Quantum obfuscation

Gorjan Alagic and Bill Fefferman

September 10, 2015

Contents

1 Introduction

1.1 Background

1.2 Summary of results

1.2.1 Quantum black-box obfuscation

1. define quantum black-box obfuscation
2. show several applications:
   - classical algorithm for it implies quantum-secure OWFs;
   - quantum obfuscator implies IND-CPA SKE for quantum states;
   - obfuscator + qOWFs implies IND-CPA PKE for quantum states;
   - obfuscator (plus what else?) implies QFHE;
   - obfuscator implies quantum money (details?);
3. impossibility results:
   - two-state black-box obfuscation impossible
   - obfuscation with cloneable outputs impossible

1.2.2 Quantum indistinguishability obfuscation

1. define quantum indistinguishability obfuscation;
2. define quantum best-possible obfuscation;
3. three variants: perfect, statistical, computational;
4. (Gorjan: can we give any applications? Even simple ones, like patching?)
5. proved each indistinguishability variant is equivalent to its corresponding best-possible variant;
6. proved impossibility of perfect and statistical indistinguishability obfuscators.

2 Preliminaries

2.1 Notation etc.

2.2 Probabilistic and quantum algorithms

We briefly review some terminology regarding probabilistic and quantum algorithms. For precise definitions, refer to [**?**]. As is standard, by a probabilistic classical algorithm $\mathcal{A}$ we will mean an infinite family of probabilistic classical circuits, at least one for each possible input size [1]. When the input register is initialized with the string x and the randomness register is initialized with the string r, the output of the relevant circuit will be denoted by $\mathcal{A}(x;r)$. We will simply write $\mathcal{A}(x)$ when the randomness register should be initialized with a uniformly random string.

A quantum algorithm $\mathcal{Q}$ will mean an infinite family of quantum circuits, at least one for each possible input size. For each circuit, the qubits it acts on are divided into an input register and an ancilla register; the former is initialized in some input state σ and the latter is always initialized in the $|0\rangle$ state. All of the qubits are also divided into an output register and a garbage register; it is always assumed that the garbage register is traced out after the circuit is applied. The (possibly mixed) state which remains in the output register is called the output of the algorithm, and is denoted $\mathcal{Q}(\sigma)$.

We will sometimes also allow for algorithms which are allowed to mix probabilistic and quantum computation in a straightforward way: a classical probabilistic circuit first uses a string of classical randomness to decide which quantum circuit to run on the given quantum input state, and the chosen quantum circuit is then executed. We will call such algorithms probabilistic-quantum and refer to output states with or without specified randomness as above. The computational power of such algorithms can already be captured by quantum algorithms alone by reversibly implementing the classical pre-processing; the classical probabilistic mixtures are then absorbed into the density operator of the quantum state as it evolves under the quantum circuit. However, the distinction does have a difference: the final density operator outputted by the resulting quantum algorithm could exhibit some important property which is *not* true for any of the outputs of the original probabilistic-quantum algorithm. We will discuss an explicit example later.

An algorithm will be referred to as polynomial-time (or efficient) if all of the relevant circuit families are polynomial-time uniform [**?**]; this applies to all classes of algorithm discussed above.

2.3 Some primitives

Definition 1. *A function $f : \{0,1\}^* \to \{0,1\}^*$ is a quantum-secure one-way function if*

1. *f can be computed exactly by a polynomial-time classical algorithm, and*
2. *for every quantum algorithm $\mathcal{A}$, every polynomial p, and all sufficiently large n,*

$$\Pr\left[\mathcal{A}(f(x), 1^n) \in f^{-1}(f(x))\right] < \frac{1}{p(n)} ,$$

where the probability is taken over uniformly random $x \in \{0,1\}^n$ and the measurement of $\mathcal{A}$.

[1] when there's more than one circuit for a given input size, there should be some efficient way to decide which inputs of that size are assigned to which circuit.

2.4 Encryption of quantum states

(Gorjan: Move the discussions of quantum encryption in here)

3 Black-box obfuscation

3.1 Definitions

We now define the notion of an interpreter, which is simply a quantum algorithm equipped with some additional data.

Definition 2. *An **interpreter** is a polynomial-time uniform family of unitary quantum circuits $\mathcal{J} = \{J_{n,m}\}_{n,m\in\mathbb{N}}$, such that for every n and m, $J_{n,m}$ has an n-qubit register A, an m-qubit register B, and an ancilla C of size $poly(n,m)$. For every $n \in \mathbb{N}$ and every m-qubit state ρ, we define a superoperator on A by*

$$\mathcal{J}_n^\rho : \sigma \mapsto \mathrm{Tr}_{BC}\left(J_{n,m}\left[\sigma \otimes \rho \otimes |0\rangle\langle 0|_C\right] J_{n,m}^\dagger\right).$$

In applications, we think of $\mathcal{J}$ as enabling an end-user to apply a superoperator to an input state σ with the help of the m-qubit "advice" state ρ, presumably provided by some other party. We say that the state ρ implements the operator $\mathcal{J}_n^\rho$. A simple example is given by universal circuits: in this case, ρ is a classical description of a quantum circuit for implementing $\mathcal{J}_n^\rho$, and $J_{n,m}$ consists of a universal sequence of gates which are applied to the first register and controlled by the second register. While this is not explicitly required by the definition, all advice states in this work will be efficiently preparable.

We now wish to consider obfuscated advice. Given a (potentially non-unitary) quantum circuit C, let U_C denote the superoperator implemented by C. We will frequently refer to "quantum adversaries" and "quantum simulators." Both will mean a polynomial-time quantum algorithm which accepts a quantum state as input (along with a polynomial-size initialized ancilla) and outputs a classical bit. We will sometimes make use of quantum simulators which have oracle access to some unitary operator; such a simulator will be denoted by, e.g., $\mathcal{S}^U$. The quantum circuits of such a simulator are allowed to make use of a "black box" gate which applies U. Each use of a black box counts towards the length of the circuit, which must remain polynomial in the input size.

Definition 3. *A **black-box quantum obfuscator** is a pair $(\mathcal{J}, \mathcal{O})$ where $\mathcal{J}$ is an interpreter and $\mathcal{O}$ is a probabilistic-quantum algorithm which, on input an n-qubit quantum circuit C, outputs an m-qubit quantum state $\mathcal{O}(C)$ satisfying*

1. *(polynomial slowdown) $m = poly(n, |C|)$;*
2. *(functional equivalence) there exists a negligible ε_1 such that $\left\|\mathcal{J}_n^{\mathcal{O}(C)} - U_C\right\|_\diamond \leq \varepsilon_1(n, |C|)$;*
3. *(virtual black-box) for every quantum adversary $\mathcal{A}$ there exists a quantum simulator $\mathcal{S}^{U_C}$ and a negligible ε_2 such that*

$$\left|\Pr[\mathcal{A}(\mathcal{O}(C)) = 1] - \Pr\left[\mathcal{S}^{U_C}\left(|0\rangle^{\otimes|C|}\right) = 1\right]\right| \leq \varepsilon_2(n, |C|).$$

We will select the functions $\varepsilon_1, \varepsilon_2$ later, to be the largest possible for which our impossibility proofs still work.

