Breaking a Classical Barrier for Classifying Arbitrary Test Examples in the Quantum Model

Khashayar Barooti
EPFL
Lausanne, Switzerland
khashayar.barooti@epfl.ch

Grzegorz Głuch
EPFL
Lausanne, Switzerland
grzegorz.gluch@epfl.ch

Ruediger Urbanke
EPFL
Lausanne, Switzerland
ruediger.urbanke@epfl.ch

Abstract

A new model for adversarial robustness was introduced by Goldwasser et al. in [GKKM20]. In this model the authors present a selective and transductive learning algorithm which guarantees a low test error and low rejection rate wrt to the original distribution. Moreover, a lower bound in terms of the VC-dimension, the standard risk and the number of samples is derived.

We show that this lower bound can be broken in the quantum world. We consider a new model, influenced by the quantum PAC-learning model introduced by [BJ95], and similar in spirit to the one in [GKKM20]. In this model we give an interactive protocol between the learner and the adversary (at test-time) that guarantees robustness. This protocol, when applied, breaks the lower bound from [GKKM20].

From the technical perspective, our protocol is inspired by recent advances in delegation of quantum computation, e.g. [Mah18]. But in order to be applicable to our task, we extend the delegation protocol to enable a new feature, e.g. by extending delegation of decision problems, i.e. BQP, to sampling problems with adversarially chosen inputs.

1 Introduction

We are interested in the task of classifying test examples that are arbitrary, by which we mean any set of examples from the input space. More formally, assume that a classifier $f : \mathcal{X} \to \{-1, 1\}$ was trained using iid samples from the training distribution $\mathcal{D}$. Then, at test time, a set of *arbitrary* examples is given to the classifier. In particular, this models the adversarial robustness setup, where an adversary applies imperceptible (think of perturbations small in ℓ_2 norm) perturbations to iid samples from $\mathcal{D}$ in order to fool f [SZS+14, NYC15]. This setup also covers a situation where an adversary is *not* limited to small perturbations. For an example of such a situation consider the case of explicit content detection [YTL+19], where an adversary produces endless variations of an image to pass the detection test.

Perhaps unsurprisingly, the task of classifying arbitrary test examples is impossible to solve in the usual settings. If f has accuracy strictly smaller than 100% and if all the test examples are chosen to correspond to inputs where f makes an error then all of them will be misclassified by f. To resolve issues of this nature a new model was recently introduced in [GKKM20]. The authors argue that one should consider *selective classifiers* and *transductive learning*. A selective classifier is allowed to abstain from prediction on certain examples, while transductive learning refers to a situation, where the test examples are presented together with training examples. In [GKKM20] it is argued that selective classifiers are necessary to obtain meaningful guarantees in the arbitrary test examples case.

The guarantees obtained in [GKKM20] give bounds on the interplay of two quantities: the risk on arbitrary test examples and the rejection rate on iid samples from $\mathcal{D}$ (training distribution). It is natural that there is a trade-off, because one could easily maximize both of this metrics separately by either: rejecting almost all inputs or just applying f without rejecting anything. One of the results in [GKKM20] is a lower bound on the possible trade-offs of these two quantities. The lower bound provides a minimum number of training samples and test examples needed for the risk on arbitrary examples + the rejection rate on $\mathcal{D}$ to be smaller than ϵ. The bound is expressed in terms of the VC-dimension and ϵ. We break this lower bound by considering a quantum model. Instead of the standard samples $x \sim \mathcal{D}$ we assume access to the many qubit quantum states $\sum_{x \in \{0,1\}^n} \sqrt{\mathcal{D}(x)} \, |x\rangle$ – similar to the quantum PAC-learning model by [BJ95].

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

On the technical side, we borrow heavily from a series of results on the delegation of quantum computation [Mah18]. These techniques allow us to "restrict the actions of the adversary." Using ideas from this line of work, we are able to design a key tool for our result. Namely a protocol between a classical verifier and a quantum prover that guarantees that the samples collected by the verifier at the end of interaction come from a distribution close to $\mathcal{D}$ - we call it a certifiable sampling protocol. This is done under the assumption that the prover cannot break a type of post-quantum cryptography - an assumption also present in previous works. Our protocol builds upon ideas from [Mah18] but it is not just plug and play: in our setting we need to collect samples from some distribution - whereas, the previous results only provide guarantees for delegating decision problems. Due to these differences a new protocol(s) is required, together with a careful analysis to verify correctness in this extended setting. For readers who are familiar with the proof in [Mah18], you can see how our extra requirement manifests itself by comparing for instance Theorem 2 and Theorem 10.

For a more in depth account of related works we refer the reader to Appendix B. For readers not familiar with basics of quantum mechanics, we have provided a brief introduction on the subject in Appendix A.

2 Model and Main Result

Assume that the training phase is finished and a classifier $f : \mathcal{X} \to \{-1, 1\}$ was learnt from iid samples from $\mathcal{D}$. We describe the dynamics of the test phase. As discussed in the introduction there are two main quantities of interest: risk on arbitrary test examples and the rejection rate on samples from $\mathcal{D}$. We think of these two quantities as arising from two modes of operation of an adversary $\mathbf{A}$, i.e., a malicious mode and an honest mode, respectively. We think of the test phase as an interaction between $\mathbf{A}$ and a second party $\mathbf{V}$ (as in verifier), i.e., $\mathbf{A}$ sends samples for classification to $\mathbf{V}$. We consider a quantum model, where an honest $\mathbf{A}$, instead of receiving a sample $\mathbf{x} \sim \mathcal{D}$, receives a quantum state $|\psi_\mathcal{D}\rangle = \sum_{x \in \{0,1\}^n} \sqrt{\mathcal{D}(x)} |x\rangle$, where we assumed that $\mathcal{X} = \{0, 1\}^n$. This model is closely related to the model that is considered in the quantum PAC-learning literature [BJ95].[1] In our model we allow an interaction between $\mathbf{A}$, that from now on we will call $\mathbf{P}$ as in prover, and $\mathbf{V}$.

More formally we assume that the interaction between $\mathbf{P}$ and $\mathbf{V}$ proceeds as follows. Upon receiving a state $|\psi_\mathcal{D}\rangle$ the adversary/prover $\mathbf{P}$ performs an arbitrary computation (quantum or classical) and starts an interaction with the learner/verifier $\mathbf{V}$. After some number of rounds of interaction a classical example might be obtained by $\mathbf{V}$ that then classifies it with f. We will consider different models depending on what type of messages can be exchanged in the protocol and what power $\mathbf{V}$ has. Ultimately, we will aim for a model where $\mathbf{V}$ is fully classical and the messages exchanged are classical also but we will also consider models were quantum messages can be exchanged.

2.1 Preliminaries

For $k \in \mathbb{N}$ we denote by $[k]$ the set $\{1, \dots, k\}$ and by $\mathfrak{D}(n)$ the family of distributions on n-bit strings. For $\mathcal{P}, \mathcal{Q} \in \mathfrak{D}(n)$ we define their Hellinger distance as $d_H(\mathcal{P}, \mathcal{Q}) := \frac{1}{\sqrt{2}} \|\sqrt{\mathcal{P}} - \sqrt{\mathcal{Q}}\|_2$. For $\mathcal{D} \in \mathfrak{D}(n)$ we define $\mathtt{O}(\mathcal{D})$ as an oracle giving access to $|\psi_\mathcal{D}\rangle := \sum_{x \in \{0,1\}^n} \sqrt{\mathcal{D}(x)} |x\rangle$, where by some abuse of notation we write $|x\rangle$ (for $x \in \{0, 1\}^n$) to denote $|x_1\rangle |x_2\rangle \dots |x_n\rangle$. In our protocols we will be interested in an interaction between $\mathbf{V}$ (Verifier) and $\mathbf{P}$ (Prover). We will write $\mathbf{P}^{\mathtt{O}(\mathcal{D})}$ to denote that $\mathbf{P}$ has access to $\mathtt{O}(\mathcal{D})$. For a quantum circuit C acting on n-qubits via the unitary transform U_C, we define $\mathcal{D}_C \in \mathfrak{D}(n)$ as the distribution arising from measuring all n qubits of $U_C|0^{\otimes n}\rangle$ in the computational (which we will also denote as Z) basis. We will sometimes abuse the notation and write $C\,|\psi\rangle$ to mean $U_C\,|\psi\rangle$. For $|\psi\rangle \in (\mathbb{C}^2)^{\otimes n}$ we say that $\mathcal{D} \in \mathfrak{D}(n)$ defined as $\mathcal{D}(x) = |\,\langle x|\psi\rangle\,|^2$ for every $x \in \{0, 1\}^n$ is the distribution associated with $|\psi\rangle$.

Definition 1 (**Standard Risk**). *For a separable binary classification task with distribution $\mathcal{D}$ and a ground truth g we define the standard risk of f as*

$$R_\mathcal{D}(f) := \mathbb{P}_{x \sim \mathcal{D}}[f(x) \neq g(x)].$$

As described before the two main quantities of interest are: the rejection rate when $\mathbf{P}$ acts honestly and the risk when $\mathbf{P}$ acts maliciously. We define these quantities more formally now. Both of them are defined with respect to a specific protocol and a classification task that will always be clear from context. In our protocols we perform various consistency checks and collect some statistics. Because of that the protocols need to be repeated some number of times to obtain meaningful guarantees. This is why rejection rates and risk on arbitrary examples are defined as values in expectation.

Definition 2 (**Rejection Rate**). *We define the rejection rate as 1 minus the expectation of the ratio of the number of samples obtained by $\mathbf{V}$ in the protocol (when an honest $\mathbf{P}$ interacts with $\mathbf{V}$) to the number of states $|\psi_\mathcal{D}\rangle$ that $\mathbf{P}$ used in the protocol. We denote it by*

$$\bot_\mathcal{D} := 1 - \mathbb{E}\left[\frac{\#\textit{examples obtained by } \mathbf{V}}{\#\textit{number of } |\psi_\mathcal{D}\rangle \textit{ used by } \mathbf{P}}\right],$$

where the expectation is over the randomness of $\mathbf{V}$ and $\mathbf{P}$ (that also includes the randomness stemming from quantum mechanics).

Definition 3 (**Risk on Arbitrary Examples**). *We define the risk on arbitrary examples as the supremum over malicious*

[1] In [BJ95] quantum samples are states of the form $\sum_{x \in \{0,1\}^n} \sqrt{\mathcal{D}(x)} |x, g(x)\rangle$. Here, g is the ground truth. Our results likely carry over to this quantum PAC-model, but as always, the details need to be verified.

provers accepted with probability 1 *of the expected risk of* f *on examples accepted by* $\mathbf{V}$*. We denote it by*

$$AR := \sup_{\mathbf{P}} \mathbb{P}_{x \sim \textit{Accepted } \mathbf{V} \leftrightarrow \mathbf{P}}[f(x) \neq g(x)],$$

where the probability is over the randomness of $\mathbf{V}$ *and* $\mathbf{P}$ *conditioned on accepted interaction and* x *is sampled at random from all obtained examples. AR stands for Arbitrary Risk but can be also thought of as Adversarial Risk in a sense that it is a risk in the presence of an adversary.*

2.2 Main Result

As discussed above, our result is applicable to the test phase. We assume that the training phase is completed and $\mathbf{V}$ has access to two objects obtained during the training phase: a classifier f and a description of a generative quantum circuit C with the following properties.

The circuit C captures the true distribution well, i.e. $d_H(\mathcal{D}_C, \mathcal{D}) = \eta \ll 1$. The classifier f is robust with respect to small changes in the distribution (i.e., it is robust to distributional shifts). This means that for all $\mathcal{D}^A \in \mathfrak{D}(n)$ such that $d_H(\mathcal{D}^A, \mathcal{D}) \leq O(\eta)$ we would have $R_{\mathcal{D}^A}(f) \approx R_{\mathcal{D}}(f)$.

We claim that if such a $\mathbf{V}^{C,f}$ ($\mathbf{V}$ having access to f and the description of C) interacts with $\mathbf{P}$ using our protocol (defined in Section 3) then this will yield a framework robust under all (computationally bounded) adversaries. Indeed there are two scenarios of interest: (1) $\mathbf{P}$ is honest, (2) $\mathbf{P}$ is malicious. In (1) $\mathbf{P}$ acts "equivalently" to just measuring $|\psi_{\mathcal{D}}\rangle$ and sending the result to $\mathbf{V}$. Our protocol guarantees that big fraction of these samples will be accepted (small rejection rate) as they came from $\mathcal{D}$ itself. Classifier f is robust wrt distributional shifts around $\mathcal{D}$, which in particular implies that it has a low risk on $\mathcal{D}$ itself. In (2) the certifiable sampling protocol (defined in Section 3) guarantees that the interaction will only be accepted if the distribution "from which $\mathbf{P}$ samples" is close to $\mathcal{D}$. Then again we know that f has low risk on samples from such a distribution, which guarantees low risk on arbitrary examples. Thus we arrive at the main theorem of our paper.

Theorem 1. *There exists a universal constant* $K \in \mathbb{N}$ *such that for every* $n \in \mathbb{N}$*, any small enough* $\eta \in (0,1)$ *, for every binary, separable classification task with a distribution* $\mathcal{D} \in \mathfrak{D}(n)$ *and a ground truth* $g : \{0,1\}^n \to \{-1,1\}$*, every classifier* $f : \{0,1\}^n \to \{-1,1\}$ *and every quantum circuit* C *with* T *gates the following conditions hold. If*

- ***($\mathcal{D}_C$ is a good approximation of $\mathcal{D}$)*** $\|\sqrt{\mathcal{D}_C} - \sqrt{\mathcal{D}}\|_2 \leq \eta$ *and*
- ***(f is robust wrt distributional shifts)*** *for all* $\mathcal{D}^A$ *such that* $\|\sqrt{\mathcal{D}^A} - \sqrt{\mathcal{D}}\|_2 \leq K \cdot \eta^{1/4}$ *we have* $R_{\mathcal{D}^A}(f) \leq O(R_{\mathcal{D}}(f))$

then there exists an efficient interactive protocol with the following properties.

- ***(Completeness / Low Rejection Rate)*** *There exists an honest quantum prover* $\mathbf{P}^{\mathbb{O}(\mathcal{D})}$ *such that*
$$\perp_{\mathcal{D}} = 1 - \Omega\left(\frac{1}{\textit{poly}(n, T, 1/\eta)}\right).$$
- ***(Soundness / Low Risk)*** *For every Quantum Polynomial Time (QPT) prover* $\mathbf{P}$ *that is accepted by the interaction with probability* 1 *we have that with high probability*
$$AR = O(R_{\mathcal{D}}(f)).$$

For a proof sketch we refer the reader to Section 3.2.

2.3 Comparison to [GKKM20]

In this section we compare Theorem 1 to the results from [GKKM20] and in particular to the lower bound presented there.

First let us discuss the similarities and differences between the model from [GKKM20] and our model. In [GKKM20] learner $\mathbf{V}$ receives as input two sets of samples: the iid samples from $\mathcal{D}$, $x_1, \ldots, x_N$ and a set of arbitrary test examples $\tilde{x}_1, \ldots, \tilde{x}_M$. Having access to both sets $\mathbf{V}$ rejects some of $\tilde{x}_i$'s and classifies the rest. In our language we think of $\tilde{x}_1, \ldots, \tilde{x}_M$ as being generated by $\mathbf{P}$. In our model, during the training phase, $\mathbf{V}$ has access to iid samples from $\mathcal{D}$ $(x_1, \ldots, x_N)$[2]. During the test phase $\mathbf{V}$ interacts with $\mathbf{P}$ over many rounds. An honest $\mathbf{P}$ in the model from [GKKM20] receives samples $x \sim \mathcal{D}$ and forwards them to $\mathbf{V}$. For us an honest $\mathbf{P}$ receives quantum states $|\psi_{\mathcal{D}}\rangle$ and starts interacting with $\mathbf{V}$ according to our protocol. For both models we measure two quantities: the risk on accepted samples and the rejection rate when $\mathbf{P}$ acts honestly.

The models are obviously different as only ours uses quantum states. We will however proceed with a comparison as if they were the same. That is we will treat iid samples from $\mathcal{D}$ and quantum states $|\psi_{\mathcal{D}}\rangle$ as an equivalent resource and compare the number of samples/states needed for meaningful guarantees. The equivalence is justified because in an idealized setting an honest $\mathbf{P}$ generates one samples from $\mathcal{D}$ from one $|\psi_{\mathcal{D}}\rangle$. Apart from this difference we note that our protocol requires an interaction between $\mathbf{V}$ and $\mathbf{P}$ while the on in [GKKM20] does not.

Now we are ready to compare Theorem 1 to the lower bound from [GKKM20, Theorem 5.5]. In there, in order to have a non-vacuous bound on the rejection rate plus the risk on

[2] $\mathbf{V}$ can also have access to states $|\psi_{\mathcal{D}}\rangle$ during the training phase. Our result is about the test phase and the exact mechanics of the training phase are not important as long as $\mathbf{V}$ has access to f and C.

accepted arbitrary examples, one requires the number of examples to be $M = \Omega(d)$, where d is the VC-dimension of the hypothesis class.

Theorem 1 guarantees a non-vacuous bound on the rejection rate plus the risk on accepted samples when $M = \text{poly}(n, T, 1/\eta)$, where we think of M as the number of states $|\psi_{\mathcal{D}}\rangle$ that was used by $\mathbf{P}$ in the protocol. The two quantities, i.e. d and $\text{poly}(n, T, 1/\eta)$, are not comparable in general but there is a crucial difference. Our bound of $\text{poly}(n, T, 1/\eta)$ depends only on the distribution $\mathcal{D}$, because n is the dimension of the input space and T is the number of gates in C. On the other hand the lower bound of d depends only on the hypothesis class. Thus there exist tasks for which $d \gg n, T$, for some circuits C with T gates for which $\|\sqrt{\mathcal{D}_C} - \sqrt{\mathcal{D}}\|_2 \ll 1$. This implies that Theorem 1 breaks the lower bound from [GKKM20] in some regimes!

For an example of a task for which a separation holds one can take a distribution and a hypothesis class constructed in [GKKM20] that certifies their lower-bound. For $d \in \mathbb{N}$ the distribution used is the uniform distribution over $\{1, \ldots, O(d)\}$ and the hypothesis class are all functions of exactly d 1's. By construction, the VC-dimension is equal to d. Moreover, this distribution can be generated exactly ($\|\sqrt{\mathcal{D}_C} - \sqrt{\mathcal{D}}\|_2 = 0$), using quantum Fourier transform, by quantum circuits acting on $n = O(\log(d))$ qubits with $T = O(\log(d))$ gates. We compare: our guarantee gives a non-vacuous bound for $M = \text{poly}(n, T, 1/\eta) = \text{polylog}(d)$ while the lower-bound requires $M = \Omega(d)$. We see an **exponential separation**. This task has an additional property. The lower-bound holds also when the classical algorithm knows $\mathcal{D}$ exactly. This shows that access to a generator C, that is required by our construction, is not a hidden source of separation.

Note 1. *Note that we need* $\mathbf{V}$ *to obtain, during the training phase, access to* f *robust to distributional shifts and a generator* C *- a requirement that does not exist in [GKKM20]. This might imply that more samples are needed during the training phase. Our focus however is on the testing phase and the number of examples (states) in this phase.*

3 Certifiable Sampling Protocols

Now we move to proving Theorem 1. To do that we show existence of a protocol, which we name a certifiable sampling protocol. The name comes from the fact that this protocol guarantees that the samples collected by $\mathbf{V}$ came from a distribution close to the requested one.

