RANDOM UNITARY REPRESENTATIONS OF SURFACE GROUPS II: THE LARGE n LIMIT

MICHAEL MAGEE

To Oona

ABSTRACT. Let Σ_g be a closed surface of genus $g \geq 2$ and Γ_g denote the fundamental group of Σ_g. We establish a generalization of Voiculescu's theorem on the asymptotic $*$-freeness of Haar unitary matrices from free groups to Γ_g. We prove that for a random representation of Γ_g into $\mathsf{SU}(n)$, with law given by the volume form arising from the Atiyah-Bott-Goldman symplectic form on moduli space, the expected value of the trace of a fixed non-identity element of Γ_g is bounded as $n \to \infty$. The proof involves an interplay between Dehn's work on the word problem in Γ_g and classical invariant theory.

CONTENTS

1. Introduction

In a foundational series of papers [Voi85, Voi86, Voi87, Voi90, Voi91], Voiculescu developed a robust theory of non-commuting random variables that became known as *free probability*. One of the initial landmarks of this theory is the following result. Let $\mathbf{F}_r$ denote the non-commutative free group of rank r. Let $\mathsf{U}(n)$ denote the group of $n \times n$ complex unitary matrices. For any $w \in \mathbf{F}_r$ we obtain a *word map* $w : \mathsf{U}(n)^r \to \mathsf{U}(n)$ by substituting matrices for generators of $\mathbf{F}_r$. Let $\mu^{\mathrm{Haar}}_{\mathsf{U}(n)^r}$ denote the probability Haar measure on $\mathsf{U}(n)^r$ and $\mathrm{Tr} : \mathsf{U}(n) \to \mathbf{C}$ the standard trace. Any integral over a compact group will be done with respect to the probability Haar measure, denoted by $d\mu$.

A simplified version of Voiculescu's result [Voi91, Thm. 3.8] can be formulated as follows[1]:

Theorem 1.1 (Voiculescu)**.** *For any non-identity $w \in \mathbf{F}_r$, as $n \to \infty$*

$$\int_{\mathsf{U}(n)^r} \mathrm{Tr}(w(x))d\mu(x) = o_w(n). \tag{1.1}$$

We describe the interpretation of Theorem 1.1 as convergence of non-commutative random variables momentarily. Before this, we explain the main result of the current paper.

Another way to think about the integral (1.1), that invites generalization, is to identify $\mathsf{U}(n)^r$ with $\mathrm{Hom}(\mathbf{F}_r, \mathsf{U}(n))$ and Haar measure as a natural probability measure on this *representation variety*. Now it is natural to ask whether there are other infinite discrete groups rather than $\mathbf{F}_r$ such that $\mathrm{Hom}(\mathbf{F}_r, \mathsf{U}(n))$ has a natural measure, and whether similar phenomena as in Theorem 1.1 may hold. *The main point of this paper is to establish the analog of Theorem 1.1 when $\mathbf{F}_r$ is replaced by the fundamental group of a compact surface of genus at least 2.*

We now explain this generalization of Theorem 1.1; for technical reasons it superficially looks slightly different as follows.

(1) The integral (1.1) is equal to 0 if $w \notin [\mathbf{F}_r, \mathbf{F}_r]$, the commutator subgroup of $\mathbf{F}_r$ [MP15, Claim 3.1], and if $w \in [\mathbf{F}_r, \mathbf{F}_r]$, the value of (1.1) is for $n \geq n_0(w)$ the same as the corresponding integral over $\mathsf{SU}(n)^r \leq \mathsf{U}(n)^r$, where $\mathsf{SU}(n)$ is the subgroup of determinant

[1] Voiculescu's result in [Voi91, Thm. 3.8] is more general than what we state here, also involving a deterministic sequence of unitary matrices.

one matrices [Mag21, Prop. 3.1]. So in all cases of interest we can replace $\mathsf{U}(n)$ by $\mathsf{SU}(n)$ in (1.1).

(2) Since $\mathrm{Tr} \circ w$ is invariant under the diagonal conjugation action of $\mathsf{SU}(n)$ on $\mathrm{Hom}(\mathbf{F}_r, \mathsf{SU}(n)) \cong \mathsf{SU}(n)^r$, the integral $\int_{\mathsf{SU}(n)^r} \mathrm{Tr}(w(x))d\mu(x)$ can be written as an integral over $\mathrm{Hom}(\mathbf{F}_r, \mathsf{SU}(n))/\mathsf{PSU}(n)$. Here $\mathsf{PSU}(n)$ is $\mathsf{SU}(n)$ modulo its center.

For $g \geq 2$ let Σ_g denote a closed topological surface of genus g. We let Γ_g denote the fundamental group of Σ_g with explicit presentation

$$\Gamma_g = \langle a_1, b_1, \ldots, a_g, b_g \,|\, [a_1, b_1] \cdots [a_g, b_g] \,\rangle.$$

The most natural measure on $\mathrm{Hom}(\Gamma_g, \mathsf{SU}(n))/\mathsf{PSU}(n)$ to replace the measure induced by Haar measure on $\mathrm{Hom}(\mathbf{F}_r, \mathsf{SU}(n))/\mathsf{PSU}(n)$ is called the Atiyah–Bott–Goldman measure. The definition of this measure involves removing singular parts of $\mathrm{Hom}(\Gamma_g, \mathsf{SU}(n))/\mathsf{PSU}(n)$. Indeed, let $\mathrm{Hom}(\Gamma_g, \mathsf{SU}(n))^{\mathrm{irr}}$ denote the collection of homomorphisms that are irreducible as linear representations. Then

$$\mathcal{M}_{g,n} \stackrel{\mathrm{def}}{=} \mathrm{Hom}(\Gamma_g, \mathsf{SU}(n))^{\mathrm{irr}}/\mathsf{PSU}(n)$$

is a smooth manifold [Gol84]. Moreover there is a symplectic form $\omega_{g,n}$ on $\mathcal{M}_{g,n}$ called the Atiyah–Bott–Goldman form after [AB83, Gol84]. This symplectic form gives, in the usual way, a volume form on $\mathcal{M}_{g,n}$ denoted by $\mathrm{Vol}_{\mathcal{M}_{g,n}}$. For many more details see Goldman [Gol84] or the prequel paper [Mag21, §§2.7].

For any $\gamma \in \Gamma$, we obtain a function $\mathrm{Tr}_\gamma : \mathrm{Hom}(\Gamma_g, \mathsf{SU}(n)) \to \mathbf{C}$ defined by

$$\mathrm{Tr}_\gamma(\phi) \stackrel{\mathrm{def}}{=} \mathrm{Tr}(\phi(\gamma)).$$

This function descends to a function $\mathrm{Tr}_\gamma : \mathcal{M}_{g,n} \to \mathbf{C}$. We are interested in the expected value

$$\mathbb{E}_{g,n}[\mathrm{Tr}_\gamma] \stackrel{\mathrm{def}}{=} \frac{\int_{\mathcal{M}_{g,n}} \mathrm{Tr}_\gamma \, d\mathrm{Vol}_{\mathcal{M}_{g,n}}}{\int_{\mathcal{M}_{g,n}} d\mathrm{Vol}_{\mathcal{M}_{g,n}}}.$$

The main theorem of this paper is the following.

Theorem 1.2. *Let $g \geq 2$. If $\gamma \in \Gamma_g$ is not the identity, then $\mathbb{E}_{g,n}[\mathrm{Tr}_\gamma] = O_\gamma(1)$ as $n \to \infty$.*

The non-commutative probabilistic consequences of Theorem 1.2 will be discussed in the next section.

1.1. **Non-commutative probability.** We follow the book [VDN92]. A *non-commutative probability space* is a pair $(\mathcal{B}, \tau)$ where $\mathcal{B}$ is a complex unital algebra and τ is a linear functional on $\mathcal{B}$ such that $\tau(1) = 1$. Let $\mathbf{C}\langle x_1, \ldots, x_r \rangle$ denote the free non-commutative unital algebra in indeterminates $x_1, \ldots, x_r$. A *random variable* in $(\mathcal{B}, \tau)$ is an element of $\mathcal{B}$. If

$(X_1,\ldots,X_r)\in\mathcal{B}^r$ are random variables in $(\mathcal{B},\tau)$, their *joint distribution* is defined to be the linear functional

$$\tilde{\tau}:\mathbf{C}\langle x_1,\ldots,x_r\rangle\to\mathbf{C}$$

defined by $\tilde{\tau}(z)\stackrel{\mathrm{def}}{=}\tau(\Phi(z))$ where $\Phi:\mathbf{C}\langle x_1,\ldots,x_r\rangle\to\mathcal{B}$ is the linear map defined by $\Phi(x_i)=X_i$. For a linear functional $\tilde{\tau}_\infty:\mathbf{C}\langle x_1,\ldots,x_r\rangle\to\mathbf{C}$ with $\tilde{\tau}_\infty(1)=1$, we say that a sequence of random variables $(X_1^{(n)},\ldots,X_r^{(n)})\in(\mathcal{B}_n,\tau_n)$ converge in distribution as $n\to\infty$ to $\tilde{\tau}_\infty$ if $\tilde{\tau}_n$ convergences pointwise to $\tilde{\tau}_\infty$ on $\mathbf{C}\langle x_1,\ldots,x_r\rangle$.

A very concrete example of this phenomenon is as follows. The function

$$\tau_n:\mathbf{F}_r\to\mathbf{C},\quad \tau_n(w)\stackrel{\mathrm{def}}{=}\frac{1}{n}\int_{\mathsf{U}(n)^r}\mathrm{Tr}(w(x))d\mu(x)$$

extends to a linear functional τ_n on the algebra $\mathbf{C}[\mathbf{F}_r]$ with $\tau_n(\mathrm{id})=1$. From this point of view, Theorem 1.1 implies the following statement.

Theorem 1.3 (Voiculescu)**.** *Let $r\geq 0$ and $X_1,\ldots,X_r$ denote fixed generators of $\mathbf{F}_r$, and $\bar{X}_1,\ldots.\bar{X}_r$ denote their inverses, i.e. $\bar{X}_i=X_i^{-1}$. The random variables $X_1,\ldots,X_r,\bar{X}_1,\ldots.\bar{X}_r$ in the non-commutative probability spaces $(\mathbf{C}[\mathbf{F}_r],\tau_n)$ converge as $n\to\infty$ to a limiting distribution*

$$\tilde{\tau}_\infty:\mathbf{C}\langle x_1,\ldots,x_r,\bar{x}_1,\ldots,\bar{x}_r\rangle\to\mathbf{C}$$

that is completely determined by (1.1). Indeed, if w is any monomial in $x_1,\ldots,x_r,\bar{x}_1,\ldots,\bar{x}_r$, then $\tilde{\tau}_\infty(w)=1$ if and only if after identifying $\bar{x}_i$ with x_i^{-1}, w reduces to the identity in $\mathbf{F}_r=\langle x_1,\ldots,x_r\rangle$, and $\tilde{\tau}_\infty(w)=0$ otherwise.

In the language of [Voi91], in the limiting non-commutative probability space $(\mathbf{C}\langle x_1,\ldots,x_r,\bar{x}_1,\ldots,\bar{x}_r\rangle,\tilde{\tau}_\infty)$, the subalgebras

$$\mathcal{A}_1\stackrel{\mathrm{def}}{=}\mathbf{C}\langle x_1,\bar{x}_1\rangle,\ldots,\mathcal{A}_r\stackrel{\mathrm{def}}{=}\mathbf{C}\langle x_r,\bar{x}_r\rangle$$

are a *free family of subalgebras*: if $a_j\in\mathcal{A}_{i_j}$ for $j\in[q]$, $i_1\neq i_2\neq\cdots\neq i_q$, and $\tilde{\tau}_\infty(a_j)=0$ for $j\in[q]$ then

$$\tilde{\tau}_\infty(a_1a_2\cdots a_q)=0.$$

Accordingly, it is said [Voi91, Thm. 3.8] that if $\{u_j(n):1\leq j\leq r\}$ are independent Haar-random elements of $\mathsf{U}(n)$, the family $\{\{u_j(n),u_j^*(n)\}:1\leq j\leq r\}$ of sets of random variables are *asymptotically free*.

Because Γ_g is not free, asymptotic freeness does not correctly capture the asymptotic behavior of the expected values $\mathbb{E}_{g,n}[\mathrm{Tr}_\gamma]$, however, an analog of Theorem 1.3 is implied by Theorem 1.2. For $\gamma\in\Gamma_g$ let

$$\tau_{g,n}(\gamma)\stackrel{\mathrm{def}}{=}\frac{1}{n}\mathbb{E}_{g,n}[\mathrm{Tr}_\gamma].$$

Corollary 1.4. *Let $g\geq 2$, $a_1,b_1,\ldots,a_g,b_g$ denote the previously fixed generators of Γ_g, and $\bar{a}_1,\bar{b}_1,\ldots,\bar{a}_g,\bar{b}_g$ denote their inverses. The random variables $a_1,b_1,\ldots,a_g,b_g$, $\bar{a}_1,\bar{b}_1,\ldots,\bar{a}_g,\bar{b}_g$ in the non-commutative probability* *spaces* $(\mathbf{C}[\Gamma_g], \tau_{g,n})$ *converge in distribution as* $n \to \infty$ *to a limiting distribution*

$$\tilde{\tau}_{g,\infty} : \mathbf{C}\langle x_1, \ldots, x_g, y_1, \ldots, y_g, \bar{x}_1, \ldots, \bar{x}_g, \bar{y}_1, \ldots, \bar{y}_g\rangle \to \mathbf{C},$$

where x_i *(resp.* $y_i, \bar{x}_i, \bar{y}_i$*) corresponds to* a_i *(resp.* $b_i, \bar{a}_i, \bar{b}_i$*). This can be described explicitly as follows. If* w *is any monomial in* $x_1, \ldots, x_g, y_1, \ldots, y_g, \bar{x}_1, \ldots, \bar{x}_g, \bar{y}_1, \ldots, \bar{y}_g$*, then* $\tilde{\tau}_{g,\infty}(w) = 1$ *if and only if* w *maps to the identity under the map*

$$\mathbf{C}\langle x_1, \ldots, x_g, y_1, \ldots, y_g, \bar{x}_1, \ldots, \bar{x}_g, \bar{y}_1, \ldots, \bar{y}_g\rangle \to \mathbf{C}[\Gamma_g]$$

obtained by identifying $x_i, y_i, \bar{x}_i, \bar{y}_i$ *with the corresponding elements of* Γ_g*. If* w *does not map to the identity under this map, then* $\tilde{\tau}_{g,\infty}(w) = 0$*.*

Notice that the estimate given in Theorem 1.2 is stronger than needed to establish Corollary 1.4.

1.2. **Related works and further questions.** The most closely related existing result to Theorem 1.2 is a theorem of the author and Puder [MP20, Thm. 1.2] that establishes Theorem 1.2 when the family of groups $\mathsf{SU}(n)$ is replaced by the family of symmetric groups S_n, and Tr is replaced by the character fix given by the number of fixed points of a permutation. In this case, the result is phrased in terms of integrating over $\mathrm{Hom}(\Gamma_g, S_n)$ with respect to the uniform probability measure. The corresponding result for $\mathrm{Hom}(\mathbf{F}_r, S_n)$ was proved much longer ago by Nica in [Nic94].

The problem of integrating geometric functions like Tr_γ over $\mathcal{M}_{g,n}$ is also connected to the work of Mirzakhani since, as Goldman explains in [Gol84, §2], the Atiyah–Bott–Goldman symplectic form generalizes the Weil–Petersson symplectic form on the Teichmüller space of genus g Riemann surfaces. In [Mir07], Mirzakhani developed a method for integrating geometric functions on moduli spaces of Riemann surfaces with respect to the Weil–Petersson volume form. Although there is certainly a similarity between *(ibid.)* and the current work, here the emphasis is on $n \to \infty$, whereas *(ibid.)* caters to the regime $g \to \infty$; the target group playing the role of $\mathsf{SU}(n)$ is always $\mathsf{PSL}(2, \mathbf{R})$.

We now take the opportunity to mention some questions that Theorem 1.2 leads to. In the paper [Voi91], Voiculescu is able to boost Theorem 1.1 from a convergence in distribution result to a result on convergence in probability, that is, for any $\epsilon > 0$, and fixed $w \in \mathbf{F}_r$, the Haar measure of the set

$$\{\, \phi \in \mathrm{Hom}(\mathbf{F}_r, \mathsf{U}(n)) \,:\, |\mathrm{Tr}(\phi(w))| \le \epsilon n \,\}$$

tends to one as $n \to \infty$ [Voi91, Thm. 3.9]. To do this, Voiculescu uses that the family of measure spaces $(\mathrm{Hom}(\mathbf{F}_r, \mathsf{U}(n)), \mu)$ form a *Levy family* in the sense of Gromov and Milman [GM83]. This latter fact relies on an estimate for the first non-zero eigenvalue of the Laplacian on $\mathrm{Hom}(\mathbf{F}_r, \mathsf{U}(n))$. It is interesting to ask whether a similar phenomenon holds for the family of measure spaces $(\mathcal{M}_{g,n}, \mu_{g,n}^{\mathrm{ABG}})$ where $\mu_{g,n}^{\mathrm{ABG}}$ is the probability measure corresponding to $\mathrm{Vol}_{\mathcal{M}_{g,n}}$. The fact that $\mathcal{M}_{g,n}$ is non-compact seems to be a significant complication in answering this question using isoperimetric inequalities.

On the other hand, as pointed out to us by a referee, the results of this paper can very likely be extended to give bounds on the variance

$$\mathbb{E}_{g,n}[|\mathrm{Tr}_\gamma|^2]$$

that can be used to improve Theorem 1.2 to the result that for $\gamma \neq \mathrm{id}$, the normalized traces $\frac{\mathrm{Tr}_\gamma}{n}$ converge in probability to zero as $n \to \infty$. To avoid adding complications to this paper, this will be pursued elsewhere.

In the prequel to this paper [Mag21] it was proved that for any fixed $\gamma \in \Gamma_g$, there is an infinite sequence of rational numbers $a_{-1}(\gamma), a_0(\gamma), a_1(\gamma), \ldots \in \mathbf{Q}$ such that for any $M \in \mathbf{N}$,

$$\mathbb{E}_{g,n}[\mathrm{Tr}_\gamma] = a_{-1}(\gamma)n + a_0(\gamma) + \frac{a_1(\gamma)}{n} + \cdots + \frac{a_{M-1}(\gamma)}{n^{M-1}} + O_{\gamma,M}\left(\frac{1}{n^M}\right) \quad (1.2)$$

as $n \to \infty$. Theorem 1.2 implies that $a_{-1}(\gamma) = 0$ if $\gamma \neq \mathrm{id}$. It is also interesting to understand the other coefficients of this series. This has been accomplished when Γ_g is replaced by $\mathbf{F}_r$ by the author and Puder in [MP19] where in fact it is proved that

$$\mathbb{E}_{\mathbf{F}_r,n}[\mathrm{Tr}_w] \stackrel{\mathrm{def}}{=} \int_{\mathsf{U}(n)^r} \mathrm{Tr}(w(x))d\mu(x)$$

is given by a *rational* function of n and in particular can be expanded as in (1.2). The corresponding coefficients of the Laurent series of $\mathbb{E}_{\mathbf{F}_r,n}[\mathrm{Tr}_w]$ are explained in terms of Euler characteristics of subgroups of mapping class groups. One corollary is that as $n \to \infty$

$$\mathbb{E}_{\mathbf{F}_r,n}[\mathrm{Tr}_w] = O\left(\frac{1}{n^{2\mathrm{cl}(w)-1}}\right) \quad (1.3)$$

where $\mathrm{cl}(w)$ is the the *commutator length of* w: the minimal number of commutators that w can be written as a product of, or ∞ if $w \notin [\mathbf{F}_r, \mathbf{F}_r]$. We guess that an estimate like (1.3) should hold for $\mathbb{E}_{g,n}[\mathrm{Tr}_\gamma]$ where commutator length in $\mathbf{F}_r$ is replaced by commutator length in Γ_g.

Another strengthening of Theorem 1.1 is the *strong asymptotic freeness* of Haar unitaries. This states that for any complex linear combination

$$\sum_w a_w w \in \mathbf{C}[\mathbf{F}_r],$$

almost surely w.r.t. Haar random $\phi \in \mathrm{Hom}(\mathbf{F}_r, \mathsf{U}(n))$ as $n \to \infty$, we have

$$\Big\| \sum_w a_w \phi(w) \Big\| \to \Big\| \sum_w a_w w \Big\|_{\mathrm{Op}(\ell^2(\mathbf{F}_r))}$$

where the left hand side is the operator norm on $\mathbf{C}^n$ with standard Hermitian inner product and the norm on the right hand side is the operator norm in the regular representation of $\mathbf{F}_r$. This result was proved by Collins and Male in [CM14]. It is probably very hard to extend this result to Γ_g; the proof of Collins and Male relies on seminal work of Haagerup and Thorbjørnsen [HT05] in a way that does not obviously extend to Γ_g.

We finally mention that the expected values $\mathbb{E}_{g,n}[\mathrm{Tr}_\gamma]$ arise as a limiting form of expected values of Wilson loops in 2D Yang-Mills theory, when the coupling constant is set to zero. This will not be discussed in detail here, we refer the reader instead to the introduction of [Mag21]. Here we just mention recent works of Lemoine [Lem22]and Dahlqvist—Lemoine [DL22] that make progress on related problems in the Yang-Mills setting.

1.3. **Overview of paper.**

Here we explain the structure of the paper.

In §§2.1-§§2.5 we give some general background to the paper not depending on the prequel [Mag21]. In §§2.6 we import results that we proved in the prequel and that are needed here.

At the beginning of §3 we state the key result (Theorem 3.1) of the remainder of the paper. To motivate things, §§3.1 contains a discussion of why the most straightforward approach does not work, and also a discussion of what will follow instead. In the remainder of §3 we explain how to augment the Weingarten calculus to arrive to a formula for the key quantity $\mathcal{J}_n(w,\mu,\nu)$ (defined in Proposition 2.9) in combinatorial terms that are 'good' for the next part of the argument.

Indeed in §§4.1 we explain how each combinatorial datum we encountered in our formula for $\mathcal{J}_n(w,\mu,\nu)$ can be used to build a decorated surface. In Corollary 4.5 we obtain a bound on $\mathcal{J}_n(w,\mu,\nu)$ in terms of the Euler characteristics of some of the surfaces that previously arose. We may restrict to certain surfaces of simplified form by performing two surgery arguments explained in §§4.2. Given that now we have reduced estimating $\mathcal{J}_n(w,\mu,\nu)$ to estimating Euler characteristics of certain surfaces, in §§4.3 we formulate a topological result (Proposition 4.8) which suffices to prove Theorem 3.1. Proposition 4.8 is proved in §§4.5 using arguments related to Dehn's algorithm and the work of Birman—Series. The necessary additional background for this proof is given in §§4.4.

In §5 we show how Theorem 3.1 in conjunction with the results of the prequel [Mag21] prove Theorem 1.2.

1.4. **Notation.**

We write $\mathbf{N}$ for the natural numbers $\{1,2,3,\ldots\}$ and $\mathbf{N}_0 \stackrel{\text{def}}{=} \mathbf{N}\cup\{0\}$. We write $[n] \stackrel{\text{def}}{=} \{1,\ldots,n\}$ for $n\in\mathbf{N}$ and $[k,\ell] \stackrel{\text{def}}{=} \{k,k+1,\ldots,\ell\}$ for $k,\ell\in\mathbf{N}$. If A and B are two sets we write $A\backslash B$ for the elements of A not in B. If H is a group and $h_1,h_2\in H$ we write $[h_1,h_2] \stackrel{\text{def}}{=} h_1h_2h_1^{-1}h_2^{-1}$. We let id denote the identity element of a group. We let $[H,H]$ be the subgroup of H generated by elements of the form $[h_1,h_2]$; this is called the commutator subgroup of H. If V is a complex vector space, for $q\in\mathbf{N}_0$ we let

$$V^{\otimes q} \stackrel{\text{def}}{=} \underbrace{V\otimes V\otimes\cdots\otimes V}_{q}.$$

We use Vinogradov notation as follows. If f and h are functions of $n \in \mathbf{N}$, we write $f \ll h$ to mean that there are constants $n_0 \geq 0$ and $C_0 \geq 0$ such that for $n \geq n_0$, $|f(n)| \leq C_0 h(n)$. We write $f = O(h)$ to mean $f \ll h$. We write $f \asymp h$ to mean both $f \ll h$ and $h \ll f$. If in any of these statements the implied constants depend on additional parameters we add these parameters as subscript to $\ll, O$, or $\asymp$. Throughout the paper we view the genus g as fixed and so any implied constant may depend on g.

