Cutting circuits with multiple two-qubit unitaries

Journal Article

Author(s):
Schmitt, Lukas; Piveteau, Christophe; Sutter, David

Publication date:
2025-02-18

Permanent link:
https://doi.org/https://doi.org/10.3929/ethz-b-000728223

Originally published in:
Quantum 9, https://doi.org/10.22331/q-2025-02-18-1634

Funding acknowledgement:
- (QuantEOM) Quantum-coherent electro-optic microwave-to-optical conversion with GaP and BaTiO3 ()

Cutting circuits with multiple two-qubit unitaries

Lukas Schmitt[1,2], Christophe Piveteau[1], and David Sutter[2]

[1]Institute for Theoretical Physics, ETH Zurich
[2]IBM Quantum, IBM Research Europe – Zurich

Quasiprobabilistic cutting techniques allow us to partition large quantum circuits into smaller subcircuits by replacing non-local gates with probabilistic mixtures of local gates. The cost of this method is a sampling overhead that scales exponentially in the number of cuts. It is crucial to determine the minimal cost for gate cutting and to understand whether allowing for classical communication between subcircuits can improve the sampling overhead. In this work, we derive a closed formula for the optimal sampling overhead for cutting an arbitrary number of two-qubit unitaries and provide the corresponding decomposition. We find that cutting several arbitrary two-qubit unitaries together is cheaper than cutting them individually and classical communication does not give any advantage.

1 Introduction

To demonstrate a quantum advantage in the close future, various challenges need to be resolved such as correcting or reducing noise, dealing with the limited connectivity (on some hardware platforms), and scaling the number of available qubits. The latter obstacle motivated the development of techniques known as *circuit knitting* or *circuit cutting*. The idea is to split large circuits into subcircuits that can be executed on smaller devices. Partitioning a circuit can be achieved by performing space-like or time-like cuts, which are typically referred to as *gate cuts* [1, 2, 3] and *wire cuts* [4, 5, 6, 7], respectively. Circuit cutting techniques typically result in a sampling overhead that scales exponentially in the number of cuts. One particular method of circuit knitting is based on a technique called *quasiprobability simulation*. This method works by probabilistically replacing the non-local gates across the various subcircuits by local operations. By performing appropriate classical post-processing, the expectation value of the original large circuit can be retrieved. More concretely, this means that we can simulate the large quantum circuit by only physically executing the small subcircuits on small quantum devices.

While quasiprobabilistic circuit cutting allows us to retrieve the expectation value of the larger circuit with arbitrary precision, the number of shots required to achieve a desired accuracy is increased compared to physically running the original circuit. Suppose we cut a non-local unitary U, then the optimal achievable sampling overhead is characterized by the quantity $\gamma(U)$ which we call the γ-factor of U. More precisely, the number of shots increases multiplicatively by $\gamma(U)^2$. If we separately apply the quasiprobability simulation technique to n different gates $(U_i)_{i=1}^n$ in the circuit, then the total sampling overhead behaves multiplicatively $\prod_{i=1}^n \gamma(U_i)$. This is the reason for the exponential scaling of the sampling overhead with respect to the number of cut gates.

Accepted in Quantum 2025-02-08, click title to verify.

A central question that has been studied in previous works [3, 8, 5] is whether allowing the smaller subcircuits to exchange classical information can improve the sampling overhead. To distinguish between the γ-factor in the two settings where classical communication is allowed or not, we introduce a corresponding subscript γ_{LO}, respectively γ_{LOCC}, where LO stands for *local operations* and LOCC for *local operations with classical communication*. In the setting of wire cutting, it was recently proven that there is a strict separation between LO and LOCC, in the sense that the γ-factor of a wire cut is strictly smaller when classical communication is allowed [5]. However, for gate cutting it is unknown if there is a separation between the two settings. Computing γ-factors has proven to be a difficult problem in practice, and explicit values are only known for a few gates. For these gates, there is no separation [3]. Hence, it is natural to ask:

Question 1. *Is* LOCC *strictly more powerful for circuit cutting than* LO*?*

An additional difficulty in analyzing sampling overheads in circuit cutting is that the γ-factor only characterizes the optimal overhead for cutting a single non-local gate. When one considers cutting a circuit with multiple non-local gates, the γ-factors of the individual gates do not fully capture the optimal overhead of cutting the complete circuit at once. For example, it has been shown [3] that for a Clifford gate U and in the setting of LOCC it is strictly cheaper to cut $n > 1$ copies of the gate that happen in the same time slice (as seen in Figure 1b), as opposed to using the optimal single-gate cutting procedure n times. Mathematically, this is captured by a strictly submultiplicative behavior of the γ-factor

$$\gamma_{\mathrm{LOCC}}(U^{\otimes n}) < \gamma_{\mathrm{LOCC}}(U)^n \,, \tag{1}$$

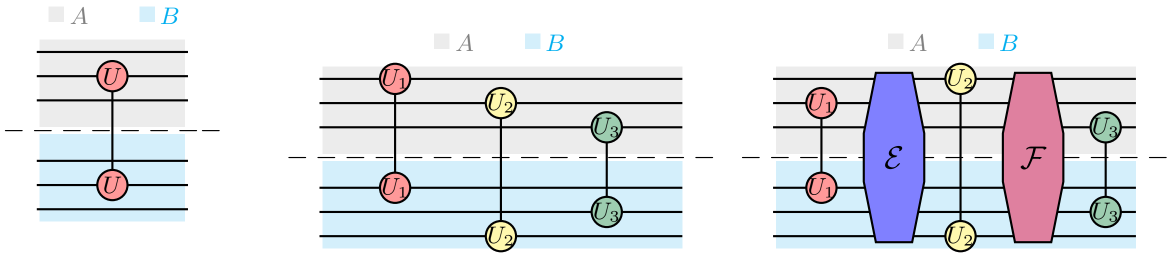
where $U^{\otimes n}$ stands for n parallel copies of the non-local gate U each acting on a distinct set of qubits. Previously to this work, it was unknown whether such a strict submultiplicative behavior also exists for non-Clifford gates U and/or in the absence of classical communication (i.e. with γ_{LO} instead of γ_{LOCC}). We thus ask:

Question 2. *Is cutting arbitrary gates simultaneously cheaper than cutting them separately?*

When applying circuit cutting in practice, one will typically be confronted with the situation that the large circuit contains multiple non-local gates which occur at possibly very distant points in time. The sub-multiplicative behavior of the γ-factor cannot be directly applied to cut multiple non-local gates that do not occur in the same time slice. However, one can still try to apply the insight that "global" cutting strategies can be more efficient by dealing with multiple gates at once instead of treating them individually.

For better illustration of the different problems at hand, we introduce three separate settings depicted in Figure 1, that we will consider throughout this work. In the *single cut* setting, we consider only one single instance of some non-local gate U that we want to cut. Here, the optimal sampling overhead is fully characterized by the γ-factor of U. In the *parallel cut* setting, we consider multiple gates $(U_i)_{i=1}^n$ that we want to cut which all occur in the same time slice of the circuit. Note that the parallel cut setting is a special case of the single-cut setting with $U = \bigotimes_{i=1}^n U_i$ and the optimal sampling overhead is thus fully characterized by the γ-factor of $\bigotimes_{i=1}^n U_i$.

The case where we consider a circuit with multiple non-local gates $(U_i)_{i=1}^n$ that occur at different time slices is a bit more involved. Clearly, the optimal procedure would be to consider the whole circuit (including local and non-local gates) as one big unitary U_{tot} and then determining the γ-factor of that. However, this is typically intractable as just determining U_{tot} is as difficult as simulating the circuit itself. For this purpose we



(a) Single cut (b) Parallel cut (c) Black box cut

Figure 1: Three different cutting scenarios, where the dashed line represents the cut. The unitaries U_i are arbitrary two-qubit gates acting on A and B. In the black box setting $\mathcal{E}$ and $\mathcal{F}$ denote arbitrary CPTP maps.

introduce a third setting, which we call the *black box* setting. Here, we consider multiple non-local gates $(U_i)_{i=1}^n$ which we want to cut, but between these gates we may have unknown operations. The circuit cutting technique is not allowed to have any information of what happens inside them (hence the name "black box") and must work for any choice of the unknown operations.

Clearly the black box setting is at least as difficult as the parallel cut setting, as the latter can be retrieved as a special choice of the black box. It is thus natural to ask:

Question 3. *In the black box setting, can cutting the gates jointly improve the sampling overhead?*

Here again, the answer of Question 3 is known for Clifford gates under LOCC [3]. The gate-based teleportation scheme used by the authors makes it possible to reduce the black box setting to the parallel cut setting, showing that the optimal sampling overhead is the same in both cases. Since gate teleportation requires classical communication and only works for Clifford gates, it was previously unknown whether the same answer to Question 3 would also hold in a more general setting.

Results In this work, we answer Questions 1 to 3 for the case where the unitaries to be cut are all two-qubit gates. It is well known, that an arbitrary two-qubit unitary U exhibits a so-called KAK decomposition

$$U = (V_1 \otimes V_2)\left(\sum_{k=0}^{3} u_k \sigma_k \otimes \sigma_k\right)(V_3 \otimes V_4)\,, \tag{2}$$

where $\sigma_0 = \mathbb{1}$, $\sigma_1, \sigma_2, \sigma_3$ are the Pauli matrices, the $u_i \in \mathbb{C}$ fulfill $\sum_{i=0}^{3}|u_i|^2 = 1$ and the V_i are single-qubit unitaries. For any such unitary U we can define the quantity

$$\Delta_U := \sum_{k \neq k'} |u_k||u_{k'}|\,, \tag{3}$$

which is nonnegative and zero only for a separable unitary. We show in Corollary 3.1 that the γ-factor of U is given by

$$\gamma_{\mathrm{LO}}(U) = \gamma_{\mathrm{LOCC}}(U) = 1 + 2\Delta_U. \tag{4}$$

Furthermore, in Corollary 4.1 we characterize the γ-factor in the parallel cut setting and show that

$$\gamma_{\mathrm{LO}}(U^{\otimes n}) = \gamma_{\mathrm{LOCC}}(U^{\otimes n}) = 2(1+\Delta_U)^n - 1 < \gamma_{\mathrm{LO}}(U)^n = \gamma_{\mathrm{LOCC}}(U)^n\,. \tag{5}$$

As a consequence, we show in Corollary 4.6 that the regularized γ-factor, i.e. the effective sampling overhead per gate in the limit of many gates, is given by

$$\lim_{n\to\infty}\left(\gamma_{\mathrm{LO}}(U^{\otimes n})\right)^{1/n} = \lim_{n\to\infty}\left(\gamma_{\mathrm{LOCC}}(U^{\otimes n})\right)^{1/n} = 1+\Delta_U\,. \tag{6}$$

We therefore have a full understanding of the optimal sampling overhead of two-qubit gates in the single-cut and parallel cut settings, and classical communication does not reduce the overhead in either of them. Our results are derived using a more general characterization of the γ-factor of general (possibly larger than two-qubit) unitaries that exhibit a form analogous to Equation (2), as seen in Theorem 5.1. Generally, our techniques can be seen as a generalization and improvement of the ideas introduced in [2].