We define the protocol in three settings (i) where the $\mathbf{V}$ has quantum capabilities (ii) where the $\mathbf{V}$ has access to a constant quantum memory (iii) where the $\mathbf{V}$ is fully classical. Because of the space restrictions we present only setting (i) and state the main result from setting (iii) in the main paper. The rest is deferred to the appendix.

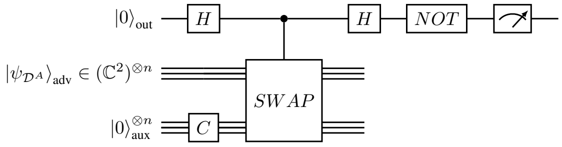
$|0\rangle_{\text{out}}$
$|\psi_{\mathcal{D}^A}\rangle_{\text{adv}} \in (\mathbb{C}^2)^{\otimes n}$
$|0\rangle^{\otimes n}_{\text{aux}}$
H
C
$SWAP$
H
NOT

Figure 1: Comparison Circuit

3.1 Quantum Verifier

In this section we present a protocol in a setting where $\mathbf{V}$ has quantum capabilities. We start with an overview and then move to a formal result.

The key component of all our protocols is a quantum circuit G, acting on three registers: out (1 qubit), adv (n qubits) and aux (n qubits), depicted in Figure 1. G is parametrized by a quantum circuit C with the associated distribution $\mathcal{D}_C$. Recall that the result of applying U_C to $0^{\otimes n}$ is the state $|\psi_{\mathcal{D}_C}\rangle$. The circuit is designed so that it measures the similarity between $\mathcal{D}^A$ and $\mathcal{D}_C$, where $\mathcal{D}^A$ is the distribution corresponding to the state $|\psi_{\mathcal{D}^A}\rangle_{\text{adv}}$. More precisely, the closer $\mathcal{D}^A$ and $\mathcal{D}_C$ are in terms of the Hellinger distance the higher the probability that G outputs 1 in the out register. We note that a circuit of this form, often referred to as the SWAP test, is a key component of many quantum algorithms [MCEM97].

Equipped with such a comparison circuit we are ready to design a protocol in a model where $\mathbf{V}$ has quantum capabilities. For now we assume that $\mathbf{P}$ acts i.i.d. in every round of the protocol (a generalization is discussed in the appendix). In the i-th round of the interaction $\mathbf{P}$ sends an n-qubit quantum state $|\psi_{\mathcal{D}^A}\rangle$ to $\mathbf{V}$, $\mathbf{V}$ samples a bit $b_i \in \{0, 1\}$ uniformly at random. If $b_i = 0$ then $\mathbf{V}$ inserts $|\psi_{\mathcal{D}^A}\rangle$ as an input to G, computes G, measures the output bit in the Z basis and records the result as γ_i. If $b = 1$ then $\mathbf{V}$ measures $|\psi_{\mathcal{D}^A}\rangle$ in the Z basis and records the outcome as $\mathbf{x}_i \in \{0, 1\}^n$. After a certain number of rounds (dependent on the desired accuracy and probability of success) $\mathbf{V}$ computes an average γ_{avg} of the set $\{\gamma_i : b_i = 0\}$. If γ_{avg} is bigger than a certain (to be determined) threshold $\mathbf{V}$ accepts the interaction and returns the set $\{\mathbf{x}_i : b_i = 1\}$.

Let us now consider the properties of this protocol. Completeness of the protocol is straightforward. An honest $\mathbf{P}$ can forward the state $|\psi_{\mathcal{D}}\rangle$ he receives to $\mathbf{V}$. For soundness of the protocol note the following facts: i) γ_{avg} is a good approximation for the probability that G outputs 1 on $|\psi_{\mathcal{D}^A}\rangle$, ii) this probability is monotonically related to $d_H(\mathcal{D}^A, \mathcal{D}_C)$ by the properties of G, iii) the samples $\{\mathbf{x}_i : b_i = 1\}$ are i.i.d. from $\mathcal{D}^A$, iv) we assumed that $d_H(\mathcal{D}, \mathcal{D}_C)$ is small. Moreover we assume that $d_H(\mathcal{D}, \mathcal{D}_C) \approx \eta$ and that η is known to $\mathbf{V}$. Combining these facts we arrive at the following conclusion. If $\mathbf{V}$ accepts the interaction then the samples it returns are i.i.d. from a distribution $\mathcal{D}^A$ such that

$$d_H(\mathcal{D}^A, \mathcal{D}) < O(d_H(\mathcal{D}_C, \mathcal{D})). \tag{1}$$

The reason the above holds is because we can set the threshold in the protocol over which $\mathbf{V}$ accepts γ_{avg} to be such that the interaction is accepted when $d_H(\mathcal{D}^A, \mathcal{D}_C) \lesssim \eta$. Then using a triangle-like inequality we arrive at (1).

Note. For the most part of the paper we assume that the $\mathbf{P}$ acts in an i.i.d. fashion. For the fully-quantum verifier we give a proof also for general setting where we drop the i.i.d. assumption. To keep the exposition manageable we do not provide general proofs for the other two models.

3.1.1 Protocol and a Proof

In this section we define the protocol formally and prove its correctness.

The protocol is defined in Figure 2. We start by assuming that $\mathbf{P}$ acts in an i.i.d. fashion and that the states $\mathbf{P}$ sends are pure. We discuss how to remove these assumptions in Appendix E.1 and E.2.

Let us prove the correctness of this protocol. The following lemma shows how the distribution of measuring the out register of G relates to the Hellinger distance of $\mathcal{D}_C$ and $\mathcal{D}^A$. The proof is deferred to Appendix F.

Lemma 1. *The probability of obtaining outcome $|1\rangle$ when measuring the out register of G executed on $|\psi_{\mathcal{D}^A}\rangle$, i.e. $\langle 0^{\otimes n}|_{aux} \langle \psi_{\mathcal{D}^A}|_{adv} \langle 0|_{out} G^\dagger \Pi^{(1)}_{out} G |0\rangle_{out} |\psi_{\mathcal{D}^A}\rangle_{adv} |0^{\otimes n}\rangle_{aux}$, is equal to $\frac{1}{2}\left(1 + (1 - d_H^2(\mathcal{D}^A, \mathcal{D}_C))^2\right)$.*

Next we show that the number of times each of the types (0 and 1) occurs is at least $N/4$ with high probability. This is a simple application of the Chernoff bound.

Lemma 2. *Let n_0, n_1 be the number of times each type occurs in the protocol from Figure 2. If $N = \Omega(\log(1/\delta))$ then $\mathbb{P}[n_0, n_1 > \frac{N}{4}] \geq 1 - \delta$.*

We are now ready to combine all the pieces and prove that the protocol from Figure 2 guarantees that if $\mathbf{V}$ accepts the interaction then with high probability the samples he collected are i.i.d. from a distribution close to $\mathcal{D}$.

Theorem 2 (Quantum Verifier)**.** *For every circuit C acting on n qubits, for every $\delta \in (0, \frac{1}{3})$, $K \in \mathbb{N}$ and all $\eta > 0$ sufficiently small there exists an interactive protocol between a quantum verifier $\mathbf{V}$ and a quantum prover $\mathbf{P}$ with the following properties. The protocol runs in $N = O\left(\frac{K}{\eta^2}\log(1/\delta)\right)$ rounds and in each round $\mathbf{P}$ sends a pure quantum state on n qubits to $\mathbf{V}$. At the end of the protocol $\mathbf{V}$ outputs $\perp$ when it rejects the interaction or it outputs $S = \{x_1, \ldots, x_{|S|}\}$, $x_i \in \{0,1\}^n$, when it accepts.*

- ***(Completeness)** There exists $\mathbf{P}^{\mathsf{O}(*)}$ such that for every $\mathcal{D} \in \mathfrak{D}(n)$ satisfying $d_H(\mathcal{D}, \mathcal{D}_C) \leq \eta$ the following holds. With probability $1 - \delta$ over the randomness in the protocol $\mathbf{P}^{\mathsf{O}(\mathcal{D})}$ succeeds, $S \sim_{\text{i.i.d.}} \mathcal{D}^{|S|}$, and $|S| \geq \Omega(K)$.*
- ***(Soundness)** For every $\mathbf{P}$ that succeeds with probability at least $\frac{2}{3}$ we have $S \sim_{\text{i.i.d.}} (\mathcal{D}^A)^{|S|}$ and $d_H(\mathcal{D}_C, \mathcal{D}^A) \leq O(\eta)$.*[3]

Note. How can we check the success probability of the prover? Assuming that the prover behaves in an i.i.d. fashion, it suffices to run the protocol $(2/\epsilon)\log(1/\delta)$ times. If the fraction of successes is bigger than $1 - \epsilon/2$ then we know with confidence $1 - \delta$ that the success probability is at least $1 - \epsilon$.

Remark 1. *We note that for certifiable sampling protocols we use the number of repetitions instead of the rejection rates and risk on arbitrary examples. This phrasing is better suited for these protocols. We stated Theorem 1 differently to easily compare it to [GKKM20]. For instance, in Theorem 2 we state that by performing $(\frac{K}{\eta^2}\log(1/\delta))$ repetitions of the protocol, when the prover acts honestly, we collect at least K samples with high probability. However, note that this theorem could also be stated as the probability of collecting a sample at each repetition beeing η^2, i.e. the rejection rate $\perp_{\mathcal{D}} = 1 - \Omega(\eta^2)$.*

Proof. We start with the completeness property and then move to soundness.

Completeness. An honest $\mathbf{P}^{\mathsf{O}(\mathcal{D})}$ obtains $|\psi_{\mathcal{D}}\rangle$ from $\mathsf{O}(\mathcal{D})$ and forwards it to $\mathbf{V}$. Lemma 2 guarantees that with probability $1 - \frac{\delta}{2}$, $n_0, n_1 = \Omega(\frac{K}{\eta^2}\log(1/\delta))$. This automatically guarantees that $|S| \geq \Omega(K)$. Moreover by Fact 3 we have that with probability $1 - \frac{\delta}{2}$

$$\left|p - \langle 0^n|_{\text{aux}} \langle \psi_{\mathcal{D}}|_{\text{adv}} \langle 0|_{\text{out}} G^\dagger \Pi^{(1)}_{\text{out}} G |0\rangle_{\text{out}} |\psi_{\mathcal{D}}\rangle_{\text{adv}} |0^n\rangle_{\text{aux}}\right| \leq \eta^2. \tag{2}$$

By Corollary 1 we thus get that $\left|p - \frac{1}{2}\left(1 + (1 - d_H^2(\mathcal{D}, \mathcal{D}_C))^2\right)\right| \leq \eta^2$ holds with probability $1 - \frac{\delta}{2}$. By assumption $d_H(\mathcal{D}, \mathcal{D}_C) \leq \eta$ so we get that $p \geq 1 - 2\eta^2$ (as a function $\frac{1}{2}(1 + (1 - x^2)^2)$ is decreasing). This means that $\mathbf{P}^{\mathsf{O}(\mathcal{D})}$ succeeds with probability $1 - \delta/2$.

By the union bound over the error events with probability $1 - \delta/2 - \delta/2 = 1 - \delta$ we have that $|S| \geq \Omega(K)$ and $\mathbf{P}^{\mathsf{O}(\mathcal{D})}$ succeeds. The property $S \sim_{\text{i.i.d.}} \mathcal{D}^{|S|}$ holds because the state sent by $\mathbf{P}$ to $\mathbf{V}$ is equal to $|\psi_{\mathcal{D}}\rangle$.

Soundness. By Corollary 1 we get that $\left|p - \frac{1}{2}\left(1 + (1 - d_H^2(\mathcal{D}^A, \mathcal{D}_C))^2\right)\right| \leq \eta^2$ with probability $1 - \frac{\delta}{2}$. $\mathbf{P}$ succeeds with probability $\frac{2}{3}$ so by

[3] $\mathcal{D}^A$ is the implicit distribution from which we collect the samples, which is the distribution corresponding to $|\psi_{\mathcal{D}_A}\rangle$

```
Verifier V                                                              Prover P
$p := 0, S := \emptyset$
for $i \in \{1, \dots N\}$ where $N = O\left(\frac{K}{\eta^2}\log(1/\delta)\right)$
 do :
  type $\overset{\$}{\leftarrow} \{0,1\}$
                                                                        $|\psi_{\mathcal{D}^A}\rangle \in (\mathbb{C}^2)^{\otimes n}$
                                          $\xleftarrow{\quad |\psi_{\mathcal{D}^A}\rangle \quad}$
   if type $= 0$ :
     Set $b_1, b_2, \dots, b_n$ to be the measurement of
      $|\psi_{\mathcal{D}^A}\rangle$ in the $Z$ basis
     $S := S \cup \{(b_j)_{j \in \{1,\dots,n\}}\}$
   if type $= 1$ :
     Set $p_i$ to be the measurement of the out register of
     $G\,|0\rangle_{\text{out}}\,|\psi_{\mathcal{D}^A}\rangle_{\text{adv}}\,|0^{\otimes n}\rangle_{\text{aux}}$ in the Z basis
     $p := p + \frac{p_i}{N}$
done :
if $p < 1 - 2\eta^2$   abort
return $S$
```

Figure 2: The interactive protocol for the model where the verifier has access to a quantum computer.

the union bound and the fact that $\frac{1}{3} + \frac{\delta}{2} < 1$ we get that $h(d_H(\mathcal{D}^A, \mathcal{D}_C)) \geq p - \eta^2 \geq 1 - 3\eta^2$, where we used h to denote the function $\frac{1}{2}(1 + (1 - x^2)^2)$. As h is a decreasing function we get that that $d_H(\mathcal{D}^A, \mathcal{D}_C) \leq \sqrt{1 - \sqrt{2(1 - 3\eta^2) - 1}} \leq 10\eta$, for sufficiently small η. □

3.2 Classical Verifier

In this section we state the certifiable sampling protocol guarantee where $\mathbf{V}$ is fully classical.

To achieve this we build on ideas for delegation of quantum computation with a classical verifier from [Mah18]. Similar ideas were used in [BCM+21] to generate certified randomness. On a high level, we design a protocol that forces $\mathbf{P}$ to commit to an n-qubit state $|\phi\rangle$, then receive instructions for measurements from $\mathbf{V}$, measure $|\phi\rangle$ accordingly and report the results back to $\mathbf{V}$. To do that we restrict $\mathbf{P}$ - as was also done in [Mah18] - to be QPT. From the technical side this procedure relies on the concept of claw-free functions, which can be understood as a post quantum cryptography scheme that allows $\mathbf{V}$ to control the computation performed by $\mathbf{P}$.

We summarize by saying that also in this model we arrive at a result similar to Theorem 2. As before, for sake of simplicity, some quantifiers of the theorem have been removed, a more in detail version of the theorem can be found in Appendix D.2, Theorem 8 (completeness) and Theorem 9 (soundness), alongside their proofs.

Theorem 3 (Classical Verifier)**.** *For a security parameter λ, every generative circuit C acting on n qubits, for every $K \in \mathbb{N}$ and all $\delta, \eta > 0$ sufficiently small there exists an interactive protocol $(\mathbf{V}, *)$ between a classical verifier $\mathbf{V}$ and a quantum prover $\mathbf{P}$ with the following properties. The protocol runs in $N = O\left(\frac{K}{\eta^4} poly(n,T)\log(1/\delta)\right)$ rounds and in each round $\mathbf{P}$ and $\mathbf{V}$ exchange $poly(n, T, \lambda)$ bits. At the end of the protocol $\mathbf{V}$ outputs $\perp$ when it rejects the interaction or it outputs $S = \{x_1, \dots, x_{|S|}\}$, $x_i \in \{0,1\}^n$, when it accepts.*

- ***(Completeness)** There exists a QPT prover $\mathbf{P}^{\mathsf{O}(*)}$ such that for every $\mathcal{D} \in \mathfrak{D}(n)$ satisfying $d_H(\mathcal{D}, \mathcal{D}_C) \leq \eta$ the following holds. With probability $1 - \delta$ over the randomness in the protocol $\mathbf{P}^{\mathsf{O}(\mathcal{D})}$ succeeds, $S \sim_{\text{i.i.d.}} \mathcal{D}^{|S|}$, and $|S| \geq \Omega(K)$.*

- ***(Soundness)** For every QPT bounded $\mathbf{P}$ that succeeds with probability 1 we have that with probability $1 - \delta - \mu(\lambda)$ the following conditions hold: $S \sim_{\text{i.i.d.}} (\mathcal{D}^A)^{|S|}$ and $d_H(\mathcal{D}_C, \mathcal{D}^A) \leq O(\eta^{1/4})$, where μ is a negligible function.*[4]

Remark 2. *Just as mentioned in remark 1, the completeness can be stated as $\perp_{\mathcal{D}} = 1 - \Omega(\eta^4 \frac{1}{poly(n,T)})$.*

Having Theorem 3 it is quite straightforward to prove Theorem 1. We provide a short proof sketch.

Theorem 1(sketch). Let us assume that there exists an efficient verifier $\mathbf{V}$ having access to classifier f and a description of circuit C, satisfying the conditions in the theorem 1,

[4] A negligible function is a function that decays faster than any inverse polynomial, e.g. $f(n) = 1/2^n$.

i.e. $\mathcal{D}_C$ is a good approximation of $\mathcal{D}$ and f is robust wrt distributional shifts of maximum distance $M \cdot \eta^{1/4}$ in Hellinger distance. Now according to Theorem 3, for this C, η, and a security parameter λ, there exists an efficient classical verifier $\mathbf{V}$ interacting with a QPT prover $\mathbf{P}$ that satisfies the soundness and completeness properties of theorem 3.

Due to the completeness statement of Theorem 3, there exists an honest prover $\mathbf{P}$ that is given access to the distribution $\mathcal{D}$, and partaking in $O(\frac{K}{\eta^4}\text{poly}(n,T)\log(1/\delta))$ repetitions of the protocol, returns samples from the same distribution and is accepted with probability $1-\delta$ for small δ, and collects K examples from $\mathcal{D}$. Now as noted in remark 2 we have,

$$\begin{aligned}\perp_{\mathcal{D}} &= 1 - \mathbb{E}\left[\frac{\#\text{number of samples collected by } \mathbf{V}}{\#\text{number of states used by } \mathbf{P}}\right] \\ &= 1 - \Omega(\eta^4 \frac{1}{\text{poly}(n,T)})\end{aligned}$$

so the completeness statement of Theorem 1 holds.

For the soundness, due to the soundness statement of Theorem 3, for any QPT bounded (in n and λ) prover $\mathbf{P}$, if $\mathbf{P}$ is accepted with probability 1, with confidence $1-\delta$ we know that if the samples given by the adversary follow a distribution $\mathcal{D}^A$, $d_H(\mathcal{D}^A, \mathcal{D}) \leq O(\eta^{1/4})$. Now using the second assumption in the statement of theorem 1, i.e. (f is robust wrt distributional shifts), as $d_H(\mathcal{D}^A, \mathcal{D}) \leq O(\eta^{1/4})$, we have that $AR = R_{\mathcal{D}^A}(f) \leq O(R_{\mathcal{D}}(f))$, which concludes the soundness proof.

□

4 Future Work

There are some important open questions that we leave for future research. First, one can ask questions of quantitative nature. Can the bounds on rejection rates and risk obtained in our work be improved? We can hope for instance to improve the rejection rate bound that depends polynomially on T (number of gates in the circuit) to depend polynomially on the depth of the citcuit A different interesting open problem would be to make the protocol non-interactive, one could imagine a prover that attaches a proof that a sample came from the right distribution. Doing this in the random oracle model, similarly to [ACGH20], might be feasible.

Second, our guarantees rely on the verifier having access to a generating circuit and a classifier robust wrt distributional shifts. It would be interesting to relax any of these assumptions to lower the sample complexity during training phase.