Note that allowing $\mathcal{O}$ to be probabilistic-quantum allows it to output different states $\mathcal{O}(C;r)$ depending on the choice of classical randomness r. Why is this different from just letting $\mathcal{O}$ be fully quantum? Suppose that we were to change the above definition to that effect, and consider the case where C is used to flip a single bit. The obfuscator could then, with equal probability, output either a state which always outputs 1 or a state which always outputs 0. Because we combined this choice into a density matrix, the functional equivalence condition would technically be satisfied. On the other hand, the actual output state of the algorithm would *never* satisfy functional equivalence – a frustrating situation for the end-user, to say the least.

(Gorjan: New work starts here. I begin with another attempt at a definition, this time without explicitly defining an interpreter.)

For shorthand, we will denote CPTP maps by, e.g., $\mathcal{A}(\cdot)$, which is to mean the map $\rho \mapsto \mathcal{A}(\rho)$. We also use "QPT" as shorthand for quantum polynomial-time algorithm, and "negl" as shorthand for some function which grows inverse-superpolynomially.

Definition 4. *A **black-box quantum obfuscator** is a pair of QPTs $(\mathcal{J}, \mathcal{O})$ such that whenever C is an n-qubit quantum circuit, the output of $\mathcal{O}$ is an m-qubit state $\mathcal{O}(C)$ satisfying*

1. *(polynomial slowdown) $m = poly(n, |C|)$;*
2. *(functional equivalence) $\left\| \mathcal{J}(\mathcal{O}(C) \otimes \cdot) - C \cdot C^\dagger \right\|_\diamond \leq \mathrm{negl}(n, |C|)$;*
3. *(virtual black-box) for every QPT adversary $\mathcal{A}$ there exists a QPT simulator $\mathcal{S}^{U_C}$ such that*

$$\left| \Pr[\mathcal{A}(\mathcal{O}(C)) = 1] - \Pr\left[\mathcal{S}^{U_C}\left(|0\rangle^{\otimes |C|}\right) = 1\right] \right| \leq \mathrm{negl}(n, |C|)\,.$$

(Gorjan: Here should discuss alternative definition (where obfuscator provides a state as well as a circuit to run), and prove that since we are only interested in existence, the two definitions are equivalent.)

3.2 Applications

3.2.1 Quantum-secure one-way functions

Proposition 1. *If there exists a classical probabilistic algorithm which is a quantum black-box obfuscator, then quantum-secure one-way functions exist.*

Proof. Essentially just as in Barak. □

The problem of constructing cryptographically useful primitives from a fully quantum obfuscator is open. This ties in nicely with the question about quantum-secure primitives from quantum encryption in the other paper.

3.2.2 Plaintext-secure private-key quantum encryption

In order to talk about some further applications, we will need to discuss some basics about *quantum encryption schemes.* These are schemes for encrypting quantum states under (in general) computational assumptions. The details will appear in another forthcoming paper. It suffices for now to say that these are schemes which consist of: (Gorjan: should define this formally, since we'll reuse it.)

- a key generation algorithm KeyGen which produces classical bitstrings k as keys;
- a QPT encryption algorithm Enc_k;
- a QPT decryption algorithm Dec_k;

such that encryption followed by decryption is (at least approximately) equivalent to the identity. One can then define suitable notions of indistinguishability of ciphertexts under chosen plaintext attacks (IND-CPA) and under non-adaptive chosen ciphertext attacks (IND-CCA1). See Anne and Stacey's paper, as well as our forthcoming paper.

Proposition 2. *If quantum black-box obfuscators exist, then so do IND-CPA-secure symmetric-key quantum encryption schemes.*

Proof. Waiting for Bill to look this up in his notes, I forgot the construction. □

3.2.3 Public-key encryption from private-key encryption

We first show the following lemma. It also appears in contemporaneous work (cite other paper.)

Lemma 1. *If quantum-secure one-way functions exist, then IND-CCA1-secure symmetric-key quantum encryption schemes exist.*

Proof. Sketch is as follows, details will appear in other paper.

1. qOWFs imply qPRFs (Zhandry);
2. using a qPRF, define encryption scheme like this: (i.) KeyGen produces random bitstrings k which index a qPRF f_k; (ii.) To encrypt an n-qubit state ρ, pick classical randomness $r \in \{0,1\}^{2n}$, and perform
$$\rho \longmapsto |r\rangle\langle r| \otimes P_{f_k(r)} \rho P_{f_k(r)}^\dagger ,$$
where P_r denotes the element of the n-qubit Pauli group indexed by r; (iii.) To decrypt, evaluate the qPRF at the randomness, and undo the Paulis.
3. to show IND-CCA1 security, prove that a polynomial number of (quantum) queries to encryption and decryption oracles is no more useful than a polynomial number of (classical) evaluations $f_k(x)$ at inputs x of the adversary's choice; it's then straightforward to show that breaking the scheme implies breaking the qPRF.

□

Proposition 3. *If quantum black-box obfuscators exist, and quantum-secure one-way functions exist, then so do IND-CPA-secure public-key encryption schemes.*

Proof. Sketch is as follows. (Gorjan: This needs to be fleshed out more carefully, especially the last step.)

1. qOWFs imply qPRFs (Zhandry);
2. qPRFs imply IND-CCA1 secure encryption (Lemma 1).
3. the private key in the public scheme is the same as the key k produced by the KeyGen algorithm of the IND-CCA1 scheme;
4. the public key is $\mathcal{O}(\mathsf{Enc}_k)$, where Enc_k is the encryption circuit of the IND-CCA1 scheme;
5. by the virtual black-box property, any interaction with $\mathcal{O}(\mathsf{Enc}_k)$ can be simulated by black-box access to the encryption map, where we are allowed to input our own randomness;
6. now show that IND-CCA1 implies that the scheme is secure even if the adversary gets to provide their own randomness in the queries; IND-CPA for the public scheme follows.

□

One curious feature of the above construction is that, in general, the public keys might be quantum, and even consumable. This is an aspect of public-key encryption that does not exist classically, and might be of interest in other applications.

