

$$\alpha_j \equiv \alpha_j(\ell_j) := (\alpha_{j,1}, \dots, \alpha_{j,\ell_j}), \quad \beta_j \equiv \beta_j(\ell_j) := (\beta_{j,1}, \dots, \beta_{j,\ell_j}).$$

In addition, we define

$$\ell := (\ell_1, \dots, \ell_W), \quad \alpha \equiv \alpha(\ell) := (\alpha_1, \dots, \alpha_W), \quad \beta \equiv \beta(\ell) := (\beta_1, \dots, \beta_W).$$

Let $\|\ell\|_1 = \sum_{j=1}^W \ell_j$ be the 1-norm of ℓ . Note that α and β are $\|\ell\|_1$ -dimensional. With these notations, using (6.12), (9.4), (9.6), (9.10) and (9.11) we have the representation

$$\begin{aligned} & \mathcal{P}(\Omega, \mathcal{E}, \hat{X}, \hat{B}, V, T) \\ &= \hat{p}_0(\hat{X}_p, \hat{B}_p) \hat{p}_0(\hat{X}_q, \hat{B}_q) \times \exp \left\{ -\sum_{j,k} \tilde{s}_{jk} Tr \Omega_j \mathcal{E}_k - \sum_{j=1}^W Tr D_{\pm}^{-1} \Omega_j D_{\pm}^{-1} \mathcal{E}_j \right\} \\ & \times \left(1 + \sum_{\ell \in \llbracket 0,4 \rrbracket^W, \|\ell\|_1 \geq 1} M^{-\frac{\|\ell\|_1}{2}} \sum_{\alpha, \beta \in \llbracket 1,4 \rrbracket^{\|\ell\|_1}} \prod_{j=1}^W \hat{q}_{\ell_j,j,\alpha_j,\beta_j} \cdot \prod_{j=1}^W \prod_{i=1}^{\ell_j} \omega_{j,\alpha_{j,i}} \xi_{j,\beta_{j,i}} \right), \end{aligned} \tag{9.13}$$

where we made the convention

$$\hat{q}_{0,j,\emptyset,\emptyset} = 1, \quad \prod_{i=1}^0 \omega_{j,\alpha_{j,i}} \xi_{j,\beta_{j,i}} = 1, \quad \forall j = 1, \dots, W. \tag{9.14}$$

According to (9.12) and (9.14), we have

$$\prod_{j=1}^W |\hat{q}_{\ell_j,j,\alpha_j,\beta_j}| \leq e^{O(\|\ell\|_1)} \prod_{j=1}^W (1 + \hat{\kappa}_j)^{\ell_j}. \tag{9.15}$$

In addition, we can decompose the sum

$$\sum_{\ell \in \llbracket 0,4 \rrbracket^W, \|\ell\|_1 \geq 1} = \sum_{m=1}^{4W} \sum_{\ell \in \llbracket 0,4 \rrbracket^W, \|\ell\|_1 = m}. \tag{9.16}$$

It is easy to see

$$\#\{\ell \in \llbracket 0,4 \rrbracket^W : \|\ell\|_1 = m\} \leq \binom{4W}{m}. \tag{9.17}$$

Moreover, it is obvious that

$$\#\{\alpha, \beta \in \llbracket 1,4 \rrbracket^{\|\ell\|_1}\} = 16^{\|\ell\|_1}. \tag{9.18}$$

Therefore, it suffices to investigate the integral

$$\begin{aligned} \mathfrak{P}_{\ell,\alpha,\beta} := & \int d\Omega d\mathcal{E} \exp \left\{ - \sum_{j,k} \tilde{s}_{jk} Tr \Omega_j \mathcal{E}_k - \sum_{j=1}^W Tr D_{\pm}^{-1} \Omega_j D_{\pm}^{-1} \mathcal{E}_j \right\} \\ & \times \prod_{j=1}^W \prod_{i=1}^{\ell_j} \omega_{j,\alpha_{j,i}} \xi_{j,\beta_{j,i}} \end{aligned}$$

for each combination (ℓ, α, β) , and then sum it up for (ℓ, α, β) to get the estimate of $P(\hat{X}, \hat{B}, V, T)$. Specifically, we have the following lemma.

Lemma 9.3 *With the notation above, we have*

$$\mathfrak{P}_{\ell,\alpha,\beta} = 0, \quad \text{if } \|\ell\|_1 = 0 \text{ or } 1. \tag{9.19}$$

Moreover, we have

$$|\mathfrak{P}_{\ell,\alpha,\beta}| \leq |\det \mathbb{A}_+|^2 \det(S^{(1)})^2 (\|\ell\|_1 - 1)! (2W^\gamma)^{(\|\ell\|_1 - 1)}, \quad \text{if } \|\ell\|_1 \geq 2. \tag{9.20}$$

We postpone the proof of Lemma 9.3 and prove Lemma 9.2 at first.

Proof of Lemma 9.2. By (4.2), (9.13) and (9.19) and the fact that $\hat{p}_0(\cdot) = O(1)$, we have

$$|\mathbb{P}(\hat{X}, \hat{B}, V, T)| \leq C \sum_{\ell \in \llbracket 0, 4 \rrbracket^W, \|\ell\|_1 \geq 2} M^{-\frac{\|\ell\|_1}{2}} \sum_{\alpha, \beta \in \llbracket 1, 4 \rrbracket^{\|\ell\|_1}} \prod_{j=1}^W |\hat{q}_{\ell_j, j, \alpha_j, \beta_j}| \cdot |\mathfrak{P}_{\ell, \alpha, \beta}|. \tag{9.21}$$

Substituting the bounds (9.15), (9.18) and (9.20) into (9.21) yields

$$\begin{aligned} |\mathbb{P}(\hat{X}, \hat{B}, V, T)| &\leq |\det \mathbb{A}_+|^2 \det(S^{(1)})^2 \\ &\times \sum_{\ell \in \llbracket 0, 4 \rrbracket^W, \|\ell\|_1 \geq 2} e^{O(\|\ell\|_1)} \cdot M^{-\frac{\|\ell\|_1}{2}} \\ &\times (\|\ell\|_1 - 1)! (2W^\gamma)^{(\|\ell\|_1 - 1)} \cdot \prod_{j=1}^W (1 + \hat{\kappa}_j)^{\ell_j}. \end{aligned} \tag{9.22}$$

Now, from (9.3) we have $\prod_{j=1}^W (1 + \hat{\kappa}_j)^{\ell_j} \leq \Theta^{\|\ell\|_1}$, which can absorb the irrelevant factor $e^{O(\|\ell\|_1)}$. Using (9.16), (9.17), we have

$$\begin{aligned} &\sum_{\ell \in \llbracket 0, 4 \rrbracket^W, \|\ell\|_1 \geq 2} e^{O(\|\ell\|_1)} \cdot M^{-\frac{\|\ell\|_1}{2}} \cdot (\|\ell\|_1 - 1)! (2W^\gamma)^{(\|\ell\|_1 - 1)} \cdot \prod_{j=1}^W (1 + \hat{\kappa}_j)^{\ell_j} \\ &\leq \sum_{m=2}^{4W} C^m \binom{4W}{m} \cdot M^{-\frac{m}{2}} \cdot \Theta^m \cdot m! W^{(m-1)\gamma} \\ &\leq \sum_{m=2}^{4W} C^m (4W)^m \cdot M^{-\frac{m}{2}} \cdot \Theta^m \cdot W^{(m-1)\gamma} = O\left(\frac{W^{2+\gamma} \Theta^2}{M}\right), \end{aligned} \tag{9.23}$$

where the last step follows from (5.30) and (5.35). Now, substituting (9.23) into (9.22), we can complete the proof of Lemma 9.2. □

Hence, what remains is to prove Lemma 9.3. We will need the following technical lemma whose proof is postponed.

Lemma 9.4 *For any index sets $I, J \subset \{1, \dots, W\}$ with $|I| = |J| = m \geq 1$, we have the following bounds for the determinants of the submatrices of S, \mathbb{A}_+ and \mathbb{A}_- defined in (8.14).*

– For $(\mathbb{A}_+)^{(I|J)}$ and $(\mathbb{A}_-)^{(I|J)}$, we have

$$|\det(\mathbb{A}_+)^{(I|J)}| / |\det \mathbb{A}_+| \leq 1, \quad |\det(\mathbb{A}_-)^{(I|J)}| / |\det \mathbb{A}_-| \leq 1. \tag{9.24}$$

– For $S^{(\mathbb{I}, \mathbb{J})}$, we have

$$|\det S^{(\mathbb{I}, \mathbb{J})}| / |\det S^{(1)}| \leq (m - 1)!(2W^\gamma)^{(m-1)}. \tag{9.25}$$

Proof of Lemma 9.3. Recall the definition in (6.16). Furthermore, we introduce the matrix

$$\mathbb{H} = (a_+^{-2} \mathbb{A}_+) \oplus S \oplus S \oplus (a_-^{-2} \mathbb{A}_-). \tag{9.26}$$

Using the fact $a_+ a_- = -1$, we can write

$$-\sum_{j,k} \tilde{s}_{jk} \text{Tr} \Omega_j \mathcal{E}_k - \sum_{j=1}^W \text{Tr} D_{\pm}^{-1} \Omega_j D_{\pm}^{-1} \mathcal{E}_j = -\mathbf{\Omega} \mathbb{H} \mathbf{\Xi}'.$$

Now, by the Gaussian integral of the Grassmann variables in (3.2), we see that

$$|\mathfrak{P}_{\ell, \alpha, \beta}| = |\det \mathbb{H}^{(\mathbb{I}, \mathbb{J})}| \tag{9.27}$$

for some index sets $\mathbb{I}, \mathbb{J} \subset \{1, \dots, 4W\}$ determined by α and β such that

$$|\mathbb{I}| = |\mathbb{J}| = \|\ell\|_1.$$

Here we mention that (9.27) may fail when at least two components in α_j coincide for some j . But $\mathfrak{P}_{\ell, \alpha, \beta} = 0$ in this case because of $\chi^2 = 0$ for any Grassmann variable χ .

