arXiv:2202.12114v2 [quant-ph] 9 Oct 2022

Faster Born probability estimation via gate merging and frame optimisation

Nikolaos Koukoulekidis[1], Hyukjoon Kwon[1,2], Hyejung H. Jee[3], David Jennings[4,1], and M. S. Kim[1]

[1]Department of Physics, Imperial College London, London SW7 2AZ, UK
[2]Korea Institute for Advanced Study, Seoul, 02455, Korea
[3]Department of Computing, Imperial College London, London SW7 2AZ, UK
[4]School of Physics and Astronomy, University of Leeds, Leeds, LS2 9JT, UK

October 11, 2022

Outcome probability estimation via classical methods is an important task for validating quantum computing devices. Outcome probabilities of any quantum circuit can be estimated using Monte Carlo sampling, where the amount of negativity present in the circuit frame representation quantifies the overhead on the number of samples required to achieve a certain precision. In this paper, we propose two classical sub-routines: circuit gate merging and frame optimisation, which optimise the circuit representation to reduce the sampling overhead. We show that the runtimes of both sub-routines scale polynomially in circuit size and gate depth. Our methods are applicable to general circuits, regardless of generating gate sets, qudit dimensions and the chosen frame representations for the circuit components. We numerically demonstrate that our methods provide improved scaling in the negativity overhead for all tested cases of random circuits with Clifford+T and Haar-random gates, and that the performance of our methods compares favourably with prior quasi-probability simulators as the number of non-Clifford gates increases.

1 Introduction

Quantum computers promise to outperform their classical counterparts [1, 2]. However, the exact boundary between quantum and classical computational power is far from being fully characterised yet [3–7]. Several works have demonstrated the difficulty in simulating certain quantum processes classically [7–17]. Such results hint towards the ingredients that may be sufficient to achieve quantum advantage. It is also possible to approach the boundary from the other side, namely by finding efficient methods to classically simulate families of quantum circuits [18–24], thereby providing insights on what ingredients are necessary for quantum advantage.

The question of efficient probability estimation has recently received vast attention due to the ongoing rapid development of quantum devices aiming to supersede classical capabilities (e.g. [25–28]). Aided by powerful error mitigation techniques [29–33], noisy intermediate-scale quantum (NISQ) [34] devices aim to deliver computational advantages, therefore fast and accurate outcome probability estimation is a necessity for quantitative benchmarking of the devices [33, 35, 36]. For example, Google's recent experimental realisation of a quantum speed-up [26] relies on classical estimation methods to predict statistical features of the outcome probabilities.

It is expected that exact classical simulation of arbitrary quantum systems is inefficient, as the resource overhead exponentially grows with the size of the system. Nevertheless, there are restricted classes of quantum circuits for which exact classical simulation is possible [37]. The most notable example is given by circuits composed only with stabilizer states and gates in the Clifford group, which can be efficiently simulated classically via the Gottesman-Knill theorem [38].

Probability estimation methods are varied and aim to explore the efficiency of circuit sampling or simulation beyond the regime of quantum cir-

Nikolaos Koukoulekidis: nk2314@imperial.ac.uk

cuits that admit tractable classical representations. The estimation methods we mention in our work can be classified under one of two leading approaches [39, 40]. The first involves stabilizer rank-based simulators [41–47], which rely on approximating the circuit components by stabilizer operators. Every state or operation is assigned an exact or approximate stabilizer rank [43] indicating the number of stabilizer operators required to perform an exact or approximate decomposition of that component. If a circuit component is non-classical, its stabilizer rank grows large thus inducing an exponential runtime cost for the estimation of the outcome probability. Algorithms based on stabilizer decompositions have been very successful in estimating outcome probabilities of circuits dominated by Clifford gates and supplemented by a few types of magic states [39, 41, 42]. These Clifford simulators are generalised from pure to noisy settings by the recently proposed density-operator stabilizer-rank simulator [40]. Furthermore, computing the stabilizer rank of arbitrary gates appears to be an intractable problem in the general case, so recent improvements on computing stabilizer rank bounds for specific non-Clifford states enhance runtimes significantly [48].

The other family of estimation methods relies on quasi-probabilistic representations of circuit components [36, 40, 49–52] and such methods are in principle directly applicable to any quantum circuit without the need for state decompositions, in particular circuits with induced noise. They are based on the notion of a *frame representation* for the components of the circuit [53, 54]. Specifically, all components are represented by quasi-probability distributions in a certain frame and sampling on these distributions can be performed. Since any state or gate admits such a representation, quasi-probability simulators naturally apply to arbitrary circuits with noise. Many such frames have been studied [51, 53–58], and the runtime depends on the total negativity that is present in the circuit representation [49]. A notable frame simulator is the dyadic frame simulator [40] which relies on operator decompositions into stabilizer dyads $|L\rangle\langle R|$, where $|L\rangle$ and $|R\rangle$ are pure stabilizer states. This method assigns dyadic negativity to non-classical elements, which quantifies the extent to which the operator's optimal linear decomposition into stabilizer dyads departs from a convex combination. The dyadic simulator is a state-of-the-art quasi-probability frame simulator for qubits, as demonstrated by its low runtime scaling $O(4^{0.228t})$ with t non-Clifford gates [40]. However, optimising the decomposition of an operator in dyads is computationally challenging.

Stabilizer rank simulators generally offer two advantages over estimation methods based on frame representations. Firstly, they can be used for sampling the circuit output probabilities, which can be viewed as a stronger notion of simulation than probability estimation. Frame representation methods produce probability estimates with additive precision, which does not suffice for sampling [17, 42]. Secondly, the stabilizer-rank algorithms developed in [40–42] are quadratically faster as they achieve a scaling of $O(2^{0.228t})$ in the asymptotic limit. However, specialised simulators (e.g. [41, 42]) suffer from additional polynomial runtime factors, which tend to be more significant compared to the exponential runtime for the experimentally relevant case of big circuits with a low number of non-Clifford elements. Recently, an algorithm of additive precision [39] has also been shown to asymptotically outperform the methods of [41, 42], at least in certain parameter regimes.

In this paper, we focus on quasi-probability estimation methods based on frame representations and look for a way to improve the performance of outcome probability estimation. Recently, there has been a proposal of a Monte Carlo sampling algorithm which allows for quasi-probability estimation of circuits that contain a bounded amount of negativity in their representation [49]. For classes of circuits in which negativity grows only polynomially in the number of input states, this estimation algorithm is efficient. The negativity of the circuit therefore indicates the hardness of the probability sampling problem. Although the negativity scales exponentially with the number of non-Clifford gates, the scaling factors hugely depend on the frame choice. Until now, however, the same fixed representation has been applied on every circuit component and the flexibility on reducing negativity has been limited.

Our aim is to explore the extent to which varying the frame representations of the components in a given circuit can lead to a reduction in the total circuit negativity. To this end, we propose a pre-processing routine for any general quantum circuit, which aims at reducing the negativity overhead required for probability estimation. Our proposed routine consists of two distinct sub-routines:

1. **Circuit gate merging:** We introduce the idea of merging gates together into new n-qudit gates for fixed n in the context of reducing sampling overhead. This sub-routine reduces the negativity of the entire circuit and is independent of the estimation method used.

   We demonstrate numerically that the average negativity reduction over a random ensemble of circuits is greater as the number of non-Clifford elements, e.g., T gates, increases and is comparable to recent asymptotic negativity bounds [40, 59, 60]. Our routine does not depend on the specifics of the circuit gate set and can therefore be used in cases of gates which are hard to decompose, e.g., Haar-random gates.

2. **Frame optimisation:** We introduce the idea of using different frames to represent the input and output phase spaces of the gates in the circuit. This is inspired by work in continuous variables [61], but our approach is novel in the context of discrete quasi-probability sampling methods.

   We argue that this sub-routine compliments gate merging as an additional source of negativity reduction when merging is no longer efficient. We then demonstrate numerically that instances of Clifford+T circuits and circuits with Haar-random gates admit significant negativity reductions by introducing additional frames in the circuit representation.

We note that a polynomial runtime for these classical sub-routines with respect to the circuit size should be guaranteed to effectively reduce the overall runtime of the sampling method. As proof of principle, we provide explicit algorithms in the main text that ensure this condition for each sub-routine.

This paper is organised as follows. In Section 2, we review the frame representation and the estimation algorithm using quasi-probability representations of a given quantum circuit. In Section 3, we outline our results within the context of the current state of quasi-probability simulator research. In Section 4 and 5 we describe the two sub-routines in more detail, before providing a summary in Section 6.