Last, our result is of a theoretical nature; it shows that a classical lower bound can be broken in the quantum model. But it would also be of great interest to find real-world applications where the underlying model is indeed of quantum nature and hence our model applies directly. As discussed, it does not seem feasible to imitate our approach when the nature does not provide superposition of samples, as our protocol innately uses principles of quantum mechanics to restrain adversarial behaviour. However, finding a new model, neither PAC nor quantum-PAC, in which some of the ideas could be emulated and results with similar guaranties could be proven, might be possible.

5 Conclusions

We introduce a new quantum model and a protocol for classifying arbitrary text examples. Our protocol breaks a classical lower bound for the amount of resources needed to achieve non-vacuous bounds on the rejection rate of clean examples and the risk on arbitrary test examples. In order to achieve that goal, we extend the delegation protocol from [Mah18] by providing a new functionality, namely the ability to collect samples from a distribution. Our results show the potential utility of quantum capabilities for addressing the adversarial robustness problem.

6 Acknowledgments

We wish to thank Thomas Vidick for helpful discussions and fantastic lecture notes which helped us significantly with the classical verification techniques. Khashayar Barooti was supported by the Swiss National Science Foundation (SNF) through the project grant 192364 on Post Quantum Cryptography.

References

[ACGH20] Gorjan Alagic, Andrew M. Childs, Alex B. Grilo, and Shih-Han Hung. Non-interactive classical verification of quantum computation. In Rafael Pass and Krzysztof Pietrzak, editors, *Theory of Cryptography - 18th International Conference, TCC 2020, Durham, NC, USA, November 16-19, 2020, Proceedings, Part III*, volume 12552 of *Lecture Notes in Computer Science*, pages 153–180. Springer, 2020.

[BCM$^+$21] Zvika Brakerski, Paul Christiano, Urmila Mahadev, Umesh Vazirani, and Thomas Vidick. A cryptographic test of quantumness and certifiable randomness from a single quantum device. *J. ACM*, 68(5), August 2021.

[BJ95] Nader H. Bshouty and Jeffrey C. Jackson. Learning dnf over the uniform distribution using a quantum example oracle. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory*, COLT '95, page 118–127, New York, NY, USA, 1995. Association for Computing Machinery.

[BL08] Jacob D. Biamonte and Peter J. Love. Realizable hamiltonians for universal adiabatic quantum computers. *Phys. Rev. A*, 78:012352, Jul 2008.

[BS21] Sebastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[Chu65] J. T. Chu. Optimal decision functions for computer character recognition. *J. ACM*, 12(2):213–226, April 1965.

[ETT$^+$19] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling CNNs with simple transformations, 2019.

[FHcvM18] Joseph F. Fitzsimons, Michal Hajdušek, and Tomoyuki Morimae. Post hoc verification of quantum computation. *Phys. Rev. Lett.*, 120:040501, Jan 2018.

[GKKM20] Shafi Goldwasser, Adam Tauman Kalai, Yael Kalai, and Omar Montasser. Beyond perturbations: Learning guarantees with arbitrary adversarial test examples. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[GU21] Grzegorz Gluch and Rüdiger L. Urbanke. Adversarial robustness: What fools you makes you stronger. *CoRR*, abs/2102.05475, 2021.

[KSV02] A. Yu. Kitaev, A. H. Shen, and M. N. Vyalyi. *Classical and Quantum Computation*. American Mathematical Society, USA, 2002.

[Mah18] Urmila Mahadev. Classical verification of quantum computations. *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 259–267, 2018.

[MCEM97] Cleve Ekert Macchiavello, B Y R. Cleve, A. Ekert, and C. Macchiavello. Quantum algorithms revisited. In *Proceedings of the Royal Society of London A*, pages 339–354, 1997.

[MHS21] Omar Montasser, Steve Hanneke, and Nathan Srebro. Adversarially robust learning with unknown perturbation sets. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3452–3482. PMLR, 15–19 Aug 2021.

[NC11] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, USA, 10th edition, 2011.

[NYC15] Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, pages 427–436. IEEE Computer Society, 2015.

[RSL18] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[SZS$^+$14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[TB19] Florian Tramer and Dan Boneh. Adversarial training and robustness for multiple perturbations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 5866–5876. Curran Associates, Inc., 2019.

[Unr16] Dominique Unruh. Computationally binding quantum commitments. In Marc Fischlin and Jean-Sébastien Coron, editors, *Advances in Cryptology - EUROCRYPT 2016 - 35th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Vienna, Austria, May 8-12, 2016, Proceedings, Part II*, volume 9666 of *Lecture Notes in Computer Science*, pages 497–527. Springer, 2016.

[WK18] Eric Wong and J. Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 5283–5292, 2018.

[YTL$^+$19] Kan Yuan, Di Tang, Xiaojing Liao, XiaoFeng Wang, Xuan Feng, Yi Chen, Menghan Sun, Haoran Lu, and Kehuan Zhang. Stealthy porn:

Understanding real-world adversarial images for illicit online promotion. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 952–966. IEEE, 2019.

Breaking a Classical Barrier for Classifying Arbitrary Test Examples in the Quantum Model: Supplementary Materials

A Background and Preliminaries

A.1 A crash course on Quantum Mechanics

In this section we introduce some basic notions of quantum mechanics, aiming to aid the reader unfamiliar with the subject. For an in depth introduction we recommend [NC11]. We start by introducing the most basic quantum system, a qubit.

Definition 4. *A qubit is a tuple $(|\psi\rangle, \mathcal{H}, X, Z)$, where $\mathcal{H}$ is a Hilbert space, $|\psi\rangle \in \mathcal{H}$ is a unit vector, and X, Z are two observables (Hermitian operators on $\mathcal{H}$) such that,*

$$(XZ - ZX)|\psi\rangle = 0 \tag{3}$$

In words, one typically refers to this property by saying that X and Z anticommute on the support of $|\psi\rangle$.

What anticommuting means on a high level, is that there are two ways to observe $|\psi\rangle$ but it is not possible to observe $|\psi\rangle$ in both ways simultaneously. The simplest type of qubits we encounter are defined over $\mathcal{H} = \mathbb{C}^2$. A standard choice of basis for $\mathbb{C}^2$ is,

$$|0\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \text{ and } |1\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{4}$$

According to this choice, the state $|\psi\rangle$ can be represented as $\alpha|0\rangle + \beta|1\rangle, \alpha, \beta \in \mathbb{C}$ such that $|\alpha|^2 + |\beta|^2 = 1$. For many reasons, that we will not delve into here, it is natural to represent states by matrices instead of vectors. Important examples of observables are the Pauli matrices, σ_X and σ_Z where,

$$\sigma_X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \text{ and } \sigma_Z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \tag{5}$$

We often refer to σ_Z as the computational basis measurement and σ_X as the Hadamard basis measurement. The reason these specific two observables play a special role is the following lemma.

Lemma 3. *Let $(|\psi\rangle, \mathcal{H}, X, Z)$ be a qubit. There exists a Hilbert space $\mathcal{H}'$ and an isometry $V : \mathcal{H} \to \mathbb{C}^2 \otimes \mathcal{H}'$, such that,*

$$VX|\psi\rangle = (\sigma_X \otimes \mathbb{I})V|\psi\rangle,$$
$$VZ|\psi\rangle = (\sigma_Z \otimes \mathbb{I})V|\psi\rangle.$$

This lemma states that up to an isometry every qubit can be seen as a state on $\mathbb{C}^2$ with two observables that are the Pauli matrices.

Next let us discuss the evolution of a quantum state. Due to laws of quantum mechanics, every operation (not containing a measurement) that is performed on a quantum state can be viewed as a unitary operator. This raises the following question: Many operations done in classical computation are not reversible. Hence, how can one perform such classical computations with a quantum computer? The following theorem bridges this gap.

Theorem 4. *Let $|\psi\rangle = \sum_{x \in \mathcal{X}} \alpha_x |x\rangle$ be a quantum state, and $f : \mathcal{X} \to \mathcal{Y}$ be a function which is efficiently computable by a classical circuit. Then there exists an efficiently computable unitary U_f such that for all $x \in \mathcal{X}, y \in \mathcal{Y}$,*

$$U_f : |x\rangle|y\rangle \mapsto |x\rangle|y + f(x)\rangle$$

This theorem allows us to evaluate functions on superposition, which is one of the main reasons quantum computing allows us to solve problems which are considered classically hard to solve.

Now that we have introduced what a qubit is and how a quantum state evolves, we can talk about the notion of measurement. Measurements allow us to observe properties of a quantum state. In the real world this can be seen as measuring the energy of a state or the spin of a particle. Measurements are often modeled as observables, let us denote one by a Hermitian operator O. As O is a Hermitian it can be decomposed as $O = \sum_i \lambda_i \Pi_i$, where Π_i is the projection onto the eigenspace corresponding to eigenvalue λ_i. The eigenvalues are referred to as measurement outcomes and the probability of observing λ_i when measuring it on $|\psi\rangle$ is given by $\langle\psi| \Pi_i |\psi\rangle$.

There are two types of measurements, namely a projector-valued measurement (PVM) and positive operator-valued measurement (POVM). A PVM is defined by a set of projection $\{\Pi_i\}$ such that $\sum_i \Pi_i = \mathbb{I}$. The probability of obtaining measurement outcome i when measuring it on $|\psi\rangle$ is given by $\langle\psi| \Pi_i |\psi\rangle$. POVMs are a generalization of PVMs. For a POVM $\{\Pi_i\}$, Π_i's are not necessarily projections but can be any positive operators. We still have the requirement that $\sum_i \Pi_i = \mathbb{I}$ and the law that the probability of observing outcome i is given by $\langle\psi| \Pi_i |\psi\rangle$.

Let us also talk about the post measurement state. For a POVM $\{\Pi_i\}$, the post measurement state of $|\psi'\rangle$, if after measuring Π on $|\psi\rangle$ the outcome is i then $|\psi'\rangle = \frac{\Pi_i |\psi\rangle}{\langle\psi|\Pi_i|\psi\rangle}$. Another interesting fact is that any POVM can be represented by a PVM on a bigger space. This is referred to as Neimark's dilation theorem.

The last topic we touch upon in this intro is the topic of mixed states and density operators. As we mentioned, a pure state is a unit vector in a Hilbert space $\mathcal{H}$. One might ask what happens when a machine/party does not have access to the whole space, e.g. $\mathcal{H} = \mathcal{H}_A \otimes \mathcal{H}_B$ where machine/party A only has access to $\mathcal{H}_A$. Now although $|\psi\rangle$ is a pure state on $\mathcal{H}$, the part of $|\psi\rangle$ that is visible to A (the part in $\mathcal{H}_A$) might behave differently, as there might be correlations between this part and the part of the state A doesn't have access to. These correlations are referred to as entanglement.

Any bipartite pure state has a Schmidt decomposition,

$$|\psi\rangle_{AB} = \sum_i \sqrt{\lambda_i}\, |u_i\rangle_A \, |v_i\rangle_B \tag{6}$$

where u_i, v_i are orthonormal bases of $\mathcal{H}_A$ and $\mathcal{H}_B$ respectively. Now looking at the view of A of the state $|\psi\rangle$, this can be represented as a density matrix $\rho_A = \sum_i \lambda_i |u_i\rangle \langle u_i|$. This is referred to as the reduced density of $|\psi\rangle$ on $\mathcal{H}_A$. Now one can consider a scenario in which the bipartite state itself is a density operator $\rho \in \text{Pos}(\mathcal{H}_{AB})$. Now the reduced density of ρ on $\mathcal{H}_A$ is defined as $\rho_A = \text{Tr}_B(\rho_{AB})$. This allows us to define the probability of getting outcome i, when measuring a POVM $\{\Pi_i\}$ on $\rho \in \mathcal{H}$. This probability is given by $\text{Tr}(\Pi_i \rho)$.

Conditioned on the measurement outcome being i, the post measurement state would be $\rho' = \frac{\Pi_i \rho \Pi_i}{\text{Tr}(\Pi_i \rho)}$.

B Related Work

The standard approach to adversarial machine learning is to consider perturbation sets. These sets describe by how much the adversary is allowed to perturb the input. The most commonly considered such perturbation sets describe perturbations that are bounded in ℓ_p norms [RSL18, WK18, BS21], but other perturbations (rotations, shifts, etc.) were also considered [ETT+19]. To date there is a considerable literature on this approach.

Whether those assumed bounds capture real-world scenarios is of course up for debate. E.g., it has been shown that defending models against one perturbation set does not necessarily improve the robustness against other perturbations and that there exist a trade-off between robustness for different perturbations [TB19].

Some of the literature tries to escape the assumption of ℓ_p bounded perturbations or fixed perturbation sets altogether. E.g., in [MHS21] it is shown how to defend against adversaries that are allowed to use perturbations from a set that is not known to the defender. Unfortunately, there are cases where the defense requires exponentially (in the VC-dimension) many samples. Another approach was considered in [GKKM20]. The authors show how to get rid of the limitations on perturbations completely. This, naturally, comes at a price: e.g., the results in [GKKM20] are based on the assumption that the learner can decide not to give an answer for some inputs (selective learning [Chu65]) and that they see the test set upfront (transductive learning), thus the capabilities of the learner are enhanced. A recent work by [GU21] proposes another approach for removing restrictions on perturbation sets. In this paper it is shown that an effective adversary can be used to design a faster learning algorithm. If one assumes that the learning problem is hard, this approach leads to a defense. The limitation of this approach is that the adversary is required to generate adversarial examples at random.

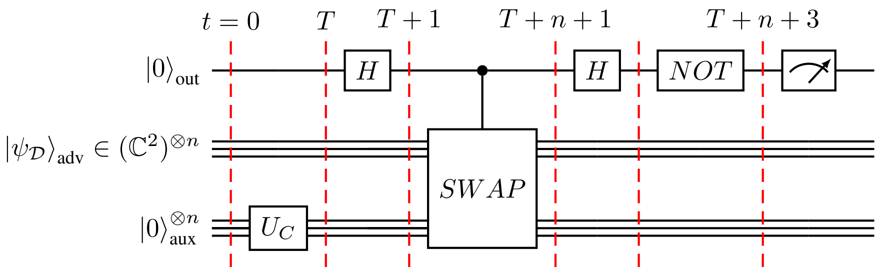
$t = 0$
T
$T + 1$
$T + n + 1$
$T + n + 3$
$|0\rangle_{\text{out}}$
$|\psi_{\mathcal{D}}\rangle_{\text{adv}} \in (\mathbb{C}^2)^{\otimes n}$
$|0\rangle^{\otimes n}_{\text{aux}}$
H
H
NOT
$SWAP$
U_C

Figure 3: Comparison Circuit with Time Slices

From the discussion above it seems that improving the state of the art of robustness for ℓ_p bounded perturbations might not guarantee a satisfying solution to the adversarial robustness puzzle in the long run. Instead, an exploration of new models is likely needed for a principled resolution of the problem.

C Overview

In this section we give an overview of how to generalize the protocol from Figure 2 to, first, the setting where $\mathbf{V}$ has access to a constant memory quantum computer and then to a setting where $\mathbf{V}$ is fully classical. We present it this way, as the protocols in the consecutive settings build on top of each other.

Constant Memory Quantum Verifier. In this model the messages in the protocol can still be quantum. (But we will see that in our protocol only $\mathbf{P}$ will send quantum states and $\mathbf{V}$ will send only classical messages.) But $\mathbf{V}$ now only has access to a constant-size quantum computer and can store only a constant number of qubits at each point in time $\mathbf{V}$. The only operation that will be required from $\mathbf{V}$ is measuring the qubits sent by $\mathbf{P}$ in either the Z or the X basis. Protocols of this form are called receive-and-measure protocols and were already previously considered in the literature, see e.g. [FHcvM18].

Our goal is to emulate the protocol that we designed in the previous step in this more restrictive constant-quantum memory model. The idea is the following. We let $\mathbf{P}$ choose an n-qubit state $|\psi_{\mathcal{D}^A}\rangle$ and then force her to create a state $|\phi\rangle$ that depends on $|\psi_{\mathcal{D}^A}\rangle$ and to send this state to $\mathbf{V}$.[5] The state $|\phi\rangle$ should satisfy the following properties. When $\mathbf{V}$ measures $|\phi\rangle$ in the Z basis then (i) with probability $\Omega(1/T)$ the distribution of outcomes of measuring one of the qubits is close to the distribution of measuring the output qubit of $G|0\rangle|\psi_{\mathcal{D}^A}\rangle|0^{\otimes n}\rangle$ in the Z basis (ii) with probability $\Omega(1/T)$ $\mathbf{V}$ can obtain $\mathbf{x}' \sim \mathcal{D}^A$.[6] These two operations emulate the steps $\mathbf{V}$ performed in the previous protocol for $b = 0$ and $b = 1$, respectively. Note that the operations succeed only with probability $\Omega(1/T)$ but this suffice for our purpose. The main question is how to force $\mathbf{P}$ to create $|\phi\rangle$ with these properties?

To solve this problem we use the well-known circuit-to-Hamiltonian reduction introduced in [KSV02]. This reduction was originally used to show that a local Hamiltonian problem is QMA-complete. Later on it was a crucial component in the the delegation of quantum computation in the constant quantum memory model [FHcvM18] and in the delegation of quantum computation with a classical verifier in [Mah18]. Unfortunately, we can not use the reduction in a black-box manner. The main issue is, the reduction is designed for decision problems, and our problem of interest is a sampling problem. Hence, in order to use the Hamiltonian model, one would need to modify the reduction to adapt sampling problems.

What is the purpose of this reduction in our context? The circuit-to-Hamiltonian reduction allows to reduce the computation of a quantum circuit G to estimating an energy of a state $|\rho\rangle$ with respect to a local Hamiltonian H_G. In particular, it allows us to build a protocol that forces $\mathbf{P}$ to prepare a so-called *history state* $|\phi\rangle$ of G. Assume that $\mathbf{P}$ chooses to evaluate G on a state $|\psi_{\mathcal{D}^A}\rangle$. Assume further that the circuit C has T gates and denote $T+n+3$ by T'. Then denote by $|\xi_0\rangle, |\xi_1\rangle, \ldots, |\xi_{T+n+3}\rangle$ the $(2n+1)$-qubit states, where $|\xi_i\rangle$ is the state after the first i gates of G are performed on $|\psi_{\mathcal{D}^A}\rangle$. We refer the reader to Figure 3, where the ξ_i's are depicted as time slices in G. With this notation the history state is defined as:

$$|\phi\rangle := \frac{1}{\sqrt{T'+1}} \left(|0\rangle_{\text{clock}} |\xi_0\rangle_{\text{comp}} + |1\rangle_{\text{clock}} |\xi_1\rangle_{\text{comp}} + |2\rangle_{\text{clock}} |\xi_2\rangle_{\text{comp}} + \cdots + |T'\rangle_{\text{clock}} |\xi_{T'}\rangle_{\text{comp}} \right).$$

[5] By sending the state to $\mathbf{V}$ we mean sending the state one qubit at a time. Whenever a qubit arrives to $\mathbf{V}$ he has a choice whether to keep it or discard it. At all times the number of qubits $\mathbf{V}$ stores cannot exceed the constant predefined number.

[6] Note that $\mathbf{x}' \in \{0,1\}^n$ but $\mathbf{V}$ has only a constant quantum memory. But it is possible to realize a protocol with these properties. Imagine that while the qubits come to $\mathbf{V}$ one by one he measures a qubit, records the result and discards the qubit making room for the next ones. In total he collects many measurement outcomes out of which he can create $\mathbf{x}'$.