Now, obviously, there exists index sets $\mathbb{I}_\alpha, \mathbb{J}_\alpha \subset \{1, \dots, W\}$ for $\alpha = 1, \dots, 4$ such that

$$\begin{aligned} \mathbb{H}^{(\mathbb{I}, \mathbb{J})} &= (a_+^{-2} \mathbb{A}_+)^{(\mathbb{I}_1 | \mathbb{J}_1)} \oplus S^{(\mathbb{I}_2 | \mathbb{J}_2)} \oplus S^{(\mathbb{I}_3 | \mathbb{J}_3)} \oplus (a_-^{-2} \mathbb{A}_-)^{(\mathbb{I}_4 | \mathbb{J}_4)}, \\ \sum_{\alpha} |\mathbb{I}_\alpha| &= \sum_{\alpha} |\mathbb{J}_\alpha| = \|\ell\|_1. \end{aligned}$$

It suffices to consider the case $|\mathbb{I}_\alpha| = |\mathbb{J}_\alpha|$ for all $\alpha = 1, 2, 3, 4$. Otherwise, $\det \mathbb{H}^{(\mathbb{I}, \mathbb{J})}$ is obviously 0, in light of the block structure of \mathbb{H} , see the definition (9.26). Now, note that, since $\det S = 0$, we have

$$\det \mathbb{H}^{(\mathbb{I}, \mathbb{J})} = 0, \quad \text{if } \|\ell\|_1 = 0, 1.$$

For more general ℓ , by Lemma 9.4, we have

$$|\det \mathbb{H}^{(\mathbb{I}, \mathbb{J})}| \leq |\det \mathbb{A}_+ \mathbb{A}_-| \det(S^{(1)})^2 (\|\ell\|_1 - 1)! (2W^\gamma)^{(\|\ell\|_1 - 1)}.$$

Then, by the fact $|\det \mathbb{A}_+ \mathbb{A}_-| = |\det \mathbb{A}_+|^2$, we can conclude the proof of Lemma 9.3. □

To prove Lemma 9.4, we will need the following lemma.

Lemma 9.5 *For the weighted Laplacian S , we have*

$$\det S^{(i|j)} = (-1)^{j-i} \det S^{(i)}, \quad \forall i, j = 1, \dots, W \tag{9.28}$$

Remark 9.6 A direct consequence of (9.28) is $\det S^{(1)} = \dots = \det S^{(W)}$.

Proof of Lemma 9.5. Without loss of generality, we assume $j > i$ in the sequel. We introduce the matrices

$$P_{ij} := I_{i-1} \oplus \begin{pmatrix} I_{j-i-1} \\ 1 \end{pmatrix} \oplus I_{W-j}, \quad E_j := I - 2\mathbf{e}_j\mathbf{e}_j^* - \sum_{\ell \neq j} \mathbf{e}_\ell\mathbf{e}_\ell^*.$$

It is not difficult to check

$$S^{(i|j)} = S^{(i)} P_{ij} E_j. \tag{9.29}$$

Then, by the fact $\det P_{ij} E_j = (-1)^{j-i}$, we can get the conclusion. □

Proof of Lemma 9.4. At first, by the definition in (8.14), (1.5) and the fact $\text{Re}a_+^2 = \text{Re}a_-^2 > 0$, it is easy to see that the singular values of \mathbb{A}_+ and \mathbb{A}_- are all larger than 1. With the aid of the rectangular matrix $(\mathbb{A}_+)^{(|I|)}$ as an intermediate matrix, we can use Cauchy interlacing property twice to see that the k th largest singular value of $(\mathbb{A}_+)^{(|J|)}$ is always smaller than the k th largest singular value of \mathbb{A}_+ . Consequently, we have the first inequality of (9.24). In the same manner, we can get the second inequality of (9.24)

Now, we prove (9.25). At first, we address the case that $I \cap J \neq \emptyset$. In light of Remark 9.6, without loss of generality, we assume that $1 \in I \cap J$. Then $S^{(|J|)}$ is a submatrix of $S^{(1)}$. Therefore, we can find two permutation matrices P and Q , such that

$$P S^{(1)} Q = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

where $D = S^{(|J|)}$. Now, by Schur complement, we know that

$$|\det S^{(|J|)}| / |\det S^{(1)}| = |\det(A - B D^{-1} C)^{-1}|.$$

Moreover, $(A - B D^{-1} C)^{-1}$ is the $(|I| - 1)$ by $(|I| - 1)$ upper-left corner of $(P S^{(1)} Q)^{-1} = Q^{-1} (S^{(1)})^{-1} P^{-1}$. That means $\det S^{(|J|)} / \det S^{(1)}$ is the determinant of a sub matrix of $(S^{(1)})^{-1}$ (with dimension $|I| - 1$), up to a sign. Then, by Assumption 1.1 (iii), we can easily get

$$|\det S^{(|J|)}| / |\det S^{(1)}| \leq (|I| - 1)! W^{(|I|-1)\gamma}.$$

Now, for the case $I \cap J = \emptyset$, we can fix one $i \in I$ and $j \in J$. Due to (9.28), it suffices to consider

$$\det S^{(I \cup J)} / \det S^{(i|j)}. \tag{9.30}$$

By similar discussion, one can see that (9.30) is the determinant of a sub matrix of $(S^{(i|j)})^{-1}$ with dimension $\|I\| - 1$. Hence, it suffices to investigate the bound of the entries of $(S^{(i|j)})^{-1}$. From (9.29) we have

$$(S^{(i|j)})^{-1} = E_j^{-1} P_{ij}^{-1} (S^{(i)})^{-1}. \tag{9.31}$$

Then, it is elementary to see that the entries of $(S^{(i|j)})^{-1}$ are bounded by $2W^\gamma$, in light of Assumption 1.1 (iii). Consequently, we have

$$|\det S^{(I \cup J)}| / |\det S^{(i|j)}| \leq (\|I\| - 1)! (2W^\gamma)^{\|I\| - 1},$$

which implies (9.25). Hence, we completed the proof of Lemma 9.4. □

9.2 $F(\hat{X}, \hat{B}, V, T)$ in the Type I' vicinity

Neglecting the $X^{[1]}$, $\mathbf{y}^{[1]}$ and $\mathbf{w}^{[1]}$ -variables in $\mathbf{Q}(\cdot)$ at first, we investigate the integral $F(\hat{X}, \hat{B}, V, T)$ in the Type I' vicinity in this section. We have the following lemma.

Lemma 9.7 *Suppose that the assumptions in Theorem 1.15 hold. In the Type I' vicinity, we have*

$$F(\hat{X}, \hat{B}, V, T) = O\left(\frac{1}{(N\eta)^{n+2}}\right). \tag{9.32}$$

Recalling the functions $\mathbb{G}(\hat{B}, T)$ and $\mathbb{F}(\hat{X}, V)$ defined in (6.20) and (6.21), we further introduce

$$\begin{aligned} \mathring{\mathbb{G}}(\hat{B}, T) &= \exp\{(a_+ - a_-)N\eta\} \mathbb{G}(\hat{B}, T), \quad \mathring{\mathbb{F}}(\hat{X}, V) \\ &= \exp\{- (a_+ - a_-)N\eta\} \mathbb{F}(\hat{X}, V). \end{aligned} \tag{9.33}$$

Then, we have the decomposition

$$F(\hat{X}, \hat{B}, V, T) = \mathring{\mathbb{G}}(\hat{B}, T) \mathring{\mathbb{F}}(\hat{X}, V). \tag{9.34}$$

Hence, we can estimate $\mathring{\mathbb{F}}(\hat{X}, V)$ and $\mathring{\mathbb{G}}(\hat{B}, T)$ separately in the sequel.

9.2.1 Estimate of $\mathring{\mathbb{F}}(\hat{X}, V)$

Lemma 9.8 *Suppose that the assumptions in Theorem 1.15 hold. In the Type I' vicinity, we have*

$$\mathring{\mathbb{F}}(\hat{X}, V) = O\left(\frac{1}{N\eta}\right). \tag{9.35}$$

Proof Using (8.9) and (8.10), we can write

$$X_j = P_1^* V_j^* \hat{X}_j V_j P_1 = P_1^* D_{\pm} P_1 + O\left(\frac{\Theta}{\sqrt{M}}\right), \tag{9.36}$$

where the remainder term represents a 2×2 matrix whose max-norm is bounded by Θ/\sqrt{M} . Using (9.36) and recalling $N = MW$ yields

$$\exp\left\{M\eta \sum_{j=1}^W \text{Tr} X_j J\right\} = \exp\left\{N\eta \text{Tr} P_1^* D_{\pm} P_1 J\right\} \left(1 + O\left(\frac{\Theta N\eta}{\sqrt{M}}\right)\right). \tag{9.37}$$

Substituting (9.37) into (3.20) and (6.21), we can write

$$\begin{aligned} \mathbb{F}(\hat{X}, V) &= \int d\mu(P_1) dX^{[1]} \exp\left\{N\eta \text{Tr} P_1^* D_{\pm} P_1 J\right\} \prod_{k=p,q} \frac{1}{\det^2(X_k^{[1]})} \\ &\times \prod_{k=p,q} \exp\left\{i \text{Tr} X_k^{[1]} J Z - \sum_j \tilde{s}_{jk} \text{Tr} X_j X_k^{[1]} J\right\} \\ &\times \prod_{k,\ell=p,q} \exp\left\{\frac{\tilde{s}_{k\ell}}{2M} \text{Tr} X_k^{[1]} J X_{\ell}^{[1]} J\right\} \cdot \left(1 + O\left(\frac{\Theta N\eta}{\sqrt{M}}\right)\right). \end{aligned}$$