In this paper, Tr denotes the standard (unnormalized) trace on square complex matrices.

Acknowledgments. We thank Benoît Collins, Doron Puder, Sanjaye Ramgoolam, Calum Shearer, and Henry Wilton for valuable discussions about this work. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 949143).

2. Background

2.1. **Representation theory of symmetric groups.** Let S_k denote the symmetric group of permutations of $[k] \stackrel{\text{def}}{=} \{1, \ldots, k\}$, and $\mathbf{C}[S_k]$ denote its group algebra. The group S_0 is by definition the group with one element.

If we refer to $S_\ell \leq S_k$ with $\ell \leq k$ we always view S_ℓ as the subgroup of permutations that fix every element of $[\ell+1, k] \stackrel{\text{def}}{=} \{\ell+1, \ldots, k\}$. We write $S'_r \leq S_k$ for the subgroup of permutations that fix every element of $[k-r]$. As a consequence we obtain fixed inclusions $\mathbf{C}[S_\ell] \subset \mathbf{C}[S_k]$ for ℓ and k as above. When we write $S_\ell \times S_{k-\ell} \leq S_k$, the first factor is S_ℓ and the second factor is $S'_{k-\ell}$.

A *Young diagram* λ is a left-aligned contiguous collection of identical square boxes in the plane, such that the number of boxes in each row is non-increasing from top to bottom. We write λ_i for the number of boxes in the i^{th} row of λ and say $\lambda \vdash k$ if λ has k boxes. We write $\ell(\lambda)$ for the number of rows of λ. For each $\lambda \vdash k$ there is a *Young subgroup*

$$S_\lambda \stackrel{\text{def}}{=} S_{\lambda_1} \times S_{\lambda_2} \times \cdots \times S_{\lambda_{\ell(\lambda)}} \leq S_k$$

where the factors are subgroups in the obvious way, according to the increasing order of $[k]$.

The equivalence classes of irreducible representations of S_k are in one-to-one correspondence with Young diagrams $\lambda \vdash k$. Given λ, the construction of the corresponding irreducible representation V^λ can be done for example using Young symmetrizers as in [FH91, Lec. 4]. We write χ_λ for the character of S_k associated to V^λ and $d_\lambda \stackrel{\text{def}}{=} \chi_\lambda(\text{id}) = \dim V^\lambda$. Given $\lambda \vdash k$, the element

$$\mathfrak{p}_\lambda \stackrel{\text{def}}{=} \frac{d_\lambda}{k!} \sum_{\sigma \in S_k} \chi_\lambda(\sigma)\sigma \in \mathbf{C}[S_k]$$

is a central idempotent in $\mathbf{C}[S_k]$.

If G is a compact group, (ρ, W) is an irreducible representation of G, and (π, V) is any finite-dimensional representation of G, the (ρ, W)-isotypic subspace of (π, V) is the invariant subspace of V spanned by all irreducible direct summands of (π, V) that are isomorphic to (ρ, W). When ρ and π can be inferred from W and V we call this simply the *W-isotypic subspace of V*. If $H \leq G$ is a subgroup, and (ρ, W) is an irreducible representation of H, then the W-isotypic subspace of V for H is the W-isotypic subspace of the restriction of (π, V) to H.

If (π, V) is any finite-dimensional unitary representation of S_k, and $\lambda \vdash k$, then V is also a module for $\mathbf{C}[S_k]$ by linear extension of π and $\pi(\mathfrak{p}_\lambda)$ is the orthogonal projection onto the V^λ-isotypic subspace of V.

For any compact group G we write $(\mathrm{triv}_G, \mathbf{C})$ for the trivial representation of G. The following lemma can be deduced for example by combining Young's rule [FH91, Cor. 4.39] with Frobenius reciprocity.

Lemma 2.1. *Let $k \in \mathbf{N}_0$, and $\lambda \vdash k$. The space of vectors in V^λ fixed by S_λ is one-dimensional.*

2.2. **Representation theory of $\mathsf{U}(n)$ and $\mathsf{SU}(n)$.** Every irreducible representation of $\mathsf{U}(n)$ restricts to an irreducible representation of $\mathsf{SU}(n)$, and all equivalence classes of irreducible representations of $\mathsf{SU}(n)$ arise in this way. The equivalence classes of irreducible representations of $\mathsf{U}(n)$ are parameterized by dominant weights, that can be thought of as non-increasing sequences

$$\Lambda = (\Lambda_1, \ldots, \Lambda_n) \in \mathbf{Z}^n,$$

also known as *signatures.* We write W^Λ for the irreducible representation of $\mathsf{U}(n)$ corresponding to the signature Λ. Two irreducible representations of $\mathsf{U}(n)$ restrict to the same one of $\mathsf{SU}(n)$ if and only if their signatures differ by a constant vector. Let $\mathbb{T}(n)$ denote the maximal torus of $\mathsf{U}(n)$ consisting of diagonal matrices. Any matrix of $\mathbb{T}(n)$ has the form $\mathrm{diag}(\exp(i\theta_1), \ldots, \exp(i\theta_n))$ where all $\theta_j \in \mathbf{R}$. Associated to the signature Λ is the character ξ_Λ of $\mathbb{T}(n)$ given by

$$\xi_\Lambda\left(\mathrm{diag}(\exp(i\theta_1), \ldots, \exp(i\theta_n))\right) \stackrel{\mathrm{def}}{=} \exp\left(i\left(\sum_{j=1}^{n} \Lambda_j \theta_j\right)\right).$$

The highest weight theory says among other things that the ξ_Λ-isotypic subspace of W^Λ for $\mathbb{T}(n)$ is one-dimensional. Any vector in this subspace is called a *highest weight vector* of W^Λ.

Given $k, \ell \in \mathbf{N}_0$ and fixed Young diagrams $\mu \vdash k$, $\nu \vdash \ell$, we define a family of representations of $\mathsf{U}(n)$ as follows. For $n \geq \ell(\mu) + \ell(\nu)$ define

$$\Lambda_{\mu,\nu}(n) \stackrel{\mathrm{def}}{=} (\mu_1, \mu_2, \ldots, \mu_{\ell(\mu)}, \underbrace{0, \ldots, 0}_{n-\ell(\mu)-\ell(\nu)}, -\nu_{\ell(\nu)}, -\nu_{\ell(\nu)-1}, \ldots, -\nu_1).$$

We let $(\rho_n^{\mu,\nu}, W_n^{\mu,\nu})$ denote the irreducible representation of $\mathsf{U}(n)$ corresponding to $\Lambda_{\mu,\nu}(n)$ when $n \geq \ell(\mu) + \ell(\nu)$. We let $D_{\mu,\nu}(n) \stackrel{\mathrm{def}}{=} \dim W_n^{\mu,\nu}$ and

$s_{\mu,\nu}(g) \stackrel{\text{def}}{=} \mathrm{Tr}(\rho_n^{\mu,\nu}(g))$ for $g \in \mathsf{U}(n)$. If $\mu \vdash k$ and $\nu \vdash \ell$ then as $n \to \infty$

$$D_{\mu,\nu}(n) \asymp n^{k+\ell} \tag{2.1}$$

by [Mag21, Cor. 2.3] (alternatively [EI16, Lem. 3.5]).

We now present a version of Schur-Weyl duality for mixed tensors due to Koike [Koi89]. The very definition of $\mathsf{U}(n)$ makes $\mathbf{C}^n$ into a unitary representation of $\mathsf{U}(n)$ for the standard Hermitian inner product. We let $\{e_1, \ldots, e_n\}$ denote the standard basis of $\mathbf{C}^n$. If (ρ, W) is any finite dimensional representation of $\mathsf{U}(n)$ we write $(\rho^\vee, W^\vee)$ for the dual representation where $W^\vee$ is the space of complex linear functionals on W. The vector space $(\mathbf{C}^n)^\vee$ has a dual basis $\{\check{e}_1, \ldots, \check{e}_n\}$ given by $\check{e}_j(v) \stackrel{\text{def}}{=} \langle v, e_j \rangle$. Throughout the paper we frequently use certain canonical isomorphisms e.g.

$$\left((\mathbf{C}^n)^{\otimes p}\right)^\vee \cong \left((\mathbf{C}^n)^\vee\right)^{\otimes p}, \quad \mathrm{End}(W) \cong W \otimes W^\vee$$

to change points of view on representations; if we use non-canonical isomorphisms we point them out.

Let $\mathcal{T}_n^{k,\ell} \stackrel{\text{def}}{=} (\mathbf{C}^n)^{\otimes k} \otimes \left((\mathbf{C}^n)^\vee\right)^{\otimes \ell}$, with the convention that $(\mathbf{C}^n)^{\otimes 0} \stackrel{\text{def}}{=} \mathbf{C}$. With the natural inner product induced by that on $\mathbf{C}^n$, this is a unitary representation of $\mathsf{U}(n)$ under the diagonal action and also a unitary representation of $S_k \times S_\ell$ where S_k acts by permuting the indices of $(\mathbf{C}^n)^{\otimes k}$ and S_ℓ acts by permuting the indices of $\left((\mathbf{C}^n)^\vee\right)^{\otimes \ell}$. We write $\pi_n^{k,\ell} : \mathsf{U}(n) \to \mathrm{End}[\mathcal{T}_n^{k,\ell}]$ and $\rho_n^{k,\ell} : \mathbf{C}[S_k \times S_\ell] \to \mathrm{End}[\mathcal{T}_n^{k,\ell}]$ for these representations. The actions of $\mathsf{U}(n)$ and $S_k \times S_\ell$ on $\mathcal{T}_n^{k,\ell}$ commute. We use the notation, for $I = (i_1, \ldots, i_k) \in [n]^k$ and $J = (j_1, \ldots, j_\ell) \in [n]^\ell$

$$e_I \stackrel{\text{def}}{=} e_{i_1} \otimes \cdots \otimes e_{i_k} \in (\mathbf{C}^n)^{\otimes k}, \quad \check{e}_J \stackrel{\text{def}}{=} \check{e}_{j_1} \otimes \cdots \otimes \check{e}_{j_\ell} \in \left((\mathbf{C}^n)^\vee\right)^{\otimes \ell},$$
$$e_I^J \stackrel{\text{def}}{=} e_I \otimes \check{e}_J \in \mathcal{T}_n^{k,\ell}.$$

We write $I \sqcup J$ for the concatenation $(i_1, \ldots, i_k, j_1, \ldots, j_\ell)$.

For $k, \ell \geq 1$ let $\dot{\mathcal{T}}_n^{k,\ell}$ denote the intersection of the kernels of the mixed contractions $c_{pq} : \mathcal{T}_n^{k,\ell} \to \mathcal{T}^{k-1,\ell-1}$, $p \in [k]$, $q \in [\ell]$ given by

$$\begin{aligned} &c_{pq}(e_{i_1} \otimes \cdots \otimes e_{i_k} \otimes \check{e}_{j_1} \otimes \cdots \otimes \check{e}_{j_\ell}) \\ &\stackrel{\text{def}}{=} \delta_{i_p j_q} e_{i_1} \otimes \cdots \otimes e_{i_{p-1}} \otimes e_{i_{p+1}} \otimes \cdots \otimes e_{i_k} \\ &\quad \otimes \check{e}_{j_1} \otimes \cdots \otimes \check{e}_{j_{q-1}} \otimes \check{e}_{j_{q+1}} \otimes \cdots \otimes \check{e}_{j_\ell}, \end{aligned} \tag{2.2}$$

where $\delta_{i_p j_q}$ is the Kronecker delta. If $k = 1$ or $\ell = 1$ then the definition is extended in the natural way, interpreting an empty tensor of e_i or $\check{e}_i$ as 1. If either $k = 0$ or $\ell = 0$ then $\dot{\mathcal{T}}_n^{k,\ell} = \mathcal{T}_n^{k,\ell}$ by convention. The space $\dot{\mathcal{T}}_n^{k,\ell}$ is an invariant subspace under $\mathsf{U}(n) \times S_k \times S_\ell$ and hence a unitary subrepresentation of $\mathcal{T}_n^{k,\ell}$. On $\dot{\mathcal{T}}_n^{k,\ell}$ there is an analog of Schur–Weyl duality due to Koike.

Theorem 2.2. [Koi89, Thm. 1.1] *There is an isomorphism of unitary representations of* $\mathsf{U}(n)\times S_k\times S_\ell$

$$\dot{\mathcal{T}}_n^{k,\ell}\cong\bigoplus_{\substack{\mu\vdash k,\nu\vdash\ell\\ \ell(\mu)+\ell(\nu)\le n}}W_n^{\mu,\nu}\otimes V^\mu\otimes V^\nu. \tag{2.3}$$

Next we explain how to construct $\mathsf{U}(n)$-subrepresentations of $\dot{\mathcal{T}}_n^{k,\ell}$ isomorphic to $W_n^{\mu,\nu}$. Suppose that $\xi\in\dot{\mathcal{T}}_n^{k,\ell}$ is a non-zero vector such that under the isomorphism (2.3),

$$\xi\cong w\otimes v \tag{2.4}$$

for $w\in W_n^{\mu,\nu}$ and $v\in V^\mu\otimes V^\nu$. Then $\mathsf{U}(n)\cdot\xi$ linearly spans a $\mathsf{U}(n)$-subrepresentation of $\dot{\mathcal{T}}_n^{k,\ell}$ isomorphic to $W_n^{\mu,\nu}$. The following argument to construct such a vector ξ, given $\mu\vdash k,\nu\vdash\ell$, appears implicitly in [Koi89] and is elaborated in [BCH+94]. For $n\ge\ell(\mu)+\ell(\nu)$ let

$$\tilde{\theta}_{\mu,\nu}^n\stackrel{\text{def}}{=}e_1^{\otimes\mu_1}\otimes\cdots\otimes e_{\ell(\mu)}^{\otimes\mu_{\ell(\mu)}}\otimes(\check{e}_n)^{\otimes\nu_1}\otimes\cdots\otimes\left(\check{e}_{n-\ell(\nu)+1}\right)^{\otimes\nu_{\ell(\nu)}}. \tag{2.5}$$

This vector is in the $\xi_{\mu,\nu}$-isotypic subspace of $\dot{\mathcal{T}}_n^{k,\ell}$ for the maximal torus $\mathbb{T}(n)$ of $\mathsf{U}(n)$ where $\xi_{\mu,\nu}$ is the character of $\mathbb{T}(n)$ corresponding to the highest weight in $W_n^{\mu,\nu}$.

Let $\mathfrak{p}_\mu\in\mathbf{C}[S_k]$, $\mathfrak{p}_\nu\in\mathbf{C}[S_\ell]$ be the projections defined in §§2.1. Let $\rho_n^k:S_k\to\mathrm{End}(\mathcal{T}_n^{k,\ell})$ denote the representation of S_k described above and $\hat{\rho}_n^k:S_\ell\to\mathrm{End}(\mathcal{T}_n^{k,\ell})$ that of S_ℓ. Clearly these two representations commute. Now let

$$\theta_{\mu,\nu}^n\stackrel{\text{def}}{=}\rho_n^k(\mathfrak{p}_\mu)\hat{\rho}_n^\ell(\mathfrak{p}_\nu)\tilde{\theta}_{\mu,\nu}^n\in\dot{\mathcal{T}}_n^{k,\ell}. \tag{2.6}$$

Now this is in the same isotypic subspace for $\mathbb{T}(n)$ as before since $S_k\times S_\ell$ commutes with $\mathsf{U}(n)$. Moreover it is in the subspace of $\dot{\mathcal{T}}_n^{k,\ell}$ corresponding to $W_n^{\mu,\nu}\otimes V^\mu\otimes V^\nu$ under the isomorphism (2.3). The intersection of the two subspaces of $\dot{\mathcal{T}}_n^{k,\ell}$ just discussed corresponds via (2.3) to $\mathbf{C}w\otimes V^\mu\otimes V^\nu$ where w is a highest weight vector in $W_n^{\mu,\nu}$ and hence $\theta_{\mu,\nu}^n$ takes the form of (2.4) as we desired.

Of course we also want to know $\theta_{\mu,\nu}^n\neq 0$.

Lemma 2.3. *Suppose that* $k,\ell\in\mathbf{N}_0$, $\mu\vdash k,\nu\vdash\ell$, *and* $\theta_{\mu,\nu}^n$ *is as in (2.6) for* $n\ge\ell(\mu)+\ell(\nu)$. *We have*

$$\|\theta_{\mu,\nu}^n\|^2=\frac{d_\mu d_\nu}{[S_k:S_\mu][S_\ell:S_\nu]}.$$

Proof. Recall the definition of Young subgroups S_μ, S_ν from §§2.1. Letting $\tilde{\theta} = \tilde{\theta}^n_{\mu,\nu}$ (as in (2.5)) and $\theta = \theta^n_{\mu,\nu}$ we have

$$\theta = \rho_n^k(\mathfrak{p}_\mu)\hat{\rho}_n^\ell(\mathfrak{p}_\nu)\tilde{\theta} = \frac{d_\mu d_\nu}{k!\ell!} \sum_{\sigma=(\sigma_1,\sigma_2)\in S_k\times S_\ell} \chi_\mu(\sigma_1)\chi_\nu(\sigma_2)\rho_n^k(\sigma_1)\hat{\rho}_n^\ell(\sigma_2)\tilde{\theta}$$

$$= \frac{d_\mu d_\nu}{k!\ell!} \sum_{\substack{[\sigma_1]\in S_k/S_\mu \\ [\sigma_2]\in S_\ell/S_\nu}} \left(\sum_{\tau_1\in S_\mu} \chi_\mu(\sigma_1\tau_1)\right)\left(\sum_{\tau_2\in S_\nu} \chi_\nu(\sigma_2\tau_2)\right) \rho_n^k(\sigma_1)\hat{\rho}_n^\ell(\sigma_2)\tilde{\theta}.$$

The second equality used that $\tilde{\theta}$ is invariant under $S_\mu \times S_\nu$.

By Lemma 2.1, there is a one dimensional subspace of invariant vectors for S_μ in V^μ. If $v_\mu \in V^\mu$ is a unit vector in this space then

$$\sum_{\tau_1\in S_\mu} \chi_\mu(\sigma_1\tau_1) = |S_\mu|\langle \sigma_1 v_\mu, v_\mu\rangle. \tag{2.7}$$

Since the vectors $\rho_n^k(\sigma_1)\hat{\rho}_n^\ell(\sigma_2)\tilde{\theta}$ for $[\sigma_1] \in S_k/S_\mu$ and $[\sigma_2] \in S_\ell/S_\nu$ are orthogonal unit vectors, this gives

$$\|\theta\|^2 = \left(\frac{d_\mu d_\nu}{k!\ell!}\right)^2 \sum_{\substack{[\sigma_1]\in S_k/S_\mu \\ [\sigma_2]\in S_\ell/S_\nu}} \left(\sum_{\tau_1\in S_\mu} \chi_\mu(\sigma_1\tau_1)\right)^2 \left(\sum_{\tau_2\in S_\nu} \chi_\nu(\sigma_2\tau_2)\right)^2$$

$$\overset{(2.7)}{=} \left(\frac{d_\mu d_\nu}{k!\ell!}\right)^2 |S_\mu|^2|S_\nu|^2 \sum_{\substack{[\sigma_1]\in S_k/S_\mu \\ [\sigma_2]\in S_\ell/S_\nu}} |\langle \sigma_1 v_\mu, v_\mu\rangle|^2 \, |\langle \sigma_2 v_\nu, v_\nu\rangle|^2$$

$$= \left(\frac{d_\mu d_\nu}{k!\ell!}\right)^2 |S_\mu||S_\nu| \sum_{\substack{\sigma_1\in S_k \\ \sigma_2\in S_\ell}} |\langle \sigma_1 v_\mu, v_\mu\rangle|^2 \, |\langle \sigma_2 v_\nu, v_\nu\rangle|^2 = \frac{d_\mu d_\nu}{[S_k : S_\mu][S_\ell : S_\nu]}.$$

The last inequality used the orthogonality relations for matrix coefficients. □

Recall that we write $\pi_n^{k,\ell} : \mathsf{U}(n) \to \mathrm{End}(\mathcal{T}_n^{k,\ell})$ for the diagonal representation of $\mathsf{U}(n)$ on $\mathcal{T}_n^{k,\ell}$. Lemma 2.3 implies that $\theta^n_{\mu,\nu}$ is a non-zero vector. By the remarks following (2.6) it is of the pure tensor form $w \otimes v$ under the Schur-Weyl isomorphism (2.3), with $w \in W_n^{\mu,\nu}$, and hence we obtain the following corollary.

Corollary 2.4. *Suppose $n \geq \ell(\mu) + \ell(\nu)$. The subspace*

$$W_n(\theta^n_{\mu,\nu}) \overset{\mathrm{def}}{=} \mathrm{span}\{\pi_n^{k,\ell}(u)\theta^n_{\mu,\nu} \, : u \in \mathsf{U}(n)\,\} \subset \dot{\mathcal{T}}_n^{k,\ell}$$

is, under $\pi_n^{k,\ell}$, a $\mathsf{U}(n)$-subrepresentation of $\dot{\mathcal{T}}_n^{k,\ell}$ isomorphic to $W_n^{\mu,\nu}$.

2.3. **The Weingarten calculus.** The Weingarten calculus is a method based on Schur–Weyl duality that allows one to calculate integrals of products of matrix coefficients in the defining representation of $\mathsf{U}(n)$ in terms of sums over permutations. It was discovered initially by Weingarten [Wei78], and developed further in works of Xu, Collins, and Collins–Śniady in [Xu97, Col03, CŚ06].

We present two formulations of the Weingarten calculus. Given $k \in \mathbf{N}$, $n \in \mathbf{N}$, the *Weingarten function* with parameters n, k is the following element[2] of $\mathbf{C}[S_k]$ [CŚ06, eq. (9)]

$$\mathrm{Wg}_{n,k} \stackrel{\text{def}}{=} \frac{1}{(k!)^2} \sum_{\substack{\lambda \vdash k \\ \ell(\lambda) \leq n}} \frac{d_\lambda^2}{D_\lambda(n)} \sum_{\sigma \in S_k} \chi_\lambda(\sigma)\sigma. \tag{2.8}$$

We write $\mathrm{Wg}_{n,k}(\sigma)$ for the coefficient of σ in (2.8). The following theorem was proved by Collins and Śniady [CŚ06, Cor. 2.4].

Theorem 2.5. *For $k \in \mathbf{N}$ and for $i_1, i'_1, j_k, j'_k, \ldots, i_k, i'_k, j_k, j'_k \in [n]$*

$$\int_{u \in \mathsf{U}(n)} u_{i_1 j_1} \cdots u_{i_k j_k} \overline{u_{i'_1 j'_1}} \cdots \overline{u_{i'_k j'_k}} d\mu(u) \tag{2.9}$$

$$= \sum_{\sigma, \tau \in S_k} \delta_{i_1 i'_{\sigma(1)}} \cdots \delta_{i_k i'_{\sigma(k)}} \delta_{j_1 j'_{\tau(1)}} \cdots \delta_{j_k j'_{\tau(k)}} \mathrm{Wg}_{n,k}(\tau\sigma^{-1}),$$

where δ_{pq} is the Kronecker delta function.

It is sometimes more flexible to reformulate Theorem 2.5 in terms of projections. Here $u \in \mathsf{U}(n)$ acts on $A \in \mathrm{End}((\mathbf{C}^n)^{\otimes k})$ by $A \mapsto \pi_n^k(u) A \pi_n^k(u^{-1})$, $\pi_n^k : \mathsf{U}(n) \to \mathrm{End}((\mathbf{C}^n)^{\otimes k})$ the diagonal action. Write $P_{n,k}$ for the orthogonal projection in $\mathrm{End}((\mathbf{C}^n)^{\otimes k})$ onto the $\mathsf{U}(n)$-invariant vectors. The following proposition is due to Collins and Śniady [CŚ06, Prop. 2.3].

Proposition 2.6 (Collins–Śniady)**.** *Let $n, k \in \mathbf{N}$. Suppose $A \in \mathrm{End}((\mathbf{C}^n)^{\otimes k})$. Then*

$$P_{n,k}[A] = \rho_n^k \left(\Phi[A] \cdot \mathrm{Wg}_{n,k} \right)$$

where

$$\Phi[A] \stackrel{\text{def}}{=} \sum_{\sigma \in S_k} \mathrm{Tr}(A \rho_n^k(\sigma^{-1}))\sigma.$$

Later we will need the following bound for the Weingarten function due to Collins and Śniady [CŚ06, Prop. 2.6]. For a permutation σ let $|\sigma|$ denote the minimum number of transpositions that σ can be written as a product of.