Hence, $|\phi\rangle$ represents a history of the evaluation of G. It is a superposition of states of the circuit after applying $0, 1, 2 \ldots, T'$ gates of the circuit tensored with a state representing a *clock*. We denoted by comp the concatenation of the three registers out, adv, aux. For instance $|\xi_0\rangle_{\text{comp}} = |0\rangle_{\text{out}} |\psi_{\mathcal{D}^A}\rangle_{\text{adv}} |0^{\otimes n}\rangle_{\text{aux}}$.

Assume for now that $\mathbf{P}$ sends $|\phi\rangle$ to $\mathbf{V}$. We will show that with such a state it is possible to realize the two properties we were hoping for. $\mathbf{V}$ measures $|\eta\rangle$ in the Z basis and depending on the outcome of measuring the clock register performs further actions.

If the outcome of measuring the clock register is equal to T', which by definition of $|\phi\rangle$ happens with probability $\frac{1}{T'+1}$, then the distribution of measuring the out register is exactly equal to the desired distribution. This is because $|\xi_{T'}\rangle_{\text{comp}}$ represents the last slice of the computation of G (see Figure 3).

If the outcome of measuring the clock register is equal to 0 then the distribution of measuring the adv register is exactly equal to $\mathcal{D}^A$. This is because $|\xi_0\rangle_{\text{comp}} = |0\rangle |\psi_{\mathcal{D}^A}\rangle |0^{\otimes n}\rangle$ represents the first slice of the computation of G (see Figure 3). Note that in the final protocol we also check if the outcomes of measuring the out and aux registers are all 0. This is done for technical reasons to simplify the proof of soundness.

We realized the two properties we were looking for. Now we can emulate the protocol described in the first step (Quantum Verifier). Thus we will obtain a result similar to Theorem 2 also in this setting. Note that in each of the cases we were succeeding only with probability $\approx 1/T'$. This will influence the guarantee of Theorem 2 in this model. In particular, this will imply that we will recover 1 sample from $\mathcal{D}$ for every T' states $|\psi_{\mathcal{D}}\rangle$ provided to an honest $\mathbf{P}$.

In Section D.1 we will explain in more detail what it formally means that we can force $\mathbf{P}$ to produce the history state. In short, the circuit-to-Hamiltonian reduction allows $\mathbf{V}$ to perform local (which means involving only few qubits) checks on the state obtained from $\mathbf{P}$ to check that it is in fact a history state. These local checks and the whole reduction has a flavor similar to the famous Cook-Levin proof that shows that 3-SAT is NP-complete.

Classical Verifier. In the last model we consider $\mathbf{V}$ that is classical and all exchanged messages are also classical. To make our protocol work we need to impose a computational restriction on $\mathbf{P}$, namely we assume that $\mathbf{P}$ is in QPT- Quantum Polynomial Time.

The goal now is to adopt the protocol from the previous step to this model. The protocol can be understood as forcing $\mathbf{P}$ to construct a history state by performing checks (measurements in the X or the Z basis) that involve only constant number of qubits. In the model where the communication is only classical we need to somehow force $\mathbf{P}$ to perform the measurements chosen by $\mathbf{V}$ and report the result of these measurements back to $\mathbf{V}$.

To achieve this we use an idea that was a crucial component in the delegation of quantum computation with a classical verifier in [Mah18]. A similar idea was used in [BCM$^+$21] to generate certified randomness with a classical verifier. On a high level, we design a protocol that forces $\mathbf{P}$ to commit to an n-qubit state $|\phi\rangle$, then receive instructions for measurements from $\mathbf{V}$, measure $|\phi\rangle$ accordingly and report the results back to $\mathbf{V}$. This procedure relies on the concept of claw-free functions, which can be understood as a post quantum cryptography scheme that allows $\mathbf{V}$ to control the computation performed by $\mathbf{P}$.

We will not go into more detail in this overview and we refer the reader to Section D.2 for a comprehensive treatment. We summarize the overview by saying that also in this model we arrive at a result similar to Theorem 2. This means that if we assume that we are able to learn generative models close to the true distribution in the Hellinger distance and our classifiers are resistant to small distributional shifts in the same distance measure then adversarial robustness is solved in a model where a classical $\mathbf{V}$ interacts with a quantum $\mathbf{P}$ who receives quantum samples from nature.

D Towards The Classical Verifier Protocol

In this section we describe the transition from the qunantum verifier protocol described in section 3.1 to the protocol with classical communication, and a fully classical verifier (learner). For the sake of simplicity we do this in 2 steps. First we introduce a protocol which still requires quantum communication and the verifier is also quantum, but the verifier requires only a constant number of qubits. This protocol is categorized as a receive and measure protocol in the literature. The second transition is done by delegating the measurements done on the verifier side in the receive and measure protocol to the prover. The transformation here is quite similar to the measurement protocol introduced in [Mah18]. We proceed by describing the constant memory quantum verifier.

Note: Often, for the sake of convenience, we drop the identity operators tensored to projection operators. For instance, instead of $I \otimes |1\rangle\langle 1| \otimes I$ we often write $|1\rangle\langle 1|$. Hence, the reader should keep in mind that there are often I matrices tensored to the projection in order to make the dimensions match.

D.1 Constant Memory Quantum Verifier

In section 3.1 we described a protocol in which a verifier $\mathbf{V}$ can certify that the distribution of the samples they get from the prover $\mathbf{P}$ is η-close to the distribution of the samples given by the nature. However, this protocol required $\mathbf{V}$ to perform computation on $2n+1$-qubit states, whereas here we assume quantum memory of $\mathbf{P}$ is constant.

We proceed by describing a protocol, achieving the same goal, in which $\mathbf{V}$ can perform operations only on constant number of qubits.[7] On a high level $\mathbf{V}$ wants to outsource the execution of the comparison circuit G to $\mathbf{P}$. Intuitively we want $\mathbf{P}$ to send to $\mathbf{V}$ a state that certifies execution of G. This is possible by modifying a well-known result called circuit-to-Hamiltonian reduction.

Circuit-to-Hamiltonian reduction. This reduction was introduced by Kitaev in the late 1990's, see [KSV02]. This reduction allows one to reduce a computation of a quantum circuit to estimating the ground energy of a local Hamiltonian. With such a tool in hand $\mathbf{V}$ can first perform the reduction to create H_G, send a classical description of H_G to $\mathbf{P}$, then $\mathbf{P}$ is supposed to send a low energy state $|\psi\rangle$ of H_G back to $\mathbf{V}$, and finally $\mathbf{V}$ estimates the energy of $|\psi\rangle$ with respect to H_G to verify that it is indeed of low energy.

For our purposes we need a slight modification of the standard reduction. Due to this fact, here we give an overview of this classical result and point to the differences needed for our setup. The main difference is, the output of the circuits we are concerned with are not single bit, and also a portion of the input $\psi_{\mathcal{D}}$, is plugged directly by the prover and $\mathbf{V}$ does not know what this input is, hence the hamiltonian can not have penalization terms based on a portion of the input and the output of the circuit, otherwise $\mathbf{V}$ would not be able to compute this hamiltonian. We follow the approach from [KSV02] and we refer the reader to this book for more details.

The starting point of the reduction is the comparison circuit G[8]. Recall that G acts on three registers: out (1 qubit), adv (n qubits), aux (n qubits) and the output of the circuit is obtained by measuring the out register in the Z basis. We want to find an object called a local Hamiltonian H_G.

Definition 5. *We say that an operator $H : (\mathbb{C}^2)^{\otimes N} \to (\mathbb{C}^2)^{\otimes N}$ on N qubits is a k-local Hamiltonian if H is expressible as $H = \sum_{r=1}^{j} H_j$, where each H_j is a Hermitian operator acting on k qubits.*

Our goal will be to define a Hamiltonian that is 5-local. As mentioned before H_G acts on a bigger number of qubits than G does. More precisely it acts on four registers *clock*, comp = (out, adv, aux) - that is there is an additional register called clock in comparison to registers of G. The standard reduction defines

$$H_G = H_{\text{in}} + H_{\text{out}} + H_{\text{prop}} + H_{\text{clock}}.$$

The high level idea is to define the terms $H_{\text{in}}, H_{\text{out}}, H_{\text{prop}}, H_{\text{clock}}$ such that G outputs 1 with high probability if and only if H_G has a small eigenvalue. In this case the minimizing vector $|\phi\rangle$ is the so called *history state*

$$\frac{1}{\sqrt{T'+1}} \sum_{j=0}^{T'} |j\rangle_{\text{clock}} \otimes G_j \dots G_1 |0\rangle_{\text{out}} |\psi\rangle_{\text{adv}} |0^n\rangle_{\text{aux}}, \tag{7}$$

where, for every j, G_j is the unitary transformation corresponding to the j-th gate in G and $|j\rangle_{\text{clock}}$ is a state in the clock state space that we will define in detail later. The terms are defined so that they impose penalties to $\langle\phi| H_G |\phi\rangle$ whenever $|\phi\rangle$ is far from the history state.

For our purposes we change the reduction by removing the H_{out} term. By doing that we will be able to say that for every $|\phi\rangle$ such that $\langle\phi| H_G |\phi\rangle$ is small there exists $|\psi_{\mathcal{D}^A}\rangle_{\text{adv}}$ such that $|\phi\rangle$ is close to the history state for $|\psi_{\mathcal{D}^A}\rangle_{\text{adv}}$. With that property in hand we can then say that if we measure $|\phi\rangle$ in the Z basis then (i) with probability $\Omega(1/T')$ the clock register is equal to $|0\rangle_{\text{clock}}$, the out is equal to $|0\rangle_{\text{out}}$, the aux register is equal to $|0^n\rangle_{\text{aux}}$ and the adv register contains a sample from $\mathcal{D}^A$ (ii) with probability $\Omega(1/T')$ the clock register is equal to $|T'\rangle_{\text{clock}}$ and the out register contains a sample from a Bernoulli

[7] This protocol is based on a circuit-to-Hamiltonian reduction. The size of this constant depends on which reduction we use

[8] The reduction can be applied to any circuit but we focus only on the comparison circuit for simplicity.

```
Verifier V                                                                          Prover P
                                          $H_G$
                                    $\xrightarrow{\hspace{3cm}}$
$\gamma = 0, p = 0$
$T' = n + T + 3$
$n' = 2n + T' + 1$
$L =$ number of terms of $H_G$
for $i \in \{1, \ldots N\}$ where $N = O\left(\frac{K(n^5 + n^2T^3 + T^5)}{\eta^4}\log(1/\delta)\right)$
do :
  $t \xleftarrow{\$} \mathsf{Terms}(H_C)$
  type $\xleftarrow{\$} \{1, 2, 3\}$
  $n_1, n_2, n_3 = 0$
                                                                     $|\phi\rangle_{AB} \in (\mathbb{C}^2)_A^{\otimes n'} \otimes \mathcal{H}_B$
                                          $|\phi\rangle_A$
                                    $\xleftarrow{\hspace{3cm}}$
  if type $= 0$ :
    # measure energy of the state with respect to the Hamiltonian
    $b_i =$ measurement of $|\phi\rangle_i$ in basis $\sigma_i^B$,   $\forall \sigma_i^B \in t$
    $\gamma = \gamma - J_t(\Pi_{i \in t} b_i)$
    $n_1 = n_1 + 1$
  if type $= 1$ :
    # obtain a sample
    $b =$ measurement of the second qubit of clock in the Z basis
    $a_1, \ldots, a_n =$ measurement of aux in the Z basis
    $b' =$ measurement of out in the Z basis
    if $b = 0, b' = 0, a_1, \ldots, a_n = 0$ :
      $b_1, \ldots, b_n =$ measurement of adv in the Z basis
      $S = S \cup \{(b_i)_{i \in [n]}\}$
    $n_2 = n_2 + 1$
  if type $= 2$ :
    # estimate the output probability
    $b_1 =$ measurement of the $T'$-th qubit of clock in the Z basis
    if $b_1 = 1$ :
      $r =$ measurement of out in the Z basis
      $p = p + r$
      $n_3 = n_3 + 1$
Done :
if $\frac{\gamma \cdot L}{n_1} > \frac{\eta^2}{2T'^2} \vee \frac{p}{n_3} < 1 - 2\eta^2$ :
  Abort
```

Figure 4: The interactive protocol, in which the verifier collects samples from a distribution close to the desired one. The verifier only requires a single qubit, as they measure one qubit at a time. H_G is the Hamiltonian corresponding to the comparison circuit, described in Figure 1. We emphasize that we send $|\phi\rangle$ one qubit a time. Note that when we measure the clock register we use the unary representation of the clock. By writing $|\phi\rangle_{AB} \in (\mathbb{C}^2)_A^{\otimes n'} \otimes \mathcal{H}_B$ and then sending $|\phi\rangle_A$ to $\mathbf{V}$ we mean that $\mathbf{P}$ might be sending a mixed state.

variable with parameter p such that p is close to the probability of G outputting 1 on $|0\rangle_{\text{out}} |\psi_{\mathcal{D}^A}\rangle_{\text{adv}} |0^n\rangle_{\text{aux}}$. Note that we can also write this probability as $\langle 0^n| \langle \psi_{\mathcal{D}^A}|_{\text{adv}} \langle 0|_{\text{out}} G^\dagger \Pi_1^{(1)} G |0\rangle_{\text{out}} |\psi_{\mathcal{D}^A}\rangle_{\text{adv}} |0^n\rangle_{\text{aux}}$, where $\Pi_s^{(\alpha)}$ is the projection onto the subspace of vectors for which the s-th qubit equals α. This notation will be useful later.

Overview of the Protocol. Assuming that the above properties hold we give a high level idea of the protocol defined in Figure 4. In each round of the protocol we perform one of the three types of operations, where the type is chosen uniformly at random (i) we estimate the energy $\langle\phi| H_G |\phi\rangle$ (ii) we measure $|\phi\rangle$ in the Z basis and if the clock register is equal to $|0\rangle_{\text{clock}}$, the out register is equal to $|0\rangle_{\text{out}}$ and the aux register is equal to $|0^n\rangle_{\text{aux}}$ then we collect a sample (iii) we measure $|\phi\rangle$ in the Z basis and if the clock register is equal to $|T'\rangle_{\text{clock}}$ we update the estimate for p. We run the protocol $\Theta(T)$ rounds thus each of the types will occur $\Omega(T)$ times with high probability and our reduction guarantees that for (ii) we successfully $\Omega(1)$ samples and for (iii) we update the estimate $\Omega(1)$ times. Overall this guarantees that the estimate for $\langle\phi| H_G |\phi\rangle$ and p will be accurate and the number of samples collected will be in $\Omega(1)$. Our reduction guarantees moreover that if $|\phi\rangle$ is in fact a low energy state of H_G then p is close to the probability of G outputting 1 on $|0\rangle_{\text{out}} |\psi_{\mathcal{D}^A}\rangle_{\text{adv}} |0^n\rangle_{\text{aux}}$ and the samples we collect come i.i.d. from distribution $\mathcal{D}^A$ that corresponds to $|\psi_{\mathcal{D}^A}\rangle$. Moreover using Lemma 9 from p we can estimate $|\langle\psi_{\mathcal{D}^A}|\psi_{\mathcal{D}_C}\rangle|$, recall that $\mathcal{D}_C$ is the distribution generated by C on $|0^n\rangle$ of which we think as being close to $\mathcal{D}$. As explained in Section 3.1 estimating $|\langle\psi_{\mathcal{D}^A}|\psi_{\mathcal{D}_C}\rangle|$ is enough to guarantee that the distribution from which we collected the samples is close to $\mathcal{D}$.

For the remainder of this section we first explain the details of the circuit-to-Hamiltonian reduction and then formalize the correctness and soundness requirements and prove the desired properties.

D.1.1 Circuit-to-Hamiltonian Reduction

We start with a quantum circuit G and want to create a Hamiltonian H_G with the properties mentioned in Section D.1. First we make our goal formal.

Lemma 4 (Circuit-to-Hamiltonian Reduction)**.** *For every comparison circuit G, for all, sufficiently small, $\epsilon > 0$ there exists an efficiently computable description of a 5-local Hamiltonian H_G with $L = O(n + T')$ many terms such that the following conditions hold. Let $\mathcal{D}^A$ be the distribution of the content of the adv register when measuring $|\phi\rangle$ in the Z basis conditioned on the clock, out and aux registers being all 0 after measurement. For every $|\phi\rangle$ such that $\langle\phi| H_G |\phi\rangle \leq \frac{\epsilon}{T'}$ if we measure $|\phi\rangle$ in the Z basis then*

- *with probability $\in \left[\frac{1-5\epsilon}{T'+1}, \frac{1+5\epsilon}{T'+1}\right]$ the clock register is equal to $|0\rangle_{clock}$, the out register is equal to $|0\rangle_{out}$, the aux register is equal to $|0^n\rangle_{aux}$,*
- *with probability $\in \left[\frac{1-5\epsilon}{T'+1}, \frac{1+5\epsilon}{T'+1}\right]$ the clock register is equal to $|T'\rangle_{clock}$ and conditioned on this event the distribution of the out register is a Bernoulli variable with parameter p such that $|p - \langle 0^n|_{aux} \langle\psi_{\mathcal{D}^A}|_{adv} \langle 0|_{out} G^\dagger \Pi_{out}^{(1)} G |0\rangle_{out} |\psi_{\mathcal{D}^A}\rangle_{adv} |0^n\rangle_{aux}| \leq 5\epsilon T'$.*

Proof. As we discussed we want to base our reduction on the standard circuit-to-Hamiltonian reduction but drop the H_{out} term. We define

$$H_G = H_{\text{in}} + H_{\text{prop}} + H_{\text{clock}}. \tag{8}$$

The term H_{in} corresponds to the condition that, at step 0, the qubits are in the right state. Formally

$$H_{\text{in}} = |0\rangle \langle 0|_{\text{clock}} \otimes \left(\sum_{j \in \text{out,aux}} \Pi_j^{(1)} \right), \tag{9}$$

where by $j \in \text{out}, \text{aux}$ we mean iterating over all the qubits in these registers. Informally speaking, we add a penalty whenever a qubit in registers out or aux is in state $|1\rangle$ while the clock is in state $|0\rangle_{\text{clock}}$.

The term H_{prop} guarantees the propagation of quantum states through the circuit. Formally

$$H_{\text{prop}} = \sum_{j=1}^{T'} H_j, \tag{10}$$

$$H_j = -\frac{1}{2}\left|j\right\rangle\left\langle j-1\right|_{\text{clock}} \otimes G_j - \frac{1}{2}\left|j-1\right\rangle\left\langle j\right|_{\text{clock}} \otimes G_j^\dagger + \frac{1}{2}(\left|j\right\rangle\left\langle j\right|_{\text{clock}} + \left|j-1\right\rangle\left\langle j-1\right|_{\text{clock}}) \otimes I.$$

We will define H_{clock} later. We could realize it with $O(\log(T'))$ qubits but then our Hamiltonian would be $O(\log(T'))$-local. But we aim for a 5-local Hamiltonian. We explain how to address this issue towards the end of this section. Because of this we will assume for now that H_{clock} does not appear in (8).