3.2.4 Fully homomorphic encryption from public-key encryption

3.2.5 Quantum money

3.2.6 Some old content

(Gorjan: formalize this fully; do we have classical or quantum security?) First, we show how black-box quantum obfuscation would allow one to generically turn any classical private-key encryption scheme into a public-key encryption scheme with quantum keys. Suppose we have a private-key encryption scheme which is secure against chosen plaintext attacks. We might try to turn it into a public-key scheme where the public key is an obfuscation $\mathcal{O}(\mathsf{Enc}_k)$ of the classical circuit "encrypt with the private key k." As shown by Barak et al., this is not possible if $\mathcal{O}$ is a classical algorithm. But what if we are willing to consider a public key which is a quantum state? Suppose that $\mathcal{O}$ is a black-box quantum obfuscator, and an adversary $\mathcal{A}$ came along and discovered the private key k from the quantum public key $|\mathcal{O}(\mathsf{Enc}_k)\rangle$. By the virtual black-box property, there would also be a simulator $\mathcal{S}$ which only uses Enc_k as a black box and *still* breaks the scheme. This would contradict the security of the private-key scheme against chosen plaintext attacks.

Next, we show that black-box obfuscators with classical outputs imply the existence of one-way functions secure against quantum adversaries. We first define a notion of (strong) quantum-secure one-way function, and a corresponding notion of hard-core predicate. These definitions are natural analogues of the classical definitions (see, e.g., Goldreich [**?**]). (Gorjan: do these definitions appear elsewhere already?)

Definition 5. *A function $f : \{0,1\}^* \to \{0,1\}^*$ is a quantum-secure one-way function if*

1. *f can be computed exactly by a polynomial-time classical algorithm, and*

2. *for every quantum algorithm $\mathcal{A}$, every polynomial p, and all sufficiently large n,*

$$\Pr\left[\mathcal{A}(f(x), 1^n) \in f^{-1}(f(x))\right] < \frac{1}{p(n)} ,$$

where the probability is taken over uniformly random $x \in \{0,1\}^n$ and the measurement of $\mathcal{A}$.

Definition 6. *A classically-polynomial-time computable predicate $b : \{0,1\}^* \to \{0,1\}$ is called hardcore for a function f if for every quantum algorithm $\mathcal{A}$, every polynomial p and all sufficiently large n,*

$$\Pr\left[\mathcal{A}(f(x)) = b(x)\right] < \frac{1}{2} + \frac{1}{p(n)} ,$$

where the probability is taken over uniformly random $x \in \{0,1\}^n$ and the measurement of $\mathcal{A}$.

For a discussion of (weak?) one-way functions secure against quantum, see Kashefi and Kerinidis [**?**]. (Gorjan: Their definition is different: they posit that there exists a polynomial p such that the success probability is less than $1/p$, but Goldreich says less than $1/p$ for *every* polynomial p.)

Proposition 1. *Suppose that there exists a quantum obfuscator $(\mathcal{J}, \mathcal{O})$ such that $\mathcal{O}$ is classical probabilistic. Then quantum-secure one-way functions exist.*

Proof. The proof is essentially identical to the classical proof given by Barak et al. [**?**] □

We remark that there are some straightforward variations on the above theme; for example, if the obfuscator in the assumption of Proposition 1 is quantum but always outputs classical states, then one can conclude that quantum-computable one-way functions (but still with classical inputs and outputs) secure against quantum adversaries exist.

3.3 Some impossibility results

3.3.1 Impossibility of two-circuit obfuscation

3.3.2 Impossibility of obfuscation for cloneable outputs

Classically, Barak et al. proved impossibility of generic black-box circuit obfuscation. We don't know how to prove the same fact for quantum circuits, but we can replicate the first step of their proof, namely the impossibility of black-box two-circuit obfuscation. We define a **black-box quantum two-circuit obfuscator** just as in the definition above, but with a different virtual black-box condition:

3. *(two-circuit virtual black-box) for every pair of quantum circuits C_1 and C_2 and every quantum adversary $\mathcal{A}$ there exists a quantum simulator $\mathcal{S}^{U_{C_1}, U_{C_2}}$ and a negligible ε_2 such that*

$$\left|\Pr[\mathcal{A}(\mathcal{O}(C_1) \otimes \mathcal{O}(C_2)) = 1] - \Pr\left[\mathcal{S}^{U_{C_1}, U_{C_2}}\left(|0\rangle^{\otimes |C_1|+|C_2|}\right) = 1\right]\right| \leq \varepsilon_2(n, |C|) .$$

Theorem 2. *There exist pairs of unitaries (U, V) such that no pair of quantum circuits is a two-circuit black-box obfuscation of (U, V).*

Proof. Let $(\mathcal{O}, \mathcal{J})$ be an obfuscator. Choose a uniformly random $z \in \{0,1\}^n$. Define the following $(n+1)$-qubit unitaries:

$$U_0 : |x\rangle|y\rangle \mapsto |x\rangle|\delta_z(x) \oplus y\rangle \qquad \text{and} \qquad U_1 : |x\rangle|y\rangle \mapsto |x\rangle|y\rangle\,,$$

where $\delta_z : x \mapsto \delta_{xz}$ is the delta function at z. Using standard methods for building reversible circuits, we can specify a unitary poly(n)-length circuit C_0 that implements U_0. We also easily build a circuit C_1 which implements U_1 (i.e., the identity) such that $|C_0| = |C_1|$. We also define a circuit

$$D_z : |s\rangle|t\rangle \mapsto |s\rangle|f_z(s) \oplus t\rangle$$

on two registers: a register of size $m := |\mathcal{O}(C_0)| = |\mathcal{O}(C_1)|$, and a one-qubit register (as well as an ancilla we omit for simplicity.) The circuit D_z initializes a portion of the ancilla to $|z\rangle|0\rangle$, and runs the interpreter circuit $J_{n+1,m}$ on $|z\rangle|0\rangle|s\rangle$. Finally, controlled by the second register of that computation, a CNOT is applied to $|t\rangle$.

