Electronic Colloquium on Computational Complexity, Report No. 114 (2018)

Time-Space Tradeoffs for Learning Finite Functions from Random Evaluations, with Applications to Polynomials

Paul Beame*
University of Washington
beame@cs.washington.edu

Shayan Oveis Gharan†
University of Washington
shayan@cs.washington.edu

Xin Yang*
University of Washington
yx1992@cs.washington.edu

June 6, 2018

Abstract

We develop an extension of recent analytic methods for obtaining time-space tradeoff lower bounds for problems of learning from uniformly random labelled examples. With our methods we can obtain bounds for learning concept classes of finite functions from random evaluations even when the sample space of random inputs can be significantly smaller than the concept class of functions and the function values can be from an arbitrary finite set.

At the core of our results, we reduce the time-space complexity of learning from random evaluations to the question of how much the corresponding evaluation matrix amplifies the 2-norms of "almost uniform" probability distributions. To analyze the latter, we formulate it as a semidefinite program, and we analyze its dual. In order to handle function values from arbitrary finite sets, we apply this norm amplification analysis to complex matrices.

As applications that follow from our new techniques, we show that any algorithm that learns n-variate polynomial functions of degree at most d over $\mathbb{F}_2$ with success at least $2^{-O(n)}$ from evaluations on randomly chosen inputs either requires space $\Omega(nm/d)$ or $2^{\Omega(n/d)}$ time where $m = (n/d)^{\Theta(d)}$ is the dimension of the space of such polynomials. These bounds are asymptotically optimal for polynomials of arbitrary constant degree since they match the tradeoffs achieved by natural learning algorithms for the problems. We extend these results to learning polynomials of degree at most d over any odd prime field $\mathbb{F}_p$ where we show that $\Omega((mn/d)\log p)$ space or time $p^{\Omega(n/d)}$ is required.

To derive our bounds for learning polynomials over finite fields, we show that an analysis of the dual of the corresponding semidefinite program follows from an understanding of the distribution of the bias of all degree d polynomials with respect to uniformly random inputs.

*Research supported in part by NSF grant CCF-1524246
†Research supported in part by NSF grant CCF-1552097 and ONR-YI grant N00014-17-1-2429

ISSN 1433-8092

1 Introduction

In supervised learning from labelled examples, the question of the sample complexity required to obtain good generalization has been of considerable interest and research. However, another important parameter is how much information from these samples needs to be kept in memory in order to learn successfully. There has been a line of work improving the memory efficiency of learning algorithms, and the question of the limits of such improvement has begun to be tackled relatively recently. Shamir [15] and Steinhardt, Valiant, and Wager [17] both obtained constraints on the space required for certain learning problems and in the latter paper, the authors asked whether one could obtain strong tradeoffs for learning from random samples that yields a superlinear threshold for the space required for efficient learning. Raz [13] showed that even given exact information, if the space of a learning algorithm is bounded by a sufficiently small quadratic function of the number of input bits, then the problem of online of learning parity functions given exact answers on random samples requires an exponential number of samples even to learn parity functions approximately.

More precisely, we consider problems of online learning from uniform random samples, in which an unknown concept x is chosen uniformly from a set X of (multivalued) concepts and a learner is given a stream of samples $(a^{(1)}, b^{(1)}, (a^{(2)}, b^{(2)}), \cdots$ where each $a^{(t)}$ is chosen uniformly at random from A and $b^{(t)} = L(a^{(t)}, x)$ for labelling function L which maps each pair (a, x) to the outcome (or label) of the value of concept $x \in X$ when given $a \in A$. The learner's goal is either that of identification "find x" or prediction "predict $L(a, x)$ for randomly chosen a with significant advantage over random guessing." In the case of learning parities, $X = A = \{0,1\}^n$ and $L(a, x) = a \cdot x \pmod 2$. With high probability $n+1$ uniformly random samples suffice to span $\{0,1\}^n$ and one can learn parities using Gaussian elimination with $(n+1)^2$ space. Alternatively, an algorithm with only $O(n)$ space can wait for a specific basis of vectors a to appear (for example the standard basis) and store the resulting values; however, this takes $\Omega(2^n)$ time. Raz [13] showed that either $\Omega(n^2)$ space or $2^{\Omega(n)}$ time is essential: even if the space is bounded by $n^2/25$, $2^{\Omega(n)}$ queries are required to learn x correctly with any probability that is $2^{-o(n)}$. In follow-on work, Kol, Raz, and Tal [9] showed that the same lower bound applies even if the input x is sparse.

We can view x as a linear function over $\mathbb{F}_2$, and, from this perspective, parity learning identifies a linear function from evaluations over uniformly random inputs. A natural generalization asks if a similar lower bound exists when we learn higher degree polynomials with bounded space. As a motivating example, consider homogeneous quadratic functions over $\mathbb{F}_2$. Let $m = \binom{n+1}{2}$ and $X = \{0,1\}^m$, which we identify with the space of homogeneous quadratic polynomials in $\mathbb{F}_2[z_1, \ldots, z_n]$ or, equivalently, the space of upper triangular Boolean matrices. This learning algorithm has $A = \{0,1\}^n$ and $X = \{0,1\}^m$, and the learning function $L(a, x) = x(a) = \sum_{i \leqslant j} x_{ij} a_i a_j \bmod 2$, or equivalently $L(a, x) = a^T x a$ when x is viewed as an upper triangular matrix.

Given $a \in \{0,1\}^n$ and $x \in \{0,1\}^m$, we can view evaluating $x(a)$ as computing $aa^T \cdot x \bmod 2$ where we interpret aa^T as an element of $\{0,1\}^m$. For $O(m)$ randomly chosen $a \in \{0,1\}^n$, the vectors aa^T almost surely span $\{0,1\}^m$ and hence we can store m samples of the form (a, b) and apply Gaussian elimination to determine x. This time, we only need $n+1$ bits to store each sample for a total space bound of $O(nm)$. An alternative algorithm using $O(m)$ space and time $2^{O(n)}$ would be to look for a specific basis such as the basis consisting of the upper triangular parts of $\{e_i e_i^T \mid 1 \leqslant i \leqslant n\} \cup \{(e_i + e_j)(e_i + e_j)^T \mid 1 \leqslant i < j \leqslant n\}$. Previous lower bounds for learning do not apply to this problem[1] because $|A| \ll |X|$. Our results imply that this tradeoff between $\Omega(nm)$ space or $2^{\Omega(n)}$ time is inherently required to learn x with probability $2^{-o(n)}$ or predict its output with at least $2^{-o(n)}$ advantage.