To tackle the black box setting, we devise an explicit protocol to perform black box cutting by reducing it to the parallel setting. Therefore, the overhead is the same as in the parallel setting (i.e. as in Equation (5)). Since the black box setting is at least as difficult than the parallel setting, our result is therefore optimal. As in the approach of [3], this protocol requires additional ancilla qubits to carry information across the black boxes. However, our protocol does not require classical communication and it also works for non-Clifford gates.

In summary, we answer the questions as follows:

Answer to Question 1: For quasiprobabilistic circuit cutting, classical communication does not improve the overhead of cutting arbitrary two-qubit gates.

Answer to Question 2: The answer is positive for all (non-local) two-qubit gates. In other words, the γ-factor behaves strictly submultiplicatively.

Answer to Question 3: For arbitrary two-qubit gates, the black box setting has the same sampling overhead as the parallel cut setting (without requiring classical communication). This comes at the cost of four additional ancilla qubits per gate.

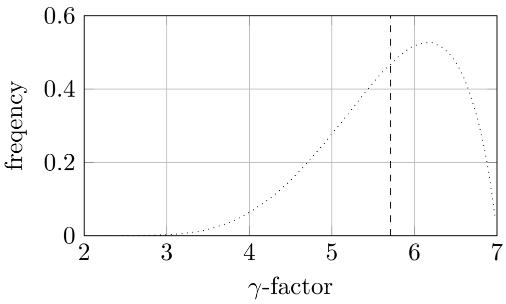
Setting	Reference	Example (Figure 1)
Single cut	Corollary 3.1	$\gamma(\mathrm{CNOT})=3$, $\gamma(\mathrm{CR}_X(\theta))=1+2\lvert\sin(\theta/2)\rvert$
Parallel cut	Corollary 4.1	$\gamma(\mathrm{CNOT}\otimes\mathrm{CR}_X(\theta))=3+4\lvert\sin(\theta/2)\rvert$
Black box cut	Theorem 6.2	$\bar{\gamma}(\mathrm{CNOT},\mathrm{CR}_X(\theta))=3+4\lvert\sin(\theta/2)\rvert$

Table 1: Summary on how to optimally cut two-qubit gates. The example considers the setting from Figure 1 with $U_1=\mathrm{CNOT}$, $U_2=\mathrm{CR}_X(\theta)$, and $U_3=\mathbb{1}$. The KAK parameters of these unitaries are given in Example 3.2. Note that for these examples, we write γ without subscript, because the overhead for LO and LOCC is the same. Our method in the black box setting requires additional ancillary qubits, however this is not the case for the other two settings.

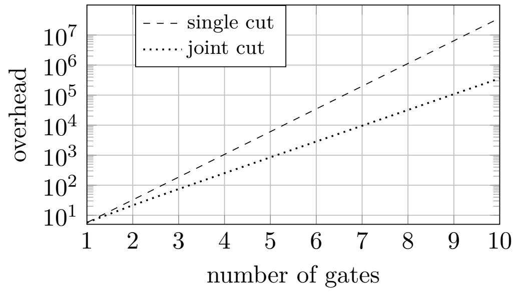
The results together with a small example are summarized in Table 1. The example given in Table 1 shows that the submultiplicative property of the γ-factor does not just appear when cutting multiple identical gates, but also when one cuts multiple distinct gates:

$$\gamma(\mathrm{CNOT}\otimes\mathrm{CR}_X(\theta)) = 3+4\lvert\sin(\theta/2)\rvert < 3+6\lvert\sin(\theta/2)\rvert) = \gamma(\mathrm{CNOT})\gamma(\mathrm{CR}_X(\theta))\,. \tag{7}$$

The difference between the two sides of Equation (7) can be substantial. The submultiplicativity of the γ-factor becomes particularly meaningful if we cut more than just two gates simultaneously. Figure 2 shows the sampling overhead for a two-qubit unitary that is sampled from the Haar measure.



(a) Distribution of γ-factor for a Haar-randomly sampled two-qubit gate U: $\gamma(U)$ is plotted on the x-axis. The y-axis shows the relative frequency out of 10^7 samples. The dashed line shows the average γ-factor which is approximately $\gamma(U) \approx 5.71$.



(b) Overhead for single and parallel cut for multiple Haar-random gates. The x-axis shows the number of gates to cut. The y-axis shows the overhead. Both methods scale exponentially in the number of gates, but the base of the parallel cut is lower.

Figure 2: Submultiplicativity of γ-factor for Haar-random two-qubit unitaries.

2 Preliminaries

2.1 Quasiprobability simulation

Here we briefly sketch the quasiprobability simulation framework on which our circuit cutting method is based. For more thorough explanations, we refer the reader to [3, 9, 10, 1, 11].

Consider that we are given a circuit that prepares some state and measures an observable, and we want to compute the expectation value of that observable. Furthermore, that circuit includes some gate $\mathcal{E}$ that we cannot physically run on our computer. In the context of gate cutting, we can only physically run operations that act locally on the two circuit bipartitions, and $\mathcal{E}$ would correspond to the channel of some non-local gate.

Quasiprobability simulation is a method of simulating a linear operation $\mathcal{E}$ via a set of available operations S that can be physically realized. Concretely, we can simulate an operation $\mathcal{E}$ if we can decompose it as

$$\mathcal{E} = \sum_{i=1}^{m} a_i \mathcal{F}_i \,, \tag{8}$$

where $a_i \in \mathbb{R}$ and $\mathcal{F}_i \in S$. This decomposition can be rewritten as

$$\mathcal{E} = \sum_{i=1}^{m} p_i \mathcal{F}_i \cdot \operatorname{sign}(a_i)(\sum_i |a_i|) \,, \tag{9}$$

for the probability distribution $p_i = |a_i|/(\sum_i |a_i|)$. Equation (9) implies that the expectation value of the circuit can be obtained with following Monte Carlo sampling technique: Every shot of the circuit, a random index i is sampled according to the probability distribution $(p_i)_i$. Then, the gate $\mathcal{E}$ is replaced by the operations $\mathcal{F}_i$ and this modified circuit is physically executed on the hardware. The final measurement outcome is multiplied by $\operatorname{sign}(a_i)(\sum_i |a_i|)$ and stored in memory. By repeating this process many times and averaging the resulting outcomes, Equation (9) implies that one will obtain the expectation value of the original circuit.

However, because the post-processing boosts the magnitude of the measurement outcomes by $(\sum_i |a_i|)$, the method incurs an additional sampling overhead if one wants to estimate the expectation value to the same accuracy. More concretely, to estimate the expectation value to an additive error of at most ϵ with a probability of at least, $1-\delta$, Hoeffding's bound implies that this can be achieved with a number of shots equal to $C \cdot (\sum_i |a_i|)^2 \epsilon^{-2} \ln\left(\frac{1}{2\delta}\right)$ for some constant C. So clearly, if one wants to use the quasiprobability simulation technique, one should use a decomposition that exhibits the smallest possible value of $(\sum_i |a_i|)$. Since the simulation overhead scales with $(\sum_i |a_i|)^2$, we assign a special name for the smallest achievable value of $(\sum_i |a_i|)$:

Definition 2.1. The γ-factor of an operation $\mathcal{E}$ over S is defined as

$$\gamma_S(\mathcal{E}) := \min\Big\{ \sum_{i=1}^{m} |a_i| : \mathcal{E} = \sum_{i=1}^{m} a_i \mathcal{F}_i \text{ where } m \geq 1, \mathcal{F}_i \in S \text{ and } a_i \in \mathbb{R} \Big\}. \tag{10}$$

A particular application of quasiprobability simulation is gate cutting [1, 3]. Consider a bipartition of the qubits in the circuit into two systems A and B corresponding to the subcircuits obtained from the circuit cutting. The operation $\mathcal{E}$ is then a unitary channel $\mathcal{E}(\rho) := U\rho U^\dagger$ for some unitary U that acts non-locally across A and B. As a slight abuse of notation, we sometimes denote the γ-factor of a unitary channel simply as $\gamma_S(U)$. The set of operations S into which we decompose our gate is then chosen to be either local operations on the systems A and B (denoted by $\mathrm{LO}(A,B)$) or local operations with classical communication ($\mathrm{LOCC}(A,B)$). If this quasiprobability simulation is applied for all gates acting non-locally across A and B, we can thus replace the large non-local circuit with a probabilistic mixture of local circuits on A and B that do not have any entangling gates between them (plus some additional classical post-processing). The method therefore allows us to obtain the expectation value of the original circuit by only physically executing smaller circuits.

Following [3], we defined the set of local operations $\mathrm{LO}(A,B)$ across the systems A, B as follows:

Definition 2.2. The set $\mathrm{LO}(A,B)$ is given by $\{\mathcal{F}_A \otimes \mathcal{F}_B \mid \mathcal{F}_A \in \mathcal{N}(A) \,\text{and}\, \mathcal{F}_B \in \mathcal{N}(B)\}$, where $\mathcal{N}(X)$ is defined as[1]

$$\mathcal{N}(X) = \left\{ \mathcal{A}^+ - \mathcal{A}^- \mid \mathcal{A}^+, \mathcal{A}^- \text{completely positive}, \mathcal{A}^+ + \mathcal{A}^- \in \mathrm{CPTP}(X) \right\}. \tag{11}$$

where $\mathrm{CPTP}(X)$ is the set of all completely-postive trace-preserving maps on the system X.

This is a slightly non-standard definition of non-local operations, since it allows for non-completely positive maps. The reason for this is that this definition captures the idea, first seen in [1] and later expanded upon in [12], that the individual operations $\mathcal{F}_i$ can be realized as quantum instruments, and depending on which component of the instrument is executed, the measurement outcome at the end of the circuit is weighted with $+1$ or -1. As an illustrative example, consider two completely-positive trace-nonincreasing maps $\mathcal{A}^+$ and $\mathcal{A}^-$ s.t. $\mathcal{A}^+ + \mathcal{A}^- \in \mathrm{CPTP}$. One can simulate the map $\mathcal{A}^+ - \mathcal{A}^-$ via the CPTP-map $\mathcal{F}$

$$\mathcal{F}(\rho) = \mathcal{A}^+(\rho) \otimes |0\rangle\langle 0| + \mathcal{A}^-(\rho) \otimes |1\rangle\langle 1| \tag{12}$$

[1] It is even sufficient to only demand $\mathcal{A}^+ + \mathcal{A}^- \in \mathrm{CPTN}(X)$, but this can be achieved in the CPTP setting by postweighting with 0.

followed by a measurement of the ancillary qubit in the Z-basis. Depending on the ancilla measurement outcome, one then weights the final measurement outcome of the circuit by ± 1. We refer to this as the *negativity* trick.

A characterization of the set of $\mathrm{LOCC}(A, B)$ is very complicated, since it can contain multiple rounds of operations and communications between A and B, so we refrain from writing the full definition here. For an in-depth definition we refer the reader to [13]. Since LOCC is usually defined as a quantum instrument, we can realize the same post-processing trick as with LO above and allow for non-completely positive maps by weighting some of the element of the instrument by -1.