We now describe the adversary $\mathcal{A}$, who will essentially run the second state it receives on the first one. More precisely, on input $\mathcal{O}(C_j) \otimes \mathcal{O}(D_z)$ where $j \in \{0,1\}$, the adversary will run the interpreter circuit $J_{m+1,|\mathcal{O}(D_z)|}$ on the input state $\mathcal{O}(C_j) \otimes |0\rangle\langle 0|_{\text{out}}$ with advice state $\mathcal{O}(D_z)$, and then measure the single-qubit "out" register. Prior to measurement, by the functional equivalence property of the obfuscator, the $m+1$-qubit output register of $J_{m+1,|\mathcal{O}(D_z)|}$ will be close (in trace distance) to the state $D_z(\mathcal{O}(C_j) \otimes |0\rangle\langle 0|_{\text{out}})$. Recall that D_z uses its input (via the interpreter) as an advice state, runs it on $|z\rangle|y\rangle$, and then flips the "out" qubit iff the computation flipped $|y\rangle$. We conclude that the "out" qubit will be (nearly) in the state $|j\rangle$ at the end of the computation; this allows the adversary to detect if it was given the obfuscation of C_0 or the obfuscation of C_1.

Now suppose $\mathcal{S}^{U_{C_j}, U_{D_z}}$ is a simulator with black-box access to the unitary applied by D_z and either C_0 or C_1. Since $\mathcal{S}$ is polynomial-time, it can only use C_j a polynomial number of times. Without knowledge of j or z, the success probability of $\mathcal{S}$ will be no better than exponentially close to $1/2$, while the success probability of $\mathcal{A}$ is exponentially close to 1. □

Next, we should address the following question: if the obfuscated states are cloneable (or if we are allowed to ask the obfuscator for more than one copy), does one-circuit black-box obfuscation become impossible? In other words, can we change the above proof so that it works even when C_0 and C_1 are functionally equivalent? The natural function choice is one that, conditioned on a control wire, applies either C_j or D_z for random $z \in \{0,1\}^n$ and random $j \in \{0,1\}$. Does this work?

4 Quantum indistinguishability obfuscation

4.1 Definitions

Definition 7. *An **indistinguishability quantum obfuscator** is a pair $(\mathcal{J}, \mathcal{O})$ where $\mathcal{J}$ is an interpreter and $\mathcal{O}$ is a quantum algorithm which on input an n-qubit quantum circuit C outputs an m-qubit quantum state $\mathcal{O}(C)$, such that*

1. *(polynomial slowdown) $m = poly(n, |C|)$.*
2. *(functional equivalence) there exists a negligible ε_1 such that $\left\| \mathcal{J}_n^{\mathcal{O}(C)} - U_C \right\|_\diamond \leq \varepsilon_1(n, |C|)$;*

3. *(indistinguishability) if a pair of circuits C_1 and C_2 satisfy $|C_1| = |C_2|$ and $\left\|U_{C_1} - U_{C_2}\right\|_\diamond \leq \varepsilon_3(n, |C|)$, then $\left\|\mathcal{O}(C_1) - \mathcal{O}(C_2)\right\|_{\mathrm{tr}} \leq \varepsilon_4(n, |C|)$.*

As before, we will select ε_3 and ε_4 appropriately later. For a definition of best-possible obfuscation, we replace condition (3) above with the following:

3. *(best-possible) for every pair of quantum circuits C_1 and C_2 that satisfy $|C_1| = |C_2|$ and $\left\|U_{C_1} - U_{C_2}\right\|_\diamond \leq \varepsilon_3(n, |C|)$ and every quantum adversary $\mathcal{A}$, there exists a quantum simulator $\mathcal{S}$ and a negligible ε_2 such that*
$$\left|\Pr[\mathcal{A}(\mathcal{O}(C_1)) = 1] - \Pr\left[\mathcal{S}\left(C_2\right) = 1\right]\right| \leq \varepsilon_2(n, |C|)\,.$$

The intuition behind the above definition is the following: any information $\mathcal{A}(\mathcal{O}(C_1))$ that is "leaked" by the obfuscation $\mathcal{O}(C_1)$ can actually be recovered from *any* functionally equivalent, similarly-sized circuit C_2. In this sense, among all such circuits, the circuit $\mathcal{O}(C_1)$ is one that leaks the least. It's not hard to see that an efficient obfuscator satisfies the best-possible condition if and only if it satisfies the indistinguishability condition. This justifies Definition 7 as a natural choice.

4.2 Equivalence of indistinguishability and best-possible

4.3 Impossibility of statistical obfuscators

Recall the following computational problems and corresponding completeness results.

Definition 8. (a, b)-Identity Check.
Input: an n-qubit quantum circuit C.
Promise: $\min_\alpha \|U - e^{i\alpha} I\|$ *is less than a or greater than b.*
Output: YES in the former case and NO in the latter.

Theorem 3. *The problem (a, b)-Identity Check is coQMA-complete if $b - a \leq 1/\mathrm{poly}(n)$.*

Note: apparently Rosgen showed that distinguishing mixed state computations is QIP-complete. Does this mean that we could have even stronger impossibility results if we asked for obfuscators that could obfuscate quantum circuits that included measurements?

Given an m-qubit state ρ, let $\mathrm{Tr}_{(l,m)}[\rho]$ denote the result of tracing out qubits l through m. Nothing is traced out if $l > m$.

Definition 9. (a, b)-Quantum State Distinguishability
Input: m-qubit quantum circuits C_1 and C_2, positive integer $k \leq m$.
Promise: let $\rho_i = \mathrm{Tr}_{(k+1,m)}[C_i|0^m\rangle\langle 0^m|C_i^\dagger]$; then $\|\rho_0 - \rho_1\|_{\mathrm{tr}}$ is less than a or greater than b.
Output: YES in the former case and NO in the latter.

Theorem 4. *The problem (a, b)-Quantum State Distinguishability is QSZK-complete if $a < b^2$.*

We will in fact only need the containment part of the above theorem.

Theorem 5. *If there exists a polynomial-time indistinguishability quantum obfuscator, then coQMA is contained in QSZK.*

Proof. (parameters should be checked.) We will actually show coQMA $\subset$ BQP$^{\text{QSZK}}$; since BQP is contained in QSZK, the result will follow. Let a and b satisfy $b - a \leq 1/\text{poly}(n)$. We will solve (a,b)-Identity Check using a subroutine that solves (α,β)-quantum state distinguishability.

Let C be the input, i.e., a classical description of an n-qubit quantum circuit. Create an identity circuit D with an equal number of inputs as C, and of equal length to C. Let O_C be a circuit that initializes a register with the classical state $|C\rangle$ containing the classical description of C, and applies the circuit of $\mathcal{O}$ which corresponds to the input length $|C|$. Likewise, let O_D be be a circuit that initializes a register with the classical state $|D\rangle$ containing the classical description of D, and applies the circuit of $\mathcal{O}$ which corresponds to the input length $|D| = |C|$. Note that, after tracing out ancillas, the outputs of these circuits are given by

$$\text{Tr}_{\text{anc.}}\left[O_C|0\rangle\langle 0|O_C^\dagger\right] = \mathcal{O}(C) \qquad \text{and} \qquad \text{Tr}_{\text{anc.}}\left[O_D|0\rangle\langle 0|O_D^\dagger\right] = \mathcal{O}(D)\,.$$

Now apply the subroutine for solving quantum state distinguishability to the pair (O_C, O_D). If it says "close", we output YES; otherwise we output NO. Let's show that this has solved (a,b)-identity-check. Note that the states $\mathcal{O}(C)$ and $\mathcal{O}(D)$ must have the same number of qubits, and denote that number by m.