The techniques in [13] and [9] were based on fairly ad-hoc simulations of the original space-bounded learning algorithm by a restricted form of linear branching program for which one can measure progress at learning x using the dimension of the consistent subspace. More recent papers, by Moshkovitz and Moshkovitz [11] using graph mixing properties and by Raz [14] using an analytic approach, considered more general tests and used a measure of progress based on 2-norms. While the method of [11] was not strong enough to reproduce the bound in [13] for the case of parity learning, the methods of Raz [14] and the later improvement [12] by Moshkovitz and Moshkovitz of [11] reproduced the parity learning bound and more.

In particular, Raz [14] defined a ± 1 matrix M that is indexed by $A \times X$. It is natural to see $M(a, x)$ as $(-1)^{L(a,x)}$ for a labelling function L that has labels in $\{0, 1\}$. The lower bound is governed by the (expectation) matrix norm of M, which is a function of the largest singular value of M, and the progress is analyzed by bounding the impact of applying the matrix to probability distributions with small expectation 2-norm. This method works if $|A| \geqslant |X|$ - i.e., the sample space of inputs is at least as large as the concept class - but it fails completely if $|A| \ll |X|$, which is precisely the situation for learning quadratic functions. Indeed, none of the prior approaches works in this case.

In our work we extend the analytic approach to capture *general* discrete problems of learning from uniform random samples in which (1) the sample space of inputs can be much smaller than the concept class and (2) members of the concept class can have values from an arbitrary finite set, which we identify with $\{0, 1, \ldots, r\}$ for convenience. Our extensions come from two different directions.

We define a property of matrices M that allows us to refine the notion of the largest singular value and extend the method of Raz [14] to the cases that $|A| \ll |X|$. This property, which we call the *norm amplification curve* of the matrix on the positive orthant, analyzes more precisely how $\|M \cdot p\|_2$ grows as a function of $\|p\|_2$ for probability vectors p on X. The key reason that this is not simply governed by the singular values is that the interior of the positive orthant can contain at most one singular vector. We give a simple condition on the 2-norm amplification curve of M that is sufficient to ensure that there is a time-space tradeoff showing that any learning algorithm for M with success probability at least $2^{-\varepsilon n}$ for some $\varepsilon > 0$ either requires space $\Omega(mn)$ or time $2^{\Omega(n)}$.

For any fixed learning problem given by a matrix M, the natural way to express the amplification curve at any particular value of $\|p\|_2$ yields an optimization problem given by a quadratic program with constraints on $\|p\|_2^2$, $\|p\|_1$ and $p \geqslant 0$, and with objective function $\|Mp\|_2^2 = \langle M^T M, pp^T \rangle$ that seems difficult to solve. Instead, we relax the quadratic program to a semi-definite program where we replace pp^T by a positive semidefinite matrix U with the analogous constraints. We can then obtain an upper bound on the amplification curve by moving to the SDP dual and evaluating the dual objective at a particular Laplacian determined by the properties of $M^T M$.

In order to handle concepts that are more than binary-valued[2], we move to matrices whose

[1] Note that in [9] the lower bound applies in a dual case when the unknown x is sparse, and hence $|X| \ll |A|$.

[2] The formalization of Moshkovitz and Moshovitz [11, 12] does include the case of multivalued outcomes, though

entries are complex r-th roots of unity. Indeed, a single matrix M does not suffice for $r > 3$ and we instead work with a family of complex matrices $M^{(1)}, \ldots, M^{(r-1)}$, each associated with a different root of unity. We use the natural generalization of the norm amplification curve to complex matrices and also generalize the semi-definite relaxation method to bound these curves using $(M^{(j)})^* M^{(j)}$ instead of $M^T M$. We then show how the overall analytic approach can be carried through with a modest number of changes from the binary-valued case.

Our lower bound shows that if the 2-norm amplification curve for M has (or, in the case of r-valued labels, matrices $M^{(1)}, \ldots, M^{(r-1)}$ have) the required property, then to achieve learning success probability for M of at least $|A|^{-\varepsilon}$ for some $\varepsilon > 0$, either space $\Omega(\log |A| \cdot \log_r |X|)$ or time $|A|^{\Omega(1)}$ is required. This matches the natural upper bounds asymptotically over a wide range of learning problems.

As applications, we focus on problems of learning polynomials of varying degrees over finite fields. For matrices M associated with polynomials over $\mathbb{F}_2$, the property of the matrices $M^T M$ required to bound the amplication curves for M correspond precisely to properties of the weight distribution of Reed-Muller codes over $\mathbb{F}_2$. In the case of quadratic polynomials, this weight distribution is known exactly. In the case of higher degree polynomials, bounds on the weight distribution of such codes, or more precisely on the bounds on the bias of random degree d polynomials over $\mathbb{F}_2$ of Ben-Eliezer, Hod, and Lovett [3] are sufficient to let us show that learning polynomials of degree at most d over $\mathbb{F}_2^n$ from random inputs with probability $2^{-\Omega(n/d)}$ either requires space $\Omega(nm/d)$ or time $2^{\Omega(n/d)}$.

We also analyze learning problems for the case of prime fields $\mathbb{F}_p$ for p odd using our multi-valued techniques involving complex matrices. For $\mathbb{F}_p$, even the cases of linear and affine polynomials are new. We relate the norm amplification curves of the associated matrices to bounds on the bias of random degree d polynomials over $\mathbb{F}_p$. We also give a precise analysis of the bias of the set of quadratic polynomials over $\mathbb{F}_p^n$ to derive tight time-space tradeoff lower bounds for learning them. For larger degrees we apply bounds on the bias that we recently proved in a companion paper ([2]).