Recalling the parametrization of P_1 in (3.25), we have

$$\text{Tr} P_1^* D_{\pm} P_1 J = (1 - 2v^2)(a_+ - a_-).$$

Consequently, we have

$$\begin{aligned} \mathring{\mathbb{F}}(\hat{X}, V) &= \int dX^{[1]} \int v dv \int \frac{d\theta}{\pi} \exp\left\{-2(a_+ - a_-)N\eta v^2\right\} \prod_{k=p,q} \frac{1}{\det^2(X_k^{[1]})} \\ &\times \prod_{k=p,q} \exp\left\{i \text{Tr} X_k^{[1]} J Z - \sum_j \tilde{s}_{jk} \text{Tr} X_j X_k^{[1]} J\right\} (1 + o(1)). \end{aligned}$$

By the fact that $X^{[1]}$ -variables are all bounded and $|\det X_k^{[1]}| = 1$ for $k = p, q$, it is easy to see that

$$|\mathring{\mathbb{F}}(\hat{X}, V)| \leq C \int_0^1 v dv \exp\left\{-2(a_+ - a_-)N\eta v^2\right\} = O\left(\frac{1}{N\eta}\right).$$

Therefore, we completed the proof. □

9.2.2 Estimate of $\mathring{\mathbb{G}}(\hat{B}, T)$

Recall the definition of $\mathring{\mathbb{G}}(\hat{B}, T)$ from (9.33), (6.20) and (3.21). In this section, we will prove the following lemma.

Lemma 9.9 *Suppose that the assumptions in Theorem 1.15 hold. In the Type I' vicinity, we have*

$$\mathring{\mathbb{G}}(\hat{B}, T) = O\left(\frac{1}{(N\eta)^{n+1}}\right). \tag{9.38}$$

Note that $y_p^{[1]}$, $y_q^{[1]}$ and t in the parametrization of Q_1 (see (3.25)) are not bounded, we shall truncate them with some appropriate bounds at first, whereby we can neglect some irrelevant terms in the integrand, in order to simplify the integral. More specifically, we will do the truncations

$$t \leq (N\eta)^{-1/4}, \quad y_p^{[1]}, y_q^{[1]} \leq (N\eta)^{\frac{1}{8}}. \tag{9.39}$$

Accordingly, we set

$$\begin{aligned} \widehat{\mathbb{G}}(\hat{B}, T) := & e^{(a_+ - a_-)N\eta} \int_{\mathbb{L}} \frac{d\sigma}{2\pi} \int_{\mathbb{L}^2} v_p^{[1]} v_q^{[1]} dv_p^{[1]} dv_q^{[1]} \int_0^{(N\eta)^{\frac{1}{8}}} dy_p^{[1]} \int_0^{(N\eta)^{\frac{1}{8}}} dy_q^{[1]} \\ & \times \int_0^{(N\eta)^{-\frac{1}{4}}} 2t dt \int_{\mathbb{L}^2} d\sigma_p^{[1]} d\sigma_q^{[1]} g(Q_1, T, \hat{B}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}), \end{aligned} \tag{9.40}$$

where we have used the parameterization of $\mathbf{w}^{[1]}$ in (3.15). We will prove the following lemma.

Lemma 9.10 *Suppose that the assumptions in Theorem 1.15 hold. In the Type I' vicinity, we have*

$$\mathring{\mathbb{G}}(\hat{B}, T) = \widehat{\mathbb{G}}(\hat{B}, T) + O(e^{-N^\varepsilon})$$

for some positive constant ε .

Proof At first, by (6.26)–(6.29), we have for any j ,

$$\text{Re} Tr B_j Y_k^{[1]} J \geq \frac{y_k^{[1]}}{(s+t)^2} \cdot \frac{\min\{\text{Re} b_{j,1}, \text{Re} b_{j,2}\}}{(s_j+t_j)^2} \geq c \frac{y_k^{[1]}}{1+2t^2}, \quad k = p, q, \tag{9.41}$$

for some positive constant c , where the last step follows from the facts that $\text{Re} b_{j,1}, \text{Re} b_{j,2} = \text{Re} a_+ + o(1)$ and $t_j = o(1)$ in the Type I' vicinity. In addition, it is not difficult to get

$$\text{Tr} B_j J = \left(a_+ - a_- + O\left(\frac{\Theta}{\sqrt{M}}\right)\right)(1+2t^2), \quad \forall j = 1, \dots, W,$$

which implies that

$$M\eta \sum_{j=1}^W \text{Tr} B_j J = (a_+ - a_-)N\eta + 2\left(a_+ - a_- + O\left(\frac{\Theta}{\sqrt{M}}\right)\right)N\eta t^2 + O\left(\frac{\Theta N\eta}{\sqrt{M}}\right). \tag{9.42}$$

Note that the second and third factors in the definition of $g(\cdot)$ in (3.21) can be bounded by 1, according to (6.23). Then, as a consequence of (9.41) and (9.42), we have

$$e^{(a_+ - a_-)N\eta} |g(\cdot)| \leq C(y_p^{[1]} y_q^{[1]})^{n+3} \exp\{-c' N\eta t^2\} \exp\left\{-c \frac{y_p^{[1]} + y_q^{[1]}}{1 + 2t^2}\right\}, \tag{9.43}$$

for some positive constants C, c and c' . By integrating $y_p^{[1]}$ and $y_q^{[1]}$ out at first, we can easily see that the first truncation in (9.39) only produces an error of order $O(\exp\{-N^\varepsilon\})$ to the integral $\mathring{\mathbb{G}}(\hat{B}, T)$, for some positive constant $\varepsilon = \varepsilon(\varepsilon_2)$ by the assumption $\eta \geq N^{-1+\varepsilon_2}$ in (1.16). Then one can substitute the first bound in (9.39) to the last factor of the r.h.s. of (9.43), thus

$$\exp\left\{-c \frac{y_p^{[1]} + y_q^{[1]}}{1 + 2t^2}\right\} \leq \exp\left\{-\frac{c}{2}(y_p^{[1]} + y_q^{[1]})\right\}.$$

We can also do the second truncation in (9.39) in the integral $\mathring{\mathbb{G}}(\hat{B}, T)$, up to an error of order $O(\exp\{-N^\varepsilon\})$, for some positive constant ε . Therefore, we completed the proof of Lemma 9.10. □

With the aid of Lemma 9.10, it suffices to work on $\widehat{\mathbb{G}}(\hat{B}, T)$ in the sequel. We have the following lemma.

Lemma 9.11 *We have*

$$\widehat{\mathbb{G}}(\hat{B}, T) = O\left(\frac{1}{(N\eta)^{n+1}}\right).$$

Proof of Lemma 9.11. Recall the parameterization of $\mathbf{w}_k^{[1]}$ in (3.15) again. To simplify the notation, we set

$$\mathbf{w}_k^{[1]} = u_k^{[1]} v_k^{[1]}, \quad k = p, q.$$

Similarly to (9.36), using $t = o(1)$ from (9.39), we have the expansion

$$B_j = Q_1^{-1} T_j^{-1} \hat{B}_j T_j Q_1 = Q_1^{-1} D_\pm Q_1 + O\left(\frac{\Theta}{\sqrt{M}}\right).$$

Consequently, we have

$$-M\eta \sum_{j=1}^W \text{Tr} B_j J = -N\eta(a_+ - a_-)(1 + 2t^2) + O\left(\frac{\Theta N\eta}{\sqrt{M}}\right). \tag{9.44}$$

In addition, for $k = p, q$, using the fact $\sum_j \tilde{s}_{jk} = 1$, we have

$$\begin{aligned} \sum_j \tilde{s}_{jk} Tr B_j Y_k^{[1]} J &= y_k^{[1]} \left((a_+ - a_-) t^2 + (a_+ (u_k^{[1]})^2 - a_- (v_k^{[1]})^2) \right) \\ &+ y_k^{[1]} \left((a_+ - a_-) (e^{-i(\sigma_k^{[1]} + \sigma)} + e^{i(\sigma_k^{[1]} + \sigma)}) \mathfrak{w}_k^{[1]} s t \right) + \frac{\Theta}{\sqrt{M}} Tr R_k Y_k^{[1]}, \end{aligned} \tag{9.45}$$

where R_k is a 2×2 matrix independent of $Y_k^{[1]}$, satisfying $\|R_k\|_{\max} = O(1)$.

