Memory-Sample Lower Bounds for Learning with Classical-Quantum Hybrid Memory*

Qipeng Liu
Simons Institute for Theory of Computing
Berkeley, USA
qipengliu0@gmail.com

Ran Raz
Princeton University
Princeton, USA
ranr@cs.princeton.edu

Wei Zhan
Princeton University
Princeton, USA
weizhan@cs.princeton.edu

ABSTRACT

In a work by Raz (J. ACM and FOCS 16), it was proved that any algorithm for parity learning on n bits requires either $\Omega(n^2)$ bits of classical memory or an exponential number (in n) of random samples. A line of recent works continued that research direction and showed that for a large collection of classical learning tasks, either super-linear classical memory size or super-polynomially many samples are needed. All these works consider learning algorithms as classical branching programs, which perform classical computation within bounded memory.

However, these results do not capture all physical computational models, remarkably, quantum computers and the use of quantum memory. It leaves the possibility that a small piece of quantum memory could significantly reduce the need for classical memory or samples and thus completely change the nature of the classical learning task. Despite the recent research on the necessity of quantum memory for intrinsic quantum learning problems like shadow tomography and purity testing, the role of quantum memory in classical learning tasks remains obscure.

In this work, we study classical learning tasks in the presence of quantum memory. We prove that any quantum algorithm with both, classical memory and quantum memory, for parity learning on n bits, requires either $\Omega(n^2)$ bits of classical memory or $\Omega(n)$ bits of quantum memory or an exponential number of samples. In other words, the memory-sample lower bound for parity learning remains qualitatively the same, even if the learning algorithm can use, in addition to the classical memory, a quantum memory of size cn (for some constant $c > 0$).

Our result is more general and applies to many other classical learning tasks. Following previous works, we represent by the matrix $M : A \times X \to \{-1, 1\}$ the following learning task. An unknown x is sampled uniformly at random from a concept class X, and a learning algorithm tries to uncover x by seeing streaming of random samples $(a_i, b_i = M(a_i, x))$ where for every i, $a_i \in A$ is chosen uniformly at random. Assume that k, ℓ, r are integers such that any submatrix of M of at least $2^{-k} \cdot |A|$ rows and at least $2^{-\ell} \cdot |X|$ columns, has a bias of at most 2^{-r}. We prove that any algorithm with classical and quantum hybrid memory for the learning problem corresponding to M needs either (1) $\Omega(k \cdot \ell)$ bits of classical memory, or (2) $\Omega(r)$ qubits of quantum memory, or (3) $2^{\Omega(r)}$ random samples, to achieve a success probability at least $2^{-O(r)}$.

Our results refute the possibility that a small amount of quantum memory significantly reduces the size of classical memory needed for efficient learning on these problems. Our results also imply improved security of several existing cryptographical protocols in the bounded-storage model (protocols that are based on parity learning on n bits), proving that security holds even in the presence of a quantum adversary with at most cn^2 bits of classical memory and cn bits of quantum memory (for some constant $c > 0$).

CCS CONCEPTS

• **Theory of computation → Quantum complexity theory**.

KEYWORDS

Learning parity, Quantum lower bounds, Time-space lower bounds

ACM Reference Format:
Qipeng Liu, Ran Raz, and Wei Zhan. 2023. Memory-Sample Lower Bounds for Learning with Classical-Quantum Hybrid Memory. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing (STOC '23), June 20–23, 2023, Orlando, FL, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3564246.3585129

1 INTRODUCTION

Memory plays an important role in learning. Starting from the seminal works by Shamir [39] and Steinhardt, Valiant and Wager [41], a sequence of works initiates and deepens the study of lower bounds for learning under memory constraints. Steinhardt, Valiant, and Wager [41] conjectured that in order to learn an unknown n-bit string from samples of random-subset parity, an algorithm needs either memory-size quadratic in n or exponentially many random samples (also in n). This conjecture was later on proved by Raz [37], showing for the first time that for some learning problems, super-linear memory size is required for efficient learning. This result was then generalized to a broad class of learning problems [6, 21–23, 29, 34, 36, 40].

Although we have a comprehensive understanding of the (in)-feasibility of learning under limitations on particular computation resource (memory), the previous works mentioned above do not

*Qipeng Liu was supported in part by the Simons Institute for the Theory of Computing, through a Quantum Postdoctoral Fellowship, by the DARPA SIEVE-VESPA grant No.HR00112020023 and by the NSF QLCI program through grant number OMA-2016245. Ran Raz, and Wei Zhan were supported by a Simons Investigator Award and by the National Science Foundation grants No. CCF-1714779, CCF-2007462.

STOC '23, June 20–23, 2023, Orlando, FL, USA

ACM ISBN 978-1-4503-9913-5/23/06…$15.00
https://doi.org/10.1145/3564246.3585129

capture all physical computational models; most notably, quantum computation and the power of quantum memory. Many researchers believe that large-scale quantum computers will eventually become viable. Recent experiments demonstrated quantum advantages, for example [3], and suggested that there are possibly no fundamental barriers to achieving quantum memory and quantum computers. Questions on the role of quantum memory in learning were proposed by Wright in the context of general state tomography [43] and by Aaronson for shadow tomography [1]. A line of works [2, 8, 11, 12, 26, 28] pioneer the idea and show either polynomial or exponential separations for learning with/without quantum memory, but all for intrinsic quantum learning tasks like state tomography, shadow tomography and purity testing.

In light of the above, it is appealing to consider classical learning tasks in the presence of quantum memory, as well as hybrid classical-quantum memory. A direct implication of all aforementioned classical results only gives trivial results. As k qubits of memory can always be efficiently simulated by $\sim 2^k$ classical bits, we can only conclude (say, for parity learning) that either $\sim 2\log n$-qubit quantum memory or exponentially many samples are needed. Prior to our work, it could have been the case that even if only a very small size quantum memory was available, it might have significantly reduced the need for classical memory and led to an efficient learning algorithm.

In this work, we prove memory-sample lower bounds in the presence of hybrid memory for a wide collection of classical learning problems. As in [23, 36], we will represent a learning problem by a matrix $M : A \times X \to \{-1, 1\}$ whose columns correspond to concepts in the concept class X and rows correspond to random samples. In the learning task, an unknown concept $x \in X$ is sampled uniformly at random and each random sample is given as $(a_i, b_i) = (a_i, M(a_i, x))$ for a uniformly picked $a_i \in A$. The learner's goal is to uncover x. In [23], it is proved that when the underlying matrix M is a (k, ℓ)-L_2 two source extractor[1] with error 2^{-r}, a learning algorithm requires either $\Omega(k \cdot \ell)$ bits of memory or $2^{\Omega(r)}$ samples to achieve a success probability at least $2^{-O(r)}$ for the learning task.

1.1 Our Results

In this work, we model a quantum learning algorithm as a program with hybrid memory consisting of q qubits of quantum memory and m bits of classical memory. At each stage, a random sample $(a_i, b_i = M(a_i, x))$ is given to the algorithm. The quantum learning algorithm applies an arbitrary quantum channel to the hybrid memory, controlled by the random sample. Although the channel can be arbitrary, we impose the outcome to be a hybrid classical-quantum state of at most q qubits and m bits. We stress that there is no limitation on the complexity of the quantum channel (and this only makes our results stronger as we are proving here lower bounds for such algorithms).

With the above model, we give the following main theorem.

Theorem 1 (Main Theorem, Informal). *Let $M : A \times X \to \{-1, 1\}$ be a matrix. If M is a (k, ℓ)-L_2 two source extractor with error 2^{-r}, a quantum learning algorithm requires either*

(1) $\Omega(k \cdot \ell)$ bits of classical memory; or,
(2) $\Omega(r)$ qubits of quantum memory; or,
(3) $2^{\Omega(r)}$ samples,

to succeed with a probability of at least $2^{-O(r)}$ in the corresponding learning task.

Our main theorem implies that for many learning problems, the availability of a quantum memory of size up to $\Omega(r)$, does not reduce the size of classical memory or the number of samples that are needed. As coherent quantum memory is challenging for near-term intermediate-scale quantum computers and is probably expensive even if and when quantum computers are widely viable, the impact of quantum memory is further limited for these learning problems.

To make the theorem more precise, let us take parity learning as an example. The above theorem says that a quantum learning algorithm needs either $\Omega(n^2)$ bits of memory, or $\Omega(n)$ qubits of quantum memory, to conduct efficient learning; otherwise, it requires $2^{\Omega(n)}$ random samples. At first glance, it seems that the constraint on quantum memory is trivial: if the target is to learn an n-bit unknown secret, a linear amount of memory always seems necessary to store the secret. However, noticing that our main theorem applies to quantum learning algorithms with hybrid memory and rules out algorithms with $n^2/1000$ bits and $n/1000$ qubits of hybrid memory for parity learning, the main theorem yields non-trivial and compelling memory-sample lower bounds. Note also that our results (and previous results) are valid even if the goal is to output only one bit of the secret. Currently, we do not know whether our main theorem is tight. For parity learning, we are not aware of any quantum learning algorithm that uses only $O(n)$ qubits of quantum memory. We leave closing the gap as a fascinating open question.

The main theorem naturally applies to other learning problems considered in [23], including learning sparse parities, learning from sparse linear equations, and many others. We do not present an exhaustive list here but refer the readers to [23] for more details.

Along the way, we propose a new approach for proving the classical memory-sample lower bounds. We call this approach, the "badness levels" method. The approach is technically equivalent to the previous approach in [23, 36] but is conceptually simpler to work with and we are able to lift it to the quantum case.

We note that proving a linear lower bound on the size of the quantum memory, without classical memory, is significantly simpler (but to the best of our knowledge such a proof has not appeared prior to our work). We present such a proof in the Appendix of the full version of this paper.

Implications to Cryptography in the Bounded-Storage Model. Since learning theory and cryptography can be viewed as two sides of the same coin, our theorem also lifts the security of many existing cryptographical protocols in the bounded-storage model (protocols that are based on parity learning) to the quantum setting. To our best knowledge, these are the first proofs of classical cryptographical protocols being secure against space-bounded quantum attackers.[2] We elaborate more below.

[1] Roughly speaking, this means that every submatrix M' of M with number of rows at least $2^{-k}|A|$ and number of columns at least $2^{-\ell}|X|$ has a relative bias at most 2^{-r}.

[2] On the other hand, there are known examples of classically-secure bounded-storage protocols that are breakable with an exponentially smaller amount of quantum memory. [24].

Cryptography in the (classical) bounded storage model was first proposed by Maurer [32]. In such a model, no computational assumption is needed. Honest execution is performed through a long streaming sequence of bits. Eavesdroppers have bounded storage and limited capability of storing conversations, thus cannot break the protocol. A line of works [4, 5, 9, 16–20, 27, 31, 33, ...] builds efficient and secure protocols for key agreement, oblivious transfer, bit commitment and time stamping in that model.

Based on the memory-sample lower bounds for parity learning of n bits, [37] suggested an encryption scheme in the bounded-storage model. Guan and Zhandry [25] proposed key agreement, oblivious transfer and bit commitment with improved rounds and better correctness, against attackers with up to $O(n^2)$ bits of memory. Following a similar idea, Liu and Vusirikala [30] showed that semi-honest multiparty computation could be achieved against attackers with up to $O(n^2)$ bits of memory. More recently, Dodis, Quach, and Wichs [18] considered message authentication in the bounded storage model based on parity learning. Our result on parity learning gives a direct lift on all the results above. When the cryptographic protocols are based on parity learning of n bits (often treated as a security parameter), our result shows that security holds even in the presence of a quantum adversary with at most $O(n^2)$ bits of classical memory and $O(n)$ qubits of quantum memory.

Despite many previous works on cryptography in the quantum bounded storage model [7, 13–15, 35, 38, 42], they all rely on streaming quantum states. Our memory-sample lower bounds give for the first time a rich class of classical cryptographical schemes (key agreement, oblivious transfer, and bit commitment) secure against space-bounded quantum attackers.

2 PROOF OVERVIEW

2.1 Recap of Proofs for Classical Lower Bounds

Since our proof builds on the previous line of works on classical memory-sample lower bounds for learning, specifically, on the proof technique of [23, 36], we provide a brief review of these proofs, using parity learning [37] as an example. In below, $M(a,x)$ denotes the inner product of a and x in $\mathbb{F}_2$.

Consider a classical branching program that tries to learn an unknown and uniformly random $x \in \{0,1\}^n$ from samples (a,b), where $a \in \{0,1\}^n$ is uniformly random and $b = M(a,x)$. We can associate every state v of the branching program with a distribution $P_{X|v}$ over $\{0,1\}^n$, indicating the distribution of x conditioned on reaching that state. At the initial state, without any information about x, the distribution is uniform (which has the smallest possible ℓ_2-norm). Along a computational path on the branching program, the distribution $P_{X|v}$ evolves and eventually gets concentrated (with large ℓ_2-norms) in order to output x correctly. Therefore, during the evolution, $P_{X|v}$ should at some stage have mildly large ℓ_2-norms ($2^{\varepsilon n}$ times larger than uniform for some small constant $\varepsilon > 0$). If we set such a distribution as a target, the distribution is hard to achieve with random samples. Only with $2^{-\Omega(n)}$ probability, the branching program can make significant progress towards the target; while most of the time a sample just splits the distributions (both the current and the target distribution) into two even parts, and that does not help much in getting closer to the target distribution (with large ℓ_2 norm).

To put it more rigorously, we examine the evolution of the inner product

$$\langle P_{X|v}, P\rangle = \sum_{x\in\{0,1\}^n} P_{X|v}(x)\cdot P(x)$$

between the distribution $P_{X|v}$ on the current state v, and a target distribution P. Receiving a sample (a,b) implies that $M(a,x)=b$, hence only the part of $P_{X|v}$ supported on such x proceeds. If this part is close to $\frac{1}{2}$ probability, we say that a divides $P_{X|v}$ evenly. Denoting the new distribution as $P^{(a,b)}_{X|v}$, after proper normalization the new inner product is

$$\langle P^{(a,b)}_{X|v}, P\rangle = \sum_{\substack{x\in\{0,1\}^n\\ M(a,x)=b}} P_{X|v}(x)\cdot P(x) \Big/ \sum_{\substack{x\in\{0,1\}^n\\ M(a,x)=b}} P_{X|v}(x). \quad (1)$$

Ideally, both $P_{X|v}$ and the point-wise product vector $P_{X|v}\cdot P$ should have reasonably small ℓ_2-norms. Due to the extractor property of M, most of $a \in \{0,1\}^n$ should divide both vectors evenly, and thus the denominator is close to $\frac{1}{2}$ while the enumerator is close to $\frac{1}{2}\langle P_{X|v}, P\rangle$. That means, given a uniformly random a, we get limited progress on the inner product. On the other hand, from $\langle U, P\rangle = 2^{-n}$ with uniform distribution U to $\langle P, P\rangle = 2^{2\varepsilon n}\cdot 2^{-n}$, the branching program needs to make multiple steps of progression. Therefore it happens with an extremely small probability.