Independent of the specific applications to learning from random examples that we obtain, the measures of matrices that we introduce, the 2-norm amplification curve on the positive orthant, and semi-definite relaxation approach seem likely to have significant applications in other contexts outside of learning.

Related work: Independently and contemporaneously with our preliminary version ([1]), Garg, Raz, and Tal [6] proved closely related results to ours for the case of binary labels. The fundamental techniques are similarly grounded in the approach of [14]. At the very high-level, they prove very similar structural properties of the matrix M, namely, they show that it is an "L_2 two-source extractor" which can be seen to be equivalent to bounding our norm amplification curve for learning matrices. More precisely, their "almost orthogonality property" essentially corresponds to upper bounding $W_\kappa(M^* M)$ for some threshold κ (see Definition 6.1 and Lemma 6.2). However, since we use duality explicitly, our proof seems more amenable to extensions, particularly, when we have more structure in the learning matrix M. Subsequently ([7]), they were also able to allow

they do not apply it to any examples and their mixing condition does not hold in the case of small input sample spaces multivalued labels by extending the extractor approach to permit correlations between the sample inputs and the concept.

2 Branching programs for learning

In order to be able to solve the learning problem given concept class X, sample space of inputs A and labelling function L on $A \times X$ exactly we require that the learning function L have the property that for all $x \neq x' \in X$ there exists an $a \in A$ such that $L(a, x) \neq L(a, x')$. Note that the set $\{0, 1, \ldots, r-1\}$ of labels allows us to model any learning situation in which r different labels are possible.

Following [13], the time and space of a learner are modelled simultaneously by expressing the learner's computation as a layered branching program: a finite rooted directed acyclic multigraph with every non-sink node having outdegree $r|A|$, with one outedge for each (a, b) with $a \in A$ and $b \in \{0, 1, \ldots, r-1\}$ that leads to a node in the next layer. Each sink node v is labelled by some $x' \in X$ which is the learner's guess of the value of the concept x. (In the case of prediction we allow the sink label to be an arbitrary function from A to $\{0, 1, \ldots, r-1\}$ denoting the best prediction of the algorithm for each $a \in A$.)

The space S used by the learning branching program is the $\log_2$ of the maximum number of nodes in any layer and the time T is the length of the longest path from the root to a sink.

The samples given to the learner $(a^{(1)}, b^{(1)}), (a^{(2)}, b^{(2)}), \ldots$ based on uniformly randomly chosen $a^{(1)}, a^{(2)}, \ldots \in A$ and a concept $x \in X$ determines a (randomly chosen) *computation* path in the branching program. When we consider computation paths we include the concept x in their description. The (expected) success probability of the learner is the probability for a uniformly random concept $x \in X$ that a random computation path given concept x reaches a sink node v with label $x' = x$ (or with sufficiently good predictions on randomly chosen $a \in A$).

Progress towards identification Following [11] and [14] we measure progress towards identifying $x \in X$ using the "expectation 2-norm" over the uniform distribution: For any set S, and $f : S \to \mathbb{R}$, define $\|f\|_2 = \left(\mathbb{E}_{s \in_R S} f^2(s)\right)^{1/2} = (\sum_{s \in S} f^2(s)/|S|)^{1/2}$. Define Δ_X to be the space of probability distributions on X. Consider the two extremes for the expectation 2-norm of elements of Δ_X: If $\mathbb{P}$ is the uniform distribution on X, then $\|\mathbb{P}\|_2 = |X|^{-1}$. This distribution represents the learner's knowledge of the concept x at the start of the branching program. On the other hand if $\mathbb{P}$ is point distribution on any x', then $\|\mathbb{P}\|_2 = |X|^{-1/2}$.

For each node v in the branching program, there is an induced probability distribution on X, $\mathbb{P}'_{x|v}$ which represents the distribution on X conditioned on the fact that the computation path passes through v. It represents the learner's knowledge of x at the time that the computation path has reached v. Intuitively, the learner has made significant progress towards identifying the concept x if $\|\mathbb{P}'_{x|v}\|_2$ is much larger than $|X|^{-1}$, say $\|\mathbb{P}'_{x|v}\|_2 \geqslant |X|^{\delta/2} \cdot |X|^{-1} = |X|^{-(1-\delta/2)}$.

The general idea will be to argue that for any fixed node v in the branching program that is at a layer t that is $|A|^{o(1)}$, the probability over a randomly chosen computation path that v is the first node on the path for which the learner has made significant progress is $|X|^{-\Omega(\log_r |A|)}$. Since by assumption of correctness the learner makes significant progress with at least $|X|^{-\varepsilon}$ probability, there must be at least $|X|^{\Omega(\log_r |A|)}$ such nodes and hence the space must be $\Omega(\log |X| \log_r |A|)$.

Given that we want to consider the first vertex on a computation path at which significant progress has been made, it is natural to truncate a computation path at v if significant progress has been already been made at v (and then one should not count any path through v towards the progress at some subsequent node w). Following [14], for technical reasons we will also truncate the computation path in other circumstances: if the concept x has too high probability at v, or if the next edge is labelled by a pair (a, b) for which the value on input a of random concepts whose computation path reaches v is significantly biased away from the uniform distribution on $\{0, 1 \ldots, r-1\}$.

Like Raz [14], we use an analytic approach to understanding the progress and the bias. In [14], only binary feedback is possible and progress is analyzed in terms of the matrix properties of a learning matrix M given by $M(a, x) = (-1)^{L(a,x)}$, which is viewed as the learning problem specification. This form is particularly convenient since it allows one to represent the predictability of outcomes under a distribution $\mathbb{P}$ on X in terms of a matrix vector product. That is, $(M \cdot \mathbb{P})(a) = \mathbf{E}_{x \sim \mathbb{P}}[(-1)^{L(a,x)}]$ is the expected bias of a concept distributed according to $\mathbb{P}$ on input a.