2 Preliminaries

2.1 Frame representation of quantum circuits

We first give a brief overview on classical circuit sampling based on the method of frame representation. Suppose that an N-qudit quantum circuit C is composed of the initial state preparation ρ, sequential quantum gates $U_1, U_2, \ldots, U_L$ and the measurement effect E. The outcome probability of the quantum circuit $p_C = \mathrm{Tr}[U_L \ldots U_2 U_1 \rho U_1^\dagger U_2^\dagger \ldots U_L^\dagger E]$ can be estimated by describing the quantum state ρ as quasi-probability distributions over phase space points $\boldsymbol{\lambda} \in \mathbb{Z}_d^{2N}$ and the quantum operations U_i as the transition matrices of the distributions. More specifically, a phase space can be constructed from a frame defined as a set of operators $\mathcal{F} := \{F(\boldsymbol{\lambda})\}$ and its dual $\mathcal{G} := \{G(\boldsymbol{\lambda})\}$ [53, 54], such that any operator O is expressed as

$$O = \sum_{\boldsymbol{\lambda}} \mathrm{Tr}[F(\boldsymbol{\lambda})O]G(\boldsymbol{\lambda}). \tag{1}$$

For a given frame, the outcome probability can be expressed in terms of the representation as

$$p_C = \sum_{\boldsymbol{\lambda}_0,\ldots,\boldsymbol{\lambda}_L} W_E(\boldsymbol{\lambda}_L) \left[\prod_{l=1}^{L} W_{U_l}(\boldsymbol{\lambda}_l|\boldsymbol{\lambda}_{l-1})\right] W_\rho(\boldsymbol{\lambda}_0), \tag{2}$$

where we define

$$W_\rho(\boldsymbol{\lambda}) = \mathrm{Tr}[F(\boldsymbol{\lambda})\rho], \tag{3}$$

$$W_U(\boldsymbol{\lambda}'|\boldsymbol{\lambda}) = \mathrm{Tr}[F(\boldsymbol{\lambda}')UG(\boldsymbol{\lambda})U^\dagger], \text{ and} \tag{4}$$

$$W_E(\boldsymbol{\lambda}) = \mathrm{Tr}[EG(\boldsymbol{\lambda})]. \tag{5}$$

In the case where 1) ρ and E are products of local initial states and measurement effects, 2) $W_\rho(\boldsymbol{\lambda}_0)$ and $W_E(\boldsymbol{\lambda}_L)$ are classical probability distributions, and 3) $W_{U_l}(\boldsymbol{\lambda}_l|\boldsymbol{\lambda}_{l-1})$, for $\ell = 1, \ldots, L$, are classical conditional probability distributions for all l, efficient classical simulation is possible, where the sampling runtime scales polynomially with N and L [62]. The simulation is performed by sampling the trajectories of $(\boldsymbol{\lambda}_0, \ldots, \boldsymbol{\lambda}_L)$ from the initial distribution $P(\boldsymbol{\lambda}_0) = W_\rho(\boldsymbol{\lambda}_0)$ and the transition matrix at each step $P_l(\boldsymbol{\lambda}_l|\boldsymbol{\lambda}_{l-1}) = W_{U_l}(\boldsymbol{\lambda}_l|\boldsymbol{\lambda}_{l-1})$, which leads to the probability estimate, $\hat{p}_C = W_E(\boldsymbol{\lambda}_L)$. Taking an average over M probability estimates converges to the Born probability as M increases.

2.2 Overhead of classical simulation

Non-classicality in the quantum process is represented by negativities in $W_\rho(\boldsymbol{\lambda})$ or $W_U(\boldsymbol{\lambda}'|\boldsymbol{\lambda})$, which gives rise to quasi-probabilities. In general, $W_\rho(\boldsymbol{\lambda})$ and $W_U(\boldsymbol{\lambda}'|\boldsymbol{\lambda})$ consist of real components that can attain negative values, while satisfying the normalisation conditions,

$$\sum_{\boldsymbol{\lambda}\in\mathbb{Z}_d^{2N}} W_\rho(\boldsymbol{\lambda}) = 1 \text{ and} \tag{6}$$

$$\sum_{\boldsymbol{\lambda}'\in\mathbb{Z}_d^{2N}} W_U(\boldsymbol{\lambda}'|\boldsymbol{\lambda}) = 1 \text{ for all } \boldsymbol{\lambda}\in\mathbb{Z}_d^{2N}. \tag{7}$$

Despite the presence of negativities in the distributions and update matrices, Monte Carlo methods can still be used with adjustments as introduced by Pashayan *et al.* [49] in order to perform probability sampling. This can be done by sampling over $P(\boldsymbol{\lambda}_0) = |W_\rho(\boldsymbol{\lambda}_0)|/\sum_{\boldsymbol{\lambda}_0}|W_\rho(\boldsymbol{\lambda}_0)|$ for the initial state preparation and taking the transition matrix of $P_l(\boldsymbol{\lambda}_l|\boldsymbol{\lambda}_{l-1}) = |W_{U_l}(\boldsymbol{\lambda}_l|\boldsymbol{\lambda}_{l-1})|/\sum_{\boldsymbol{\lambda}_l}|W_{U_l}(\boldsymbol{\lambda}_l|\boldsymbol{\lambda}_{l-1})|$ for the quantum gate, while keep track of the signs. In this case, the probability estimate is modified to

$$\begin{aligned}\hat{p}_C = \text{Sign}\left(W_\rho(\boldsymbol{\lambda}_0)\prod_{l=1}^{L} W_{U_l}(\boldsymbol{\lambda}_l|\boldsymbol{\lambda}_{l-1})\right)\times \\ N_\rho\left(\prod_{l=1}^{L} N_{U_l}(\boldsymbol{\lambda}_{l-1})\right) W_E(\boldsymbol{\lambda}_L),\end{aligned} \tag{8}$$

where we have defined

$$N_\rho := \sum_{\boldsymbol{\lambda}_0}|W_\rho(\boldsymbol{\lambda}_0)| \tag{9}$$

$$N_{U_l}(\boldsymbol{\lambda}_{l-1}) := \sum_{\boldsymbol{\lambda}_l}|W_{U_l}(\boldsymbol{\lambda}_l|\boldsymbol{\lambda}_{l-1})|\,. \tag{10}$$

In order to converge to the Born probability, one can similarly take the average of increasingly many probability estimates sampled over trajectories $(\boldsymbol{\lambda}_0,\ldots,\boldsymbol{\lambda}_L)$ using distributions $P(\boldsymbol{\lambda}_0)$ and $P_l(\boldsymbol{\lambda}_l|\boldsymbol{\lambda}_{l-1})$.

This directly relates the total amount of circuit negativity to the computational overhead: the larger the negativity in the circuit, the more samples required for an accurate estimation.

Observation 1 (Pashayan *et al.* [49])**.** *The outcome probability p_C of the quantum circuit C can be estimated by $\hat{p}_C$ from the number of samples*

$$M \geq M(\epsilon,\delta) = \frac{2}{\epsilon^2}N_C^2\ln(2/\delta), \tag{11}$$

with at least probability $1-\delta$ of having error less than ϵ. Here,

$$N_C = N_\rho\times\left[\prod_{l=1}^{L}\max_{\boldsymbol{\lambda}_0,\ldots,\boldsymbol{\lambda}_{L-1}} N_{U_l}(\boldsymbol{\lambda}_{l-1})\right]\times\max_{\boldsymbol{\lambda}_L}|W_E(\boldsymbol{\lambda}_L)|\,, \tag{12}$$

is the (maximum) circuit negativity.

As is clear by Eq. (11), the negativity of the circuit acts as an overhead for the convergence time of the sampling algorithm, therefore it is desirable to reduce it before executing the sampling by considering different frame choices.

3 Main results

In this work, we develop a pre-processing routine to reduce the negativity of the circuit, which in turn reduces the number of samples required to estimate the outcome probability of the circuit. Our routine is applicable to any general circuit consisting of a product input state and product measurement, but independently of the input state dimension, the gate set (e.g. Clifford unitaries or Haar-random gates) and adaptive operations based on intermediate measurement outcomes.

3.1 Frame parametrisation

The central focus of our work is to consider frame parametrisations that are allowed to vary across the circuit. It is clear from the definitions that the circuit negativity of a given circuit in Eq. (12) depends on the choice of the frame $\mathcal{F}$ and its dual $\mathcal{G}$. Note that $\mathcal{F}$ and $\mathcal{G}$ are uniquely defined by each other for phase space dimension equal to d^2, where d is the qudit dimension. We therefore make the dependence clear by labelling the representation functions by $\mathcal{G}$:

$$W_\rho^{\mathcal{G}}(\boldsymbol{\lambda}) = \text{Tr}[F(\boldsymbol{\lambda})\rho], \tag{13}$$

$$W_U^{\mathcal{G}'|\mathcal{G}}(\boldsymbol{\lambda}'|\boldsymbol{\lambda}) = \text{Tr}[F'(\boldsymbol{\lambda}')UG(\boldsymbol{\lambda})U^\dagger], \text{ and} \tag{14}$$

$$W_E^{\mathcal{G}}(\boldsymbol{\lambda}) = \text{Tr}[EG(\boldsymbol{\lambda})], \tag{15}$$

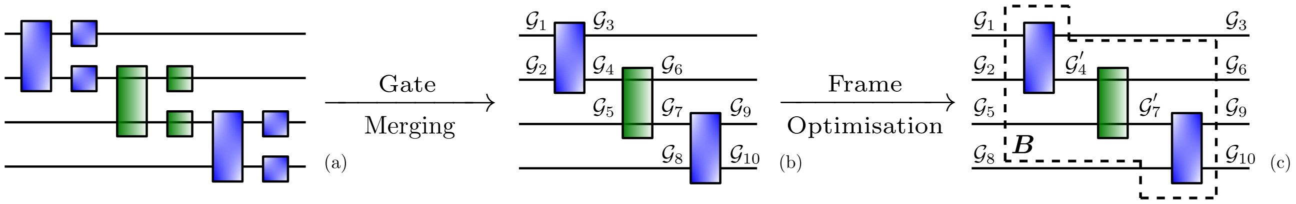

Figure 1: Sketch of routine on a toy circuit. The first step (a) → (b) is gate merging, here implemented with $n = 2$. Gates that share input and output wires merge in the schematic way depicted in the figure. The second step (b) → (c) is frame optimisation, here implemented with $\ell = 2$ in a block $\mathbf{B}$ comprising the three merged gates. The optimisation results into updated frames $\mathcal{G}_4', \mathcal{G}_7'$, while the remaining frames that connect the block $\mathbf{B}$ to the rest of the circuit components are left unchanged at this optimisation cycle.

where we used different frames, $\mathcal{G}$ and $\mathcal{G}'$, for the input and output wires respectively in the definition of $W_U^{\mathcal{G}'|\mathcal{G}}$.