- **completness.** In this case, the obfuscated states satisfy $\|\mathcal{O}(C) - \mathcal{O}(D)\|_{\text{tr}} \leq \alpha$. By the definition of the induced trace norm, this implies that $\|\mathcal{J}^n_{\mathcal{O}(C)} - \mathcal{J}^n_{\mathcal{O}(D)}\|_\diamond \leq \alpha$. By functional equivalence for C and D and the triangle inequality, it follows that $\|U_C - U_D\|_\diamond = \|U_C - I\|_\diamond \leq \alpha$, as desired.

- **soundness.** In this case, the obfuscated states satisfy $\|\mathcal{O}(C) - \mathcal{O}(D)\|_{\text{tr}} \geq \beta$. We claim that this implies $\|U_C - U_D\|_\diamond > b$. Suppose this is not the case, i.e., that these operators are in fact close; then by the indistinguishability property, it would follow that $\mathcal{O}(C)$ and $\mathcal{O}(D)$ are close as well, a contradiction.

The above amounts to a BQP$^{\text{QSZK}}$ protocol for a coQMA-hard problem, thus placing coQMA in QSZK. □

5 [OLD NOTES]

5.1 Preliminaries

Given a probability distribution X on a finite set S and an element $s \in S$, let $s \sim X$ denote the experiment of sampling s according to the distribution X. For example, $\Pr_{s\sim X}[s \in S']$ denotes the probability that a sample of X belongs to some subset $S' \subset S$. The total variation distance between two probability distributions X and Y taking values in S is defined by

$$|X - Y| = \frac{1}{2}\sum_{s\in S}\left|\Pr[X = s] - \Pr[Y = s]\right|.$$

If A and B are random variables with the same range, the notation $|A - B|$ will mean the total variation distance between the distributions of A and B.

We will often refer to circuits as deciding some problem, in the following sense. Let $\{C_n\}_{n\in\mathbb{N}}$ be a uniform family of classical probabilistic circuits. Fix $x \in \{0,1\}^n$, and let $C_n x$ denote the random variable determined by running C_n on the input x (with each remaining input bit set to either 0 or to the outcome of a uniformly random coinflip) and reading out the value of the first output bit. If the input x is selected according to some probability distribution A, then the acceptance probability is $\Pr_{x\sim A}[C_n x = 1]$. It is implicit that the probability is now taken over both the choice of x and the coins of C_n.

We can also view quantum circuits in this way. In the remainder of these notes, "quantum circuit" will always mean "unitary quantum circuit." Any measurements will be specified explicitly, and performed after the quantum circuit is applied. This is sufficient to describe arbitrary quantum computations (which in general may include many rounds of unitary operations, adapted measurements, and classical pre- and post-processing.) Set $\{C_n\}_{n\in\mathbb{N}}$ to be a uniform family of quantum circuits, and let $p(n)$ denote the number of qubits acted on by C_n. Given a quantum state $|\psi\rangle$ on n qubits, we apply the circuit C_n to the state $|\psi\rangle|0\rangle^{\otimes p(n)-n}$, and then measure the first qubit in the computational basis. This procedure can be described by a $\{0,1\}$-valued random variable, which we will denote by $M(C_n|\psi\rangle)$. Specifically, for $a \in \{0,1\}$,

$$\Pr[M(C_n|\psi\rangle) = a] = \sum_{x\in\{0,1\}^{p(n)}:x_1=a}\left|\langle x|C_n|\psi\rangle|0\rangle^{\otimes p(n)-n}\right|^2.$$

- it is worthwhile to discuss here why there is no "more powerful" way to use circuit families to solve decision problems (e.g., by appealing to PromiseBQP/PromiseBPP-hardness).

A *probability ensemble* D is a sequence $\{D_n\}_{n\in\mathbb{N}}$ of bitstring-valued random variables, such that for some polynomial ℓ, each D_n takes values in $\{0,1\}^{\ell(n)}$. We may sometimes need such ensembles to be polynomial-time constructible, meaning that one can sample from D via a uniform family of probabilistic circuits. Recall that a function $\phi : \mathbb{N} \to [0,\infty)$ is *negligible* if it is smaller than inverse-polynomial in n. We identify four distinct notions of indistinguishability of two probability ensembles A and B:

1. *perfectly indistinguishable*: $A_n = B_n$ for all sufficiently large n;
2. *statistically indistinguishable*: there exists a negligible function ϕ such that $|A_n - B_n| \leq \phi(n)$ for all sufficiently large n;

3. *quantumly indistinguishable*: there exists a negligible function ϕ such that, given any uniform family $\{C_n\}_{n\in\mathbb{N}}$ of quantum circuits, for all sufficiently large n we have

$$\left| \Pr_{x\sim A_n}[M(C_n|x\rangle)=1] - \Pr_{x\sim B_n}[M(C_n|x\rangle)=1] \right| \leq \phi(n)\,.$$

4. *classically indistinguishable*: there exists a negligible function ϕ such that, given any uniform family $\{C_n\}_{n\in\mathbb{N}}$ of classical probabilistic circuits, for all sufficiently large n we have

$$\left| \Pr_{x\sim A_n}[C_n x=1] - \Pr_{x\sim B_n}[C_n x=1] \right| \leq \phi(n)\,;$$

The quantum case can also be expressed naturally in terms of density operators. Let us write $M(C_n\rho)$ for the random variable corresponding to the same decision experiment as before, but now starting with a density operator ρ. Given a probability distribution A on $\{0,1\}^n$, set

$$\rho_A = \sum_{x\in\{0,1\}^n} \Pr[A=x]|x\rangle\langle x|\,.$$

The random variable $M(C_n\rho_A)$ then exactly captures the outcome of selecting a string at random according to A, and then running the decision experiment corresponding to C_n. In other words,

$$\Pr_{x\sim A}[M(C_n|x\rangle)=a] = \Pr[M(C_n\rho_A)=a]\,.$$

Proposition 2. *Indistinguishability of probability ensembles satisfies*

$$\textit{perfect} \Rightarrow \textit{statistical} \Rightarrow \textit{quantum} \Rightarrow \textit{classical}\,.$$