For the analysis we follow [KSV02]. It will be useful to consider a change of basis given by

$$W = \sum_{j=0}^{T'} \left|j\right\rangle\left\langle j\right|_{\text{clock}} \otimes G_j \dots G_1.$$

What we mean is that we represent the vector $\left|\phi\right\rangle$ in the form $\left|\phi\right\rangle = W\left|\tilde{\phi}\right\rangle$. Under this change the Hamiltonian is transformed into its conjugate $\widetilde{H}_G = W^\dagger H_G W$. Simple calculation verifies that $\widetilde{H}_{\text{in}} = H_{\text{in}}$ and $\widetilde{H}_{\text{prop}} = E \otimes I$, where

$$E = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & \\ -\frac{1}{2} & 1 & -\frac{1}{2} & 0 & 0 & \\ 0 & -\frac{1}{2} & 1 & -\frac{1}{2} & 0 & \\ 0 & 0 & -\frac{1}{2} & 1 & -\frac{1}{2} & \\ 0 & 0 & 0 & -\frac{1}{2} & 1 & \\ & & & & & \ddots \end{pmatrix}$$

Let $\left|\tilde{\phi}\right\rangle$ be such that $\left\langle\tilde{\phi}\right|\widetilde{H}_G\left|\tilde{\phi}\right\rangle \leq \frac{\epsilon}{T'}$. We will show that it is close to a history state of $\left|\psi_{\mathcal{D}^A}\right\rangle_{\text{adv}}$. Let's write $\tilde{\phi} = \sum_{j=0}^{T'} \alpha_j \left|j\right\rangle_{\text{clock}} \left|\xi_j\right\rangle_{\text{comp}}$, for $\alpha_j \in \mathbb{R}_{\geq 0}$ and $\left|\xi_0\right\rangle_{\text{comp}} = \sum_{s \in \{0,1\}^{n+1}} \beta_s \left|s[1]\right\rangle_{\text{out}} \left|\psi_s\right\rangle_{\text{adv}} \left|s[2,n+1]\right\rangle_{\text{aux}}$, where $s[i,j]$ denotes the substring of s from i to j. Then by the fact that $\widetilde{H}_{\text{prop}} = E \otimes I$ we have

$$\begin{aligned} \left\langle\tilde{\phi}\right|\widetilde{H}_{\text{prop}}\left|\tilde{\phi}\right\rangle &= \frac{1}{2}\sum_{j=1}^{T'} \|\alpha_{j-1}\left|\xi_{j-1}\right\rangle_{\text{comp}} - \alpha_j\left|\xi_j\right\rangle_{\text{comp}}\|^2 \\ &\geq \frac{1}{2}\sum_{j=1}^{T'} |\alpha_{j-1} - \alpha_j|^2, \frac{1}{2}\sum_{j=1}^{T'} \min(|\alpha_{j-1})|^2, |\alpha_j|^2) \cdot \| \left|\xi_{j-1}\right\rangle_{\text{comp}} - \left|\xi_j\right\rangle_{\text{comp}}\|^2. \end{aligned} \tag{11}$$

Note that the bound above gives two inequalities. Thus we get that $\max_{j\in[T']} |\alpha_{j-1} - \alpha_j|^2 \leq \frac{2\epsilon}{T'}$, which combined with the fact that $\sum_{j=0}^{T'} |\alpha_j|^2 = 1$ gives us that

$$\max_{j\in\{0,\dots,T'\}} \left| |\alpha_j|^2 - \frac{1}{T'+1} \right| \leq \frac{2\epsilon}{T'}. \tag{12}$$

Using (12) and the bound for $\frac{1}{2}\sum_{j=1}^{T'} \min(|\alpha_{j-1})|^2, |\alpha_j|^2) \cdot \| \left|\xi_{j-1}\right\rangle_{\text{comp}} - \left|\xi_j\right\rangle_{\text{comp}}\|^2$ from (11) we get that for $\epsilon \leq 1$

$$\sum_{j=1}^{T'} \| \left|\xi_{j-1}\right\rangle_{\text{comp}} - \left|\xi_j\right\rangle_{\text{comp}}\|^2 \leq \frac{\frac{2\epsilon}{T'}}{\frac{1}{T'+1} - \frac{2\epsilon}{T'}} \leq 4\epsilon,$$

which implies that

$$\| \left|\xi_0\right\rangle_{\text{comp}} - \left|\xi_{T'}\right\rangle_{\text{comp}}\|^2 \leq 4\epsilon T'. \tag{13}$$

Using the second term from H_G we also have

$$\left\langle\tilde{\phi}\right|\widetilde{H}_{\text{in}}\left|\tilde{\phi}\right\rangle = \sum_{j=1}^{n} \sum_{s\in\{0,1\}^n : s[j]=1} \beta_s^2 \leq \frac{\epsilon}{T'}. \tag{14}$$

Note that the distribution corresponding to $\left|\psi_{0^n}\right\rangle$ is $\mathcal{D}^A$. Observe moreover that (12) guarantees that for small enough ϵ if we measure $\left|\tilde{\phi}\right\rangle$ in the Z basis then with probability $\in \left[\frac{1-3\epsilon}{T'+1}, \frac{1+3\epsilon}{T'+1}\right]$ the clock register is equal to $\left|0\right\rangle_{\text{clock}}$ and with probability

$\in \left[\frac{1-3\epsilon}{T'+1}, \frac{1+3\epsilon}{T'+1}\right]$ the clock register is equal to $|T'\rangle_{\text{clock}}$. Moreover conditioned on the clock register being $|0\rangle_{\text{clock}}$ probability of out and aux register being $|0\rangle_{\text{out}}, |0^n\rangle_{\text{aux}}$ respectively is, by (14), lower bounded by $1 - \frac{\epsilon}{T'}$. Thus we collect a sample from $\mathcal{D}^A$ with probability $\in \left[\frac{1-3\epsilon}{T'+1}(1 - \frac{\epsilon}{T'}), \frac{1-3\epsilon}{T'+1}\right] \subseteq \left[\frac{1-5\epsilon}{T'+1}, \frac{1+5\epsilon}{T'+1}\right]$.

For the second condition observe that

$$\begin{aligned}
&|p - \langle 0^n|_{\text{aux}} \langle \psi_{\mathcal{D}^A}|_{\text{adv}} \langle 0|_{\text{out}} G^\dagger \Pi^{(1)}_{\text{out}} G |0\rangle_{\text{out}} |\psi_{\mathcal{D}^A}\rangle_{\text{adv}} |0^n\rangle_{\text{aux}} | \\
&= | \langle \xi_{T'}|_{\text{comp}} W^\dagger \Pi^{(1)}_{\text{out}} W |\xi_{T'}\rangle_{\text{comp}} - \langle 0^n|_{\text{aux}} \langle \psi_{\mathcal{D}^A}|_{\text{adv}} \langle 0|_{\text{out}} G^\dagger \Pi^{(1)}_{\text{out}} G |0\rangle_{\text{out}} |\psi_{\mathcal{D}^A}\rangle_{\text{adv}} |0^n\rangle_{\text{aux}} | \\
&\leq | \langle \xi_0|_{\text{comp}} W^\dagger \Pi^{(1)}_{\text{out}} W |\xi_0\rangle_{\text{comp}} - \langle 0^n|_{\text{aux}} \langle \psi_{\mathcal{D}^A}|_{\text{adv}} \langle 0|_{\text{out}} G^\dagger \Pi^{(1)}_{\text{out}} G |0\rangle_{\text{out}} |\psi_{\mathcal{D}^A}\rangle_{\text{adv}} |0^n\rangle_{\text{aux}} | + 4\epsilon T' \\
&\leq \frac{\epsilon}{T'} + 4\epsilon T' \leq 5\epsilon T',
\end{aligned}$$

where in the first inequality we used (13) and the fact that the largest eigenvalue of $W^\dagger \Pi^{(1)}_{\text{out}} W$ is at most of norm 1 and in the second inequality we used (14) and again the fact that the largest eigenvalue of $W^\dagger \Pi^{(1)}_{\text{out}} W$ is at most of norm 1.

Realizing the clock. As we mentioned we also need to specify how to realize the clock register. The naive implementation would result in a $O(\log(T'))$-local Hamiltonian. To obtain a 5-local Hamiltonian we use a unary representation. That is we embed the counter space in a larger space in the following way

$$|j\rangle_{\text{clock}} \mapsto |\underbrace{1,\ldots,1}_{j}, \underbrace{0,\ldots,0}_{T'-j}\rangle.$$

We need to now change H_{in} and H_{prop} to be consistent with this change. But more importantly we need to also penalize incorrect configurations in the clock register. This is what the H_{clock} term is responsible for. We refer the readeer to [KSV02] for details. The proof of Lemma 4 extends naturally to this case. □

We will need a slight extension of Lemma 4 to the case where $\mathbf{P}$ sends mixed states. For the standard use cases of the reduction this extension is trivial but our purposes require more careful treatment. The difference of our setup in comparison to the standard reduction is that we also collect samples that need to satisfy a specific requirement and this is the reason why the analysis is more involved.

Corollary 1 (Circuit-to-Hamiltonian Reduction for Mixed States)**.** *For every comparison circuit G, if $d_H(\mathcal{D}, \mathcal{D}_C) = \eta$ is sufficiently small then there exists an efficiently computable description of a 5-local Hamiltonian H_G with $L = O(n + T')$ many terms such that the following conditions hold. Let $\mathcal{D}^A$ be the distribution of the content of the adv register when measuring ρ_A in the Z basis conditioned on the clock, out and aux registers being all 0 after measurement. For every density matrix ρ_A such that $Tr(H_G \rho_A) \leq \frac{\eta^2}{T'^3}$ if we measure ρ_A in the Z basis then*

- *with probability $\in \left[\frac{1-7\eta}{T'+1}, \frac{1+7\eta}{T'+1}\right]$ the clock register is equal to $|0\rangle_{clock}$, the out register is equal to $|0\rangle_{out}$, the aux register is equal to $|0^n\rangle_{aux}$,*
- *with probability $\in \left[\frac{1-7\eta}{T'+1}, \frac{1+7\eta}{T'+1}\right]$ the clock register is equal to $|T'\rangle_{clock}$ and if conditioned on this event the distribution of the out register is a Bernoulli variable with parameter $p \geq 1 - 3\eta^2$ then $d_H(\mathcal{D}_C, \mathcal{D}^A) \leq O(\eta^{1/4})$.*

Proof. Let $\epsilon = \frac{\eta^2}{T'^2}$. By the ensemble interpretation of density matrices we can express

$$\rho_A = \sum_{i=1}^{k} q_i |\phi_i\rangle \langle \phi_i|_{\text{comp}}.$$

Thus we can write

$$\sum_{i=1}^{k} q_i \langle \phi_i| H_G |\phi_i\rangle \leq \frac{\epsilon}{T'}.$$

By Markov inequality we have

$$\sum_{i=1}^{k} q_i \mathbb{1}_{\left\{\langle \phi_i|H_G|\phi_i\rangle > \frac{\sqrt{\epsilon}}{T'}\right\}} \leq \frac{\sqrt{\epsilon}}{T'}. \tag{15}$$

For $i \in [k]$ let $\mathcal{D}_i^A$ be the distribution of contents of adv conditioned on clock, out, and aux registers being all 0 when measuring $|\phi_i\rangle$ in the Z basis. Note that for all i such that $\langle\phi_i| H_G |\phi_i\rangle \leq \frac{\sqrt{\epsilon}}{T'}$ Lemma 4 guarantees that $\mathcal{D}_i^A$ satisfies the conditions of the reduction.

To see the the first condition note that by (15) we get that the probability that the clock register is $|0\rangle_{\text{clock}}$ is $\in \left[\frac{1-5\epsilon-2\sqrt{\epsilon}}{T'+1}, \frac{1+5\epsilon+2\sqrt{\epsilon}}{T'+1}\right] \subseteq \left[\frac{1-7\sqrt{\epsilon}}{T'+1}, \frac{1+7\sqrt{\epsilon}}{T'+1}\right] \subseteq \left[\frac{1-7\eta}{T'+1}, \frac{1+7\eta}{T'+1}\right]$. Same bound on probability holds also for the clock register being equal to $|T'\rangle_{\text{clock}}$.

For $i \in [k]$ let p_i be the probability of obtaining outcome 1 in the out register when measuring $|\phi_i\rangle$ in the Z basis conditioned on clock register being in state $|T'\rangle_{\text{clock}}$. Then for the second condition observe that

$$\begin{aligned}
p &= \sum_{i=1}^{k} q_i p_i \\
&\leq \sum_{i=1}^{k} q_i p_i \mathbb{1}_{\left\{\langle\phi_i|H_G|\phi_i\rangle \leq \frac{\sqrt{\epsilon}}{T'}\right\}} + \frac{\sqrt{\epsilon}}{T'} \\
&\leq \sum_{i=1}^{k} q_i \langle 0^n|_{\text{aux}} \left\langle \psi_{\mathcal{D}_i^A} \right|_{\text{adv}} \langle 0|_{\text{out}} G^\dagger \Pi_{\text{out}}^{(1)} G |0\rangle_{\text{out}} \left|\psi_{\mathcal{D}_i^A}\right\rangle_{\text{adv}} |0^n\rangle_{\text{aux}} \mathbb{1}_{\left\{\langle\phi_i|H_G|\phi_i\rangle \leq \frac{\sqrt{\epsilon}}{T'}\right\}} + 6\sqrt{\epsilon}T' \\
&\leq \sum_{i=1}^{k} q_i \langle 0^n|_{\text{aux}} \left\langle \psi_{\mathcal{D}_i^A} \right|_{\text{adv}} \langle 0|_{\text{out}} G^\dagger \Pi_{\text{out}}^{(1)} G |0\rangle_{\text{out}} \left|\psi_{\mathcal{D}_i^A}\right\rangle_{\text{adv}} |0^n\rangle_{\text{aux}} + 6\sqrt{\epsilon}T' + \frac{\sqrt{\epsilon}}{T'} \\
&\leq \sum_{i=1}^{k} q_i f(d_H(\mathcal{D}_C, \mathcal{D}_i^A)) + 7\sqrt{\epsilon}T' \qquad (16)
\end{aligned}$$

where in the first inequality we used (15), in the second inequality we used properties of $\mathcal{D}_i^A$ guaranteed by Lemma 4, in the fourth we used Corollary 1.

By (16) and the assumption $p \geq 1 - 3\eta^2$ we get that

$$\sum_{i=1}^{k} q_i f(d_H(\mathcal{D}_C, \mathcal{D}_i^A)) \geq 1 - 3\eta^2 - 7\sqrt{\epsilon}T' \geq 1 - 10\eta^2,$$

where in the last inequality we used that $\epsilon = \frac{\eta^2}{T'^2}$. We conclude by applying Lemma 8.

□

D.1.2 Correctness of the Protocol

Recall that protocol from Figure 4 builds upon the protocol from Figure 2. Now $\mathbf{V}$, instead of running G itself, outsources its execution to $\mathbf{P}$. On a high level correctness of this new protocol is a consequence of correctness of the quantum verifier protocol (Theorem 2) and circuit-to-Hamiltonian reduction (Lemma 4). One, however, needs to be careful as the guarantees about the protocol will change slightly and some details in the proof need to be verified.

Lemma 5. *Let n_1, n_2, n_3 be the number of times each type occurs in protocol from Figure 4. If $N = \Omega(\log(1/\delta))$ then $\mathbb{P}[n_1, n_2, n_3 > \frac{N}{6}] \geq 1 - \delta$.*

Proof. For $b \in \{1, 2, 3\}$, n_b can be seen as sum of random Bernoulli variables $\{x_i\}_{i\in[N]}$ with parameter $1/3$. Then by Fact 3 we get that $\mathbb{P}[|\frac{n_b}{N} - \frac{1}{3}| > \frac{1}{6}] \leq 2e^{-\frac{N}{72}} \leq \frac{\delta}{3}$. We finish by applying the union bound to the error events.

□

Lemma 6. *Let ρ_A be the reduced density of the first n' qubits of $|\phi\rangle_{AB}$, γ, p, n_1, n_2, n_3, S be as in the protocol defined in*

Figure 4. Let p^*, q^* and λ be defined as,

$$\lambda = Tr(H_G \rho_A),$$
$$q^* = Tr(|0\rangle\langle 0|_{clock} \otimes |0\rangle\langle 0|_{out} \otimes |0^n\rangle\langle 0^n|_{aux}\, \rho_A),$$
$$p^* = \frac{Tr(|T'\rangle\langle T'|_{clock} \otimes |1\rangle\langle 1|_{out}\, \rho_A)}{Tr(|1\rangle\langle 1|_{out}\, \rho_A)}.$$

We define the event $\mathcal{F}$ to be $\left|\frac{\gamma \cdot L}{n_1} - \lambda\right| \leq \epsilon, \left|\frac{|S|}{n_2} - q^*\right| \leq \epsilon \left|\frac{p}{n_3} - p^*\right| \leq \epsilon$. If $N = \Omega(\frac{n^2+T'^2}{\epsilon^2}\log(1/\delta))$ then $\mathbb{P}[\mathcal{F}] \geq 1-\delta$.

Note 2. *For the sake of convenience, we often write $|0\rangle\langle 0|_{clock} \otimes |0\rangle\langle 0|_{out} \otimes |0^n\rangle\langle 0^n|_{aux}$, when we actually mean $|0\rangle\langle 0|_{clock} \otimes |0\rangle\langle 0|_{out} \otimes I_{adv} \otimes |0^n\rangle\langle 0^n|_{aux}$.*

Proof. Note that for every term $t \in H_G$ we have $|J_t| \leq 1$. Then if $n_1 = \Omega(\frac{L^2}{\epsilon^2}\log\frac{1}{\delta})$ then Fact 3 guarantees that $\mathbb{P}[|\frac{\gamma \cdot L}{n_1} - \lambda| > \epsilon] \leq \delta$.

Next we define Bernoulli variables $\{s_i\}_{i \in [n_2]}$ to indicate whether $|S|$ increases in a given round, i.e. $|S| = \sum_{i=1}^{n_2} s_i$. By definition $\mu = E[s_i] = \text{Tr}(|0\rangle\langle 0|_{\text{clock}} \otimes |0\rangle\langle 0|_{\text{out}} \otimes |0^n\rangle\langle 0^n|_{\text{aux}}\, \rho_A)$. Using Fact 3 we get that if $n_2 = \Omega(\frac{1}{\epsilon^2}\log(1/\delta))$ then $\mathbb{P}\left[\left|\frac{|S|}{n_2} - q^*\right| > \epsilon\right] \leq \delta$. The exact same argument can be used for $\frac{p}{n_3}$.

To conclude we note that, by the union bound, if $n_1 = \Omega(\frac{L^2}{\epsilon^2}\log(1/\delta))$ and $n_2, n_3 = \Omega(\frac{1}{\epsilon^2}\log(1/\delta))$ then $\mathbb{P}[\mathcal{F}] \geq 1-\delta$. By Lemma 5 and the union bound we get that if $N = \Omega(\frac{L^2}{\epsilon^2}\log(1/\delta))$ then $\mathbb{P}[\mathcal{F}] \geq 1-\delta$. As Lemma 4 guarantees that $L = O(n+T')$ we can also set $N = \Omega(\frac{n^2+T'^2}{\epsilon^2}\log(1/\delta))$.

□

Intuitively Lemma 6 guaranties that with a high probability, the estimates $\frac{\gamma \cdot L}{n_1}, \frac{|S|}{n_2}, \frac{p}{n_3}$ are accurate enough. With that fact in hand we proceed by stating the main theorem of this section.