In order to ensure that the number of frames does not grow exponentially with the number of qudits N, we restrict to *product* frames that are constructed as tensor products of single qudit frames. This allows us to parametrise each single qudit phase space separately, rather than the entire N-qudit phase space. Therefore, we reserve the label $\mathcal{G}$ for denoting single qudit frames and the boldface symbol $\boldsymbol{\mathcal{G}}$ for denoting a set of single qudit frames. The negativity of each circuit component can now be expressed as

$$N_\rho^{\boldsymbol{\mathcal{G}}} = \sum_{\boldsymbol{\lambda}} \left| W_\rho^{\boldsymbol{\mathcal{G}}}(\boldsymbol{\lambda}) \right| , \tag{16}$$

$$N^{\boldsymbol{\mathcal{G}}'|\boldsymbol{\mathcal{G}}} = \max_{\boldsymbol{\lambda}} \left[\sum_{\boldsymbol{\lambda}'} \left| W_U^{\boldsymbol{\mathcal{G}}'|\boldsymbol{\mathcal{G}}}(\boldsymbol{\lambda}'|\boldsymbol{\lambda}) \right| \right] , \text{ and} \tag{17}$$

$$N_E^{\boldsymbol{\mathcal{G}}} = \max_{\boldsymbol{\lambda}} \left| W_E^{\boldsymbol{\mathcal{G}}}(\boldsymbol{\lambda}) \right| , \tag{18}$$

where $\boldsymbol{\mathcal{G}}, \boldsymbol{\mathcal{G}}'$ contain elements from the complete set of frames required to represent the circuit. In practice, each circuit component is parametrised only via the frames that correspond to its input and output wires. For example, in Fig.1(b), when parametrising the first gate in the sequence, we can simply consider $\boldsymbol{\mathcal{G}}$ as the set $\{\mathcal{G}_1, \mathcal{G}_2\}$ and $\boldsymbol{\mathcal{G}}'$ as the set $\{\mathcal{G}_3, \mathcal{G}_4\}$. If only a unique frame representation $\mathcal{G}$ is used for all circuit components, then $\boldsymbol{\mathcal{G}} = \boldsymbol{\mathcal{G}}' = \{\mathcal{G}\}$ in the expressions above and the label can be dropped, simplifying to the notation of the previous section. The total negativity of the parametrised circuit can now be expressed as a function of the circuit frame set $\boldsymbol{\mathcal{G}}$:

$$N_C(\boldsymbol{\mathcal{G}}) = N_\rho^{\boldsymbol{\mathcal{G}}} \times \left[\prod_{l=1}^{L} N_{U_l}^{\boldsymbol{\mathcal{G}}'|\boldsymbol{\mathcal{G}}} \right] \times N_E^{\boldsymbol{\mathcal{G}}}. \tag{19}$$

We note that Observation 1 still holds by replacing N_C with the more general form $N_C(\boldsymbol{\mathcal{G}})$. Our main objective is to study the reduction of this circuit negativity by tuning $\boldsymbol{\mathcal{G}}$.

3.2 Examples of frame parametrisations

While our results are general and applicable to any family of parametrised frames, in this work we provide two examples of explicit, product frame parametrisations: (i) parametrised Wigner frames and (ii) rotated Pauli frames.

Parametrised Wigner frames employ the conventional phase space of the discrete Wigner function [54, 55]. Let us define the discrete displacement operator for a d-dimensional system as

$$D(p,q) = \chi(-2^{-1}pq) Z^p X^q, \tag{20}$$

where $\chi(q) = e^{i(2\pi/d)q}$. For a qubit system ($d = 2$), this takes the form $D(p,q) = i^{pq} Z^p X^q$. It can be generalised to an N-qudit system as

$$D(\boldsymbol{\lambda}) = \bigotimes_{i=1}^{N} D(p_i, q_i) , \tag{21}$$

where $\boldsymbol{\lambda} := (p_1, q_1, p_2, q_2, \ldots, p_N, q_N)^T \in \mathbb{Z}_d^{2N}$ denotes a phase space point of the whole system. We then define the frame $\mathcal{F} = \{F(\boldsymbol{\lambda})\} := \{D(\boldsymbol{\lambda}) F_0 D^\dagger(\boldsymbol{\lambda})\}$ and its dual frame $\mathcal{G} = \{G(\boldsymbol{\lambda})\} := \{D(\boldsymbol{\lambda}) G_0 D^\dagger(\boldsymbol{\lambda})\}$ using the following reference operators:

$$F_0 = \frac{1}{d} \sum_{\boldsymbol{\lambda}} \left[\frac{1}{g(\boldsymbol{\lambda})} \right] D(\boldsymbol{\lambda}) \tag{22}$$

$$G_0 = \frac{1}{d} \sum_{\boldsymbol{\lambda}} g(\boldsymbol{\lambda}) D(-\boldsymbol{\lambda}) , \tag{23}$$

where we introduced the parametrisation function $g(\boldsymbol{\lambda})$. Note that the following relation holds:

$$g(\boldsymbol{\lambda}) = \mathrm{Tr}\left[G_0 D(\boldsymbol{\lambda}) \right] , \tag{24}$$

so the parametrisation function $g(\boldsymbol{\lambda}) : \mathbb{Z}_d^{2N} \mapsto \mathbb{C}\setminus\{0\}$ can be fully characterised by the reference operator G_0. In order to impose that $W_\rho^{\mathcal{G}}(\boldsymbol{\lambda})$ is real-valued and that $\sum_{\boldsymbol{\lambda}} W_\rho^{\mathcal{G}}(\boldsymbol{\lambda}) = 1$, we need the additional conditions $g^*(\boldsymbol{\omega}) = g(-\boldsymbol{\omega})$ and $g(\mathbf{0}) = 1$, which are equivalent to $G_0^\dagger = G_0$ and $\mathrm{Tr}[G_0] = 1$ respectively. By taking $g(\boldsymbol{\lambda}) = 1$ for all $\boldsymbol{\lambda}$, the conventional discrete Wigner function [55] is recovered. One can calculate the quasi-probability distributions of circuit elements via Eq. (13)-(15) using the defined frame and dual frame. In odd dimensions, the parametrised Wigner frame is a good choice for Clifford dominated circuits as Clifford gates do not possess any negativity in the conventional Wigner distribution. Therefore, $g(\boldsymbol{\lambda}) = 1$ is already optimal for most circuit elements when considered in isolation and constitutes an obvious starting point for frame optimisation.

In the qubit case, it is known that the Hadamard and CNOT gates have non-zero negativity even in the conventional Wigner distribution [63], which motivates us to introduce the next frame parametrisation, valid only for qubits: the rotated Pauli frames.

Rotated Pauli frames are based on the Bloch decomposition of a quantum operator. Consider the set of displacement operators for a single qubit $\{D(\boldsymbol{\lambda})\}$ as defined in Eq. (20) for $\boldsymbol{\lambda} \in \mathbb{Z}_2^2 = \{(0,0),(0,1),(1,0),(1,1)\}$. The usual Bloch vector for a single-qubit state ρ can be written as

$$W_\rho(\boldsymbol{\lambda}) = \frac{1}{2}\mathrm{Tr}\left[\rho D(\boldsymbol{\lambda})\right], \tag{25}$$

and this defines a valid quasi-probability distribution with frame $\{\frac{1}{2}D(\boldsymbol{\lambda})\}$. We can define a new frame by applying a rotation to the space of the Bloch vector. Let us consider a rotational angle vector $\boldsymbol{\theta} := (\theta_X, \theta_Y, \theta_Z)$ and a corresponding rotation operator $R(\boldsymbol{\theta}) := R(\theta_Z)R(\theta_Y)R(\theta_X)$, where $R(\theta_X) := e^{-i\theta X/2}$ and similarly for Y, Z. Applying this to the Bloch vector in Eq. (25) results in a set of rotated displacement operators, parametrised by $\boldsymbol{\theta}$:

$$D^{\boldsymbol{\theta}}(0,0) := \frac{1}{2}\mathbb{1} \tag{26}$$

$$D^{\boldsymbol{\theta}}(0,1) := \begin{pmatrix} -\sin\theta_Y & e^{-i\theta_X}\cos\theta_Y \\ e^{+i\theta_X}\cos\theta_Y & \sin\theta_Y \end{pmatrix} \tag{27}$$

$$D^{\boldsymbol{\theta}}(1,0) := \begin{pmatrix} \cos\theta_Y\cos\theta_Z & e^{-i\theta_X}(\sin\theta_Y\cos\theta_Z + i\sin\theta_Z) \\ e^{+i\theta_X}(\sin\theta_Y\cos\theta_Z - i\sin\theta_Z) & -\cos\theta_Y\cos\theta_Z \end{pmatrix} \tag{28}$$

$$D^{\boldsymbol{\theta}}(1,1) := \begin{pmatrix} \cos\theta_Y\sin\theta_Z & e^{-i\theta_X}(\sin\theta_Y\sin\theta_Z + i\cos\theta_Z) \\ e^{+i\theta_X}(\sin\theta_Y\sin\theta_Z - i\cos\theta_Z) & -\cos\theta_Y\sin\theta_Z \end{pmatrix}. \tag{29}$$