Observe that the term in (9.44) is obviously independent of $\mathbf{w}^{[1]}$ -variables. In addition, for $k = p$ or q , we have

$$i Tr Y_k^{[1]} J Z = (-\eta + iE(1 - 2(v_k^{[1]})^2)) y_k^{[1]}, \tag{9.46}$$

and for $k, \ell = p$ or q , we have

$$Tr Y_k^{[1]} J Y_\ell^{[1]} J = y_k^{[1]} y_\ell^{[1]} \left((\mathfrak{w}_k^{[1]} \mathfrak{w}_\ell^{[1]})^2 + \mathfrak{w}_k^{[1]} \mathfrak{w}_\ell^{[1]} \left(e^{i(\sigma_k^{[1]} - \sigma_\ell^{[1]})} - e^{i(\sigma_\ell^{[1]} - \sigma_k^{[1]})} \right) \right). \tag{9.47}$$

Moreover, we have

$$\left(\left(\mathbf{w}_q^{[1]} (\mathbf{w}_q^{[1]})^* \right)_{12} \left(\mathbf{w}_p^{[1]} (\mathbf{w}_p^{[1]})^* \right)_{21} \right)^n = \left(\mathfrak{w}_p^{[1]} \mathfrak{w}_q^{[1]} \right)^n e^{in(\sigma_p^{[1]} - \sigma_q^{[1]})}. \tag{9.48}$$

Substituting (9.44), (9.45) and (9.46)–(9.48) to the definition of $g(\cdot)$ in (3.21) and reordering the factors properly, we can write the integrand in (9.40) as

$$\begin{aligned} \exp\{(a_+ - a_-)N\eta\} g(\cdot) &= \exp\{in(\sigma_p^{[1]} - \sigma_q^{[1]})\} \\ &\times \exp \left\{ -(a_+ - a_-)st \sum_{k=p,q} y_k^{[1]} \mathfrak{w}_k^{[1]} \left(e^{-i(\sigma_k^{[1]} + \sigma)} + e^{i(\sigma_k^{[1]} + \sigma)} \right) \right\} \\ &\times \exp \left\{ -\frac{\Theta}{\sqrt{M}} \sum_{k=p,q} Tr R_k Y_k^{[1]} \right\} \\ &\times \exp \left\{ -\frac{1}{M} \tilde{s}_{pq} y_p^{[1]} y_q^{[1]} \mathfrak{w}_p^{[1]} \mathfrak{w}_q^{[1]} \left(e^{i(\sigma_p^{[1]} - \sigma_q^{[1]})} - e^{i(\sigma_q^{[1]} - \sigma_p^{[1]})} \right) \right\} \\ &\times \prod_{k=p,q} (y_k^{[1]})^{n+3} (\mathfrak{w}_k^{[1]})^n \cdot \prod_{k,\ell=p,q} \exp \left\{ -\frac{1}{2M} \tilde{s}_{k\ell} y_k^{[1]} y_\ell^{[1]} \left(\mathfrak{w}_k^{[1]} \mathfrak{w}_\ell^{[1]} \right)^2 \right\} \\ &\times \exp \left\{ -2N\eta(a_+ - a_-)t^2 \right\} \prod_{k=p,q} \exp \left\{ -y_k^{[1]} \left((a_+ (u_k^{[1]})^2 - a_- (v_k^{[1]})^2) \right) \right. \\ &\left. + (a_+ - a_-)t^2 + \eta - iE \left(1 - 2(v_k^{[1]})^2 \right) \right\} \times \left(1 + O\left(\frac{\Theta N \eta}{\sqrt{M}}\right) \right), \end{aligned} \tag{9.49}$$

where the last factor is independent of the $\mathbf{w}^{[1]}$ -variables. Here, we put the factors containing $\sigma_p^{[1]}$ and $\sigma_q^{[1]}$ together, namely, the first two lines on the r.h.s. of (9.49). For further discussion, we write for $k = p, q$

$$Tr R_k Y_k^{[1]} = y_k^{[1]} (\mathfrak{r}_k^+ e^{i\sigma_k} + \mathfrak{r}_k^- e^{-i\sigma_k} + \mathfrak{r}_k), \tag{9.50}$$

where $\mathfrak{r}_k^+, \mathfrak{r}_k^-$ and \mathfrak{r}_k are all polynomials of $u_k^{[1]}$ and $v_k^{[1]}$, with bounded degree and bounded coefficients, in light of $\|R_k\|_{\max} = O(1)$, the definition of $Y_k^{[1]}$ in (3.14) and the parametrization in (3.15).

Now, we start to estimate the integral (9.40) by using (9.49). We deal with the integral over $\sigma_p^{[1]}$ and $\sigma_q^{[1]}$ at first. These variables are collected in the integral of the form

$$\begin{aligned} \mathcal{I}_\sigma(\ell_1, \ell_2) &:= \int_{\mathbb{L}^2} d\sigma_p^{[1]} d\sigma_q^{[1]} \exp\{\mathbf{i}(n + \ell_1)\sigma_p^{[1]}\} \exp\{-\mathbf{i}(n + \ell_2)\sigma_q^{[1]}\} \\ &\times \exp\left\{-\frac{\Theta}{\sqrt{M}} \sum_{k=p,q} Tr R_k Y_k^{[1]}\right\} \\ &\times \exp\left\{-(a_+ - a_-)st \sum_{k=p,q} y_k^{[1]} \mathfrak{w}_k^{[1]} (e^{-\mathbf{i}(\sigma_k^{[1]} + \sigma)} + e^{\mathbf{i}(\sigma_k^{[1]} + \sigma)})\right\} \\ &\times \exp\left\{-\frac{1}{M} \tilde{s}_{pq} y_p^{[1]} y_q^{[1]} \mathfrak{w}_p^{[1]} \mathfrak{w}_q^{[1]} (e^{i(\sigma_p^{[1]} - \sigma_q^{[1]})} - e^{i(\sigma_q^{[1]} - \sigma_p^{[1]})})\right\} \end{aligned}$$

with integers ℓ_1 and ℓ_2 independent of n . Note that according to (9.49), it suffices to consider $\mathcal{I}_\sigma(0, 0)$ for the proof of (9.38). We study $\mathcal{I}_\sigma(\ell_1, \ell_2)$ for general ℓ_1 and ℓ_2 here, which will be used later.

Now, we set

$$\begin{aligned} c_{p,q} &:= \tilde{s}_{pq} y_p^{[1]} y_q^{[1]} \mathfrak{w}_p^{[1]} \mathfrak{w}_q^{[1]}, \\ c_{k,1} &:= -(a_+ - a_-)st y_k^{[1]} \mathfrak{w}_k^{[1]} e^{-i\sigma} - \frac{\Theta}{\sqrt{M}} y_k^{[1]} \mathfrak{r}_k^-, \quad k = p, q, \\ c_{k,2} &:= -(a_+ - a_-)st y_k^{[1]} \mathfrak{w}_k^{[1]} e^{i\sigma} - \frac{\Theta}{\sqrt{M}} y_k^{[1]} \mathfrak{r}_k^+, \quad k = p, q. \end{aligned} \tag{9.51}$$

In addition, we introduce

$$d_{p,q} := y_p^{[1]} y_q^{[1]}, \quad d_k := \left(t + \frac{\Theta}{\sqrt{M}}\right) y_k^{[1]}, \quad k = p, q. \tag{9.52}$$

Obviously, when (9.39) is satisfied, we have

$$c_{p,q} = O(d_{p,q}), \quad c_{k,1} = O(d_k), \quad c_{k,2} = O(d_k), \quad k = p, q. \tag{9.53}$$

With the aid of the notation defined in (9.50) and (9.51), we can write

$$\begin{aligned} \mathcal{I}_\sigma(\ell_1, \ell_2) &= \exp \left\{ -\frac{\Theta}{\sqrt{M}}(y_p^{[1]}\tau_p + y_q^{[1]}\tau_q) \right\} \int_{\mathbb{L}^2} d\sigma_p d\sigma_q \\ &\quad \times \exp \left\{ \mathbf{i}(n + \ell_1)\sigma_p^{[1]} \right\} \exp \left\{ -\mathbf{i}(n + \ell_2)\sigma_q^{[1]} \right\} \\ &\quad \times \prod_{k=p,q} \exp \left\{ c_{k,1}e^{-\mathbf{i}\sigma_k^{[1]}} + c_{k,2}e^{\mathbf{i}\sigma_k^{[1]}} \right\} \\ &\quad \times \exp \left\{ -\frac{c_{p,q}}{M}e^{\mathbf{i}(\sigma_p^{[1]}-\sigma_q^{[1]})} + \frac{c_{p,q}}{M}e^{\mathbf{i}(\sigma_q^{[1]}-\sigma_p^{[1]})} \right\}. \end{aligned} \tag{9.54}$$

We have the following lemma.

Lemma 9.12 *Under the truncation (9.39), we have*

$$|\mathcal{I}_\sigma(\ell_1, \ell_2)| \leq C \left(\left(\frac{d_{p,q}}{M} \right)^{n+\ell_3} + d_p^{2(n+\ell_3)} + d_q^{2(n+\ell_3)} \right), \quad \ell_3 := \frac{\ell_1 + \ell_2}{2}$$

for some positive constant C .

Proof At first, by Taylor expansion, we have

$$\begin{aligned} &\exp \left\{ \mathbf{i}(n + \ell_1)\sigma_p^{[1]} \right\} \exp \left\{ -\mathbf{i}(n + \ell_2)\sigma_q^{[1]} \right\} \exp \left\{ -\frac{c_{p,q}}{M}e^{\mathbf{i}(\sigma_p^{[1]}-\sigma_q^{[1]})} + \frac{c_{p,q}}{M}e^{\mathbf{i}(\sigma_q^{[1]}-\sigma_p^{[1]})} \right\} \\ &= \sum_{n_1, n_2=0}^{\infty} \frac{(-1)^{n_1}}{(n_1)!(n_2)!} \left(\frac{c_{p,q}}{M} \right)^{n_1+n_2} \exp \left\{ \mathbf{i}(n + \ell_1 + n_1 - n_2)\sigma_p^{[1]} \right\} \\ &\quad \times \exp \left\{ -\mathbf{i}(n + \ell_2 + n_1 - n_2)\sigma_q^{[1]} \right\}. \end{aligned} \tag{9.55}$$