2.2 Representation of two-qubit unitaries

It is well known that a semi-simple Lie algebra $\mathfrak{g}$ can be decomposed as $\mathfrak{g} = \mathfrak{m} \oplus \mathfrak{k}$, $\mathfrak{m} = \mathfrak{k}^{\perp}$, with the following relations

$$[\mathfrak{k}, \mathfrak{k}] \subset \mathfrak{k}\,, \qquad [\mathfrak{k}, \mathfrak{m}] = \mathfrak{m}\,, \qquad \text{and} \qquad [\mathfrak{m}, \mathfrak{m}] \subset \mathfrak{k}\,. \tag{13}$$

This decomposition is called Cartan decomposition. For such a pair, one can find the corresponding Cartan subalgebra $\mathfrak{h}$, which is the largest subalgebra that is contained in $\mathfrak{m}$. For these algebras, it is well known, that the Lie group G can be written as

$$G = KAK\,, \tag{14}$$

where $K = \exp(\mathfrak{k})$ and $A = \exp(\mathfrak{h})$ (see for instance [14]). This decomposition is also called the KAK decomposition.

Applied to $\mathfrak{g} = \mathfrak{su}(2^n)$, we find $\mathfrak{k} = \mathfrak{su}(2^{n-1}) \oplus \mathfrak{su}(2^{n-1})$, which is in general not useful for circuit cutting. However, in the case of $\mathfrak{g} = \mathfrak{su}(4)$, this simplifies to $\mathfrak{k} = \mathfrak{su}(2) \oplus \mathfrak{su}(2)$ and therefore $K = SU(2) \otimes SU(2)$. These elements can be implemented with local operations. For the cartan subalgebra one finds $\mathfrak{h} = \{\mathrm{i}\sigma_1 \otimes \sigma_1, \mathrm{i}\sigma_2 \otimes \sigma_2, \mathrm{i}\sigma_3 \otimes \sigma_3\}$, where we denoted the Pauli matrices by $\sigma_1 = X$, $\sigma_2 = Y$, $\sigma_3 = Z$. Furthermore let $\sigma_0 = \mathbb{1}$ denote the identity matrix.

Using the above, it follows that any two-qubit unitary U(4) has a KAK decomposition

$$U = (V_1 \otimes V_2) \exp\left(\sum_{k=1}^{3} \mathrm{i}\theta_k \sigma_k \otimes \sigma_k\right) (V_3 \otimes V_4)\,, \tag{15}$$

for some single-qubit unitaries V_1, V_2, V_3, V_4, and $\theta_k \in \mathbb{R}$ such that $|\theta_3| \leq \theta_2 \leq \theta_1 \leq \pi/4$. Evaluating the exponential function, this can be rewritten as

$$U = (V_1 \otimes V_2) \left(\sum_{k=0}^{3} u_k \sigma_k \otimes \sigma_k\right) (V_3 \otimes V_4) =: (V_1 \otimes V_2)\, W\, (V_3 \otimes V_4)\,, \tag{16}$$

where $u_k \in \mathbb{C}$ depend on $\theta_1, \theta_2, \theta_3$ and fulfill $\sum_{i=0}^{3} |u_i|^2 = 1$. Equation (16) is central in our analysis of optimal gate cutting of two-qubit unitaries, as cutting U is equivalent to cutting the interaction unitary W. In particular we have $\gamma_S(U) = \gamma_S(W)$ and in the proof of Theorem 5.1, we thus focus on cutting W. Since there does not exist an analogous decomposition to Equation (16) for general unitaries, optimal cutting still remains an open problem.

3 Cutting one single two-qubit unitary

In this section we show how to optimally cut an arbitrary two-qubit gate.

Corollary 3.1. *Let U be a two-qubit unitary with KAK parameters $(u_0, u_1, u_2, u_3) \in \mathbb{C}^4$. Then,*

$$\gamma_{\mathrm{LO}}(U) = \gamma_{\mathrm{LOCC}}(U) = 1 + 2\sum_{i\neq j} |u_i||u_j| \,. \tag{17}$$

Proof. Corollary 3.1 follows as a special case of Theorem 5.1. Since every two-qubit unitary can be written as in Equation (16), we can directly apply Theorem 5.1 with $L_k = \sigma_k$ and $R_k = \sigma_k$. The coefficients u_k are directly given through the KAK coefficients of U. The optimal decomposition of U can be found in the proof of Theorem 5.1. □

The result proves that there is no advantage for cutting a two-qubit unitary with LOCC compared to the LO setting. In the following, whenever this is the case, we will drop the subscript of the γ-factor. The previously best known decomposition [1] had a sampling overhead of $1 + \sum_{i\neq j}(|u_i u_j^* + u_j u_i^*| + |u_i u_j^* - u_j u_i^*|)$, which is only optimal for certain gates.

Example 3.2. Let us apply Corollary 3.1 to some well-known two-qubit unitaries to check consistency with [3]:

(i) $\gamma(\mathrm{CNOT}) = 3$, since $u = \left(\frac{1}{\sqrt{2}}, \frac{\mathrm{i}}{\sqrt{2}}, 0, 0\right)$.

(ii) $\gamma(\mathrm{SWAP}) = 7$, since $u = \left(\frac{\sqrt{2}(1+\mathrm{i})}{4}, \frac{\sqrt{2}(1+\mathrm{i})}{4}, \frac{\sqrt{2}(1+\mathrm{i})}{4}, \frac{\sqrt{2}(1+\mathrm{i})}{4}\right)$.

(iii) $\gamma(\mathrm{CR}_X(\theta)) = 1 + 2|\sin(\theta/2)|$, since $u_{\mathrm{CR}_X(\theta)} = (\cos(\frac{\theta}{4}), \mathrm{i}\sin(\frac{\theta}{4}), 0, 0)$.[2]

Corollary 3.3. *Let U be a two-qubit unitary, then $1 \leq \gamma(U) \leq 7$.*

Proof. Clearly $\gamma(U)$ cannot be smaller than 1. To see that it is bounded from above by 7, consider KAK parameters $u := (u_0, u_1, u_2, u_3) \in \mathbb{C}^4$. As a consequence of unitarity, these fulfill $\|u\|_2 = 1$. Corollary 3.1 allows us to write

$$\gamma(U) = 1 + 2\sum_{i\neq j} |u_i||u_j| \tag{18}$$

$$= \sum_{i=0}^{3} |u_i|^2 + 2\sum_{i\neq j} |u_i||u_j| \tag{19}$$

$$= 2\left(\sum_{i=0}^{3} |u_i|\right)^2 - \sum_i |u_i|^2 \tag{20}$$

$$= 2\,\|u\|_1^2 - 1 \tag{21}$$

$$\leq 7\,, \tag{22}$$

where the final step uses the Cauchy-Schwarz inequality, i.e., $\|u\|_1 \leq \sqrt{4}\,\|u\|_2$. □

[2] $\mathrm{CR}_X(\theta)$ denotes a controlled $\mathrm{R}_X(\theta)$ gate with rotation angle θ, where $\mathrm{R}_X(\theta) := \mathrm{e}^{-\mathrm{i}\frac{\theta}{2}X}$.

4 Cutting parallel two-qubit unitaries

In Section 3 we have seen how to optimaly cut an arbitrary two-qubit unitary. In this section we would like to extend the setting to cutting multiple two-qubit unitaries which are arranged in parallel (cf. Figure 1b).

Corollary 4.1. *Let $n \in \mathbb{N}$ and $(U^{(i)}_{A_iB_i})_{i=1}^n$ be a family of two-qubit unitaries with KAK parameters $(u_0^{(i)}, u_1^{(i)}, u_2^{(i)}, u_3^{(i)})_{i=1}^n$. Then, for $U_{AB} = \otimes_{i=1}^n U^{(i)}_{A_iB_i}$, where $A = \otimes_{i=1}^n A_i$ and $B = \otimes_{i=1}^n B_i$ we have*

$$\gamma_{\mathrm{LO}(A:B)}(U_{AB}) = \gamma_{\mathrm{LOCC}(A:B)}(U_{AB}) = 1 + 2\sum_{\mathbf{k}\neq\mathbf{k}'} |u_{\mathbf{k}}||u_{\mathbf{k}'}|\,, \tag{23}$$

where $\mathbf{k} \in \{0,1,2,3\}^n$ and $u_{\mathbf{k}} = \prod_{i=1}^n u^{(i)}_{k_i}$.

Proof. Corollary 4.1 follows as a consequence of Theorem 5.1. The KAK decomposition from Equation (16) allows us to write for any $i \in \{1,\dots,n\}$

$$U^{(i)}_{A_iB_i} = \left(V^{(1,i)}_{A_i} \otimes V^{(2,i)}_{B_i}\right)\left(\sum_{k=0}^{3} u^{(i)}_k (\sigma_k)_{A_i} \otimes (\sigma_k)_{B_i}\right)\left(V^{(3,i)}_{A_i} \otimes V^{(4,i)}_{B_i}\right)\,. \tag{24}$$

This implies

$$\begin{aligned} &U_{AB} \\ &= \bigotimes_{i=1}^n U^{(i)}_{A_iB_i} &(25)\\ &= \left(\bigotimes_{i=1}^n V^{(1,i)}_{A_i} \otimes \bigotimes_{i=1}^n V^{(2,i)}_{B_i}\right) \sum_{k_1,\dots,k_n=0}^{3} \left[\left(\prod_{j=1}^n u^{(j)}_{k_j}\right)\left(\bigotimes_{i=1}^n \sigma_{k_i}\right)_A \otimes \left(\bigotimes_{i=1}^n \sigma_{k_i}\right)_B\right] \left(\bigotimes_{i=1}^n V^{(3,i)}_{A_i} \otimes \bigotimes_{i=1}^n V^{(4,i)}_{B_i}\right), &(26)\end{aligned}$$

and fulfills therefore the assumptions of Theorem 5.1. □

The expression of the γ-factor simplifies further in the case that many of the unitaries $U^{(i)}$ are identical. For instance, when there are three different types of unitaries, the γ-factor behaves as follows (the same result holds analogously for a number of gates different than three):

Corollary 4.2 (Calculating the overhead)**.** *Let U, V, W be two-qubit unitaries and $n, m, p \in \mathbb{N}$. Then,*

$$\gamma(U^{\otimes n} \otimes V^{\otimes m} \otimes W^{\otimes p}) = 2\Big(\sum_{i=0}^{3} |u_i|\Big)^{2n}\Big(\sum_{i=0}^{3} |v_i|\Big)^{2m}\Big(\sum_{i=0}^{3} |w_i|\Big)^{2p} - 1\,, \tag{27}$$

where u_i, v_i and w_i are the corresponding KAK parameters of U, V, and W, respectively.