Then, we define the frame $\mathcal{F} = \{F(\boldsymbol{\lambda})\}$ and its dual frame $\mathcal{G} = \{G(\boldsymbol{\lambda})\}$ as

$$F(\boldsymbol{\lambda}) := \frac{1}{2^N}D^{\boldsymbol{\theta}}(\boldsymbol{\lambda}), \tag{30}$$

$$G(\boldsymbol{\lambda}) := D^{\boldsymbol{\theta}}(\boldsymbol{\lambda}), \tag{31}$$

which provide a parametrised frame representation for a qubit. This can be generalised to an N-qubit system via

$$D^{\boldsymbol{\theta}}(\boldsymbol{\lambda}) = \bigotimes_{i=1}^{N} D^{\boldsymbol{\theta}_i}(\boldsymbol{\lambda}_i) \tag{32}$$

with $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N)$, where $\boldsymbol{\theta}_i$ is the rotational angle vector for the i-th qubit. The rotated Pauli frames possess the desired property that all stabilizer states and Clifford gates have zero negativity in the conventional Bloch frame representation with $\boldsymbol{\theta} = (0,0,0)$. Thus, when a given qubit circuit is dominated by Clifford gates, it can be advantageous to employ the rotated Pauli frame.

3.3 Pre-processing routine for negativity reduction

The central idea of our pre-processing routine for negativity reduction can now be expressed by the following lower bounds on gate negativity.

Theorem 1. *For two consecutive gates U and V, the following bounds on negativity hold:*

$$N_V^{\boldsymbol{\mathcal{G}}|\boldsymbol{\mathcal{G}}} N_U^{\boldsymbol{\mathcal{G}}|\boldsymbol{\mathcal{G}}} \geq \min_{\boldsymbol{\mathcal{G}}'} N_V^{\boldsymbol{\mathcal{G}}|\boldsymbol{\mathcal{G}}'} N_U^{\boldsymbol{\mathcal{G}}'|\boldsymbol{\mathcal{G}}} \geq N_{VU}^{\boldsymbol{\mathcal{G}}|\boldsymbol{\mathcal{G}}}, \quad (33)$$

where $\boldsymbol{\mathcal{G}}$ and $\boldsymbol{\mathcal{G}}'$ are frame sets that represent gates U, V and UV.

Proof. The first inequality holds since $\boldsymbol{\mathcal{G}}$ is one specific choice of the optimisation variable set $\boldsymbol{\mathcal{G}}'$. The second inequality is due to Observation 2 in the next section. □

Theorem 1 motivates us to introduce two sub-routines applicable to any quasi-probability estimation algorithm with runtime cost determined by the circuit negativity.

The second inequality in Theorem 1 suggests that merging two gates into one is generally advantageous in minimising the total negativity. This leads to the first pre-processing sub-routine, *gate merging*. The inequality is independent of the specific frame parametrisation and can be directly extended to an arbitrary number of gates. The trade-off is that the merged gate may be of a larger size. For example, if U and V are 2-qudit gates sharing one wire between them, gate VU will be a 3-qudit gate. The dimension of the merged gate increases exponentially as the number of qudits involved becomes larger, hence one should truncate the maximum number of qudits acted on by the merged gates, which we define as the spatial parameter n.

The first inequality in Theorem 1 states that, unless the frames between two gates in sequence are already optimal, we can always reduce the total negativity of the two gates by optimising the frames they share. This leads to the second sub-routine, *frame optimisation*. The optimisation can be directly generalised to a circuit block $\mathbf{B}$ containing a sequence of ℓ frames $\boldsymbol{\mathcal{G}}$ by simultaneously optimising all the frames in the block, $\min_{\boldsymbol{\mathcal{G}}} N_{\mathbf{B}}(\boldsymbol{\mathcal{G}})$. The temporal parameter ℓ is the number of frames to be optimised in one optimisation cycle. The optimisation takes place iteratively in the sense that every optimisation cycle optimises the frames within a block, taking as an initial state the optimised frames obtained from the previous cycle. This ensures that negativity cannot increase above its initial value, no matter how many optimisation cycles occur.

Given fixed values for the truncation parameters n, ℓ, we show in the following two sections that the total runtime τ of our routine is polynomial in the number of circuit components,

$$\tau = O(N, L^2). \quad (34)$$

In general, larger n or ℓ give larger negativity reduction at the cost of additional classical computation.

We note that gate merging yields lower negativity than any frame optimisation between the gates. However, fixing $n < N$ prevents us from merging gates indefinitely, so frame optimisation can then be used for further negativity reduction.

Algorithm 1 Outcome Probability Estimation with Merging and Optimisation

```
Input: An $N$-qudit quantum circuit $C$ with a product input state $\rho = \rho_1 \otimes \cdots \otimes \rho_N$, the list of gates $\mathcal{U} = \{U_1, ..., U_L\}$, and the product measurement operator $E = E_1 \otimes \cdots \otimes E_N$; the spatial parameter $n$; the temporal parameter $\ell$; the desired accuracy $\epsilon$.
Run gate merging (Sub-routine 1) with the input gate sequence and $n$ and return the merged gate sequence $\{V_1, ..., V_{L'}\}$ with $L' \leq L$ consisting of gates acting on at most $n$ qudits.
Run frame optimisation (Sub-routine 2) with the merged circuit and $\ell$ and return the optimised frame sequence $\boldsymbol{\mathcal{G}}_{\text{opt}}$.
Run a sampling algorithm to achieve the input accuracy $\epsilon$ according to Eq. (11) using the quasi-probability representations of the merged circuit obtained with the optimised frame sequence $\boldsymbol{\mathcal{G}}_{\text{opt}}$.
Output: $p_{\text{est}}$, the estimated outcome probability.
```

We present an algorithm for Born probability estimation, including our complete pre-processing routine and sampling, in Algorithm 1 and illustrate its implementation on a toy circuit in Fig. 1. In the following two sections, we discuss in more detail how the two sub-routines, gate merging and frame optimisation, can be implemented. For clarity, we focus on qubit circuits and on the frame parametrisations introduced in the previous section, although our methods are general.

4 Gate merging

The central idea of our first sub-routine, gate merging, is that the sampling cost of a merged circuit block consisting of multiple quantum gates is in general lower than sequential sampling of each gate. More precisely, this can be summarised as the following observation:

Observation 2. *Let $\{U_1, U_2, \ldots, U_k\}$ be a sequence of quantum gates. The negativity of the merged gate $U = U_k \ldots U_2 U_1$ is always less or equal to the product of the individual negativities, i.e.,*

$$\mathcal{N}_U^{\mathcal{G}} \leq \prod_{i=1}^{k} \mathcal{N}_{U_i}^{\mathcal{G}}, \tag{35}$$

for any frame set $\mathcal{G}$ assigned to the gate sequence.

Proof. It is sufficient to prove the statement for two gates U and V. By noting that the quasi-probability of the merged gate is expressed as

$$W_{VU}^{\mathcal{G}}(\boldsymbol{\lambda}_3|\boldsymbol{\lambda}_1) = \sum_{\boldsymbol{\lambda}_2} W_V^{\mathcal{G}}(\boldsymbol{\lambda}_3|\boldsymbol{\lambda}_2) W_U^{\mathcal{G}}(\boldsymbol{\lambda}_2|\boldsymbol{\lambda}_1), \tag{36}$$

the negativity of the gate can be bounded as

$$\begin{aligned}
\mathcal{N}_{VU}^{\mathcal{G}} &= \max_{\boldsymbol{\lambda}_1} \sum_{\boldsymbol{\lambda}_3} \left| W_{VU}^{\mathcal{G}}(\boldsymbol{\lambda}_3|\boldsymbol{\lambda}_1) \right| \\
&= \max_{\boldsymbol{\lambda}_1} \sum_{\boldsymbol{\lambda}_3} \left| \sum_{\boldsymbol{\lambda}_2} W_V^{\mathcal{G}}(\boldsymbol{\lambda}_3|\boldsymbol{\lambda}_2) W_U^{\mathcal{G}}(\boldsymbol{\lambda}_2|\boldsymbol{\lambda}_1) \right| \\
&\leq \max_{\boldsymbol{\lambda}_1} \sum_{\boldsymbol{\lambda}_2} \left| W_U^{\mathcal{G}}(\boldsymbol{\lambda}_2|\boldsymbol{\lambda}_1) \right| \sum_{\boldsymbol{\lambda}_3} \left| W_V^{\mathcal{G}}(\boldsymbol{\lambda}_3|\boldsymbol{\lambda}_2) \right| \\
&\leq \mathcal{N}_U^{\mathcal{G}} \max_{\boldsymbol{\lambda}_2} \sum_{\boldsymbol{\lambda}_3} \left| W_V^{\mathcal{G}}(\boldsymbol{\lambda}_3|\boldsymbol{\lambda}_2) \right| \\
&= \mathcal{N}_V^{\mathcal{G}} \mathcal{N}_U^{\mathcal{G}}.
\end{aligned} \tag{37}$$