Proposition 2.7. *For any fixed $\sigma \in S_k$, $\mathrm{Wg}_{n,k}(\sigma) \ll_k n^{-k-|\sigma|}$ as $n \to \infty$.*

[2] Although not relevant here, classically the Weingarten function arises as the multiplicative inverse of $\sum_{\sigma \in S_k} n^{\#\text{cycles}(\sigma)} \sigma$ in $\mathbf{C}[S_k]$,whenever $n \geq k$.

2.4. **Free groups and surface groups.** Let $\mathbf{F}_{2g} \stackrel{\text{def}}{=} \langle a_1, b_1, \ldots, a_g, b_g \rangle$ be the free group on $2g$ generators $a_1, b_1, \ldots, a_g, b_g$ and $R_g \stackrel{\text{def}}{=} [a_1, b_1] \cdots [a_g, b_g] \in \mathbf{F}_{2g}$. There is a quotient map $\mathbf{F}_{2g} \to \Gamma_g$ given by reduction modulo R_g. We say that $w \in \mathbf{F}_{2g}$ *represents the conjugacy class of* $\gamma \in \Gamma_g$ if the projection of w to Γ_g is in the conjugacy class of γ in Γ_g.

Given $w \in \mathbf{F}_{2g}$, we view w as a combinatorial word in $a_1, a_1^{-1}, b_1, b_1^{-1}, \ldots, a_g, a_g^{-1}, b_g, b_g^{-1}$ by writing it in reduced (shortest) form; i.e., a_1 does not follow a_1^{-1} etc. We say that w is *cyclically reduced* if the first letter of its reduced word is not the inverse of the last letter. The length $|w|$ of $w \in \mathbf{F}_{2g}$ is the length of its reduced form word. We say $w \in \mathbf{F}_{2g}$ is a *shortest element* representing the conjugacy class of $\gamma \in \Gamma_g$ if it has minimal length among all elements representing the conjugacy class of γ. If w is a shortest element representing some conjugacy class in Γ_g then w is cyclically reduced.

For any group H, the commutator subgroup $[H, H] \leq H$ is the subgroup generated by all elements of the form $[h_1, h_2] \stackrel{\text{def}}{=} h_1 h_2 h_1^{-1} h_2^{-1}$ with $h_1, h_2 \in H$. If $\gamma \in [\Gamma_g, \Gamma_g]$, and w represents the conjugacy class of γ, then $w \in [\mathbf{F}_{2g}, \mathbf{F}_{2g}]$ (see [Mag21, §2.6]).

2.5. **Witten zeta functions.** Witten zeta functions appeared first in Witten's work [Wit91] and were named by Zagier in [Zag94]. The *Witten zeta function of* $\mathsf{SU}(n)$ is defined, for s in a half-plane of convergence, by

$$\zeta(s; n) \stackrel{\text{def}}{=} \sum_{(\rho, W) \in \widehat{\mathsf{SU}(n)}} \frac{1}{(\dim W)^s} \tag{2.10}$$

where $\widehat{\mathsf{SU}(n)}$ denotes the equivalence classes of irreducible representations of $\mathsf{SU}(n)$. Indeed, the series (2.10) converges for $\mathrm{Re}(s) > \frac{2}{n}$ by a result of Larsen and Lubotzky [LL08, Thm. 5.1] (see also [HS19, §2]). Also relevant to this work is a result of Guralnick, Larsen, and Manack [GLM12, Thm 2., also eq. (7)] that states for fixed $s > 0$

$$\lim_{n \to \infty} \zeta(s; n) = 1. \tag{2.11}$$

2.6. **Results of the prequel paper.** By [Mag21, Prop. 1.5], if $\gamma \notin [\Gamma_g, \Gamma_g]$, then $\mathbb{E}_{g,n}[\mathrm{Tr}_\gamma] = 0$ for $n \geq n_0(\gamma)$. This proves Theorem 1.2 in this case. *Hence in the rest of the paper we need only consider* $\gamma \in [\Gamma_g, \Gamma_g]$ *and hence* $w \in [\mathbf{F}_{2g}, \mathbf{F}_{2g}]$ *if* $w \in \mathbf{F}_{2g}$ *represents the conjugacy class of* γ.

For each $w \in \mathbf{F}_{2g}$, we have a *word map* $w : \mathsf{U}(n)^{2g} \to \mathsf{U}(n)$ obtained by substituting matrices for the generators of $\mathbf{F}_{2g}$. For example, if $u_1, v_1, \ldots, u_g, v_g \in \mathsf{U}(n)$ then $R_g(u_1, v_1, \ldots, u_g, v_g) = [u_1, v_1] \cdots [u_g, v_g]$. We begin with the following result from the prequel paper [Mag21, Cor. 1.8].

Proposition 2.8. *Suppose that $g \geq 2$, $\gamma \in \Gamma_g$, and $w \in \mathbf{F}_{2g}$ represents the conjugacy class of γ. For any $B \in \mathbf{N}$ we have as $n \to \infty$*

$$\mathbb{E}_{g,n}[\mathrm{Tr}_\gamma] = \zeta(2g-2;n)^{-1} \sum_{\substack{\mu,\nu \text{ Young diagrams} \\ \ell(\mu),\ell(\nu) \leq B, \mu_1,\nu_1 \leq B^2}} D_{\mu,\nu}(n)\, \mathcal{I}_n(w,\mu,\nu) \quad (2.12)$$
$$+ O_{B,w,g}\left(n^{|w|} n^{-2\log B}\right).$$

where

$$\mathcal{I}_n(w,\mu,\nu) \stackrel{\text{def}}{=} \int_{\mathsf{SU}(n)^{2g}} \mathrm{Tr}(w(x)) \overline{s_{\mu,\nu}(R_g(x)))} d\mu(x). \quad (2.13)$$

Notice that for $n \geq 2B$ the right hand side of (2.12) makes sense, i.e. $D_{\mu,\nu}, s_{\mu,\nu}$ are well-defined. We also have the following proposition that follows from [Mag21, Prop. 3.1] together with $\overline{s_{\mu,\nu}} = s_{\nu,\mu}$.

Proposition 2.9. *Let $w \in [\mathbf{F}_{2g}, \mathbf{F}_{2g}]$. Then for any fixed μ,ν and $n \geq \ell(\mu) + \ell(\nu)$*

$$\mathcal{I}_n(w,\mu,\nu) = \mathcal{J}_n(w,\nu,\mu) \stackrel{\text{def}}{=} \int_{\mathsf{U}(n)^{2g}} \mathrm{Tr}(w(x)) s_{\nu,\mu}(R_g(x)) d\mu(x).$$

This is convenient as it will allow us to use the Weingarten calculus directly as it is presented in §§2.3 for $\mathsf{U}(n)$ rather than $\mathsf{SU}(n)$. By using Proposition 2.9, taking a representative $w \in \mathbf{F}_{2g}$ of the conjugacy class of γ and taking B such that $|w| - 2\log B \leq -1$ in Proposition 2.8 we obtain the following result from which we begin the new arguments of this paper.

Corollary 2.10. *Let $\gamma \in [\Gamma_g, \Gamma_g]$ and $w \in [\mathbf{F}_{2g}, \mathbf{F}_{2g}]$ be a representative of the conjugacy class of $\gamma \in \Gamma$. Then there exists a* finite *set $\tilde{\Omega}$ of pairs (μ,ν) of Young diagrams such that*

$$\mathbb{E}_{g,n}[\mathrm{Tr}_\gamma] = \zeta(2g-2;n)^{-1} \sum_{(\mu,\nu) \in \tilde{\Omega}} D_{\mu,\nu}(n) \mathcal{J}_n(w,\nu,\mu) + O_{w,g}\left(\frac{1}{n}\right).$$

As we know $\lim_{n\to\infty} \zeta(2g-2,n) = 1$ by (2.11), we have now reduced the proof of Theorem 1.2 to establishing suitable bounds for the integrals $\mathcal{J}_n(w,\mu,\nu)$ where we can view μ,ν as *fixed* Young diagrams since $\tilde{\Omega}$ is finite.

3. Combinatorial Integration

3.1. **Setup and motivation.**

The main result of the rest of the paper is the following.

Theorem 3.1. *Let $\gamma \in \Gamma_g$ with $\gamma \neq \mathrm{id}$. Let $w \in \mathbf{F}_{2g}$ be a shortest element representing the conjugacy class of γ. For each $k,\ell \in \mathbf{N}_0$ there is a constant $C(w,k,\ell) > 0$ such that for any $\mu \vdash k$, $\nu \vdash \ell$*

$$|D_{\mu,\nu}(n)\mathcal{J}_n(w,\mu,\nu)| \leq C(w,k,\ell)$$

for all $n \in \mathbf{N}$.

Accordingly, since we know the large n behavior of $D_{\mu,\nu}(n)$ from (2.1), in this section we wish to estimate

$$\mathcal{J}_n(w,\mu,\nu) = \int_{\mathsf{U}(n)^{2g}} \mathrm{Tr}(w(x)) s_{\mu,\nu}(R_g(x)) d\mu(x)$$

for fixed $\mu \vdash k, \nu \vdash \ell$.

What doesn't work. We begin by discussing why the most straightforward approach to this problem leads to serious complications. It is possible to approach the problem by writing $s_{\mu,\nu}(h)$ as a fixed finite linear combinations of functions

$$p_{\mu'}(h) p_{\nu'}(h^{-1})$$

where $p_{\mu'}(h)$ (resp. $p_{\nu'}(h^{-1})$) is a power sum symmetric polynomial of the eigenvalues of h (resp. h^{-1} or $\bar{h}$). See for example [Mag21, §§3.3] for one way to do this. The coefficients of this expansion are fixed, but not transparent, since they involve Littlewood–Richardson coefficients. In any case, this approach leads to writing $\mathcal{J}_n(w,\mu,\nu)$ as a finite linear combination of integrals of the form

$$\begin{aligned} &\int_{\mathsf{U}(n)^{2g}} \mathrm{Tr}(w(x)) \mathrm{Tr}\left(R_g(x)^{k_1}\right) \cdots \mathrm{Tr}\left(R_g(x)^{k_p}\right) \\ &\cdot \mathrm{Tr}\left(R_g(x)^{-\ell_1}\right) \cdots \mathrm{Tr}\left(R_g(x)^{-\ell_q}\right) d\mu(x) \end{aligned} \tag{3.1}$$

where $\sum k_j = |\mu|$ and $\sum \ell_j = |\nu|$.

The work of the author and Puder in [MP19] gives a full asymptotic expansion for (3.1) as $n \to \infty$. However these estimates are not sufficient for the current paper and to motivate the rest of this §3 we explain briefly the issues involved. However, this discussion is not needed to understand the arguments that we will make to prove Theorem 3.1.

The main result of [MP19] gives a full 'genus' expansion of (3.1) in terms of surfaces and maps on surfaces dictated by $w \in \mathbf{F}_{2g}$. Roughly speaking, every term in this expansion comes from a homotopy class of map f from an orientable surface Σ_f to $\bigvee_{i=1}^{2g} S^1$; to contribute to (3.1) the surface Σ_f has one boundary component that maps to w at the level of the fundamental groups, p boundary components that map respectively to $R_g^{k_1}, \ldots, R_g^{k_p}$ at the level of fundamental groups, and q boundary components that map respectively to $R_g^{-\ell_1}, \ldots, R_g^{-\ell_q}$ at the level of fundamental groups. The contribution of the pair (f, Σ_f) to (3.1) is of the form $c(f,\Sigma_f) n^{\chi(\Sigma_f)}$; the coefficient $c(f,\Sigma_f)$ is an Euler characteristic of a symmetry group of (f,Σ_f) and is not easy to calculate in general. However, one could still hope to get decay of (3.1) by controlling the possible $\chi(\Sigma_f)$ that could appear.

There are two issues with this. The first one is that if w is not the shortest element representing the conjugacy class of γ then we get bounds that are not helpful. For a very simple example, let $w = R_g^\ell$, $\gamma = \mathrm{id}_{\Gamma_g}$, and consider the potential contribution from $p = 0, q = 1, \ell_1 = \ell$. Then for any ν with $|\nu| = \ell$ there is contribution to $\mathcal{J}_n(w, \emptyset, \nu)$ that is a multiple of

$$\int_{\mathsf{U}(n)^{2g}} \mathrm{Tr}\left(R_g(x)^{\ell}\right) \mathrm{Tr}\left(R_g(x)^{-\ell}\right) d\mu(x).$$

Here, in the theory of [MP19] there is a (Σ_f, f) that is an annulus, one boundary component corresponding to $w = R_g^{\ell}$ and one corresponding to $R_g^{-\ell}$, so we can only bound the corresponding contribution to $D_{\emptyset,\nu}(n)\mathcal{J}_n(w, \emptyset, \nu)$ by using [MP19] on the order of $D_{\emptyset,\nu}(n) \asymp n^{\ell}$. On the other hand, any approach that works to establish Theorem 3.1 (for $\gamma \neq \mathrm{id}$) should extend to show when $\gamma = \mathrm{id}$, $D_{\emptyset,\nu}(n)\mathcal{J}_n(w, \emptyset, \nu) \ll n$ as $\mathbb{E}_{g,n}[\mathrm{Tr}_{\mathrm{id}}] = n$.

Indeed, this phenomenon extends to words of the form $w_0 R_g^{\ell}$ and more generally to words that are not shortest representatives of some conjugacy class in Γ_g. It means that even if we use something similar in spirit to [MP19], to prove Theorem 3.1 we must incorporate the theory of shortest representative words. This indeed takes place in §§4.3-§§4.5; the topological result proved there hinges on this theory.

The second issue is a little more subtle and only appears for 'mixed' representations, i.e., both $\mu, \nu \neq \emptyset$. In this case, suppose w is a shortest element representing some conjugacy class in Γ_g and $w \in [\mathbf{F}_{2g}, \mathbf{F}_{2g}]$. This means that there is a pair (f_0, Σ_{f_0}) where Σ_{f_0} has one boundary component that maps to w at the level of the fundamental groups. Let us take $\mu, \nu = (k), (k)$, i.e each Young diagram has one row of k boxes. This means we get a potential contribution to $D_{\mu,\nu}(n)\mathcal{J}_n(w, \mu, \nu)$ that is a constant multiple of

$$D_{(k),(k)}(n) \int_{\mathsf{U}(n)^{2g}} \mathrm{Tr}(w(x))\mathrm{Tr}\left(R_g(x)^{k}\right) \mathrm{Tr}\left(R_g(x)^{-k}\right) d\mu(x) \tag{3.2}$$

Now, for every $k \in \mathbf{N}$, there is (f, Σ_f) contributing to (3.2) with one component that is (f_0, Σ_{f_0}) and the other is an annulus with boundary components corresponding to R_g^k, R_g^{-k}. Since the annulus has Euler characteristic 0, and $D_{(k),(k)} \asymp n^{2k}$, the order of this contribution to $D_{(k),(k)}(n)\mathcal{J}_n(w, (k), (k))$ is potentially $\gg n^{2k} n^{\chi(\Sigma_{f_0})}$. For large enough k the exponent here is arbitrarily large, which is clearly catastrophic. In reality, this contribution must cancel with some other contribution but we do not know how to see these cancellations.

This ends the discussion of the difficulties of the most straightforward approach to the problems of this paper.

What does work. To bypass the previous issues we produce a refined version of the Weingarten calculus that leads to a restricted set of surfaces, for instance, not including the ones causing the problem above as well as all generalizations of this issue.

The basic approach is the following. Instead of trying to deal with a complicated formula for $s_{\mu,\nu}(R_2(x))$ (as above) we instead use the copy $W_n(\theta^n_{\mu,\nu})$

of $W_n^{\mu,\nu}$ in $\dot{\mathcal{T}}_n^{k,\ell}$ that we found in Corollary 2.4. In §§3.3 we compute the orthogonal projection $\mathfrak{q}_\theta$ from $\mathcal{T}_n^{k,\ell}$ (note; not $\dot{\mathcal{T}}_n^{k,\ell}$) onto $W_n(\theta_{\mu,\nu}^n)$ (Proposition 3.2). In the formula we obtain, we give bounds on the coefficients appearing therein (Lemma 3.3). In addition, we also remember that $q_\theta \in \mathrm{End}(\dot{\mathcal{T}}_n^{k,\ell})$; this fact is not obvious from our formula but turns out to be vital going forward.

The calculation of $\mathfrak{q}_\theta$ is extra to, but in the same spirit as, the vanilla Weingarten calculus, which is why we claim to have refined the Weingarten calculus here.

In the expression for $\mathcal{J}_n(w,\mu,\nu)$ we now write

$$s_{\mu,\nu}(R_2(x)) = \mathrm{Tr}_{\mathcal{T}_n^{k,\ell}}(A\mathfrak{q}_\theta B\mathfrak{q}_\theta A^{-1}\mathfrak{q}_\theta B^{-1}\mathfrak{q}_\theta C\mathfrak{q}_\theta D\mathfrak{q}_\theta C^{-1}\mathfrak{q}_\theta D^{-1}\mathfrak{q}_\theta)$$

where A,B,C,D are the images of the generators of Γ_2 under x. Then the entire integral of $\mathrm{Tr}(w(x))s_{\mu,\nu}(R_2(x))$ is done using the usual Weingarten calculus. The fact that $q_\theta \in \mathrm{End}(\dot{\mathcal{T}}_n^{k,\ell})$ intervenes at a critical point to show that certain contributions from the classical Weingarten calculus cancel and lead to restrictions on the non-zero contributions. Precisely, the restriction we obtain is summarized in the **forbidden matching** property below (§§3.4) and property **P4** (§§4.3).

3.2. **Proof of Theorem 3.1 when $k=\ell=0$.** Here we give a proof of Theorem 3.1 when $k=\ell=0$. This will allow us to bypass the slightly confusing issue of using the Weingarten function $\mathrm{Wg}_{n,k+\ell}$ when $k+\ell=0$ in §§3.3.

If $k=\ell=0$ then the only possible $\mu\vdash k$, $\nu\vdash\ell$ are empty Young diagrams $\mu=\nu=\emptyset$, and $W_n^{\emptyset,\emptyset}$ is the trivial representation of $\mathsf{U}(n)$, so $D_{\emptyset,\emptyset}(n)=1$ for all $n\geq 1$ and $s_{\emptyset,\emptyset}(h)=1$ for all $h\in\mathsf{U}(n)$. We then have

$$D_{\emptyset,\emptyset}(n)\mathcal{J}_n(w,\emptyset,\emptyset) = \mathcal{J}_n(w,\emptyset,\emptyset) = \int_{\mathsf{U}(n)^{2g}} \mathrm{Tr}(w(x))d\mu(x). \tag{3.3}$$

If $w\in\mathbf{F}_{2g}$ is a cyclically shortest word representing the conjugacy class of $\gamma\in\Gamma_g$ with $\gamma\neq\mathrm{id}$, then $w\neq\mathrm{id}$. It then follows from (1.1) that $D_{\emptyset,\emptyset}(n)\mathcal{J}_n(w,\emptyset,\emptyset)=o_w(n)$ as $n\to\infty$, but in fact, (3.3) is given by a rational function of n for $n\geq n_0(w)$ by a straightforward application of the Weingarten calculus [MP19]. This implies $D_{\emptyset,\emptyset}(n)\mathcal{J}_n(w,\emptyset,\emptyset)=O_w(1)$ as $n\to\infty$, as required.

This proves Theorem 3.1 when $k=\ell=0$. Hence in the rest of this §3 we can assume $k+\ell>0$.

3.3. **A projection formula.** Here we develop an integral calculus that is more powerful than the usual Weingarten calculus and allows us to directly tackle $\mathcal{J}_n(w,\mu,\nu)$ without writing it in terms of integrals as in (3.1). The key point is that our method leads to the **forbidden matchings** property of §§3.4 and property **P4** of §§4.3.

We now view k, ℓ, $\mu \vdash k, \nu \vdash \ell$ as fixed, assume $k+\ell > 0$, $n \geq \ell(\mu)+\ell(\nu)$, and write $\theta = \theta^n_{\mu,\nu}$ as in (2.6), suppressing the dependence on n. Let $W_n(\theta)$ be defined as in Corollary 2.4. Thus $W_n(\theta)$ is an irreducible summand of $\dot{\mathcal{T}}^{k,\ell}_n$ isomorphic to $W^{\mu,\nu}_n$ for the group $\mathsf{U}(n)$.

In the remainder of the paper we drop the dependence of our notation on n whenever it adds clarity.

Our first task is to compute the orthogonal projection $\mathfrak{q}_\theta$ onto $W(\theta)$. Let P_θ denote the orthogonal projection in $\mathcal{T}^{k,\ell}_n$ onto θ. We also view P_θ as an element of $\mathrm{End}(\dot{\mathcal{T}}^{k,\ell}_n)$ by restriction.

Under the canonical isomorphism $\mathrm{End}(\dot{\mathcal{T}}^{k,\ell}_n) \cong \dot{\mathcal{T}}^{k,\ell}_n \otimes \left(\dot{\mathcal{T}}^{k,\ell}_n\right)^\vee$ we have $P_\theta \cong \frac{\theta\otimes\theta^\vee}{\|\theta\|^2}$ and also from (2.6)

$$P_\theta = \frac{1}{\|\theta\|^2}\rho^k(\mathfrak{p}_\mu)\hat{\rho}^\ell(\mathfrak{p}_\nu)[\tilde{\theta}_{\mu,\nu} \otimes \tilde{\theta}^\vee_{\mu,\nu}]\rho^k(\mathfrak{p}_\mu)\hat{\rho}^\ell(\mathfrak{p}_\nu); \tag{3.4}$$

here the inner square bracket is interpreted as an element of $\mathrm{End}(\dot{\mathcal{T}}^{k,\ell}_n)$. By Schur's lemma we have

$$\mathfrak{q}_\theta = D_{\mu,\nu}(n)\int_{h\in\mathsf{U}(n)} \pi(h)P_\theta\pi(h^{-1})d\mu(h) \tag{3.5}$$

since the right hand side is an element of $\mathrm{End}(W(\theta)) \subset \mathrm{End}(\mathcal{T}^{k,\ell}_n)$ that commutes with $\pi^{k,\ell}(\mathsf{U}(n))$, so it is a multiple of $\mathfrak{q}_\theta$, and it has the correct trace.

On the other hand, we can view $\mathcal{T}^{k,\ell}_n \otimes \left(\dot{\mathcal{T}}^{k,\ell}_n\right)^\vee \cong \mathcal{T}^{k+\ell,k+\ell}_n$ by the canonical isomorphism

$$\mathcal{T}^{k,\ell}_n \otimes \left(\dot{\mathcal{T}}^{k,\ell}_n\right)^\vee \cong (\mathbf{C}^n)^{\otimes k} \otimes \left((\mathbf{C}^n)^{\otimes \ell}\right)^\vee \otimes \left((\mathbf{C}^n)^{\otimes k}\right)^\vee \otimes (\mathbf{C}^n)^{\otimes \ell}$$

followed by the following fixed isomorphism

$$\varphi : e^J_I \otimes \check{e}^{J'}_{I'} \mapsto e_{I\sqcup J'} \otimes \check{e}_{I'\sqcup J}. \tag{3.6}$$

Finally, there is a canonical isomorphism $\mathcal{T}^{k+\ell,k+\ell}_n \cong \mathrm{End}((\mathbf{C}^n)^{\otimes k+\ell})$. So combining these we fix isomorphisms

$$\mathrm{End}(\mathcal{T}^{k,\ell}_n) \cong \dot{\mathcal{T}}^{k,\ell}_n \otimes \left(\dot{\mathcal{T}}^{k,\ell}_n\right)^\vee \xrightarrow[\varphi]{\sim} \mathcal{T}^{k+\ell,k+\ell}_n \cong \mathrm{End}((\mathbf{C}^n)^{\otimes k+\ell}). \tag{3.7}$$

We view the outer two isomorphisms as fixed identifications. These isomorphisms are of unitary representations of $\mathsf{U}(n)$ when everything is given its natural inner product. Moreover for $\sigma = (\sigma_1, \sigma_2) \in S_k \times S_\ell$ and $\tau = (\tau_1, \tau_2) \in S_k \times S_l$ we have for $A \in \mathrm{End}(\mathcal{T}^{k,\ell}_n)$

$$\varphi[\rho^k(\sigma_1)\hat{\rho}^\ell(\sigma_2)A\rho^k(\tau_1)\hat{\rho}^\ell(\tau_2)] = \rho^{k+\ell}(\sigma_1, \tau_2^{-1})\varphi[A]\rho^{k+\ell}(\tau_1, \sigma_2^{-1}), \tag{3.8}$$

recalling that $\rho^{k+\ell} : \mathbf{C}[S_{k+\ell}] \to \mathrm{End}((\mathbf{C}^n)^{\otimes k+\ell})$ is the representation by permuting coordinates.