To ensure that the above statement goes smoothly, we require the following properties for every state v in the branching program:

- The ℓ_2-norm $\left\|P_{X|v}\right\|_2$ is small.
- The ℓ_2-norm $\left\|P_{X|v}\cdot P\right\|_2$ is small, which is implied when the ℓ_∞-norm $\left\|P_{X|v}\right\|_\infty$ is small.
- The denominator in Eq. (1) is bounded away from 0 for every sample (a,b).

These properties do not hold by themselves. Instead, we execute a *truncation* procedure on the branching program *before* choosing a target distribution. More specifically, the branching program is modified so that it stops whenever it:

- (ℓ_2 truncation): Reaches a state v with large $\left\|P_{X|v}\right\|_2$;
- (ℓ_∞ truncation): Reaches a state v with large $P_{X|v}(x)$ when the unknown concept is x;
- (Sample truncation): Or, for the next sample (a,b), a does not divide $P_{X|v}$ evenly.

It turns out that after ℓ_2 truncation, the other two truncation steps add $2^{-\Omega(n)}$ error in each stage of the branching program. Therefore the proof boils down to proving a $2^{-\Omega(n^2)}$ bound on the probability of reaching a state with large $\left\|P_{X|v}\right\|_2$, from which by a standard union bound, we can prove the memory-sample lower bounds for parity learning: either $2^{\Omega(n)}$ samples or $\Omega(n^2)$ bits of memory are necessary.

2.2 Badness Levels

As mentioned above, to bound the probability of reaching a state with a large ℓ_2-norm, the basic idea is to fix its distribution as the target distribution P, and bound the increment of the inner product $\langle P_{X|v}, P\rangle$. This was done in [23, 36] by designing a potential

function that tracks the average of $\langle P_{X|v}, P\rangle^k$ for some $k = \Theta(n)$, where the average is over states v in the same stage of the branching program. Here we propose another approach using the concept of *badness levels*. Although it is technically equivalent to the potential function approach in the classical case, it is more pliable and easier to be adapted to the quantum case. We view this approach as a separate contribution of our work.

We first define a *bad event* to be a pair (v, a) of the state v and the upcoming part of the sample a, such that $\langle P_{X|v}, P\rangle \geq 2^{-n}$, and for one of the two possible outcomes b,

$$\sum_{\substack{x\in\{0,1\}^n \\ M(a,x)=b}} P_{X|v}(x)\cdot P(x) \geq \left(\frac{1}{2}+2^{-\delta n}\right)\cdot\langle P_{X|v}, P\rangle \qquad (2)$$

with some small constant δ. In other words, the inner product $\langle P_{X|v}, P\rangle$ is large enough, while not being divided evenly by a. From Eq. (1) we know that the inner product gets at most roughly doubled through a bad event. In contrast, in the good case, the inner product either gets a mere $(1 + 2^{-\delta n})$ multiplicative factor or is already smaller than the baseline 2^{-n}. Also, the extractor property of M ensures that for every state v, over uniformly random a, the bad event happens with at most $2^{-\Omega(n)}$ probability.

Now, the badness level $\beta(v)$ of a state v keeps track of how many times the computational path went through bad events before reaching v.[3] The above observations on the bad events imply that (omitting the smaller factors):

- For every state v, $\langle P_{X|v}, P\rangle$ is bounded by $2^{\beta(v)}\cdot 2^{-n}$;
- Heading to the next stage, $\beta(v)$ increases by 1 with probability $2^{-\Omega(n)}$.

Therefore at each stage, the total weight of states with badness level β is at most $2^{-\Omega(\beta n)}$. Thus any state with $\langle P_{X|v}, P\rangle \geq 2^{2\varepsilon n}\cdot 2^{-n}$ must have $2^{-\Omega(n^2)}$ probability.

2.3 Obstacles for Proving Quantum Lower Bounds

In this section, we present an attempt to prove the same $2^{\Omega(n)}$-sample or $\Omega(n^2)$-quantum-memory lower bound for the pure quantum case. Along the way we identify some obstacles to proving memory-sample lower bounds for *quantum* learning algorithms, and in the next section we show how to overcome these obstacles while proving lower bounds for *hybrid* learning algorithms, with quadratic-size classical-memory and linear-size quantum-memory.

Following the same framework as the above described proof for the classical case, we first need to transfer all the notions to a quantum algorithms:

- The state v is a quantum state in the Hilbert space of quantum memory;
- The distribution $P_{X|v}$ is still well-defined: It is the distribution of x when the quantum memory is measured to v (see Section 3.4 and Eq. (3));
- We are still able to implement ℓ_2 truncation: If $P_{X|v}$ has large ℓ_2-norm, project the entire system to the orthogonal subspace $v^\perp$ of v and repeat, until there is no such state v (see Section 4.1 for details).
- We are also able to implement sample truncation, in a similar manner to ℓ_2 truncation. As the criteria here depends on a, we separately create a copy of the current system for each a, truncate the states v using projection when $P_{X|v}$ is not evenly divided by a in each copy, and then merge them back together. We prove that the error introduced by this truncation is small.

Here comes the first major obstacle: ℓ_∞ truncation. In the classical case, ℓ_∞ truncation is implemented for each individual x, in contrast to ℓ_2 truncation where the states are removed altogether. Relying on the fact that it is already known that the ℓ_2 norm of the distribution is small, using Markov inequality, one can prove that the error introduced by the ℓ_∞ truncation is small.

However, when we try to emulate the classical implementation of ℓ_∞ truncation with quantum truncation, that is, to only project to $v^\perp$ the system *conditioned on* the specific x where $P_{X|v}(x)$ is large, instead of for every x, it may lead to huge changes to the distributions $P_{X|u}$ on states u non-orthogonal to v. The following example illustrates such a scenario:

Example. Consider a quantum learning algorithm, and assume that at some stage of the computation, for each $x \in \{0,1\}^n$, the quantum memory is in some pure state $v(x)$. We pick each $v(x)$ uniformly at random in a Hilbert space of dimension $d \approx 2^{n/2}$ and consider a typical configuration of $v(x)$. Now the ℓ_2-norms are bounded for every quantum state v: the worst ones happen when $v = v(x)$ for some x, where $\|P_{X|v(x)}\|_2$ is typically around $d\cdot 2^{-n}$, close to the ℓ_2-norm of uniform distribution. However, those worst distributions also have ℓ_∞-norms close to $d\cdot 2^{-n}$, which is much larger than the ℓ_∞-norm of the uniform distribution, and needs to be truncated. But truncating $v(x)$ off for x means that x is completely erased, and we end up removing everything.

Moving on, we fix a target state v with a target distribution $P_{X|v}$ which exceeds the ℓ_2-norm threshold, and the goal is again to prove a $2^{-\Omega(n^2)}$ amplitude bound on v. The bad event should still be defined as a pair (v, a) satisfying Eq. (2), with v now being a quantum state. We then run into the second major obstacle: it is not clear how to define badness levels.

If we define the badness level $\beta(v)$ for each state v individually by examining the bad events over the historical states, then it is not clear how to measure the total weight of a badness level β. In the classical case, we simply define the total weight as the total probability of states with badness level β. But here in the quantum case, it turns out that such a definition either depends on the choice of basis, which might have large increment in each stage, or completely fails to imply the desired amplitude bound on the target state.

The other choice is to have a more *operational* definition of badness levels, and it is indeed tempting to define β as another register whose updates are controlled by the quantum memory. The problem with such definitions is that the bad event (Eq. (2)) is not linear in v. Therefore an operational definition of badness level, which is a linear operator, inevitably introduces error that escalates fast with the number of stages.

[3] For now we think of $\beta(v)$ as a natural number. In the actual proof, $\beta(v)$ is a distribution on natural numbers, as for different computational paths reaching the same state, the count of bad events can be different.

2.4 Hybrid Memory Lower Bounds with Small Quantum Memory

The obstacles in the previous section are for proving quadratic quantum memory lower bound. We note that proving linear quantum memory lower bound (without classical memory) is not hard: the proof can be entirely information theoretical, as with very limited memory, say, $\frac{1}{2}n$ qubits, the information gained from each sample is exponentially small, despite the memory being quantum. We present such a proof in the Appendix of the full version of this paper.

The lower bounds that we prove here are with hybrid memory: To learn parity with both classical and quantum memory, an algorithm needs either $2^{\Omega(n)}$ samples, or $\Omega(n^2)$ classical memory, or $\Omega(n)$ quantum memory (Theorem 1). We now describe how we overcome the previously mentioned obstacles.

ℓ_∞ Truncation. When there is only small quantum memory and no classical memory, the treatment for ℓ_∞ truncation is straightforward. We remove all quantum states v with distributions of large ℓ_∞-norm, by projecting the system to the orthogonal subspace $v^\perp$, just like the process of ℓ_2 truncation. As the overall distribution on x is uniform, any state v with $\|P_{X|v}\|_\infty \geq 2^{\delta n} \cdot 2^{-n}$ must have weight at most $2^{-\delta n}$. Therefore, as long as the dimension of the Hilbert space is much smaller than δn, the error introduced in this truncation is small. [4]

With classical memory in presence, the actual ℓ_∞ truncation step (see Section 4.2, Step 2) is more complicated. We first apply the original classical ℓ_∞ truncation on the classical memory W. Now that $\|P_{X|w}\|_\infty$ is bounded for each classical memory state w, we can remove the quantum states v with large $\|P_{X|v,w}\|_\infty$ by projection as stated above. Since the classical ℓ_∞ truncation depends on x, it could change the distributions $P_{X|v,w}$. However, as in the classical case, $P_{X|w}$ will not change a lot. Thus, wherever $P_{X|v,w}$ changes drastically, it must have a small weight and can also be removed by projection. This removal corresponds to truncation by G_t in Section 4.2.

Badness Levels. Interestingly, we are able to avoid the problems of defining the badness level on quantum memory altogether, by keeping it a property on the classical memory only. To do so we need to alter the definition of a bad event: it is now a pair (w, a) of classical memory state w and sample a, such that *there exists* some quantum memory state v with $P_{X|v,w}$ satisfying Eq. (2).

For each fixed classical memory state w, we still need to ensure that bad events happen with a small probability. We prove it (Lemma 5.2) by showing that, if there are many different samples a, each associated with some quantum state v_a satisfying Eq. (2), then there is some quantum state v that simultaneously satisfies Eq. (2) with most of such a (which is impossible because of the extractor property). This is ultimately due to the continuous nature of Eq. (2): Under some proper congruent transformation, Eq. (2) becomes a simple threshold inequality on quadratic forms over v. Now if it is satisfied by some v_a, it is going to be satisfied by most v for a much smaller threshold parameter δ, and hence the existence of a simultaneously satisfying v.[5] In this argument, we use Lemma 3.1, which is derived from the anti-concentration bound for Gaussian quadratic forms, and crucially relies on the fact that the dimension is at most $2^{\varepsilon n}$ for some small ε.

Another technical problem is that to use the extractor property, we need to ensure that $\langle P_{X|v,w}, P\rangle \geq 2^{-n}$ for the simultaneously satisfying v. Thus, what we do in Lemma 5.2 is to first conceptually remove the parts where $\langle P_{X|v,w}, P\rangle$ is too small, using projection similarly to the truncation steps. After the removal, we are left with a subspace $\mathcal{V}'$ where $\langle P_{X|v,w}, P\rangle$ is always lower bounded, and we show that for every state v that satisfies Eq. (2), the inequality is still close to being satisfied after projecting v onto $\mathcal{V}'$. Therefore we could still apply the above argument and find a simultaneously satisfying v within the subspace.

3 PRELIMINARIES

3.1 Vectors and Matrices

For a vector $v \in \mathbb{C}^d$ and $p \in [1, \infty]$, we define the ℓ_p norm of v as

$$\|v\|_p = \left(\sum_{i=1}^{d} |v_i|^p\right)^{1/p}.$$

For two vectors $u, v \in \mathbb{C}^d$, define their inner product as $\langle u, v\rangle = u^\dagger v = \sum_{i=1}^{d} \overline{u_i} v_i$. So $\|v\|_2^2 = \langle v, v\rangle$. We also view every distribution P over a set $\mathcal{X}$ as a non-negative real vector with $\|P\|_1 = 1$.

We specifically use Dirac notation to denote unit vectors, $|v\rangle \in \mathbb{C}^d$ implies that $\||v\rangle\|_2 = 1$. For a non-zero vector $u \in \mathbb{C}^d$, let $|v\rangle \sim u$ be the normalization of u, that is, $|v\rangle = u/\|u\|_2$.

For every vector $v \in \mathbb{C}^d$, let $\operatorname{Diag} v \in \mathbb{C}^{d\times d}$ be the diagonal matrix whose diagonal entries represent v. Conversely, for every square matrix M, let $\operatorname{diag} M$ be the vector consisting of the diagonal entries of M. For a matrix (or generally a linear operator) M, we use $\|M\|_{\mathrm{Tr}}$ and $\|M\|_2$ to denote its trace norm and spectral norm respectively, that is,

$$\|M\|_{\mathrm{Tr}} = \operatorname{Tr}\left[\sqrt{MM^\dagger}\right], \quad \|M\|_2 = \max_{v\neq 0} \|Mv\|_2/\|v\|_2.$$

For an Hermitian $M \in \mathbb{C}^{d\times d}$, we say it is a positive semi-definite operator if for every $v \in \mathbb{C}^d$, $v^\dagger M v \geq 0$. A (partial) density operator is a positive semi-definite operator with its trace being 1 (or at most 1, respectively).

Viewing a Learning Problem as a Matrix. Let $M : \mathcal{A} \times \mathcal{X} \to \{-1, 1\}$ be a matrix. The matrix M corresponds to the following learning problem. There is an unknown element $x \in \mathcal{X}$ that was chosen uniformly at random. A learner tries to learn x from samples (a, b), where $a \in \mathcal{A}$ is chosen uniformly at random and $b = M(a, x)$. That is, the learning algorithm is given a stream of samples, $(a_1, b_1), (a_2, b_2), \ldots$, where each a_t is uniformly distributed and for every t, $b_t = M(a_t, x)$. For each $a \in \mathcal{A}$, we use $M_a : \mathcal{X} \to \{-1, 1\}$ to denote the vector corresponding to the a-th row of M.