We then apply this argument iteratively to any sequence of quantum gates $\{U_1, U_2, \ldots, U_k\}$ to obtain Eq. (35), which completes the proof. □

Such a negativity reduction can be exemplified by considering the Toffoli gate, which can be optimally decomposed into four T gates [64] along with Clifford gates and Pauli measurements. We compare the negativity of the Toffoli gate itself and its decomposed gate sequence using the Pauli frame, where the negativity only comes from non-Clifford gates. One can readily observe that the Toffoli gate negativity $\mathcal{N}_{\text{Toffoli}}^{\text{Pauli}} = 2$ is lower than the total negativity of the decomposed gate sequence $\left[\mathcal{N}_T^{\text{Pauli}}\right]^4 = 4$.

The idea of reducing the negativity of quantum gates by merging (Eq. (35)) can be compared to the submultiplicativity of magic state negativity characterised by the robustness measure ($\mathcal{R}$), which obeys $\mathcal{R}(\rho_1 \otimes \rho_2) \leq \mathcal{R}(\rho_1)\mathcal{R}(\rho_2)$ [59]. In particular, the robustness of the T state is equivalent to the negativity of the T gate from the sampling cost viewpoint, as one T gate can be "gadgetised" via Cliffords and a single T state [41]. In Ref. [59], the asymptotic negativity per single T gate is $\lim_{t\to\infty} \left[\mathcal{R}\left(|T\rangle^{\otimes t}\right)\right]^{1/t} \approx 2^{0.272}$ which provides a lower bound on their sampling runtime $\Omega(4^{0.272t})$.

In order to compare this with the gate merging method, we consider an n-qubit block consisting of Clifford+T gates (see Fig. 2(f) for an example with $n = 5$). This can be compared to considering n T states in the robustness measure, having the same number of qubits (i.e., the size of Hilbert space) in the block to evaluate the negativity. Figs. 2(a-d) show the distribution of the negativity of 1000 random n-qubit blocks consisting of 100 Clifford gates and t T gates. We observe that the negativity per T gate after merging the gate sequences in a random n-qubit block can be occasionally lower than the robustness measure of n T states [59]. We also note that the negativity reduction works efficiently when the number of T gate in the block, t, increases. For example, when $n = 5$ and $t = 15$, 95% of the randomly chosen merged blocks yield negativity per T state lower than the robustness measure. We also plot in Fig. 2(e) the average negativity per T gate versus t, demonstrating that it is decreasing, which implies that our appoach can prove efficient when the structure of the gate block considered becomes more complicated.

The main advantage of our approach is that it is not limited to a particular type of gate set, e.g. Clifford+T circuits, but can be directly applied to any types of quantum gates. The aforementioned approaches using stabilizer rank, robustness and generalised robustness rely on the gadgetisation of a non-Clifford gate using magic states. Therefore, evaluating the classical overhead should be preceded by finding an optimal Clifford gadget with minimum resource of magic. On the other hand, gate merging does not have such a limitation, so it can be useful when the efficient decomposition of a quantum circuit into non-stabilizer states and Clifford gates is non-

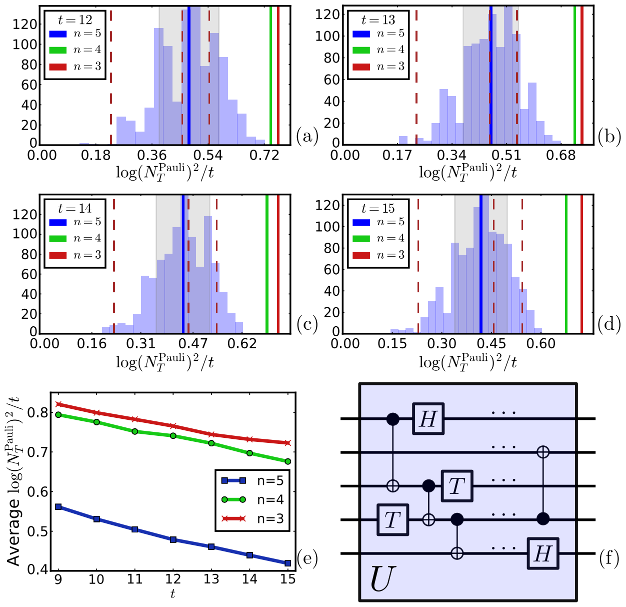

Figure 2: Histograms of 1000 random Clifford+T circuits with $N = 5$ consisting of 100 1-qubit and 2-qubit Clifford gates, supplemented by t T gates and merged using spatial parameter $n = 5$. The leftmost (blue) solid line with the gray region depict the average and standard deviation of each histogram. The brown and green solid lines (from right to left) represent the higher averages of the corresponding histograms for $n = 3$ and 4 respectively. Vertical dashed lines provide some state-of-the art scalings, more specifically from left to right: $O(2^{0.228t})$ of the Bravyi-Gosset algorithm from [41] based on the stabilizer rank, $O(4^{0.228t})$ of the dyadic frame simulator from [40] and the lower bound $\Omega(4^{0.272t})$ based on the robustness of magic from [59]. As t increases, we observe a higher frequency of circuits with log negativity squared per T gate lower than the robustness lower bound: (a) 71%, (b) 81%, (c) 89%, (d) 95%.
(e) Histogram average for $n = 3, 4, 5$ against t.
(f) Example 5-qubit merged gate U made up from Clifford gate ($CNOT$s and H) and T gates.

trivial. We also highlight that merging gates reduces the negativity independently of the choice of frames.

We now describe the gate merging method for a generic N-qubit quantum circuit with L gates. This can be done by grouping the quantum circuit into n-qubit blocks (see Fig. 1(a)→(b)), then Observation 2 guarantees that the negativity of each block is reduced after merging the gate sequences in it. There are various ways of grouping the circuit into n-qubit blocks, but we introduce the iterative Sub-routine 1 for concreteness. The broad idea of the sub-routine is to iteratively connect any yet unmerged (disjoint) gates. All gates remain in the set $\mathcal{U}_{\text{disj}}$ until they either finally act on n qubits or cannot connect to other gates anymore, when they are move to the output set $\mathcal{U}_{\text{merged}}$.

Sub-routine 1 Gate merging

```
Input: List of gates $\mathcal{U} = \{U_1, ..., U_L\}$ in
  qudit quantum circuit $C$ and spatial
  parameter $n$.
Define list of merged gates $\mathcal{U}_{\text{merged}} \leftarrow \{\}$, and
  list of disjoint gates $\mathcal{U}_{\text{disj}} \leftarrow \{\}$
for $U_i \in \mathcal{U}$ do
    Set target gate $U_{\text{target}} \leftarrow U_i$
    for $V \in \mathcal{U}_{\text{disj}}$ do
        if $U_{\text{target}}$ shares a wire with $V$ then
          Remove $V$ from $\mathcal{U}_{\text{disj}}$.
            if $\text{rank}(U_{\text{target}}V) > d^n$ then
                Add $V$ to $\mathcal{U}_{\text{merged}}$.
            else if $\text{rank}(U_{\text{target}}V) \le d^n$ then
                $U_{\text{target}} \leftarrow U_{\text{target}}V$.
    Add $U_{\text{target}}$ to $\mathcal{U}_{\text{disj}}$.
for $U_i \in \mathcal{U}_{\text{merged}}$ do
    Set target gate $U_{\text{target}} \leftarrow U_i$
    for $V \in \mathcal{U}_{\text{disj}}$ do
        if $\text{rank}(U_{\text{target}}V) \le d^n$ then
                $U_{\text{target}} \leftarrow U_{\text{target}}V$.
    Add $U_{\text{target}}$ to $\mathcal{U}_{\text{disj}}$.
Append $\mathcal{U}_{\text{disj}}$ to $\mathcal{U}_{\text{merged}}$.
Output: $\mathcal{U}_{\text{merged}}$.
```

At every step, a target gate U_{target}, the algorithm searches through the disjoint gates to find the next one that is connected to U_{target}. We therefore require to search less than L gates for every U_{target}, while the cost of merging two gates (i.e., multiplying) is $O(2^{2n})$, a constant as we fix $n < N$. So the full gate merging sub-routine scales as $O(2^{2n}L^2)$. The computational cost to compute the transition matrix $W_U^{\mathcal{G}}$ for n-qubit unitary U and its negativity also exponentially scales with n as there are $\mathcal{O}(2^{2n})$ possible phase space points for a n-qubit system.