We now return to the calculation of $\mathfrak{q}_\theta$ in (3.5). We have

$$\mathfrak{q}_\theta = D_{\mu,\nu}(n)\varphi^{-1}[P_{n,k+\ell}[\varphi(P_\theta)]] \tag{3.9}$$

where $P_{n,k+\ell}$ is the projection onto the $\mathsf{U}(n)$-invariant vectors (by conjugation) in $\mathrm{End}((\mathbf{C}^n)^{\otimes k+\ell})$. This can now be done using the classical Weingarten calculus. By Proposition 2.6 we have

$$P_{n,k+\ell}[\varphi(P_\theta)] = \rho^{k+\ell}\left(\Phi[\varphi(P_\theta)]\cdot \mathrm{Wg}_{n,k+\ell}\right) \tag{3.10}$$

where

$$\Phi[\varphi(P_\theta)] = \sum_{\sigma\in S_{k+\ell}} \mathrm{Tr}(\varphi(P_\theta)\rho^{k+\ell}(\sigma^{-1}))\sigma.$$

By (3.8) and (3.4), and since e.g. $\chi_\mu(g) = \chi_\mu(g^{-1})$ we obtain

$$\begin{aligned}\varphi(P_\theta) &= \frac{1}{\|\theta\|^2}\varphi\left(\rho^k(\mathfrak{p}_\mu)\hat\rho^\ell(\mathfrak{p}_\nu)[\tilde\theta_{\mu,\nu}\otimes\tilde\theta^\vee_{\mu,\nu}]\rho^k(\mathfrak{p}_\mu)\hat\rho^\ell(\mathfrak{p}_\nu)\right)\\ &= \frac{1}{\|\theta\|^2}\rho^{k+\ell}(\mathfrak{p}_{\mu\otimes\nu})\varphi\left(\tilde\theta_{\mu,\nu}\otimes\tilde\theta^\vee_{\mu,\nu}\right)\rho^{k+\ell}(\mathfrak{p}_{\mu\otimes\nu})\end{aligned}$$

where

$$\mathfrak{p}_{\mu\otimes\nu} \stackrel{\text{def}}{=} \frac{d_\mu d_\nu}{k!\ell!}\sum_{\sigma=(\sigma_1,\sigma_2)\in S_k\times S_\ell}\chi_\mu(\sigma_1)\chi_\nu(\sigma_2)\sigma \in \mathbf{C}[S_{k+\ell}].$$

Now using that Φ is a $\mathbf{C}[S_{k+\ell}]$-bimodule morphism [CŚ06, Prop. 2.3 (1)] we obtain

$$\begin{aligned}\Phi[\varphi(P_\theta)] &= \frac{1}{\|\theta\|^2}\mathfrak{p}_{\mu\otimes\nu}\Phi\left[\varphi\left(\tilde\theta_{\mu,\nu}\otimes\tilde\theta^\vee_{\mu,\nu}\right)\right]\mathfrak{p}_{\mu\otimes\nu}\\ &= \frac{1}{\|\theta\|^2}\mathfrak{p}_{\mu\otimes\nu}\left(\sum_{\sigma\in S_{k+\ell}}\mathrm{Tr}\left(\varphi\left(\tilde\theta_{\mu,\nu}\otimes\tilde\theta^\vee_{\mu,\nu}\right)\rho^{k+\ell}\left(\sigma^{-1}\right)\right)\sigma\right)\mathfrak{p}_{\mu\otimes\nu}.\end{aligned}$$

Now, $\mathrm{Tr}(\varphi\left(\tilde\theta_{\mu,\nu}\otimes\tilde\theta^\vee_{\mu,\nu}\right)\rho^{k+\ell}(\sigma^{-1}))$ is equal to 1 if and only if σ is in $S_\mu\times S_\nu\le S_k\times S_\ell$, and is 0 otherwise. So we obtain

$$\Phi[\varphi(P_\theta)] = \frac{1}{\|\theta\|^2}\mathfrak{p}_{\mu\otimes\nu}\left(\sum_{\sigma\in S_\mu\times S_\nu}\sigma\right)\mathfrak{p}_{\mu\otimes\nu},$$

hence from (3.10)

$$P_{n,k+\ell}[\varphi(P_\theta)] = \rho^{k+\ell}\left(z_\theta\right)$$

where

$$z_\theta \stackrel{\text{def}}{=} \sum_{\tau\in S_{k+\ell}} z_\theta(\tau)\tau \stackrel{\text{def}}{=} \frac{1}{\|\theta\|^2}\mathfrak{p}_{\mu\otimes\nu}\left(\sum_{\sigma\in S_\mu\times S_\nu}\sigma\right)\mathfrak{p}_{\mu\otimes\nu}\mathrm{Wg}_{n,k+\ell}\in\mathbf{C}[S_{k+\ell}]. \tag{3.11}$$

Therefore we obtain the following proposition.

Proposition 3.2. *We have*

$$\mathfrak{q}_\theta = D_{\mu,\nu}(n)\varphi^{-1}[\rho^{k+\ell}(z_\theta)].$$

We can use the bound for the coefficients of $\mathrm{Wg}_{n,k+\ell}$ from Proposition 2.7 to infer a bound on the coefficients $z_\theta(\tau)$. For $\sigma \in S_{k+\ell}$, let $\|\sigma\|_{k,\ell}$ denote the minimum m for which

$$\sigma = \sigma_0 t_1 t_2 \cdots t_m$$

where $\sigma_0 \in S_k \times S_\ell$ and $t_1, \ldots, t_m$ are transpositions in $S_{k+\ell}$.

Lemma 3.3. *For all $\tau \in S_{k+\ell}$ and $\theta = \theta_{\mu,\nu}$ as above, $z_\theta(\tau) = O_{k,\ell}(n^{-k-\ell-\|\tau\|_{k,\ell}})$ as $n \to \infty$.*

Proof. Referring to (3.11), as $n \to \infty$, $\|\theta\|^{-2} = O_{k,\ell}(1)$ by Lemma 2.3 and the coefficients of
$\mathfrak{p}_{\mu\otimes\nu}\left(\sum_{\sigma\in S_\mu\times S_\nu}\sigma\right)\mathfrak{p}_{\mu\otimes\nu}$ are clearly $O_{k,\ell}(1)$, so z_θ has the form

$$\left(\sum_{\sigma\in S_k\times S_\ell} A(\sigma)\sigma\right)\mathrm{Wg}_{n,k+\ell}$$

where each $A(\sigma)$ is $O_{k,\ell}(1)$. This means

$$z_\theta(\tau) = \sum_{\sigma\in S_k\times S_\ell,\, \sigma'\in S_{k+\ell}\,:\,\sigma\sigma'=\tau} A(\sigma)\mathrm{Wg}_{n,k+\ell}(\sigma').$$

The order of any of the finitely many summands above is $n^{-k-\ell-|\sigma'|}$ by Proposition 2.7, and the minimum possible value of $|\sigma'|$ is $\|\tau\|_{k,\ell}$. □

Before moving on, it is useful to explain the operator $\varphi^{-1}[\rho^{k+\ell}(\pi)]$ for $\pi \in S_{k+\ell}$. For $I = (i_1, \ldots, i_{k+\ell})$ let $I'(I;\pi) \stackrel{\text{def}}{=} i_{\pi(1)}, \ldots, i_{\pi(k)}$ and $J'(I;\pi) \stackrel{\text{def}}{=} i_{\pi(k+1)}, \ldots, i_{\pi(k+\ell)}$. As an element of $(\mathbf{C}^n)^{\otimes k+\ell} \otimes \left((\mathbf{C}^n)^\vee\right)^{\otimes k+\ell}$, $\rho^{k+\ell}(\pi)$ is given by

$$\sum_{I=(i_1,\ldots,i_k),J=(j_{k+1},\ldots,j_{k+\ell})} e_{I'(I\sqcup J;\pi)\sqcup J'(I\sqcup J;\pi)} \otimes \check{e}_{I\sqcup J},$$

so from (3.6)

$$\varphi^{-1}[\rho^{k+\ell}(\pi)] = \sum_{I=(i_1,\ldots,i_k),J=(j_{k+1},\ldots,j_{k+\ell})} e^J_{I'(I\sqcup J;\pi)} \otimes \check{e}_I^{J'(I\sqcup J;\pi)}. \tag{3.12}$$

3.4. **A combinatorial integration formula.** In this rest of this §3 we assume $g = 2$. All proofs extend to $g \geq 3$. We write $\{a,b,c,d\}$ for the generators of $\mathbf{F}_4$ and $R \stackrel{\text{def}}{=} [a,b][c,d]$. Assume both γ and w are not the identity and $w \in [\mathbf{F}_4, \mathbf{F}_4]$ according to the remarks at the beginning of §§2.6. We write w in reduced form:

$$w = f_1^{\epsilon_1} f_2^{\epsilon_2} \ldots f_{|w|}^{\epsilon_{|w|}}, \quad \epsilon_u \in \{\pm 1\},\ f_u \in \{a,b,c,d\}, \tag{3.13}$$

where if $f_u = f_{u+1}$, then $\epsilon_u = \epsilon_{u+1}$. For $f \in \{a,b,c,d\}$ let p_f denote the number of occurrences of f^{+1} in (3.13). The expression (3.13) implies that for $h \stackrel{\text{def}}{=} (h_a, h_b, h_c, h_d) \in \mathsf{U}(n)^4$,

$$\operatorname{Tr}(w(h)) = \sum_{i_j \in [n]} (h_{f_1}^{\epsilon_1})_{i_1 i_2} (h_{f_2}^{\epsilon_2})_{i_2 i_3} \cdots (h_{f_{|w|}}^{\epsilon_{|w|}})_{i_{|w|} i_1}. \tag{3.14}$$

Working with this expression will be cumbersome so we explain a diagrammatic way to think about (3.14). This will be the starting point for how we eventually understand $\mathcal{J}_n(w, \mu, \nu)$ in terms of decorated surfaces. We begin with a collection of intervals as follows.

w-intervals and the w-loop

Firstly, for every $j \in [w]$, with $f_j = f$ as in (3.13) and $\epsilon_j = 1$ we take a copy of $[0,1]$ and direct it from 0 to 1.

In our constructions, every interval will have two directions: the *intrinsic direction* (which is the direction from 0 to 1) and the *assigned direction.* In the case just discussed, these agree, but in general they will not.

We write $[0,1]_{f,j,w}$ for such an interval and $\mathfrak{I}_{f,w}^{+}$ for the collection of these intervals.

For every $j \in [w]$, with $f_j = f$ as in (3.13) and $\epsilon_j = -1$ we take a copy of $[0,1]$ and direct this interval from 1 to 0. We write $[0,1]_{f^{-1},j,w}$ for such an interval and $\mathfrak{I}_{f,w}^{-}$ for the collection of these intervals.

All the intervals described above are called *w-intervals.* There are $|w|$ of these intervals in total.

w-intermediate-intervals.

Between each $[0,1]_{f_j^{\epsilon_j},j,w}$ and $[0,1]_{f_{j+1}^{\epsilon_{j+1}},j+1,w}$ we add a new interval connecting $1_{f_j^{\epsilon_j},j,w}$ to $0_{f_{j+1}^{\epsilon_{j+1}},j+1,w}$, where the indices j run mod $|w|$. These intervals added are called *w-intermediate-intervals.* Note that these intervals together with the w-intervals now form a closed cycle that is paved by $2|w|$ intervals alternating between w-intervals and w-intermediate-intervals. Starting at $[0,1]_{f_1^{\epsilon_1},1,w}$, reading the directions and f-labels of the w-intervals so that every w-interval is traversed from 0 to 1 spells out the word w. The resulting circle is called the w-loop and the previously defined orientation of this loop is now fixed. See Figure 3.1 for an illustration of the w-loop in a particular example.

We now view the indices i_j as an assignment

$$\mathbf{a} : \{\text{end-points of } w\text{-intervals}\} \to [n],$$

$$\mathbf{a}(0_{f,j,w}) \stackrel{\text{def}}{=} i_j, \mathbf{a}(1_{f,j,w}) = i_{j+1},\ \mathbf{a}(0_{f^{-1},j,w}) = i_j,\ \mathbf{a}(1_{f^{-1},j,w}) = i_{j+1}.$$

The condition that $\mathbf{a}$ comes from a single collection of i_j is precisely that *if two end points of w-intervals are connected by a w-intermediate-interval, they are assigned the same value by* $\mathbf{a}$. Let $\mathcal{A}(w)$ denote the collection of such $\mathbf{a}$. If I is any copy of $[0,1]$ we write 0_I for the copy of 0 and 1_I for the

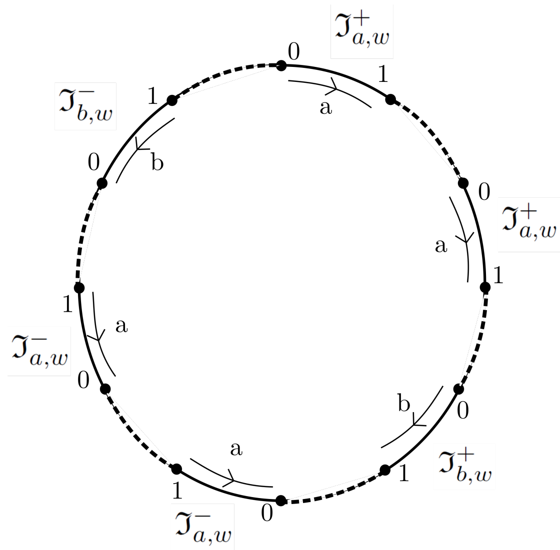

FIGURE 3.1. Illustration of the w-loop for $w = a^2ba^{-2}b^{-1}$. The solid intervals are w-intervals and the dashed intervals are w-intermediate-intervals. We also label each interval by the set e.g. $\mathfrak{I}^+_{a,w}$ to which they belong.

copy of 1 in I. We can now write

$$\mathrm{Tr}(w(h)) = \sum_{\mathbf{a}\in\mathcal{A}(w)} \prod_{f\in\{a,b,c,d\}} \left(\prod_{\mathfrak{i}\in\mathfrak{I}^+_{f,w}} h_{\mathbf{a}(0_\mathfrak{i})\mathbf{a}(1_\mathfrak{i})} \right) \left(\prod_{\mathfrak{j}\in\mathfrak{I}^-_{f,w}} \overline{h_{\mathbf{a}(1_\mathfrak{j})\mathbf{a}(0_\mathfrak{j})}} \right).$$

Now let v_p be an orthonormal basis for $W_n(\theta)$. We have

$$s_{\mu,\nu}(R_g(h_a,h_b,h_c,h_d)) = \sum_{p_i} \langle h_a v_{p_2}, v_{p_1}\rangle \langle h_b v_{p_3}, v_{p_2}\rangle \langle h_a^{-1} v_{p_4}, v_{p_3}\rangle \langle h_b^{-1} v_{p_5}, v_{p_4}\rangle$$
$$\langle h_c v_{p_6}, v_{p_5}\rangle \langle h_d v_{p_7}, v_{p_6}\rangle \langle h_c^{-1} v_{p_8}, v_{p_7}\rangle \langle h_d^{-1} v_{p_1}, v_{p_8}\rangle.$$

Here we have written e.g. $h_a v_{p_2}$ for $\pi_n^{k,\ell}(h_a) v_{p_2}$ to make things easier to read. Next we write each $v_p = \sum_{I,J} \beta^J_{pI} e^J_I$, where $\beta^J_{pI} \stackrel{\text{def}}{=} \langle v_p, e^J_I \rangle$. We then have

$$
\begin{aligned}
&\langle h_a v_{p_2}, v_{p_1}\rangle\langle h_b v_{p_3}, v_{p_2}\rangle\langle h_a^{-1} v_{p_4}, v_{p_3}\rangle\langle h_b^{-1} v_{p_5}, v_{p_4}\rangle \\
&\times \langle h_c v_{p_6}, v_{p_5}\rangle\langle h_d v_{p_7}, v_{p_6}\rangle\langle h_c^{-1} v_{p_8}, v_{p_7}\rangle\langle h_d^{-1} v_{p_1}, v_{p_8}\rangle \\
= & \sum_{\mathbf{r}_f,\mathbf{R}_f,\mathbf{V}_f,\mathbf{v}_f,\mathbf{U}_f,\mathbf{u}_f,\mathbf{s}_f,\mathbf{S}_f} \beta^{\mathbf{V}_a}_{p_2\mathbf{s}_a}\bar{\beta}^{\mathbf{U}_a}_{p_1\mathbf{r}_a}\beta^{\mathbf{V}_b}_{p_3\mathbf{s}_b}\bar{\beta}^{\mathbf{U}_b}_{p_2\mathbf{r}_b}\beta^{\mathbf{u}_a}_{p_4\mathbf{R}_a}\bar{\beta}^{\mathbf{v}_a}_{p_3\mathbf{S}_a}\beta^{\mathbf{u}_b}_{p_5\mathbf{R}_b}\bar{\beta}^{\mathbf{v}_b}_{p_4\mathbf{S}_b} \\
&\beta^{\mathbf{V}_c}_{p_6\mathbf{s}_c}\bar{\beta}^{\mathbf{U}_c}_{p_5\mathbf{r}_c}\beta^{\mathbf{V}_d}_{p_7\mathbf{s}_d}\bar{\beta}^{\mathbf{U}_d}_{p_6\mathbf{r}_d}\beta^{\mathbf{u}_c}_{p_8\mathbf{R}_c}\bar{\beta}^{\mathbf{v}_c}_{p_7\mathbf{S}_c}\beta^{\mathbf{u}_d}_{p_1\mathbf{R}_d}\bar{\beta}^{\mathbf{v}_d}_{p_8\mathbf{S}_d} \\
&\langle h_a e^{\mathbf{V}_a}_{\mathbf{s}_a}, e^{\mathbf{U}_a}_{\mathbf{r}_a}\rangle\langle h_b e^{\mathbf{V}_b}_{\mathbf{s}_b}, e^{\mathbf{U}_b}_{\mathbf{r}_b}\rangle\langle h_a^{-1} e^{\mathbf{u}_a}_{\mathbf{R}_a}, e^{\mathbf{v}_a}_{\mathbf{S}_a}\rangle\langle h_b^{-1} e^{\mathbf{u}_b}_{\mathbf{R}_b}, e^{\mathbf{v}_b}_{\mathbf{S}_b}\rangle \\
&\langle h_c e^{\mathbf{V}_c}_{\mathbf{s}_c}, e^{\mathbf{U}_c}_{\mathbf{r}_c}\rangle\langle h_d e^{\mathbf{V}_d}_{\mathbf{s}_d}, e^{\mathbf{U}_d}_{\mathbf{r}_d}\rangle\langle h_c^{-1} e^{\mathbf{u}_c}_{\mathbf{R}_c}, e^{\mathbf{v}_c}_{\mathbf{S}_c}\rangle\langle h_d^{-1} e^{\mathbf{u}_d}_{\mathbf{R}_d}, e^{\mathbf{v}_d}_{\mathbf{S}_d}\rangle. \qquad (3.15)
\end{aligned}
$$

We calculate

$$
\begin{aligned}
&\langle h_f e^{\mathbf{V}_f}_{\mathbf{s}_f}, e^{\mathbf{U}_f}_{\mathbf{r}_f}\rangle\langle h_f^{-1} e^{\mathbf{u}_f}_{\mathbf{R}_f}, e^{\mathbf{v}_f}_{\mathbf{S}_f}\rangle \\
=&\langle h_f e_{\mathbf{s}_f}, e_{\mathbf{r}_f}\rangle\overline{\langle h_f e_{\mathbf{V}_f}, e_{\mathbf{U}_f}\rangle\langle h_f e_{\mathbf{S}_f}, e_{\mathbf{R}_f}\rangle}\langle h_f e_{\mathbf{v}_f}, e_{\mathbf{u}_f}\rangle \\
=&\langle h_f e_{\mathbf{s}_f\sqcup\mathbf{v}_f}, e_{\mathbf{r}_f\sqcup\mathbf{u}_f}\rangle\overline{\langle h_f e_{\mathbf{S}_f\sqcup\mathbf{V}_f}, e_{\mathbf{R}_f\sqcup\mathbf{U}_f}\rangle}. \qquad (3.16)
\end{aligned}
$$

We now want a diagrammatic interpretation of (3.15) similarly to before. We make the following constructions.

R-intervals.

For each $j \in [k]$ and $f \in \{a,b,c,d\}$, we make a copy of $[0,1]$, direct it from 0 to 1, label it by f, and also number it by j. We write $\mathfrak{I}^{+}_{f,R}$ for the collection of these intervals. These correspond to occurrences of f in R.

For each $j \in [k]$ and $f \in \{a,b,c,d\}$, we make a copy of $[0,1]$, direct it from 1 to 0, label it by f, and also number it by j. We write $\mathfrak{I}^{-}_{f,R}$ for the collection of these intervals. These correspond to occurrences of f^{-1} in R.

(These two constructions of k intervals correspond to the presence of f and f^{-1} each exactly once in R.)

These intervals are called *R-intervals.* There are $8k$ R-intervals in total (for general g, there are $4gk$ of these intervals).

R^{-1}-intervals.

For each $j \in [k+1, k+\ell]$ and $f \in \{a,b,c,d\}$, we make a copy of $[0,1]$, direct it from 0 to 1, label it by f, and also number it by j. We write $\mathfrak{I}^{+}_{f,R^{-1}}$ for the collection of these intervals. These correspond to occurrences of f in R^{-1}.

For each $j \in [k+1, k+\ell]$ and $f \in \{a,b,c,d\}$, we make a copy of $[0,1]$, direct it from 1 to 0, label it by f, and also number it by j. We write $\mathfrak{I}^{-}_{f,R^{-1}}$ for the collection of these intervals. These correspond to occurrences of f^{-1} in R^{-1}.

These intervals are called *R^{-1}-intervals.* There are 8ℓ R^{-1} intervals in total (for general g, there are $4g\ell$ of these intervals). See Figure 3.2 for an illustration of the R and R^{-1} intervals.

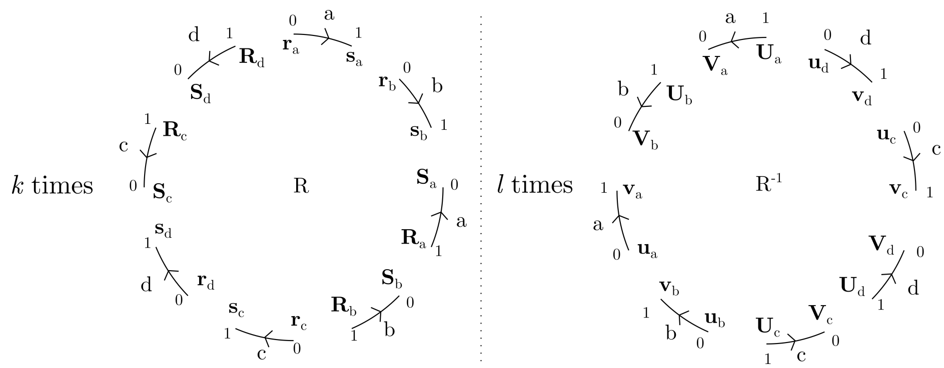

FIGURE 3.2. Here is shown the R-intervals (left) and the R^{-1}-intervals (right). We have indicated their assigned direction and label (which f they correspond to). We have also, for each endpoint of an interval, indicated which index function, e.g. $\mathbf{r}_a$, has this endpoint in its domain.