Extractors. A matrix $M : \mathcal{A} \times \mathcal{X} \to \{-1, 1\}$ with $n = \log_2 |\mathcal{X}|$ is a (k, ℓ)-L_2 extractor with error 2^{-r}, if for every distribution P over

[4] The example in the previous section that shows the infeasibility of treating ℓ_∞ truncation the same way as ℓ_2 truncation does not work here, as it requires $n/2$ qubits of memory while here we have a smaller memory size.

[5] We note that the error bound for sample truncation (Lemma 4.11) is also proved using this argument.

$\mathcal{X}$ with $\|P\|_2 \le 2^{\ell} \cdot 2^{-n/2}$, there are at most $2^{-k} \cdot |\mathcal{A}|$ rows $a \in \mathcal{A}$ such that

$$|\langle M_a, P\rangle| \ge 2^{-r}.$$

3.2 Anti-Concentration Bound for Quadratic Form on Unit Vectors

Lemma 3.1. *There exists an absolute constant c such that following holds. Let σ be a Hermitian operator over the Hilbert space $\mathcal{V} = \mathbb{C}^d$, and let v be a uniformly random unit vector in $\mathcal{V}$. Then for every $\varepsilon > 0$, we have*

$$\Pr\left[|v^{\dagger}\sigma v| \le \frac{\varepsilon\|\sigma\|_2}{d}\right] \le c\sqrt{\varepsilon} + e^{-d}.$$

The proof is based on Carbery–Wright inequality [10] and can be found in the full version of this paper.

3.3 Multipartite Quantum Systems

The state of q qubits can be represented in a Hilbert space $\mathcal{V} = (\mathbb{C}^2)^{\otimes q} = \mathbb{C}^{2^q}$. In a product of m Hilbert spaces $\mathcal{V}_{[m]} = \mathcal{V}_1 \otimes \cdots \otimes \mathcal{V}_m$, a multipartite partial system $V_1, \ldots, V_m$ is represented by a partial density operator $\rho_{V_{[m]}}$. For a subset $I \subseteq [m]$ of indices, the subsystem on $\{V_i\}_{i\in I}$ (or V_I for short) is defined by tracing out $j \notin I$, that is,

$$\rho_{V_I} = \mathrm{Tr}_{V_{j\notin I}}[\rho_{V_{[m]}}].$$

Now for any two disjoint subsets $I, J \subset [m]$, given some $|v_J\rangle \in \mathcal{V}_J = \bigotimes_{j\in J}\mathcal{V}_j$, the conditional system on V_I is defined as

$$\rho_{V_I|v_J} = \left(\mathbb{I}_{V_I} \otimes \langle v_J|\right)\rho_{V_{I\cup J}}\left(\mathbb{I}_{V_I} \otimes |v_J\rangle\right),$$

which is a partial density operator on V_I. Note that the trace

$$\mathrm{Tr}\left[\rho_{V_I|v_J}\right] = \langle v_J|\rho_{V_J}|v_J\rangle$$

only depends on the system ρ and $|v_J\rangle$, while being *independent* of the choice of I.

Another simple fact that will be repeatedly used later on is that for an *orthogonal basis* $\mathcal{B}$ of V_J, we have

$$\rho_{V_I} = \mathrm{Tr}_{V_J}[\rho_{V_{I\cup J}}] = \sum_{|v_J\rangle\in\mathcal{B}} \rho_{V_I|v_J}.$$

3.4 Classical-Quantum Systems

In the underlying space $\mathcal{V}_1 \otimes \cdots \otimes \mathcal{V}_m$ of the multipartite system, we say $\mathcal{V}_i$ is classical if there is a fixed orthogonal basis $\mathcal{B}_i$ of $\mathcal{V}_i$, such that for every multipartite system $\rho_{V_{[m]}}$, every pair of distinct $|v_i\rangle \ne |v_i'\rangle \in \mathcal{B}_i$ and every two states $|v\rangle, |v'\rangle \in \bigotimes_{j\ne i}\mathcal{V}_j$, we have

$$\langle v_i, v|\rho_{V_{[m]}}|v_i', v'\rangle = 0.$$

Without loss of generality, in the rest of the work we always assume $\mathcal{B}_i$ is the set of computational basis states. We also identify $\mathcal{V}_i$ with the discrete set $\mathcal{B}_i$, and remove the Dirac brackets when we talk about the classical elements in $\mathcal{V}_i$. In this case every multipartite system $\rho_{V_{[m]}}$ can be written as a direct sum

$$\rho_{V_{[m]}} = \bigoplus_{v_i\in\mathcal{V}_i} \rho_{V_{[m]\setminus\{i\}}|v_i}.$$

The reader may find this direct sum viewpoint easier to handle in some later scenarios.

When V_I is classical, conditioned on any $|v_J\rangle \in \mathcal{V}_J$ with J disjoint from I, the system $\rho_{V_I|v_J}$ is represented as a diagonal matrix on V_I. If $\mathrm{Tr}[\rho_{V_I|v_J}] > 0$, it induces a distribution over the computation basis states of V_I, defined as

$$P^{\rho}_{V_I|v_J} = \mathrm{diag}\,\rho_{V_I|v_J}/\mathrm{Tr}[\rho_{V_I|v_J}]. \tag{3}$$

In the rest of this paper, whenever we use this notation $P^{\rho}_{V_I|v_J}$, it is always implicitly assumed that $\mathrm{Tr}[\rho_{V_I|v_J}] > 0$ and the distribution exists.

In this work we typically consider the following scenario: There is a quantum memory register V ranging in the Hilbert space $\mathcal{V}$, and a classical memory register W ranging in the set of memory states $\mathcal{W}$, along with some classical information $X \in \mathcal{X}$ (later in the work, it is the concept to be learned) that is correlated with V and W. We will make use of the following fact:

Claim 3.2. *Let ρ_{XVW} be a classical-quantum system over classical X, W and quantum V. For every $w \in \mathcal{W}$, $P^{\rho}_{X|w}$ is a convex combination of $P^{\rho}_{X|v,w}$ for some $\{|v\rangle\} \subseteq \mathcal{V}$.*

Proof. Let $\mathcal{B}$ be an orthogonal basis of $\mathcal{V}$, so that we have (from the end of last section)

$$\rho_{X|w} = \sum_{|v\rangle\in\mathcal{B}} \rho_{X|v,w}.$$

Therefore $P^{\rho}_{X|w}$ is a linear combination of $P^{\rho}_{X|v,w}$ for $|v\rangle \in \mathcal{B}$, with non-negative coefficients. Since they are all distributions, it must be a convex combination. □

Characterization of operators over classical-quantum hybrid systems. We identify all possible operators on the classical-quantum hybrid memory space $\mathcal{V} \otimes \mathcal{W}$. A priori to the assumption that W is classical, we think of a quantum channel operating on the system as working on the underlying space $\mathcal{V} \otimes \mathbb{C}^{|\mathcal{W}|}$. Now we denote $\mathscr{T}_{\mathcal{V}\otimes\mathcal{W}}$ to be the set of all such quantum channels Φ that satisfy the following: for every classical-quantum system ρ_{VW} in $\mathcal{V} \otimes \mathcal{W}$, W is still classical in $\Phi(\rho_{VW})$. That is, for every two states $|v\rangle, |v'\rangle \in \mathcal{V}$ and every pair of distinct $w, w' \in \mathcal{W}$, we have

$$\langle v, w|\Phi(\rho_{VW})|v', w'\rangle = 0.$$

Note that not all channels in $\mathscr{T}_{\mathcal{V}\otimes\mathcal{W}}$ are physically realizable. For instance, with one-bit classical memory and no quantum memory, the channel

$$\begin{pmatrix} a & c \\ \bar{c} & b \end{pmatrix} \mapsto \begin{pmatrix} a & ic \\ -i\bar{c} & b \end{pmatrix}$$

is not a classical operator. However, since we are constrained to classical quantum systems, this channel is effectively equivalent to an identity channel on one-bit classical memory. Generally speaking, every channel in $\mathscr{T}_{\mathcal{V}\otimes\mathcal{W}}$ is equivalent to a channel controlled by $\mathcal{W}$ that maps $\mathcal{V}$ to $\mathcal{V} \otimes \mathcal{W}$. This observation and the following claim are proved in the full version of this paper:

Claim 3.3. *Let ρ_{XVW} be a classical-quantum system over classical X, W and quantum V. Let $\Phi \in \mathscr{T}_{\mathcal{V}\otimes\mathcal{W}}$, and we use $\Phi(\rho)$ to denote the system after applying Φ to VW and identity to X. Then for every*

$|v\rangle \in \mathcal{V}$ and $w \in \mathcal{W}$, $P^{\Phi(\rho)}_{X|v,w}$ is a convex combination of $P^{\rho}_{X|v',w'}$ for some $\{|v'\rangle\} \subseteq \mathcal{V}$ and $\{w'\} \subseteq \mathcal{W}$.

3.5 Branching Program with Hybrid Memory

For a learning problem that corresponds to the matrix M, a branching program of hybrid memory with m-bit classical memory, q-qubit quantum memory and length T is specified as follows.

At each stage $0 \le t \le T$, the memory state of the branching program is described as a classical-quantum system $\rho^{(t)}_{VW}$ over quantum memory space $\mathcal{V} = (\mathbb{C}^2)^{\otimes q}$ and classical memory space $\mathcal{W} = \{0,1\}^m$. The memory state evolves based on the samples that the branching program receives, and therefore depends on the unknown element $x \in_R \mathcal{X}$. We can then interpret the overall systems over XVW, in which X consists of an unknown concept x, resulting in a classical-quantum system $\rho^{(t)}_{XVW}$. It always holds that the distribution of x is uniform, i.e.,

$$\rho^{(t)}_X = \mathrm{Tr}_{VW}[\rho^{(t)}_{XVW}] = \frac{1}{2^n}\mathbb{I}_X.$$

Initially the memory VW is independent of X and can be arbitrarily initialized. We assume that

$$\rho^{(0)}_{XVW} = \frac{1}{2^n}\mathbb{I}_X \otimes \frac{1}{2^q}\mathbb{I}_V \otimes \frac{1}{2^m}\mathbb{I}_W.$$

At each stage $0 \le t < T$, the branching program receives a sample (a,b), where $a \in_R \mathcal{A}$ and $b = M(a,x)$, and applies an operation $\Phi_{t,a,b} \in \mathscr{T}_{\mathcal{V}\otimes\mathcal{W}}$ over its memory state. Thus the evolution of the entire system can be written as

$$\rho^{(t+1)}_{XVW} = \mathop{\mathbf{E}}_{a\in_R\mathcal{A}}\left[\sum_{x\in\mathcal{X}} |x\rangle\langle x| \otimes \Phi_{t,a,M(a,x)}\big(\rho^{(t)}_{VW|x}\big)\right].$$

Finally, at stage $t = T$, a measurement over the computational bases is applied on $\rho^{(T)}_{VW}$, and the branching program outputs an element $\widetilde{x} \in X$ as a function of the measurement result $(v,w) \in \{0,1\}^{q+m}$. The success probability of the program is the probability that $\widetilde{x} = x$ which can be formulated as

$$\sum_{\substack{x\in\mathcal{X}, v\in\{0,1\}^q, w\in\mathcal{W}\\ \widetilde{x}(v,w)=x}} \langle x,v,w|\rho^{(T)}_{XVW}|x,v,w\rangle.$$

4 MAIN RESULT

Theorem 2. *Let $\mathcal{X}, \mathcal{A}$ be two finite sets with $n = \log_2|\mathcal{X}|$. Let $M : \mathcal{A}\times\mathcal{X} \to \{-1,1\}$ be a matrix which is a (k',ℓ')-L_2 extractor with error $2^{-r'}$ for sufficiently large k', ℓ' and r', where $\ell' \le n$. Let*

$$r = \min\left\{\frac{1}{4}r', \frac{1}{26}\ell' + \frac{1}{6}, \frac{1}{2}(k'-1)\right\}.$$

Let ρ be a branching program for the learning problem corresponding to M, described by classical-quantum systems $\rho^{(t)}_{XVW}$, with q-qubit quantum memory V, m-bit classical memory W and length T. If $m \le \frac{1}{44}(k'-1)\ell'$, $q \le r-7$ and $T \le 2^{r-2}$, the success probability of ρ is at most $O(2^{q-r})$.

From now on we let $k = k'-1$ and $\ell = \frac{1}{5}(\ell' - 13r - 2)$. Then we have the following inequalities to be used later:

$$q + r + 1 - r' \le -2r. \tag{4}$$

$$2\ell + 9r - n \le -r. \tag{5}$$

$$(k-r)\ell \ge 2m + 4r + 1. \tag{6}$$

4.1 Truncated Classical-Quantum Systems

Here we describe how to truncate a partial classical-quantum system ρ_{XVW} according to some property $G(v,w)$ of desire on $\rho_{X|v,w}$. The goal is to remove the parts of ρ_{XVW} where G is not satisfied. We execute the following procedure:

(1) Maintain a partial system ρ'_{XVW} initialized as ρ_{XVW}, and subspaces $\mathcal{V}_w \subseteq \mathcal{V}$ initialized as $\mathcal{V}$ for each $w \in \mathcal{W}$.
(2) Pick $w \in \mathcal{W}$ and $|v\rangle \in \mathcal{V}_w$ such that $\mathrm{Tr}[\rho'_{X|v,w}] > 0$ and $G(v,w)$ is false.
(3) Change the partial system ρ'_{XVW} into $\Pi_{v,w}\rho'_{XVW}\Pi_{v,w}$ by the projection
$$\Pi_{v,w} = \mathbb{I}_X \otimes (\mathbb{I}_{VW} - |v,w\rangle\langle v,w|),$$
and change $\mathcal{V}_w$ to its subspace orthogonal to $|v\rangle$, that is
$$\{|v'\rangle \in \mathcal{V}_w \mid \langle v|v'\rangle = 0\}.$$
(4) Repeat from step 2 until there is no such w and $|v\rangle$. Denote the final system as $\rho^{|G}_{XVW}$.