As we can observe from the scaling, the limiting factor of gate merging is the spatial parameter n, which stems from the exponential growth of the dimension of Hilbert space by increasing the number of qubits. We find numerically that a practical choice for the spatial parameter n is $n \le 5$. As this is a fundamental property of a quantum system, a similar issue arises in the robustness measure as evaluating the robustness of $\mathcal{R}(|T\rangle^{\otimes n})$ and finding its optimal decomposition among $\mathcal{O}(2^{n^2})$ stabilizer states is in general a challenging task for a large n [59].

Due to the computational need to truncate the spatial parameter $n < N$, a question arises of whether there exist new methods of manipulating the circuit frames and further reducing the total negativity, after gate merging is completed. We provide a positive answer to this question in the following section, where we describe our second sub-routine, frame optimisation.

5 Frame optimisation

Frame optimisation aims to reduce the total circuit negativity by optimally choosing frames for different circuit components. As we discussed in Section 3.1, we can introduce specific frame parametrisations, such as parametrised Wigner frames or rotated Pauli frames, and iteratively choose the frames throughout the circuit.

Sub-routine 2 Frame optimisation

```
   Input: Quantum circuit $C$ and temporal parameter $\ell$.
1: Determine the total number of frames, $|\boldsymbol{\mathcal{G}}_{\text{opt}}|$, in the circuit $C$.
2: Define the set of reference frames,
     $\boldsymbol{\mathcal{G}}_{\text{opt}} \leftarrow \{\mathcal{G}_1, \ldots, \mathcal{G}_{|\boldsymbol{\mathcal{G}}_{\text{opt}}|}\}$.
3: Fix the number of optimisation cycles $c$.
4: for $i = 1, \ldots, c$ do
5:     Choose a subset $\boldsymbol{\mathcal{G}}^{(i)}_{\text{target}} \subset \boldsymbol{\mathcal{G}}_{\text{opt}}$ with at most $\ell$ frames.
6:     Find a circuit block $\boldsymbol{B}$ containing the frames in $\boldsymbol{\mathcal{G}}^{(i)}_{\text{target}}$.
7:     Find $\overline{\boldsymbol{\mathcal{G}}}^{(i)}_{\text{target}} = \text{argmin}_{\boldsymbol{\mathcal{G}}^{(i)}_{\text{target}}} N_{\boldsymbol{B}}\left(\boldsymbol{\mathcal{G}}^{(i)}_{\text{target}}\right)$.
8:     Update the corresponding frames in $\boldsymbol{\mathcal{G}}_{\text{opt}}$ with $\overline{\boldsymbol{\mathcal{G}}}^{(i)}_{\text{target}}$.
   Output: $\boldsymbol{\mathcal{G}}_{\text{opt}}$.
```

In principle, the best strategy in terms of achieving the highest negativity reduction would be to carry out global optimisation over all circuit frames, requiring that the number of parameters to be optimised should scale with the number of qubits N and circuit length L. In this work, we show that a *local* optimisation, with only a fixed number of parameters, is sufficient to achieve considerable negativity reduction and scales only linearly in N and L. This optimisation sub-routine is implemented by dividing the circuit into blocks containing at most ℓ frames to be optimised, for a fixed temporal parameter ℓ.

To perform the frame optimisation on a quantum circuit C consisting of an input state ρ, a gate sequence $\{U_1, ..., U_L\}$ and a measurement effect E, we need to start from an initial frame parametrisation. We denote this parametrisation as $\boldsymbol{\mathcal{G}}_{\text{opt}} = \{\mathcal{G}_1, \ldots, \mathcal{G}_{|\boldsymbol{\mathcal{G}}_{\text{opt}}|}\}$, where $|\boldsymbol{\mathcal{G}}_{\text{opt}}|$ is the number of frames to be optimised. The procedure is outline in Sub-routine 2 and explained here. We take a subset $\boldsymbol{\mathcal{G}}^{(1)}_{\text{target}} \subset \boldsymbol{\mathcal{G}}_{\text{opt}}$ with up to ℓ frames (either sequentially or randomly) and create the block $\mathbf{B}$ of circuit components which are attached to those ℓ frames. Keeping all other frames in the block $\mathbf{B}$ fixed with the corresponding frames in $\boldsymbol{\mathcal{G}}_{\text{opt}}$, we want to minimise the total negativity of the block $N_{\mathbf{B}}$ over all possible choices for $\boldsymbol{\mathcal{G}}_{\text{target}}$, so that the minimum

$$\min_{\boldsymbol{\mathcal{G}}_{\text{target}}} N_{\mathbf{B}}(\boldsymbol{\mathcal{G}}_{\text{target}}). \tag{38}$$

occurs at $\overline{\boldsymbol{\mathcal{G}}}^{(1)}_{\text{target}}$ allowing us to update the corresponding frames in $\boldsymbol{\mathcal{G}}_{\text{opt}}$, which is the end of the first cycle in our frame optimisation. We repeat this process c times by choosing another set of ℓ frames as the new $\boldsymbol{\mathcal{G}}^{(i)}_{\text{target}}$, $i = 1, \ldots, c$. The number of optimisation cycles c can be chosen arbitrarily, for example it can be chosen as $c \geq |\boldsymbol{\mathcal{G}}_{\text{opt}}|/\ell$, with the aim of optimising all frames in the circuit at least once. The order in which frames are optimised can also be chosen arbitrarily and can potentially result in a different overall negativity reduction.

We demonstrate the local frame optimisation method with an example. Let us consider the initial part of a simple general circuit depicted in Fig. 3 for the case of $n = 2$ and $\ell = 2$. To perform the (i)-th optimisation cycle we consider $\boldsymbol{\mathcal{G}}^{(i)}_{\text{target}} = \{\mathcal{G}_1, \mathcal{G}_2\}$, we consider the corresponding block $\mathbf{B}_1 = \{\rho_1, \rho_2, U_1, U_2\}$, which is a set of all circuit components connected to the frames in $\boldsymbol{\mathcal{G}}^{(i)}_{\text{target}}$. Then, the explicit optimisation we perform is

$$\begin{aligned}&\min_{\boldsymbol{\mathcal{G}}^{(i)}_{\text{target}}} N_{\mathbf{B}_1}\left(\boldsymbol{\mathcal{G}}^{(i)}_{\text{target}}\right) = \\ &\min_{\{\mathcal{G}_1, \mathcal{G}_2\}} N_{\rho_1}(\mathcal{G}_1) N_{\rho_2}(\mathcal{G}_2) N_{U_1}(\mathcal{G}_2) N_{U_2}(\mathcal{G}_1),\end{aligned} \tag{39}$$

where $N_X(\mathcal{G}_X)$ is the negativity of component X as a function of $\mathcal{G}_X$ with all other frames

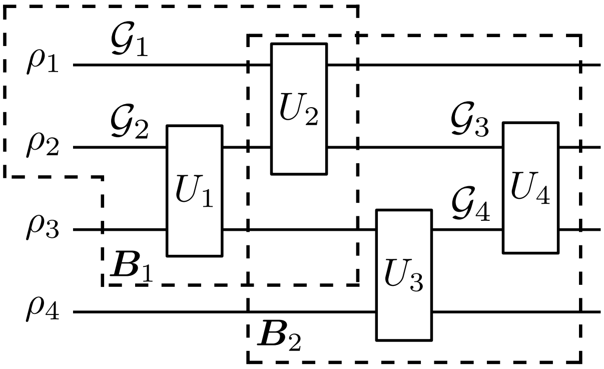

Figure 3: Example of how to form a block when $\boldsymbol{\mathcal{G}}_{\text{target}}$ is given in the case of $n = 2$ and $\ell = 2$. Only relevant frames are shown. When $\boldsymbol{\mathcal{G}}_{\text{target}} = \{\mathcal{G}_1, \mathcal{G}_2\}$, the corresponding block $\mathbf{B}_1$, which contains all circuit elements connected to the frames in $\boldsymbol{\mathcal{G}}_{\text{target}}$, is $\mathbf{B}_1 = \{\rho_1, \rho_2, U_1, U_2\}$. When $\boldsymbol{\mathcal{G}}_{\text{target}} = \{\mathcal{G}_3, \mathcal{G}_4\}$, then the corresponding block is $\mathbf{B}_2 = \{U_2, U_3, U_4\}$.

fixed to the corresponding ones in $\boldsymbol{\mathcal{G}}_{\text{opt}}$. As an additional example, we could have considered the set $\boldsymbol{\mathcal{G}}^{(i)}_{\text{target}} = \{\mathcal{G}_3, \mathcal{G}_4\}$ corresponding to the block $\mathbf{B}_2 = \{U_2, U_3, U_4\}$ in Fig. 3. Then, the block negativity we optimise is $N_{\mathbf{B}_2}(\boldsymbol{\mathcal{G}}^{(i)}_{\text{target}}) = N_{U_2}(\mathcal{G}_3)N_{U_3}(\mathcal{G}_4)N_{U_4}(\mathcal{G}_3, \mathcal{G}_4)$.