Now, for any $m_1, m_2 \in \mathbb{Z}$, we denote

$$\begin{aligned} \tilde{\mathcal{I}}_\sigma(m_1, m_2) &:= \int_{\mathbb{L}^2} d\sigma_p^{[1]} d\sigma_q^{[1]} \exp \{ \mathbf{i}m_1\sigma_p^{[1]} \} \exp \{ -\mathbf{i}m_2\sigma_q^{[1]} \} \\ &\quad \times \prod_{k=p,q} \exp \left\{ c_{k,1}e^{-\mathbf{i}\sigma_k^{[1]}} + c_{k,2}e^{\mathbf{i}\sigma_k^{[1]}} \right\} \\ &= 4\pi^2 \sum_{n_3=0}^{\infty} \mathbf{1}(n_3 + m_1 \geq 0) \frac{(c_{p,1})^{n_3+m_1} (c_{p,2})^{n_3}}{n_3!(n_3 + m_1)!} \\ &\quad \times \sum_{n_4=0}^{\infty} \mathbf{1}(n_4 + m_2 \geq 0) \frac{(c_{q,1})^{n_4} (c_{q,2})^{n_4+m_2}}{n_4!(n_4 + m_2)!}. \end{aligned} \tag{9.56}$$

Setting

$$m_1 := n + \ell_1 + n_1 - n_2, \quad m_2 := n + \ell_2 + n_1 - n_2, \tag{9.57}$$

and using (9.55), we can rewrite (9.54) as

$$\mathcal{I}_\sigma(\ell_1, \ell_2) = \exp \left\{ -\frac{\Theta}{\sqrt{M}}(y_p^{[1]}\mathbf{r}_p + y_q^{[1]}\mathbf{r}_q) \right\} \sum_{n_1, n_2=0}^\infty \frac{(-1)^{n_1}}{(n_1)!(n_2)!} \left(\frac{c_{p,q}}{M}\right)^{n_1+n_2} \tilde{\mathcal{I}}_\sigma(m_1, m_2). \tag{9.58}$$

For simplicity, we employ the notation

$$m_3 := m_3(\ell_1, n_1, n_2, n_3) = m_1 + n_3, \quad m_4 := m_4(\ell_2, n_1, n_2, n_4) = m_2 + n_4. \tag{9.59}$$

Consequently, by (9.58) and (9.56) we obtain

$$\begin{aligned} |\mathcal{I}_\sigma(\ell_1, \ell_2)| &\leq 4\pi^2 \left| \exp \left\{ -\frac{\Theta}{\sqrt{M}}(y_p^{[1]}\mathbf{r}_p + y_q^{[1]}\mathbf{r}_q) \right\} \right| \sum_{n_1, n_2=0}^\infty \frac{1}{(n_1)!(n_2)!} \left| \frac{c_{p,q}}{M} \right|^{n_1+n_2} \\ &\quad \times \sum_{n_3=0}^\infty \mathbf{1}(m_3 \geq 0) \frac{|c_{p,1}|^{m_3} |c_{p,2}|^{n_3}}{n_3! m_3!} \cdot \sum_{n_4=0}^\infty \mathbf{1}(m_4 \geq 0) \frac{|c_{q,1}|^{n_4} |c_{q,2}|^{m_4}}{n_4! m_4!} \\ &\leq C \max_{n_1, n_2, n_3, n_4 \in \mathbb{N}} \left| \left(\frac{c_{p,q}}{M}\right)^{n_1+n_2} (c_{p,1})^{m_3} (c_{p,2})^{n_3} (c_{q,1})^{n_4} (c_{q,2})^{m_4} \right| \\ &\leq C \max_{n_1, n_2, n_3, n_4 \in \mathbb{N}} \left| \left(\frac{c_{p,q}}{M}\right)^{n_1+n_2} (c_{p,1})^{m_3} (c_{q,2})^{m_4} \right| \end{aligned} \tag{9.60}$$

for some positive constant C , where in the last step we used the fact $|c_{k,1}| < 1, |c_{k,2}| < 1$, which can be seen directly from the definition in (9.53), the truncations in (9.39) and the assumption $\eta \leq M^{-1}N^{\varepsilon_2}$. Analogously, we also have $|c_{p,q}/M| < 1$. According to the definitions (9.57) and (9.59), we have

$$2(n_1 + n_2) + m_3 + m_4 \geq 2n + \ell_1 + \ell_2.$$

Hence, by using $|c_{k,1}| < 1, |c_{k,2}| < 1$ and $|c_{p,q}/M| < 1$, we have the trivial bound

$$\begin{aligned} &\max_{n_1, n_2, n_3, n_4 \geq 0} \left| \left(\sqrt{\frac{|c_{p,q}|}{M}}\right)^{2(n_1+n_2)} (c_{p,1})^{m_3} (c_{q,2})^{m_4} \right| \\ &\leq \left(\frac{|c_{p,q}|}{M}\right)^{n+\ell_3} + |c_{p,1}|^{2(n+\ell_3)} + |c_{q,2}|^{2(n+\ell_3)}. \end{aligned}$$

Therefore, we completed the proof by using (9.53). □

Now, we return to the proof of Lemma 9.11. Using (9.49) and Lemma 9.12 with $\ell_1 = \ell_2 = 0$ to (9.40), and integrating the bounded variables $v_p^{[1]}, v_q^{[1]}$ and σ out, we

can get

$$\begin{aligned}
 |\widehat{\mathbb{G}}(\widehat{B}, T)| &\leq C \int_0^{(N\eta)^{\frac{1}{8}}} dy_p^{[1]} \int_0^{(N\eta)^{\frac{1}{8}}} dy_q^{[1]} \int_0^{(N\eta)^{-\frac{1}{4}}} 2t dt \cdot (d_{p,q})^{n+3} \\
 &\quad \times \left(\left(\frac{d_{p,q}}{M} \right)^n + d_p^{2n} + d_q^{2n} \right) \exp \{ -2N\eta(a_+ - a_-)t^2 \} \\
 &\quad \times \exp \left\{ -\frac{\sqrt{4 - E^2}}{2} \sum_{k=p,q} y_k^{[1]} \right\} (1 + o(1))
 \end{aligned}$$

where the last two factors come from the facts

$$\begin{aligned}
 &\left| \exp \left\{ -\sum_{k=p,q} y_k^{[1]} \left(a_+ (u_k^{[1]})^2 - a_- (v_k^{[1]})^2 \right) \right\} \right| = \exp \left\{ -\frac{\sqrt{4 - E^2}}{2} \sum_{k=p,q} y_k^{[1]} \right\}, \\
 &\left| \exp \left\{ -\sum_{k=p,q} y_k^{[1]} \left((a_+ - a_-)t^2 + \eta - \mathbf{i}E \left(1 - 2(v_k^{[1]})^2 \right) \right) \right\} \left(1 + O \left(\frac{\Theta N \eta}{\sqrt{M}} \right) \right) \right| \\
 &= 1 + o(1). \tag{9.61}
 \end{aligned}$$

In (9.61) we used the fact $(u_k^{[1]})^2 + (v_k^{[1]})^2 = 1$. Now, we integrate $y_p^{[1]}$ and $y_q^{[1]}$ out. Consequently, by the definition in (9.52), we have

$$\begin{aligned}
 |\widehat{\mathbb{G}}(\widehat{B}, T)| &\leq C \int_0^{(N\eta)^{-\frac{1}{4}}} 2t dt \left(\frac{1}{M^n} + \left(\frac{\Theta}{\sqrt{M}} \right)^{2n} + t^{2n} \right) \exp \{ -2N\eta(a_+ - a_-)t^2 \} \\
 &= O \left(\frac{1}{(N\eta)^{n+1}} \right),
 \end{aligned}$$

where in the last step we have used the assumption $\eta \leq M^{-1}N^{\varepsilon_2}$ in (1.16), Assumption 1.14, the definition of Θ in (5.30) and the fact $N = MW$. Hence, we completed the proof of Lemma 9.11. □

Finally, we can prove Lemma 9.9, and further prove Lemma 9.7.

Proof of Lemma 9.9. This is a direct consequence of Lemmas 9.10 and 9.11. □

Proof of Lemma 9.7. This is a direct consequence of (9.34), Lemma 9.8 and Lemma 9.9. □

9.3 Summing up: Proof of Lemma 9.1

In this section, we slightly modify the discussions in Sects. 9.1 and 9.2 to prove Lemma 9.1. The combination of Lemmas 9.2 and 9.7 would directly imply Lemma 9.1 if the $\mathbb{Q}(\cdot)$ factor were not present in the definition of $\mathbb{A}(\cdot)$. Now we should take $\mathbb{Q}(\cdot)$ into account. This argument is similar to the corresponding discussion in Sect. 6.4.

Proof of Lemma 9.1. At first, we observe that κ_1, κ_2 and κ_3 in (6.5) are obviously independent of n . Then, by the fact $\kappa_1 = W^{O(1)}$, it suffices to consider one monomial of the form

$$p_1\left(t, s, (y_p^{[1]})^{-1}, (y_q^{[1]})^{-1}\right) p_2\left(\left\{e^{i\sigma_k^{[1]}}, e^{-i\sigma_k^{[1]}}\right\}_{k=p,q}\right) q\left(\left\{\frac{\omega_{i,a}\xi_{j,b}}{M}\right\}_{i,j=1,\dots,W}, a,b=1,\dots,4\right),$$

where the degrees of $p_1(\cdot), p_2(\cdot)$ and $q(\cdot)$ are all $O(1)$, and independent of n , in light of the fact $\kappa_3 = O(1)$ in (6.5). Especially, the order of $(y_p^{[1]})^{-1}$ and $(y_q^{[1]})^{-1}$ are not larger than 2, which can be easily seen from the definition of $\mathcal{Q}(\cdot)$ in (3.19).