In step 2 we pick w and $|v\rangle$ arbitrarily as long as it satisfies the requirements, however we could always think of it as iterating over $w \in \mathcal{W}$ and processing each $\rho_{XV|w}$ separately. The choices of $|v\rangle$ for each w do affect the final system $\rho^{|G}_{XVW}$; Yet as we will see later, these choices are irrelevant to our proof.

Below, we give two useful lemmas on truncated systems whose proofs are omitted in this version of the paper:

Lemma 4.1. *For every $|v\rangle \in \mathcal{V}$ and $w \in \mathcal{W}$ with $\mathrm{Tr}[\rho^{|G}_{X|v,w}] > 0$, there exists $|v'\rangle$ in the remaining subspace $\mathcal{V}_w$ such that*

$$P^{\rho^{|G}}_{X|v,w} = P^{\rho}_{X|v',w} = P^{\rho^{|G}}_{X|v',w}.$$

A direct corollary of the above lemma is that if $G(v,w)$ only depends on the distribution $P^{\rho}_{X|v,w}$, then $G(v,w)$ holds for every $|v\rangle \in \mathcal{V}$ and $w \in \mathcal{W}$ in the truncated system $\rho^{|G}_{XVW}$, even when $|v\rangle$ is not in the remaining subspace $\mathcal{V}_w$.

Lemma 4.2. *For each $w \in W$, let $|v_1\rangle, \ldots, |v_d\rangle$ be the states picked in step 2 within $\mathcal{V}_w$. Then*

$$\left\|\rho_{XV|w} - \rho^{|G}_{XV|w}\right\|_{\mathrm{Tr}} \le 3\sum_{i=1}^{d}\sqrt{\mathrm{Tr}[\rho_{X|v_i,w}]\mathrm{Tr}[\rho_{XV|w}]}.$$

Since $\mathrm{Tr}[\rho_{XV|w}] \le 1$ always holds, by summing over all $w \in \mathcal{W}$ we get the following corollary:

Corollary 4.3. *Let $|v_1,w_1\rangle, \ldots, |v_d,w_d\rangle$ be all of the memory states picked in step 2. Then*

$$\left\|\rho_{XVW} - \rho^{|G}_{XVW}\right\|_{\mathrm{Tr}} \le 3\sum_{i=1}^{d}\sqrt{\mathrm{Tr}[\rho_{X|v_i,w_i}]}.$$

4.2 Truncated Branching Program

The properties that we desire for the partial system ρ_{XVW} consist of three parts:

- Small L_2 norm: Let $G_2(v,w)$ be the property that
$$\left\|P^{\rho}_{X|v,w}\right\|_2 \le 2^{\ell}\cdot 2^{-n/2}.$$
- Small L_∞ norm: Let $G_\infty(v,w)$ be the property that
$$\left\|P^{\rho}_{X|v,w}\right\|_\infty \le 2^{2\ell+9r}\cdot 2^{-n}.$$
- Even division: For every $a\in\mathcal{A}$, let $G_a(v,w)$ be the property that
$$|\langle M_a, P^{\rho}_{X|v,w}\rangle| \le 2^{-r}.$$

Now we define the truncated branching program, by specifying the truncated partial classical-quantum system $\tau^{(t)}_{XVW}$ for each stage t. Initially let $\tau^{(0)}_{XVW}=\rho^{(0)}_{XVW}$. For each stage $0\le t\le T$, the truncation consists of three ingredients (below we ignore the superscripts on P for convenience):

(1) Remove parts where $\left\|P_{X|v,w}\right\|_2$ is large. That is, let $\tau^{(t,\star)}_{XVW}=\tau^{(t)|G_2}_{XVW}$.
(2) Remove parts where $\left\|P_{X|v,w}\right\|_\infty$ is large. This is done by two steps.
   - First, let $g\in\{0,1\}^{\mathcal{X}\otimes\mathcal{W}}$ be an indicator vector such that $g(x,w)=1$ if and only if
   $$\mathrm{Tr}[\tau^{(t,\star)}_{X|w}]>0 \text{ and } P^{\tau^{(t,\star)}}_{X|w}(x)\le 2^{2\ell+5r}\cdot 2^{-n}.$$
   Let $\tau^{(t,\circ)}_{XVW}=(gg^\dagger\otimes\mathbb{I}_V)\tau^{(t,\star)}_{XVW}(gg^\dagger\otimes\mathbb{I}_V)$, where $gg^\dagger$ is the projection operator acting on $\mathcal{X}\otimes\mathcal{W}$.
   - To make sure that the distributions did not change a lot after the projection $gg^\dagger$, for each $0\le t<T$, let $G_t(v,w)$ be the property that
   $$\mathrm{Tr}[\tau^{(t,\circ)}_{X|v,w}]\ge(1-2^{-r})\mathrm{Tr}[\tau^{(t,\star)}_{X|v,w}].$$
   Let $\tau^{(t,\infty)}_{XVW}=\tau^{(t,\circ)|G_\infty\wedge G_t}_{XVW}$.
(3) For each $a\in\mathcal{A}$, remove (only for this a) parts where $P_{X|v,w}$ is not evenly divided by a. That is, for each $a\in\mathcal{A}$, let $\tau^{(t,a)}_{XVW}=\tau^{(t,\infty)|G_a}_{XVW}$.

Then, if $t<T$, for each $a\in_R\mathcal{A}$ we evolve the system by applying the sample operations $\Phi_{t,a,b}$ as the original branching program on $\tau^{(t,a)}_{XVW}$, so that we have
$$\tau^{(t+1)}_{XVW}=\mathop{\mathbb{E}}_{a\in_R\mathcal{A}}\left[\sum_{x\in\mathcal{X}}|x\rangle\langle x|\otimes\Phi_{t,a,M(a,x)}\left(\tau^{(t,a)}_{VW|x}\right)\right].$$

4.3 Bounding the Truncation Difference

In order to show that the success probability of the original branching program $\rho^{(t)}$ is low, the plan is to prove an upper bound on the success probability of the truncated branching program $\tau^{(t)}$, and bound the difference between the two probabilities.

Here we bound the difference by the trace distance between the two systems $\rho^{(t)}_{XVW}$ and $\tau^{(t)}_{XVW}$. We will show that the contribution to the trace distance from each one of the truncation ingredients is small, and in addition the evolution preserves the trace distance.

4.3.1 Truncation by G_2.

Lemma 4.4. *For every $0\le t\le T$, $|v\rangle\in\mathcal{V}$ and $w\in\mathcal{W}$ such that $G_2(v,w)$ is violated (that is, $\left\|P^{\tau^{(t)}}_{X|v,w}\right\|_2>2^{\ell}\cdot 2^{-n/2}$), we must have $\mathrm{Tr}[\tau^{(t)}_{X|v,w}]<2^{-2m}\cdot 2^{-4r}$.*

The lemma says, for any direction $|v,w\rangle$ picked by the truncation procedure, the weight will be small and the truncation will not change the state significantly.

Proof. This is our main technical lemma and we defer the proof to Section 5. □

Since there are at most 2^{q+m} such directions picked in the truncation procedure, we conclude the following corollary.

Corollary 4.5. *For every $0\le t\le T$, we have*
$$\left\|\tau^{(t,\star)}_{XVW}-\tau^{(t)}_{XVW}\right\|_{\mathrm{Tr}}\le 3\cdot 2^{q-2r}.$$

4.3.2 Truncation by G_∞.

Lemma 4.6. *For every $0\le t\le T$ and $w\in\mathcal{W}$ we have*
$$\sum_{\substack{x\in\mathcal{X}\\ g(x,w)=0}}P^{\tau^{(t,\star)}}_{X|w}(x)\le 2^{-5r}.$$

Proof. By Claim 3.2, $P^{\tau^{(t,\star)}}_{X|w}$ is a convex combination of $P^{\tau^{(t,\star)}}_{X|v,w}$. From Lemma 4.1 we know that $G_2(P^{\tau^{(t,\star)}}_{X|v,w})$ holds for every $|v\rangle$ and w, and thus by convexity of ℓ_2-norms we know that $G_2(P^{\tau^{(t,\star)}}_{X|w})$ also holds. That means
$$\mathop{\mathbb{E}}_{x\sim P^{\tau^{(t,\star)}}_{X|w}}\left[P^{\tau^{(t,\star)}}_{X|w}(x)\right]=\left\|P^{\tau^{(t,\star)}}_{X|w}\right\|_2^2\le 2^{2\ell}\cdot 2^{-n}.$$
Therefore, by Markov's inequality we have
$$\sum_{\substack{x\in\mathcal{X}\\ g(x,w)=0}}P^{\tau^{(t,\star)}}_{X|w}(x)=\Pr_{x\sim P^{\tau^{(t,\star)}}_{X|w}}\left[P^{\tau^{(t,\star)}}_{X|w}(x)>2^{2\ell+5r}\cdot 2^{-n}\right]\le 2^{-5r}.$$
□

Corollary 4.7. *For every $0\le t\le T$ and every $w\in\mathcal{W}$, we have $\tau^{(t,\circ)}_{XV|w}\le\tau^{(t,\star)}_{XV|w}$, and*
$$\mathrm{Tr}[\tau^{(t,\circ)}_{XV|w}]\ge(1-2^{-5r})\cdot\mathrm{Tr}[\tau^{(t,\star)}_{XV|w}].$$
Moreover, we have $\left\|\tau^{(t,\circ)}_{XVW}-\tau^{(t,\star)}_{XVW}\right\|_{\mathrm{Tr}}\le 2^{-5r}$.

Lemma 4.8. *For every $0\le t\le T$, $|v\rangle\in\mathcal{V}$ and $w\in\mathcal{W}$ such that $G_\infty(v,w)$ is violated (that is, $\left\|P^{\tau^{(t,\circ)}}_{X|v,w}\right\|_\infty>2^{2\ell+9r}\cdot 2^{-n}$) or $G_t(v,w)$ is violated (that is, $\mathrm{Tr}[\tau^{(t,\circ)}_{X|v,w}]<(1-2^{-r})\mathrm{Tr}[\tau^{(t,\star)}_{X|v,w}]$), we must have $\mathrm{Tr}[\tau^{(t,\circ)}_{X|v,w}]<2\cdot 2^{-4r}\cdot\mathrm{Tr}[\tau^{(t,\circ)}_{X|w}]$.*

Proof. If $G_\infty(v,w)$ is violated, let $x\in\mathcal{X}$ be the one such that $P^{\tau^{(t,\circ)}}_{X|v,w}(x)>2^{2\ell+9r}\cdot 2^{-n}$. If $g(x,w)=0$ then $P^{\tau^{(t,\circ)}}_{X|w}(x)=0$, while if $g(x,w)=1$ then by Corollary 4.7,
$$P^{\tau^{(t,\circ)}}_{X|w}(x)\le\frac{\mathrm{Tr}[\tau^{(t,\star)}_{X|w}]}{\mathrm{Tr}[\tau^{(t,\circ)}_{X|w}]}\cdot 2^{2\ell+5r}\cdot 2^{-n}\le(1-2^{-5r})^{-1}\cdot 2^{2\ell+5r}\cdot 2^{-n}.$$

Hence we always have

$$\mathrm{Tr}[\tau^{(t,\circ)}_{X|v,w}] \le \frac{P^{\tau^{(t,\circ)}}_{X|w}(x)}{P^{\tau^{(t,\circ)}}_{X|v,w}(x)} \cdot \mathrm{Tr}[\tau^{(t,\circ)}_{X|w}] \le 2 \cdot 2^{-4r} \cdot \mathrm{Tr}[\tau^{(t,\circ)}_{X|w}],$$

where the first inequality comes from the fact that $\tau^{(t,\circ)}_{X|w} \ge \tau^{(t,\circ)}_{X|v,w}$ and Equation (3).

If $G_t(v,w)$ is violated, since we know from Corollary 4.7 that

$$\begin{aligned}\left|\mathrm{Tr}[\tau^{(t,\circ)}_{X|v,w}] - \mathrm{Tr}[\tau^{(t,\star)}_{X|v,w}]\right| &\le \left\|\tau^{(t,\circ)}_{XV|w} - \tau^{(t,\star)}_{XV|w}\right\|_{\mathrm{Tr}} \\ &\le 2^{-5r} \cdot \mathrm{Tr}[\tau^{(t,\star)}_{XV|w}] \\ &\le 2^{-5r} \cdot (1-2^{-5r})^{-1} \cdot \mathrm{Tr}[\tau^{(t,\circ)}_{XV|w}],\end{aligned}$$

therefore from $\mathrm{Tr}[\tau^{(t,\circ)}_{X|v,w}] < (1-2^{-r})\mathrm{Tr}[\tau^{(t,\star)}_{X|v,w}]$ we deduce that

$$\begin{aligned}\mathrm{Tr}[\tau^{(t,\circ)}_{X|v,w}] &< (2^r-1)\cdot\left(\mathrm{Tr}[\tau^{(t,\star)}_{X|v,w}] - \mathrm{Tr}[\tau^{(t,\circ)}_{X|v,w}]\right) \\ &\le (2^r-1)\cdot 2^{-5r}\cdot(1-2^{-5r})^{-1}\cdot\mathrm{Tr}[\tau^{(t,\circ)}_{XV|w}] \\ &< 2\cdot 2^{-4r}\cdot\mathrm{Tr}[\tau^{(t,\circ)}_{X|w}]. \qquad \square\end{aligned}$$

Corollary 4.9. *For every $0 \le t \le T$, we have*

$$\left\|\tau^{(t,\infty)}_{XVW} - \tau^{(t,\circ)}_{XVW}\right\|_{\mathrm{Tr}} \le 5 \cdot 2^{q-2r}.$$

4.3.3 Truncation by G_a. Notice that in the truncation step from $\tau^{(t,\star)}$ to $\tau^{(t,\circ)}$, the distribution $P^{\tau^{(t,\star)}}_{X|v,w}$ might change and not satisfy G_2 anymore. However, with the truncation by G_t, any such distribution that changes too much is eliminated, and we have the following guarantee.