It would be natural to try to extend this analytic approach for $r > 2$ by replacing $(-1)^{L(a,x)}$ by $\omega^{L(a,x)}$ where $\omega = e^{2\pi i/r}$ is a primitive r-th root of unity. However, for $r > 3$, simply having $\mathbf{E}_{x \in_R X}[\omega^{f(x)}]$ small does not imply that f is close to uniformly distributed on $\{0, 1, \ldots, r-1\}$. Nonetheless, by setting $g_k = \mathbf{Pr}_{x \in_R X}[f(x) = k]$ we can apply the following proposition, which is a standard method using exponential sums, to show that bounding $|\mathbf{E}_{x \in_R X}[\omega^{j \cdot f(x)}]|$ for all $j \in \{1, \ldots, r-1\}$ is sufficient to show that f is close to uniformly distributed. We include its proof for completeness.

Proposition 2.1. *Suppose that $\sum_{k=0}^{r-1} g_k = 1$ and define $g(z) = \sum_{k=0}^{r-1} g_k z^k$. If $|g(\omega^j)| < \varepsilon$ for all $j \in \{1, \ldots, r-1\}$ then for all $k \in \{0, 1, \ldots, r-1\}$, $|g_k - \frac{1}{r}| \leqslant \varepsilon$.*

Proof. Write $s(z) = \sum_{k=0}^{r-1} z^k$ and observe that $s(1) = r$ but $s(\omega^j) = 0$ for all $j \in \{1, \ldots, r-1\}$. Define $h(z) = g(z) - \frac{1}{r} \cdot s(z)$. Observe that $h(z) = \sum_{k=0}^{r-1} h_k z^k$ where $h_k = g_k - \frac{1}{r}$ so it suffices to prove that $|h_k| \leqslant \varepsilon$ for all $k \in \{0, 1, \ldots, r\}$. Note that $h(1) = 0$, and $h(\omega^j) = g(\omega^j)$ for all $j \in \{1, \ldots, r-1\}$. Therefore $|h(\omega^j)| \leqslant \varepsilon$ for all $j \in \{0, 1 \ldots, r-1\}$. If we let $\mathbf{h} = (h_0, \ldots, h_{r-1})^T$ be the vector of coefficients and $\mathbf{v} = (h(1), h(\omega), \ldots, h(\omega^{r-1}))^T$ be the vector of values of h, we have $V(\omega) \cdot \mathbf{h} = \mathbf{v}$ where $V(\omega)$ is the usual $r \times r$ Vandermonde matrix for the powers of ω. Now $V(\omega)^{-1} = V(\omega^{-1})/r$, the matrix of the discrete Fourier transform; so in particular, for every $k \in \{0, 1, \ldots, r-1\}$,

$$h_k = \frac{1}{r} \cdot \sum_{j=0}^{r-1} \omega^{-jk} v_j.$$

Hence $|h_k| \leqslant \frac{1}{r} \cdot \sum_{j=0}^{r-1} |v_j| \leqslant \varepsilon$ since every $|v_j| \leqslant \varepsilon$, which suffices to prove the proposition. □

Therefore, instead of the single ± matrix M given by $M(a, x) = (-1)^{L(a,x)}$, we will analyze the learning problem given by L using $r - 1$ different[3] complex matrices $M^{(j)} \in \mathbb{C}^{A \times X}$ for $j \in \{1, \ldots, r-1\}$ given by $M^{(j)}(a, x) = \omega^{j \cdot L(a,x)}$. We now define the probability distributions and truncation process for computation paths inductively as follows:

[3] In Proposition 2.1 one can observe that $|g(\omega^j)| = |g(\omega^{r-j})|$ so $\lceil (r-1)/2 \rceil$ matrices suffice, but we find it convenient to argue using all $r - 1$ matrices; however, this does imply that a single matrix suffices when $r = 3$.

Definition 2.2. *We define probability distributions $\mathbb{P}_{x|v} \in \Delta_X$ and the (δ, α, γ)-truncation of the computation paths inductively as follows:*

- *If v is the root, then $\mathbb{P}_{x|v}$ is the uniform distribution on X.*
- *(Significant Progress) If $\|\mathbb{P}_{x|v}\|_2 \geqslant |X|^{-(1-\delta/2)}$ then truncate all computation paths at v. We call vertex v* significant *in this case.*
- *(High Probability) Truncate the computation paths at v for all concepts x' for which $\mathbb{P}_{x|v}(x') \geqslant |X|^{-\alpha}$. Let* $\mathrm{High}(v)$ *be the set of such concepts.*
- *(High Bias) Truncate any computation path at v if it follows an outedge e of v with label (a, b) for which $|(M^{(j)} \cdot \mathbb{P}_{x|v})(a)| \geqslant |A|^{-\gamma}$ for some $j \in \{1, \ldots, r-1\}$. That is, we truncate the paths at v if the label outcome for the next sample for $a \in A$ is too predictable given the knowledge that the path was not truncated previously and arrived at v.*
- *If v is not the root then define $\mathbb{P}_{x|v}$ to be the conditional probability distribution on x over all computation paths that have not previously been truncated and arrive at v.*

For an edge $e = (v, w)$ of the branching program, we also define a probability distribution $\mathbb{P}_{x|e} \in \Delta_X$, which is the conditional probability distribution on X induced by the truncated computation paths that pass through edge e.

With this definition, it is no longer immediate from the assumption of correctness that the truncated path reaches a significant node with at least $|A|^{-\varepsilon}$ probability. However, we will see that a single assumption about the matrices $M^{(j)}$ will be sufficient to prove both that this holds and that the probability is $|X|^{-\log_r |A|}$ that the path reaches any specific node v at which significant progress has been made.