Theorem 5 (Constant Memory Quantum Verifier)**.** *For every circuit C acting on n qubits, with T gates, for every $\delta \in (0, \frac{1}{3})$, $K \in \mathbb{N}$ and all $\eta > 0$ small enough there exists an interactive protocol between a verifier with constant quantum memory $\mathbf{V}$ and a quantum prover $\mathbf{P}$ with the following properties. The protocol runs in $N = O\left(\frac{K \cdot (n^5 + n^2T^3 + T^5)}{\eta^4}\log(1/\delta)\right)$ rounds, in each round $\mathbf{P}$ sends a (potentially mixed) quantum state on $O(n+T)$ qubits to $\mathbf{V}$. At the end of the protocol $\mathbf{V}$ outputs $\perp$ when it rejects the interaction or $S = \{x_1, \dots, x_{|S|}\}$, where $x_i \in \{0,1\}^n$, when it accepts.*

- *(Completeness) There exists $\mathbf{P}^{\mathsf{0}(*)}$ such that for every $\mathcal{D} \in \mathfrak{D}(n)$ satisfying $d_H(\mathcal{D}, \mathcal{D}_C) \leq \eta$ the following holds. With probability $1-\delta$ over the randomness in the protocol $\mathbf{P}^{\mathsf{0}(\mathcal{D})}$ succeeds, $S \sim_{\text{i.i.d.}} \mathcal{D}^{|S|}$ and $|S| \geq \Omega(K)$.*
- *(Soundness) For every $\mathbf{P}$ that succeeds with probability at least $\frac{2}{3}$ we have $S \sim_{\text{i.i.d.}} (\mathcal{D}^A)^{|S|}$ and $d_H(\mathcal{D}_C, \mathcal{D}^A) \leq O(\eta^{1/4})$.*

Proof. We first address completeness of the protocol and then move to soundness.

Completeness. Recall that the $\mathbf{P}$ that was guaranteed to exist in Theorem 2 was just sending state $|\psi_{\mathcal{D}}\rangle_{\text{adv}}$ to $\mathbf{V}$. Recall that we denote by $T' = n + T + 3$ the number of gates in G and by n' the number of qubits that are sent by $\mathbf{P}$ in each round. As we discussed the natural extension of this strategy to the constant memory model is for $\mathbf{P}$ to prepare the history state $|\phi_{\mathcal{D}}\rangle_{\text{comp}}$ of $|\psi_{\mathcal{D}}\rangle_{\text{adv}}$ and send it to $\mathbf{V}$. As $N = O\left(\frac{K \cdot (n^2T'^3 + T'^5)}{\eta^4}\log(1/\delta)\right) = O\left(\frac{K \cdot (n^5 + n^2T^3 + T^5)}{\eta^4}\log(1/\delta)\right)$ we get by Lemma 6 that with probability $1-\delta$

- the estimate of the energy $\frac{\gamma \cdot L}{n_1} \leq \frac{\eta^2}{4T'^3}$ as $\langle\phi_{\mathcal{D}}| H_G |\phi_{\mathcal{D}}\rangle = 0$,
- $|S| = \Omega(K)$ as in this case $\text{Tr}(|0\rangle\langle 0|_{\text{clock}} \otimes |0\rangle\langle 0|_{\text{out}} \otimes |0^n\rangle\langle 0^n|_{\text{aux}}\, \rho_A)$, which is the probability of getting a sample if the type is 1 is equal to $\langle\phi_{\mathcal{D}}| \Pi^{(T')}_{\text{clock}} |\phi_{\mathcal{D}}\rangle = \frac{1}{T'+1}$,
- $p \geq \frac{\langle\phi_{\mathcal{D}}|\Pi^{(0)}_{\text{clock}}\Pi^{(1)}_{\text{out}}|\phi_{\mathcal{D}}\rangle}{\langle\phi_{\mathcal{D}}|\Pi^{(0)}_{\text{clock}}|\phi_{\mathcal{D}}\rangle} - \frac{\eta^2}{4} \geq f(d_H(\mathcal{D}_C, \mathcal{D})) - \frac{\eta^2}{4} \geq 1 - 2\eta^2$, thus the two checks are verified and the interaction is accepted. By definition $S \sim_{\text{i.i.d.}} (\mathcal{D})^{|S|}$. Thus completeness is verified.

Soundness. We follow the structure of the proof of Theorem 11, which is the analog of this theorem for a fully quantum verifier. Let ρ_A be the density matrix representing the state sent by $\mathbf{P}$. By Lemma 6 we know that with probability $1-\delta/2$ the energy estimate is within an additive error of $\frac{\eta^2}{4T'^3}$ and p is estimated within an additive error of $\frac{\eta^2}{4}$. So as $\mathbf{P}$ succeeds with probability $\frac{2}{3}$ then by the union bound and the fact that $\frac{1}{3}+\frac{\delta}{2}<1$ we get that $\text{Tr}(H_G\rho_A) \leq \frac{\eta^2}{2T'^3}+\frac{\eta^2}{4T'^3}=\frac{\eta^2}{T'^3}$ and $p \geq 1-2\eta^2-\frac{\eta^2}{4} \geq 1-3\eta^2$. With that we can apply Corollary 1 and conclude that $d_H(\mathcal{D}^A,\mathcal{D}_C) \leq O(\eta^{1/4})$. □

D.2 Classical Verifier

Now we are ready to move to the last model we consider in this work, namely the one where $\mathbf{V}$ is fully classical and the communication is also classical. Recall that in Section D.1 we designed the protocol by forcing $\mathbf{P}$ to send to $\mathbf{V}$ a history state $|\phi_{\mathcal{D}^A}\rangle_{\text{comp}}$ corresponding to a distribution satisfying $d_H(\mathcal{D}_C,\mathcal{D}^A) \leq O(\eta^{1/4})$. To extend this protocol to the classical model we first force $\mathbf{P}$ to commit to a state ρ, a state that will in some sense correspond to $|\phi_{\mathcal{D}^A}\rangle_{\text{comp}}$ and then force $\mathbf{P}$ to measure this state in the basis chosen by $\mathbf{V}$. By making the prover to measure his qubits honestly we get a version of constant quantum memory Protocol (Figure 4) in which all the quantum computation is done on the prover side and the verifier and the communication is completely classical.

To achieve our goal we will use cryptographic tools. As the protocol will rely on hardness of computational problems, our soundness results will only address provers who are computationally bounded, namely only provers in the QPT class. Recall that the QPT class is defined as follows. There exists a classical algorithm running in poly(λ) time that for an input size 1^λ, generates the prover's circuit of size poly(λ).

Next we give a high level overview of the protocol. An honest prover $\mathbf{P}$ is given a local Hamiltonian correspoding to G and computes the ground state of the Hamiltonian, i.e. the history state $|\phi_{\text{hist}}\rangle$. Later the prover is asked to commit to this state before the protocol proceeds with the interactive stage, in which the prover is asked to measure qubits of the state he has committed to either in computational or the Hadamard basis, and send the outcomes to the verifier. At each iteration, the verifier decides to do one of the following 3,

- estimate the energy of the state the prover has committed to,
- estimate the probability of the output of the circuit being 1,
- collect a sample from the distribution corresponding to the prover's state.

The description of this protocol is given in Figure 5. We note that the results presented in this section heavily rely on [Mah18]. Some of the technical lemmas are not proven here. We refer the reader to [Mah18] for said proofs.

Note 3. *We stress that with this protocol one can only retrieve samples from measurements done in the Z basis. The distribution of samples collected in the protocol when* $\mathbf{V}$ *asks for the X basis measurements are* ***not*** *in general equal to the distribution of measuring the state of* $\mathbf{P}$ *in the X basis. This means that if our protocol required samples from the distribution corresponding to the X measurements it is not clear if it could be realized in the fully classical model.*

Similar to [Mah18] we require a more refined version of the circuit-to-Hamiltonian reduction, namely we require our Hamiltonians to be 2-local and of the form $\sum_{i,j} -\frac{J_{i,j}}{2}(\sigma_{X,i}\sigma_{X,j}+\sigma_{Z,i}\sigma_{Z,j})$.

Theorem 6 ([BL08]). *For any integer* $n \geq 1$ *there exists* $n' = poly(n)$, $a(n)$ *and* $\delta \geq 1/poly(n)$ *such that given a* T*-gate quantum circuit* G*, there exists an efficiently computable real-weighted Hamiltonian* H_G *in* $XX-ZZ$ *form, such that,*

- *(completeness) If* G *accepts* x *with probability at least* $2/3$*, then* $\lambda_0(H_G) \leq a$.
- *(soundness) If* G *accepts* x *with probability at most* $1/3$*, then* $\lambda_0(H_G) > a+\delta$.

As proved in Section D.1, by modifying the standard circuit-to-Hamiltonian reduction, we can show that "for any $|\phi\rangle$ such that $\langle\phi| H_G |\phi\rangle < \epsilon$ the distribution of the measurement outcome of the first qubit of $|\phi\rangle$ (conditioned on the clock register being T') is ϵ close to the distribution of what G would output". For the sake of simplicity we skip reproving this statement for the 2-local Hamiltonian.

We proceed by stating the completeness and soundness properties of this protocol and providing a proof sketch.

```
Verifier V                                                                 Prover P
                                   $H_C = \sum_{ij} -\frac{J_{ij}}{2}(\sigma_{X,i}\sigma_{X,j} + \sigma_{Z,i}\sigma_{Z,j})$
                                   $\xrightarrow{\qquad\qquad}$
for $i \in [N]$ where $N = O\left(\frac{K}{\eta^4}\text{poly}(n,T)\log(1/\delta)\right)$ do
$\gamma, p, n_1, n_2, n_3 = 0$
$T' = n + T + 3, n' = 2n + T' + 1, S = \emptyset$
type $\overset{\$}{\leftarrow} \{1, 2, 3\}$
# generate the functions
for $\ell \in [n']$, set $(pk_\ell, td_\ell) \leftarrow \mathsf{Gen}(1^\lambda)$
                                   $\xrightarrow{pk_1, \ldots, pk_{n'}}$
                                   $\xleftarrow{y_1, \ldots, y_{n'}}$
$b \overset{\$}{\leftarrow} \{0, 1\}$
                                   $\xrightarrow{b}$
                                   $\{x_i\}_{i\in[n']}=(o,t_1,\ldots,t_T,a_1,\ldots,a_n,i_1,\ldots,i_n)$
                                   $\xleftarrow{o, t_j, a_j, i_j \in \{0,1\}^{m+1}}$
if $b = 0$ :
    #the computational basis
    if $f_{pk_i}(x_i) \neq y_i$ :
        abort()
    if type $= 1$ :
        #Estimate $\gamma$
        $n_1 = n_1 + 1$
        $\gamma \leftarrow \gamma - J_{ij}(-1)^{b(x_i)+b(x_j)}$
    if type $= 2$ :
        #Estimate $p$
        #check if clock is = $T'$
        if $(b(t_0), b(t_1), \ldots, b(t_{T'})) = (1, \ldots, 1)$ :
            $p = p + b(o)$
            $n_2 = n_2 + 1$
    if type $= 3$ :
        #Obtain a sample
        $n_3 = n_3 + 1$
        #check if clock is = 0
        if$(b(t_0), \ldots, b(t_{T'})) = (0, 0, \ldots, 0)$
            $\wedge\, b(o) = 0$
            $\wedge\, b(i_0), \ldots, b(i_{n'}) = (0, \ldots, 0)$ :
            $s = b(a_1)|| \ldots ||b(a_{n'})$
            $S = S \cup \{s\}$
if $b = 1$ :
    #the Hadamard basis
                                   $\xleftarrow{\{d_i\}_{i\in[n']}}$
    if type $= 1$ :
        $n_1 = n_1 + 1$
        $\gamma = \gamma - J_{i,j}(-1)^{(d_i(x_{i,1}+x_{i,0})+d_j(x_{j,1}+x_{j,0}))}$
else :
    continue
return $\gamma, p, S, n_1, n_2, n_3$
```

Figure 5: The description of the classical verifier protocol. Notice that x_i values are $m + 1$ bits long each, e.g. o contains the measurement outcome of the output register, plus the remaining m bits of the input to f_{pk_1}.

D.2.1 Cryptographic Assumptions, Claw-Free Functions

In this section we review the cryptographic assumptions that the soundness of our protocol relies on.

The protocol starts with $\mathbf{V}$ sending description of n' functions $f_{pk_1}, \ldots, f_{pk_{n'}}$ to $\mathbf{P}$. These functions are a family of 2-to-1 functions called *claw-free*. Intuitively the prover is asked to commit to an image y_i for each f_{pk_i}. Let us denote the two preimages of y_i under f_{pk_i} by $x_{i,0}$ and $x_{i,1}$. Based on the challenge bit, the prover is asked to either reveal one of $x_{i,b}$ values or reveal d_i such that $d_i(x_{i,0} + x_{i,1}) = 0$. We also assume that for all i there exists an efficiently computable function b_i, which given $x_{i,b}$ outputs the bit b. The existence function b is inherent in most claw-free family constructions.

For an efficiently computable 2-to-1 function f, $\mathbf{P}$ can do the aforementioned task as follows,

1. $\mathbf{P}$ prepares the state $\frac{1}{\sqrt{2^{m+1}}} \sum_{x \in \{0,1\}^{m+1}} (|x\rangle)$
2. evaluates f_{pk} on the super-position to get the state, $\frac{1}{\sqrt{2^{m+1}}} \sum_{y \in \{0,1\}^m} (|x_{0,y}\rangle + |x_{1,y}\rangle) |y\rangle$
3. measures the last register to get y^*, the state after this step would be $\frac{1}{\sqrt{2}}(|x_{0,y^*}\rangle + |x_{1,y^*}\rangle)$ and send y^* to the verifier
4. if the challenge is 0, measures the state in the computational basis to obtain one of the preimages x_{b,y^*}, and if the challenge is 1, measures in the Hadamard basis to obtain d s.t. $d(x_{0,y^*} + x_{1,y^*}) = 0$

The cryptographic property that we want to capture with our family of functions is that, although based on the challenge bit, $\mathbf{P}$ can either return one preimage x_b or d s.t. $d(x_1 + x_2) = 0$ for $b \in \{0,1\}$ and $f^{-1}(y) = \{x_0, x_1\}$, but they should not succeed in both tasks simultaneously.

Definition 6 (Adaptive hardcore bit property)**.** *For parameter λ, a family of functions $\{f_{pk}\}_{pk} : \{0,1\}^{m(\lambda)} \rightarrow \{0,1\}^{m(\lambda)-1}$, is called an adaptive hardcore family if,* [9]

1. *for all pk, f_{pk} is 2-to-1,*
2. *there exists a classically, efficiently computable bijective function, $b_{pk,y} : f^{-1}(y) \rightarrow \{0,1\}$*
3. *$\forall pk$, $\forall \mathbf{P} \in$ QPT, if $(y, x^*, d) \leftarrow \mathbf{P}(pk)$, let $\{x_0, x_1\} = f_{pk}^{-1}(y)$ then $|1/2 - \mathbb{P}[f(x^*) = y \wedge d(x_0+x_1) = 0]| \leq ngl(\lambda)$.*

For the sake of convenience we only consider a specific construction of claw-free familes. We change the second requirement of definition 6 to,

$2'$. For any pk and any $y \in \text{Img}(f_{pk})$, the preimages of y take the form (b, x_b), meaning that $b(x)$ is actually the first bit of x.

D.2.2 The Canonical Isometry

In order to prove γ is an accurate estimate of the energy using a similar argument to Lemma 6, we have to prove the $\mathbb{E}[\gamma] = \text{Tr}(H_G \rho)$, where in a sense ρ is the prover's state. Letting (X_i, Z_i) be the observables of $\mathbf{P}$ which determin the value of the i^{th} response, we require an isometry which *teleports* these observables to $\sigma_{X,i}, \sigma_{Z,i}$ as the Hamiltonian is penalizing the bad configurations of the state with respect to $\sigma_{X,i}, \sigma_{Z,i}$. Let us assume the prover's state is in a hilbert space $H \otimes H_{\text{env}}$, where he might share some entanglement with the enviroment.

Based on how the estimates are updated in the protocol the natural way to define the observables that $\mathbf{P}$ measures, would be,[10]

$$Z(a) = \textstyle\sum_{x_1,\ldots,x_{n'} \in \{0,1\}^m} (-1)^{b(x)\cdot a} |x_1\rangle\langle x_1| \otimes \ldots |x_{n'}\rangle\langle x_{n'}| \otimes I_P$$
$$X(a) = \textstyle\sum_{d_1,\ldots,d_{n'}} (-1)^{\sum a_i(d_i(x_{i,0}+x_{i,1}))} U^\dagger(|d_1\rangle\langle d_1| \otimes \ldots |d_{n'}\rangle\langle d_{n'}| \otimes I_P)U$$

Basically in our modelling of actions of $\mathbf{P}$, if the challenge bit is $b = 0$, $\mathbf{P}$ measures his state in the computational basis in order to get the preimages, and if $b = 1$, he applies an arbitrary unitary U, followed by a Hadamard measurement to retrieve the d values.

[9] m is a polynomial

[10] here we will use b by absuing the notation instead of $(b_i(x_i))_{i \in [n]}$

As mentioned before we would like to see these obserables as σ_X and σ_Z up to some isometry. The canonical choice of isometry here is $V : H \to (\mathbb{C}^2)^{\otimes n'} \otimes (\mathbb{C}^2)^{\otimes n'} \otimes H$ defined as,

$$V\left|\psi\right\rangle = (\tfrac{1}{2^{n'}} \textstyle\sum_{a,b} I \otimes \sigma_X(a)\sigma_Z(b) \otimes X(a)Z(b)) \left|\phi^+\right\rangle^{\otimes n'} \left|\psi\right\rangle$$

Where $\left|\phi^+\right\rangle$ is an EPR pair.

Definition 7 (Extracted Qubits)**.** *Let* $\mathbf{P}$ *be a prover playing in Protocol 5.* X, Z *defined and let* V *be the canonical isometry sending* (X, Z) *to* (σ_X, σ_Z)*. Let* $\left|\phi\right\rangle \in H \otimes H_{env}$ *be the state of the prover after sending the* y_i *values. We call the reduced density of* $V\left|\phi\right\rangle$ *on* H *the extracted qubits of the prover, and we denote it by* ρ*.*

Fact 1 ([Mah18])**.** *Let* $\mathbf{P}$ *be any* QPT *prover,* ρ *be their extracted qubit. We have,*[11]

- $\forall b \in \{0,1\}^{n'}, \quad Tr(\sigma_Z(b)\rho) = \left\langle\psi\right| Z(b) \left|\psi\right\rangle$
- $\forall b \in \{0,1\}^{n'}, \quad Tr(\sigma_X(b)\rho) = \frac{1}{2^{n'}} \sum_a (-1)^{a\cdot b} \left\langle\psi\right| Z(b)X(a)Z(b) \left|\psi\right\rangle$

Previously we mentioned that one can retrieve samples by asking the prover measure their state in computational basis, the first bullet exactly corresponds to this scenario. Intuitively what it tells us is that the distribution of the $b(x)$, for x values returned by $\mathbf{P}$, is identical to the distribution of the measurement outcomes of the extracted qubit ρ in the computational basis. However, in the case of the Hadamard basis, the matter is more subtle as the distribution is "twirled". As long as we only care about collecting samples via Z measurements, the twirl operator does not cause us any issues, as we will show that it would not affect the energy estimate in the protocol.

In order to follow the proofs done in [Mah18] we require our function family $\mathcal{F}$ to have an stronger property than the adaptive hardcore bit property; namely for it to be a *collapsing* family, defined in [Unr16].

Let $\mathcal{F} = \{f_{pk}\}$ be a family of functions and consider the following game,

1. The challenger picks $pk \leftarrow Gen(1^\lambda)$ and sends it to the adversary
2. The adversary prepares a state $\left|\phi\right\rangle = \sum_x \alpha_x \left|x\right\rangle$ and sends it to the challenger
3. The challenger evaluates f_{pk} in super position of $\left|\phi\right\rangle$
4. The challenger measures the image register and obtains y and a state
$$\left|\phi'\right\rangle = (\sum_{x: f_{pk}(x)=y} \alpha_x \left|x\right\rangle) \left|y\right\rangle$$
5. The challenger flips a bit b and based on that either measures the first register of $\left|\phi'\right\rangle$ in the computational basis or does not
6. The challenger sends the state after step 5 to the prover (either $\left|\phi'\right\rangle$ or a (classical) probabilistic mixture $\sum_{x, f_{pk}(x)=y} |\alpha_x|^2 \left|x\right\rangle \left\langle x\right|$)
7. The adversary outputs a bit b' based on the state he has received.
8. If $b = b'$ the challenger accepts.