Lemma 4.10. *For every $0 \le t \le T$, $|v\rangle \in \mathcal{V}$ and $w \in \mathcal{W}$, we have*

$$\left\|P^{\tau^{(t,\infty)}}_{X|v,w}\right\|_2 \le (1-2^{-r})^{-1} \cdot 2^{\ell} \cdot 2^{-n/2}.$$

Proof. By Lemma 4.1, there exists $|v'\rangle \in \mathcal{V}$ such that $P^{\tau^{(t,\infty)}}_{X|v,w} = P^{\tau^{(t,\infty)}}_{X|v',w} = P^{\tau^{(t,\circ)}}_{X|v',w}$. The truncation by G_t ensures that

$$\mathrm{Tr}[\tau^{(t,\circ)}_{X|v',w}] \ge (1-2^{-r})\mathrm{Tr}[\tau^{(t,\star)}_{X|v',w}],$$

and therefore

$$\begin{aligned}\left\|P^{\tau^{(t,\infty)}}_{X|v,w}\right\|_2 = \left\|P^{\tau^{(t,\circ)}}_{X|v',w}\right\|_2 &= \frac{\left\|\mathrm{diag}\,\tau^{(t,\circ)}_{X|v',w}\right\|_2}{\mathrm{Tr}[\tau^{(t,\circ)}_{X|v',w}]} \\ &\le \frac{\left\|\mathrm{diag}\,\tau^{(t,\star)}_{X|v',w}\right\|_2}{(1-2^{-r})\mathrm{Tr}[\tau^{(t,\star)}_{X|v',w}]} \le (1-2^{-r})^{-1}\cdot 2^{\ell}\cdot 2^{-n/2}. \qquad \square\end{aligned}$$

Lemma 4.11. *For every partial classical-quantum system τ_{XV} over $\mathcal{X}\otimes\mathcal{V}$ such that $\left\|P^{\tau}_{X|v}\right\|_2 \le 2^{\ell'}\cdot 2^{-n/2}$ holds for every $|v\rangle \in \mathcal{V}$, we have*

$$\Pr_{a\in_R\mathcal{A}}\left[\exists |v\rangle\in\mathcal{V}, |\langle M_a, P^{\tau}_{X|v}\rangle| \ge 2^{-r}\right] \le 2^{-2r}.$$

Proof. Notice that we can think of $\tau_V = \mathrm{Tr}_X[\tau_{XV}]$ to be $\mathbb{I}_V$. This is because we can first assume that τ_V is full rank (otherwise change $\mathcal{V}$ to its subspace and the conclusion in this lemma still holds), and if we have diagonalization $Q^\dagger \tau_V Q = \mathbb{I}_V$ for some non-singular Q, then consider the new system

$$\tau'_{XV} = (\mathbb{I}_X \otimes Q^\dagger)\tau_{XV}(\mathbb{I}_X \otimes Q),$$

and the set of distributions $\{P^{\tau}_{X|v}\}$ and $\{P^{\tau'}_{X|v}\}$ over $|v\rangle \in \mathcal{V}$ are the same, since $P^{\tau'}_{X|v} = P^{\tau}_{X|v'}$ for $|v'\rangle \sim Q|v\rangle$. With $\tau_V = \mathbb{I}_V$ we have $\mathrm{Tr}[\tau_{X|v}] = 1$ for every $|v\rangle \in \mathcal{V}$, and thus $P^{\tau}_{X|v} = \mathrm{diag}\,\tau_{X|v}$.

Let $\mathcal{A}' \subseteq \mathcal{A}$ be the set of $a \in \mathcal{A}$ such that there exists $|v\rangle \in \mathcal{V}$ with $|\langle M_a, P^{\tau}_{X|v}\rangle| \ge 2^{-r}$. For each $a \in \mathcal{A}'$, let

$$\sigma_a = \mathrm{Tr}_X[(\mathrm{Diag}\, M_a \otimes \mathbb{I}_V)\tau_{XV}]$$

which is a Hermitian operator on $\mathcal{V}$. There exists $|v\rangle \in \mathcal{V}$ such that

$$|\langle v|\sigma_a|v\rangle| = |\langle M_a, \mathrm{diag}\,\tau_{X|v}\rangle| = |\langle M_a, P^{\tau}_{X|v}\rangle| \ge 2^{-r},$$

which means that $\|\sigma_a\|_2 \ge 2^{-r}$. Now let $|u\rangle$ be a uniformly random unit vector in $\mathcal{V}$, and by Lemma 3.1 we know that for some absolute constant c,

$$\begin{aligned}\Pr_{|u\rangle}\left[|\langle u|\sigma_a|u\rangle| \ge 2^{-r'}\right] &\ge 1 - 2^{(q+r-r')/2}c - e^{-2^q} \\ &\ge 1 - 2^{-r}c - e^{-1} \ge 1/2.\end{aligned}$$

The second last inequality comes from Eq. (4), while the last inequality is because of the assumption that r is sufficiently large.

Since the above holds for every $a \in \mathcal{A}'$, it implies that

$$\Pr_{a\in\mathcal{A}', |u\rangle}[|\langle u|\sigma_a|u\rangle| \ge 2^{-r'}] \ge 1/2.$$

It means that there exists some $|u\rangle \in \mathcal{V}$ such that $|\langle u|\sigma_a|u\rangle| \ge 2^{-r'}$ for at least half of $a \in \mathcal{A}'$. On the other hand, since M is a (k',ℓ')-extractor with error $2^{-r'}$, and $\left\|P^{\tau}_{X|u}\right\|_2 \le 2^{\ell'}\cdot 2^{-n/2}$, there are at most $2^{-k'}$ fraction of $a \in \mathcal{A}$ such that $|\langle u|\sigma_a|u\rangle| = |\langle M_a, P^{\tau}_{X|u}\rangle| \ge 2^{-r'}$. That means

$$\Pr_{a\in_R\mathcal{A}}\left[a \in \mathcal{A}'\right] \le 2\cdot 2^{-k'} \le 2^{-2r}.$$

Here $k' - 1 \ge 2r$, by the definition of r. $\square$

Corollary 4.12. *For every $0 \le t \le T$, we have $\mathop{\mathbf{E}}_{a\in_R\mathcal{A}}\left\|\tau^{(t,a)}_{XVW} - \tau^{(t,\infty)}_{XVW}\right\|_{\mathrm{Tr}} \le 2^{-2r}$.*

4.3.4 Evolution preserves trace distance.

Lemma 4.13. *For every $0 \le t < T$, we have*

$$\left\|\tau^{(t+1)}_{XVW} - \rho^{(t+1)}_{XVW}\right\|_{\mathrm{Tr}} \le \mathop{\mathbf{E}}_{a\in_R\mathcal{A}}\left\|\tau^{(t,a)}_{XVW} - \rho^{(t)}_{XVW}\right\|_{\mathrm{Tr}}.$$

The proof is by triangle inequality and contractivity of quantum channels under trace norms, and thus omitted.

4.4 Proof of Theorem 2

Proof. First, combining Corollaries 4.5, 4.7, 4.9 and 4.12 and Lemma 4.13 we have

$$\begin{aligned}&\left\|\tau^{(t+1)}_{XVW} - \rho^{(t+1)}_{XVW}\right\|_{\mathrm{Tr}} \\ \le\ &\left\|\tau^{(t)}_{XVW} - \rho^{(t)}_{XVW}\right\|_{\mathrm{Tr}} + 8\cdot 2^{q-2r} + 2^{-5r} + 2^{-2r}.\end{aligned}$$

Since $\tau^{(0)}_{XVW} = \rho^{(0)}_{XVW}$, by triangle inequality we know that

$$\left\|\tau^{(T)}_{XVW} - \rho^{(T)}_{XVW}\right\|_{\mathrm{Tr}} \le T \cdot 10 \cdot 2^{q-2r} \le 10 \cdot 2^{q-r},$$

and thus

$$\left\|\tau^{(T,\infty)}_{XVW} - \rho^{(T)}_{XVW}\right\|_{\mathrm{Tr}} \le 10 \cdot 2^{q-r} + 8 \cdot 2^{q-2r} + 2^{-5r}.$$

This bounds the difference between the measurement probabilities of $\tau^{(T,\infty)}_{XVW}$ and $\rho^{(T)}_{XVW}$ under any measurement, specifically the difference between the success probability of the branching program ρ and the following value on τ:

$$\begin{aligned}&\sum_{\substack{x\in\mathcal{X},v\in\{0,1\}^q,w\in\mathcal{W}\\ \widetilde{x}(v,w)=x}} \langle x,v,w|\tau^{(T,\infty)}_{XVW}|x,v,w\rangle\\ =&\sum_{v\in\{0,1\}^q,w\in\mathcal{W}} \mathrm{Tr}[\tau^{(T,\infty)}_{X|v,w}]\cdot P^{\tau^{(T,\infty)}}_{X|v,w}(\widetilde{x}(v,w)).\end{aligned}$$

Since $\left\|P^{\tau^{(T,\infty)}}_{X|v,w}\right\|_\infty \le 2^{2\ell+9r}\cdot 2^{-n}$ and $\mathrm{Tr}[\tau^{(T,\infty)}_{XVW}] \le 1$, the above value is at most $2^{2\ell+9r}\cdot 2^{-n}$. Therefore the success probability of the branching program ρ is at most (recall that $2\ell+9r-n\le -r$)

$$10\cdot 2^{q-r} + 8\cdot 2^{q-2r} + 2^{-5r} + 2^{2\ell+9r}\cdot 2^{-n} = O(2^{q-r}). \qquad \square$$

5 PROOF OF LEMMA 4.4

The first step towards proving Lemma 4.4 is to analyze how $P^{\tau^{(t)}}_{X|v,w}$ evolves according to the rule

$$\tau^{(t+1)}_{XVW} = \mathop{\mathbf{E}}_{a\in_R\mathcal{A}}\left[\sum_{x\in\mathcal{X}}|x\rangle\langle x|\otimes\Phi_{t,a,M(a,x)}\left(\tau^{(t,a)}_{VW|x}\right)\right].$$

We introduce the following notations. For every $a\in\mathcal{A}$ and $b\in\{-1,1\}$, let

$$\mathbb{1}_{a,b} = \frac{1}{2}(\vec{1} + b\cdot M_a),$$

which is a 0-1 vector that indicates whether $M(a,x)=b$. Let

$$\tau^{(t,a,b)}_{XVW} = (\mathrm{Diag}\,\mathbb{1}_{a,b}\otimes\mathbb{I}_{VW})\tau^{(t,a)}_{XVW}, \tag{7}$$

so that we can write

$$\begin{aligned}\tau^{(t+1)}_{XVW} = \mathop{\mathbf{E}}_{a\in_R\mathcal{A}}\Big[&(\mathbb{I}_X\otimes\Phi_{t,a,1})\left(\tau^{(t,a,1)}_{XVW}\right)\\ &+(\mathbb{I}_X\otimes\Phi_{t,a,-1})\left(\tau^{(t,a,-1)}_{XVW}\right)\Big].\end{aligned} \tag{8}$$

Thus Claim 3.3 implies that $P^{\tau^{(t+1)}}_{X|v,w}$ is a convex combination of $P^{\tau^{(t,a,b)}}_{X|v',w'}$ for some a,b,w' and $|v'\rangle$.

5.1 Target Distribution and Badness

Before considering the target distribution, let us first establish that the ℓ_2-norms of $P^{\tau^{(t)}}_{X|v,w}$ cannot be too large, using Lemma 4.1 and Lemma 4.10:

Lemma 5.1. *For every $0\le t\le T$, $|v\rangle\in\mathcal{V}$, $w\in\mathcal{W}$, we have*

$$\left\|P^{\tau^{(t)}}_{X|v,w}\right\|_2 \le 4\cdot 2^{\ell}\cdot 2^{-n/2}.$$

From now on we use P to denote a fixed target distribution (which we will later choose to be the distribution in Lemma 4.4), such that

$$2^{\ell}\cdot 2^{-n/2} \le \|P\|_2 \le 4\cdot 2^{\ell}\cdot 2^{-n/2}.$$

We want to bound the progress of $\langle P^{\tau^{(t)}}_{X|v,w}, P\rangle$, which starts off as 2^{-n} at $t=0$, and becomes at least $2^{2\ell}\cdot 2^{-n}$ when $P^{\tau^{(t)}}_{X|v,w} = P$. Note that by Cauchy-Schwarz we always have

$$\langle P^{\tau^{(t)}}_{X|v,w}, P\rangle \le \left\|P^{\tau^{(t)}}_{X|v,w}\right\|_2\|P\|_2 \le 16\cdot 2^{2\ell}\cdot 2^{-n}. \tag{9}$$

In order to bound the progress, we introduce some new notations. For any superscript (such as (t,a)) on the partial systems, we use σ_{XVW} to denote $\tau_{XVW}(\mathrm{Diag}\,P\otimes\mathbb{I}_{VW})$. Notice that

$$\mathrm{Tr}[\sigma_{X|v,w}] = \mathrm{Tr}[\tau_{X|v,w}\,\mathrm{Diag}\,P] = \mathrm{Tr}[\tau_{X|v,w}]\cdot\langle P^{\tau}_{X|v,w}, P\rangle.$$

Similarly, $P^{\sigma}_{X|v,w}$ can be deduced from $P^{\tau}_{X|v,w}$ via

$$P^{\sigma}_{X|v,w}(x) = \frac{\mathrm{Tr}[\tau_{X|v,w}]}{\mathrm{Tr}[\sigma_{X|v,w}]}\cdot P^{\tau}_{X|v,w}(x)\cdot P(x) = \frac{P^{\tau}_{X|v,w}(x)\cdot P(x)}{\langle P^{\tau}_{X|v,w}, P\rangle}. \tag{10}$$

Therefore we can bound the ℓ_2 norm of $P^{\sigma}_{X|v,w}$ as

$$\left\|P^{\sigma}_{X|v,w}\right\|_2 \le \frac{1}{\langle P^{\tau}_{X|v,w}, P\rangle}\cdot\left\|P^{\tau}_{X|v,w}\right\|_\infty\cdot\|P\|_2.$$

Now we can identity the places where $\langle P^{\tau^{(t)}}_{X|v,w}, P\rangle$ increases by a lot, which happens when the *inner product* is not evenly divided by some $a\in\mathcal{A}$ (we will see the reason in the analysis later). Formally, at stage $0\le t<T$, we say (w,a) is *bad* if

$$\exists|v\rangle\in\mathcal{V}, \text{ s.t. } |\langle M_a, P^{\sigma^{(t,a)}}_{X|v,w}\rangle| > 2^{-r} \text{ and } \langle P^{\tau^{(t,a)}}_{X|v,w}, P\rangle \ge \frac{1}{2}\cdot 2^{-n}. \tag{11}$$

Lemma 5.2. *For every $0\le t<T$ and $w\in\mathcal{W}$, we have*

$$\Pr_{a\in_R\mathcal{A}}[(w,a) \text{ is bad}] \le 2^{-k}.$$

Proof. Since $\tau^{(t,a)}_{XVW}$ is truncated from $\tau^{(t,\infty)}_{XVW}$, Lemma 4.1 shows that for every $|v\rangle\in\mathcal{V}$, $w\in\mathcal{W}$ and $a\in\mathcal{A}$ there is $|v'\rangle\in\mathcal{V}$ such that

$$P^{\tau^{(t,a)}}_{X|v,w} = P^{\tau^{(t,\infty)}}_{X|v',w}$$

and by Eq. (10) it also implies that

$$P^{\sigma^{(t,a)}}_{X|v,w} = P^{\sigma^{(t,\infty)}}_{X|v',w}.$$

Now fix some $w\in\mathcal{W}$, and let $\mathcal{A}'\subseteq\mathcal{A}$ be the set of of $a\in\mathcal{A}$ such that

$$\exists|v\rangle\in\mathcal{V}, \text{ s.t. } |\langle M_a, P^{\sigma^{(t,\infty)}}_{X|v,w}\rangle| > 2^{-r} \text{ and } \langle P^{\tau^{(t,\infty)}}_{X|v,w}, P\rangle \ge \frac{1}{2}\cdot 2^{-n}.$$

Then $\mathcal{A}'$ contains all a such that (w,a) is bad, and our goal is to bound the fraction of $\mathcal{A}'$ in $\mathcal{A}$.