Now, we reuse the notation $P_q(\hat{X}, \hat{B}, V, T)$ and $F_p(\hat{X}, \hat{B}, V, T)$ in (6.36), by redefining them as

$$P_q(\hat{X}, \hat{B}, V, T) := \int d\Omega d\mathcal{E} \mathcal{P}(\cdot) \cdot q\left(\left\{\frac{\omega_{i,a}\xi_{j,b}}{M}\right\}_{i,j=1,\dots,W}, a,b=1,\dots,4\right),$$

$$F_p(\hat{X}, \hat{B}, V, T) := \int dX^{[1]} dy^{[1]} d\mathbf{w}^{[1]} d\mu(P_1) d\nu(Q_1) \mathcal{F}(\cdot) \\ \times p_1\left(t, s, (y_p^{[1]})^{-1}, (y_q^{[1]})^{-1}\right) p_2\left(\left\{e^{i\sigma_k^{[1]}}, e^{-i\sigma_k^{[1]}}\right\}_{k=p,q}\right).$$

It is easy to check $\mathcal{P}(\cdot)q(\cdot)$ also has an expansion of the form in (9.13). Hence, the bound in (9.2) holds for $P_q(\cdot)$ as well. For $F_p(\cdot)$, the main modification is to use Lemma 9.12 with general ℓ_1 and ℓ_2 independent of n , owing to the function $p_2(\cdot)$. In addition, by the truncations in (9.39), we can bound $p_1(\cdot)$ by some constant C . Hence, it suffices to replace n by $n + \ell_3$ in the proof of Lemma 9.11. Finally, we can get

$$F_p(\hat{X}, \hat{B}, V, T) = O\left(\frac{1}{(N\eta)^{n+\ell_3}}\right),$$

with some finite integer ℓ_3 independent of n . Consequently, we completed the proof of Lemma 9.1. □

10 Integral over the Type II and III vicinities

In this section, we prove Lemma 5.9. We only present the discussion for $\mathcal{I}(\mathcal{Y}_+^b, \mathcal{Y}_+^b, \mathcal{Y}_+^x, \mathcal{Y}_+^x, \mathcal{Y}_S, \mathbb{I}^{W-1})$, i.e. integral over the Type II vicinity. The discussion on $\mathcal{I}(\mathcal{Y}_+^b, \mathcal{Y}_-^b, \mathcal{Y}_-^x, \mathcal{Y}_-^x, \mathcal{Y}_S, \mathbb{I}^{W-1})$ is analogous. We start from (8.46). Similarly, we shall provide an estimate for the integrand. At first, under the parameterization (8.2) with $\varkappa = +$, we see that

$$\prod_{j=1}^W (x_{j,1} - x_{j,2})^2 (b_{j,1} + b_{j,2})^2 = \frac{(-a_+^2)^W}{M^W} (a_+ - a_-)^{2W} \left(1 + O\left(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}}\right)\right) \\ \times \prod_{j=1}^W \left(\hat{x}_{j,1} - \hat{x}_{j,2} + O\left(\frac{\Theta}{\sqrt{M}}\right)\right)^2. \quad (10.1)$$

Then, what remains is to estimate $\mathbf{A}(\hat{X}, \hat{B}, V, T)$. Our aim is to prove the following lemma.

Lemma 10.1 *Suppose that the assumptions in Theorem 1.15 hold. In the Type II vicinity, we have*

$$|\mathbf{A}(\hat{X}, \hat{B}, V, T)| \leq e^{-cN\eta} |\det \mathbb{A}_+|^2 \det(S^{(1)})^2 \tag{10.2}$$

for some positive constant c .

With the aid of (10.1) and Lemma 10.1, we can prove Lemma 5.9.

Proof of Lemma 5.9. Recall (8.46). At first, by the definition of \mathbb{A}_+^v in (8.33), (5.24) and the fact $\text{Re}a_+^2 > 0$, we can see that

$$\text{Re}(\hat{\mathbf{x}}' \mathbb{A}_+^v \hat{\mathbf{x}}) \geq \|\hat{\mathbf{x}}\|_2^2 \tag{10.3}$$

for all $\{V_j\}_{j=2}^W \in (\hat{U}^{(2)})^{W-1}$. Substituting (5.21), (10.1), (10.2), (10.3) and the estimates in Proposition 8.4 into (8.46) yields

$$\begin{aligned} & |\mathcal{I}(\gamma_+^b, \gamma_-^b, \gamma_+^x, \gamma_-^x, \gamma_S, \mathbb{I}^{W-1})| \\ & \leq e^{-cN\eta} \cdot \frac{(a_+ - a_-)^{2W}}{8^W \pi^{3W-1}} \cdot |\det S^{(1)}|^2 \cdot |\det \mathbb{A}_+| \cdot \int_{\mathbb{L}^{W-1}} \prod_{j=2}^W \frac{d\theta_j}{2\pi} \int_{\mathbb{I}^{W-1}} \prod_{j=2}^W 2v_j dv_j \\ & \quad \times \int_{\mathbb{R}^{W-1}} \prod_{j=2}^W d\tau_{j,1} \int_{\mathbb{R}^{W-1}} \prod_{j=2}^W d\tau_{j,2} \int_{\mathbb{R}^W} \prod_{j=1}^W dc_{j,1} \int_{\mathbb{R}^W} \prod_{j=1}^W dc_{j,2} \int_{\mathbb{R}^W} \prod_{j=1}^W d\hat{x}_{j,1} \\ & \quad \times \int_{\mathbb{R}^W} \prod_{j=1}^W d\hat{x}_{j,2} \exp\{(a_+ - a_-)^2 \boldsymbol{\tau}'_1 S^{(1)} \boldsymbol{\tau}_1\} \exp\{(a_+ - a_-)^2 \boldsymbol{\tau}'_2 S^{(1)} \boldsymbol{\tau}_2\} \\ & \quad \times \exp\left\{-\frac{1}{2} \|\hat{\mathbf{c}}_1\|_2^2 - \frac{1}{2} \|\hat{\mathbf{c}}_2\|_2^2\right\} \exp\left\{-\frac{1}{2} \|\hat{\mathbf{x}}_1\|_2^2 - \frac{1}{2} \|\hat{\mathbf{x}}_2\|_2^2\right\} \\ & \quad \times \prod_{j=1}^W \left(\hat{x}_{j,1} - \hat{x}_{j,2} + O\left(\frac{\Theta}{\sqrt{M}}\right)\right)^2, \end{aligned}$$

where we absorbed several factors by $\exp\{-cN\eta\}$. We also enlarged the domains to the full ones. Then, using the trivial facts

$$\int_{\mathbb{L}^{W-1}} \prod_{j=2}^W \frac{d\theta_j}{2\pi} \int_{\mathbb{I}^{W-1}} \prod_{j=2}^W 2v_j dv_j = 1$$

and performing the Gaussian integral for the remaining variables, we can get

$$|\mathcal{I}(\gamma_+^b, \gamma_-^b, \gamma_+^x, \gamma_-^x, \gamma_S, \mathbb{I}^{W-1})| \leq C |\det S^{(1)}| \cdot |\det \mathbb{A}_+| \cdot \left(1 + O\left(\frac{\Theta}{\sqrt{M}}\right)\right)^W.$$

$$(10.4)$$

Observe that

$$|\det \mathbb{A}_+| \leq |1 + a_+^2|^W \leq 2^W. \tag{10.5}$$

Moreover, by Assumption 1.1 (ii), we see that $|\mathfrak{s}_{ii}| \leq (1 - c_0)/2$ for some small positive constant c_0 . Consequently, since $S^{(1)}$ is negative definite, we have

$$|\det S^{(1)}| \leq \prod_{i \neq 1} |\mathfrak{s}_{ii}| \leq \left(\frac{1 - c_0}{2}\right)^W \tag{10.6}$$

by Hadamard’s inequality. Substituting (10.5) and (10.6) into (10.4) yields

$$|\mathcal{I}(\mathcal{R}_+^b, \mathcal{R}_-^b, \mathcal{R}_+^x, \mathcal{R}_+^x, \mathcal{R}_S, \mathbb{I}^{W-1})| = O(e^{-cW}) \tag{10.7}$$

for some positive constant δ . Hence, we proved the first part of Lemma 5.9. The second part can be proved analogously. \square

In the sequel, we prove Lemma 10.1. We also ignore the factor $\mathbf{Q}(\cdot)$ from the discussion at first.

10.1 $\mathbf{P}(\hat{X}, \hat{B}, V, T)$ in the Type II vicinity

Lemma 10.2 *Suppose that the assumptions in Theorem 1.15 hold. In the Type II vicinity, we have*

$$|\mathbf{P}(\hat{X}, \hat{B}, V, T)| \leq \frac{W^{2+\gamma} \Theta^2}{M} |\det \mathbb{A}_+|^2 \det(S^{(1)})^2. \tag{10.8}$$

Proof We will follow the strategy in Sect. 9.1. We regard all V -variables as fixed parameters. Now, we define the function

$$\mathfrak{i} \equiv \mathfrak{i}_j(\hat{X}, \hat{B}, T) := |\mathfrak{x}_{j,1}| + |\mathfrak{x}_{j,2}| + |\mathfrak{b}_{j,1}| + |\mathfrak{b}_{j,2}| + |\mathfrak{t}_j|.$$

Then, we recall the representation (9.4) and the definition of $\Delta_{\ell,j}$ in (9.5). We still adopt the representation (9.8). It is easy to see that in the Type II vicinity, we also have the bound (9.9) for $\mathfrak{p}_{\ell,j,\alpha,\beta}$. The main difference is the first factor of the r.h.s. of (9.4). We expand it around the saddle point as

$$\begin{aligned} \exp \left\{ -Tr V_j^* \hat{X}_j^{-1} V_j \Omega_j T_j^{-1} \hat{B}_j^{-1} T_j \mathfrak{E}_j \right\} &= \exp \left\{ -Tr D_+^{-1} \Omega_j D_{\pm}^{-1} \mathfrak{E}_j \right\} \\ &\times \exp \left\{ -\frac{1}{\sqrt{M}} \hat{\Delta}_j \right\}. \end{aligned}$$