Proof. As seen in Corollary 4.1, we can calculate $\gamma(T)$ as

$$\gamma(T) = 1 + 2\sum_{\mathbf{k}\neq\mathbf{k}'} |t_{\mathbf{k}}||t_{\mathbf{k}'}| = 2\left(\sum_{\mathbf{k}} |t_{\mathbf{k}}|\right)^2 - 1\,. \tag{28}$$

If $T = U^{\otimes n} \otimes V^{\otimes m} \otimes W^{\otimes p}$, then $|t_{\mathbf{k}}|$ is given by

$$|t_{\mathbf{k}}| = \Big(\prod_{r=1}^{n} |u_{i_r}|\Big)\Big(\prod_{r=1}^{m} |v_{j_r}|\Big)\Big(\prod_{r=1}^{p} |w_{l_r}|\Big)\,. \tag{29}$$

We then calculate

$$\sum_{\mathbf{k}} |t_{\mathbf{k}}| = \sum_{i_1,\ldots,i_n,j_1,\ldots,j_m,l_1,\ldots,l_p=0}^{3} |u_{i_1}|\cdots|u_{i_n}|\cdot|v_{j_1}|\cdots|v_{j_m}|\cdot|w_{l_1}|\cdots|w_{l_p}| \tag{30}$$

$$= \left(\sum_{i_1=0}^{3}|u_{i_1}|\right)\cdots\left(\sum_{i_n=0}^{3}|u_{i_n}|\right)\left(\sum_{j_1=0}^{3}|v_{j_1}|\right)\cdots\left(\sum_{j_m=0}^{3}|v_{j_m}|\right)\left(\sum_{l_1=0}^{3}|w_{l_1}|\right)\cdots\left(\sum_{l_p=0}^{3}|w_{l_p}|\right) \tag{31}$$

$$= \left(\sum_{i=0}^{3}|u_i|\right)^n\left(\sum_{j=0}^{3}|v_j|\right)^m\left(\sum_{l=0}^{3}|w_l|\right)^p, \tag{32}$$

which concludes the proof. □

Example 4.3. Let us apply Corollary 4.2 to some interesting examples:

(i) $\gamma(\mathrm{CNOT}^{\otimes n}) = 2^{n+1}-1$, since $u_{\mathrm{CNOT}} = (\frac{1}{\sqrt{2}}, \frac{\mathrm{i}}{\sqrt{2}}, 0, 0)$ (see Example 3.2). This is consistent with the LOCC result from [3].

(ii) $\gamma(\mathrm{SWAP}^{\otimes n}) = 2\times 4^n - 1$, since $u_{\mathrm{SWAP}} = (\frac{\sqrt{2}(1+\mathrm{i})}{4}, \frac{\sqrt{2}(1+\mathrm{i})}{4}, \frac{\sqrt{2}(1+\mathrm{i})}{4}, \frac{\sqrt{2}(1+\mathrm{i})}{4})$ (see Example 3.2). This is consistent with the LOCC result from [3].

(iii) $\gamma(\mathrm{CR}_X(\theta)^{\otimes n}) = 2(|\cos(\frac{\theta}{4})| + |\sin(\frac{\theta}{4})|)^{2n} - 1$, since $u_{\mathrm{CR}_X(\theta)} = (\cos(\frac{\theta}{4}), \mathrm{i}\sin(\frac{\theta}{4}), 0, 0)$.[3] This answers an open question from [3].

Example 3.2 and Example 4.3 show that n parallel cuts can be considerably cheaper than cutting n-times a gate separately.

Corollary 4.4. *In general, Corollary 4.2 implies that for any two-qubit gate with $\gamma(U) > 1$ the optimal sampling overhead is strictly submultiplicative under the tensor product, i.e.*

$$\gamma(U^{\otimes n}) < \gamma(U)^n\,. \tag{33}$$

Proof. Introducing $x = (\sum_i |u_i|)^2$, Corollary 4.2 is equivalent to

$$\gamma(U^{\otimes n}) = 2x^n - 1 < (2x-1)^n = \gamma(U)^n\,. \tag{34}$$

Since $\gamma(U) > 1$, we have $x > 1$. The inequality then follows by induction

$$\begin{aligned}(2x-1)^{n+1} &= (2x-1)^n(2x-1) &(35)\\ &> (2x^n-1)(2x-1) &(36)\\ &= 4x^{n+1} - 2x - 2x^n + 1 &(37)\\ &= 2x^{n+1} - 1 + 2x^{n+1} - 2x - 2x^n + 2 &(38)\\ &= \gamma(U^{\otimes(n+1)}) + 2x^n(x-1) - 2(x-1) &(39)\\ &= \gamma(U^{\otimes(n+1)}) + 2(x^n-1)(x-1) &(40)\\ &> \gamma(U^{\otimes(n+1)})\,. &(41)\end{aligned}$$

□

[3] $\mathrm{CR}_X(\theta)$ denotes a controlled $\mathrm{R}_X(\theta)$ gate with rotation angle θ, where $\mathrm{R}_X(\theta) := \mathrm{e}^{-\mathrm{i}\frac{\theta}{2}X}$.

In the following, we want to consider the asymptotic difference between the single-cut and parallel cut settings. We therefore introduce the following quantity:

Definition 4.5 (Regularized γ-factor)**.** The regularized γ-factor is defined as

$$\gamma_R(U) = \lim_{n\to\infty} (\gamma(U^{\otimes n}))^{\frac{1}{n}} \,. \tag{42}$$

$\gamma_R(U)$ can be considered as the "effective γ-factor" of a single gate in the parallel cut setting, in the limit of infinite gates.

Corollary 4.6. *Using Corollary 4.2, $\gamma_R(U)$ evaluates to*

$$\gamma_R(U) = \lim_{n\to\infty} \left(2(\sum_i |u_i|)^{2n} - 1\right)^{\frac{1}{n}} = \left(\sum_i |u_i|\right)^2 = 1 + \sum_{i\neq j} |u_i||u_j| \,. \tag{43}$$

5 Cutting KAK-like unitaries

Previous chapters relied on the KAK-decomposition of two-qubit unitaries. As discussed in Section 2.2, larger unitaries can in general not be represented in such a form. However, one can still consider the class of unitaries that are in a similar form to the two-qubit KAK decomposition. For such unitaries, we formulate the following theorem:

Theorem 5.1 (Cutting KAK-like unitaries)**.** *Let U_{AB} be a unitary of the form*

$$U_{AB} = \left(V_A^{(1)} \otimes V_B^{(2)}\right) \sum_k u_k (L_k)_A \otimes (R_k)_B \left(V_A^{(3)} \otimes V_B^{(4)}\right) \,, \tag{44}$$

where $V_A^{(1)}, V_B^{(2)}, V_A^{(3)}, V_B^{(4)}, (L_k)_A$ and $(R_k)_B$ are unitaries on A and B.

Then it holds that

$$\gamma_{\mathrm{LO}}(U) \leq \sum_k |u_k|^2 + 2 \sum_{k\neq k'} |u_k||u_{k'}| \,. \tag{45}$$

Furthermore, if $(L_k)_A$ and $(R_k)_B$ are orthogonal [4] for different k, then Equation (45) simplifies, holds with equality and we find $\gamma_{\mathrm{LO}}(U) = \gamma_{\mathrm{LOCC}}(U)$:

$$\gamma_{\mathrm{LO}}(U) = 1 + 2 \sum_{k\neq k'} |u_k||u_{k'}| \,. \tag{46}$$

The assertion of Theorem 5.1 follows from the following two lemmas, since $\gamma_{\mathrm{LOCC}}(U) \leq \gamma_{\mathrm{LO}}(U)$.

Lemma 5.2. *Let U_{AB} be of the same form as in Theorem 5.1. It then holds that*

$$\gamma_{\mathrm{LO}}(U) \leq \sum_k |u_k|^2 + 2 \sum_{i\neq j} |u_i||u_j| \,. \tag{47}$$

Proof. We prove this statement by explicitly constructing a quasiprobability decomposition of U into LO which exhibits a one-norm of the quasiprobability coefficients given by $\sum_k |u_k|^2 + 2\sum_{i\neq j} |u_i||u_j|$. This proof is an extension of the decomposition basis derived in [1], which is optimal for some two-qubit unitaries but not for all of them. As explained

[4] Orthogonal with respect to the Hilbert-Schmidt inner product $\langle A, B\rangle = \mathrm{tr}(A^\dagger B)$

in the introduction Section 2.2, we can directly consider the interaction part W of the unitary U (up to local unitaries U and W are equal). The induced channel of W is given by

$$\mathcal{W}(\rho_{AB}) = \sum_{k,k'} u_k u^*_{k'} (L_k)_A \otimes (R_k)_B \rho_{AB} (L_{k'})^\dagger_A \otimes (R_{k'})^\dagger_B \,. \tag{48}$$

By writing the complex factors by its argument and absolute value $u_k = |u_k| \mathrm{e}^{\mathrm{i}\phi_k}$, we get

$$\begin{aligned}\mathcal{W}(\rho) = &\sum_k |u_k|^2 (L_k)_A \otimes (R_k)_B \rho_{AB} (L_k)^\dagger_A \otimes (R_k)^\dagger_B \\ &+ \sum_{k<k'} |u_k||u_{k'}| (\mathrm{e}^{\mathrm{i}\phi_{k,k'}} (L_k)_A \otimes (R_k)_B \rho_{AB} (L_{k'})^\dagger_A \otimes (R_{k'})^\dagger_B \\ &\quad + \mathrm{e}^{-\mathrm{i}\phi_{k,k'}} (L_{k'})_A \otimes (R_{k'})_B \rho_{AB} (L_k)^\dagger_A \otimes (R_k)^\dagger_B) \,,\end{aligned} \tag{49}$$

where $\phi_{k,k'} = \phi_k - \phi_{k'}$[5]. Implementing the first sum with LO is trivial, and the difficulty will be the implementation of the term:

$$\mathrm{e}^{\mathrm{i}\phi_{k,k'}} (L_k)_A \otimes (R_k)_B \rho_{AB} (L_{k'})^\dagger_A \otimes (R_{k'})^\dagger_B + \mathrm{e}^{-\mathrm{i}\phi_{k,k'}} (L_{k'})_A \otimes (R_{k'})_B \rho_{AB} (L_k)^\dagger_A \otimes (R_k)^\dagger_B \tag{50}$$

To do so, we define the expressions

$$(l^\phi_{k,k'})_\pm = \frac{L_k \pm \mathrm{e}^{-\mathrm{i}\phi} L_{k'}}{2} \qquad \text{and} \qquad (r^\phi_{k,k'})_\pm = \frac{R_k \pm \mathrm{e}^{-\mathrm{i}\phi} R_{k'}}{2} \,, \tag{51}$$

and based on them the maps

$$(\mathcal{C}^\phi_{k,k'})_\pm(\rho) = (l^\phi_{k,k'})_\pm \, \rho \, (l^\phi_{k,k'})^\dagger_\pm \qquad \text{and} \qquad (\mathcal{D}^\phi_{k,k'})_\pm(\rho) = (r^\phi_{k,k'})_\pm \, \rho \, (r^\phi_{k,k'})^\dagger_\pm \,. \tag{52}$$