We now view (by identifying endpoints of intervals with the given numbers of intervals in $[k+\ell]$)

$$\begin{aligned} \mathbf{r}_f &: \{0_{\mathfrak{i}} : \mathfrak{i} \in \mathfrak{I}^+_{f,R}\} \to [n], \quad \mathbf{R}_f : \{1_{\mathfrak{i}} : \mathfrak{i} \in \mathfrak{I}^-_{f,R}\} \to [n], \\ \mathbf{s}_f &: \{1_{\mathfrak{i}} : \mathfrak{i} \in \mathfrak{I}^+_{f,R}\} \to [n], \quad \mathbf{S}_f : \{0_{\mathfrak{i}} : \mathfrak{i} \in \mathfrak{I}^-_{f,R}\} \to [n], \\ \mathbf{U}_f &: \{1_{\mathfrak{i}} : \mathfrak{i} \in \mathfrak{I}^-_{f,R^{-1}}\} \to [n], \quad \mathbf{u}_f : \{0_{\mathfrak{i}} : \mathfrak{i} \in \mathfrak{I}^+_{f,R^{-1}}\} \to [n], \\ \mathbf{V}_f &: \{0_{\mathfrak{i}} : \mathfrak{i} \in \mathfrak{I}^-_{f,R^{-1}}\} \to [n], \quad \mathbf{v}_f : \{1_{\mathfrak{i}} : \mathfrak{i} \in \mathfrak{I}^+_{f,R^{-1}}\} \to [n]. \end{aligned}$$

We obtain from (3.16)

$$\begin{aligned} &\langle h_a e^{\mathbf{V}_a}_{\mathbf{s}_a}, e^{\mathbf{U}_a}_{\mathbf{r}_a}\rangle \langle h_b e^{\mathbf{V}_b}_{\mathbf{s}_b}, e^{\mathbf{U}_b}_{\mathbf{r}_b}\rangle \langle h_a^{-1} e^{\mathbf{u}_a}_{\mathbf{R}_a}, e^{\mathbf{v}_a}_{\mathbf{S}_a}\rangle \langle h_b^{-1} e^{\mathbf{u}_b}_{\mathbf{R}_b}, e^{\mathbf{v}_b}_{\mathbf{S}_b}\rangle \\ &\langle h_c e^{\mathbf{V}_c}_{\mathbf{s}_c}, e^{\mathbf{U}_c}_{\mathbf{r}_c}\rangle \langle h_d e^{\mathbf{V}_d}_{\mathbf{s}_d}, e^{\mathbf{U}_d}_{\mathbf{r}_d}\rangle \langle h_c^{-1} e^{\mathbf{u}_c}_{\mathbf{R}_c}, e^{\mathbf{v}_c}_{\mathbf{S}_c}\rangle \langle h_d^{-1} e^{\mathbf{u}_d}_{\mathbf{R}_d}, e^{\mathbf{v}_d}_{\mathbf{S}_d}\rangle \\ = &\prod_f \prod_{\mathfrak{i}^+ \in \mathfrak{I}^+_{f,R}} \prod_{\mathfrak{i}^- \in \mathfrak{I}^-_{f,R}} \prod_{j^+ \in \mathfrak{I}^+_{f,R^{-1}}} \prod_{j^- \in \mathfrak{I}^-_{f,R^{-1}}} \\ &h_{\mathbf{r}_f(0_{\mathfrak{i}^+})\mathbf{s}_f(1_{\mathfrak{i}^+})} h_{\mathbf{u}_f(0_{\mathfrak{j}^+})\mathbf{v}_f(1_{\mathfrak{j}^+})} \bar{h}_{\mathbf{R}_f(1_{\mathfrak{i}^-})\mathbf{S}_f(0_{\mathfrak{i}^-})} \bar{h}_{\mathbf{U}_f(1_{\mathfrak{j}^-})\mathbf{V}_f(0_{\mathfrak{j}^-})}. \end{aligned}$$

With this formalism we obtain

$$\begin{aligned}
&\mathcal{J}_n(w,\mu,\nu)\\
=&\sum_{p_i}\sum_{\mathbf{r}_f,\mathbf{R}_f,\mathbf{V}_f,\mathbf{v}_f,\mathbf{U}_f,\mathbf{u}_f,\mathbf{s}_f,\mathbf{S}_f}\sum_{\mathbf{a}\in\mathcal{A}(w)}\\
&\beta^{\mathbf{V}_a}_{p_2\mathbf{s}_a}\bar{\beta}^{\mathbf{U}_a}_{p_1\mathbf{r}_a}\beta^{\mathbf{V}_b}_{p_3\mathbf{s}_b}\bar{\beta}^{\mathbf{U}_b}_{p_2\mathbf{r}_b}\beta^{\mathbf{u}_a}_{p_4\mathbf{R}_a}\bar{\beta}^{\mathbf{v}_a}_{p_3\mathbf{S}_a}\beta^{\mathbf{u}_b}_{p_5\mathbf{R}_b}\bar{\beta}^{\mathbf{v}_b}_{p_4\mathbf{S}_b}\\
\times&\beta^{\mathbf{V}_c}_{p_6\mathbf{s}_c}\bar{\beta}^{\mathbf{U}_c}_{p_5\mathbf{r}_c}\beta^{\mathbf{V}_d}_{p_7\mathbf{s}_d}\bar{\beta}^{\mathbf{U}_d}_{p_6\mathbf{r}_d}\beta^{\mathbf{u}_c}_{p_8\mathbf{R}_c}\bar{\beta}^{\mathbf{v}_c}_{p_7\mathbf{S}_c}\beta^{\mathbf{u}_d}_{p_1\mathbf{R}_d}\bar{\beta}^{\mathbf{v}_d}_{p_8\mathbf{S}_d}\\
&\prod_{f\in\{a,b,c,d\}}\\
&\int_{h\in\mathsf{U}(n)}\prod_{\mathfrak{i}\in\mathfrak{I}^+_{f,w}}\prod_{\mathfrak{j}\in\mathfrak{I}^-_{f,w}}\prod_{\mathfrak{i}^+\in\mathfrak{I}^+_{f,R}}\prod_{\mathfrak{i}^-\in\mathfrak{I}^-_{f,R}}\prod_{j^+\in\mathfrak{I}^+_{f,R^{-1}}}\prod_{j^-\in\mathfrak{I}^-_{f,R^{-1}}}\\
&h_{\mathbf{r}_f(0_{\mathfrak{i}^+})\mathbf{s}_f(1_{\mathfrak{i}^+})}h_{\mathbf{u}_f(0_{\mathfrak{j}^+})\mathbf{v}_f(1_{\mathfrak{j}^+})}\bar{h}_{\mathbf{R}_f(1_{\mathfrak{i}^-})\mathbf{S}_f(0_{\mathfrak{i}^-})}\bar{h}_{\mathbf{U}_f(1_{\mathfrak{j}^-})\mathbf{V}_f(0_{\mathfrak{j}^-})}dh.
\end{aligned}\tag{3.17}$$

For each f, the integral in (3.17) can be done using the Weingarten calculus (Theorem 2.5). To do this, fix bijections for each $f\in\{a,b,c,d\}$

$$\mathfrak{I}^+_f\stackrel{\text{def}}{=}\mathfrak{I}^+_{f,R}\cup\mathfrak{I}^+_{f,R^{-1}}\cup\mathfrak{I}^+_{f,w}\cong[k+\ell+p_f]$$

$$\mathfrak{I}^-_f\stackrel{\text{def}}{=}\mathfrak{I}^-_{f,R}\cup\mathfrak{I}^-_{f,R^{-1}}\cup\mathfrak{I}^-_{f,w}\cong[k+\ell+p_f]$$

such that

$$\mathfrak{I}^+_{f,w}\cong[k+\ell+1,k+\ell+p_f],\ \mathfrak{I}^-_{f,w}\cong[k+\ell+1,k+\ell+p_f]$$

and

$$\mathfrak{I}^+_{f,R}\cong[k],\ \mathfrak{I}^-_{f,R}\cong[k],\ \mathfrak{I}^+_{f,R^{-1}}\cong[k+1,k+\ell],\ \mathfrak{I}^-_{f,R^{-1}}\cong[k+1,k+\ell]\tag{3.18}$$

correspond to the original numberings of $\mathfrak{I}^+_{f,R}$, $\mathfrak{I}^-_{f,R}$, $\mathfrak{I}^+_{f,R^{-1}}$, $\mathfrak{I}^-_{f,R^{-1}}$.

Hence if $\sigma_f,\tau_f\in S_{k+\ell+p_f}$ we view $\sigma_f,\tau_f:\mathfrak{I}^+_f\to\mathfrak{I}^-_f$ by the above fixed bijections. For each $f\in\{a,b,c,d\}$ we say $(\mathbf{a},\mathbf{r}_f,\mathbf{u}_f,\mathbf{R}_f,\mathbf{U}_f)\to\sigma_f$ if for all $\mathfrak{i}\in\mathfrak{I}^+_f$, $\mathfrak{i}'\in\mathfrak{I}^-_f$ with $\sigma_f(\mathfrak{i})=\mathfrak{i}'$, we have

$$[\mathbf{r}_f\sqcup\mathbf{u}_f\sqcup\mathbf{a}](0_{\mathfrak{i}})=[\mathbf{R}_f\sqcup\mathbf{U}_f\sqcup\mathbf{a}](1_{\mathfrak{i}'});$$

here we wrote e.g. $[\mathbf{r}_f\sqcup\mathbf{u}_f\sqcup\mathbf{a}]$ for the function that $\mathbf{a},\mathbf{r}_f,\mathbf{u}_f$ induce on $\{0_{\mathfrak{i}}:\mathfrak{i}\in\mathfrak{I}^+_f\}$. Similarly we say $(\mathbf{a},\mathbf{s}_f,\mathbf{v}_f,\mathbf{S}_f,\mathbf{V}_f)\to\tau_f$ if for all $\mathfrak{i}\in\mathfrak{I}^+_f,\mathfrak{i}'\in\mathfrak{I}^-_f$ with $\tau_f(\mathfrak{i})=\mathfrak{i}'$ we have

$$[\mathbf{s}_f\sqcup\mathbf{v}_f\sqcup\mathbf{a}](1_{\mathfrak{i}})=[\mathbf{S}_f\sqcup\mathbf{V}_f\sqcup\mathbf{a}](0_{\mathfrak{i}'}).$$

Theorem 2.5 translates to

$$\int_{h\in\mathsf{U}(n)} \prod_{\mathfrak{i}\in\mathfrak{I}^+_{f,w}} \prod_{\mathfrak{j}\in\mathfrak{I}^-_{f,w}} \prod_{\mathfrak{i}^+\in\mathfrak{I}^+_{f,R}} \prod_{\mathfrak{i}^-\in\mathfrak{I}^-_{f,R}} \prod_{j^+\in\mathfrak{I}^+_{f,R^{-1}}} \prod_{j^-\in\mathfrak{I}^-_{f,R^{-1}}}$$
$$h_{\mathbf{r}_f(0_{\mathfrak{i}^+})\mathbf{s}_f(1_{\mathfrak{i}^+})} h_{\mathbf{u}_f(0_{\mathfrak{j}^+})\mathbf{v}_f(1_{\mathfrak{j}^+})} \bar{h}_{\mathbf{R}_f(1_{\mathfrak{i}^-})\mathbf{S}_f(0_{\mathfrak{i}^-})} \bar{h}_{\mathbf{U}_f(1_{\mathfrak{j}^-})\mathbf{V}_f(0_{\mathfrak{j}^-})} dh$$
$$= \sum_{\sigma_f,\tau_f\in S_{k+\ell+p_f}} \mathrm{Wg}_{n,k+\ell+p_f}(\sigma_f\tau_f^{-1})$$
$$\times \mathbf{1}\{(\mathbf{a},\mathbf{r}_f,\mathbf{u}_f,\mathbf{R}_f,\mathbf{U}_f)\to\sigma_f, (\mathbf{a},\mathbf{s}_f,\mathbf{v}_f,\mathbf{S}_f,\mathbf{V}_f)\to\tau_f\},$$

so putting this into (3.17) gives

$$\mathcal{J}_n(w,\mu,\nu) = \sum_{\sigma_f,\tau_f\in S_{k+\ell+p_f}} \left(\prod_{f\in\{a,b,c,d\}} \mathrm{Wg}_{n,k+\ell+p_f}(\sigma_f\tau_f^{-1})\right)$$
$$\sum_{p_i} \sum_{\substack{\mathbf{a}\in\mathcal{A}(w),\mathbf{r}_f,\mathbf{R}_f,\mathbf{V}_f,\mathbf{v}_f,\mathbf{U}_f,\mathbf{u}_f,\mathbf{s}_f,\mathbf{S}_f\\ (\mathbf{a},\mathbf{r}_f,\mathbf{u}_f,\mathbf{R}_f,\mathbf{U}_f)\to\sigma_f\\ (\mathbf{a},\mathbf{s}_f,\mathbf{v}_f,\mathbf{S}_f,\mathbf{V}_f)\to\tau_f}}$$
$$\beta^{\mathbf{V}_a}_{p_2\mathbf{s}_a}\bar{\beta}^{\mathbf{U}_a}_{p_1\mathbf{r}_a}\beta^{\mathbf{V}_b}_{p_3\mathbf{s}_b}\bar{\beta}^{\mathbf{U}_b}_{p_2\mathbf{r}_b}\beta^{\mathbf{u}_a}_{p_4\mathbf{R}_a}\bar{\beta}^{\mathbf{v}_a}_{p_3\mathbf{S}_a}\beta^{\mathbf{u}_b}_{p_5\mathbf{R}_b}\bar{\beta}^{\mathbf{v}_b}_{p_4\mathbf{S}_b}$$
$$\times \beta^{\mathbf{V}_c}_{p_6\mathbf{s}_c}\bar{\beta}^{\mathbf{U}_c}_{p_5\mathbf{r}_c}\beta^{\mathbf{V}_d}_{p_7\mathbf{s}_d}\bar{\beta}^{\mathbf{U}_d}_{p_6\mathbf{r}_d}\beta^{\mathbf{u}_c}_{p_8\mathbf{R}_c}\bar{\beta}^{\mathbf{v}_c}_{p_7\mathbf{S}_c}\beta^{\mathbf{u}_d}_{p_1\mathbf{R}_d}\bar{\beta}^{\mathbf{v}_d}_{p_8\mathbf{S}_d}.$$

Here we make our main improvement over the classical Weingarten calculus. We introduce the following beneficial property that the σ_f,τ_f possibly have.

Forbidden matchings property: For every $f\in\{a,b,c,d\}$ the following hold: neither σ_f nor τ_f map any element of $\mathfrak{I}^+_{f,R}$ to an element of $\mathfrak{I}^-_{f,R^{-1}}$, or map an element of $\mathfrak{I}^+_{f,R^{-1}}$ to an element of $\mathfrak{I}^-_{f,R}$.

We have the following key lemma.

Lemma 3.4. *If for some $f\in\{a,b,c,d\}$, σ_f and τ_f do not have the **forbidden matchings** property, then for any choice of $p_1,\ldots,p_8$*

$$\sum_{\substack{\mathbf{a}\in\mathcal{A}(w),\mathbf{r}_f,\mathbf{R}_f,\mathbf{V}_f,\mathbf{v}_f,\mathbf{U}_f,\mathbf{u}_f,\mathbf{s}_f,\mathbf{S}_f\\ (\mathbf{a},\mathbf{r}_f,\mathbf{u}_f,\mathbf{R}_f,\mathbf{U}_f)\to\sigma_f\\ (\mathbf{a},\mathbf{s}_f,\mathbf{v}_f,\mathbf{S}_f,\mathbf{V}_f)\to\tau_f}} \tag{3.19}$$
$$\beta^{\mathbf{V}_a}_{p_2\mathbf{s}_a}\bar{\beta}^{\mathbf{U}_a}_{p_1\mathbf{r}_a}\beta^{\mathbf{V}_b}_{p_3\mathbf{s}_b}\bar{\beta}^{\mathbf{U}_b}_{p_2\mathbf{r}_b}\beta^{\mathbf{u}_a}_{p_4\mathbf{R}_a}\bar{\beta}^{\mathbf{v}_a}_{p_3\mathbf{S}_a}\beta^{\mathbf{u}_b}_{p_5\mathbf{R}_b}\bar{\beta}^{\mathbf{v}_b}_{p_4\mathbf{S}_b}$$
$$\times \beta^{\mathbf{V}_c}_{p_6\mathbf{s}_c}\bar{\beta}^{\mathbf{U}_c}_{p_5\mathbf{r}_c}\beta^{\mathbf{V}_d}_{p_7\mathbf{s}_d}\bar{\beta}^{\mathbf{U}_d}_{p_6\mathbf{r}_d}\beta^{\mathbf{u}_c}_{p_8\mathbf{R}_c}\bar{\beta}^{\mathbf{v}_c}_{p_7\mathbf{S}_c}\beta^{\mathbf{u}_d}_{p_1\mathbf{R}_d}\bar{\beta}^{\mathbf{v}_d}_{p_8\mathbf{S}_d}$$
$$=0.$$

Proof. Indeed suppose σ_a matches an element $\mathfrak{i}\in\mathfrak{I}^+_{a,R}$ with $\mathfrak{j}\in\mathfrak{I}^-_{a,R^{-1}}$; $\sigma_a(\mathfrak{i})=\mathfrak{j}$. With our given fixed bijections (3.18), $\mathfrak{i}$ corresponds to an element of $[k]$ and $\mathfrak{j}$ corresponds to an element of $[k+1,k+\ell]$. Without loss of generality in the argument suppose that $0_\mathfrak{i}$ corresponds to 1 and $0_\mathfrak{j}$ corresponds to $k+1$. The condition $\sigma_a(\mathfrak{i})=\mathfrak{j}$ and $(\mathbf{a},\mathbf{r}_a,\mathbf{u}_a,\mathbf{R}_a,\mathbf{U}_a)\to\sigma_f$ means that as functions on $[k]$ and $[k+1,k+\ell]$, $\mathbf{r}_a(1)=\mathbf{U}_a(k+1)$. There are no other constraints on these values.

Then for all variables in (3.19) fixed apart from $\mathbf{r}_a$ and $\mathbf{U}_a$, and all values of $\mathbf{r}_a,\mathbf{U}_a$ fixed other than $\mathbf{r}_a(1)$ and $\mathbf{U}_a(k+1)$ the ensuing sum over $\mathbf{r}_a,\mathbf{U}_a$ is

$$\sum_{\mathbf{r}_a(1)=\mathbf{U}_a(k+1)}\beta_{p_2\mathbf{r}_a}^{\mathbf{U}_a}.$$

But recalling the contraction operators from (2.2), this sum is the coordinate of $e_{\mathbf{r}_a(2)}\otimes\cdots\cdots e_{\mathbf{r}_a(k)}\otimes\check{e}_{\mathbf{U}_a(k+2)}\otimes\cdots\otimes\check{e}_{\mathbf{U}_a(k+\ell)}$ in $c_{1,1}(v_{p_2})$. But $c_{1,1}(v_{p_2})=0$ because $v_{p_2}\in\dot{\mathcal{T}}_n^{k,\ell}$. □

We henceforth write $\sum^*_{\sigma_f,\tau_f}$ *to mean the sum is restricted to* σ_f,τ_f *satisfying the* ***forbidden matchings*** *property.* Lemma 3.4 now implies

$$\begin{aligned}\mathcal{J}_n(w,\mu,\nu)=&\sum^*_{\sigma_f,\tau_f\in S_{k+\ell+p_f}}\left(\prod_{f\in\{a,b,c,d\}}\mathrm{Wg}_{n,k+\ell+p_f}(\sigma_f\tau_f^{-1})\right)\\&\sum_{p_i}\sum_{\substack{\mathbf{a}\in\mathcal{A}(w),\mathbf{r}_f,\mathbf{R}_f,\mathbf{V}_f,\mathbf{v}_f,\mathbf{U}_f,\mathbf{u}_f,\mathbf{s}_f,\mathbf{S}_f\\(\mathbf{a},\mathbf{r}_f,\mathbf{u}_f,\mathbf{R}_f,\mathbf{U}_f)\to\sigma_f\\(\mathbf{a},\mathbf{s}_f,\mathbf{v}_f,\mathbf{S}_f,\mathbf{V}_f)\to\tau_f}}\\&\beta^{\mathbf{V}_a}_{p_2\mathbf{s}_a}\bar\beta^{\mathbf{U}_a}_{p_1\mathbf{r}_a}\beta^{\mathbf{V}_b}_{p_3\mathbf{s}_b}\bar\beta^{\mathbf{U}_b}_{p_2\mathbf{r}_b}\beta^{\mathbf{u}_a}_{p_4\mathbf{R}_a}\bar\beta^{\mathbf{v}_a}_{p_3\mathbf{S}_a}\beta^{\mathbf{u}_b}_{p_5\mathbf{R}_b}\bar\beta^{\mathbf{v}_b}_{p_4\mathbf{S}_b}\\&\times\beta^{\mathbf{V}_c}_{p_6\mathbf{s}_c}\bar\beta^{\mathbf{U}_c}_{p_5\mathbf{r}_c}\beta^{\mathbf{V}_d}_{p_7\mathbf{s}_d}\bar\beta^{\mathbf{U}_d}_{p_6\mathbf{r}_d}\beta^{\mathbf{u}_c}_{p_8\mathbf{R}_c}\bar\beta^{\mathbf{v}_c}_{p_7\mathbf{S}_c}\beta^{\mathbf{u}_d}_{p_1\mathbf{R}_d}\bar\beta^{\mathbf{v}_d}_{p_8\mathbf{S}_d}.\end{aligned}\tag{3.20}$$

Moreover, we can significantly tidy up (3.20). For everything in (3.20) fixed except for e.g. p_2, the ensuing sum over p_2 is

$$\sum_{p_2}\beta^{\mathbf{V}_a}_{p_2\mathbf{s}_a}\bar\beta^{\mathbf{U}_b}_{p_2\mathbf{r}_b}=\sum_{p_2}\langle e^{\mathbf{U}_b}_{\mathbf{r}_b},v_{p_2}\rangle\langle v_{p_2},e^{\mathbf{V}_a}_{\mathbf{s}_a}\rangle=\langle\mathfrak{q}_\theta e^{\mathbf{U}_b}_{\mathbf{r}_b},e^{\mathbf{V}_a}_{\mathbf{s}_a}\rangle.$$

Therefore executing the sums over p_i in (3.20) we replace the sum over p_i and the product over β-terms by

$$\begin{aligned}&\langle\mathfrak{q}_\theta e^{\mathbf{U}_b}_{\mathbf{r}_b},e^{\mathbf{V}_a}_{\mathbf{s}_a}\rangle\langle\mathfrak{q}_\theta e^{\mathbf{v}_a}_{\mathbf{S}_a},e^{\mathbf{V}_b}_{\mathbf{s}_b}\rangle\langle\mathfrak{q}_\theta e^{\mathbf{v}_b}_{\mathbf{S}_b},e^{\mathbf{u}_a}_{\mathbf{R}_a}\rangle\langle\mathfrak{q}_\theta e^{\mathbf{U}_c}_{\mathbf{r}_c},e^{\mathbf{u}_b}_{\mathbf{R}_b}\rangle\times\\&\langle\mathfrak{q}_\theta e^{\mathbf{U}_d}_{\mathbf{r}_d},e^{\mathbf{V}_c}_{\mathbf{s}_c}\rangle\langle\mathfrak{q}_\theta e^{\mathbf{v}_c}_{\mathbf{S}_c},e^{\mathbf{V}_d}_{\mathbf{s}_d}\rangle\langle\mathfrak{q}_\theta e^{\mathbf{v}_d}_{\mathbf{S}_d},e^{\mathbf{u}_c}_{\mathbf{R}_c}\rangle\langle\mathfrak{q}_\theta e^{\mathbf{U}_a}_{\mathbf{r}_a},e^{\mathbf{u}_d}_{\mathbf{R}_d}\rangle.\end{aligned}\tag{3.21}$$

By Proposition 3.2 we have e.g.

$$\langle\mathfrak{q}_\theta e^{\mathbf{U}_b}_{\mathbf{r}_b},e^{\mathbf{V}_a}_{\mathbf{s}_a}\rangle=D_{\mu,\nu}(n)\sum_{\pi\in S_{k+\ell}}z_\theta(\pi)\langle\varphi^{-1}[\rho_n^{k,\ell}(\pi)]e^{\mathbf{U}_b}_{\mathbf{r}_b},e^{\mathbf{V}_a}_{\mathbf{s}_a}\rangle$$

Now recall from (3.12) that

$$\varphi^{-1}[\rho_n^{k+\ell}(\pi)]=\sum_{I=(i_1,\ldots,i_k),J=(j_{k+1},\ldots,j_{k+\ell})}e^J_{I'(I\sqcup J;\pi)}\otimes\check{e}_I^{J'(I\sqcup J;\pi)}.$$