Note that at each optimisation step, previously optimised frames in $\boldsymbol{\mathcal{G}}_{\text{opt}}$ are used in the next optimisation cycle. This ensures that the negativity never increases compared to the initial frame choice $\{\mathcal{G}_1, \ldots, \mathcal{G}_{|\boldsymbol{\mathcal{G}}_{\text{opt}}|}\}$ between optimisation cycles.

The presented local optimisation method is efficient in the number of circuit components. Consider an N-qubit circuit of length L where each of L gates acts on at most n qubits. Then, there are at most $N + nL$ different frames to be optimised. Since ℓ is fixed, each optimisation cycle takes a constant amount of time $O(1)$. Therefore, the frame optimisation of the entire circuit scales as $(N + nL) \times O(1) = O(N, L)$. Note that the exact value depends on truncation parameters n and ℓ as well as the specifics of the circuit and its frame parametrisation.

Fig. 4 shows the performance of the frame optimisation for a circuit with $N = 6$ and $L = 15$ consisting of 2-qubit Haar-random unitaries, which are in general difficult to be simulated with stabiliser-based simulators because they do not admit efficient decompositions. In Fig. 4(a), we use rotated Pauli frames as our frame parametrisation and initialised each frame in the circuit to the set of standard qubit Pauli operators. In Fig. 4(b), we choose parametrised Wigner frames as our frame parametrisation and initialise each frame in the circuit to the set of conventional phase-space operators corresponding to $g(\boldsymbol{\lambda}) = 1$ (see Sec. 3.2). We can observe that the largest negativity reduction comes from gate merging with higher n, but the frame optimisation also achieves a significant negativity reduction. In general, larger ℓ results in lower negativity after optimisation of all frames with fixed n. In the case of parametrised Wigner frames, together with gate merging, we could considerably decrease the initial log-negativity from ∼27.3 to ∼8.9 with truncation parameters $n = 4$ and

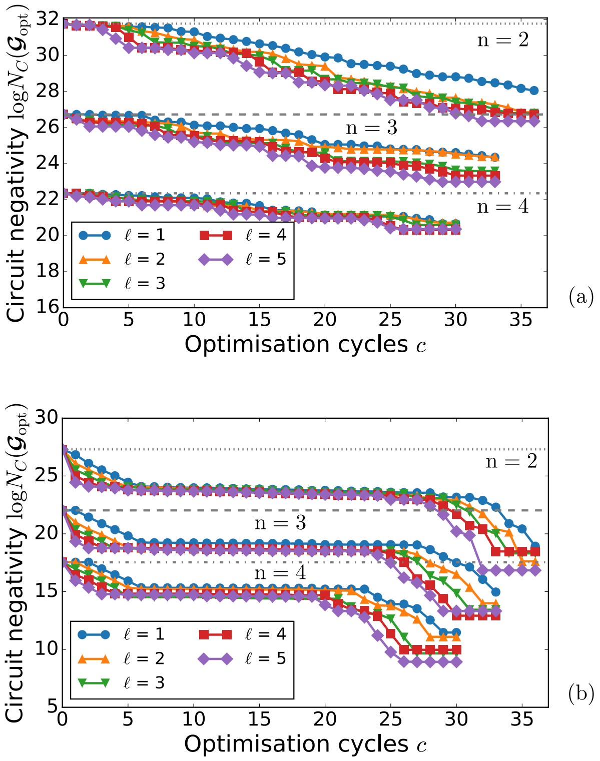

Figure 4: Plots showing negativity reduction of a circuit consisting of 2-qubit Haar-random gates with $N = 6$ and $L = 15$ after each frame optimisation cycle with different spatial and temporal parameters, n and ℓ. The optimisation is carried out sequentially from the first frame to the last frame. Optimisation is performed via the basin-hopping algorithm as introduced in [65]. (a) Results after frame optimisation with rotated Pauli frames. The reference frame is the standard Pauli operators. The total negativity continuously decreases as we optimise more frames. (b) Results after frame optimisation with parametrised Wigner frames. The reference frame is the conventional phase-space operators for the Wigner function. The most of negativity reduction occurs near the initial states and the measurements.

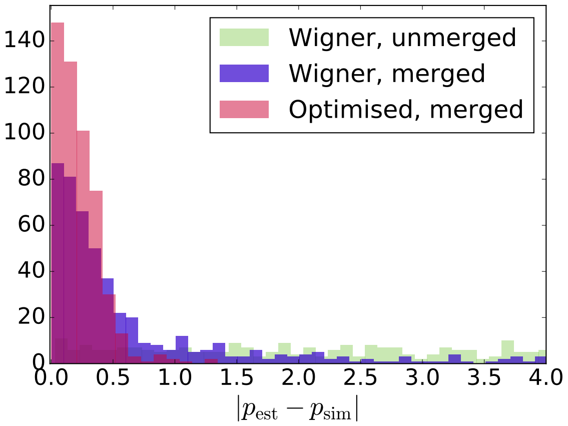

Figure 5: Histograms of the deviation of estimated probability p_{est} from actual outcome probability p_{sim} (as calculated by Qiskit [66]) for 500 circuits consisting of 2-qubit Haar-random gates with $N = 3$, $L = 8$ and $\ell = 1$. The number of samples taken for each circuit is 10^6, which took around 10 seconds on a standard computer. The plot shown is truncated at $|p_{\mathrm{est}} - p_{\mathrm{sim}}| = 4.0$ to demonstrate the advantage of our routines clearly. The advantage is amplified as N and L increase.

$\ell = 5$, which means that we need $\sim 2^{2\times 18}$ times less samples to reach a given accuracy for probability estimation.

We demonstrate the practical significance of our routine, by sampling 500 circuits consisting of Haar-random gates, with the results presented in Fig. 5. Unmerged circuits represented entirely by Wigner frames do not show any signs of convergence to the actual probability distribution. Merged circuits clearly converge a lot better, especially when their frame representation is optimised.

6 Conclusion

We introduce two classical sub-routines, gate merging and frame parametrisation, which reduce the total negativity in the quasi-probability representation of a quantum circuit, hence leading to sampling overhead reduction. We emphasise that our methods are very general; they are not restricted to specific choices of frames or frame parametrisations, and can be applicable to any circuit independently of generating gate sets or the purity and dimension of its input qudits. Both sub-routines are efficient in the sense that the runtime scales polynomially in the circuit size N and number of gates L.

We numerically demonstrate that both methods improve the exponential scaling of the circuit negativity by testing them on Clifford+T circuits and circuits with Haar-random gates. Specifically, gate merging is shown to compete on average with the quasi-probability simulators based on dyadic frames and the robustness of magic. Frame optimisation can further compliment gate merging in reducing negativity, when merging gates in the circuits is no longer practical due to the growing size of the gates.

A clear direction for our work is to improve the classical optimisation performed for the frame representation. Our parametrisation resembles variational techniques used in near-term quantum algorithms [67], although our cost function, circuit negativity, is calculated classically. Our optimisation could therefore potentially benefit by research on variational techniques, such as identifying "good" circuit-inspired ansatze for initialising frames or investigating barren plateaus in order to improve optimisation convergence. Such methods could shed light on what families of circuits are hardest to sample from using quasi-probability techniques.

One can also investigate the possibility of performing frame optimisation analytically, at least for particular classes of quantum circuits. Additional assumptions will likely be required for the circuit structure, but finding optimal frames analytically would eliminate the hidden constant runtime costs of "black-box" classical algorithms currently employed for the optimisation. For example, it would be particularly useful to investigate the existence of a finite set of frames as a function of circuit components resulting in minimum negativity for Clifford+T circuits.

Acknowledgements

This work is supported by the KIST Open Research Program and the UK Hub in Quantum Computing and Simulation, part of the UK National Quantum Technologies Programme with funding from UKRI EPSRC grant EP/T001062/1. NK is supported by the EPSRC Centre for Doctoral Training on Controlled Quantum Dynamics. HK is supported by the KIAS Individual Grant No. CG085301 at Korea Institute for Advanced Study. HHJ is supported by the Centre for Doctoral Training on

Controlled Quantum Dynamics funded by the EPSRC (EP/L016524/1). DJ is supported by the Royal Society and a University Academic Fellowship. MSK acknowledges the Samsung GRC grant.

Code availability

The code that implements the sub-routines and supports the presented numerical findings can be accessed at NK's GitHub repository: `https://github.com/nkoukou/parameterised_negativity`.

References

[1] R. P. Feynman, International Journal of Theoretical Physics **21**, 467 (1982).

[2] J. Preskill, "Quantum computing 40 years later," (2021), arXiv:2106.10522 .

[3] R. Jozsa and M. V. den Nest, Quantum Inf. Comput. **14**, 633 (2014).

[4] D. E. Koh, Quantum Info. Comput. **17**, 262–282 (2017).

[5] S. Aaronson, A. Bouland, G. Kuperberg, and S. Mehraban (Association for Computing Machinery, New York, NY, USA, 2017) p. 317–327.

[6] M. Hebenstreit, R. Jozsa, B. Kraus, and S. Strelchuk, Phys. Rev. A **102**, 052604 (2020).

[7] M. Yoganathan, R. Jozsa, and S. Strelchuk, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences **475**, 20180427 (2019).

[8] S. Aaronson and A. Arkhipov, "The computational complexity of linear optics," (2010), arXiv:1011.3245 .