In the rest of the proof we temporarily omit the super script and write $\tau^{(t,\infty)}$ and $\sigma^{(t,\infty)}$ simply as τ and σ. For the same reason as in Lemma 4.11 we can assume that $\tau_{V|w} = \mathbb{I}_V$, and thus

$$\begin{aligned}\langle v|\sigma_{V|w}|v\rangle &= \mathrm{Tr}[\sigma_{X|v,w}] = \langle P^{\tau}_{X|v,w}, P\rangle,\\ \text{and}\quad \mathrm{Tr}[\sigma_{XV|w}] &= \langle P^{\tau}_{X|w}, P\rangle \le 16\cdot 2^{2\ell}\cdot 2^{-n}.\end{aligned}$$

where the last inequality is by Lemma 4.10 and Cauchy-Schwarz, in the same way as Eq. (9).

Suppose that we have diagonalization $\sigma_{V|w} = U^\dagger DU$, where U is unitary and D is diagonal and non-negative. Let $\mathcal{V}'\subseteq\mathcal{V}$ be the subspace spanned by $U^\dagger|e\rangle$ over the computational basis vectors

$|e\rangle \in \mathcal{V}$ such that $\langle e|D|e\rangle \geq 2^{-4r} \cdot 2^{-2\ell} \cdot 2^{-n}$. So for every $|v\rangle \in \mathcal{V}'$ we have

$$\langle P^{\tau}_{X|v,w}, P\rangle = \mathrm{Tr}[\sigma_{X|v,w}] \geq 2^{-4r} \cdot 2^{-2\ell} \cdot 2^{-n}.$$

We claim that for every $a \in \mathcal{A}'$, there exists $|v\rangle \in \mathcal{V}'$ such that $|\langle M_a, P^{\sigma}_{X|v,w}\rangle| > \frac{1}{2} \cdot 2^{-r}$. To prove the claim, let Π be the projection operator from $\mathcal{V}$ to $\mathcal{V}'$, and then $(\mathbb{I}_X \otimes \Pi)\sigma_{XV|w}(\mathbb{I}_X \otimes \Pi)$ can be conceptually seen as a truncated partial system $\sigma^{|G}_{XV|w}$ where $G(v, w)$ holds when $\mathrm{Tr}[\sigma_{X|v,w}] \geq 2^{-4r-2\ell} \cdot 2^{-n}$ for the fixed w. By Lemma 4.2 we have

$$\begin{aligned}\left\|\sigma^{|G}_{XV|w} - \sigma_{XV|w}\right\|_{\mathrm{Tr}} &\leq 3 \cdot 2^{q} \cdot \sqrt{2^{-4r-2\ell-n} \cdot \mathrm{Tr}[\sigma_{XV|w}]} \\ &\leq 12 \cdot 2^{q-2r} \cdot 2^{-n}.\end{aligned}$$

Since $a \in \mathcal{A}'$, assume for $|u\rangle \in \mathcal{V}$ we have $|\langle M_a, P^{\sigma}_{X|u,w}\rangle| > 2^{-r}$ and $\mathrm{Tr}[\sigma_{X|u,w}] = \langle P^{\tau}_{X|u,w}, P\rangle \geq \frac{1}{2} \cdot 2^{-n}$. Let $|v\rangle \sim \Pi|u\rangle$, then we have

$$\begin{aligned}&\left\|P^{\sigma}_{X|u,w} - P^{\sigma}_{X|v,w}\right\|_1 = \left\|P^{\sigma}_{X|u,w} - P^{\sigma^{|G}}_{X|u,w}\right\|_1 \\ \leq &\left\|\frac{\sigma_{X|u,w}}{\mathrm{Tr}[\sigma_{X|u,w}]} - \frac{\sigma^{|G}_{X|u,w}}{\mathrm{Tr}[\sigma^{|G}_{X|u,w}]}\right\|_{\mathrm{Tr}} \\ \leq &\left\|\frac{\sigma_{X|u,w}}{\mathrm{Tr}[\sigma_{X|u,w}]} - \frac{\sigma^{|G}_{X|u,w}}{\mathrm{Tr}[\sigma_{X|u,w}]}\right\|_{\mathrm{Tr}} + \left\|\frac{\sigma^{|G}_{X|u,w}}{\mathrm{Tr}[\sigma_{X|u,w}]} - \frac{\sigma^{|G}_{X|u,w}}{\mathrm{Tr}[\sigma^{|G}_{X|u,w}]}\right\|_{\mathrm{Tr}} \\ = &\frac{\left\|\sigma_{X|u,w} - \sigma^{|G}_{X|u,w}\right\|_{\mathrm{Tr}}}{\mathrm{Tr}[\sigma_{X|u,w}]} + \frac{\left|\mathrm{Tr}[\sigma^{|G}_{X|u,w}] - \mathrm{Tr}[\sigma_{X|u,w}]\right|}{\mathrm{Tr}[\sigma_{X|u,w}]} \\ \leq &\frac{2\left\|\sigma_{X|u,w} - \sigma^{|G}_{X|u,w}\right\|_{\mathrm{Tr}}}{\mathrm{Tr}[\sigma_{X|u,w}]} \leq \frac{2\left\|\sigma_{XV|w} - \sigma^{|G}_{XV|w}\right\|_{\mathrm{Tr}}}{\mathrm{Tr}[\sigma_{X|u,w}]} \\ \leq &48 \cdot 2^{q-2r} \leq \frac{1}{2} \cdot 2^{-r},\end{aligned}$$

where the last step is due to $q \leq r - 7$. Thus

$$|\langle M_a, P^{\sigma}_{X|v,w}\rangle| \geq |\langle M_a, P^{\sigma}_{X|u,w}\rangle| - \left\|P^{\sigma}_{X|u,w} - P^{\sigma}_{X|v,w}\right\|_1 > \frac{1}{2} \cdot 2^{-r}.$$

Similarly to the proof for Lemma 4.11, for each $a \in \mathcal{A}'$ let

$$\pi_a = \mathrm{Tr}_X[(\mathrm{Diag}\, M_a \otimes U^{\dagger}D^{-1/2}U) \cdot \sigma_{XV|w} \cdot (\mathbb{I}_X \otimes U^{\dagger}D^{-1/2}U)]$$

which is a Hermitian operator on $\mathcal{V}$. For each $|v\rangle \in \mathcal{V}$, let $|v'\rangle \sim U^{\dagger}D^{1/2}U|v\rangle$. Recall that $\sigma_{V|w} = U^{\dagger}DU$, and therefore

$$\begin{aligned}P^{\sigma}_{X|v,w} &= \frac{\mathrm{diag}\,(\mathbb{I}_X \otimes \langle v|)\sigma_{XV|w}(\mathbb{I}_X \otimes |v\rangle)}{\langle v|\sigma_{V|w}|v\rangle} \\ &= \frac{\mathrm{diag}\,(\mathbb{I}_X \otimes \langle v'|U^{\dagger}D^{-1/2}U)\sigma_{XV|w}(\mathbb{I}_X \otimes U^{\dagger}D^{-1/2}U|v'\rangle)}{\langle v'|U^{\dagger}D^{-1/2}U\sigma_{V|w}U^{\dagger}D^{-1/2}U|v'\rangle} \\ &= \mathrm{diag}\,(\mathbb{I}_X \otimes \langle v'|U^{\dagger}D^{-1/2}U)\sigma_{XV|w}(\mathbb{I}_X \otimes U^{\dagger}D^{-1/2}U|v'\rangle).\end{aligned}$$

And that means

$$\langle v'|\pi_a|v'\rangle = \Big\langle M_a, \mathrm{diag}\,(\mathbb{I}_X \otimes \langle v'|U^{\dagger}D^{-1/2}U)\, \sigma_{XV|w}(\mathbb{I}_X \otimes U^{\dagger}D^{-1/2}U|v'\rangle)\Big\rangle = \langle M_a, P^{\sigma}_{X|v,w}\rangle.$$

We showed above that there exists $|v\rangle \in \mathcal{V}'$, and thus $|v'\rangle \in \mathcal{V}'$ such that

$$|\langle v'|\pi_a|v'\rangle| = \left|\langle M_a, P^{\sigma}_{X|v,w}\rangle\right| \geq \frac{1}{2} \cdot 2^{-r},$$

which means that for $\Pi\pi_a\Pi$, the restriction of π_a on $\mathcal{V}'$, we have $\|\Pi\pi_a\Pi\|_2 \geq \frac{1}{2} \cdot 2^{-r}$. Now consider a uniformly random unit vector $|v'\rangle$ in $\mathcal{V}'$, and by Lemma 3.1 we know that for some absolute constant c,

$$\Pr_{|v\rangle'}\left[|\langle v'|\sigma_a|v'\rangle| \geq 2^{-r'}\right] \geq 1 - 2^{(q+r+1-r')/2}c - e^{-2^q} \geq \frac{1}{2}.$$

Therefore, for the random vector $|v\rangle \sim U^{\dagger}D^{-1/2}U|v'\rangle$ where $|v'\rangle$ is uniform in $\mathcal{V}'$, we conclude that

$$\Pr_{|v\rangle}\left[|\langle M_a, P^{\sigma}_{X|v,w}\rangle| \geq 2^{-r'}\right] \geq \frac{1}{2}.$$

On the other hand, as $|v'\rangle \in \mathcal{V}'$, it also holds that $|v\rangle \in \mathcal{V}'$, therefore $\langle P^{\tau}_{X|v,w}, P\rangle \geq 2^{-4r} \cdot 2^{-2\ell} \cdot 2^{-n}$ is always true. Thus there exists a $|v\rangle \in \mathcal{V}$ that simultaneously satisfies

$$\langle P^{\tau}_{X|v,w}, P\rangle \geq 2^{-4r} \cdot 2^{-2\ell} \cdot 2^{-n} \quad \text{and} \quad |\langle M_a, P^{\sigma}_{X|v,w}\rangle| \geq 2^{-r'}$$

for at least $1/2$ of $a \in \mathcal{A}'$. Since

$$\left\|P^{\sigma}_{X|v,w}\right\|_2 \leq \frac{1}{\langle P^{\tau}_{X|v,w}, P\rangle} \cdot \left\|P^{\tau}_{X|v,w}\right\|_{\infty} \cdot \|P\|_2 \leq 2^{\ell'} \cdot 2^{-n/2},$$

and M is a (k', ℓ')-extractor with error $2^{-r'}$, there are at most $2^{-k'}$ fraction of $a \in \mathcal{A}$ such that $|\langle M_a, P^{\sigma}_{X|v',w}\rangle| \geq 2^{-r'}$, which means that

$$\Pr_{a \in_R \mathcal{A}}[(w, a) \text{ is bad}] \leq \Pr_{a \in_R \mathcal{A}}[a \in \mathcal{A}'] \leq 2 \cdot 2^{-k'} = 2^{-k}. \qquad \square$$

5.2 Badness Levels

At stage t, for each classical memory state $w \in \mathcal{W}$ we count how many times the path to it has been bad, which is a random variable depending on the previous random choices of $a \in \mathcal{A}$. This is stored in another classical register B, which we call *badness level* and takes values $\beta \in \{0, \ldots, T\}$. It is initially set to be 0, that is, we let

$$\tau^{(0)}_{XVWB} = \tau^{(0)}_{XVW} \otimes |0\rangle\langle 0|_B.$$

We ensure that the distribution of B always only depends on W and is independent of X and V conditioned on W, using the following updating rules on the combined system τ_{XVWB} for each stage $0 \leq t < T$:

- The truncation steps are executed independently of B. Therefore, for each $a \in \mathcal{A}$ we let
$$\tau^{(t,a)}_{XVWB} = \sum_{w \in \mathcal{W}} \tau^{(t,a)}_{XV|w} \otimes |w\rangle\langle w| \otimes \mathrm{Diag}\, P^{\tau^{(t)}}_{B|w}. \tag{12}$$
- The value of B updates before the evolution step, where for each $a \in \mathcal{A}$ and $b \in \{-1, 1\}$ we let
$$\tau^{(t,a,b)}_{XVWB} = (\mathrm{Diag}\, \mathbb{1}_{a,b} \otimes \mathbb{I}_V \otimes U_a)\tau^{(t,a)}_{XVWB}(\mathbb{I}_{XV} \otimes U_a^{\dagger}).$$
Here U_a is a permutation operator, depending on $\tau^{(t,a)}_{XVW}$, acting on $\mathcal{W} \otimes \{0, \ldots, T\}$ such that
$$U_a|w\rangle|\beta\rangle = \begin{cases} |w\rangle|(\beta+1) \bmod (T+1)\rangle & \text{if } (w, a) \text{ is bad,} \\ |w\rangle|\beta\rangle & \text{otherwise.}\end{cases}$$

- For the evolution step, we apply the channels $\Phi_{t,a,b}$ on the memories W and V to get

$$\tau_{XVWB}^{(t+1)} = \mathop{\mathbf{E}}_{a\in_R\mathcal{A}}\Big[(\mathbb{I}_X\otimes\Phi_{t,a,1}\otimes\mathbb{I}_B)(\tau_{XVWB}^{(t,a,1)}) + (\mathbb{I}_X\otimes\Phi_{t,a,-1}\otimes\mathbb{I}_B)(\tau_{XVWB}^{(t,a,-1)})\Big].$$

Notice that the evolution step might introduce dependencies between X, V and B. However, such dependencies are eliminated later due to how we handle the truncation steps (12), and thus do not affect our proof.

