Constructing approximately Haar-random unitaries from GUEs An application of random matrix theory to quantum information theory

Jennifer Hritz

Contents

1	Introduction	1
2	Background	1
	2.1 Quantum information theory	
	2.2 Quantum computing and errors	
	2.3 Randomized benchmarking	
	2.4 Unitary T -designs	2
3	Paper overview	2
	3.1 Approximating the Haar measure	2
	3.2 Construction of the T -design	
4	Paper details	4
	4.1 Importance of small moments of the construction	4
	4.2 Finding the moments of the construction	5
	4.3 Bounding the moments of the construction	
5	Conclusion	8

1 Introduction

Random matrix theory has a wide range of applications from physics to computer science. This project aims give an example of how random matrices are used in the field of quantum information theory. Specifically, we cover the topic of T-designs, an approximation of unitary matrices (unitaries) distributed according to the Haar measure. We will explain T-designs through breaking down the paper Efficient $Unitary\ T$ -designs from E Random E Sums [Che+24], which gives a particular construction of E designs hinging around Gaussian Unitary Ensembles (GUE). We additionally use plots made in Julia to better understand this construction.

2 Background

This projects brings together many concepts from a variety of fields—especially random matrix theory and quantum information theory. Assuming the reader is already familiar with the basics of random matrix theory, this section aims to give some contextual background on the out-of-field concepts.

2.1 Quantum information theory

Quantum information theory is the study of encoding information in quantum states. Analogous to classical information theory, quantum bits (qubits) are individual units of information. Unlike a classic bit, which encodes information as a 0 or 1, a qubit can encode a 0, 1, or some superposition of the two, where the state is either a 0 or 1 with some probability and phase. Taking advantage of this quantum property, qubits are believed to be able to provide significant speedups for many problems that will take an infeasible amount of time to complete with classical computers. Qubits also have quantum gates, unitary operators which can change the state of the qubit. Quantum gates will be at the center of motivating the work of this project.

2.2 Quantum computing and errors

Experimental implementations of qubits has growing significantly in the past two decades. To make a qubit, experimentalists only need to have a quantum system capable of being put in a superposition state and then manipulated and measured. There are several major architectures which accomplish this, such as superconducting qubits, which use non-linear affects in a tiny superconducting circuit to create an artificial atom. Different qubit architectures have advantages and disadvantages, but they all share a common struggle with errors. Current qubits are very error-prone, making it hard to actually run any algorithms on them. Much of the work in the field is going into reducing these errors, such as through engineering improvements in the architecture surrounding qubits as well as by creating robust error-correction codes. There are several metrics by which the quality of a qubit can be characterized. One is coherence time, the time it takes for a qubit to lose the state it has been placed in. Another is randomized benchmarking, as explained below.

2.3 Randomized benchmarking

Randomized benchmarking is a measure of the average error rate of performing a quantum gate on a qubit. In theory, if we set a qubit in a starting state, perform a series of random quantum gates, perform the inverse of these gates, and measure the ending state; it should be identical to the starting state. However, due to errors in the qubit, the ending state will have drifted from the initial state. Randomized benchmarking quantifies the frequency and degree of this drift. A key part of randomized benchmarking is creating the set of random quantum gates. We want to sample for a uniform distribution of gates, so as to not over represent any gates, but computers cannot actually generate Haar-random unitaries. We seek a way to approximate the Haar measure, and the focus of this project is on constructions which do this.

2.4 Unitary T-designs

Unitary T-designs are ensembles of unitaries which reproduce the first T moments of the Haar measure. These have many applications in quantum information, including generating random quantum gates for randomized benchmarking. Most current unitary T-design constructions rely on products of unitaries drawn from a random walk. This draws on the concept of universality, where by combining many independent sources of randomness, you get something approximately uniform. However, these product unitary T-designs are inefficient, mainly because multiplying matrices is inefficient. One way to increase efficiency would be to replace the products with sums of random matrices.

3 Paper overview

In Efficient Unitary T-designs from Random Sums [Che+24], the authors multiply two exponentiated sums of i.i.d. random Hermitian matrices to create a approximate unitary T-design. This is more efficient than prior constructions, which use products of random matrices. This paper provides an interesting look into an application of random matrix theory to quantum information. The purpose of this section is to give a feeling for the content of the paper, before more rigorously digging into the underlying math governing the concepts covered.