[9] M. J. Bremner, R. Jozsa, and D. J. Shepherd, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences **467**, 459 (2011).

[10] T. Morimae, K. Fujii, and J. F. Fitzsimons, Phys. Rev. Lett. **112**, 130502 (2014).

[11] M. J. Bremner, A. Montanaro, and D. J. Shepherd, Phys. Rev. Lett. **117**, 080501 (2016).

[12] X. Gao, S.-T. Wang, and L.-M. Duan, Phys. Rev. Lett. **118**, 040502 (2017).

[13] J. Bermejo-Vega, D. Hangleiter, M. Schwarz, R. Raussendorf, and J. Eisert, Phys. Rev. X **8**, 021010 (2018).

[14] K. Fujii, H. Kobayashi, T. Morimae, H. Nishimura, S. Tamate, and S. Tani, Phys. Rev. Lett. **120**, 200502 (2018).

[15] A. Bouland, J. F. Fitzsimons, and D. E. Koh (Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, DEU, 2018).

[16] S. Boixo, S. V. Isakov, V. N. Smelyanskiy, R. Babbush, N. Ding, Z. Jiang, M. J. Bremner, J. M. Martinis, and H. Neven, Nature Physics **14**, 595 (2018).

[17] H. Pashayan, S. D. Bartlett, and D. Gross, Quantum **4**, 223 (2020).

[18] L. G. Valiant, SIAM Journal on Computing **31**, 1229 (2002).

[19] B. M. Terhal and D. P. DiVincenzo, Phys. Rev. A **65**, 032325 (2002).

[20] S. D. Bartlett, B. C. Sanders, S. L. Braunstein, and K. Nemoto, Phys. Rev. Lett. **88**, 097904 (2002).

[21] S. Aaronson and D. Gottesman, Phys. Rev. A **70**, 052328 (2004).

[22] D. Gross, S. T. Flammia, and J. Eisert, Phys. Rev. Lett. **102**, 190501 (2009).

[23] D. J. Brod, Phys. Rev. A **93**, 062332 (2016).

[24] T. Haug and M. S. Kim, "Scalable measures of magic for quantum computers," (2022).

[25] H.-S. Zhong, H. Wang, Y.-H. Deng, M.-C. Chen, L.-C. Peng, Y.-H. Luo, J. Qin, D. Wu, X. Ding, Y. Hu, and et al., Science **370**, 1460–1463 (2020).

[26] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, D. A. Buell, B. Burkett, Y. Chen, Z. Chen, B. Chiaro, R. Collins, W. Courtney, A. Dunsworth, E. Farhi, B. Foxen, A. Fowler, C. Gidney, M. Giustina, R. Graff, K. Guerin, S. Habegger, M. P. Harrigan, M. J. Hartmann, A. Ho, M. Hoffmann, T. Huang, T. S. Humble, S. V. Isakov, E. Jeffrey, Z. Jiang, D. Kafri, K. Kechedzhi, J. Kelly, P. V. Klimov, S. Knysh, A. Korotkov, F. Kostritsa, D. Landhuis, M. Lindmark, E. Lucero, D. Lyakh, S. Mandrà, J. R. McClean, M. McEwen, A. Megrant, X. Mi, K. Michielsen, M. Mohseni, J. Mutus, O. Naaman, M. Neeley, C. Neill, M. Y.

Niu, E. Ostby, A. Petukhov, J. C. Platt, C. Quintana, E. G. Rieffel, P. Roushan, N. C. Rubin, D. Sank, K. J. Satzinger, V. Smelyanskiy, K. J. Sung, M. D. Trevithick, A. Vainsencher, B. Villalonga, T. White, Z. J. Yao, P. Yeh, A. Zalcman, H. Neven, and J. M. Martinis, Nature **574**, 505 (2019).

[27] H. Bernien, S. Schwartz, A. Keesling, H. Levine, A. Omran, H. Pichler, S. Choi, A. S. Zibrov, M. Endres, M. Greiner, and et al., Nature **551**, 579–584 (2017).

[28] C. Neill, P. Roushan, K. Kechedzhi, S. Boixo, S. V. Isakov, V. Smelyanskiy, A. Megrant, B. Chiaro, A. Dunsworth, K. Arya, and et al., Science **360**, 195–199 (2018).

[29] J. P. Bonilla Ataides, D. K. Tuckett, S. D. Bartlett, S. T. Flammia, and B. J. Brown, Nature Communications **12**, 2172 (2021).

[30] S. Bravyi, S. Sheldon, A. Kandala, D. C. Mckay, and J. M. Gambetta, Phys. Rev. A **103**, 042605 (2021).

[31] A. Kandala, K. Temme, A. D. Córcoles, A. Mezzacapo, J. M. Chow, and J. M. Gambetta, Nature **567**, 491 (2019).

[32] S. Endo, S. C. Benjamin, and Y. Li, Phys. Rev. X **8**, 031027 (2018).

[33] K. Temme, S. Bravyi, and J. M. Gambetta, Phys. Rev. Lett. **119**, 180509 (2017).

[34] J. Preskill, Quantum **2**, 79 (2018).

[35] A. W. Harrow and A. Montanaro, Nature **549**, 203 (2017).

[36] R. S. Bennink, E. M. Ferragut, T. S. Humble, J. A. Laska, J. J. Nutaro, M. G. Pleszkoch, and R. C. Pooser, Phys. Rev. A **95**, 062337 (2017).

[37] V. Veitch, C. Ferrie, D. Gross, and J. Emerson, New Journal of Physics **14**, 113011 (2012).

[38] D. Gottesman, in *Encyclopedia of Mathematical Physics*, edited by J.-P. Françoise, G. L. Naber, and T. S. Tsun (Academic Press, Oxford, 2006) pp. 196 – 201.

[39] H. Pashayan, O. Reardon-Smith, K. Korzekwa, and S. D. Bartlett, "Fast estimation of outcome probabilities for quantum circuits," (2021), arXiv:2101.12223 .

[40] J. R. Seddon, B. Regula, H. Pashayan, Y. Ouyang, and E. T. Campbell, PRX Quantum **2**, 010345 (2021).

[41] S. Bravyi and D. Gosset, Phys. Rev. Lett. **116**, 250501 (2016).

[42] S. Bravyi, D. Browne, P. Calpin, E. Campbell, D. Gosset, and M. Howard, Quantum **3**, 181 (2019).

[43] S. Bravyi, G. Smith, and J. A. Smolin, Phys. Rev. X **6**, 021043 (2016).

[44] H. Qassim, J. J. Wallman, and J. Emerson, Quantum **3**, 170 (2019).

[45] Y. Huang and P. Love, Phys. Rev. A **99**, 052307 (2019).

[46] L. Kocia and M. Sarovar, Phys. Rev. A **103**, 022603 (2021).

[47] A. Kissinger, J. van de Wetering, and R. Vilmart, "Classical simulation of quantum circuits with partial and graphical stabiliser decompositions," (2022), arXiv:2202.09202 .

[48] H. Qassim, H. Pashayan, and D. Gosset, Quantum **5**, 606 (2021).

[49] H. Pashayan, J. J. Wallman, and S. D. Bartlett, Phys. Rev. Lett. **115**, 070501 (2015).

[50] D. Stahlke, Phys. Rev. A **90**, 022302 (2014).

[51] P. Rall, D. Liang, J. Cook, and W. Kretschmer, Phys. Rev. A **99**, 062337 (2019).

[52] J. R. Seddon and E. T. Campbell, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences **475**, 20190251 (2019).

[53] C. Ferrie and J. Emerson, Journal of Physics A: Mathematical and Theoretical **41**, 352001 (2008).

[54] C. Ferrie and J. Emerson, New Journal of Physics **11**, 063040 (2009).

[55] D. Gross, Journal of Mathematical Physics **47**, 122107 (2006).

[56] M. Ruzzi, M. A. Marchiolli, and D. Galetti, Journal of Physics A: Mathematical and General **38**, 6239 (2005).

[57] M. A. Marchiolli, M. Ruzzi, and D. Galetti, Phys. Rev. A **72**, 042308 (2005).

[58] D. S. França, S. Strelchuk, and M. Studziński, Phys. Rev. Lett. **126**, 210502 (2021).

[59] M. Howard and E. Campbell, Phys. Rev. Lett. **118**, 090501 (2017).

[60] M. Heinrich and D. Gross, Quantum **3**, 132 (2019).

[61] S. Rahimi-Keshari, T. C. Ralph, and C. M. Caves, Phys. Rev. X **6**, 021039 (2016).

[62] A. Mari and J. Eisert, Phys. Rev. Lett. **109**, 230503 (2012).

[63] R. Raussendorf, D. E. Browne, N. Delfosse, C. Okay, and J. Bermejo-Vega, Physical Review A **95**, 052334 (2017).

[64] C. Jones, Phys. Rev. A **87**, 022328 (2013).

[65] D. J. Wales and J. P. K. Doye, The Journal of Physical Chemistry A **101**, 5111–5116 (1997).

[66] M. S. A. et al., "Qiskit: An open-source framework for quantum computing," (2021).

[67] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles, Nature Reviews Physics **3**, 625 (2021).