We take the formula above as the definition of $\widehat{\Delta}_j$, which is of the form

$$\widehat{\Delta}_j = \sum_{\alpha, \beta=1}^4 \hat{p}_{j, \alpha, \beta} \cdot \omega_{j, \alpha} \xi_{j, \beta},$$

where $\hat{p}_{j, \alpha, \beta}$ is a function of \hat{X}, \hat{B}, V and T -variables, satisfying

$$\hat{p}_{j, \alpha, \beta} = O(\iota).$$

Let $\widehat{\mathbb{H}} = (a_+^{-2} \mathbb{A}_+) \oplus S \oplus (a_+^{-2} \mathbb{A}_+) \oplus S$. Recalling the notation in (6.16), we can write

$$-\sum_{j, k} \tilde{s}_{jk} Tr \Omega_j \Xi_k - \sum_{j=1}^W Tr D_+^{-1} \Omega_j D_{\pm}^{-1} \Xi_j = -\Omega \widehat{\mathbb{H}} \Xi'.$$

Now, via replacing $\Delta_{1, j}$ by $\widehat{\Delta}_{1, j}$, \hat{k}_j by \hat{i}_j , \mathbb{H} by $\widehat{\mathbb{H}}$ in the proof of Lemma 9.2, we can perform the proof of Lemma 10.2 in the same way. We leave the details to the reader. \square

10.2 $F(\hat{X}, \hat{B}, V, T)$ in the Type II vicinity

Lemma 10.3 *Suppose that the assumptions in Theorem 1.15 hold. In the Type II vicinity, we have*

$$F(\hat{X}, \hat{B}, V, T) = O\left(\frac{\exp\{-(a_+ - a_-)N\eta\}}{(N\eta)^{n+1}}\right). \tag{10.9}$$

Proof Recall the decomposition (9.34). Note that Lemma 9.9 is still applicable. Hence, it suffices to estimate $\mathbb{F}(\hat{X}, V)$. Now, note that in the Type II vicinity, it is easy to see that

$$\sum_{j=1}^W Tr X_j J = O\left(\frac{\|\hat{x}_1\|_1 + \|\hat{x}_2\|_1}{\sqrt{M}}\right) = O\left(\frac{\Theta}{\sqrt{M}}\right).$$

Consequently, by the assumption on η , we have

$$\exp\left\{M\eta \sum_{j=1}^W Tr X_j J\right\} = \exp\{O(\Theta\sqrt{M}\eta)\} = 1 + o(1).$$

From (3.20) we can also see that all the other factors of $f(P_1, V, \hat{X}, X^{[1]})$ are $O(1)$. Hence, by the definition (9.33), we have $\mathbb{F}(\hat{X}, V) = O(\exp\{-(a_+ - a_-)N\eta\})$, which together with Lemma 9.9 yields the conclusion. \square

10.3 Summing up: Proof of Lemma 10.1

Analogously, we shall slightly modify the proofs of Lemma 10.2 and Lemma 10.3, in order to take $\mathbf{Q}(\cdot)$ into account. The proof can then be performed in the same manner as Lemma 9.1. We omit the details.

11 Proof of Theorem 1.15

The conclusion for Case 1 is a direct consequence of the discussions in Sects. 3.5–10. More precisely, by using Lemmas 5.1, 5.6, 5.8 and 5.9, we can get (1.21) immediately.

The proofs of Case 2 and Case 3 can be performed analogously, with slight modifications, which will be stated below. In Case 2, we shall slightly modify the discussions in Sects. 3.5–10 for Case 1, according to the decomposition of supermatrices in (3.9). Now, at first, in (3.12) and (3.13), for $A = \check{S}, \check{X}, \check{Y}, \check{\Omega}$ or \check{E} , we replace $A_p^{(1)}$ and $A_q^{(1)}$ by $A_p^{(1,2)}$ and A_q respectively, and replace $A_q^{[1]}$ by $A_p^{[2]}$. In addition, in the last three lines of (3.13), we shall also replace \tilde{s}_{jq} by \tilde{s}_{jp} , and replace \tilde{s}_{pq} and \tilde{s}_{qp} by \tilde{s}_{pp} , and in the first line, we replace $\bar{\phi}_{1,q,1}\phi_{1,p,1}\bar{\phi}_{2,p,1}\phi_{2,q,1}$ by $\bar{\phi}_{1,p,2}\phi_{1,p,1}\bar{\phi}_{2,p,1}\phi_{2,p,2}$. Then, in (3.14) and (3.15), for $A = X, Y, \Omega, \mathcal{E}, \omega, \xi, \mathbf{w}, y, \tilde{u}, \tilde{v}$ or σ , we replace $A_q^{[1]}$ by $A_p^{[2]}$. With these modifications, it is easy to check the proof in Sects. 3.5–10 applies to Case 2 as well. The main point is we can still gain the factor $1/(N\eta)^{n+1}$ from integral of $g(\cdot)$ defined in (3.21) (with $y_q^{[1]}$ and $\mathbf{w}_q^{[1]}$ replaced by $y_p^{[2]}$ and $\mathbf{w}_p^{[2]}$). Heuristically, we can go back to (4.4), and replace $\sigma_q^{[1]}$ by $\sigma_p^{[2]}$ therein. It is then quite clear the same estimate holds. Consequently, Lemmas 5.1, 5.6, 5.8 and 5.9 still hold under the replacement of the variables described above. Hence, (1.21) holds in Case 2.

In Case 3, we can also mimic the discussions for Case 1 with slight modifications. We also start from (3.12) and (3.13). For $A = \check{S}, \check{X}, \check{Y}, \check{\Omega}, \check{E}, \omega$ and ξ , we replace $A_q^{(1)}$ by A_q , and replace $A_q^{[1]}$ by 0. In addition, in the first line of (3.13), we replace $\bar{\phi}_{1,q,1}\phi_{1,p,1}\bar{\phi}_{2,p,1}\phi_{2,q,1}$ by $\phi_{1,p,1}\phi_{1,p,1}\bar{\phi}_{2,p,1}\phi_{2,p,1}$. Consequently, after using superbosonization formula, we will get the factor $(y_p^{[1]}|(\mathbf{w}_p^{[1]}(\mathbf{w}_p^{[1]})^*)_{12}|)^{2n}$ instead of $(y_p^{[1]}y_q^{[1]}|(\mathbf{w}_q^{[1]}(\mathbf{w}_q^{[1]})^*)_{12}(\mathbf{w}_p^{[1]}(\mathbf{w}_p^{[1]})^*)_{21}|)^n$ in (3.16). Then, for the superdeterminant terms

$$\prod_{k=p,q} \frac{\det(X_k - \Omega_k(Y_k)^{-1}\mathcal{E}_k)}{\det Y_k}, \quad \prod_{k=p,q} \frac{y_k^{[1]} \left(y_k^{[1]} - \xi_k^{[1]}(X_k^{[1]})^{-1}\omega_k^{[1]} \right)^2}{\det^2(X_k^{[1]})}.$$

we shall only keep the factors with $k = p$ and delete those with $k = q$. Moreover, we shall also replace $A_q^{[1]}$ by 0 for $A = X, Y, \Omega, \mathcal{E}, \omega, \xi, \mathbf{w}, y, \tilde{u}, \tilde{v}$ or σ in (3.16). In addition, $dA^{[1]}$ shall be redefined as the differential of $A_p^{[1]}$ -variables only, for $A = X, \mathbf{y}, \mathbf{w}, \omega$ and ξ . One can check step by step that such a modification does not require any essential change of our discussions for Case 1. Especially, note that our modification has nothing to do with the saddle point analysis on the Gaussian measure $\exp\{-M(K(\hat{X}, V) + L(\hat{B}, T))\}$. Moreover, the term $\mathcal{P}(\cdot)$ in (3.29) can be redefined by deleting the factor with $k = q$ in the last term therein. Such a modification does not

change our analysis of $\mathcal{P}(\cdot)$. In addition, the irrelevant term $\mathcal{Q}(\cdot)$ can also be defined accordingly. Specifically, we shall delete the factor with $k = q$ in the last term of (3.30) and replace $A_q^{[1]}$ by 0 for $A = \Omega, \mathcal{E}, \omega, \xi, \mathbf{w}, y$. It is routine to check that Lemma 6.3 still holds under such a modification. Analogously, we can redefine the functions $\mathcal{F}(\cdot), f(\cdot)$ and $g(\cdot)$ in (3.19)–(3.21). Now, the main difference between Case 3 and Case 1 or 2 is that the factor $(y_p^{[1]}|(\mathbf{w}_p^{[1]}(\mathbf{w}_p^{[1]})^*)_{12}|)^{2n}$ does not produce oscillation in the integral of $g(\cdot)$ any more. Heuristically, the counterpart of (4.4) in Case 3 reads

$$\begin{aligned}
 & e^{(a_+ - a_-)N\eta} \int d\mathbf{y}^{[1]} d\mathbf{w}^{[1]} d\nu(Q_1) \cdot g(\hat{B}, T, Q_1, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}) \\
 & \sim \int_0^\infty 2t dt \int_{\mathbb{L}} d\sigma_p^{[1]} \cdot e^{-cN\eta t^2 + c_1 e^{-i\sigma_p^{[1]} t}} \sim \frac{1}{N\eta}.
 \end{aligned}$$

Hence, (1.21) holds for Case 3. Therefore, we completed the proof of Theorem 1.15.