3.1 Approximating the Haar measure

Remembering that moments describe the shape of a distribution and that the Haar measure is uniform, it follows that all of its moments will be zero. Explicitly, given an $N \times N$ Haar-random unitary U, it will have trace moments $\mathbb{E}[\overline{\text{Tr}}[U^p]] \approx 0$ for all p. We will claim—and later show—that a T-design is approximately Haar if its first T moments are sufficiently small. As we will see, actually evaluating the moments of our constructed T-design is quite difficult, mainly because it involves careful and intricate canceling of terms in a process known as Weingarten calculus. Instead, we will draw on a concept from complexity theory known as the polynomial method. By some careful definitions, we can relate W to a polynomial such that by bounding the polynomial, we will give a bound of the trace moments.

3.2 Construction of the T-design

We seek to first construct a GUE of Hermitian matrices then convert them into Haar unitaries. Start with a finite collection of i.i.d. Hermitian matrices H_j whose sum matches the first q-moments of the GUE, G,

$$\boldsymbol{H} \sim \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \boldsymbol{H}_{j} \text{ where } \mathbb{E}[\boldsymbol{H}_{j}^{\otimes k}] \approx \mathbb{E}[\boldsymbol{G}^{\otimes k}] \text{ for each } k = 1, \dots, q.$$

According to the central limit theorem, such matrices converge to the GUE. ¹

$$e^{i\theta H}e^{i\theta H'} \approx e^{i\theta G}e^{i\theta G'} =: W_{GUE}$$

Noting that we seek to approximate the trace moments of Haar-random unitaries—which are 0, we will choose a θ which roughly minimizes the moments of \mathbf{W}_{GUE} . The GUE is distributed according to the semicircle law (Figure 1) while the eigenvalues of Haar-random unitaries are evenly distributed on the unit circle.

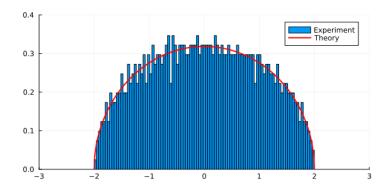
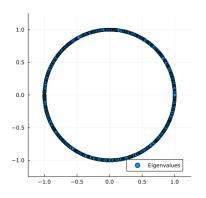
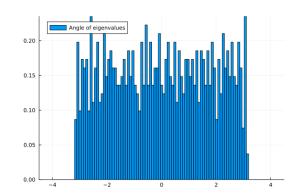


Figure 1: Semicircle law of GUE.

By exponentiating the GUE, we turn our Hermitian matrices into unitaries, whose eigenvalues are distributed on the unit circle (Figure 2a).





- (a) Eigenvalues plotted on real vs imaginary axis.
- (b) Distribution of the angle of eigenvalues around the unit circle.

Figure 2: The eigenvalues of the exponentiated GUE.

However, while the eigenvectors of the exponentiated GUE and Haar-random matrices have the same distribution, the eigenvalues have different distributions. In Figure 2b, we see that the eigenvalues of

 $^{^{1}}$ While presenting on this project, it was brought to my attention that the central limit theorem is not the most efficient way to generate GUEs, and given that the paper is about creating efficient T-designs, it naturally brings into question why the authors chose to generate a GUE this way. I did not have a good answer at the time, but after digging more into the paper and surrounding literature, I found that it is because they were proposing to implement this GUE generation with quantum algorithms. The details of this is quite messy, but my impression is that they use the central limit theorem because there is pre-existing literature that already shows that Hermitian matrices can be efficiently created with quantum algorithms.

the exponentiated GUE are not uniformly distributed around the unit circle. We solve this with the aforementioned product, which we will show approximates the Haar-random unitaries (Figure 3).

$$oldsymbol{W}_{GUE}pprox oldsymbol{U}_{Haar}$$

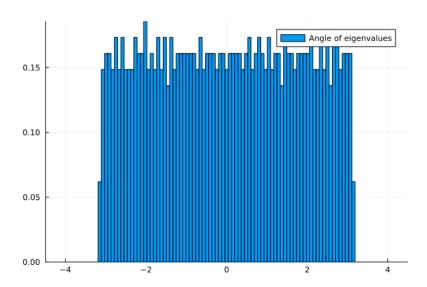


Figure 3: Distribution of the angle of the exponentiated GUE product eigenvalues around the unit circle.