Under usage of the *negativity* trick, we then define the channels

$$\mathcal{C}^\phi_{k,k'}(\rho) = (\mathcal{C}^\phi_{k,k'})_+(\rho) - (\mathcal{C}^\phi_{k,k'})_-(\rho) \quad \text{and} \quad \mathcal{D}^\phi_{k,k'}(\rho) = (\mathcal{D}^\phi_{k,k'})_+(\rho) - (\mathcal{D}^\phi_{k,k'})_-(\rho) \,, \tag{53}$$

which evaluate to

$$\mathcal{C}^\phi_{k,k'}(\rho) = \frac{1}{2}(\mathrm{e}^{\mathrm{i}\phi} L_k \rho L^\dagger_{k'} + \mathrm{e}^{-\mathrm{i}\phi} L_{k'} \rho L^\dagger_k) \quad \text{and} \quad \mathcal{D}^\phi_{k,k'}(\rho) = \frac{1}{2}(\mathrm{e}^{\mathrm{i}\phi} R_k \rho R^\dagger_{k'} + \mathrm{e}^{-\mathrm{i}\phi} R_{k'} \rho R^\dagger_k) \,. \tag{54}$$

This expression already resembles Equation (50), but is so far only a local channel.[6] In order to obtain Equation (50), we apply these channels in parallel. This gives us

$$\begin{aligned}(\mathcal{C}^\phi_{k,k'} \otimes \mathcal{D}^\phi_{k,k'})(\rho) = &\frac{1}{4}\mathrm{e}^{2\mathrm{i}\phi}(L_k \otimes R_k)\rho(L^\dagger_{k'} \otimes R^\dagger_{k'}) + \frac{1}{4}\mathrm{e}^{-2\mathrm{i}\phi}(L_{k'} \otimes R_{k'})\rho(L^\dagger_k \otimes R^\dagger_k) \\ &+ \frac{1}{4}(L_k \otimes R_{k'})\rho(L^\dagger_{k'} \otimes R^\dagger_k) + \frac{1}{4}(L_{k'} \otimes R_k)\rho(L^\dagger_k \otimes R^\dagger_{k'}) \,,\end{aligned} \tag{55}$$

which resembles Equation (50) even more, but we still have mixed terms we want to get rid of. We do so by considering the operation $\mathcal{C}^{\phi+\frac{\pi}{2}}_{k,k'}$ and $\mathcal{D}^{\phi+\frac{\pi}{2}}_{k,k'}$. This introduces a global and a relative phase shift. For instance for $\mathcal{C}^{\phi+\frac{\pi}{2}}_{k,k'}$, we get

$$\mathcal{C}^{\phi+\frac{\pi}{2}}_{k,k'}(\rho) = \frac{1}{2}\mathrm{e}^{\mathrm{i}\frac{\pi}{2}}(L_k \rho L^\dagger_{k'} \mathrm{e}^{\mathrm{i}\phi} + \mathrm{e}^{-i\pi} L_{k'} \rho L^\dagger_k \mathrm{e}^{-\mathrm{i}\phi}) = \frac{1}{2}\mathrm{e}^{\mathrm{i}\frac{\pi}{2}}(L_k \rho L^\dagger_{k'} \mathrm{e}^{\mathrm{i}\phi} - L_{k'} \rho L^\dagger_k \mathrm{e}^{-\mathrm{i}\phi}) \,, \tag{56}$$

[5] Note that in contrast to [1], we describe $e^{i\phi_{k,k'}} L_k \otimes R_k \rho L_{k'} \otimes R_{k'} + e^{-i\phi_{k,k'}} L_{k'} \otimes R_{k'} \rho L_k \otimes R_k$ with local operations instead of $\sigma_k \otimes \sigma_k \rho \sigma_{k'} \otimes \sigma_{k'}$. This will allow us to consider the phase of the u_k to find the optimal decomposition.

[6] One can check that $(\mathcal{C}^\phi_{k,k'})_\pm, (\mathcal{D}^\phi_{k,k'})_\pm \succeq 0$ and $(\mathcal{C}^\phi_{k,k'})_+ + (\mathcal{C}^\phi_{k,k'})_-, (\mathcal{D}^\phi_{k,k'})_+ + (\mathcal{D}^\phi_{k,k'})_- \in \mathrm{CPTP}$ are fulfilled.

which extended to both parties yields

$$(\mathcal{C}_{k,k'}^{\phi+\frac{\pi}{2}} \otimes \mathcal{D}_{k,k'}^{\phi+\frac{\pi}{2}})(\rho) = -\frac{1}{4}\mathrm{e}^{2\mathrm{i}\phi}(L_k \otimes R_k)\rho(L_{k'}^\dagger \otimes R_{k'}^\dagger) - \frac{1}{4}\mathrm{e}^{-2\mathrm{i}\phi}(L_{k'} \otimes R_{k'})\rho(L_k^\dagger \otimes R_k^\dagger)$$
$$+ \frac{1}{4}(L_k \otimes R_{k'})\rho(L_{k'}^\dagger \otimes R_k^\dagger) + \frac{1}{4}(L_{k'} \otimes R_k)\rho(L_k^\dagger \otimes R_{k'}^\dagger)\,. \tag{57}$$

This allows us to formulate the desired term as

$$2\left(\mathcal{C}_{k,k'}^{\frac{\phi_{k,k'}}{2}} \otimes \mathcal{D}_{k,k'}^{\frac{\phi_{k,k'}}{2}} - \mathcal{C}_{k,k'}^{\frac{\phi_{k,k'}}{2}+\frac{\pi}{2}} \otimes \mathcal{D}_{k,k'}^{\frac{\phi_{k,k'}}{2}+\frac{\pi}{2}}\right)$$
$$= \mathrm{e}^{\mathrm{i}\phi_{k,k'}}(L_k \otimes R_k)\rho(L_{k'}^\dagger \otimes R_{k'}^\dagger) + \mathrm{e}^{-\mathrm{i}\phi_{k,k'}}(L_{k'} \otimes R_{k'})\rho(L_k^\dagger \otimes R_k^\dagger)\,. \tag{58}$$

Combining this with Equation (49) gives

$$\mathcal{W}(\rho) = \sum_k |u_k|^2 (L_k)_A \otimes (R_k)_B \rho_{AB} (L_k)_A^\dagger \otimes (R_k)_B^\dagger$$
$$+ \sum_{k<k'} |u_k||u_{k'}|\, 2\left(\mathcal{C}_{k,k'}^{\frac{\phi_{k,k'}}{2}} \otimes \mathcal{D}_{k,k'}^{\frac{\phi_{k,k'}}{2}} - \mathcal{C}_{k,k'}^{\frac{\phi_{k,k'}}{2}+\frac{\pi}{2}} \otimes \mathcal{D}_{k,k'}^{\frac{\phi_{k,k'}}{2}+\frac{\pi}{2}}\right)\,, \tag{59}$$

which implies

$$\gamma_{\mathrm{LO}}(U_{AB}) \leq \sum_k |u_k|^2 + 4\sum_{k<k'} |u_k||u_{k'}| = \sum_k |u_k|^2 + 2\sum_{k\neq k'} |u_k||u_{k'}|\,. \tag{60}$$

□

This construction that handles the interference terms is also known as Hadamard test and is used with a similar aim in [15]. It differs however from our manuscript, in that it decomposes the unitary in the Pauli basis which is not necessarily the best decomposition. Therefore it only provides an upper bound on the optimal simulation overhead, which is not tight for most instances.

For convenience, we present the circuits that implement the channels $(\mathcal{C}_{k,k'}^{\phi})_\pm(\rho)$ and $(\mathcal{D}_{k,k'}^{\phi})_\pm(\rho)$ in Appendix A and explain how to apply them.

Lemma 5.3. *Let U_{AB} be of the form described in Theorem 5.1 and furthermore require orthogonality of L_k and R_k in the following sense:*

$$\mathrm{tr}(L_k^\dagger L_{k'}) = \dim(A)\delta_{k,k'} \qquad \text{and} \qquad \mathrm{tr}(R_k^\dagger R_{k'}) = \dim(B)\delta_{k,k'}\,. \tag{61}$$

Then

$$\gamma_{\mathrm{LOCC}}(U) \geq 1 + 2\sum_{i\neq j} |u_k||u_{k'}|\,. \tag{62}$$

Proof. Let $|\Phi_U\rangle$ denote the Choi state of a unitary U_{AB}.[7] Simulating the channel U_{AB} is at least as expensive as preparing a state that can be prepared using the channel — in particular the Choi state. We therefore have

$$\gamma_{\mathrm{LOCC}}(U_{AB}) = \gamma_{\mathrm{LOCC}}(W_{AB}) \geq \gamma_{\mathrm{LOCC}}(|\Phi_W\rangle)\,, \tag{63}$$

[7] We define the Choi state of U_A as $|\Phi_U\rangle = \mathrm{id}_{A'} \otimes U_A|\psi\rangle_{A'A}$, where A' is a copy of the Hilbert space A and $|\psi\rangle_{A'A}$ is the maximally entangled state on the joint system $A'A$.

where we introduced the γ-factor for states and W_{AB} denotes the interaction unitary. Furthermore, the Choi state of W_{AB} is denoted by $|\Phi_W\rangle$. The γ-factor for states is defined by

$$\gamma_S(\rho_{AB}) \coloneqq \min\left\{\gamma_S(\mathcal{E}) | \mathcal{E}(|0\rangle\langle 0|_{AB}) = \rho_{AB}\right\}, \tag{64}$$

but as was shown in [3, Lemma 4.1], for pure states it can be written as a function of the Schmidt coefficients. One finds

$$\gamma_{\mathrm{LOCC}}(|\Phi_W\rangle) = 2\left(\sum_i \alpha_i\right)^2 - 1\,, \tag{65}$$

where the α_i are the Schmidt coefficients of $|\Phi_W\rangle$. We can calculate these concretely in terms of u_i. The Choi state is then given as

$$\begin{aligned}
|\Phi_W\rangle_{A'B'AB} &= \frac{1}{\sqrt{\dim(A)\dim(B)}} \sum_{i=1}^{\dim(A)} \sum_{j=1}^{\dim(B)} |i,j\rangle_{A'B'} \otimes W|i,j\rangle_{AB} & (66)\\
&= \frac{1}{\sqrt{\dim(A)\dim(B)}} \sum_{i=1}^{\dim(A)} \sum_{j=1}^{\dim(B)} \sum_k u_k |i,j\rangle_{A'B'} \otimes (L_k \otimes R_k)|i,j\rangle_{AB} & (67)\\
&= \sum_k |u_k| |\varphi_k^{(1)}\rangle_{A'A} \otimes |\varphi_k^{(2)}\rangle_{B'B}\,, & (68)
\end{aligned}$$

where $|\varphi_k\rangle_{A'A}$ and $|\varphi_k\rangle_{B'B}$ are orthonormal vectors given by

$$\begin{aligned}
|\varphi_k^{(1)}\rangle_{A'A} &= \frac{1}{\sqrt{\dim(A)}} \sum_{j=1}^{\dim(A)} |j\rangle_{A'} L_k |j\rangle_A \qquad \text{and} & (69)\\
|\varphi_k^{(2)}\rangle_{B'B} &= \frac{1}{\sqrt{\dim(B)}} \arg(u_k) \sum_{j=1}^{\dim(B)} |j\rangle_{B'} R_k |j\rangle_B\,. & (70)
\end{aligned}$$

Orthonormality follows from the above trace properties of L_k and R_k. In this form, the Choi state is in its Schmidt decomposition and we can identify the Schmidt coefficients α_i as $|u_i|$. Using Equation (63) and an analogous calculation to the proof of Corollary 3.3, we get

$$\gamma_{\mathrm{LOCC}} \geq 2\left(\sum_{i=0}^{3} |u_i|\right)^2 - 1 = 1 + 2\sum_{i\neq j} |u_i||u_j|\,, \tag{71}$$

which concludes the proof. □

As an additional application of Theorem 5.1 and our complete understanding of the parallel cut setting, we show that one can leverage this theorem to explicitly find decompositions of unitaries that consist of multiple two-qubit unitaries interleaved by arbitrary, but known[8] operations as depicted in Figure 3. In general, we will not be able to give the optimal decomposition for these scenarios as this would rely on our ability to cut arbitrary unitaries, however we can still provide a (not necessarily optimal) decomposition with an overhead that is given by the optimal decomposition of the parallel cut. This decomposition will depend on the interleaved operations.