3 Norm amplification by matrices on the positive orthant

By definition, for $\mathbb{P} \in \Delta_X$, and $M \in \mathbb{C}^{A \times X}$, $\|M \cdot \mathbb{P}\|_2^2 = \mathbf{E}_{a \in_R A}[|(M \cdot \mathbb{P})(a)|^2]$. Observe that for $\mathbb{P} = \mathbb{P}_{x|v}$ and $M = M^{(j)}$ for $j \in \{1, \ldots, r-1\}$, the values $|(M^{(j)} \cdot \mathbb{P}_{x|v})(a)|$ are the quantities that we test to determine whether an edge labelled a is a high bias edge that causes the truncation of the computation path. Therefore $\|M^{(j)} \cdot \mathbb{P}_{x|v}\|_2^2$ is the expected square of this bias value for uniformly random inputs at v.

If we have not learned the concept x, we would not expect to be able to predict its value on a random input; moreover, since any path that would follow a high bias input is truncated, it is essential to argue that $\|M^{(j)} \cdot \mathbb{P}_{x|v}\|_2$ remains small at any node v where there has not been significant progress.

In [14] there is a single ± 1 matrix M and $\|M \cdot \mathbb{P}_{x|v}\|_2$ is bounded using the matrix norm $\|M\|_2$ given by $\|M\|_2 = \sup_{\substack{f: X \to \mathbb{R} \\ f \neq 0}} \|M \cdot f\|_2 / \|f\|_2$, where the numerator is an expectation 2-norm over A and the denominator is an expectation 2-norm over X. Thus $\|M\|_2 = \sqrt{|X|/|A|} \cdot \sigma_{\max}(M)$, where $\sigma_{\max}(M)$ is the largest singular value of M and $\sqrt{|X|/|A|}$ is a normalization factor.

In the case of the matrix M associated with parity learning, $|A| = |X| = 2^n$ and all the singular values are equal to $\sqrt{|X|}$ so $\|M\|_2 = \sqrt{|X|} = 2^{n/2}$. With this bound, if v is not a node of significant progress then $\|\mathbb{P}_{x|v}\|_2 \leqslant 2^{-(1-\delta/2)n}$ and hence $\|M \cdot \mathbb{P}_{x|v}\|_2 \leqslant 2^{-(1-\delta)n/2}$ which is $1/|A|^{(1-\delta)/2}$ and hence small.

However, even in the case of learning quadratic functions over $\mathbb{F}_2$, the largest singular value of the matrix M is still $\sqrt{|X|}$ (the uniform distribution on X is a singular vector) and so $\|M\|_2 = |X|/\sqrt{|A|}$. But in that case, when $\|\mathbb{P}_{x|v}\|_2$ is $|X|^{-(1-\delta/2)}$ we conclude that $\|M\|_2 \cdot \|\mathbb{P}_{x|v}\|_2$ is at most $|X|^{\delta/2}/\sqrt{|A|}$ which is much larger than 1 and hence a useless bound on $\|M \cdot \mathbb{P}_{x|v}\|_2$.

Indeed, the same kind of problem occurs in using the method of Raz [14] for any learning problem for which $|A|$ is $|X|^{o(1)}$: If v is a child of the root of the branching program at which the more likely outcome b of a single randomly chosen input $a \in A$ is remembered, then $\|\mathbb{P}_{x|v}\|_2 \leqslant \sqrt{2}/|X|$. However, in this case $|(M \cdot \mathbb{P}_{x|v})(a)| = 1$ and so $\|(M \cdot \mathbb{P}_{x|v})\|_2 \geqslant |A|^{-1/2}$. It follows that $\|M\|_2 \geqslant |X|/(2|A|)^{1/2}$ and when $|A|$ is $|X|^{o(1)}$ the derived upper bound on $\|M \cdot \mathbb{P}_{x|v'}\|_2$ at nodes v' where $\|\mathbb{P}_{x|v'}\|_2 \geqslant 1/|X|^{1-\delta/2}$ will be larger than 1 and therefore useless.

We need a more precise way to bound $\|M \cdot \mathbb{P}\|_2$ as a function of $\|\mathbb{P}\|_2$ than using the single number $\|M\|_2$. To do this we will need to use the fact that $\mathbb{P} \in \Delta_X$ – it has a fixed ℓ_1 norm and (more importantly) it is non-negative and therefore lies in the positive orthant.

Definition 3.1. *For $M \in \mathbb{C}^{A\times X}$ the 2-*norm amplification curve *of M, $\tau_M : [0,1] \to \mathbb{R}$ is given by*

$$\tau_M(\delta) = \sup_{\substack{\mathbb{P}\in\Delta_X \\ \|\mathbb{P}\|_2 \leqslant 1/|X|^{1-\delta/2}}} \log_{|A|}(\|M \cdot \mathbb{P}\|_2).$$

In other words, whenever $\|\mathbb{P}\|_2$ is at most $|X|^{-(1-\delta/2)}$, $\|M \cdot \mathbb{P}\|_2$ is at most $|A|^{\tau_M(\delta)}$. To prove our lower bounds we will bound the norm amplification curves $\tau_{M^{(j)}}$ for all $j \in \{1, \ldots, r-1\}$.

4 Theorems

Our general lower bound for learning problems over arbitrary finite label sets is given by following theorem.

Theorem 4.1. *There are constants $c_1, c_2, c_3 > 0$ such that the follow holds. Let $L : A \times X \to \{0, 1, \ldots, r-1\}$ be a labelling function and for $j = 1, \cdots, r-1$ define the matix $M^{(j)} \in \mathbb{C}^{A\times X}$ by $M^{(j)}(a, x) = \omega^{j\cdot L(a,x)}$ where $\omega = e^{2\pi i/r}$ and assume*[4] *that $|A| \leqslant |X|$. Suppose that for $0 < \delta' < 1$ we have $\tau_{M^{(j)}}(\delta') \leqslant -\gamma' < 0$ for all $j \in \{1, \cdots, r-1\}$. Then, for $\varepsilon \geqslant c_1 \min(\delta', \gamma') > 0$, $\beta \geqslant c_2 \min(\delta', \gamma') > 0$, and $\eta \geqslant c_3 \delta' \gamma' > 0$, any algorithm that solves the learning problem for L with success probability at least $|A|^{-\varepsilon}$ or advantage $\geqslant |A|^{-\varepsilon/2}$ either requires space at least $\eta \log_2 |A| \log_r |X|$ or time at least $|A|^{\beta}$.*

Applications to learning polynomials There are many potential applications of the above theorem but for this paper we focus learning polynomials from their evaluations over finite fields of various sizes. The bounds are derived using the semidefinite programming approach given in Section 6 together with analyses for polynomials given in Section 7.