4 Paper details

Given the length and depth of the paper of focus, with much of it requiring background outside of the scope of this project, we will focus on just a few key sections of proving the construction. Our aim is to give an intuitive understanding for why the constructions works as a T-design, through primarily random matrix theory, without getting into the details of the exact bounds and convergence rates.

4.1 Importance of small moments of the construction

Before finding and bounding the moments of our construction, we first seek confirm that small moments lead to approximating the Haar measure. First, we will introduce the concept of an "adaptive T-query quantum decision algorithm," which essentially takes T samples from a distribution and tells us if it is Haar-random or not. If our construction passes this algorithm, it is equivalent to confirming it is a T-design. In the paper, they state [Che+24, Lemma IV.2]

Lemma IV.2. Let **W** be a random unitary that is invariant under unitary conjugation:

$$oldsymbol{W} \sim oldsymbol{U} oldsymbol{D} oldsymbol{U}^\dagger$$

where D is deterministically unitary and U is Haar-random. The first T trace moments are

$$\alpha_p := \frac{1}{N} \overline{Tr}(\boldsymbol{W}^p)$$

for $1 \le p \le T$, and let $\alpha_T := (\alpha_1, \dots, \alpha_T)$. Consider an adaptive T-query quantum decision algorithm A. We state that

$$||\alpha_T|| < \frac{\delta}{32 \cdot T^{7/2}} \quad implies \quad |\mathbb{E}[\boldsymbol{\mathcal{A}}(\boldsymbol{W})] - \mathbb{E}[\boldsymbol{\mathcal{A}}(\boldsymbol{U})]| < \delta.$$

In other words, we state that small trace moments imply that a our algorithm can distinguish our construction W from a Haar-random unitary U with probability at most δ .

To prove this, we start by more formally describing the outcome of our algorithm [Che+24, Lemma VI.1]. Note that $|v\rangle$ is a "ket" in the Bra-Ket notation. A ket v is a vector in a complex vector space representing some quantum state of—in this case—our qubit.

Lemma VI.1. Let \mathcal{A} be a quantum algorithm that makes T queries to a black-box unitary $X \in U(N)$. The final state of the algorithm will be

$$\sum_{k \in K} p_k(\boldsymbol{X}) |k\rangle.$$

where p_k describes a real-valued multilinear polynomial.

Without getting into the details, this lemma allows us to write the outcome as a polynomial. Using off-the-shelf methods of bounding derivatives of polynomials, we can bound this outcome and thereby bound the probability of distinguishing our construction \boldsymbol{W} from a Haar-random unitary \boldsymbol{U} .

4.2 Finding the moments of the construction

Strictly speaking, in this subsection we will not find the moments, but rather optimize the parameters and form of our construction such that the moments can "found" through bounding in the next section. Given two different GUEs G_1 and G_2 , we have our construction

$$W = e^{iG_1\theta} \cdot e^{iG_2\theta}$$
 for precisely chosen θ such that $\overline{\text{Tr}}[e^{iG\theta}] \approx 0$.

As mentioned in the paper overview, to show that it approximates Haar-random unitaries, we need to show that it has small moments. To start, we will use two tricks that will help us bound our construction, starting with the easier infinite-dimension case then discussing how that applies to our finite-dimension construction:

- 1. Breaking the matrix into unitaries and diagonals using ensemble invariance under unitary conjugation, to isolate the eigenvalues.
- 2. Choosing a θ such that the trace moments of the matrix are minimized.

These are, respectively, covered by **Lemma VII.1** and **Lemma VII.2** in the paper [Che+24, Lemma IV.1, 2]. We start with the simpler case of an infinite-dimensional exponentiated GUE, $e^{iG\theta}$.

Lemma VII.1. A GUE G is invariant under conjugation of Haar-random unitaries. Formally,

$$e^{i\boldsymbol{G}\theta} \sim \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{\dagger},$$

where U is a Haar-random unitary and D is an independent random diagonal matrix with the same spectrum as the exponentiated GUE.