[8]Note that we are not referring to the black box setting here.

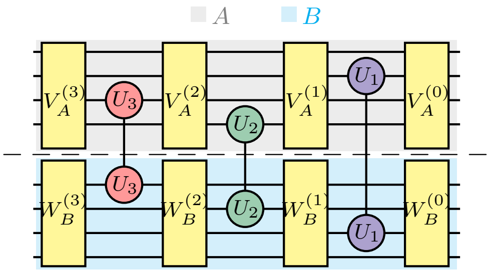


Figure 3: Cut of multiple two-qubit unitaries (here denoted by U_1, U_2, U_3) with interleaved operations (denoted by $V_A^{(i)}, W_B^{(j)}$).

Corollary 5.4 (Parallel cut as upper bound)**.** *Let U be a unitary of the form*

$$U = (V_A^{(0)} \otimes W_B^{(0)}) \prod_{i=1}^{n} \left(U_{A_i B_i}^{(i)} (V_A^{(i)} \otimes W_B^{(i)}) \right) , \tag{72}$$

where $V_A^{(i)}$ and $W_B^{(i)}$ are arbitrary unitaries on the subsystems A and B, respectively and $U_{A_i B_i}$ acts as a two-qubit unitary between A and B. We then have

$$\gamma_{\mathrm{LO}}(U) \leq \gamma(\bigotimes_{i=1}^{n} U_i) \,. \tag{73}$$

Since cutting the gates U_i jointly is strictly cheaper than cutting them individually, cutting the whole circuit U is also cheaper than cutting the U_i individually.

Proof. We show the upper bound by constructing an explicit decomposition with overhead equal to the parallel cut scenario. Take a unitary U of the form of Corollary 5.4 with two-qubit unitaries $U_{A_i B_i}^{(i)}$. To simplify the notation, we will assume that $U_{A_i B_i}^{(i)}$ acts on the i^{th} qubit of A and B. We then perform a KAK decomposition of all $U_{A_i B_i}^{(i)}$, i.e.

$$U_{A_i B_i}^{(i)} = \left(V_{A_i}^{(1,i)} \otimes W_{B_i}^{(2,i)} \right) \left(\sum_{k=0}^{3} u_k^{(i)} (\sigma_k)_{A_i} \otimes (\sigma_k)_{B_i} \right) \left(V_{A_i}^{(3,i)} \otimes W_{B_i}^{(4,i)} \right) , \tag{74}$$

and absorb the local one-qubit unitaries $V_{A_i}^{(1,i)}, V_{A_i}^{(2,i)}, W_{B_i}^{(1,i)}, W_{B_i}^{(2,i)}$, into the unitaries $V_A^{(i-1)}, W_B^{(i-1)}$ and $V_A^{(i)}, W_B^{(i)}$. This yields

$$\begin{aligned} U &= \sum_{k_1,\ldots,k_n=0}^{3} (V_A^{(0)} \otimes W_B^{(0)}) \prod_{i=1}^{n} u_{k_i}^{(i)} \left((\sigma_{k_i})_{A_i} \otimes (\sigma_{k_i})_{B_i} (V_A^{(i)} \otimes W_B^{(i)}) \right) && (75) \\ &= \sum_{k_1,\ldots,k_n=0}^{3} \prod_{i=0}^{n} u_{k_i}^{(i)} (L_{k_1,\ldots,k_n})_A \otimes (R_{k_1,\ldots,k_n})_B && (76) \\ &= \sum_{\mathbf{k}} u_{\mathbf{k}} \, (L_{\mathbf{k}})_A \otimes (R_{\mathbf{k}})_B \,, && (77) \end{aligned}$$

where we introduced $u_{\mathbf{k}} = \prod_{i=0}^{n} u_{k_i}^{(i)}$ and the unitaries

$$(L_{\mathbf{k}})_A = (L_{k_1,\ldots,k_n})_A = V_A^{(0)} \prod_{i=1}^{n} (\sigma_{k_i})_{A_i} V_A^{(i)} \tag{78}$$

$$(R_{\mathbf{k}})_B = (R_{k_1,\dots,k_n})_B = W_B^{(0)} \prod_{i=1}^{n} (\sigma_{k_i})_{B_i} V_B^{(i)} \,. \tag{79}$$

In this form, we see that the first part of Theorem 5.1 applies. Since $V_A^{(i)}$ and $V_B^{(i)}$ are arbitrary, the unitaries $L_{\mathbf{k}}$ and $R_{\mathbf{k}}$ are in general not orthogonal and therefore optimality is not necessarily achieved. However, since $u_{\mathbf{k}}$ is a product of KAK coefficients, we still find $\sum_{\mathbf{k}} |u_{\mathbf{k}}|^2 = 1$. □

Remark 5.5. The following comments concerning Corollary 5.4 are worth being mentioned.

(i) Corollary 5.4 shows that performing local operations between two-qubit gates only reduces the overhead of the unitary compared to the parallel cut case ($V_A^{(i)} = W_B^{(i)} = \mathbb{1}$ for all i). Intuitively, we can understand this since the case that all two-qubit gates act on different qubits is already the case where we create the most entanglement between both partitions A and B. Local unitaries cannot increase this entanglement. However, they can reduce it. Consider the case where $V_A^{(i)}, W_B^{(i)}$ are SWAP gates that make all two-qubit unitaries unitaries between A and B effectively act on the same two qubits. In this case the overall circuit is equivalent to a single two-qubit unitary and cutting it costs a maximum of 7 and therefore considerably less than cutting all two-qubit gates simultaneously.

(ii) The channel $(\mathcal{C}^{\phi}_{\mathbf{k},\mathbf{k}'} \otimes \mathcal{D}^{\phi}_{\mathbf{k},\mathbf{k}'})(\rho)$ depends on the interleaved operations $V_A^{(i)}, W_B^{(i)}$. If the gates $V_A^{(i)}, W_B^{(i)}$ are simple, the calculation of the channel decomposition is feasible. However, if the interleaved gates are complicated, e.g. in the situation where one two-qubit gate at the beginning and one at the end of a long circuit are to be cut simultaneously, finding the LO decomposition might be difficult.

6 Cutting in the black box setting

As discussed in Remark 5.5 (Item ii), evaluating the optimal decomposition of a unitary consisting of two-qubit gates interleaved with arbitrary unitaries might be difficult if the evaluation of the interleaved elements is complicated. Therefore this section will focus on the construction of a decomposition that is independent of the form of the interleaved element. This is what we call the black box cut as shown in Figure 1c. We believe that this scenario is also the most useful in practice as it allows us to jointly cut multiple two-qubit gates, which reduces the sampling overhead, without making any assumptions on where in the circuit these gates may lie.

Before we present the theorem that covers this case, let us define the minimal overhead in the black box setting. For simplicity, we will first restrict to the case of two two-qubit gates, but the case with arbitrary number of gates follows analogously.

Definition 6.1 (Black box γ-factor)**.** We define the minimal overhead of two channels that are separated by an arbitrary black box channel using operations $S(A, B) = \{\mathrm{LO}(A, B), \mathrm{LOCC}(A, B)\}$ as

$$\overline{\gamma}_S(\mathcal{E}^1_{AB}, \mathcal{E}^2_{AB}) \coloneqq \min \Big\{ \sum_{i=1}^{m} |a_i| : \mathcal{E}^1_{AB} X_{AB} \mathcal{E}^2_{AB} = \mathrm{tr}_{E_A E_B} \sum_{i=1}^{m} a_i \mathcal{F}_i X_{AB} \mathcal{G}_i, \\ \forall X_{AB} \in \mathrm{CPTP},\ m \geq 1,\ \mathcal{F}_i, \mathcal{G}_i \in S(AE_A, BE_B) \text{ and } a_i \in \mathbb{R} \Big\}. \tag{80}$$

Here, we allow the introduction of an ancillary system E_A and E_B on which the black box X is not acting. This allows to correlate channels before and after the black box computation.

We can then formulate our theorem.

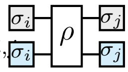
Theorem 6.2. *Two two-qubit unitaries U_{A_1,B_1} and V_{A_2,B_2} have a decomposition independent of a black box channel, with optimal black box overhead*

$$\overline{\gamma}_{\mathrm{LOCC}}(U_{A_1,B_1}, V_{A_2,B_2}) = \overline{\gamma}_{\mathrm{LO}}(U_{A_1,B_1}, V_{A_2,B_2}) = 1 + 2 \sum_{\mathbf{k} \neq \mathbf{k}'} |w_{\mathbf{k}}||w_{\mathbf{k}'}|\,, \tag{81}$$

where $\mathbf{k} \in \{0,1,2,3\}^2$ and $w_{\mathbf{k}} = u_{k_1} v_{k_2}$ is the product of the KAK-coefficients of U_{A_1,B_1} and V_{A_2,B_2}.