Proof. Let A and B be probability ensembles. The first implication is immediate from the definition. For the second, recall the definition of trace norm $\|\rho\|_1 = \mathrm{Tr}\sqrt{\rho\rho^\dagger}$, and note that the trace distance between the two relevant density operators is

$$\|\rho_A - \rho_B\|_1 = 2|A-B|\,.$$

It's easy to check that the trace norm is unitarily invariant, so applying the same quantum circuit to both ρ_A and ρ_B does not affect the trace distance. The final measurement is just a projection to some subspace, and so the difference in the acceptance probabilities is bounded above by twice the trace distance. For the third implication, by standard arguments we can replace any classical circuit family that distinguishes two ensembles with a classical reversible circuit family that does the same. Reversible circuits are a special case of quantum circuits. □

- examples of why the implications are strict: (1) trivial, (2) large statistical difference but no quantum distinguisher (graph isomorphism?), and (3) quantum distinguisher but no classical distinguisher (factoring, or Bill's idea?).

By the triangle inequality, all four notions of indistinguishability are transitive, i.e. if A is indistinguishable from B and B is indistinguishable from C, then A is indistinguishable from C. All four notions of indistinguishability are also closed under applying polynomial-time operations to both ensembles; the exception is that classically indistinguishable ensembles may become classically distinguishable after an efficient quantum algorithm is applied.

5.2 Black-box Quantum circuit obfuscation

Given a (not necessarily uniform) family of circuits $\mathcal{C}$, let $\mathcal{C}_n$ denote the subset of $\mathcal{C}$ consisting of all circuits that act on exactly n qubits. If each $\mathcal{C}_n$ consists of one circuit only, then C_n will refer to that unique circuit, and the expression $\mathcal{C}|x\rangle$ will mean $C_n|x\rangle$ where n is the number of qubits of the state $|x\rangle$.

For a quantum circuit C, let U_C denote the unitary operator implemented by C. The notation S^C will stand for a quantum circuit S which, in addition to a universal set of quantum gates, can also make use of an additional black-box gate which implements U_C. The black-box gate can be used as many times as needed, although each use does count toward the total length of S^C.

We are now ready to define a few different notions of quantum circuit obfuscation. Our definitions closely follow the classical ones in Goldwasser and Rothblum.

Definition 10. *A classical probabilistic algorithm $\mathcal{O}$ that takes as input a quantum circuit C and outputs another quantum circuit $\mathcal{O}(C)$ is a quantum* **black-box obfuscator** *for the circuit family $\mathcal{C}$ if it satisfies:*

1. *preserving functionality: there is a negligible function ϕ such that for any n and any $C \in \mathcal{C}_n$,*

$$\Pr[U_C \neq U_{\mathcal{O}(C)}] \leq \phi(n)\,.$$

2. *polynomial slowdown: there is a polynomial p such that for any $C \in \mathcal{C}$, $|\mathcal{O}(C)| \leq p(|C|)$.*

3. *virtual black-box: For any adversary (uniform quantum circuit family) $\mathcal{A}$, there is a simulator (uniform quantum circuit family) $\mathcal{S}$ and a negligible ϕ such that*

$$\left|\Pr[M(\mathcal{A}|\mathcal{O}(C)\rangle) = 1] - \Pr[M(\mathcal{S}^C|0\rangle) = 1]\right| \leq \phi(n)$$

for every n and every $C \in \mathcal{C}_n$.

- in GR there aren't four versions of the last property – just the computational one. Why?
- we may later wish to relax the functionality-preserving condition, so that two unitaries are considered functionally equivalent so long as (say) there is no polynomial-length proof of their inequality. This would affect later definitions too.
- are the classically not-black-box-obfuscatable functions also not quantum black-box obfuscatable?
- if not, are there other examples of not-quantum-black-box-obfuscatable functions? In order for these examples to be interesting, I guess they shouldn't be "learnable," i.e., you can't figure out exactly what they are with a polynomial number of black-box uses.
- is there an example family of quantum circuits which *is* black-box obfuscatable?

5.3 Best-possible

In what follows, for the sake of simplicity we omit the perfect, statistical, and classical variants of the definitions; one can arrive at these versions simply by replacing quantum indistinguishability of the relevant ensembles to one of the other notions. We will always be obfuscating quantum

circuits, so when the word "quantum" appears in front of "obfuscator", this refers to the type of indistinguishability. We say that two uniform quantum circuit families $\mathcal{C}'$ and $\mathcal{C}''$ are equivalent if they consist of functionally equivalent circuits of the same size; more precisely, for every n, $|\mathcal{C}'_n| = |\mathcal{C}''_n| = 1$ and $|C'_n| = |C''_n|$ and $U_{C'_n} = U_{C''_n}$.

- the exact-same-length condition seems too strong, but it does appear in GR too, along with a later comment about how it can be removed. I guess some care is needed.

Definition 11. *A classical probabilistic algorithm $\mathcal{O}$ that takes as input a quantum circuit C and outputs another quantum circuit $\mathcal{O}(C)$ is a quantum* **best-possible obfuscator** *for the family $\mathcal{C}$ if it satisfies properties (1) and (2) from Definition 10, as well as the following property:*

3. *for any learner (uniform quantum circuit family) $\mathcal{L}$, there is a simulator (uniform quantum circuit family) $\mathcal{S}$ and a negligible ϕ such that, for all uniform equivalent subfamilies $\mathcal{C}', \mathcal{C}''$ of $\mathcal{C}$, the two ensembles $\mathcal{L}(\mathcal{O}(\mathcal{C}'))$ and $\mathcal{S}(\mathcal{C}'')$ are quantumly indistinguishable.*

Example: quantum witness encryption. The classical idea of witness encryption is from a paper by Sahai, Garg and others, and the idea of solving it with obfuscation is from the big paper by Sahai et al. In the quantum case, we set up the problem as follows. Suppose Alice wishes to encrypt a quantum plaintext $|x\rangle$, but not to a particular key or for a particular person; instead, the encryption is tied to a challenge question, and anyone that can answer the question correctly can decrypt the plaintext. Alice outputs a ciphertext $F_\phi|x\rangle$ where ϕ is a quantum 3-SAT formula, such that there exists an efficient algorithm Eval with the property that $\text{Eval}(F_\phi|x\rangle, |y\rangle) = |x\rangle$ if $|y\rangle$ is a satisfying assignment for ϕ. The security requirement is that if ϕ does not have a satisfying assignment, then the ensembles $F_\phi|x\rangle$ and $F_\phi|x'\rangle$ are quantum indistinguishable (formally, this now requires a definition of distinguishing *quantum* ensembles) whenever $|x\rangle$ and $|x'\rangle$ are quantum states on the same number of qubits. Note that the definition says nothing about the case where ϕ is satisfiable but a satisfying assignment is not known. While this may seem counterintuitive, Sahai and Garg etc. are nonetheless able to construct various interesting encryption schemes (like public-key encryption and identity encryption) from witness encryption.