12 Comment on the prefactor N^{C_0} in (1.21)

In the proof of (1.21), we have used N^{C_0} to replace $M\Theta^2 W^{C_0}/(N\eta)^\ell$ (see the proof of Lemma 5.8). However, the latter is also artificial. It can be improved to some n -dependent constant C_n via a more delicate analysis on $\mathbf{A}(\cdot)$, i.e. the integral of $\mathcal{P}(\cdot)\mathcal{Q}(\cdot)\mathcal{F}(\cdot)$. Such an improvement stems from the cancellation in the Gaussian integral. At first, a finer analysis will show that the factor $\mathcal{Q}(\cdot)$ can really be ignored, in the sense that it does not play any role in the estimate of the order of $\mathbb{E}|G_{ij}(z)|^{2n}$. Hence, for simplicity, we just focus on the product $\mathbf{P}(\cdot)\mathbf{F}(\cdot)$ instead of $\mathbf{A}(\cdot)$. Then, we go back to Lemmas 9.2 and 9.7. Recall the decomposition (9.34). A more careful analysis on $\mathbf{F}(\cdot)$ leads us to the following expansion, up to the subleading order terms of the factors $\mathring{\mathbb{G}}(\cdot)$ and $\mathring{\mathbb{F}}(\cdot)$,

$$\begin{aligned}
 \mathbf{F}(\cdot) = \mathring{\mathbb{G}}(\cdot)\mathring{\mathbb{F}}(\cdot) & \sim \frac{1}{(N\eta)^{n+2}} \left(1 + \frac{M\eta}{\sqrt{M}} \sum_{j=1}^W l_j(\hat{x}_{j,1}, \hat{x}_{j,2}, \hat{v}_j) + \dots \right) \\
 & \times \left(1 + \frac{M\eta}{\sqrt{M}} \sum_{j=1}^W l'_j(\hat{b}_{j,1}, \hat{b}_{j,2}, \hat{t}_j) + \dots \right), \tag{12.1}
 \end{aligned}$$

where $l_j(\cdot)$'s and $l'_j(\cdot)$'s are some linear combinations of the arguments. Analogously, we shall write down the leading order term of $\mathbf{P}(\cdot)$ in terms of $\hat{\mathbf{x}}, \hat{\mathbf{b}}, \hat{\mathbf{t}}$ and $\hat{\mathbf{v}}$ explicitly. Then it can be seen that the leading order term of $\mathbf{P}(\cdot)$ is a linear combination of $\hat{x}_{j,1}\hat{x}_{k,2}, \hat{b}_{j,1}\hat{b}_{k,2}, \hat{x}_{j,\alpha}\hat{b}_{k,\beta}, \nu_{j,\alpha}\tau_{k,\beta}$ for $j, k = 1, \dots, W$ and $\alpha, \beta = 1, 2$, in which all the coefficients are of order $1/M$. Observe that the Gaussian integral in (8.45) will kill the linear terms. Consequently, in the expansion (12.1), the first term that survives

after the Gaussian integral is actually

$$\frac{1}{(N\eta)^{n+2}} \cdot \frac{M\eta}{\sqrt{M}} \sum_{j=1}^W l_j(\hat{x}_{j,1}, \hat{x}_{j,2}, \hat{v}_j) \cdot \frac{M\eta}{\sqrt{M}} \sum_{j=1}^W l'_j(\hat{b}_{j,1}, \hat{b}_{j,2}, \hat{l}_j). \tag{12.2}$$

Replacing $A(\cdot)$ by the product of the leading order term of $P(\cdot)$ and (12.2) in the integral (8.45) and taking the Gaussian integral over $\mathbf{c}, \mathbf{d}, \boldsymbol{\tau}$ and \mathbf{v} -variables yield the true order $1/(N\eta)^n$, without additional N -dependent prefactors.

Table of symbols

For the convenience of the reader, in the following table we collect some frequently used symbols followed by the locations where they are defined.

a_+, a_-	(1.27)	\hat{X}_j, \hat{B}_j	(3.23)	\mathbb{S}, \mathbb{S}^v	(5.23)
u, s, u_j, s_j	(3.25)	P_1, Q_1	(3.25)	Θ	(5.30)
$\tau_{j,1}, \tau_{j,2}$	(8.20)	V_j, T_j	(3.24)	$\mathbb{I}, \mathbb{L}, \Sigma, \Gamma$	(1.30)
$v_{j,1}, v_{j,2}$	(8.23)	$\mathbb{A}_+, \mathbb{A}_-$	(8.14)	$\gamma_{\pm}^b, \gamma_{\pm}^x, \gamma_S$	(5.32)
$\mathbb{k}(a)$	(5.4)	$\mathbb{A}_+^v, \mathbb{A}_-^v$	(8.33)	$\tilde{\gamma}, \tilde{\gamma}_S$	(8.4)
$D_{\pm}, D_{\mp}, D_+, D_-$	(1.28)	\mathbb{H}	(9.26)	γ_{∞}	(8.39)

Acknowledgments Open access funding provided by Institute of Science and Technology (IST Austria). The authors are very grateful to the anonymous referees for careful reading and valuable comments, which helped to improve the organization.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Aizenman, M., Molchanov, S.: Localization at large disorder and at extreme energies: an elementary derivation. *Commun. Math. Phys.* **157**, 245–278 (1993)
2. Anderson, P.: Absences of diffusion in certain random lattices. *Phys. Rev.* **109**, 1492–1505 (1958)
3. Bunder, J.E., Efetov, K.B., Kravtsov, V.E., Yevtushenko, O.M., Zirnbauer, M.R.: Superbosonization formula and its application to random matrix theory. *J. Stat. Phys.* **129**(5–6), 809–832 (2007)
4. Disertori, M., Pinson, H., Spencer, T.: Density of states for random band matrices. *Commun. Math. Phys.* **232**(1), 83–124 (2002)
5. Disertori, M., Spencer, T.: Anderson localization for a supersymmetric sigma model. *Commun. Math. Phys.* **300**(3), 659–671 (2010)
6. Disertori, M., Spencer, T., Zirnbauer, M.R.: Quasi-diffusion in a 3D supersymmetric hyperbolic sigma model. *Commun. Math. Phys.* **300**(2), 435–486 (2010)
7. Efetov, K.: *Supersymmetry in Disorder and Chaos*. Cambridge University Press, Cambridge (1997)
8. Ellis, R.B.: *Discrete Green’s Functions for Products of Regular Graphs* (2003). [arXiv:math/0309080](https://arxiv.org/abs/math/0309080)
9. Erdős, L., Knowles, A., Yau, H.-T.: Averaging fluctuations in resolvents of random band matrices. *Ann. Henri Poincaré* **14**(8), 1837–1926 (2013)
10. Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: Spectral statistics of Erdős-Rényi graphs I: local semicircle law. *Ann. Prob.* **41**(3B), 2279–2375 (2013)
11. Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: Delocalization and diffusion profile for random band matrices. *Commun. Math. Phys.* **323**, 367–416 (2013)

12. Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: The local semicircle law for a general class of random matrices. *Electron. J. Probab* **18**(59), 1–58 (2013)
13. Erdős, L., Schlein, B., Yau, H.-T.: Local semicircle law and complete delocalization for Wigner random matrices. *Commun. Math. Phys.* **287**, 641–655 (2009)
14. Erdős, L., Yau, H.-T., Yin, J.: Bulk universality for generalized Wigner matrices. *Probab. Theory Relat. Fields* **154**(1–2), 341–407 (2012)
15. Fröhlich, J., Spencer, T.: Absence of diffusion in the Anderson tight binding model for large disorder or low energy. *Commun. Math. Phys.* **88**, 151–184 (1983)
16. Fyodorov, Y.V., Mirlin, A.D.: Scaling properties of localization in random band matrices: a σ -model approach. *Phys. Rev. Lett.* **67**, 2405–2409 (1991)
17. Klein, A.: Extended states in the Anderson model on the Bethe lattice. *Adv. Math.* **133**(1), 163–184 (1998)
18. Littellmann, P., Sommers, H.J., Zirnbauer, M.R.: Superbosonization of invariant random matrix ensembles. *Commun. Math. Phys.* **283**(2), 343–395 (2008)
19. Shcherbina, T.: On the second mixed moment of the characteristic polynomials of 1D band matrices. *Commun. Math. Phys.* **328**(1), 45–82 (2014)
20. Shcherbina, T.: Universality of the local regime for the block band matrices with a finite number of blocks. *J. Stat. Phys.* **155**(3), 466–499 (2014)
21. Shcherbina, T.: Universality of the second mixed moment of the characteristic polynomials of the 1D band matrices: real symmetric case (2014). [arXiv:1410.3084](https://arxiv.org/abs/1410.3084)
22. Schäfer, L., Wegner, F.: Disordered system with n orbitals per site: Lagrange formulation, hyperbolic symmetry, and Goldstone modes. *Z. Phys. B* **38**, 113–126 (1980)
23. Schenker, J.: Eigenvector localization for random band matrices with power law band width. *Commun. Math. Phys.* **290**, 1065–1097 (2009)
24. Sodin, S.: An estimate for the average spectral measure of random band matrices. *J. Stat. Phys.* **144**(1), 46–59 (2011)
25. Sodin, S.: The spectral edge of some random band matrices. *Ann. Math.* **172**(3), 2223–2251 (2010)
26. Spencer, T.: *SUSY Statistical Mechanics and Random Band Matrices, Quantum Many Body Systems*. Springer, Berlin Heidelberg (2012)
27. Spencer, T., Zirnbauer, M.R.: Spontaneous symmetry breaking of a hyperbolic sigma model in three dimensions. *Commun. Math. Phys.* **252**(1–3), 167–187 (2004)
28. Tao, T., Vu, V.: Random matrices: universality of local eigenvalue statistics. *Acta Math.* **206**(1), 127–204 (2011)
29. Wang, W.M.: Mean field upper and lower bounds on Lyapunov exponents. *Am. J. Math.* 851–878 (2002)
30. Wegner, F.J.: Disordered system with n orbitals per site: $n \rightarrow \infty$ limit. *Phys. Rev. B* **19**, 783–792 (1979)
31. Wigner, E.: Characteristic vectors of bordered matrices with infinite dimensions. *Ann. Math.* **62**, 548–564 (1955)