For some intuition, consider the diagonalization $e^{iG\theta} = BDB^{\dagger}$ of the exponentiated GUE. By invariance under conjugation, this is the same as $UBDB^{\dagger}U^{\dagger}$, which has the same distribution as UDU^{\dagger} .

The motivation for this is to effectively split up the eigenvalues and eigenvectors of our construction. The eigenvalues of the exponentiated GUE are not the same as the eigenvalues of a Haar unitary, but the eigenvectors are, as shown by it being invariant under conjugation of Haar-random unitaries. This isolates the eigenvalues, such that we only need to check if the eigenvalues of the construction are Haar-random. Since the trace moments describe the distribution of the eigenvalues, that means that by bounding the moments, we will show that the eigenvalues are Haar-random, as desired.

Lemma VII.2. For a GUE G,

$$\lim_{N \to \infty} \mathbb{E}[\overline{Tr}[e^{iG\theta}]] = \frac{J_1(2\theta)}{\theta}$$

where J_1 is a Bessel function of the first kind.

Proof. Giving a sketch of the proof, we remember that, for the GUE G, the odd moments vanish and the even moments follow the Catalan numbers. Specifically, the 2k-th moment gives us

$$\mathbb{E}[\overline{\text{Tr}}[\boldsymbol{G}^{2k}]] = C_k + \mathcal{O}(N^{-2}) \text{ where } C_n = \frac{1}{n+1} \binom{2n}{n}.$$

The $\mathcal{O}(N^{-2})$ comes from our GUE being of finite-dimension N. As N approaches infinity, this error washes out. As $N \to \infty$, the exponentiated GUE gives

$$\lim_{N \to \infty} \mathbb{E}[\overline{\text{Tr}}[e^{i\boldsymbol{G}\theta}]] = \sum_{n=0}^{\infty} \frac{1}{n!} \overline{\text{Tr}}[(i\boldsymbol{G}\theta)^n]$$
$$= \sum_{n=0}^{\infty} \frac{(i\theta)^{2n}}{(2n)!} \text{Cat}_n$$
$$= \frac{J_1(2\theta)}{\theta}$$

In the limit, the zeros of the Bessel function will minimize the moments. Precisely, this tells us that we can choose some θ such that

$$\lim_{n \to \infty} \mathbb{E}\left[\overline{\mathrm{Tr}}[e^{i\boldsymbol{G}\boldsymbol{\theta}}]\right] = 0.$$

With these two lemmas in place, we now use **Lemma VII.4** from the paper to account for the finite-dimension N [Che+24, Lemma VII.4].

Lemma VII.4. For a N-dimensional GUE G, we have

$$\left| \mathbb{E}[\overline{Tr}[e^{i\mathbf{G}\theta}]] - \frac{J_1(2\theta)}{\theta} \right| \le \frac{K_0\theta}{N},$$

where $K_0 > 0$ is some constant.

Proof. To summarize the proof, start by taking a sample E from G. The empirical spectral density (ESD) is

$$p(E) = \frac{1}{N} \sum_{i=1}^{N} \delta(E - \lambda_i(E))$$

where $\lambda_i(E)$ is the eigenvalue of the instance E. Remember that the infinite dimension spectral density of the semicircle is

$$p_{sc}(E) = \frac{1}{2\pi} \sqrt{4 - E^2}$$
 on $[-2, 2]$.

We can use these two definitions to find the difference in spectrums between a finite N- and infinite-dimension GUE. We take the difference of the two as follows

$$\left| \mathbb{E}\left[\overline{\text{Tr}}[e^{i\theta G}] \right] - \lim_{N \to \infty} \left[\mathbb{E}\left[\overline{\text{Tr}}[e^{i\theta G}] \right] \right] \right| \le \left| \mathbb{E}\left[\int_{-\infty}^{\infty} e^{i\theta E} (p(E) - p_{sc}(E)) dE \right] \right|$$

Using a bound which controls the distance between the GUE ESD and semicircle and another bound which controls the distribution of the maximum eigenvalue of the GUE, we are able to split this integral into the bulk and the outliers, and bound both components.