This means that $\langle \varphi^{-1}[\rho_n^{k+\ell}(\pi)]e_{\mathbf{r}_b}^{\mathbf{U}_b}, e_{\mathbf{s}_a}^{\mathbf{V}_a}\rangle$ is either equal to 0 or 1 and $\langle \varphi^{-1}[\rho_n^{k+\ell}(\pi)]e_{\mathbf{r}_b}^{\mathbf{U}_b}, e_{\mathbf{s}_a}^{\mathbf{V}_a}\rangle = 1$ if and only if, letting (3.18) induce identifications

$$\{1_{\mathfrak{i}} : \mathbf{i} \in \mathfrak{I}^+_{a,R}\} \cong [k], \{1_{\mathfrak{i}} : \mathbf{i} \in \mathfrak{I}^-_{b,R^{-1}}\} \cong [k+1, k+\ell],$$
$$\{0_{\mathfrak{i}} : \mathbf{i} \in \mathfrak{I}^+_{b,R}\} \cong [k], \{0_{\mathfrak{i}} : \mathbf{i} \in \mathfrak{I}^-_{a,R^{-1}}\} \cong [k+1, k+\ell],$$

via their given indexing of intervals, we have $[\mathbf{s}_a \sqcup \mathbf{U}_b] \circ \pi = [\mathbf{r}_b \sqcup \mathbf{V}_a]$, where e.g. $\mathbf{s}_a \sqcup \mathbf{U}_b$ is the function either on endpoints of intervals or on $[k+\ell]$ induced by the union of $\mathbf{s}_a$ and $\mathbf{U}_b$. Hence, repeating this argument,

$$\begin{aligned}(3.21) = D_{\mu,\nu}(n)^8 \sum_{\pi_1,\ldots,\pi_8 \in S_{k+\ell}} \left(\prod_{i=1}^{8} z_\theta(\pi_i)\right) \\ \mathbf{1}\{[\mathbf{s}_a \sqcup \mathbf{U}_b] \circ \pi_1 = [\mathbf{r}_b \sqcup \mathbf{V}_a], [\mathbf{s}_b \sqcup \mathbf{v}_a] \circ \pi_2 = [\mathbf{S}_a \sqcup \mathbf{V}_b], \\ [\mathbf{R}_a \sqcup \mathbf{v}_b] \circ \pi_3 = [\mathbf{S}_b \sqcup \mathbf{u}_a], [\mathbf{R}_b \sqcup \mathbf{U}_c] \circ \pi_4 = [\mathbf{r}_c \sqcup \mathbf{u}_b], \\ [\mathbf{s}_c \sqcup \mathbf{U}_d] \circ \pi_5 = [\mathbf{r}_d \sqcup \mathbf{V}_c], [\mathbf{s}_d \sqcup \mathbf{v}_c] \circ \pi_6 = [\mathbf{S}_c \sqcup \mathbf{V}_d], \\ [\mathbf{R}_c \sqcup \mathbf{v}_d] \circ \pi_7 = [\mathbf{S}_d \sqcup \mathbf{u}_c], [\mathbf{R}_d \sqcup \mathbf{U}_a] \circ \pi_8 = [\mathbf{r}_a \sqcup \mathbf{u}_d]\}.\end{aligned}$$

Putting all these arguments together gives

$$\begin{aligned}&\mathcal{J}_n(w,\mu,\nu) \\ =&D_{\mu,\nu}(n)^8 \sum^{*}_{\sigma_f,\tau_f \in S_{p_f+k+\ell}} \sum_{\pi_1,\ldots,\pi_8 \in S_{k+\ell}} \left(\prod_{f\in\{a,b,c,d\}} \mathrm{Wg}_{n,k+\ell+p_f}(\sigma_f \tau_f^{-1})\right)\left(\prod_{i=1}^{8} z_\theta(\pi_i)\right) \\ &\sum_{p_i} \sum_{\substack{\mathbf{a}\in\mathcal{A}(w),\mathbf{r}_f,\mathbf{R}_f,\mathbf{V}_f,\mathbf{v}_f,\mathbf{U}_f,\mathbf{u}_f,\mathbf{s}_f,\mathbf{S}_f \\ (\mathbf{a},\mathbf{r}_f,\mathbf{u}_f,\mathbf{R}_f,\mathbf{U}_f)\to\sigma_f \\ (\mathbf{a},\mathbf{s}_f,\mathbf{v}_f,\mathbf{S}_f,\mathbf{V}_f)\to\tau_f}} \\ &\mathbf{1}\{[\mathbf{s}_a \sqcup \mathbf{U}_b] \circ \pi_1 = [\mathbf{r}_b \sqcup \mathbf{V}_a], [\mathbf{s}_b \sqcup \mathbf{v}_a] \circ \pi_2 = [\mathbf{S}_a \sqcup \mathbf{V}_b], \\ &[\mathbf{R}_a \sqcup \mathbf{v}_b] \circ \pi_3 = [\mathbf{S}_b \sqcup \mathbf{u}_a], [\mathbf{R}_b \sqcup \mathbf{U}_c] \circ \pi_4 = [\mathbf{r}_c \sqcup \mathbf{u}_b], \\ &[\mathbf{s}_c \sqcup \mathbf{U}_d] \circ \pi_5 = [\mathbf{r}_d \sqcup \mathbf{V}_c], [\mathbf{s}_d \sqcup \mathbf{v}_c] \circ \pi_6 = [\mathbf{S}_c \sqcup \mathbf{V}_d], \\ &[\mathbf{R}_c \sqcup \mathbf{v}_d] \circ \pi_7 = [\mathbf{S}_d \sqcup \mathbf{u}_c], [\mathbf{R}_d \sqcup \mathbf{U}_a] \circ \pi_8 = [\mathbf{r}_a \sqcup \mathbf{u}_d]\}.\end{aligned}$$

This formula says that we can calculate $\mathcal{J}_n(w,\mu,\nu)$ by summing over some combinatorial data of matchings (the σ_f, τ_f, π_i) a quantity that we can understand well times a count of the number of indices that satisfy the prescribed matchings. To formalize this point of view we make the following definition.

Definition 3.5. A *matching datum* of the triple (w,k,ℓ) is a pair $(\sigma_f,\tau_f) \in S_{k+\ell+p_f} \times S_{k+\ell+p_f}$ as above, satisfying the **forbidden matchings** property for each $f \in \{a,b,c,d\}$, together with $(\pi_1,\ldots,\pi_8) \in (S_{k+\ell})^8$. We write

$$\mathsf{MATCH}(w,k,\ell)$$

for the finite collection of all matching data for (w,k,ℓ).

Given a matching datum $\{\sigma_f, \tau_f, \pi_i\}$, we write $\mathcal{N}(\{\sigma_f, \tau_f, \pi_i\})$ for the number of choices of $\mathbf{a} \in \mathcal{A}(w), \mathbf{r}_f, \mathbf{R}_f, \mathbf{V}_f, \mathbf{v}_f, \mathbf{U}_f, \mathbf{u}_f, \mathbf{s}_f, \mathbf{S}_f$ such that

$$\begin{aligned}&(\mathbf{a}, \mathbf{r}_f, \mathbf{u}_f, \mathbf{R}_f, \mathbf{U}_f) \to \sigma_f, (\mathbf{a}, \mathbf{s}_f, \mathbf{v}_f, \mathbf{S}_f, \mathbf{V}_f) \to \tau_f,\\ &[\mathbf{s}_a \sqcup \mathbf{U}_b] \circ \pi_1 = [\mathbf{r}_b \sqcup \mathbf{V}_a], [\mathbf{s}_b \sqcup \mathbf{v}_a] \circ \pi_2 = [\mathbf{S}_a \sqcup \mathbf{V}_b],\\ &[\mathbf{R}_a \sqcup \mathbf{v}_b] \circ \pi_3 = [\mathbf{S}_b \sqcup \mathbf{u}_a], [\mathbf{R}_b \sqcup \mathbf{U}_c] \circ \pi_4 = [\mathbf{r}_c \sqcup \mathbf{u}_b],\\ &[\mathbf{s}_c \sqcup \mathbf{U}_d] \circ \pi_5 = [\mathbf{r}_d \sqcup \mathbf{V}_c], [\mathbf{s}_d \sqcup \mathbf{v}_c] \circ \pi_6 = [\mathbf{S}_c \sqcup \mathbf{V}_d],\\ &[\mathbf{R}_c \sqcup \mathbf{v}_d] \circ \pi_7 = [\mathbf{S}_d \sqcup \mathbf{u}_c], [\mathbf{R}_d \sqcup \mathbf{U}_a] \circ \pi_8 = [\mathbf{r}_a \sqcup \mathbf{u}_d]. \end{aligned} \tag{3.22}$$

With this notation, we have proved the following theorem.

Theorem 3.6. *For $k + \ell > 0$, $\mu \vdash k$ and $\nu \vdash \ell$, $w \in [\mathbf{F}_4, \mathbf{F}_4]$, we have*

$$\begin{aligned}\mathcal{J}_n(w, \mu, \nu) = &D_{\mu,\nu}(n)^8 \sum_{\{\sigma_f,\tau_f,\pi_i\} \in \mathsf{MATCH}(w,k,\ell)} \left(\prod_{i=1}^{8} z_\theta(\pi_i) \right) \\ &\left(\prod_{f \in \{a,b,c,d\}} \mathrm{Wg}_{n,k+\ell+p_f}(\sigma_f \tau_f^{-1}) \right) \mathcal{N}(\{\sigma_f, \tau_f, \pi_i\}). \end{aligned} \tag{3.23}$$

We conclude this section by bounding the terms $z_\theta(\pi_i)$ and $\mathrm{Wg}_{n,k+\ell+p_f}(\sigma_f \tau_f^{-1})$ using Proposition 2.7 and Lemma 3.3, recalling also (2.1). Note that $\sum_{f \in \{a,b,c,d\}} p_f = \frac{|w|}{2}$. This yields

Corollary 3.7. *For $k + \ell > 0$, $\mu \vdash k$ and $\nu \vdash \ell$, $w \in [\mathbf{F}_4, \mathbf{F}_4]$, we have*

$$\begin{aligned}\mathcal{J}_n(w, \mu, \nu) \ll_{k,\ell,w} &n^{-4k-4\ell-\frac{|w|}{2}} \sum_{\{\sigma_f,\tau_f,\pi_i\} \in \mathsf{MATCH}(w,k,\ell)} \\ &n^{-\sum_f |\sigma_f \tau_f^{-1}| - \sum_{i=1}^{8} \|\pi_i\|_{k,\ell}} \mathcal{N}(\{\sigma_f, \tau_f, \pi_i\}). \end{aligned} \tag{3.24}$$

We will proceed in the next section to understand all the quantities in (3.24) in topological terms by constructing a surface from each $\{\sigma_f, \tau_f, \pi_i\}$.

4. Topology

4.1. **Construction of surfaces from matching data.** We now show how a datum in $\mathsf{MATCH}(w, k, \ell)$ can be used to construct a surface such that the terms appearing in (3.23) can be bounded by topological features of the surface. This construction is similar to the constructions of [MP19, MP15], but with the presence of additional π_i adding a new aspect. We continue to assume $g = 2$ for simplicity. We can still assume that $\gamma \in [\Gamma_2, \Gamma_2]$ and hence $w \in [\mathbf{F}_4, \mathbf{F}_4]$.

Construction of the 1-skeleton. *π-intervals.* The identifications of the previous section mean that we view

$$\begin{aligned}
\pi_1 &: \{\, 0_{\mathfrak{i}} \,:\, \mathfrak{i} \in \mathfrak{I}^{+}_{b,R} \cup \mathfrak{I}^{-}_{a,R^{-1}} \} \to \{\, 1_{\mathfrak{i}'} \,:\, \mathfrak{i}' \in \mathfrak{I}^{+}_{a,R} \cup \mathfrak{I}^{-}_{b,R^{-1}} \,\},\\
\pi_2 &: \{\, 0_{\mathfrak{i}} \,:\, \mathfrak{i} \in \mathfrak{I}^{-}_{a,R} \cup \mathfrak{I}^{-}_{b,R^{-1}} \} \to \{\, 1_{\mathfrak{i}'} \,:\, \mathfrak{i}' \in \mathfrak{I}^{+}_{b,R} \cup \mathfrak{I}^{+}_{a,R^{-1}} \,\},\\
\pi_3 &: \{\, 0_{\mathfrak{i}} \,:\, \mathfrak{i} \in \mathfrak{I}^{-}_{b,R} \cup \mathfrak{I}^{+}_{a,R^{-1}} \} \to \{\, 1_{\mathfrak{i}'} \,:\, \mathfrak{i}' \in \mathfrak{I}^{-}_{a,R} \cup \mathfrak{I}^{+}_{b,R^{-1}} \,\},\\
\pi_4 &: \{\, 0_{\mathfrak{i}} \,:\, \mathfrak{i} \in \mathfrak{I}^{+}_{c,R} \cup \mathfrak{I}^{+}_{b,R^{-1}} \} \to \{\, 1_{\mathfrak{i}'} \,:\, \mathfrak{i}' \in \mathfrak{I}^{-}_{b,R} \cup \mathfrak{I}^{-}_{c,R^{-1}} \,\},\\
\pi_5 &: \{\, 0_{\mathfrak{i}} \,:\, \mathfrak{i} \in \mathfrak{I}^{+}_{d,R} \cup \mathfrak{I}^{-}_{c,R^{-1}} \} \to \{\, 1_{\mathfrak{i}'} \,:\, \mathfrak{i}' \in \mathfrak{I}^{+}_{c,R} \cup \mathfrak{I}^{-}_{d,R^{-1}} \,\},\\
\pi_6 &: \{\, 0_{\mathfrak{i}} \,:\, \mathfrak{i} \in \mathfrak{I}^{-}_{c,R} \cup \mathfrak{I}^{-}_{d,R^{-1}} \} \to \{\, 1_{\mathfrak{i}'} \,:\, \mathfrak{i}' \in \mathfrak{I}^{+}_{d,R} \cup \mathfrak{I}^{+}_{c,R^{-1}} \,\},\\
\pi_7 &: \{\, 0_{\mathfrak{i}} \,:\, \mathfrak{i} \in \mathfrak{I}^{-}_{d,R} \cup \mathfrak{I}^{+}_{c,R^{-1}} \} \to \{\, 1_{\mathfrak{i}'} \,:\, \mathfrak{i}' \in \mathfrak{I}^{-}_{c,R} \cup \mathfrak{I}^{+}_{d,R^{-1}} \,\},\\
\pi_8 &: \{\, 0_{\mathfrak{i}} \,:\, \mathfrak{i} \in \mathfrak{I}^{+}_{a,R} \cup \mathfrak{I}^{+}_{d,R^{-1}} \} \to \{\, 1_{\mathfrak{i}'} \,:\, \mathfrak{i}' \in \mathfrak{I}^{-}_{d,R} \cup \mathfrak{I}^{-}_{a,R^{-1}} \,\}.
\end{aligned} \tag{4.1}$$

We add an arc between any two interval endpoints that are mapped to one another by some π_i. All the intervals added here are called *π-intervals.* The purpose of this construction is that the conditions concerning π_i in (3.22) correspond to the fact that *two end-points of intervals connected by a π-interval are assigned the same value in $[n]$ by the relevant functions out of* $\mathbf{r}_f, \mathbf{R}_f, \mathbf{V}_f, \mathbf{v}_f, \mathbf{U}_f, \mathbf{u}_f, \mathbf{s}_f, \mathbf{S}_f$ *(at most one of these functions has any given interval endpoint in its domain).*

The π-intervals together with the R-intervals and R^{-1} intervals form a collection of loops that we call $R^{\pm}$-π-loops.

σ-arcs and τ-arcs. Recall from the previous sections that we view

$$\sigma_f, \tau_f : \mathfrak{I}^{+}_{f} \to \mathfrak{I}^{-}_{f}.$$

We add an arc between each $0_{\mathfrak{i}}$ and $1_{\mathfrak{i}'}$ with $\sigma_f(\mathfrak{i}) = \mathfrak{i}'$ and between each $1_{\mathfrak{i}}$ and $0_{\mathfrak{i}'}$ with $\tau_f(\mathfrak{i}) = \mathfrak{i}'$. These arcs are called σ_f-arcs and τ_f-arcs respectively. Any σ_f-arc (resp. τ_f-arc) is also called a σ-arc (resp. τ-arc). Notice even though an arc is formally the same as an interval, we distinguish these types of objects. The only arcs that exist are σ-arcs and τ-arcs. The purpose of this construction is that the conditions pertaining to σ_f, τ_f in (3.22) are equivalent to the fact that *two end-points of intervals connected by a σ-arc or τ-arc are assigned the same value in $[n]$ by the relevant functions out of* $\mathbf{a}, \mathbf{r}_f, \mathbf{R}_f, \mathbf{V}_f, \mathbf{v}_f, \mathbf{U}_f, \mathbf{u}_f, \mathbf{s}_f, \mathbf{S}_f$.

After adding these arcs, every endpoint of an interval has exactly one arc emanating from it. We have therefore now constructed a trivalent graph

$$G(\{\sigma_f, \tau_f, \pi_i\}).$$

Each vertex of the graph is an endpoint of two intervals and one arc. The number of vertices of this graph is twice the total number of w-intervals, R-intervals, and R^{-1}-intervals which is $2(|w| + 8(k+\ell))$. Therefore we have

$$\chi(G(\{\sigma_f, \tau_f, \pi_i\})) = -(|w| + 8(k+\ell)). \tag{4.2}$$

(For general g, we have $\chi(G(\{\sigma_f, \tau_f, \pi_i\})) = -(|w| + 4g(k+\ell))$.) Moreover, *the conditions in (3.22) are now interpreted purely in terms of the combinatorics of this graph.*

Gluing in discs. There are two types of cycles in $G(\{\sigma_f, \tau_f, \pi_i\})$ that we wish to consider:

- Cycles that alternate between following either a w-intermediate interval or a π-interval and then either a σ-arc or a τ-arc. These cycles are disjoint from one another, and every σ or τ-arc is contained in exactly one such cycle. We call these cycles *type-I cycles.* For every type-I cycle, we glue a disc to $G(\{\sigma_f, \tau_f, \pi_i\})$ along its boundary, following the cycle. These discs will be called *type-I discs.* (These are analogous to the o-discs of [MP19].)
- Cycles that alternate between following either a w-interval, an R-interval, or an R^{-1}-interval and then either a σ-arc or a τ-arc. Again, these cycles are disjoint, and every σ or τ-arc is contained in exactly one such cycle. We call these cycles *type-II cycles.* For every type-II cycle, we glue a disc to $G(\{\sigma_f, \tau_f, \pi_i\})$ identifying the boundary of the disc with the cycle. These discs will be called *type-II discs.* (These are similar to the z-discs of [MP19].)

Because every interior of an interval meets exactly one of the glued-in discs, and every arc has two boundary segments of discs glued to it, the object resulting from gluing in these discs is a decorated topological surface that we denote by

$$\Sigma(\{\sigma_f, \tau_f, \pi_i\}).$$

An example of this construction is depicted in Figure 4.1.

The boundary components of $\Sigma(\{\sigma_f, \tau_f, \pi_i\})$ consist of the w-loop and the $R^{\pm}$-π-loops. It is not hard to check that $\Sigma(\{\sigma_f, \tau_f, \pi_i\})$ is orientable with an orientation compatible with the fixed orientations of the boundary loops corresponding to traversing every w-interval or $R^{\pm 1}$-interval from 0 to 1.

We view the given CW-complex structure, and the assigned labelings and directions of the intervals that now pave $\partial\Sigma$ as part of the data of $\Sigma(\{\sigma_f, \tau_f, \pi_i\})$. The number of discs of $\Sigma(\{\sigma_f, \tau_f, \pi_i\})$ is connected to the quantities appearing in Proposition 3.6 as follows.

Lemma 4.1. $\mathcal{N}(\{\sigma_f, \tau_f, \pi_i\}) = n^{\#\{\textit{type-I discs of } \Sigma(\{\sigma_f, \tau_f, \pi_i\})\}}$.

Proof. The constraints on the functions $\mathbf{a}, \mathbf{r}_f, \mathbf{R}_f, \mathbf{V}_f, \mathbf{v}_f, \mathbf{U}_f, \mathbf{u}_f, \mathbf{s}_f, \mathbf{S}_f$ in (3.22) now correspond to the fact that altogether, they assign the same value in $[n]$ to every interval end-point in the same type-I-cycle, and there are no other constraints between them. □

The quantities $|\sigma_f \tau_f^{-1}|$ in (3.24) can also be related to $\Sigma(\{\sigma_f, \tau_f, \pi_i\})$ as follows.

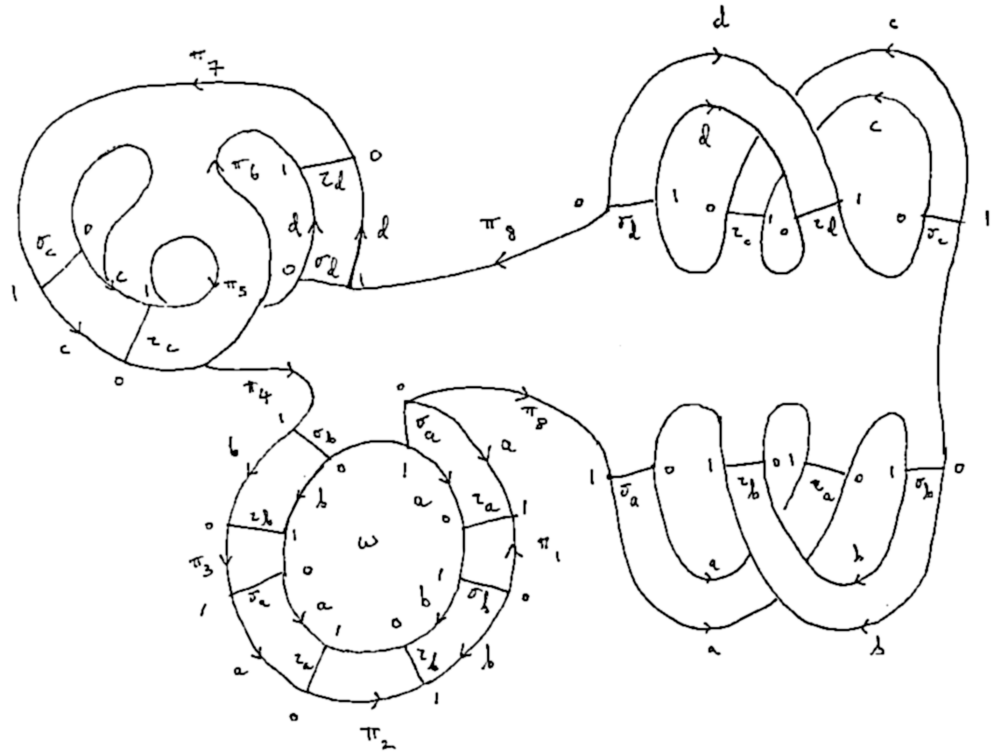

FIGURE 4.1. This is a depiction of an example $\Sigma(\{\sigma_f, \tau_f, \pi_i\})$ for $w = ab^{-1}a^{-1}b$. The σ, τ, and some of the π_i-arcs are labeled along with the numbers (0 or 1) of the points being matched in the w-intervals. Each w-interval is also labeled with its corresponding letter. Here $k = \ell = 1$; π_8 is a transposition and all other π_i are the identity. There is one resulting $R^{\pm}$-π-loop. In this example, for each $f \in \{a, b, c, d\}$, $\sigma_f = \tau_f$. This means that all type II discs are rectangles.

Lemma 4.2. *We have*

$$\prod_{f \in \{a,b,c,d\}} n^{-|\sigma_f \tau_f^{-1}|} = n^{-4(k+\ell) - \frac{|w|}{2}} n^{\#\{\textit{type-II discs of } \Sigma(\{\sigma_f, \tau_f, \pi_i\})\}}.$$

Proof. Recalling the definition of $|\sigma_f \tau_f^{-1}|$ from Proposition 2.7, we can also write

$$|\sigma_f \tau_f^{-1}| = k + \ell + p_f - \#\{\text{cycles of } \sigma_f \tau_f^{-1}\}.$$

The cycles of $\{\sigma_f\tau_f^{-1} : f \in \{a,b,c,d\}\}$ are in 1:1 correspondence with the type-II cycles of $\Sigma(\{\sigma_f,\tau_f,\pi_i\})$ and hence also the type-II discs. Therefore

$$\prod_{f\in\{a,b,c,d\}} n^{-|\sigma_f\tau_f^{-1}|} = n^{-4(k+\ell)} n^{\sum_{f\in\{a,b,c,d\}}(-p_f+\#\{\text{cycles of } \sigma_f\tau_f^{-1}\})}$$
$$= n^{-4(k+\ell)-\frac{|w|}{2}} n^{\#\{\text{type-II discs of } \Sigma(\{\sigma_f,\tau_f,\pi_i\})\}}.$$

□

We are now able to prove the following.