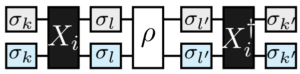
To prevent expression from becoming too convoluted, we will introduce a graphical notations of tensor products for the following section. In this notation we always consider channels which act from both sides on a density matrix. These expression have to be read from the inside, usually ρ, to the outside. The vertical dimension denotes that operations act on different subsystems.

$$\textstyle\sum_{i,j} (\sigma_i)_A \otimes (\sigma_i)_B \rho_{AB} (\sigma_j)_A \otimes (\sigma_j)_B \overset{\text{notation}}{=} \sum_{i,j} \tag{82}$$



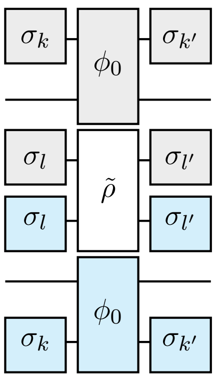
Proof. The essential ingredient for achieving submultiplicativity is to correlate the channels before and after the black box. We achieve this, by reducing the problem to the parallel cut and by making use of ancillary qubits. These are used to perform Bell measurements similar to the gate teleportation protocol. We will illustrate the procedure using the above introduced notation. The overall channel we want to implement is given by

$$\mathcal{E}(\rho) = \mathcal{U}(\mathcal{X}(\mathcal{V}(\rho))) = \textstyle\sum_{k,k',i,l,l'} u_k u_{k'}^* v_l v_{l'} \tag{83}$$

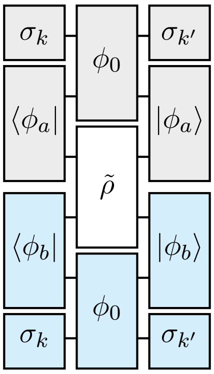


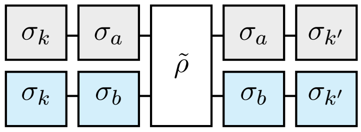
where we introduce Kraus operators X_i for the black box channel $\mathcal{X}$. To correlate $\mathcal{U}$ and $\mathcal{V}$, we introduce additional qubits in the Bell-state $|\phi_0\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$, ($\phi_0 = |\phi_0\rangle\langle\phi_0|$) for both parties. Then we perform the channel we used for parallel two-qubit gates (cf. Corollary 4.1) on system and ancillary qubits.

$$\mathcal{G}(\rho) = \textstyle\sum_{k,l,k',l'} u_k u_{k'}^* v_l v_{l'}^* \tag{84}$$



Applying this channel comes at a sampling overhead of $1 + 2\sum_{\mathbf{k}\neq\mathbf{k}'} |w_{\mathbf{k}}||w_{\mathbf{k}'}|$ and requires only local operations. After that, we can apply any black box channel $\mathcal{X}$ as long as it does not act on the ancillary qubits. We will call the so obtained state $\tilde{\rho}$. In order to teleport the gate from the quasi Choi-state, we perform a measurement in the Bell-basis between the unused part of the ancillary qubit and the qubit on which we want the second unitary to act. We will illustrate the measurement in the Bell-basis as $\langle\phi_a|$, where the index a corresponds to the measurement outcome. The post-measurement state corresponding to the measurement outcomes a and b is then given by

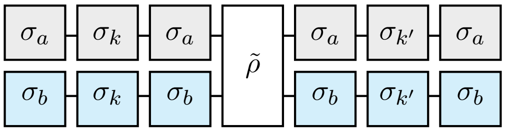
$\sum_{k,k'} u_k u_{k'}^*$ 

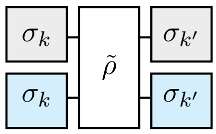
$=$ $\sum_{k,k'} u_k u_{k'}^*$ 

(85)

where the Pauli operations σ_a and σ_b appear as a consequence of projecting into the Bell basis. In the normal gate-teleportation scenario, one can only recover from these Pauli gates, if the teleported gate was a Clifford gate. Even then, one needs to classical communicate the local measurement of a or b to the other side to perform the recovery operation $U(\sigma_a \otimes \sigma_b)U^\dagger$. This operation is only local for Clifford gates. The quasiprobability simulation setting is however different. Here, we are not actually performing the non-local unitary, but rather we just sample from local channels. This gives us additional information to work with.

We recover the original channel by first applying a local operation σ_x based on the measurement result x of each party. This modifies the post-measurement state to

$\sum_{k,k'} u_k u_{k'}^*$ 

$= \sum_{k,k'} u_k u_{k'}^* f(a,k) f(a,k') f(b,k) f(b,k')$ 

(86)

with functions $f(i,j)$ that are either $+1$ or -1. The sign follows from the the anti-commutation relations between the Pauli matrices $\sigma_a, \sigma_k, \sigma_b$ and $\sigma_{k'}$. Since, we are doing a quasiprobability simulation, we are not performing the whole sum over k, k' and only apply a specific channel that depends on k, k'. We either have the case $k = k'$, but then the recovery is trivial or the case $k \neq k'$. In this case, we have knowledge of k, k' and a, b and can in the post processing weight the result with a factor of $f(a,k)f(a,k')f(b,k)f(b,k')$, which is either $+1$ or -1. Such a weighting does not affect the γ-factor, but gives the right estimator for the expectation value. Since all used channels only require local operations, we present a concrete construction with $\overline{\gamma}_{\mathrm{LOCC}}(U_{A_1,B_1}, V_{A_2,B_2}) \leq \overline{\gamma}_{\mathrm{LO}}(U_{A_1,B_1}, V_{A_2,B_2}) \leq 1 + 2\sum_{\mathbf{k}\neq\mathbf{k}'} |w_{\mathbf{k}}||w_{\mathbf{k}'}|$.

This construction works independently of the black box channel and has the same overhead as the parallel cut (the black box being the identity). It this therefore optimal according to our definition Definition 6.1. We show the procedure as a quantum circuit in Figure 4. □

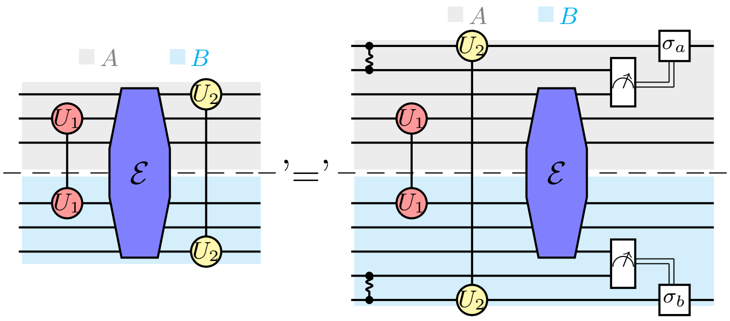


Figure 4: Protocol of the black box cut. Both parties A and B apply a Pauli operation σ_x based on their Bell-measurement outcome.

In the scenario of normal gate teleportation, it is critical that we have to send classical information to the other party. Otherwise we would be able to signal faster than light. In the QPD setting, this is not an issue, since each individual channel cannot signal.

6.1 Cutting in the black box setting for N two-qubit gates

For completeness, we will present how to generalize the previous theorem to n gates. For this, we first define the black box overhead for multiple channels.

Definition 6.3 (Black box γ-factor)**.** The minimal black box overhead of channels $\{\mathcal{E}_{AB}^{(j)}\}_j$ that are separated by arbitrary black box channels $\{X_{AB}^{(j)}\}_j$ is given by

$$\overline{\gamma}_S(\{\mathcal{E}_{AB}^j\}_j) \coloneqq \min \Big\{ \sum_{i=1}^{m} |a_i| : \prod_j X_{AB}^{(j)} \mathcal{E}_{AB}^{(j)} = \mathrm{tr}_{E_A E_B} \sum_{i=1}^{m} a_i (\prod_j X_{AB}^{(j)} \mathcal{F}_i^{(j)}), \\ \forall X_{AB}^{(j)} \in \mathrm{CPTP},\, m \geq 1,\, \mathcal{F}_i^{(j)} \in S(AE_A, BE_B) \text{ and } a_i \in \mathbb{R} \Big\}, \tag{87}$$

with $S(AE_A, BE_B) = \{\mathrm{LO}(AE_A, BE_B), \mathrm{LOCC}(AE_A, BE_B)\}$.

Now we extend the previous theorem.

Theorem 6.4. *N two-qubit unitaries $\{U_{A_i,B_i}^{(i)}\}_{i=1}^N$ have a decomposition independent of interleaving black box channel $X_{AB}^{(i)}$, with optimal black box overhead*

$$\overline{\gamma}_{\mathrm{LOCC}}(\{U_{A_i,B_i}^{(i)}\}_i^N) = \overline{\gamma}_{\mathrm{LO}}(\{U_{A_i,B_i}^{(i)}\}_i^N) = 1 + 2 \sum_{\mathbf{k} \neq \mathbf{k}'} |w_{\mathbf{k}}||w_{\mathbf{k}'}|\,, \tag{88}$$

where $\mathbf{k} \in \{0,1,2,3\}^N$ and $w_{\mathbf{k}} = \prod_{i=1}^N u_{k_i}^{(i)}$ is the product of the KAK-coefficients of $U_{A_i,B_i}^{(i)}$.

Note that Theorem 6.2 and Theorem 6.4 have the same simulation overhead as Corollary 4.1. Arbitrary cuts are therefore, apart from a requirement for ancillary qubits, not more expensive than parallel cuts.

Proof. We proof the assertion by explicitly constructing operations $\mathcal{F}_i^{(j)} \in \mathrm{LO}(AE_A, BE_B)$. Optimality then follows from the black boxes being identity maps and Corollary 4.1. The procedure is analogous to the two gate case. First we implement the channel used for the joint parallel cut of all n two-qubit gates (cf. Section 4) via $\mathcal{F}_i^{(j)}$. This comes at a sampling overhead of

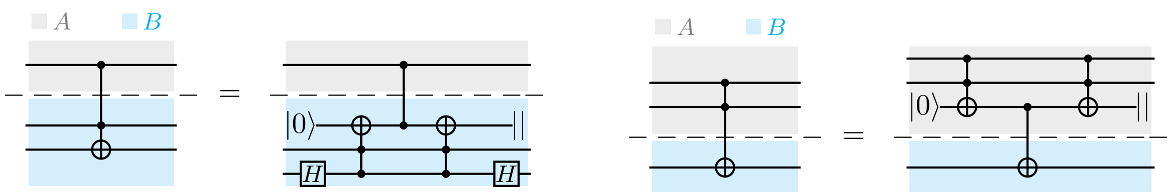
$$1 + 2\sum_{\mathbf{k}\neq\mathbf{k}'} |w_{\mathbf{k}}||w_{\mathbf{k}'}|\,, \tag{89}$$

where $\mathbf{k} \in \{0,1,2,3\}^N$ and $w_{\mathbf{k}} = \prod_{i=1}^N u_{k_i}^{(i)}$ is the product of the KAK-coefficients of $U_{A_i,B_i}^{(i)}$. Importantly, instead of implementing the two-qubit unitaries $U_{A_i,B_i}^{(i)}$ at their position A_i, B_i we apply them to Bell states in the environment E_A and E_B as seen in the previous part of this section. These "virtual" Choi states can then be used to apply the two-qubit gate at a later time. This is done via the operations $\{\mathcal{F}_i^{(j)}\}_{i=1}^{j-1}$. As seen before, the strategy consists of a bell measurement between E_A and A_i and E_B and B_i together with a corresponding local correction. The overall sign of the sampling run can be calculated based on the measurement results. Since all this can be done locally and in the postprocessing, these operations do not change the overhead. □

7 Open questions

While our work now provides an extended understanding of circuit cutting for two-qubit unitaries, it still remains an open question to characterize the optimal overhead for more general unitaries. For some specific unitaries like the Toffoli gate, our results can be naturally extended, as seen in Remark 7.1 below.

Remark 7.1 (Toffoli gate)**.** We can apply Corollary 3.1 to obtain optimal cuts for the Toffoli gate via the two identities given in Figure 5. We immediately see that $\gamma(\text{Toffoli}) \leq$



(a) Cutting a Toffoli between qubits 1 and 2.

(b) Cutting a Toffoli between qubits 2 and 3.

Figure 5: These two circuits allow us to prove that $\gamma(\text{Toffoli}) = 3$ as described in the text below.