[4] We could write the statement of the theorem to apply to all A and X by replacing each occurrence of $|A|$ in the lower bounds with $\min(|A|, |X|)$. When $|A| \geqslant |X|$ and $r = 2$, we can use $\|M\|_2$ to bound $\tau_M(\delta')$ which yields the bound given in [14]

Learning polynomials over $\mathbb{F}_2$ We first consider the case of polynomials over $\mathbb{F}_2$ which yield a binary labelling set. In this case $\omega = -1$ and there is only one matrix M, whose entries are $M(a,x) = (-1)^{L(a,x)}$ as in [14].

The case of linear functions over $\mathbb{F}_2$ is just the parity learning problem. For learning higher degree polynomials over $\mathbb{F}_2$ we obtain the following bounds on the norm amplification curves of their associated matrices:

Theorem 4.2. *The following norm amplification bounds hold:*

(a) *For all $\delta \in [0,1]$, the matrix M for learning quadratic functions over $\mathbb{F}_2^n$ satisfies*
$$\tau_M(\delta) \leqslant \tfrac{-(1-\delta)}{8} + \tfrac{5+\delta}{8n}.$$

(b) *For any $\zeta > 0$, there are constants δ, γ with $0 < \delta < 1/2$ and $\gamma > 0$ such that for $d \leqslant (1-\zeta)n$ the matrix M for learning functions of degree $\leqslant d$ over $\mathbb{F}_2^n$ satisfies $\tau_M(\delta) \leqslant -\gamma/d$.*

Theorem 4.2 is proved in Section 7. The case for quadratic polynomials over $\mathbb{F}_2$ follows from properties of the weight distribution of Reed-Muller codes $RM(n,2)$ shown by [16] and [10]. The case for higher degree polynomials over $\mathbb{F}_2$ follows from tail bounds on the bias of $\mathbb{F}_2$ polynomials given by [3].

Using these bounds together with Theorem 4.1 yields the following:

Theorem 4.3. *There are constants $\varepsilon, \zeta > 0$ such that the following hold:*

(a) *Let $m = \binom{n+1}{2}$ for positive integer n. Any algorithm for learning quadratic functions over $\mathbb{F}_2^n$ that succeeds with probability at least $2^{-\varepsilon n}$ requires space $\Omega(nm)$ or time $2^{\Omega(n)}$.*

(b) *Let $n > 0$ and $d > 0$ be integers such that $d \leqslant (1-\zeta) \cdot n$ and let $m = \sum_{i=0}^{d} \binom{n}{i}$. Any algorithm for learning polynomial functions of degree at most d over $\mathbb{F}_2^n$ that succeeds with probability at least $2^{-\varepsilon n/d}$ requires space $\Omega(nm/d)$ or time $2^{\Omega(n/d)}$.*

These bounds are tight for constant d since they match the resources used by the natural learning algorithms described in the introduction up to constant factors in the space bound and in the exponent of the time bound.

Learning polynomials over $\mathbb{F}_p$ for odd prime p. The following theorem bounds the norm amplification curves for polynomials of various degrees over odd prime fields.

Theorem 4.4. *Let p be an odd prime. For all $\delta \in (0,1)$ and for all $j \in \mathbb{F}_p^*$,*

(a) *the matrices $M^{(j)}$ for learning linear functions over $\mathbb{F}_p^n$ satisfy $\tau_{M^{(j)}}(\delta) \leqslant -\frac{1-\delta}{2}$,*

(b) *the matrices $M^{(j)}$ for learning affine functions over $\mathbb{F}_p^n$ satisfy $\tau_{M^{(j)}}(\delta) \leqslant -\frac{1-\delta}{2} + \frac{\delta}{2n}$,*

(c) *the matrices $M^{(j)}$ for learning quadratic functions over $\mathbb{F}_p^n$ satisfy $\tau_{M^{(j)}}(\delta) \leqslant \frac{-(1-\delta)}{4} + \frac{2}{n}$, and*

(d) *for any $0 < \zeta < 1/2$, there are δ, γ with $0 < \delta < 1/2$ and $\gamma > 0$ such that for $d \leqslant \zeta n$, the matrices $M^{(j)}$ for learning functions of degree $\leqslant d$ over $\mathbb{F}_p^n$ satisfy $\tau_{M^{(j)}}(\delta) \leqslant -\gamma/d$.*

The proof of Theorem 4.4 is in Section 7. Parts (a) and (b) are almost immediate. The proof of part (c) involves a tight structural characterization of quadratic polynomials over $\mathbb{F}_p$. The proof of part (d) for $d \geqslant 3$ uses tail bounds on the bias of polynomials of degree at most d over $\mathbb{F}_p$ recently proved by the authors in a companion paper ([2]).

Using the bounds on the norm amplification curves of Theorem 4.4 together with Theorem 4.1, we immediately obtain the time-space tradeoff lower bounds in following theorem.

Theorem 4.5. *Let p be an odd prime. There is an $\varepsilon > 0$ such that the following hold:*

(a) Any algorithm for learning linear or affine functions over $\mathbb{F}_p^n$ from their evaluations that succeeds with probability at least $p^{-\varepsilon n}$ requires time $p^{\Omega(n)}$ or space $\Omega(n^2 \log p)$.

(b) Let $m = \binom{n+2}{2}$. Any algorithm for learning quadratic functions over $\mathbb{F}_p^n$ that succeeds with probability at least $p^{-\varepsilon n}$ requires space $\Omega(nm \log p)$ or time $p^{\Omega(n)}$.