Theorem 4.3. *For $k+\ell>0$, $\mu \vdash k$ and $\nu \vdash \ell$, $w \in [\mathbf{F}_4,\mathbf{F}_4]$, we have*

$$\mathcal{J}_n(w,\mu,\nu) \ll_{w,k,\ell} \sum_{\{\sigma_f,\tau_f,\pi_i\}\in \mathsf{MATCH}(w,k,\ell)} n^{-\sum_{i=1}^{8}\|\pi_i\|_{k,\ell}} n^{\chi(\Sigma(\{\sigma_f,\tau_f,\pi_i\}))}.$$

Proof. Combining Lemmas 4.1 and 4.2 with Corollary 3.7 gives

$$\mathcal{J}_n(w,\mu,\nu) \ll_{w,k,\ell} n^{-8k-8\ell-|w|} \times \sum_{\{\sigma_f,\tau_f,\pi_i\}\in \mathsf{MATCH}(w,k,\ell)} n^{-\sum_{i=1}^{8}\|\pi_i\|_{k,\ell}} n^{\#\{\text{discs of } \Sigma(\{\sigma_f,\tau_f,\pi_i\})\}}.$$

Then from (4.2) we obtain

$$\mathcal{J}_n(w,\mu,\nu) \ll_{w,k,\ell} \sum_{\{\sigma_f,\tau_f,\pi_i\}\in \mathsf{MATCH}(w,k,\ell)} n^{-\sum_{i=1}^{8}\|\pi_i\|_{k,\ell}} n^{\chi(G(\{\sigma_f,\tau_f,\pi_i\}))+\#\{\text{discs of } \Sigma(\{\sigma_f,\tau_f,\pi_i\})\}}$$
$$= \sum_{\{\sigma_f,\tau_f,\pi_i\}\in \mathsf{MATCH}(w,k,\ell)} n^{-\sum_{i=1}^{8}\|\pi_i\|_{k,\ell}} n^{\chi(\Sigma(\{\sigma_f,\tau_f,\pi_i\}))}.$$

□

4.2. **Two simplifying surgeries.** Theorem 4.3 suggests that we now bound

$$\chi(\Sigma(\{\sigma_f,\tau_f,\pi_i\})) - \sum_{i=1}^{8}\|\pi_i\|_{k,\ell}$$

for all $\{\sigma_f,\tau_f,\pi_i\} \in \mathsf{MATCH}(w,k,\ell)$. To do this, we make some observations that simplify the task. If C is a simple closed curve in a surface S, then *compressing S along C* means that we cut S along C and then glue discs to cap off any new boundary components created by the cut.

Suppose that we are given $\{\sigma_f,\tau_f,\pi_i\} \in \mathsf{MATCH}(w,k,\ell)$. Then $\{\sigma_f,\sigma_f,\pi_i\}$ is also in $\mathsf{MATCH}(w,k,\ell)$ (the **forbidden matching** property continues to hold). It is not hard to see that

$$\chi(\Sigma(\{\sigma_f,\sigma_f,\pi_i\})) \geq \chi(\Sigma(\{\sigma_f,\tau_f,\pi_i\})).$$

Indeed, the τ_f arcs can be replaced by σ_f-parallel arcs inside the type-II discs of $\Sigma(\{\sigma_f,\tau_f,\pi_i\})$. The resulting surface's arcs may not cut the surface into discs, but this can be fixed by (possibly repeatedly) compressing the surface along simple closed curves disjoint from the arcs, leaving the combinatorial data of the arcs unchanged but only potentially increasing the Euler characteristic.

It remains to deal with the sum $\sum_{i=1}^{8} \|\pi_i\|_{k,\ell}$.

Suppose again that an arbitrary $\{\sigma_f, \tau_f, \pi_i\} \in \mathsf{MATCH}(w, k, \ell)$ is given. For each $i \in [8]$ write

$$\pi_i = \pi_i^* \sigma_i$$

where $\pi_i^* \in S_k \times S_\ell$, $\sigma_i = (\pi_i^*)^{-1}\pi_i \in S_{k+\ell}$, and $|\sigma_i| = \|\pi_i\|_{k,\ell}$. Let $X_0 \stackrel{\mathrm{def}}{=} \Sigma(\{\sigma_f, \tau_f, \pi_i\})$.

Take $\Sigma(\{\sigma_f, \tau_f, \pi_i\})$ and add to it all the π_i^*-intervals that would have been added if π_i was replaced by π_i^* for each $i \in [8]$ in its construction. The resulting object X_1 is the decorated surface X_0 together with a collection of π_i^*-intervals with endpoints in the boundary of X_0, and interiors disjoint from X_0. This adds $8(k+\ell)$ edges to X_0 and hence

$$\chi(X_1) = \chi(\Sigma(\{\sigma_f, \tau_f, \pi_i\})) - 8(k+\ell).$$

Now we consider all cycles that for any fixed $i \in [8]$, alternate between π_i-intervals and π_i^*-intervals. The number of these cycles is the total number of cycles of the permutations $\{\, (\pi_i^*)^{-1}\pi_i \,:\, i \in [8] \,\}$. On the other hand, the number of cycles of $(\pi_i^*)^{-1}\pi_i$ is

$$k + \ell - |(\pi_i^*)^{-1}\pi_i| = k + \ell - |\sigma_i| = k + \ell - \|\pi_i\|_{k,\ell}$$

So in total there are $8(k+\ell) - \sum_i \|\pi_i\|_{k,\ell}$ of these cycles. For every such cycle, we glue a disc along its boundary to the cycle. The resulting object is denoted X_2. Now, X_2 is a topological surface, and we added $8(k+\ell) - \sum_i \|\pi_i\|_{k,\ell}$ discs to X_1 to form X_2, so

$$\chi(X_2) = \chi(X_1) + 8(k+\ell) - \sum_i \|\pi_i\|_{k,\ell} = \chi(\Sigma(\{\sigma_f, \tau_f, \pi_i\})) - \sum_i \|\pi_i\|_{k,\ell}.$$

Now 'forget' all the original π_i-intervals from X_2 to form X_3. The surface X_3 is a decorated surface in the same sense as X_0, except the connected components of $X_3 - \{\text{arcs}\}$ may not be discs. Similarly to before, by sequentially compressing X_3 along non-nullhomotopic simple closed curves disjoint from arcs, if they exist, we obtain a new decorated surface X_4. See Figure 4.2 for an illustration of this surgery taking place. Moreover, and this is the main point, X_4 is the same as $\Sigma(\{\sigma_f, \tau_f, \pi_i^*\})$ in the sense that they are related by a decoration-respecting cellular homeomorphism. Compression can only increase the Euler characteristic, so we obtain

$$\chi(\Sigma(\{\sigma_f, \tau_f, \pi_i^*\})) \geq \chi(X_3) = \chi(X_2) = \chi(\Sigma(\{\sigma_f, \tau_f, \pi_i\}) - \sum_i \|\pi_i\|_{k,\ell}.$$

Combining these two arguments proves the following proposition.

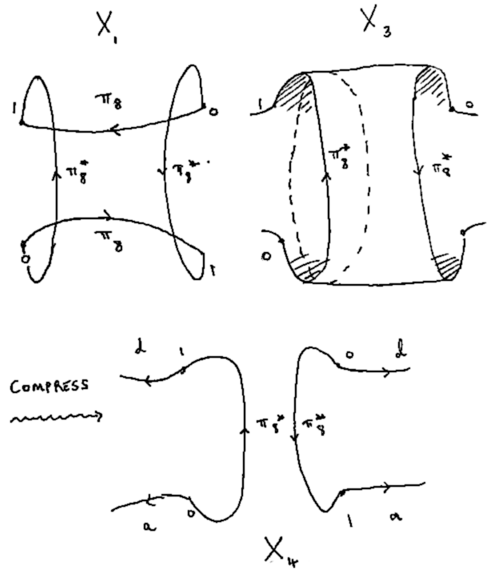

FIGURE 4.2. This is a local illustration of the second type of simplifying surgery, precisely in the context of Figure 4.1. The dashed simple closed curve in X_3 is disjoint from any arcs, and cutting along this curve and gluing in two discs yields X_4. Going back to Figure 4.1 again, the net effect of this surgery is to cut the left half from the right half.

Proposition 4.4. *For any given* $\{\sigma_f, \tau_f, \pi_i\}$*, there exist* $\pi_i^* \in S_k \times S_\ell$ *for* $i \in [8]$ *such that*

$$\chi(\Sigma(\{\sigma_f, \sigma_f, \pi_i^*\})) - \sum_{i=1}^{8} \|\pi_i^*\|_{k,\ell} = \chi(\Sigma(\{\sigma_f, \sigma_f, \pi_i^*\}))$$
$$\geq \chi(\Sigma(\{\sigma_f, \tau_f, \pi_i\})) - \sum_{i=1}^{8} \|\pi_i\|_{k,\ell}.$$

This has the following immediate corollary when combined with Theorem 4.3. Let

$$\mathsf{MATCH}^*(w, k, \ell)$$

denote the subset of $\mathsf{MATCH}(w, k, \ell)$ consisting of $\{\sigma_f, \sigma_f, \pi_i\}$ (i.e. $\sigma_f = \tau_f$ for each $f \in \{a, b, c, d\}$) with $\pi_i \in S_k \times S_\ell$ for each $i \in [8]$.

Corollary 4.5. *For* $k + \ell > 0$*,* $\mu \vdash k$ *and* $\nu \vdash \ell$*,* $w \in [\mathbf{F}_4, \mathbf{F}_4]$*, we have*

$$\mathcal{J}_n(w, \mu, \nu) \ll_{w,k,\ell} n^{\max_{\{\sigma_f, \sigma_f, \pi_i\} \in \mathsf{MATCH}^*(w,k,\ell)} \chi(\Sigma(\{\sigma_f, \sigma_f, \pi_i\}))}.$$

The benefit to having $\pi_i \in S_k \times S_\ell$ for $i \in [8]$ is the following. Suppose now that $\{\sigma_f, \sigma_f, \pi_i\} \in \mathsf{MATCH}^*(w, k, \ell)$. Recall that the boundary loops of $\Sigma(\{\sigma_f, \sigma_f, \pi_i\})$ consist of one w-loop and some number of $R^{\pm}$-π-loops. The condition that each $\pi_i \in S_k \times S_\ell$ means that no π-interval ever connects an endpoint of a R-interval with an endpoint of an R^{-1}-interval. So every boundary component of $\Sigma(\{\sigma_f, \sigma_f, \pi_i\})$ that is not the w-loop contains either only R-intervals or only R^{-1}-intervals, and in fact, when following the boundary component and reading the directions and labels of the intervals according to traversing each from 0 to 1, reads out a positive power of R (in the former case of only R-intervals) or a negative power of R^{-1} (in the latter case of only R^{-1}-intervals). The sum of the positive powers of R in boundary loops is k, and the sum of the negative powers of R is $-\ell$. Knowing this boundary structure is extremely important for the arguments in the next sections.

4.3. **A topological result that proves Theorem 3.1.** Here, in the spirit of Culler [Cul81], we explain another way to think about the surfaces $\Sigma(\{\sigma_f, \sigma_f, \pi_i\})$ for $\{\sigma_f, \sigma_f, \pi_i\} \in \mathsf{MATCH}^*(w, k, \ell)$ that is easier to work with than the construction we gave. At this point we also show how things work for general $g \geq 2$. An *arc* in a surface Σ is a properly embedded interval in Σ with endpoints in the boundary $\partial\Sigma$.

Definition 4.6. For $w \in \mathbf{F}_{2g}$, we define $\mathsf{surfaces}(w, k, \ell)$ to be the set of all decorated surfaces Σ^* as follows. A decorated surface $\Sigma^* \in \mathsf{surfaces}(w, k, \ell)$ is an oriented surface with boundary, with compatibly oriented boundary components, together with a collection of disjoint embedded arcs that cut Σ^* into topological discs. One boundary component is assigned to be a w-loop, and every other boundary component is assigned to be either a R-loop or an R^{-1}-loop. Each arc is assigned a transverse direction and a label in $\{a_1, b_1, \ldots, a_g, b_g\}$. Every arc-endpoint in $\partial\Sigma^*$ inherits a transverse direction and label from the assigned direction and label of its arc. We require that Σ^* satisfy the following properties.

P1: When one follows the w-loop according to its assigned orientation, and reads f when an f-labeled arc-endpoint is traversed in its given direction, and f^{-1} when an f-labeled arc-endpoint is traversed counter to its given direction, one reads a cyclic rotation of w in reduced form, depending on where one begins to read.

P2: When one follows any R-loop according to its assigned orientation in the same way as before, one reads (a cyclic rotation) of some positive power of R_g in reduced form. The sum of these positive powers over all R-loops is k.

P3: When one follows any R^{-1}-loop according to its assigned orientation in the same way as before, one reads (a cyclic rotation) of some negative power of R_g in reduced form. The sum of these negative powers over all R^{-1}-loops is $-\ell$.

P4: No arc connects an R-loop to an R^{-1}-loop.

Given a surface $\Sigma(\{\sigma_f, \sigma_f, \pi_i\})$ with $\{\sigma_f, \sigma_f, \pi_i\} \in \mathsf{MATCH}^*(w, k, \ell)$, all the type-II discs of the surface are rectangles. Hence, by collapsing each w-interval, R-interval, and R^{-1}-interval to a point, and collapsing every type-II rectangle to an arc, we obtain a CW-complex that is a surface with boundary, cut into discs by arcs. Every arc inherits a transverse direction and label from the compatible assigned directions and labels of the intervals in the boundary of its originating type-II rectangle. We call this modified surface $\Sigma^* = \Sigma^*(\{\sigma_f, \pi_i\})$. It clearly satisfies **P1-P3** and **P4** follows from the **forbidden matchings** property. (Of course, when $g = 2$, we identify $\{a, b, c, d\}$ with $\{a_1, b_1, a_2, b_2\}$.) We also have $\chi(\Sigma(\{\sigma_f, \sigma_f, \pi_i\})) = \chi(\Sigma^*(\{\sigma_f, \pi_i\}))$. With Definition 4.6 and the remarks proceeding it, we can now state a further consequence of Corollary 4.5 as it extends to general $g \geq 2$.

Corollary 4.7. *For $k + \ell > 0$, $\mu \vdash k$, $\nu \vdash \ell$, $w \in [\mathbf{F}_{2g}, \mathbf{F}_{2g}]$, as $n \to \infty$*

$$\mathcal{J}_n(w, \mu, \nu) \ll_{w,k,\ell} n^{\max\{\,\chi(\Sigma^*)\,:\,\Sigma^* \in \mathsf{surfaces}(w,k,\ell)\,\}}.$$

In order for Corollary 4.7 to give us strong enough results it needs to be combined with the following non-trivial topological bound.

Proposition 4.8. *If $w \in [\mathbf{F}_{2g}, \mathbf{F}_{2g}]$ is a shortest element representing the conjugacy class of $\gamma \in \Gamma_g$, $w \neq \mathrm{id}$, and $\Sigma^* \in \mathsf{surfaces}(w, k, \ell)$ then $\chi(\Sigma^*) \leq -(k + \ell)$.*

Remark 4.9. Proposition 4.8 is by no means a trivial statement and one has to use that w is a shortest element representing the conjugacy class of some element of Γ_g. For example, if $w = R_g$, then w represents the conjugacy class of id_{Γ_g}, but for $k = 0$ and $\ell = 1$ there is an 'obvious' annulus in $\mathsf{surfaces}(w, 0, 1)$. This has $\chi = 0 > -(k + \ell) = -1$. Proposition 4.8 also requires $w \neq \mathrm{id}$; if $w = \mathrm{id}$ then for $k = 0$ and $\ell = 1$ one can take a disc with no arcs as a valid element of $\mathsf{surfaces}(\mathrm{id}, 0, 0)$. This has $\chi = 1 > -(k+\ell) = 0$. In fact this disc is ultimately responsible for $\mathbb{E}_{g,n}[\mathrm{Tr}_{\mathrm{id}}] = n$.

The proof of Proposition 4.8 is self-contained and given in §§4.5. Before doing this, we prove Theorem 3.1.

Proof of Theorem 3.1 given Proposition 4.8. Since Theorem 3.1 was proved when $k = \ell = 0$ in §§3.2, we can assume $k+\ell > 0$. Then combining Corollary 4.7 and Proposition 4.8 gives

$$\mathcal{J}_n(w, \mu, \nu) \ll_{w,k,\ell} n^{-(k+\ell)}.$$

On the other hand, $D_{\mu,\nu}(n) = O\left(n^{k+\ell}\right)$ from (2.1). Therefore $D_{\mu,\nu}(n)\mathcal{J}_n(w, \mu, \nu) \ll_{w,k,\ell} 1$. □

4.4. **Work of Dehn and Birman-Series.** As we mentioned in §§3.1, to prove Proposition 4.8 we have to use the fact that $w \in [\mathbf{F}_{2g}, \mathbf{F}_{2g}]$ is a shortest element representing the conjugacy class of $\gamma \in \Gamma_g$. We use a combinatorial characterization of such words that stems from Dehn's algorithm [Deh12] for solving the problem of whether a given word represents the identity in

Γ_g. The ideas of Dehn's algorithm were refined by Birman and Series in [BS87]. In [MP20], the author and Puder used Birman and Series' results (alongside other methods) to obtain the analog of Theorem 1.2 when the family of groups $\mathsf{SU}(n)$ is replaced by the family of symmetric groups S_n. Similar consequences of the work of Dehn, Birman, and Series that we used in *(ibid.)* will be used here.

We now follow the language of [MP20] to state the results we need in this paper. These results are simple and direct consequences of the work of Birman and Series.

We view the universal cover of Σ_g as a disc tiled by $4g$-gons that we call U. We assume every edge of this tiling is directed and labeled by some element of $\{a_1, b_1, \ldots, a_g, b_g\}$ such that when we read counter-clockwise along the boundary of any octagon we read the reduced cyclic word $[a_1, b_1] \cdots [a_g, b_g]$. By fixing a basepoint $u \in U$ we obtain a free cellular action of Γ_g on U that respects the labels and directions of edges and identifies the quotient $\Gamma_g \backslash U$ with Σ_g; this gives a description of Σ_g as a $4g$-gon with glued sides as is typical.

Now suppose that $\gamma \in \Gamma$ is not the identity. The quotient $A_\gamma \stackrel{\text{def}}{=} \langle \gamma \rangle \backslash U$ of U by the cyclic group generated by γ is an open annulus tiled by infinitely many $4g$-gons. The edges of A_γ inherit directions and labels from those of the edges of U. The point $u \in U$ maps to some point denoted by $x_0 \in A_\gamma$.

Now let $w \in \mathbf{F}_{2g}$ be an element that represents γ, and identify w with a combinatorial word by writing w in reduced form. Beginning at x_0, and following the path spelled out by w beginning at x_0, we obtain an oriented closed loop L_w in the one-skeleton of A_γ. If w is a shortest element representing the conjugacy class of γ, then this loop L_w must not have self-intersections. In this case, that we from now assume, L_w is therefore a topologically embedded circle in the annulus A_γ that is non-nullhomotopic and cuts A_γ into two annuli $A_\gamma^\pm$.

Every vertex of A_γ has $4g$ incident half-edges each of which has an orientation and direction given by the edge they are in. Going clockwise, the cyclic order of the half-edges incident at any vertex is

'a_1-outgoing, b_1-incoming, a_1-incoming, b_1-outgoing, ... , a_g-outgoing, b_g-incoming, a_g-incoming, b_g-outgoing'.

We define $\hat{L}_w$ to be the loop L_w with all incident half edges in A_γ attached. We call the new half-edges added *hanging half-edges.*

Moreover, we thicken up $\hat{L}_w$ by viewing each edge of L_w as a rectangle, each hanging half-edge as a half-rectangle, and each vertex replaced by a disc. In other words, we take a small neighborhood of $\hat{L}_w$ in A_γ. We now think of $\hat{L}_w$ as the thickened version. This is a topological annulus, where the hanging half-edges have become stubs hanging off. A *piece of* $\hat{L}_w$ is a contiguous collection of hanging half-rectangles and rectangle sides following edges of L_w in the boundary of $\hat{L}_w$. Such a piece is in either A_γ^+ or A_γ^-. Given a piece P of $\hat{L}_w$ we write $\mathfrak{e}(P)$ for the number of rectangle sides

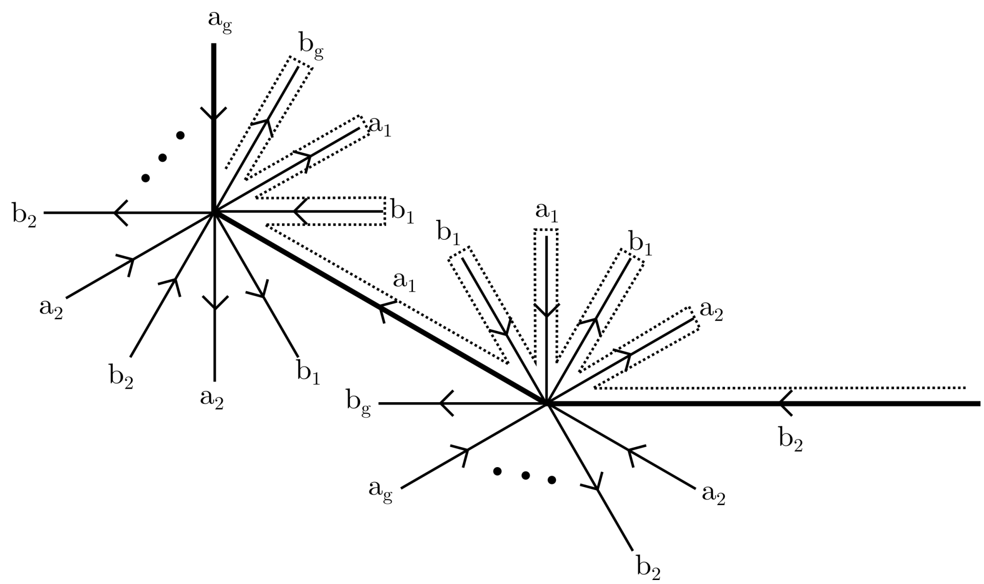

FIGURE 4.3. Illustration of a piece P of $\hat{L}_w$ in the case when the reduced form of w contains $a_g a_1^{-1} b_2^{-1}$ as a subword. The edges of L_w are in bold. The piece is indicated by the dotted lines. This piece P has $\mathfrak{e}(P) = 2$, $\mathfrak{he}(P) = 7$, and $\chi(P) = 1$. Note that a piece may also run along the other side of L_w.

following edges of L_w, and $\mathfrak{he}(P)$ for the number of hanging-half edges in P. We say that a piece P has Euler characteristic $\chi(P) = 0$ if it follows an entire boundary component of $\hat{L}_w$, and $\chi(P) = 1$ otherwise as we view it as an interval running along the rectangle sides and around the sides of the hanging half-rectangles. See Figure 4.3 for an illustration of a piece of $\hat{L}_w$.

Birman and Series prove in [BS87, Thm. 2.12(a)] that if w is a shortest element representing the conjugacy class of $\gamma \in \Gamma_g$ then there are strong restrictions on the pieces of $\hat{L}_w$ that can appear. This has the following consequence which is given by[3] [MP20, Proof of Lem. 5.18].

Lemma 4.10. *If w is a shortest element representing the conjugacy class of $\gamma \in \Gamma_g$, and both γ and hence w are non-identity, then for any piece P of $\hat{L}_w$, we have*

$$\mathfrak{e}(P) \le (2g-1)\mathfrak{he}(P) + 2g\chi(P).$$

Proof. Since w is a shortest element representing some non-identity conjugacy class in Γ_g, in the language of [MP20], L_w is a boundary reduced tiled surface. Then the proof of [MP20, Lem. 5.18] contains the result stated in the lemma. The basic idea of the proof is not complicated and goes back to

[3] We stress that Lemma 4.10 is a straightforward consequence of Birman and Series' work, so even though we cite [MP20], this paper does not depend on [MP20] in any significant way.

Dehn [Deh12]: if there are too many edges (i.e. $\mathfrak{e}(P)$ is large) then one can find a string of letters in the reduced word of w (e.g. $aba^{-1}b^{-1}c$) that can be shortened using the relator R (e.g. $aba^{-1}b^{-1}c = dcd^{-1}$). ☐

This inequality plays a crucial role in the next section.

4.5. **Proof of Proposition 4.8.** Suppose that $g \geq 2$ and $w \in [\mathbf{F}_{2g}, \mathbf{F}_{2g}]$ is a non-identity shortest element representing the conjugacy class of $\gamma \in \Gamma_g$. In particular, w is cyclically reduced. We let $R = R_g$. Now fix $k, \ell \in \mathbf{N}_0$ and suppose $\Sigma^* \in \mathsf{surfaces}(w, k, \ell)$. The arcs of Σ^* are of three different types:

WR: An arc with one endpoint in the w-loop and one endpoint in an R or R^{-1}-loop.
RR: An arc with both endpoints in R or R^{-1} loops. By property **P4,** the endpoints of such an arc are both in R-loops or both in R^{-1}-loops.
WW: An arc with both endpoints in the w-loop.