Combing the all the above lemmas as well as factoring in some additional fluctuations of the trace moments, we now state our construction in a form with effectively isolated eigenvalues, a well-chosen θ to minimize moments, and an error to represent the imperfections in our finite-dimension GUE

$$m{W} = m{U} m{D_1} m{U}^\dagger m{V} m{D_2} m{V}^\dagger \quad ext{such that} \quad \overline{ ext{Tr}} [m{D}_i] pprox \mathcal{O}(rac{1}{\sqrt{N}})$$

where U and V are Haar-random unitaries and $D_i \in D_1, D_2$ are diagonal matrices with spectra matching the respective exponentiated GUEs $e^{iG_1\theta}$, $e^{iG_2\theta}$. With this established, we can now work on establishing a bound on the trace moments of our construction.

4.3 Bounding the moments of the construction

We seek to bound the moments of the exponentiated GUE product. We get this from **Lemma IV.3** in the paper [Che+24, Lemma IV.3].

Lemma IV.3. The expectation of the pth trace moment is [Che+24, Lemma IV.3]

$$\left| \mathbb{E}\left[\overline{Tr} [\boldsymbol{W}^p] \right] \right| = \mathcal{O}\left(\frac{p^{15/4}}{\sqrt{N}} \right) + p \, \mathbb{E}\left[\left| \overline{Tr} \boldsymbol{D}_2 \right| + \left| \overline{Tr} \boldsymbol{D}_1 \right| + \left| \overline{Tr} \boldsymbol{D}_1 \right| \left| \overline{Tr} \boldsymbol{D}_2 \right| \right].$$

We will now break down the steps to proving this lemma. Starting where we left off in the previous section, can expand our expectation

$$\mathbb{E}\left[\overline{\text{Tr}}[\boldsymbol{W}^p]\right] = \mathbb{E}\left[\left(e^{i\theta G_1}e^{i\theta G_2}\right)^p\right] = \mathbb{E}_{\boldsymbol{D}_1,\boldsymbol{D}_2}\left[\mathbb{E}_{\boldsymbol{U},\boldsymbol{V}}\left[\overline{\text{Tr}}\left[\left(\boldsymbol{U}\boldsymbol{D}_1\boldsymbol{U}^{\dagger}\boldsymbol{V}\boldsymbol{D}_2\boldsymbol{V}^{\dagger}\right)^p\right]\right]\right].$$

We split up the expectation because we seek to address the randomness of the Haar-random basis (U, V) separately from the randomness of the eigenvalues (D_1, D_2) . This allows us to fix the diagonals and focus on the two Haar ensembles. For some fixed (D_1, D_2) , let us pull out the second expectation

$$f(\boldsymbol{D}_1, \boldsymbol{D}_2) := \mathbb{E}_{\boldsymbol{U}, \boldsymbol{V}} \left[\overline{\mathrm{Tr}} [\boldsymbol{W}^p] \right].$$

We now come to a crux of the paper: by bounding this $f(\mathbf{D}_1, \mathbf{D}_2)$ we will be able to bound the moments of our construction. We bound this using the polynomial method [Che+24, Lemma VII.6]

Lemma VII.6. There exists a bound for the derivative near zero an algebraic function of the form

$$r_n(x) = \frac{p_n(x)}{\sqrt{t_{2n}(x)}}$$

where $p_n(x)$ is an algebraic polynomial of degree at most n with complex coefficients and $t_{2n}(x)$ is defined by very specific conditions.

For intuition, the idea of the lemma is that we get a bound on the derivative of $\lim_{x\to 0} r_n(x)$ in terms of bounds on $r_n(x)$ at finitely many points x. Thinking of r_n as an algebraic function in x=1/N, this tells us how fast it coverages to the limit value as N goes to infinity. Since we want to compute the average of $f(\mathbf{D}_1, \mathbf{D}_2)$ over all $\mathbf{D}_1, \mathbf{D}_2$ for large finite N, we need this lemma to get a precise numerical estimate on our function. For simplicity, we will blackbox parts of this lemma and the subsequent proof of it, mainly the exact details of the very specific conditions of the algebraic function and the unwieldy bound put on $r_n(x)$. Most importantly, we state that our current function $f(\mathbf{D}_1, \mathbf{D}_2)$ does not match these very specific conditions, but it does look like a rational function of N that could reasonably match these conditions. The authors spend most of the rest of this section of the paper massaging it to get something we can apply to **Lemma VII.6**. To make use of the lemma, we need to be able to do the following