(c) There are constants $\zeta, \varepsilon > 0$ such that for $3 \leqslant d \leqslant (1 - \zeta) \cdot n$ and for m equal to the number of monomials of degree at most d over $\mathbb{F}_p^n$, any algorithm for learning polynomial functions of degree at most d over $\mathbb{F}_p^n$ that succeeds with probability at least $p^{-\varepsilon n/d}$ requires space $\Omega(\log p \cdot nm/d)$ or time $p^{\Omega(n/d)}$.

5 Lower Bounds for Learning Finite Functions from Random Samples

In this section we prove Theorem 4.1. Let $0 < \delta' < 1$ be the value given in the statement of the theorem. To do this we define several positive quantities based on δ' that will be useful:

- $\delta = \delta'/6$,
- $\alpha = 1 - 2\delta$,
- $\gamma = \min_j\{-\tau_{M^{(j)}}(\delta')/2\}$,
- $\beta = \min(\gamma, \delta)/8$, and
- $\varepsilon = \beta/2$.

Let B be a learning branching program for L with length at most $|A|^\beta - 1$ and success probability at least $|A|^{-\varepsilon}$ of identifying the concept (or producing a prediction advantage of more than $|A|^{-\varepsilon/2}$).

We will prove that B must have width $|X|^{\Omega(\delta\gamma \, \log_r |A|)}$. We first apply the (δ, α, γ)-truncation procedure given in Definition 2.2 to yield $\mathbb{P}_{x|v}$ and $\mathbb{P}_{e|v}$ for all vertices v in B.

The following simple technical lemmas are analogues of ones proved in [14], though we structure our argument somewhat differently. The first uses the bound on the amplification curve of the matrices $M^{(j)}$ for $j \in [r-1]$ in place of its matrix norm.

Lemma 5.1. *Suppose that vertex v in B is not significant. Then*

$$\mathbf{Pr}_{a \in_R A}[\exists j \in \{1, \cdots, r-1\}, |(M^{(j)} \cdot \mathbb{P}_{x|v})(a)| \geqslant |A|^{-\gamma}] \leqslant (r-1) \cdot |A|^{-2\gamma}.$$

Proof. Since v is not significant $\|\mathbb{P}_{x|v}\|_2 \leqslant |X|^{-(1-\delta/2)}$. For fixed $j \in \{1, \cdots, r-1\}$, by definition of $\tau_{M^{(j)}}$,

$$\mathbf{E}_{a \in_R A}[|(M^{(j)} \cdot \mathbb{P}_{x|v})(a)|^2] = \|M^{(j)} \cdot \mathbb{P}_{x|v}\|_2^2 \leqslant |A|^{2\tau_{M^{(j)}}(\delta)} \leqslant |A|^{2\tau_{M^{(j)}}(\delta')} \leqslant |A|^{-4\gamma}.$$

Therefore, by Markov's inequality,

$$\mathbf{Pr}_{a \in_R A}[|(M^{(j)} \cdot \mathbb{P}_{x|v})(a)| \geqslant |A|^{-\gamma}] = \mathbf{Pr}_{a \in_R A}[|(M^{(j)} \cdot \mathbb{P}_{x|v})(a)|^2 \geqslant |A|^{-2\gamma}] \leqslant |A|^{-2\gamma}.$$

Hence by a union bound,

$$\mathbf{Pr}_{a \in_R A}[\exists j \in \{1, \cdots, r-1\}, |(M^{(j)} \cdot \mathbb{P}_{x|v})(a)| \geqslant |A|^{-\gamma}] \leqslant (r-1) \cdot |A|^{-2\gamma}.$$

□

The second is trivial in the case that $r = 2$ but requires a proof for larger r.

Lemma 5.2. *Suppose that vertex v in B is not significant and that $a \in A$ has the property that for all $j \in [r-1]$, $|(M^{(j)} \cdot \mathbb{P}_{x|v})(a)| < |A|^{-\gamma}$. Then for all $b \in \{0, 1, \ldots, r-1\}$,*

$$\left|\mathbf{Pr}_{x' \sim \mathbb{P}_{x|v}}[L(a, x') = b] - \tfrac{1}{r}\right| \leqslant |A|^{-\gamma}.$$

Proof. We apply Proposition 2.1: For $b \in \{0, 1, \ldots, r-1\}$, write $g_b = \mathbf{Pr}_{x' \sim \mathbb{P}_{x|v}}[L(a, x') = b]$ and define $g(z) = \sum_{b=0}^{r-1} g_b z^b$. Observe that for $j \in \{1, \ldots, r-1\}$,

$$\begin{aligned} g(\omega^j) &= \sum_{b=0}^{r-1} \mathbf{Pr}_{x' \sim \mathbb{P}_{x|v}}[L(a, x') = b] \cdot \omega^{jb} \\ &= \sum_{b=0}^{r-1} \sum_{x' \in X} \mathbb{P}_{x|v}(x') \cdot \mathbf{1}_{L(a,x')=b} \cdot \omega^{jb} \\ &= \sum_{x' \in X} \mathbb{P}_{x|v}(x') \cdot \sum_{b=0}^{r-1} \mathbf{1}_{L(a,x')=b} \cdot \omega^{jb} \\ &= \sum_{x' \in X} \mathbb{P}_{x|v}(x') \cdot \omega^{j \cdot L(a,x')} \\ &= (M^{(j)} \cdot \mathbb{P}_{x|v})(a). \end{aligned}$$

Therefore $|g(\omega^j)| = |(M^{(j)} \cdot \mathbb{P}_{x|v})(a)| < A^{-\gamma}$. Applying Proposition 2.1 immediately yields the lemma. □

Lemma 5.3. *Suppose that vertex v in B is not significant. Then*

$$\mathbf{Pr}_{x' \sim \mathbb{P}_{x|v}}[x' \in \mathrm{High}(v)] \leqslant |X|^{-\delta}.$$