The boundary of any disc of Σ^* alternates between segments of $\partial\Sigma^*$ and arcs. A disc is a *pre-piece disc* if its boundary contains exactly one segment of the w-loop. A disc is called a *junction disc* if it is not a pre-piece disc. We say that a junction disc is *piece-adjacent* if it meets a WR-arc-side.

To be precise, we view all discs as open discs, and hence not containing any arcs. A disc meets certain arc-sides along its boundary; it is possible for a disc to meet both sides of the same arc and we view this scenario as the disc meeting two separate arc-sides. We say an arc-side has the same type WR/RR/WW as its corresponding arc.

Note that any pre-piece disc cannot meet any WW-arc-side: if it did, the disc could only meet this one arc-side together with one segment of the w-loop and this would contradict the fact that w is cyclically reduced since the arc matches a letter f with a cyclically adjacent letter f^{-1} of w. It is also clear that any pre-piece disc meets exactly 2 WR-arc-sides: the ones that emanate from the sole segment of the w-loop. So in light of **P4** a pre-piece disc takes one of the forms shown in Figure 4.4.

We define a *piece of* Σ^* to be a connected component of

$$\{\text{pre-piece discs}\} \cup \{\text{WR-arcs}\}.$$

A piece of Σ^* is therefore either a contiguous collection of pre-piece discs that meet only along WR-arcs, or a single WR-arc. If P is a piece of Σ^*, either $\chi(P) = 1$, or $\chi(P) = 0$, in which case P meets the entire w-loop and is the unique piece.

We now have **two** definitions of pieces; pieces of $\hat{L}_w$ and pieces of Σ^*. These are, as the names suggest, closely related, and this is the key observation in the proof of Proposition 4.8. Indeed, the reader should carefully consider Figure 4.5 that leads to the following lemma. In analogy to pieces of $\hat{L}_w$, if P is any piece of Σ^*, we write $\mathfrak{e}(P)$ for the number of WR-arcs in P, and $\mathfrak{he}(P)$ for the number of RR-arc sides that meet P (this is zero if P is a single WR-arc).

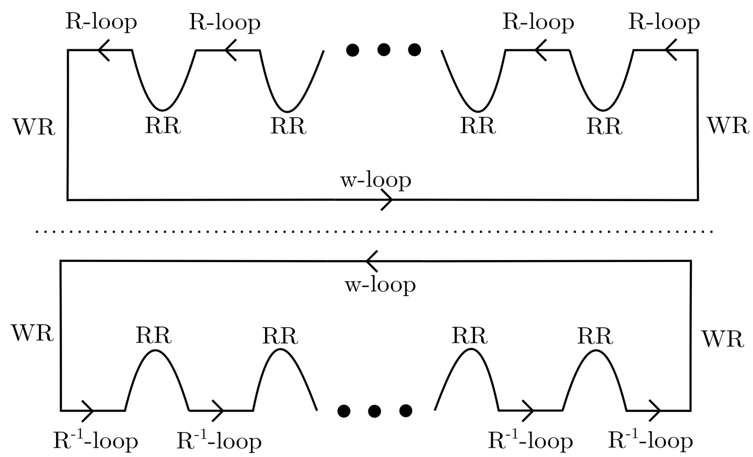

FIGURE 4.4. Possible forms of pre-piece discs. The number of R-loop segments or R^{-1}-loop segments is at least 1 and bounded given k and ℓ. The arrows denote the orientations of the boundary loops.

Lemma 4.11. *If w is a shortest element representing the conjugacy class of $\gamma \in \Gamma_g$, $k, \ell \in \mathbf{N}_0$, and $\Sigma^* \in \mathsf{surfaces}(w,k,\ell)$ then for any piece P of Σ^*, we have*

$$\mathfrak{e}(P) \leq (2g-1)\mathfrak{he}(P) + 2g\chi(P).$$

Proof. Given any piece P of Σ^*, it contains a consecutive (possibly cyclic) series of WR-arcs that correspond to a contiguous collection of edges in the loop L_w. The discs of P correspond to certain vertices of L_w; each of these vertices has two emanating half-edges belonging to the edges defined by WR-arcs of P. The piece P can either meet only R-loops or meet only R^{-1}-loops.

We define a piece P' of $\hat{L}_w$ corresponding to P as follows. If P meets R-loops, then P' consists of rectangle sides along the edges of L_w corresponding to the WR-arcs of P together with all hanging half-edges at vertices corresponding to discs of P that are on the left of L_w as it is traversed in its assigned orientation (corresponding to reading w along L_w). If P' meets R^{-1}-loops, then P' is defined similarly with the modification that we include instead hanging half-edges on the right of L_w. Figure 4.5 together with its captioned discussion now shows that

$$\mathfrak{he}(P') \leq \mathfrak{he}(P),$$

and $\mathfrak{e}(P) = \mathfrak{e}(P')$ by construction. We also have $\chi(P') = \chi(P)$. Therefore Lemma 4.10 applied to P' implies

$$\mathfrak{e}(P) = \mathfrak{e}(P') \leq (2g-1)\mathfrak{he}(P') + 2g\chi(P') \leq (2g-1)\mathfrak{he}(P) + 2g\chi(P).$$

□

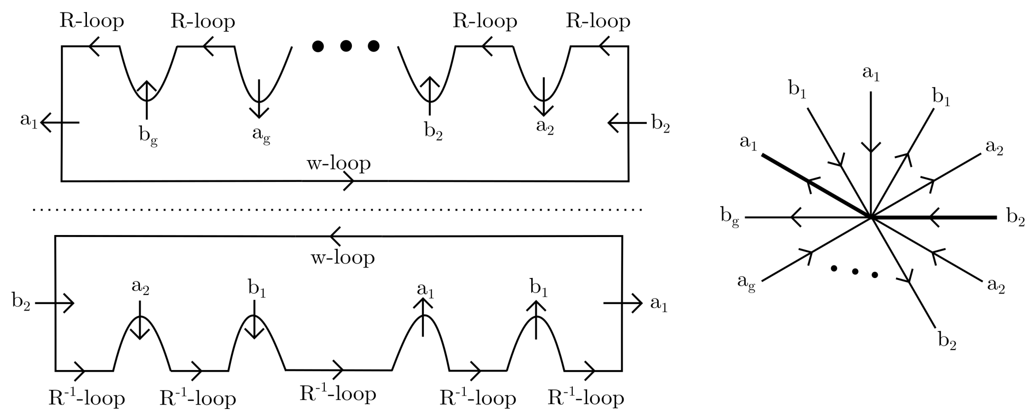

FIGURE 4.5. Given a segment of the w-loop corresponding to a juncture between letters $a_1^{-1}b_2^{-1}$ in w, if this segment is part of a pre-piece disc then some possible forms of that disc are shown above. This juncture between letters of w corresponds to a vertex in L_w. The right hand illustration shows the neighborhood of this vertex in the annulus A_γ, where the bold lines correspond to half-edges of L_w. The right hand picture actually almost determines the left hand pictures. Indeed, given the a_1 arc on the top-left, the next arc has to be a b_g arc with the given direction, since only b_g^{-1} cyclically precedes a_1 in R_g or any power of R_g. Then the next arc a_g with its direction is determined since only a_g cyclically precedes b_g in R_g. This continues until an arc labeled by b_2 and with an incoming direction is reached, as in the right arc of the top-left picture. At this point, the boundary of the disc may close up. (This is analogous to what happens in the bottom picture, where an analogous pattern occurs.) The only indeterminacy is that after reaching a b_2 arc with an incoming direction for the first time, the entire pattern shown in the right hand picture may repeat any number of times, as long as k and ℓ allow it. The upshot of this is that any pre-piece disc has at least as many incident RR-arc-sides as there are hanging half-edges on the corresponding side of L_w, at the corresponding vertex.

Let N_{RR} be the number of RR-arcs, N_{WR} the number of WR-arcs, and N_{WW} the number of WW-arcs in Σ^*. In the following we refer to discs of Σ^* simply as discs. Since there are $4g(k+\ell)$ incidences between arcs and R-loops or R^{-1} loops we have

$$2N_{RR} + N_{WR} = 4g(k+\ell). \tag{4.3}$$

Let Σ_1 be the surface formed by cutting Σ^* along all RR-arcs. We have

$$\chi(\Sigma_1) = \sum_{\text{discs } D} \left(1 - \frac{d'(D)}{2}\right)$$

where $d'(D)$ is the number of arc-sides meeting D that are <u>not</u> of type RR. This formula holds because $d'(D)$ is the degree of the disc D in the dual graph G_1 of Σ_1, the right hand side is easily seen to be $\chi(G_1) = V(G_1) - E(G_1)$, and since Σ_1 deformation retracts to an obvious embedded copy of G_1, $\chi(G_1) = \chi(\Sigma_1)$. We partition the sum above according to

$$\begin{aligned}
\chi(\Sigma_1) =& S_0 + S_1 + S_2,\\
S_0 \stackrel{\text{def}}{=}& \sum_{\text{pre-piece discs } D} \left(1 - \frac{d'(D)}{2}\right),\\
S_1 \stackrel{\text{def}}{=}& \sum_{\text{piece-adjacent junction discs } D} \left(1 - \frac{d'(D)}{2}\right),\\
S_2 \stackrel{\text{def}}{=}& \sum_{\text{not piece-adjacent junction discs } D} \left(1 - \frac{d'(D)}{2}\right).
\end{aligned}$$

Note first that a pre-piece disc has $d'(D) = 2$ (cf. Fig 4.4). Hence $S_0 = 0$. We deal with S_1 next. For a disc D of Σ^*, let $d_{WR}(D)$ denote the number of WR-arc-sides meeting D. Note that a piece-adjacent junction disc D has $d_{WR}(D) > 0$ by definition. We rewrite S_1 as

$$\begin{aligned}
S_1 =& \sum_{\text{piece-adjacent junction discs } D} \left(1 - \frac{d'(D)}{2}\right) \frac{1}{d_{WR}(D)} \sum_{\text{incidences between } D \text{ and WR-arc-sides}} 1\\
=& \sum_{\text{pieces } P} \sum_{\text{incidences between } P \text{ and some junction disc } D \text{ along WR-arc}} Q(D) \qquad (4.4)
\end{aligned}$$

where for a piece-adjacent junction disc D

$$Q(D) \stackrel{\text{def}}{=} \frac{1}{d_{WR}(D)}\left(1 - \frac{d'(D)}{2}\right).$$

Suppose that D is a piece-adjacent junction disc. By parity considerations, $d_{WR}(D)$ is even. We estimate $Q(D)$ by splitting into two cases. If $d_{WR}(D) = 2$ then $d'(D) \geq 3$ since otherwise, D would meet only 2 WR arc-sides and other RR arc-sides, hence be a pre-piece disc and not be a junction disc. In this case

$$Q(D) = \frac{1}{2}\left(1 - \frac{d'(D)}{2}\right) \leq \frac{1}{2}\left(1 - \frac{3}{2}\right) = -\frac{1}{4}.$$

Otherwise, $d_{WR}(D) \geq 4$ and since $d'(D) \geq d_{WR}(D)$, we have

$$Q(D) \leq \frac{1}{d_{WR}(D)}\left(1 - \frac{d_{WR}(D)}{2}\right) = \frac{1}{d_{WR}(D)} - \frac{1}{2} \leq \frac{1}{4} - \frac{1}{2} = -\frac{1}{4}.$$

So we have proved that for all piece-adjacent junction discs D, $Q(D) \leq -\frac{1}{4}$. Putting this into (4.4) gives

$$\begin{aligned} S_1 &\leq -\frac{1}{4} \sum_{\text{pieces } P} \sum_{\text{incidences between } P \text{ and some junction disc } D \text{ along WR-arc}} 1 \\ &= -\frac{1}{4} \sum_{\text{pieces } P} 2\chi(P) = -\frac{1}{2} \sum_{\text{pieces } P} \chi(P). \end{aligned} \tag{4.5}$$

We now turn to S_2. *Here is the key moment where* $w \neq \mathrm{id}$ *is used*[4]. Since $w \neq \mathrm{id}$, any disc must meet an arc. Indeed, the only other possibility is that the boundary of the disc is an entire boundary loop that has no emanating arcs. This hypothetical boundary loop cannot be an R or R^{-1}-loop, so it has to be the w-loop. But this would entail $w = \mathrm{id}$.

Hence any disc contributing to S_2 meets no WR-arc-side, but meets some arc-side. Therefore it meets only WW-arcs or only RR-arcs. Every disc D contributing to S_2 meeting only WW-arcs gives a non-positive contribution since w is cyclically reduced hence $d'(D) \geq 2$. Every disc D contributing to S_2 meeting only RR-arcs, which we will call an *RR-disc*, has $d'(D) = 0$ and hence contributes 1 to S_2.

This shows

$$S_2 \leq \#\{\text{RR-discs}\}. \tag{4.6}$$

In total combining $S_0 = 0$ with (4.5) and (4.6) we get

$$\chi(\Sigma_1) \leq \#\{\text{RR-discs}\} - \frac{1}{2} \sum_{\text{pieces } P \text{ of } \Sigma^*} \chi(P).$$

To obtain Σ^* from Σ_1 we have to glue all cut RR-arcs, of which there are N_{RR}. Each gluing decreases χ by 1 so

$$\chi(\Sigma^*) \leq \#\{\text{RR-discs}\} - N_{RR} - \frac{1}{2} \sum_{\text{pieces } P \text{ of } \Sigma^*} \chi(P).$$

[4] Although technically, $w \neq \mathrm{id}$ was used to define L_w and pieces etc, if w is the identity the proof of Proposition 4.8 could, a priori, circumvent these definitions.

Using Lemma 4.11 with the above gives

$$\begin{aligned}\chi(\Sigma^*) \leq & \#\{\text{RR-discs}\} - N_{RR} - \frac{1}{2}\sum_{\text{pieces } P \text{ of } \Sigma^*} \chi(P) \\ \leq & \#\{\text{RR-discs}\} - N_{RR} - \frac{1}{4g}\sum_{\text{pieces } P \text{ of } \Sigma^*} \mathfrak{e}(P) \\ & + \frac{(2g-1)}{4g}\sum_{\text{pieces } P \text{ of } \Sigma^*} \mathfrak{he}(P) \qquad (4.7) \\ = & \#\{\text{RR-discs}\} - N_{RR} - \frac{N_{WR}}{4g} + \frac{(2g-1)}{4g}\sum_{\text{pieces } P \text{ of } \Sigma^*} \mathfrak{he}(P). \qquad (4.8)\end{aligned}$$

Let $\mathfrak{he}'(\Sigma^*)$ denote the total number of RR-arc-sides meeting RR-discs. Every RR-disc has to meet at least $4g$ arc-sides; this observation is similar to the reasoning in Figure 4.5. Therefore

$$\mathfrak{he}'(\Sigma^*) \geq 4g\#\{\text{RR-discs}\}. \qquad (4.9)$$

Every RR-arc-side either meets a piece P and contributes to $\mathfrak{he}(P)$ or a disc meeting only RR-arc-sides and contributes to $\mathfrak{he}'(\Sigma^*)$. Hence

$$\mathfrak{he}'(\Sigma^*) + \sum_{\text{pieces } P \text{ of } \Sigma^*} \mathfrak{he}(P) = 2N_{RR}. \qquad (4.10)$$

Combining (4.3), (4.9), and (4.10) with (4.8) gives

$$\begin{aligned}\chi(\Sigma^*) & \overset{(4.9)}{\leq} \frac{\mathfrak{he}'(\Sigma^*)}{4g} - N_{RR} - \frac{N_{WR}}{4g} + \frac{(2g-1)}{4g}\sum_{\text{pieces } P \text{ of } \Sigma^*} \mathfrak{he}(P) \\ & \overset{(4.10)}{=} \frac{\mathfrak{he}'(\Sigma^*)}{4g} - N_{RR} - \frac{N_{WR}}{4g} + \frac{(2g-1)N_{RR}}{2g} - \frac{(2g-1)}{4g}\mathfrak{he}'(\Sigma^*) \\ & = -\frac{1}{4g}(2N_{RR}+N_{WR}) - \frac{2g-2}{4g}\mathfrak{he}'(\Sigma^*) \\ & \leq -\frac{1}{4g}(2N_{RR}+N_{WR}) \overset{(4.3)}{=} -\frac{4g(k+\ell)}{4g} = -(k+\ell).\end{aligned}$$

This completes the proof of Proposition 4.8. □

5. Proof of main theorem

5.1. **Proof of Theorem 1.2.**

Proof of Theorem 1.2. Assume $\gamma \in [\Gamma_g, \Gamma_g]$ is not the identity and that $w \in [\mathbf{F}_{2g}, \mathbf{F}_{2g}]$ is a shortest element representing the conjugacy class of γ, hence also not the identity. By Corollary 2.10 we have

$$\mathbb{E}_{g,n}[\mathrm{Tr}_\gamma] = \zeta(2g-2;n)^{-1}\sum_{(\mu,\nu)\,\in\tilde{\Omega}} D_{\mu,\nu}(n)\mathcal{J}_n(w,\nu,\mu) + O_{w,g}\left(\frac{1}{n}\right),$$

where $\tilde{\Omega}$ is a finite collection of pairs of Young diagrams. We know $\lim_{n\to\infty}\zeta(2g-2;n) = 1$ from (2.11) and for each fixed (μ,ν), $D_{\mu,\nu}(n)\mathcal{J}_n(w,\nu,\mu) = D_{\nu,\mu}(n)\mathcal{J}_n(w,\nu,\mu) = O_{w,\mu,\nu}(1)$ by Theorem 3.1. Hence $\mathbb{E}_{g,n}[\mathrm{Tr}_\gamma] = O_\gamma(1)$ as $n\to\infty$ as required. $\square$

References

[AB83] M. F. Atiyah and R. Bott. The Yang-Mills Equations over Riemann Surfaces. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 308(1505):523–615, 1983. 3

[BCH+94] G. Benkart, M. Chakrabarti, T. Halverson, R. Leduc, C. Lee, and J. Stroomer. Tensor product representations of general linear groups and their connections with Brauer algebras. *J. Algebra*, 166(3):529–567, 1994. 11

[BS87] J. S. Birman and C. Series. Dehn's algorithm revisited, with applications to simple curves on surfaces. In *Combinatorial group theory and topology (Alta, Utah, 1984)*, volume 111 of *Ann. of Math. Stud.*, pages 451–478. Princeton Univ. Press, 1987. 38, 39

[CM14] B. Collins and C. Male. The strong asymptotic freeness of Haar and deterministic matrices. *Ann. Sci. Éc. Norm. Supér. (4)*, 47(1):147–163, 2014. 6

[Col03] B. Collins. Moments and cumulants of polynomial random variables on unitary groups, the Itzykson-Zuber integral, and free probability. *International Mathematics Research Notices*, 2003(17):953–982, 2003. 12

[CŚ06] B. Collins and P. Śniady. Integration with respect to the Haar measure on unitary, orthogonal and symplectic group. *Comm. Math. Phys.*, 264(3):773–795, 2006. 12, 13, 20

[Cul81] M. Culler. Using surfaces to solve equations in free groups. *Topology*, 20(2):133–145, 1981. 36

[Deh12] M. Dehn. Transformation der Kurven auf zweiseitigen Flächen. *Math. Ann.*, 72(3):413–421, 1912. 37, 39

[DL22] Antoine Dahlqvist and Thibaut Lemoine. Large n limit of yang-mills partition function and wilson loops on compact surfaces, 2022. 7

[EI16] T. Enomoto and M. Izumi. Indecomposable characters of infinite dimensional groups associated with operator algebras. *J. Math. Soc. Japan*, 68(3):1231–1270, 2016. 10

[FH91] W. Fulton and J. Harris. *Representation theory*, volume 129 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1991. A first course, Readings in Mathematics. 8, 9

[GLM12] R. Guralnick, M. Larsen, and C. Manack. Low degree representations of simple Lie groups. *Proc. Amer. Math. Soc.*, 140(5):1823–1834, 2012. 14

[GM83] M. Gromov and V. D. Milman. A topological application of the isoperimetric inequality. *Amer. J. Math.*, 105(4):843–854, 1983. 5

[Gol84] W. M. Goldman. The symplectic nature of fundamental groups of surfaces. *Advances in Mathematics*, 54(2):200 – 225, 1984. 3, 5

[HS19] J. Häsä and A. Stasinski. Representation growth of compact linear groups. *Trans. Amer. Math. Soc.*, 372(2):925–980, 2019. 14

[HT05] U. Haagerup and S. Thorbjørnsen. A new application of random matrices: $\mathrm{Ext}(C^*_{\mathrm{red}}(F_2))$ is not a group. *Ann. of Math. (2)*, 162(2):711–775, 2005. 6

[Koi89] K. Koike. On the decomposition of tensor products of the representations of the classical groups: by means of the universal characters. *Adv. Math.*, 74(1):57–86, 1989. 10, 11

[Lem22] Thibaut Lemoine. Large N behaviour of the two-dimensional Yang-Mills partition function. *Combin. Probab. Comput.*, 31(1):144–165, 2022. 7

[LL08] M. Larsen and A. Lubotzky. Representation growth of linear groups. *J. Eur. Math. Soc. (JEMS)*, 10(2):351–390, 2008. 14

[Mag21] M. Magee. Random Unitary Representations of Surface Groups I: Asymptotic expansions, 2021. Communications in Mathematical Physics, to appear, available at arXiv:2101.00252. 2, 3, 6, 7, 10, 14, 15

[Mir07] M. Mirzakhani. Simple geodesics and Weil-Petersson volumes of moduli spaces of bordered Riemann surfaces. *Invent. Math.*, 167(1):179–222, 2007. 5

[MP15] M. Magee and D. Puder. Word measures on unitary groups. arXiv preprint 1509.07374 v2, 2015. 2, 29

[MP19] M. Magee and D. Puder. Matrix group integrals, surfaces, and mapping class groups I: $\mathcal{U}(n)$. *Invent. Math.*, 218(2):341–411, 2019. 6, 16, 18, 29, 31

[MP20] M. Magee and D. Puder. The asymptotic statistics of random covering surfaces, 2020. Preprint, arXiv:2003.05892v1. 5, 38, 39

[Nic94] A. Nica. On the number of cycles of given length of a free word in several random permutations. *Random Structures & Algorithms*, 5(5):703–730, 1994. 5

[VDN92] D. V. Voiculescu, K. J. Dykema, and A. Nica. *Free random variables*, volume 1 of *CRM Monograph Series*. American Mathematical Society, Providence, RI, 1992. 3

[Voi85] D. Voiculescu. Symmetries of some reduced free product C^*-algebras. In *Operator algebras and their connections with topology and ergodic theory (Buşteni, 1983)*, volume 1132 of *Lecture Notes in Math.*, pages 556–588. Springer, Berlin, 1985. 2

[Voi86] D. Voiculescu. Addition of certain noncommuting random variables. *J. Funct. Anal.*, 66(3):323–346, 1986. 2

[Voi87] D. Voiculescu. Multiplication of certain noncommuting random variables. *J. Operator Theory*, 18(2):223–235, 1987. 2

[Voi90] D. Voiculescu. Noncommutative random variables and spectral problems in free product C^*-algebras. In *Proceedings of the Seventh Great Plains Operator Theory Seminar (Lawrence, KS, 1987)*, volume 20, pages 263–283, 1990. 2

[Voi91] D. Voiculescu. Limit laws for random matrices and free products. *Invent. Math.*, 104(1):201–220, 1991. 2, 4, 5

[Wei78] D. Weingarten. Asymptotic behavior of group integrals in the limit of infinite rank. *Journal of Mathematical Physics*, 19(5):999–1001, 1978. 12

[Wit91] E. Witten. On quantum gauge theories in two dimensions. *Comm. Math. Phys.*, 141(1):153–209, 1991. 14

[Xu97] F. Xu. A random matrix model from two dimensional Yang-Mills theory. *Communications in mathematical physics*, 190(2):287–307, 1997. 12

[Zag94] D. Zagier. Values of zeta functions and their applications. In *First European Congress of Mathematics, Vol. II (Paris, 1992)*, volume 120 of *Progr. Math.*, pages 497–512. Birkhäuser, Basel, 1994. 14

Michael Magee,
Department of Mathematical Sciences,
Durham University, Lower Mountjoy, DH1 3LE Durham, United Kingdom
michael.r.magee@durham.ac.uk