We can check that the combined partial system $\tau_{XVWB}^{(t)}$ defined above is consistent with the partial system $\tau_{XVW}^{(t)}$ that we discussed in previous sections, in the sense that $\mathrm{Tr}_B[\tau_{XVWB}^{(t)}] = \tau_{XVW}^{(t)}$ always holds:

- For the truncation step, it is straightforward to check that

$$\mathrm{Tr}_B[\tau_{XVWB}^{(t,a)}] = \sum_{w\in\mathcal{W}}\tau_{XV|w}^{(t,a)}\otimes|w\rangle\langle w| = \tau_{XVW}^{(t,a)}.$$

- The permutation operator U_a acts on $\mathcal{W}$ as identity since

$$\mathrm{Tr}_B\left[U_a|w,\beta\rangle\langle w,\beta|U_a^\dagger\right] = |w\rangle\langle w|.$$

  Recalling Eq. (7) that $\tau_{XVW}^{(t,a,b)} = (\mathrm{Diag}\,\mathbb{1}_{a,b}\otimes\mathbb{I}_V)\tau_{XVW}^{(t,a)}$, we have $\mathrm{Tr}_B[\tau_{XVWB}^{(t,a,b)}] = \tau_{XVW}^{(t,a,b)}$.
- The evolution step can be checked directly from the formula without B (Eq. (8)):

$$\tau_{XVW}^{(t+1)} = \mathop{\mathbf{E}}_{a\in_R\mathcal{A}}\Big[(\mathbb{I}_X\otimes\Phi_{t,a,1})(\tau_{XVW}^{(t,a,1)}) + (\mathbb{I}_X\otimes\Phi_{t,a,-1})(\tau_{XVW}^{(t,a,-1)})\Big].$$

So all previously proved properties about $\tau_{XVW}^{(t)}$ are preserved. In addition, we prove the following two properties about badness levels.

Lemma 5.3. *For every $0\le t\le T$, $|v\rangle\in\mathcal{V}$ and $w\in\mathcal{W}$, we have*

$$\langle P_{X|v,w}^{\tau^{(t)}}, P\rangle \le \sum_{\beta=0}^{T} P_{B|w}^{\tau^{(t)}}(\beta)\cdot 2^{\beta}\cdot 2^{-n}\cdot(1-2^{-r})^{-3t}.$$

Proof. We prove it by induction on t. For $t=0$ the lemma is true as $\langle P_{X|v,w}^{\tau^{(t)}}, P\rangle = 2^{-n}$ and $P_{B|w}^{\tau^{(t)}}(0) = 1$.

Suppose the lemma holds true for some $t < T$. By a similar argument as in Lemma 4.10 and applying Lemma 4.1 multiple times, we know that for every $|v\rangle\in\mathcal{V}$, $w\in\mathcal{W}$ and $a\in\mathcal{A}$, there exists $|v'\rangle$and $|v''\rangle\in\mathcal{V}$ such that

$$\langle P_{X|v,w}^{\tau^{(t,a)}}, P\rangle = \langle P_{X|v',w}^{\tau^{(t,\circ)}}, P\rangle \le (1-2^{-r})^{-1}\langle P_{X|v',w}^{\tau^{(t,\star)}}, P\rangle = (1-2^{-r})^{-1}\langle P_{X|v'',w}^{\tau^{(t)}}, P\rangle,$$

and therefore

$$\langle P_{X|v,w}^{\tau^{(t,a)}}, P\rangle \le \sum_{\beta=0}^{T} P_{B|w}^{\tau^{(t)}}(\beta)\cdot 2^{\beta}\cdot 2^{-n}\cdot(1-2^{-r})^{-3t-1}. \tag{13}$$

Also, the truncation step by G_a implies that $|\langle M_a, P_{X|v,w}^{\tau^{(t,a)}}\rangle| \le 2^{-r}$. That is, for both $b\in\{-1,1\}$,

$$1-2^{-r} \le 2\left\|\mathbb{1}_{a,b}\cdot P_{X|v,w}^{\tau^{(t,a)}}\right\|_1 \le 1+2^{-r}.$$

Therefore we have, unconditionally

$$\langle P_{X|v,w}^{\tau^{(t,a,b)}}, P\rangle = \frac{\langle\mathbb{1}_{a,b}\cdot P_{X|v,w}^{\tau^{(t,a)}}, P\rangle}{\left\|\mathbb{1}_{a,b}\cdot P_{X|v,w}^{\tau^{(t,a)}}\right\|_1} \le 2(1-2^{-r})^{-1}\cdot\langle P_{X|v,w}^{\tau^{(t,a)}}, P\rangle. \tag{14}$$

When the inner product is evenly divided, i.e. $|\langle M_a, P_{X|v,w}^{\sigma^{(t,a)}}\rangle| \le 2^{-r}$, we further have

$$\langle\mathbb{1}_{a,b}\cdot P_{X|v,w}^{\tau^{(t,a)}}, P\rangle \le \frac{1}{2}(1+2^{-r})\langle P_{X|v,w}^{\tau^{(t,a)}}, P\rangle \le \frac{1}{2}(1-2^{-r})^{-1}\langle P_{X|v,w}^{\tau^{(t,a)}}, P\rangle,$$

which means that

$$\langle P_{X|v,w}^{\tau^{(t,a,b)}}, P\rangle = \frac{\langle\mathbb{1}_{a,b}\cdot P_{X|v,w}^{\tau^{(t,a)}}, P\rangle}{\left\|\mathbb{1}_{a,b}\cdot P_{X|v,w}^{\tau^{(t,a)}}\right\|_1} \le (1-2^{-r})^{-2}\cdot\langle P_{X|v,w}^{\tau^{(t,a)}}, P\rangle. \tag{15}$$

Now there are three cases to discuss:

- If (w,a) is bad, we have $P_{B|w}^{\tau^{(t,a,b)}}(\beta) = P_{B|w}^{\tau^{(t)}}(\beta-1)$ for every $\beta>0$. Notice that $P_{B|w}^{\tau^{(t)}}(T) = 0$ as $t<T$, and thus Eq. (13) and Eq. (14) imply that

$$\langle P_{X|v,w}^{\tau^{(t,a,b)}}, P\rangle \le \sum_{\beta=0}^{T-1} P_{B|w}^{\tau^{(t)}}(\beta)\cdot 2^{\beta+1}\cdot 2^{-n}\cdot(1-2^{-r})^{-3t-2} \le \sum_{\beta=0}^{T} P_{B|w}^{\tau^{(t,a,b)}}(\beta)\cdot 2^{\beta}\cdot 2^{-n}\cdot(1-2^{-r})^{-3(t+1)}.$$

- If (w,a) is not bad and $|\langle M_a, P_{X|v,w}^{\sigma^{(t,a)}}\rangle| \le 2^{-r}$, we have $P_{B|w}^{\tau^{(t,a,b)}}(\beta) = P_{B|w}^{\tau^{(t)}}(\beta)$ for every $\beta\ge 0$. Then Eq. (13) and Eq. (15) imply that

$$\langle P_{X|v,w}^{\tau^{(t,a,b)}}, P\rangle \le \sum_{\beta=0}^{T} P_{B|w}^{\tau^{(t)}}(\beta)\cdot 2^{\beta}\cdot 2^{-n}\cdot(1-2^{-r})^{-3t-3} = \sum_{\beta=0}^{T} P_{B|w}^{\tau^{(t,a,b)}}(\beta)\cdot 2^{\beta}\cdot 2^{-n}\cdot(1-2^{-r})^{-3(t+1)}.$$

- If (w,a) is not bad and $|\langle M_a, P_{X|v,w}^{\sigma^{(t,a)}}\rangle| > 2^{-r}$, by the definition of badness (11) we must have $\langle P_{X|v,w}^{\tau^{(t,a)}}, P\rangle < \frac{1}{2}\cdot 2^{-n}$. Thus by Eq. (14),

$$\langle P_{X|v,w}^{\tau^{(t,a,b)}}, P\rangle < (1-2^{-r})^{-1}\cdot 2^{-n} \le \sum_{\beta=0}^{T} P_{B|w}^{\tau^{(t,a,b)}}(\beta)\cdot 2^{\beta}\cdot 2^{-n}\cdot(1-2^{-r})^{-3(t+1)}.$$

The last inequality follows from $\sum_{\beta=0}^{T} P_{B|w}^{\tau^{(t,a,b)}}(\beta)\cdot 2^{\beta}\cdot 2^{-n}\cdot(1-2^{-r})^{-3(t+1)} \ge 2^{-n}(1-2^{-r})^{-3(t+1)}$. Hence we obtain the same conclusion from all three cases.

For the evolution step, since B is classical we can view X and B as a whole and apply Claim 3.3 on $P^{\tau^{(t+1)}}_{XB|v,w}$, which asserts that $P^{\tau^{(t+1)}}_{XB|v,w}$ is a convex combination of $P^{\tau^{(t,a,b)}}_{XB|v',w'}$ for some a, b, w' and $|v'\rangle$. Then by linearity we conclude that [6]

$$\langle P^{\tau^{(t+1)}}_{X|v,w}, P\rangle \le \sum_{\beta=0}^{T} P^{\tau^{(t+1)}}_{B|w}(\beta)\cdot 2^{\beta}\cdot 2^{-n}\cdot(1-2^{-r})^{-3(t+1)}. \qquad \square$$

Lemma 5.4. *For every $0 \le \beta \le t \le T$ we have*

$$\langle\beta|\tau^{(t)}_B|\beta\rangle \le 2^{-k\beta}\binom{t}{\beta}.$$

Proof. We prove it by induction on t. For $t = 0$ the lemma holds as $\tau^{(0)}_B = |0\rangle\langle 0|_B$. Also notice that the lemma is trivially true for every t when $\beta = 0$.

Now suppose the lemma holds for some t. By definition we have

$$\tau^{(t+1)}_B = \mathop{\mathbf{E}}_{a\in_R\mathcal{A}}[\tau^{(t,a,1)}_B + \tau^{(t,a,-1)}_B] = \mathop{\mathbf{E}}_{a\in_R\mathcal{A}} \mathrm{Tr}_W[U_a\tau^{(t,a)}_{WB}U_a^\dagger].$$

Therefore

$$\langle\beta|\tau^{(t+1)}_B|\beta\rangle = \sum_{w\in\mathcal{W}} \mathop{\mathbf{E}}_{a\in_R\mathcal{A}}\left[\langle w,\beta|U_a\tau^{(t,a)}_{WB}U_a^\dagger|w,\beta\rangle\right].$$

By Lemma 5.2 we know that for every $w \in \mathcal{W}$, the probability that (w, a) is bad for $a \in_R \mathcal{A}$ is at most 2^{-k}. In other words, for every $\beta > 0$,

$$U_a^\dagger|w,\beta\rangle = \begin{cases} |w,\beta\rangle, & \text{w.p. } \ge 1-2^{-k} \\ |w,\beta-1\rangle, & \text{w.p. } \le 2^{-k}\end{cases}$$

where the probability is taken over the random choice of a. It means that

$$\begin{aligned}\langle\beta|\tau^{(t+1)}_B|\beta\rangle &\le \sum_{w\in\mathcal{W}}\langle w,\beta|\tau^{(t,a)}_{WB}|w,\beta\rangle \\ &\quad + 2^{-k}\sum_{w\in\mathcal{W}}\langle w,\beta-1|\tau^{(t,a)}_{WB}|w,\beta-1\rangle \\ &= \langle\beta|\tau^{(t,a)}_B|\beta\rangle + 2^{-k}\cdot\langle\beta-1|\tau^{(t,a)}_B|\beta-1\rangle.\end{aligned}$$

Notice that

$$\begin{aligned}\tau^{(t,a)}_B &= \sum_{w\in\mathcal{W}}\mathrm{Tr}[\tau^{(t,a)}_{XV|w}]\cdot\mathrm{Diag}\,P^{\tau^{(t)}}_{B|w} \\ &\le \sum_{w\in\mathcal{W}}\mathrm{Tr}[\tau^{(t)}_{XV|w}]\cdot\mathrm{Diag}\,P^{\tau^{(t)}}_{B|w} = \tau^{(t)}_B,\end{aligned}$$

and thus we conclude that

$$\begin{aligned}\langle\beta|\tau^{(t+1)}_B|\beta\rangle &\le \langle\beta|\tau^{(t)}_B|\beta\rangle + 2^{-k}\cdot\langle\beta-1|\tau^{(t)}_B|\beta-1\rangle \\ &\le 2^{-k\beta}\binom{t}{\beta} + 2^{-k}\cdot 2^{-k(\beta-1)}\binom{t}{\beta-1} = 2^{-k\beta}\binom{t+1}{\beta}. \qquad \square\end{aligned}$$

With the lemmas above in hand, we can finally prove Lemma 4.4.

[6] It should be noted that in $\tau^{(t+1)}$, X and B are not independent. (In $\tau^{(t,a,b)}$ they are independent (conditioned on v', w')). Nevertheless, independence of X, B (in $\tau^{(t+1)}$) is not needed or used here and we can conclude the final inequality by linearity by taking the corresponding convex combination of all inequalities.

Proof for Lemma 4.4. For the target distribution $P = P^{\tau^{(t)}}_{X|v,w}$ we have $\langle P^{\tau^{(t)}}_{X|v,w}, P\rangle > 2^{2\ell}\cdot 2^{-n}$, so by Lemma 5.3,

$$\sum_{\beta=0}^{T} P^{\tau^{(t)}}_{B|w}(\beta)\cdot 2^{\beta}\cdot(1-2^{-r})^{-3t} > 2^{2\ell}.$$

Since $t \le T \le 2^{r-2}$, we have $(1-2^{-r})^{-3t} \le 2$, and thus

$$\sum_{\beta=\ell}^{T} P^{\tau^{(t)}}_{B|w}(\beta)\cdot 2^{\beta} > \frac{1}{2}\left(2^{2\ell} - 2\cdot\sum_{\beta=0}^{\ell-1}2^{\beta}\right) > 2^{\ell}.$$

On the other hand, for every $\beta \ge \ell$, by Lemma 5.4,

$$\mathrm{Tr}[\tau^{(t)}_{B|w}]\cdot P^{\tau^{(t)}}_{B|w}(\beta) \le \langle\beta|\tau^{(t)}_B|\beta\rangle \le (2^{-k}t)^{\beta} < 2^{-(k-r)\beta},$$

and thus by Eq. (6),

$$\begin{aligned}\mathrm{Tr}[\tau^{(t)}_{X|v,w}] \le \mathrm{Tr}[\tau^{(t)}_{B|w}] &< 2^{-\ell}\sum_{\beta=\ell}^{T}2^{-(k-r)\beta}\cdot 2^{\beta} \\ &\le 2\cdot 2^{-(k-r)\ell} \le 2^{-2m}\cdot 2^{-4r}. \qquad \square\end{aligned}$$

ACKNOWLEDGMENTS

We are grateful to Uma Girish for many important discussions and suggestions on the draft of this paper, and to the anonymous reviewers for their helpful comments.