The problem of quantum witness encryption can be solved using a quantum best-possible obfuscator $\mathcal{O}$, as follows. First, Alice selects a random Clifford (or Pauli) circuit C. She then writes down a quantum circuit M_C which accepts two registers (and some ancillas), such that $M|z\rangle|y\rangle|0\rangle = |C^{-1}z\rangle|y\rangle|0\rangle$ when $|y\rangle$ is a satisfying assignment for ϕ, and $M|z\rangle|y'\rangle|0\rangle = |z\rangle|y'\rangle|0\rangle$ for $|y'\rangle$ not a satisfying assignment for ϕ. The ciphertext $F_\phi|x\rangle$ will consist of the pair $(C|x\rangle, \mathcal{O}(M_C))$. A recipient with a satisfying assignment $|y\rangle$ can decrypt by computing $\mathcal{O}(M_C)|Cx\rangle|y\rangle|0\rangle$. On the other hand, if no satisfying assignment exists, then M_C acts like the identity operator on every input. By the definition of best-possible, a quantum adversary can learn nothing more from $\mathcal{O}(M_C)$ than she could from the trivial circuit with no gates. Moreover, by the design property of Cliffords (or Paulis) the adversary also observes $|Cx\rangle$ to be a maximally mixed state.

- Stephen has a description of how to build the circuit M_C, and that should be added.
- I guess the state $C|x\rangle$ and the circuit M_C are correlated. Is this a problem? This probably has to be addressed by defining quantum indistinguishability of quantum ensembles, and then showing that quantum indistinguishability of the classical ensemble $\mathcal{O}(M_C)$ plus 2-design property on $C|x\rangle$ implies quantum indistinguishability of the quantum ensemble $(C|x\rangle, \mathcal{O}(M_C))$.

- what does M_C do if you feed in a state that has a little bit of projection into a satisfying assignment? I guess that, unless the size of the projection is $1/\text{poly}$, it's still indistinguishable from identity...
- I have some ideas on why the above is exactly the right definition (e.g., weakening to ϕ being just a 3-SAT formula opens it up to being solved by classical obfuscation.)

5.4 Indistinguishability

Definition 12. *A classical probabilistic algorithm $\mathcal{O}$ that takes as input a quantum circuit C and outputs another quantum circuit $\mathcal{O}(C)$ is a quantum* **indistinguishability obfuscator** *for the family $\mathcal{C}$ if it satisfies properties (1) and (2) from Definition 10, as well as the following property:*

3. *for all uniform equivalent subfamilies $\mathcal{C}', \mathcal{C}''$ of $\mathcal{C}$, the two ensembles $\mathcal{O}(\mathcal{C}')$ and $\mathcal{O}(\mathcal{C}'')$ are quantumly indistinguishable.*

- in all of the above, we could have considered obfuscating quantum states, or even using quantum algorithms to obfuscate classical descriptions of a quantum circuit. Why is this the "right" case (or at least an interesting one)?

5.5 Relationships between the definitions

With the definitions set up as above, many of the proofs of Goldwasser and Rothblum go through with little to no changes.

Proposition 3. *There exists an inefficient perfect indistinguishability obfuscator for all quantum circuits.*

Proof. The obfuscator just picks the lexicographically first circuit which implements the same unitary as the given circuit. Looping through lexicographically ordered circuits can be done in PSPACE, and equivalence-checking can be done in QMA $\subset$ QIP = PSPACE too. □

- what's the smallest class that one can do this in?

Proposition 4. *If $\mathcal{O}$ is a best-possible quantum obfuscator for a circuit family $\mathcal{C}$, then it is also a quantum indistinguishability obfuscator for $\mathcal{C}$.*

Proof. Let $\mathcal{C}'$ and $\mathcal{C}''$ be uniform equivalent subfamilies of $\mathcal{C}$, and let $\mathcal{L}$ be the trivial learner that simply implements the identity operator. By the best-possible property, there is a simulator $\mathcal{S}$ such that $\mathcal{S}(\mathcal{C}'')$ is quantum indistinguishable from $\mathcal{L}(\mathcal{O}(\mathcal{C}')) = \mathcal{O}(\mathcal{C}')$. By the same property, we also have that $\mathcal{S}(\mathcal{C}'')$ is quantum indistinguishable from $\mathcal{L}(\mathcal{O}(\mathcal{C}'')) = \mathcal{O}(\mathcal{C}'')$. By the transitivity property of indistinguishability, it follows that $\mathcal{O}(\mathcal{C}')$ is indistinguishable from $\mathcal{O}(\mathcal{C}'')$. □

Proposition 5. *If $\mathcal{O}$ is an efficient quantum indistinguishability obfuscator for a circuit family $\mathcal{C}$, then it is also an efficient quantum best-possible obfuscator for $\mathcal{C}$.*

Proof. Let $\mathcal{C}'$ and $\mathcal{C}''$ be equivalent subfamilies of $\mathcal{C}$, and let $\mathcal{L}$ be a (quantum) learner whose output on $\mathcal{C}'$ is the ensemble $\mathcal{L}(\mathcal{O}(\mathcal{C}'))$. We define a (quantum) simulator by setting $\mathcal{S} = \mathcal{L} \circ \mathcal{O}$; its output on $\mathcal{C}''$ is then the ensemble $\mathcal{L}(\mathcal{O}(\mathcal{C}''))$. Since the ensembles $\mathcal{O}(\mathcal{C}')$ and $\mathcal{O}(\mathcal{C}'')$ are quantum indistinguishable, so are their images under $\mathcal{L}$. □

5.6 Example: Clifford circuits

Recall that the single-qubit Pauli operators are defined by

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Each Pauli operator is self-adjoint and unitary. A few useful relations are

$$X^2 = Y^2 = Z^2 = I \qquad XY = -YX = iZ \qquad XZ = -ZX = -iY \qquad YZ = -ZY = iX.$$

From these relations, it's easy to see that the set of matrices αM where $\alpha \in \{\pm 1, \pm i\}$ and $M \in \{I, X, Y, Z\}$ forms a group under matrix multiplication. This group is generated by $\{X, Y, Z\}$ and $\{\pm 1, \pm i\}$. In the n-qubit case, we first set

$$X_j = I^{\otimes j-1} \otimes X \otimes I^{\otimes n-j}$$

and likewise for Y_j and Z_j. We define the n-qubit Pauli group $\mathcal{P}_n$ to be the group generated by $\{X_j, Y_j, Z_j : j = 1, \ldots, n\}$ and $\{\pm 1, \pm i\}$.