$\mathcal{F}$ is called a collapsing family if for any QPT adversary $\mathbf{A}$, the probability of $\mathbf{A}$ winning in the aforementioned game is at most $1/2 + ngl(\lambda)$.

We proceed by stating and proving the completeness of the protocol.

[11] $\sigma_W(a) = \Pi_{i \text{ s.t. } a_i=1} \sigma_{W,i}$ the X or Z measurement of indices such that $a_i = 1$

D.2.3 Completeness

In this section we describe an honest prover strategy. We describe a prover $\mathbf{P}^{\mathsf{O}(\mathcal{D})}$ that wins in protocol 5 with probability 1, and provides us with samples from $\mathcal{D}$. Recall that in the constant quantum memory protocol, $\mathbf{P}$ first creates a history state $|\phi_{\mathcal{D}}\rangle$ for $|\psi_{\mathcal{D}}\rangle$ and then sends $|\phi_{\mathcal{D}}\rangle$ to $\mathbf{V}$. This prover satisfies the completeness property. For the classical model we will show that a prover who commits to the same history state also satisfies completeness.

Theorem 7 (Completeness)**.** *There exists a QPT prover $\mathbf{P}^{\mathsf{O}(*)}$, such that for any distribution $\mathcal{D} \in \mathfrak{D}(n)$, any λ-collapsing claw-free family $\mathcal{F}$, $\mathbf{P}^{\mathsf{O}(\mathcal{D})}$ wins in Protocol 5 with probability 1 and we have:*

- $S \sim \mathcal{D}^{|S|}$

We note that completeness for this protocol is in some sense easier to prove than the completeness of the protocols described in the previous sections. The reason for this is that the protocol does not abort when the estimates are not sattisfying to desired bounds. We proceed by describing the strategy for the honest prover and the proof of the completeness.

Proof. Let us denote $2n + T' + 1$ by n'. The honest prover will create a state such that each bit b_i would correspond to the measurement of a qubit from the history state. They extend the state with zeros in the following way.

$$|\phi_{hist}\rangle \left|0^{mn'}\right\rangle_X = \sum_{b_1,\dots,b_{n'}} \alpha_{b_1,\dots,b_{n'}} |b_1\rangle |0^m\rangle \dots |b_{n'}\rangle |0^m\rangle \tag{17}$$

By applying QFT on the 0 registers we get the state,

$$|\phi'\rangle = \frac{1}{\sqrt{2^{mn'}}} \sum_{b_1,\dots,b_{n'}} \alpha_b \Big(\sum_{z \in \{0,1\}^{mn'}} |b_1\rangle |z_1\rangle \dots |b_{n'}\rangle |z_{n'}\rangle \Big) \tag{18}$$

We add a zero register to $|\psi\rangle$ and evaluate f_{pk_i} on the superpositions to get the state,

$$|\phi''\rangle = \frac{1}{\sqrt{2^{mn'}}} \sum_{b_1,\dots,b_{n'}} \alpha_b \Big(\sum_{x \in \{0,1\}^{mn'}} |b_1\rangle |z_1\rangle |f_{pk_1}(b_1, z_1)\rangle \dots |b_{n'}\rangle |z_{n'}\rangle \left|f_{pk_{n'}}(b_{n'}, z_{n'})\right\rangle \Big) \tag{19}$$

$\mathbf{P}$ proceeds by measuring the image registers to get values $y_1, \dots, y_{n'}$. The state after obtaining this measurement outcome will be,

$$|\phi_{\mathbf{P}}\rangle = \sum_b \alpha_b |b_1\rangle |x_{1,b_1}\rangle |y_1\rangle \dots |b_{n'}\rangle \left|x_{n',b_{n'}}\right\rangle |y_{n'}\rangle \tag{20}$$

where (b_i, x_{i,b_i}) is the b_i-labeled preimage of y_i under f_{pk_i}.

Upon receiving challenge 0, $\mathbf{P}$ measure the state $|\phi_{\mathbf{P}}\rangle$ in computational basis and sends $x_1, \dots, x_n = (b_1, x_{1,b_1}), \dots, (b_{n'}, x_{n',b_{n'}})$ values to $\mathbf{V}$. From equation 20 one can deduce that $b_1(x_1), \dots, b_{n'}(x_{n'})$ as in the protocol is distributed identically to the outcome of the measurement of the history state in computational basis. By construction $\mathbf{P}$ always succeeds in the preimage check, hence, wins with probability 1.

As the outcomes of measuring the b registers of state $|\phi_{\mathbf{P}}\rangle$ are distributed identically to the measurement outcomes of the history state $|\phi_{\text{hist}}\rangle$ we have $S \sim \mathcal{D}^{|S|}$, following the completeness proof from Theorem 5. □

In fact the completeness can be modified so that it captures the fact that the estimates computed in the protocol are close to the actual energy and outcome probability. We state the theorem here, but as similar statements are proven in the soundness we avoid repeating the proof here.

Theorem 8 (Completeness 2)**.** *There exists a QPT prover $\mathbf{P}^{\mathsf{O}(*)}$, such that for any distribution $\mathcal{D} \in \mathfrak{D}(n)$, any family of λ-collapsing claw-free family $\mathcal{F}$, $\mathbf{P}^{\mathsf{O}(\mathcal{P})}$ wins $N = O(\frac{K}{\eta^4} poly(n, T) \log(1/\delta))$ iterations of Protocol 5 with probability 1, we have that with probability at least $1 - \delta$,*

- $|S| \geq \Omega(K)$,

- $p/n_2 \geq 1 - 2d_H^2(\mathcal{D}, \mathcal{D}_C)$,[12]
- $\gamma/n_1\binom{n'}{2} \in [q - \frac{\eta^2}{2T'^3}, q + \frac{\eta^2}{2T'^3}]$, *where* $q = \langle\phi_{\mathit{hist}}| H_G |\phi_{\mathit{hist}}\rangle$,
- $S \sim (\mathcal{D})^{|S|}$.

D.2.4 Soundness

Now that we have established an honest prover strategy, the only thing left is to prove that for any prover who wins the game with a high probability, the verifier $\mathbf{V}$ would collect samples from a distribution close to the $\mathcal{D}_C$.

The key fact to prove the soundness of the protocol is that the values x_i and d_i are somewhat correlated with the measurement outcomes of the i^{th} qubit of the prover's extracted qubit.

Fact 2 ([Mah18])**.** *Let $\mathcal{F}$ be a collapsing claw-free family and let $\mathbf{P}$ be any QPT prover who wins in Protocol 5 with probability 1, let ρ be the prover's extracted qubits, B_i and D_i the outcome of measuring the i^{th} qubit of ρ in the computational and the Hadamard basis respectively. For any parity $\chi : \{0,1\}^{n'} \to \{-1,+1\}$ we have,*

- *(computational basis measurement) $\chi(B_1, \ldots, B_{n'})$ is identically distributed to*

$$\chi(b_1(x_1), \ldots, b_{n'}(x_{n'})).$$

- *(the Hadamard basis measurement) $\chi(D_1, \ldots, D_{n'})$ is computationally indistinguishable from $\chi(d_1 \cdot (x_{1,0} + x_{1,1}), \ldots, d_{n'} \cdot (x_{n',0} + x_{n',1}))$.*

Lemma 7. *For any T-gate quantum circuit C, and its corresponding T'-gate comparison circuit G, let $\mathcal{F} = \{f_{pk}\}$ be a family of claw-free functions satisfying the collapsing property, and let H_G be an $n' = 2n + T' + 1$ qubit Hamiltonian corresponding to G. For any QPT prover $\mathbf{P}$ let ρ be the reduced density of the extracted qubits of $\mathbf{P}$, and let $\mathcal{D}^A$ be the distribution of outcomes of measuring the adv register of ρ conditioned on the measurement outcome of clock, out and aux register being 0 in the computational basis. If $\mathbf{P}$ wins in protocol 5 with probability 1 we have,*

- $\mathbb{E}[\gamma/n_1\binom{n'}{2}] \approx \mathit{Tr}(H_G\rho)$,
- $\mathbb{E}[p/n_2] = \frac{\mathit{Tr}(|T'\rangle\langle T'|_{\mathit{clock}} \otimes |1\rangle\langle 1|_{\mathit{out}}\rho)}{\mathit{Tr}(|T'\rangle\langle T'|_{\mathit{clock}}\rho)}$,
- $S \sim (\mathcal{D}^A)^{|S|}$.

Proof. To prove this theorem we will be using Fact 2. First we prove the properties only relying on the computational measurements, namely properties about p and S. Let us focus on distribution of S first.

A sample is collected if $\mathsf{type} = 1$, the challenge bit b is equal to 0 and $b(o), b(\text{clock}), b(\text{aux})$ are all 0. Due to Fact 2 this is equivalent to when the outcome of measuring the aux, clock and out registers of ρ are 0. Conditioned on this happening the sample collected has the exact same distribution as measuring the adv register of ρ, which is equivalent to $\mathcal{D}^A$.

The estimate p is increased by $b(o)$, when $\mathsf{type} = 2$, the challenge bit is 0 and

$$b(t_0), \ldots, b(t_{T'}) = (1, \ldots, 1).$$

Conditioned on the clock being T', The expectation of $b(o)$ is $\mathrm{Tr}(|T'\rangle\langle T'|_{\text{clock}} \otimes |1\rangle\langle 1|_{\text{out}}\,\rho)$ due to fact 2. Hence we have $\mathbb{E}[p] = n_2 \frac{\mathrm{Tr}(|T'\rangle\langle T'|_{\text{clock}} \otimes |1\rangle\langle 1|_{\text{out}}\rho)}{\mathrm{Tr}(|T'\rangle\langle T'|_{\text{clock}}\rho)}$.

The next thing to prove is that the energy estimate has the desired expectation. If we consider the n_1 rounds in which we change the energy estimate, the expectation of the amount of change done to γ is equal to:

$$-\frac{1}{2\binom{n'}{2}} \sum_{i,j} J_{i,j}(-1^{b_i(x_i)+b_j(x_j)} + -1^{d_i(x_{i,0}+x_{i,1})+d_j(x_{j,0}+x_{j,1})})$$

[12] This is done similar to the protocol described in Figure 4

For $b_i(x_i)$ and $b_j(x_j)$, we know that these random variables are distributed identically to measurement of ρ in computational basis. The only issue is that $d_i(x_{i,0} + x_{i,1})$ is not distributed identically to Hadamard measurement of ρ, but rather is computationally indistinguishable from it.

However for any parity χ if the distance between the expectations of $\chi(d_i(x_{i,0} + x_{i,1}))$ and χ applied on the measurement outcomes of ρ in the Hadamard basis is negligible in λ; as otherwise an adversary could distinguish between the two by random sampling using only $O(1/poly(\mu))$ samples. Hence we have,

$$\mathbb{E}[J_{i,j}(-1)^{b_i(x_i)+b_j x_j}] = J_{i,j}\mathrm{Tr}(\sigma_{Z,i}\sigma_{Z,j}\rho)$$
$$\mathbb{E}[J_{i,j}(-1)^{d_i(x_{i,0}+x_{i,1})+d_j(x_{j,0}+x_{j,1})}] = J_{i,j}(\mathrm{Tr}(\sigma_{X,i}\sigma_{X,j}\rho) \pm ngl(\lambda))$$

Hence we have that $\mathbb{E}[\gamma] \approx n_1 \frac{1}{\binom{n'}{2}}\mathrm{Tr}(H_G\rho)$ as desired. □

Theorem 9 (Perfect Prover Soundness)**.** *For any security parameter λ, any T-gate circuit C acting on n-qubits, the protocol defined in Figure 5 has the following properties. It is an interactive protocol $(\mathbf{V}, *)$ between a classical verifier $\mathbf{V}$ and a quantum prover. For any QPT prover $\mathbf{P}$, let ρ be the reduced density of the extracted qubits of $\mathbf{P}$, and $\mathcal{D}^A$ be the distribution of outcomes of measuring the adv register of ρ in the computational basis conditioned on the measurement outcome of clock, out and aux registers being 0. If $\mathbf{P}$ wins $N = O(\frac{K}{\eta^4}poly(n,T)\log(1/\delta))$ iterations of the protocol with probability 1 and $\frac{\gamma}{n_1}\binom{2n+T+1}{2} \leq \frac{\eta^2}{2T'^3}$ and $\frac{p}{n_2} \geq 1 - 2\eta^2$, then with probability $1-\delta-\mu(\lambda)$, we have,*

- $d_H(\mathcal{D}_C, \mathcal{D}^A) \leq O(\eta^{1/4})$,
- $S \sim (\mathcal{D}^A)^{|S|}$.

where μ is a negligible function.

Note. The guarantee expressed in the last sentence of Theorem 9 might seem mysterious at first. Note however that the conditions contained there are equivalent to the checks performed at the last step in the constant quantum memory protocol from Figure 4. The fact that the checks are contained in the statement of the theorem and not in the protocol itself allows us to analyze perfect provers only and simplifies the presentation considerably.

Proof of Theorem 9. Let G be the T' comparison circuit of C and let H_G be the corresponding 2-local Hamiltonian acting on $n' = 2n + T' + 1$ qubits.

Applying Lemma 5 we have that with probability $1 - e^{-\frac{N}{18}}$ we have $n_1 \geq \Omega(N)$. From Lemma 7 we have,

$$\mathbb{E}\left[\frac{\gamma}{n_1}\binom{n'}{2}\right] \approx \mathrm{Tr}(H_G\rho) \tag{21}$$
$$\mathbb{E}\left[\frac{p}{n_2}\right] = \frac{\mathrm{Tr}(|T'\rangle\langle T'|_{\text{clock}} \otimes |1\rangle\langle 1|_{\text{out}}\,\rho)}{\mathrm{Tr}(|T'\rangle\langle T'|_{\text{clock}}\,\rho)} \tag{22}$$

Using Fact 3 we get,

$$\mathbb{P}\left[\left|\frac{\gamma}{n_1}\binom{n'}{2} - \mathrm{Tr}(H_G\rho)\right| \geq \frac{\eta^2}{2T'^3}\right] \leq 2e^{-\frac{\eta^4 n_1}{8\binom{n'}{2}^2 T'^6 J}} \tag{23}$$
$$\mathbb{P}\left[\left|\frac{p}{n_2} - \frac{\mathrm{Tr}(|T'\rangle\langle T'|_{\text{clock}} \otimes |1\rangle\langle 1|_{\text{out}}\,\rho)}{\mathrm{Tr}(|T'\rangle\langle T'|_{\text{clock}}\,\rho)}\right| \geq \frac{3\eta^2}{2}\right] \leq 2e^{-\frac{9\eta^4 n_2}{8}}, \tag{24}$$

where $J = \sup_{i \neq j \in [n']}\{|J_{i,j}|\}$.

If we use the hypothesis of the theorem, (23) and (24) we get that with probability $1 - \frac{\delta}{8}$,

$$\mathrm{Tr}(H_G\rho) \leq \frac{\eta^2}{T'^3} \tag{25}$$
$$\frac{\mathrm{Tr}(|T'\rangle\langle T'|_{\text{clock}} \otimes |1\rangle\langle 1|_{\text{out}}\,\rho)}{\mathrm{Tr}(|T'\rangle\langle T'|_{\text{clock}}\,\rho)} \geq 1 - 2\eta^2 - \frac{3\eta^2}{2} \geq 1 - \frac{7\eta^2}{2} \tag{26}$$

Note that (26) implies that probability of the measurement outcome of the out register being 1, when the clock is T' is at least $1 - \frac{7\eta^2}{2}$. By employing Corollary 1 we have that with probability $\frac{1 \pm 7\eta}{T'+1}$, $d_H(\mathcal{D}_C, \mathcal{D}^A) \leq O(\eta^{1/4})$.

By Fact 3 we have that $n_2 = \Omega(\frac{N}{T'})$ with probability $1 - \frac{\delta}{20}$ so it's enough for $N = O(\frac{1}{\eta^4} \log(1/\delta) \binom{n'}{2}^2 T'^7 J) = O(\frac{1}{\eta^4} \log(\frac{1}{\delta}) \mathrm{poly}(T, n))$ for (24) to hold with probability $\leq \delta/10$. If we apply the union bound over all failure events we get that all the conditions will be satisfied with probability at least $1 - \delta$, hence with probability $1 - \delta$ we get $d_H(\mathcal{D}_C, \mathcal{D}^A) \leq O(\eta^{1/4})$.

The second bullet follows directly from Lemma 7. □

Discussion We note that we have proven the soundness of our protocol only in the perfect prover setting. The problem with this statement is that, it can not be verified that the prover is winning with probability 1. Also the soundness guaranty is different from the previous sections as the game does not abort when the estimates do not satisfy the bound. The reason we modified the game in this manner is that if the game aborted after checking the bounds, even the honest prover would not have won the game with probability 1, as there is a small probability that the estimates computed in the protocol are far from the expected value.

However, it is possible to achieve a stronger soundness guaranties, similar to Theorem 5. This requires more adjustments to the protocol which allows one to prove the soundness for a non-perfect prover by following a similar path as the one in Claim 7.1 of [Mah18], where a reduction from the non-perfect prover to a perfect prover is given.

E Generalized Setting For the Quantum Verifier Protocol

E.1 Non i.i.d. Quantum Verifier

Let us now relax the assumption that $\mathbf{P}$ acts i.i.d., i.e. that $\mathbf{P}$ sends the same $|\psi_{\mathcal{D}^A}\rangle$ in every round. We still assume at this point that the states sent by $\mathbf{P}$ are pure. For a discussion about mixed states see Section E.2. First we state a slightly changed theorem.

Theorem 10 (Quantum Verifier). *For every circuit C acting on n qubits, for every $\delta \in (0, \frac{1}{3})$ and all $\eta > 0$ sufficiently small there exists an interactive protocol between a quantum verifier $\mathbf{V}$ and a quantum prover $\mathbf{P}$ with the following properties. The protocol runs in $N = O(\frac{1}{\eta^2} \log(1/\delta))$ rounds, in each round $\mathbf{P}$ sends a pure quantum state on n qubits to $\mathbf{V}$. At the end of the protocol $\mathbf{V}$ outputs $\perp$ when it rejects the interaction or $x \in \{0,1\}^n$ when it accepts.*

- *(Completeness) There exists $\mathbf{P}^{\mathsf{O}(*)}$ such that for every $\mathcal{D} \in \mathfrak{D}(n)$ satisfying $d_H(\mathcal{D}, \mathcal{D}_C) \leq \eta$ the following holds.. With probability $1 - \delta$ over the randomness in the protocol $\mathbf{P}^{\mathsf{O}(\mathcal{D})}$ succeeds and $x \sim_{i.i.d.} \mathcal{D}$.*
- *(Soundness) For every $\mathbf{P}$ that succeeds with probability $\geq 1 - \frac{\delta}{2}$ we have that with probability $1 - \delta$ over the randomness in the protocol $x \sim_{i.i.d.} \mathcal{D}^A$ and $d_H(\mathcal{D}_C, \mathcal{D}^A) \leq O\left(\eta^{1/4}\right)$.*

Before going to the proof of the theorem we first state a technical lemma.