REFERENCES

[1] Scott Aaronson. 2020. Shadow Tomography of Quantum States. *SIAM J. Comput.* 49, 5 (2020). https://doi.org/10.1137/18M120275X

[2] Dorit Aharonov, Jordan Cotler, and Xiao-Liang Qi. 2022. Quantum algorithmic measurement. *Nature communications* 13, 1 (2022), 1–9. https://doi.org/10.1038/s41467-021-27922-0

[3] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando GSL Brandao, David A Buell, et al. 2019. Quantum supremacy using a programmable superconducting processor. *Nature* 574, 7779 (2019), 505–510. https://doi.org/10.1038/s41586-019-1666-5

[4] Yonatan Aumann, Yan Zong Ding, and Michael O. Rabin. 2002. Everlasting security in the bounded storage model. *IEEE Trans. Inf. Theory* 48, 6 (2002), 1668–1680. https://doi.org/10.1109/TIT.2002.1003845

[5] Yonatan Aumann and Michael O. Rabin. 1999. Information Theoretically Secure Communication in the Limited Storage Space Model. In *Advances in Cryptology - CRYPTO '99, 19th Annual International Cryptology Conference, Santa Barbara, California, USA, August 15-19, 1999, Proceedings (Lecture Notes in Computer Science, Vol. 1666)*, Michael J. Wiener (Ed.). Springer, 65–79. https://doi.org/10.1007/3-540-48405-1_5

[6] Paul Beame, Shayan Oveis Gharan, and Xin Yang. 2018. Time-Space Tradeoffs for Learning Finite Functions from Random Evaluations, with Applications to Polynomials. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018 (Proceedings of Machine Learning Research, Vol. 75)*, Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (Eds.). PMLR, 843–856.

[7] Anne Broadbent and Peter Yuen. 2021. Device-Independent Oblivious Transfer from the Bounded-Quantum-Storage-Model and Computational Assumptions. arXiv:2111.08595 [quant-ph]

[8] Sébastien Bubeck, Sitan Chen, and Jerry Li. 2020. Entanglement is Necessary for Optimal Quantum Property Testing. In *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020, Durham, NC, USA, November 16-19, 2020*, Sandy Irani (Ed.). IEEE, 692–703. https://doi.org/10.1109/FOCS46700.2020.00070

[9] Christian Cachin and Ueli M. Maurer. 1997. Unconditional Security Against Memory-Bounded Adversaries. In *Advances in Cryptology - CRYPTO '97, 17th Annual International Cryptology Conference, Santa Barbara, California, USA, August 17-21, 1997, Proceedings (Lecture Notes in Computer Science, Vol. 1294)*, Burton S. Kaliski Jr. (Ed.). Springer, 292–306. https://doi.org/10.1007/BFb0052243

[10] Anthony Carbery and James Wright. 2001. Distributional and L^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical Research Letters* 8, 3 (2001), 233–248. https://doi.org/10.4310/mrl.2001.v8.n3.a1

[11] Sitan Chen, Jordan Cotler, Hsin-Yuan Huang, and Jerry Li. 2021. Exponential Separations Between Learning With and Without Quantum Memory. In *62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2021, Denver, CO, USA, February 7-10, 2022*. IEEE, 574–585. https://doi.org/10.1109/FOCS52979.2021.00063

[12] Sitan Chen, Jerry Li, and Ryan O'Donnell. 2022. Toward Instance-Optimal State Certification With Incoherent Measurements. In *Conference on Learning Theory, 2-5 July 2022, London, UK (Proceedings of Machine Learning Research, Vol. 178)*, Po-Ling Loh and Maxim Raginsky (Eds.). PMLR, 2541–2596.

[13] Ivan Damgård, Serge Fehr, Renato Renner, Louis Salvail, and Christian Schaffner. 2007. A Tight High-Order Entropic Quantum Uncertainty Relation with Applications. In *Advances in Cryptology - CRYPTO 2007, 27th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 19-23, 2007, Proceedings (Lecture Notes in Computer Science, Vol. 4622)*, Alfred Menezes (Ed.). Springer, 360–378. https://doi.org/10.1007/978-3-540-74143-5_20

[14] Ivan Damgård, Serge Fehr, Louis Salvail, and Christian Schaffner. 2008. Cryptography in the Bounded-Quantum-Storage Model. *SIAM J. Comput.* 37, 6 (2008), 1865–1890. https://doi.org/10.1137/060651343

[15] Ivan Damgård, Serge Fehr, Louis Salvail, and Christian Schaffner. 2014. Secure identification and QKD in the bounded-quantum-storage model. *Theor. Comput. Sci.* 560 (2014), 12–26. https://doi.org/10.1016/j.tcs.2014.09.014

[16] Yan Zong Ding and Michael O. Rabin. 2002. Hyper-Encryption and Everlasting Security. In *STACS 2002, 19th Annual Symposium on Theoretical Aspects of Computer Science, Antibes - Juan les Pins, France, March 14-16, 2002, Proceedings (Lecture Notes in Computer Science, Vol. 2285)*, Helmut Alt and Afonso Ferreira (Eds.). Springer, 1–26. https://doi.org/10.1007/3-540-45841-7_1

[17] Yevgeniy Dodis, Willy Quach, and Daniel Wichs. 2021. Speak Much, Remember Little: Cryptography in the Bounded Storage Model, Revisited. Cryptology ePrint Archive, Paper 2021/1270. https://eprint.iacr.org/2021/1270

[18] Yevgeniy Dodis, Willy Quach, and Daniel Wichs. 2022. Authentication in the Bounded Storage Model. In *Advances in Cryptology - EUROCRYPT 2022 - 41st Annual International Conference on the Theory and Applications of Cryptographic Techniques, Trondheim, Norway, May 30 - June 3, 2022, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 13277)*, Orr Dunkelman and Stefan Dziembowski (Eds.). Springer, 737–766. https://doi.org/10.1007/978-3-031-07082-2_26

[19] Stefan Dziembowski and Ueli M. Maurer. 2002. Tight security proofs for the bounded-storage model. In *Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montréal, Québec, Canada*, John H. Reif (Ed.). ACM, 341–350. https://doi.org/10.1145/509907.509960

[20] Stefan Dziembowski and Ueli M. Maurer. 2004. On Generating the Initial Key in the Bounded-Storage Model. In *Advances in Cryptology - EUROCRYPT 2004, International Conference on the Theory and Applications of Cryptographic Techniques, Interlaken, Switzerland, May 2-6, 2004, Proceedings (Lecture Notes in Computer Science, Vol. 3027)*, Christian Cachin and Jan Camenisch (Eds.). Springer, 126–137. https://doi.org/10.1007/978-3-540-24676-3_8

[21] Sumegha Garg, Pravesh K. Kothari, Pengda Liu, and Ran Raz. 2021. Memory-Sample Lower Bounds for Learning Parity with Noise. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2021, August 16-18, 2021, University of Washington, Seattle, Washington, USA (Virtual Conference) (LIPIcs, Vol. 207)*, Mary Wootters and Laura Sanità (Eds.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 60:1–60:19. https://doi.org/10.4230/LIPIcs.APPROX/RANDOM.2021.60

[22] Sumegha Garg, Pravesh K. Kothari, and Ran Raz. 2020. Time-Space Tradeoffs for Distinguishing Distributions and Applications to Security of Goldreich's PRG. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2020, August 17-19, 2020, Virtual Conference (LIPIcs, Vol. 176)*, Jaroslaw Byrka and Raghu Meka (Eds.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 21:1–21:18. https://doi.org/10.4230/LIPIcs.APPROX/RANDOM.2020.21

[23] Sumegha Garg, Ran Raz, and Avishay Tal. 2018. Extractor-Based Time-Space Lower Bounds for Learning. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing* (Los Angeles, CA, USA) *(STOC 2018)*. Association for Computing Machinery, New York, NY, USA, 990–1002. https://doi.org/10.1145/3188745.3188962

[24] Dmitry Gavinsky, Julia Kempe, Iordanis Kerenidis, Ran Raz, and Ronald de Wolf. 2008. Exponential Separation for One-Way Quantum Communication Complexity, with Applications to Cryptography. *SIAM J. Comput.* 38, 5 (2008), 1695–1708. https://doi.org/10.1137/070706550

[25] Jiaxin Guan and Mark Zhandry. 2019. Simple Schemes in the Bounded Storage Model. In *Advances in Cryptology - EUROCRYPT 2019 - 38th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Darmstadt, Germany, May 19-23, 2019, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 11478)*, Yuval Ishai and Vincent Rijmen (Eds.). Springer, 500–524. https://doi.org/10.1007/978-3-030-17659-4_17

[26] Jeongwan Haah, Aram W. Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. 2017. Sample-Optimal Tomography of Quantum States. *IEEE Trans. Inf. Theory* 63, 9 (2017), 5628–5641. https://doi.org/10.1109/TIT.2017.2719044

[27] Danny Harnik and Moni Naor. 2006. On Everlasting Security in the *Hybrid* Bounded Storage Model. In *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 4052)*, Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener (Eds.). Springer, 192–203. https://doi.org/10.1007/11787006_17

[28] Hsin-Yuan Huang, Richard Kueng, and John Preskill. 2020. Predicting many properties of a quantum system from very few measurements. *Nature Physics* 16, 10 (2020), 1050–1057. https://doi.org/10.1038/s41567-020-0932-7

[29] Gillat Kol, Ran Raz, and Avishay Tal. 2017. Time-space hardness of learning sparse parities. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, Hamed Hatami, Pierre McKenzie, and Valerie King (Eds.). ACM, 1067–1080. https://doi.org/10.1145/3055399.3055430

[30] Jiahui Liu and Satyanarayana Vusirikala. 2021. Secure Multiparty Computation in the Bounded Storage Model. In *Cryptography and Coding - 18th IMA International Conference, IMACC 2021, Virtual Event, December 14-15, 2021, Proceedings (Lecture Notes in Computer Science, Vol. 13129)*, Maura B. Paterson (Ed.). Springer, 289–325. https://doi.org/10.1007/978-3-030-92641-0_14

[31] Chi-Jen Lu. 2002. Hyper-encryption against Space-Bounded Adversaries from On-Line Strong Extractors. In *Advances in Cryptology - CRYPTO 2002, 22nd Annual International Cryptology Conference, Santa Barbara, California, USA, August 18-22, 2002, Proceedings (Lecture Notes in Computer Science, Vol. 2442)*, Moti Yung (Ed.). Springer, 257–271. https://doi.org/10.1007/3-540-45708-9_17

[32] Ueli M. Maurer. 1992. Conditionally-Perfect Secrecy and a Provably-Secure Randomized Cipher. *J. Cryptol.* 5, 1 (1992), 53–66. https://doi.org/10.1007/BF00191321

[33] Tal Moran, Ronen Shaltiel, and Amnon Ta-Shma. 2009. Non-interactive Timestamping in the Bounded-Storage Model. *J. Cryptol.* 22, 2 (2009), 189–226. https://doi.org/10.1007/s00145-008-9035-9

[34] Dana Moshkovitz and Michal Moshkovitz. 2018. Entropy Samplers and Strong Generic Lower Bounds For Space Bounded Learning. In *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA (LIPIcs, Vol. 94)*, Anna R. Karlin (Ed.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 28:1–28:20. https://doi.org/10.4230/LIPIcs.ITCS.2018.28

[35] S. Pironio, Ll. Masanes, A. Leverrier, and A. Acín. 2013. Security of Device-Independent Quantum Key Distribution in the Bounded-Quantum-Storage Model. *Phys. Rev. X* 3 (Aug 2013), 031007. Issue 3. https://doi.org/10.1103/PhysRevX.3.031007

[36] Ran Raz. 2017. A Time-Space Lower Bound for a Large Class of Learning Problems. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, Chris Umans (Ed.). IEEE Computer Society, 732–742. https://doi.org/10.1109/FOCS.2017.73

[37] Ran Raz. 2018. Fast Learning Requires Good Memory: A Time-Space Lower Bound for Parity Learning. *J. ACM* 66, 1, Article 3 (dec 2018), 18 pages. https://doi.org/10.1145/3186563

[38] Christian Schaffner. 2007. Cryptography in the Bounded-Quantum-Storage Model. arXiv:0709.0289 [quant-ph]

[39] Ohad Shamir. 2014. Fundamental Limits of Online and Distributed Algorithms for Statistical Learning and Estimation. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (Eds.). 163–171.

[40] Vatsal Sharan, Aaron Sidford, and Gregory Valiant. 2019. Memory-sample tradeoffs for linear regression with small error. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019*, Moses Charikar and Edith Cohen (Eds.). ACM, 890–901. https://doi.org/10.1145/3313276.3316403

[41] Jacob Steinhardt, Gregory Valiant, and Stefan Wager. 2016. Memory, Communication, and Statistical Queries. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016 (JMLR Workshop and Conference Proceedings, Vol. 49)*, Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir (Eds.). JMLR.org, 1490–1516.

[42] Stephanie Wehner and Jürg Wullschleger. 2008. Composable Security in the Bounded-Quantum-Storage Model. In *Automata, Languages and Programming, 35th International Colloquium, ICALP 2008, Reykjavik, Iceland, July 7-11, 2008, Proceedings, Part II - Track B: Logic, Semantics, and Theory of Programming & Track C: Security and Cryptography Foundations (Lecture Notes in Computer Science, Vol. 5126)*, Luca Aceto, Ivan Damgård, Leslie Ann Goldberg, Magnús M. Halldórsson, Anna Ingólfsdóttir, and Igor Walukiewicz (Eds.). Springer, 604–615. https://doi.org/10.1007/978-3-540-70583-3_49

[43] John Wright. 2016. *How to learn a quantum state.* Ph. D. Dissertation. Carnegie Mellon University.

Received 2022-11-07; accepted 2023-02-06