The Clifford group on n qubits is defined to be the normalizer of the Pauli group inside the unitary group, i.e.,

$$\mathcal{C}_n = \{U \in U(2^n) : UPU^\dagger \in \mathcal{P}_n \text{ for all } P \in \mathcal{P}_n\}.$$

By direct computation on the Pauli generators, it's easy to check that the following gates are elements of $\mathcal{C}_n$ for any $n \geq 2$:

$$H = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \qquad P = \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix}, \qquad CNOT = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

It is a theorem (see Gottesman's papers) that the above gates (when applied to arbitrary qubits or pairs of qubits) actually generate the entire Clifford group. A Clifford circuit is any circuit which is made up of gates from the above gate set. It is well-known that Clifford circuit computations can be efficiently simulated by a classical computer, but that adding any gate outside the Clifford group yields a quantum-universal set. In spite of their lack of computational power, Clifford circuits are quite relevant in quantum information, e.g., in quantum error correction and quantum cryptography.

In this section, we show how to put any Clifford circuit into a unique normal form. Something like this is already discussed in Gottesman's PI lectures. Selinger also provides a unique normal form (as well as generators and relations for $\mathcal{C}_n$) but he uses a different gate set. The approach below also seems more natural, as it's closely related to how Cliffords are usually discussed in the QI literature.

For us, a "unique normal form" is a map f from Clifford circuits to Clifford circuits, such that *(i.)* C and $f(C)$ always implement the same unitary operator, and *(ii.)* whenever C_1 and C_2 are circuits which implement the same unitary operator, $f(C_1)$ and $f(C_2)$ are identical as circuits. We will sketch out how this can be done using a polynomial-time classical algorithm. By definition, this immediately gives an indistinguishability obfuscator for Clifford circuits.

Moreover, by a result of Richard Low, given a black box that implements a Clifford group element U, we can "learn" the action of U on the Pauli generators in polynomial time. As our algorithm will make clear, knowing the action of U on the generators suffices to produce the normal form. This means that any learner that has access to a normal-form Clifford circuit for U can be simulated by a learner with black-box access to U. This obfuscation scheme thus also satisfies the conditions of black-box obfuscation.

Unfortunately, this obfuscation is in some sense trivial; while it is true that the precise form of the initial circuit is not learnable from the obfuscated circuit, it is nonetheless easy to learn the full functionality.

We can map each element of the n-qubit Pauli group to a $2n$-bit string by ignoring the phase and setting

$$X_i \mapsto (\underbrace{0,\ldots,0}_{i-1},1,0,\ldots,0) \qquad \text{and} \qquad Z_i \mapsto (\underbrace{0,\ldots,0}_{n+i-1},1,0,\ldots,0)\,.$$

By checking the relations on the generating set, one sees that this map yields an isomorphism

$$f : \mathcal{P}_n/\{\pm 1,\pm I\} \to \mathbb{Z}_2^{2n}\,.$$

It's also easy to compute how the conjugation action of a Clifford gate on a Pauli generator affects the corresponding binary string. Since conjugation is linear, this is described by a matrix. For example,

$$H \mapsto \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad P \mapsto \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \qquad CNOT \mapsto \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

In general, for any fixed n, applying any of the above gates to a particular qubit (or pair of qubits for CNOT) will correspond to some easily computable $2n \times 2n$ binary matrix. Given a Clifford circuit C, we can multiply the matrices corresponding to each gate in C to get a matrix $M(C)$. This matrix satisfies the property

$$M(C)f(P) = f(M(CPC^\dagger))$$

for every Pauli $P \in \mathcal{P}_n$. In fact, it is also the case that $M(C_1) = M(C_2)$ whenever C_1, C_2 are two Clifford circuits that implement the same element of the Clifford group. This follows from the isomorphism

$$\mathcal{C}_n' \cong \mathrm{Sp}(2n,\mathbb{F}_2)\,,$$

where $\mathcal{C}_n'$ denotes $\mathcal{C}_n$ modulo $\mathcal{P}_n$ and arbitrary phases, and $\mathrm{Sp}(2n,\mathbb{F}_2)$ denotes the group of $2n \times 2n$ symplectic matrices over $\mathbb{F}_2$. Why symplectic? Well, because Clifford elements preserve both commutation and anti-commutation of Pauli group elements, and whether two Pauli group elements commute or anti-commute is captured by a symplectic form of their corresponding binary strings:

$$PQ = (-1)^{\omega(f(P),f(Q))}QP$$

where

$$\omega(x,y) = (x_1,\ldots,x_n|y_{n+1},\ldots,y_{2n}) + (y_1,\ldots,y_n|x_{n+1},\ldots,x_{2n})$$

and $(a|b)$ denotes the dot product modulo 2.

It now remains to produce a unique Clifford circuit from $M(C)$, and append the right element of $\mathcal{P}_n$. The former is done through a row reduction procedure. The key observation is that row

reduction operations correspond to left-multiplication by matrices corresponding to gates. Once we have row-reduced $M(C)$ to the identity, we then invert the sequence of gates we applied to output a circuit C'. We then know that

$$C^{-1}C' = P$$

for some $P \in \mathcal{P}_n$. By applying each gate of CC' to the Pauli generators, we can compute P and append its inverse to C'. This constitutes a unique circuit which is equivalent to C up to overall phases.

- the above is clearly just a sketch, which we can flesh out if we decide this is really important stuff.

Why is this uninteresting Note that any canonical form obfuscator is not, in general, a black-box obfuscator. A learner which is given the canonical form of a circuit can, in general, learn something that a learner with only black-box access cannot: namely, the canonical form itself! It's useful here to think about what such an obfuscator does on a family of circuits which are *already* in canonical form.

Now suppose all of the functions computed by the relevant class of circuits are black-box learnable, in the sense that there is an efficient algorithm which can use black-box access to a function f to output a description of any circuit (and hence also the canonical circuit) for computing f. Strictly speaking, the canonical-form obfuscator is now also a black-box obfuscator. But now again consider a uniform family of circuits which are already in canonical form. In this case, black-box access can be used to recover the entire original circuit perfectly. This should mean that, in an intuitive sense, obfuscation is completely impossible for this circuit family. This explains why our definitions (as well as the classical ones) are meaningless when we talk about efficiently learnable functions.