Lemma 8. *For every $\eta > 0, k \in \mathbb{N}$, every set of distributions $\mathcal{D}, \mathcal{D}_C, \mathcal{D}_1^A, \ldots, \mathcal{D}_k^A \in \mathfrak{D}(n)$ and every $q_1, \ldots, q_k \in [0,1]$ such that $\sum_{i=1}^k q_1 = 1$ the following holds. Let $f(x) = \frac{1}{2}(1 + (1 - x^2)^2)$. If $\sum_{i=1}^k q_i f(d_H(\mathcal{D}_C, \mathcal{D}_i^A)) \geq 1 - 50\eta^2$ then*

$$d_H\left(\sum_{i=1}^k q_i \mathcal{D}_i^A, \mathcal{D}_C\right) \leq O\left(\eta^{1/4}\right).$$

Proof. We bound the quantity

$$\begin{aligned}
&\sum_{i=1}^k q_i d_H(\mathcal{D}_C, \mathcal{D}_i^A) \\
&\leq \sum_{i=1}^k q_i \left(d_H(\mathcal{D}_C, \mathcal{D}_i^A) \mathbb{1}_{\{d_H(\mathcal{D}_C, \mathcal{D}_i^A) \leq \sqrt{\eta}\}} + \mathbb{1}_{\{d_H(\mathcal{D}_C, \mathcal{D}_i^A) > \sqrt{\eta}\}} \right) \\
&\leq \sqrt{\eta} + \sum_{i=1}^k q_i \mathbb{1}_{\{d_H(\mathcal{D}_C, \mathcal{D}_i^A) > \sqrt{\eta}\}} \qquad (27)
\end{aligned}$$

```
Verifier V                                                              Prover P
$p := 0, S := \emptyset$
for $i \in \{1, \dots N\}$ where $N = O\left(\frac{1}{\eta^2}\log(1/\delta)\right)$
 do :
   type $\xleftarrow{\$} \{0,1\}$
                                                          $|\psi_{\mathcal{D}^A}\rangle = (\mathbb{C}^2)^{\otimes n}$
                                   $\xleftarrow{|\psi_{\mathcal{D}^A}\rangle}$
   if type $= 0$ :
     Set $b_1, b_2, \dots, b_n$ to be the measurement of
     $|\psi_{\mathcal{D}^A}\rangle$ in the $Z$ basis
     $S := S \cup \{(b_j)_{j \in \{1,\dots,n\}}\}$
   if type $= 1$ :
     Set $p_i$ to be the measurement of the out register of
     $G\,|0\rangle_{\text{out}}\,|\psi_{\mathcal{D}^A}\rangle_{\text{adv}}\,|0^{\otimes n}\rangle_{\text{aux}}$ in the Z basis
     $p := p + \frac{p_i}{N}$
done :
if $p < 1 - 2\eta^2$   abort
Set $x$ to be an element of $S$ chosen uniformly at random
return $x$
```

Figure 6: The interactive protocol for the model where the verifier has access to a quantum computer and the prover doesn't need to act in an i.i.d. fashion.

Let $l = \sum_{i=1}^k q_i \mathbb{1}_{\{d_H(\mathcal{D}_C, \mathcal{D}_i^A) > \sqrt{\eta}\}}$. By definition and the fact that $f(x) \leq 1 - x^2/2$ we have

$$\left(1 - \frac{\eta}{2}\right) l + (1 - l) \geq \sum_{i=1}^k q_i f(d_H(\mathcal{D}_C, \mathcal{D}_i^A)).$$

Using the assumption $\sum_{i=1}^k q_i f(d_H(\mathcal{D}_C, \mathcal{D}_i^A)) \geq 1 - 50\eta^2$ we get $l \leq 100\eta$. Plugging it in (27) we get

$$\begin{aligned} 101\sqrt{\eta} &\geq \sqrt{\eta} + 100\eta \\ &\geq \sum_{i=1}^k q_i d_H(\mathcal{D}_C, \mathcal{D}_i^A) \\ &\geq \sqrt{2}\sum_{i=1}^k q_i \triangle(\mathcal{D}_C, \mathcal{D}_i^A) && \text{As } d_H(\mathcal{P}, \mathcal{Q}) \geq \sqrt{2}\triangle(\mathcal{P}, \mathcal{Q}) \\ &\geq \sqrt{2}\triangle\left(\mathcal{D}_C, \sum_{i=1}^k q_i \mathcal{D}_i^A\right) && \text{Triangle inequality and identity } \triangle(\mathcal{P}, \mathcal{Q}) = \frac{1}{2}\|\mathcal{P} - \mathcal{Q}\|_1. \end{aligned} \tag{28}$$

Now we can bound the quantity of interest

$$\begin{aligned} & d_H\left(\sum_{i=1}^k q_i \mathcal{D}_i^A, \mathcal{D}_C\right) \\ &\leq \sqrt{\triangle\left(\sum_{i=1}^k q_i \mathcal{D}_i^A, \mathcal{D}_C\right)} && \text{By } d_H(\mathcal{P}, \mathcal{Q}) \leq \sqrt{\triangle(\mathcal{P}, \mathcal{Q})} \\ &\leq O\left(\eta^{1/4}\right) && \text{By (28)} \end{aligned}$$

□

Proof of Theorem 10. The modified protocol is given in Figure 6. In each run at most one sample is generated. The number of iterations is changed from $O\left(\frac{K}{\eta^2}\log(1/\delta)\right)$ to $O\left(\frac{1}{\eta^2}\log(1/\delta)\right)$. The biggest change is in the very last step of the protocol, where instead of returning the whole set S we return a random element from S. The reason behind this change will hopefully become clear at the end of the proof.

It suffices to prove the soundness as the completeness proof is analogous to the proof of Theorem 2.

Assume that $\mathbf{P}$ sends the states $\left|\psi_{\mathcal{D}_1^A}\right\rangle, \left|\psi_{\mathcal{D}_2^A}\right\rangle, \ldots, \left|\psi_{\mathcal{D}_N^A}\right\rangle$ to $\mathbf{V}$. Let the rounds in which the type is 1 be $I \subseteq [N]$ and denote $|I|$ by k. Then for every $i \in I$ we have that $\mathbf{V}$ gets a sample according to a Bernoulli variable with parameter

$$\langle 0^n|_{\text{aux}} \left\langle \psi_{\mathcal{D}_i^A}\right|_{\text{adv}} \langle 0|_{\text{out}} G^\dagger \Pi^{(1)}_{\text{out}} G |0\rangle_{\text{out}} \left|\psi_{\mathcal{D}_i^A}\right\rangle_{\text{adv}} |0^n\rangle_{\text{aux}}.$$

Thus by Fact 3 and Corollary 1 we have that with probability $1 - \frac{\delta}{2}$

$$\left| p - \frac{1}{k}\sum_{i \in I} f(d_H(\mathcal{D}_C, \mathcal{D}_i^A)) \right| \leq \eta^2, \tag{29}$$

where $f(x) = \frac{1}{2}(1 + (1-x^2)^2)$.

$\mathbf{P}$ succeeds with probability $1 - \frac{\delta}{2}$ so by (29) and the union bound we get that with probability $1 - \delta$

$$\frac{1}{k}\sum_{i \in I} f(d_H(\mathcal{D}_C, \mathcal{D}_i^A)) \geq 1 - 2\eta^2 - \eta^2 \geq 1 - 3\eta^2. \tag{30}$$

By Lemma 8 we get then

$$\mathbb{P}_I\left[d_H\left(\frac{1}{k}\sum_{i\in I} \mathcal{D}_i^A, \mathcal{D}_C \right) \geq O(\eta^{1/4}) \right] \leq \delta.$$

As I and $[N] \setminus I$ have the same distribution we also get

$$\mathbb{P}_I\left[d_H\left(\frac{1}{N-k}\sum_{i\notin I} \mathcal{D}_i^A, \mathcal{D}_C \right) \geq O(\eta^{1/4}) \right] \leq \delta.$$

Finally note that the samples we collected in S came exactly from the distribution $S \sim \Pi_{i \notin I}\mathcal{D}_i^A$, so if we choose the sample to return x as a uniformly random element of S then $x \sim \frac{1}{N-k}\sum_{i\notin I}\mathcal{D}_i^A$. To conclude note that $\mathcal{D}^A = \frac{1}{N-k}\sum_{i \notin I}\mathcal{D}_i^A$.

□

E.2 Prover sending mixed states

In this section we explain what happens when instead of sending a pure state $|\psi_{\mathcal{D}^A}\rangle$, $\mathbf{P}$ is allowed to send a mixed state ρ_A. This means that $\mathbf{P}$ can prepare a state $|\psi\rangle_{\text{E,F}}$ in a bigger space $(\mathbb{C}^2)^{\otimes n}_{\text{E}} \otimes H_{\text{F}}$ and send only the E part of the system to $\mathbf{V}$. We still assume here that $\mathbf{P}$ acts in an i.i.d. fashion. In this setting the guarantee for soundness will deteriorate (as in Theorem 10) to $d_H(\mathcal{D}_C, \mathcal{D}^A) \leq O(\eta^{1/4})$ in comparison to $d_H(\mathcal{D}_C, \mathcal{D}^A) \leq O(\eta)$ as in Theorem 2. The slightly changed theorem becomes

Theorem 11 (Quantum Verifier with Mixed States)**.** *For every circuit C acting on n qubits, for every $\delta \in (0, \frac{1}{3})$, $K \in \mathbb{N}$ and all $\eta > 0$ sufficiently small there exists an interactive protocol between a quantum verifier $\mathbf{V}$ and a quantum prover $\mathbf{P}$ with the following properties. The protocol runs in $N = O(\frac{K}{\eta^2}\log(1/\delta))$ rounds and in each round $\mathbf{P}$ sends a (potentially mixed) quantum state on n qubits to $\mathbf{V}$. At the end of the protocol $\mathbf{V}$ outputs $\perp$ when it rejects the interaction or it outputs $S = \{x_1, \ldots, x_{|S|}\}$, $x_i \in \{0,1\}^n$, when it accepts.*

- *(Completeness) There exists $\mathbf{P}^{\mathsf{O}(*)}$ such that for every $\mathcal{D} \in \mathfrak{D}(n)$ satisfying $d_H(\mathcal{D}, \mathcal{D}_C) \leq \eta$ the following holds. With probability $1 - \delta$ over the randomness in the protocol $\mathbf{P}^{\mathsf{O}(\mathcal{D})}$ succeeds, $S \sim_{\text{i.i.d.}} \mathcal{D}^{|S|}$ and $|S| \geq \Omega(K)$.*

- *(Soundness) For every* $\mathbf{P}$ *that succeeds with probability at least* $\frac{2}{3}$ *we have* $S \sim_{\textit{i.i.d.}} (\mathcal{D}^A)^{|S|}$ *and* $d_H(\mathcal{D}_C, \mathcal{D}^A) \leq O(\eta^{1/4})$*.*

Proof of Theorem 11. Note that we only need to verify the soundness property as for the completeness we know that $\mathbf{P}$ sends pure states. By the ensemble interpretation of density matrices we can express

$$\rho_A = \sum_{j=1}^{k} q_i \left|\psi_{\mathcal{D}_i^A}\right\rangle \left\langle\psi_{\mathcal{D}_i^A}\right|, \tag{31}$$

where $\left|\psi_{\mathcal{D}_i^A}\right\rangle \in (\mathbb{C}^2)^{\otimes n}$. This expression is not unique but it will not play a role for us. We observe that measuring ρ_A in the Z basis and collecting a sample is equivalent to collecting a sample from a distribution $\sum_{j=1}^k q_i \mathcal{D}_i^A$. By Corollary 1 we know that the probability of obtaining outcome 1 when running G on $\left|\psi_{\mathcal{D}_i^A}\right\rangle$ and measuring out register is equal to the Bernoulli variable with parameter $f(d_H(\mathcal{D}_C, \mathcal{D}_i^A))$, for $f(x) = \frac{1}{2}(1 + (1 - x^2)^2)$. The distribution of measuring the out register when running G on ρ_A is thus equal to

$$\sum_{j=1}^{k} q_i \cdot f(d_H(\mathcal{D}_C, \mathcal{D}_i^A)).$$

By Fact 3 and the setting of N we have that with probability $1 - \frac{\delta}{2}$

$$\left| p - \sum_{j=1}^{k} q_i \cdot f(d_H(\mathcal{D}_C, \mathcal{D}_i^A)) \right| \leq \eta^2$$

$\mathbf{P}$ succeeds with probability $\frac{2}{3}$ so by the union bound and the fact that $\frac{1}{3} + \frac{\delta}{2} < 1$ we get that $\sum_{j=1}^k q_i \cdot f(d_H(\mathcal{D}_C, \mathcal{D}_i^A)) \geq p - \eta^2 \geq 1 - 3\eta^2$. Application of Lemma 8 finishes the proof. □

F Basic Facts and Omitted Proofs

We denote the total variation distance of $\mathcal{P}, \mathcal{Q}$ as $\triangle(\mathcal{P}, \mathcal{Q}) = \frac{1}{2}\|\mathcal{P} - \mathcal{Q}\|_1$. The two similarity measures satisfy $d_H^2(\mathcal{P}, \mathcal{Q}) \leq \triangle(\mathcal{P}, \mathcal{Q}) \leq \sqrt{2} d_H(\mathcal{P}, \mathcal{Q})$. A direct calculation yields the useful identity $1 - d_H^2(\mathcal{P}, \mathcal{Q}) = \sum_{x \in \{0,1\}^n} \sqrt{\mathcal{P}(x)\mathcal{Q}(x)}$.

Fact 3 (Chernoff-Hoeffding)**.** *Let* $X_1, \ldots, X_k$ *be independent Bernoulli variables with parameter* p*. Then for every* $0 < \epsilon < 1$

$$\mathbb{P}\left[\left|\frac{1}{k}\sum_{i=1}^{k} X_i - p\right| > \epsilon\right] \leq 2e^{-\frac{\epsilon^2 k}{2}}.$$

Lemma 9. *For every* $|\psi_{\mathcal{D}^A}\rangle$ *and* C *the probability of obtaining outcome* $|1\rangle$ *when measuring the out register of* $G\,|0\rangle_{\textit{out}}\,|\psi_{\mathcal{D}^A}\rangle_{\textit{adv}}\,|0^{\otimes n}\rangle_{\textit{aux}}$ *is equal to*

$$\frac{1}{2}\left(1 + |\langle\psi_{\mathcal{D}^A}|\psi_{\mathcal{D}_C}\rangle|^2\right).$$

Proof. We analyze the evolution of the state

$$\begin{aligned}
&(\text{NOT} \otimes \mathbb{I} \otimes \mathbb{I})(H \otimes \mathbb{I} \otimes \mathbb{I})(\text{CSWAP})(H \otimes \mathbb{I} \otimes \mathbb{I})(\mathbb{I} \otimes \mathbb{I} \otimes U_C)\,|0\rangle\,|\psi_{\mathcal{D}^A}\rangle\,\left|0^{\otimes n}\right\rangle \\
&= (\text{NOT} \otimes \mathbb{I} \otimes \mathbb{I})(H \otimes \mathbb{I} \otimes \mathbb{I})(\text{CSWAP})(H \otimes \mathbb{I} \otimes \mathbb{I})\,|0\rangle\,|\psi_{\mathcal{D}^A}\rangle\,|\psi_{\mathcal{D}_C}\rangle \\
&= (\text{NOT} \otimes \mathbb{I} \otimes \mathbb{I})(H \otimes \mathbb{I} \otimes \mathbb{I})(\text{CSWAP})\left(\frac{|0\rangle + |1\rangle}{\sqrt{2}}\right)|\psi_{\mathcal{D}^A}\rangle\,|\psi_{\mathcal{D}_C}\rangle \\
&= (\text{NOT} \otimes \mathbb{I} \otimes \mathbb{I})(H \otimes \mathbb{I} \otimes \mathbb{I})\frac{1}{\sqrt{2}}\left(|0\rangle\,|\psi_{\mathcal{D}^A}\rangle\,|\psi_{\mathcal{D}_C}\rangle + |1\rangle\,|\psi_{\mathcal{D}_C}\rangle\,|\psi_{\mathcal{D}^A}\rangle\right) \\
&= (\text{NOT} \otimes \mathbb{I} \otimes \mathbb{I})\frac{1}{2}\left((|0\rangle + |1\rangle)\,|\psi_{\mathcal{D}^A}\rangle\,|\psi_{\mathcal{D}_C}\rangle + (|0\rangle - |1\rangle)\,|\psi_{\mathcal{D}_C}\rangle\,|\psi_{\mathcal{D}^A}\rangle\right) \\
&= \frac{1}{2}\left(|1\rangle\left[|\psi_{\mathcal{D}^A}\rangle\,|\psi_{\mathcal{D}_C}\rangle + |\psi_{\mathcal{D}_C}\rangle\,|\psi_{\mathcal{D}^A}\rangle\right] + |0\rangle\left[|\psi_{\mathcal{D}^A}\rangle\,|\psi_{\mathcal{D}_C}\rangle - |\psi_{\mathcal{D}_C}\rangle\,|\psi_{\mathcal{D}^A}\rangle\right]\right).
\end{aligned}$$

The probability of obtaining outcome $|1\rangle$ when measuring the out register in the Z basis is then

$$
\begin{aligned}
&\frac{1}{4}\left[\left(\langle\psi_{\mathcal{D}^A}|\,\langle\psi_{\mathcal{D}_C}| + \langle\psi_{\mathcal{D}_C}|\,\langle\psi_{\mathcal{D}^A}|\right)\left(|\psi_{\mathcal{D}^A}\rangle\,|\psi_{\mathcal{D}_C}\rangle + |\psi_{\mathcal{D}_C}\rangle\,|\psi_{\mathcal{D}^A}\rangle\right)\right]\\
&= \frac{1}{4}[\langle\psi_{\mathcal{D}^A}|\psi_{\mathcal{D}^A}\rangle\,\langle\psi_{\mathcal{D}_C}|\psi_{\mathcal{D}_C}\rangle + \langle\psi_{\mathcal{D}^A}|\psi_{\mathcal{D}_C}\rangle\,\langle\psi_{\mathcal{D}_C}|\psi_{\mathcal{D}^A}\rangle +\\
&+ \langle\psi_{\mathcal{D}_C}|\psi_{\mathcal{D}^A}\rangle\,\langle\psi_{\mathcal{D}^A}|\psi_{\mathcal{D}_C}\rangle + \langle\psi_{\mathcal{D}_C}|\psi_{\mathcal{D}_C}\rangle\,\langle\psi_{\mathcal{D}^A}|\psi_{\mathcal{D}^A}\rangle]\\
&= \frac{1}{4}\left[2\|\psi_{\mathcal{D}^A}\|^2\|\psi_{\mathcal{D}_C}\|^2 + 2\,\langle\psi_{\mathcal{D}_C}|\psi_{\mathcal{D}^A}\rangle\,\langle\psi_{\mathcal{D}^A}|\psi_{\mathcal{D}_C}\rangle\right]\\
&= \frac{1}{2}\left[1 + |\,\langle\psi_{\mathcal{D}^A}|\psi_{\mathcal{D}_C}\rangle\,|^2\right].
\end{aligned}
$$

□

Lemma 1. Apply Lemma 9 and the $1 - d_H^2(\mathcal{P}, \mathcal{Q}) = \sum_{x\in\{0,1\}^n}\sqrt{\mathcal{P}(x)\mathcal{Q}(x)}$ identity. □

Lemma 2. For $b \in \{0, 1\}$, n_b can be seen as a sum of random Bernoulli variables $\{x_i\}_{i\in[N]}$ with parameter $1/2$. Then, by Fact 3, we get that $\mathbb{P}[|\frac{n_b}{N} - \frac{1}{2}| > \frac{1}{4}] \leq 2e^{-\frac{N}{32}} \leq \frac{\delta}{2}$. We finish by applying the union bound over the error events. □