3, since $\gamma(\text{CZ}) = \gamma(\text{CNOT}) = 3$ as ensured by Corollary 3.1. The other direction, i.e. $\gamma(\text{Toffoli}) \geq 3$ follows since $\gamma(\text{Toffoli}) \geq \gamma(|\psi_{\text{Toffoli}}^{(\phi)}\rangle\langle\psi_{\text{Toffoli}}^{(\phi)}|)$, where

$$|\psi_{\text{Toffoli}}^{(\phi)}\rangle = (\text{Toffoli}_{AAB} \otimes \mathbb{1}_B)(|+\rangle_A|1\rangle_A \otimes |0\rangle_B|0\rangle_B)\,. \tag{90}$$

A small calculation then shows

$$\gamma(\text{Toffoli}) \geq \gamma(|\psi_{\text{Toffoli}}^{(\phi)}\rangle\langle\psi_{\text{Toffoli}}^{(\phi)}|) = 3\,. \tag{91}$$

Surprisingly, $\gamma(|\text{Choi}_{\text{Toffoli}}\rangle) = 2.76$, indicating, that the Choi state is in general not a tight bound from below. Similar identities can be found for larger multi-controlled gates based on the symmetry of the CCZ gate.

If similar arguments can be made for arbitrary unitaries is unclear. An interesting step in this direction could be to isolate interaction parts. For instance, for the Toffoli gate, it is possible to isolate the non-local interaction, even though it is not of KAK-like form and Theorem 5.1 does not apply directly. It would be interesting to see for which other unitaries this is the case.

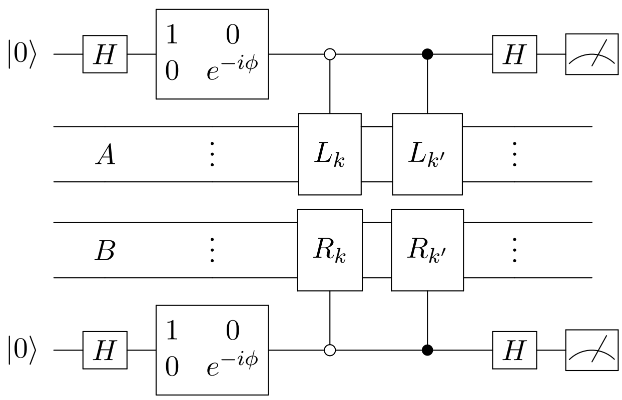
Another important topic for further research is the characterization of the optimal sampling overhead for non-unitary channels. Here it seems likely that there is a strict separation between γ_{LO} and γ_{LOCC}, at least on channels that themselves involve classical communication between the bipartitions.

Relation to other works During the preparation of this work, we became aware of two other independent efforts to optimally cut two-qubit unitaries which were published at almost the same time [16, 17]. The authors in [16] obtain the same optimality result for the single-cut, parallel cut, and black box setting, however they only consider controlled rotation gates and not general two-qubit gates. In [17], the authors derive the same technical result as Theorem 5.1 and thus obtain the same optimality result for single cuts and parallel cuts of arbitrary two-qubit gates. However, they do not investigate the optimal overhead for the black box setting.

Acknowledgements We thank Stefan Woerner for help with Remark 7.1 and Angus Lowe for useful discussions. This work was supported by the ETH Quantum Center, the National Centres for Competence in Research in Quantum Science and Technology (QSIT) and The Mathematics of Physics (SwissMAP), and the Swiss National Science Foundation Sinergia grant CRSII5 186364.

A Circuits for implementing cutting protocol

To obtain the optimal sampling overhead presented in Theorem 5.1, one has to apply the channel $(\mathcal{C}^{\phi}_{k,k'})_{\pm}(\rho) \otimes \mathcal{D}^{\phi}_{k,k'})_{\pm}(\rho)$. This channel is implemented by the following circuit:



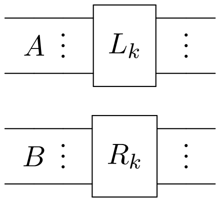
The first and the last qubit are ancillary qubits that are required for the negativity trick. Their outcome a_1 and a_2 ($a_i \in \{0,1\}$) determines the sign $(-1)^{a_1+a_2}$ with which the shot is weighted. With this the total workflow can be described by the following two steps:

(1) Determine the KAK-like decomposition of the unitary U_{AB} that is to be cut.

$$U_{AB} = \left(V_A^{(1)} \otimes V_B^{(2)}\right) \sum_{k=1}^{N} u_k (L_k)_A \otimes (R_k)_B \left(V_A^{(3)} \otimes V_B^{(4)}\right)$$

In the case of parallel two-qubit unitaries, this can be easily achieved by multiplying their single KAK-decompositions.

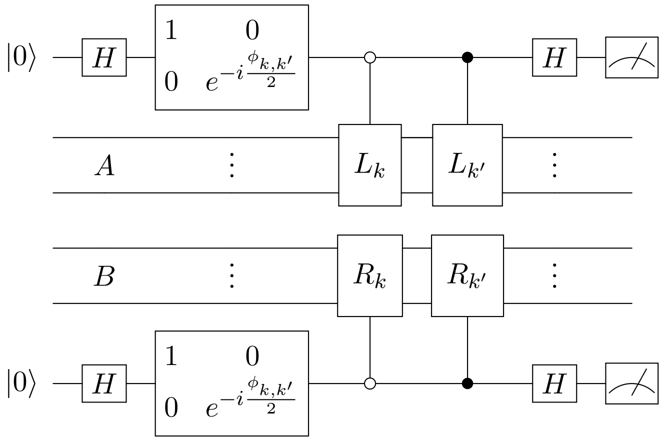
(2) Randomly run the following circuits and weight the measurement outcome with the corresponding factor in post processing:

- 

,

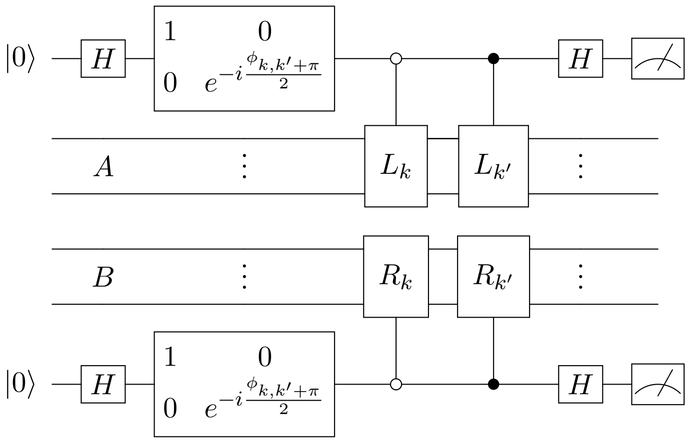
  The measurement outcome of each of these N circuits is weighted by the KAK-coefficient $|u_k|^2$.

- There are $N(N-1)/2$ circuits of the form

  with $k < k'$, that have to be weighted by $2|u_k||u_{k'}|(-1)^{a_1+a_2}$.

- There $N(N-1)/2$ circuits of the form

  with $k < k'$, that have to be weighted by $-2|u_k||u_{k'}|(-1)^{a_1+a_2}$.

The phases $\phi_{k,k'}$ are determined via the phases of the coefficients $u_k = |u_k|\mathrm{e}^{\mathrm{i}\phi_k}$ and $\phi_{k,k'} = \phi_k - \phi_k'$.

References

[1] K. Mitarai and K. Fujii. Constructing a virtual two-qubit gate by sampling single-qubit operations. *New Journal of Physics*, 23(2):023021, 2021. DOI: 10.1088/1367-2630/abd7bc.

[2] K. Mitarai and K. Fujii. Overhead for simulating a non-local channel with local channels by quasiprobability sampling. *Quantum*, 5:388, 2021. DOI: 10.22331/q-2021-01-28-388.

[3] C. Piveteau and D. Sutter. Circuit knitting with classical communication. *IEEE Transactions on Information Theory*, pages 1–1, 2023. DOI: 10.1109/TIT.2023.3310797.

[4] T. Peng, A. W. Harrow, M. Ozols, and X. Wu. Simulating large quantum circuits on a small quantum computer. *Phys. Rev. Lett.*, 125:150504, 2020. DOI: 10.1103/PhysRevLett.125.150504.

[5] L. Brenner, C. Piveteau, and D. Sutter. Optimal wire cutting with classical communication, 2023. Available online: https://arxiv.org/abs/2302.03366.

[6] H. Harada, K. Wada, and N. Yamamoto. Doubly optimal parallel wire cutting without ancilla qubits. *PRX Quantum*, 5:040308, 2024. DOI: 10.1103/PRXQuantum.5.040308.

[7] E. Pednault. An alternative approach to optimal wire cutting without ancilla qubits, 2023. Available online: https://arxiv.org/abs/2303.08287.

[8] A. Lowe, M. Medvidović, A. Hayes, L. J. O'Riordan, T. R. Bromley, J. M. Arrazola, and N. Killoran. Fast quantum circuit cutting with randomized measurements. *Quantum*, 7:934, 2023. DOI: 10.22331/q-2023-03-02-934.

[9] K. Temme, S. Bravyi, and J. M. Gambetta. Error mitigation for short-depth quantum circuits. *Phys. Rev. Lett.*, 119:180509, 2017. DOI: 10.1103/PhysRevLett.119.180509.

[10] S. Endo, S. C. Benjamin, and Y. Li. Practical quantum error mitigation for near-future applications. *Phys. Rev. X*, 8:031027, 2018. DOI: 10.1103/PhysRevX.8.031027.

[11] C. Piveteau. Advanced methods for quasiprobabilistic quantum error mitigation. DOI: 10.3929/ethz-b-000504508. Master thesis, ETH Zurich, September 2020.

[12] X. Zhao, L. Zhang, B. Zhao, and X. Wang. Power of quantum measurement in simulating unphysical operations, 2023. Available online: https://arxiv.org/abs/2309.09963.

[13] E. Chitambar, D. Leung, L. Mančinska, M. Ozols, and A. Winter. Everything you always wanted to know about LOCC (but were afraid to ask). *Communications in Mathematical Physics*, 328(1):303–326, 2014. DOI: 10.1007/s00220-014-1953-9.

[14] N. Khaneja and S. J. Glaser. Cartan decomposition of su(2n) and control of spin systems. *Chemical Physics*, 267(1):11–23, 2001. DOI: https://doi.org/10.1016/S0301-0104(01)00318-4.

[15] S. Bravyi, G. Smith, and J. A. Smolin. Trading classical and quantum computational resources. *Phys. Rev. X*, 6:021043, 2016. DOI: 10.1103/PhysRevX.6.021043.

[16] C. Ufrecht, L. S. Herzog, D. D. Scherer, M. Periyasamy, S. Rietsch, A. Plinge, and C. Mutschler. Optimal joint cutting of two-qubit rotation gates. *Phys. Rev. A*, 109:052440, 2024. DOI: 10.1103/PhysRevA.109.052440.

[17] A. W. Harrow and A. Lowe. Optimal quantum circuit cuts with application to clustered Hamiltonian simulation. *PRX Quantum*, 6:010316, 2025. DOI: 10.1103/PRXQuantum.6.010316.