- 1. Express some variant of $f(\mathbf{D}_1, \mathbf{D}_2)$ as a low-degree rational function of 1/N with manageable poles
- 2. Compute the large-N limit i.e. the limit of this rational function at 0
- 3. Obtain a bound for this rational function at a certain finite set of points x

We use Weingarten calculus to produce a rational function $f(\mathbf{D}_1, \mathbf{D}_2)$ such that

$$f(\boldsymbol{D}_1, \boldsymbol{D}_2) = f_{\boldsymbol{D}_1, \boldsymbol{D}_2}(N).$$

Weingarten calculus is a tedious (at least in this application) method to compute moments of products of Haar-random matrices. Hand waving the details, the key part that allowed us to create this equality is through fixing the trace moments while allowing the dimension N to vary. Weingarten calculus will eventually give us both a limiting value and control over the poles.

To compute the large-N limit, from the leading order behavior in the infinite N limit from Weingarten calculus, we get

$$\lim_{N \to \infty} |f_{\boldsymbol{D}_1, \boldsymbol{D}_2}(N)| \le p(\left|\overline{\operatorname{Tr}}[\boldsymbol{D}_1]\right| + \left|\overline{\operatorname{Tr}}[\boldsymbol{D}_2]\right| + \left|\overline{\operatorname{Tr}}[\boldsymbol{D}_1]\right| \left|\overline{\operatorname{Tr}}[\boldsymbol{D}_2]\right|).$$

This bound comes from taking a small perturbation of f_{D_1,D_2} and showing that it goes to zero in the limit using the Weingarten calculus. Then, by showing that this small perturbation does not significantly change the limit f_{D_1,D_2} , it implies that f_{D_1,D_2} is small in the limit as well. Control over the poles can be accomplished by stretching the domain and tweaking some unwanted dependencies.

Finally, we seek to bound $f_{\mathcal{D}_1,\mathcal{D}_2}(N)$ at finitely many inputs N. However, we only know $f_{\mathcal{D}_1,\mathcal{D}_2}(N)$ comes from trace moments of actual unitaries for the one specific value of N corresponding to \mathcal{D}_1 and \mathcal{D}_2 . As a result, we do not know much about $f_{\mathcal{D}_1,\mathcal{D}_2}$ away from this specific value of N. We solve this by creating a family of unitaries $(\mathcal{D}'_{1,N},\mathcal{D}'_{2,N})$ of different dimension N such that $f_{\mathcal{D}_1,\mathcal{D}_2}(N) = f(\mathcal{D}'_{1,N},\mathcal{D}'_{2,N})$. As a result, we can bound $f_{\mathcal{D}_1,\mathcal{D}_2}(N)$ using the trace moments from random matrix theory. The designs $\mathcal{D}'_{1,N}$, $\mathcal{D}'_{2,N}$ are described in more detail in the paper [Che+24, Lemma V]. They are part of a large question known as the unitary moment problem, which essentially says that given some moments, we can construct unitaries from them.

Having completed all the tasks in the list, we finally can use **Lemma VII.6** on the derivative to bound the difference between $\lim_{N\to\infty} f_{D_1,D_2}(N)$ and $f_{D_1,D_2}(N)$ for finite N. We will then interpolate this to a bound on $f_{D_1,D_2}(N)$ for large finite N.

Armed with our bound on $f_{D_1,D_2}(N)$, we can effectively conclude that our construction has small moments, thereby allowing us to conclude that it is a T-design.

5 Conclusion

Through methods from random matrix theory, we have constructed a T-design, an ensemble reproducing the first T moments of the Haar measure. This was done by taking the product of two exponentiated GUEs, which we computationally and mathematically confirmed matches Haar-random unitaries. T-designs have a number of applications, most notably in quantum information theory. An example application is in randomized benchmarking, where they can be used to generate sets random quantum gates.

References

[Che+24] C-F. Chen et al. "Efficient Unitary T-designs from Random Sums". In: 65th Annual Symposium on Foundations of Computer Science (FOCS 2024). 2024.