Proof. Since v is not significant,

$$\mathbf{E}_{x' \sim \mathbb{P}_{x|v}}[\mathbb{P}_{x|v}(x')] = \sum_{x' \in X} (\mathbb{P}_{x|v}(x'))^2 = |X| \cdot \|\mathbb{P}_{x|v}\|_2^2 \leqslant |X|^{-(1-\delta)} = |X|^{-(\alpha+\delta)}.$$

Therefore since $\alpha = 1 - 2\delta$, by Markov's inequality,

$$\mathbf{Pr}_{x' \sim \mathbb{P}_{x|v}}[x' \in \mathrm{High}(v)] = \mathbf{Pr}_{x' \sim \mathbb{P}_{x|v}}[\mathbb{P}_{x|v}(x') \geqslant |X|^{-\alpha}] \leqslant |X|^{-\delta}.$$

□

Lemma 5.4. *The probability, over uniformly random $x' \in X$ and uniformly random computation path C in B given concept x', that the truncated version T of C reaches a significant vertex of B is at least $\frac{1}{2}|A|^{-\beta/2}$.*

Proof. Let x' be chosen uniformly at random from X and consider the truncated path T. T will not reach a significant vertex of B only if one of the following holds:

1. T is truncated at a vertex v where $\mathbb{P}_{x|v}(x') \geqslant |X|^{-\alpha}$.
2. T is truncated at a vertex v because the next edge of C is labeled by (a, b) where $|(M^{(j)} \cdot \mathbb{P}_{x|v})(a)| \geqslant |A|^{-\gamma}$ for some $j \in \{1, \cdots, r-1\}$.
3. T ends at a leaf that is not significant.

By Lemma 5.3, for each vertex v on C, conditioned on the truncated path reaching v, the probability that $\mathbb{P}_{x|v}(x') \geqslant |X|^{-\alpha}$ is at most $|X|^{-\delta}$. Similarly, by Lemma 5.1, for each v on the path, conditioned on the truncated path reaching v, the probability that $|(M^{(j)} \cdot \mathbb{P}_{x|v})(a)| \geqslant |A|^{-\gamma}$ for any $j \in [r-1]$ is at most $(r-1) \cdot |A|^{-2\gamma}$. Therefore, since T has length at most $|A|^{\beta}$, the probability that T is truncated at v for either reason is at most $|A|^{\beta}((r-1) \cdot |A|^{-2\gamma} + |X|^{-\delta}) < r \cdot |A|^{-\beta}$ since $|A| \leqslant |X|$ and $\beta < \min(\gamma, \delta/2)$.

(*Readers who wish to focus on identification may find it easier on first reading to skip to the alternative proof at the end of this argument.*) For any sink node v of B, let $\mathrm{Trunc}(v)$ denote the probability that the random computation path C for a random concept x' chosen uniformly from X is truncated, conditioned on the computation on x' ending at v. By Markov's inequality, the probability that the computation path for a random concept x' ends at vertex v with $\mathrm{Trunc}(v) > 2r \cdot |A|^{-\beta/2}$ is less than $\frac{1}{2}|A|^{-\beta/2}$.

Let $f_v : A \to \{0, 1, \ldots, r-1\}$ denote the function labelling node v which encapsulates the best prediction of the algorithm for the label of each point in A. (This will simply be some $x'' \in X$ in the case of identification rather than prediction.) We argue that if v is not significant and $\mathrm{Trunc}(v) \leqslant 2r|A|^{-\beta/2}$ then f_v provides little advantage in predicting $L(\cdot, x')$.

By Lemmas 5.1 and 5.2, if v is not significant then

$$\mathbf{Pr}_{x' \sim \mathbb{P}_{x|v},\, a \in_R A}[L(a, x') = f_v(a)] \leqslant \frac{1}{r} + |A|^{-\gamma} + (r-1) \cdot |A|^{-2\gamma}.$$

Now $\mathbb{P}_{x|v}$ is the distribution on concepts $x' \in X$ conditioned on their (randomly chosen) computation path reaching v and not being truncated. On the other hand, correctness is defined with respect to randomly chosen computation paths independent of truncation. However, if $\mathrm{Trunc}(v) \leqslant 2r \cdot |A|^{-\beta/2}$, then over the distribution independent of truncation we obtain that the probability conditioned on x' reaching v of predicting $L(a, x')$ is at most

$$\frac{1}{r} + |A|^{-\gamma} + (r-1) \cdot |A|^{-2\gamma} + 2r \cdot |A|^{-\beta/2} \leqslant \frac{1}{r} + |A|^{-\varepsilon/2}$$

which is a prediction advantage of at most $|A|^{-\varepsilon/2}$. Therefore none of the nodes non-significant nodes v with small $\mathrm{Trunc}(v)$ can contribute to the $|A|^{-\varepsilon}$ success probability and hence the probability that the computation reaches a significant node must be at least the success probability $|A|^{-\varepsilon}$ of having a large advantage minus the probability that the computation on x' reaches a sink vertex v with $\mathrm{Trunc}(v) > 2r \cdot |A|^{-\beta/2}$, which is $|A|^{-\varepsilon} - \frac{1}{2}|A|^{-\beta/2} = \frac{1}{2}|A|^{-\beta/2}$ as required.

(*An alternative simpler argument that may be a bit more intuitive for the case of identification.*) If T reaches a leaf v that is not significant then, conditioned on arriving at v, the probability that the concept x' equals the label of v is at most $\max_{x'' \in X} \mathbb{P}_{x|v}(x'')$. Now

$$\frac{\max_{x'' \in X} \mathbb{P}_{x|v}(x'')}{|X|^{1/2}} \leqslant \|\mathbb{P}_{x|v}\|_2 < |X|^{-(1-\delta/2)}$$

since v is not significant, so we have $\max_{x'' \in X} \mathbb{P}_{x|v}(x'') < |X|^{-(1-\delta)/2} = |X|^{-(\alpha+\delta)/2}$ and the probability that B is correct conditioned on the truncated path reaching a leaf vertex that is not significant is less than $|X|^{-(\alpha+\delta)/2} \leqslant |X|^{-\beta} \leqslant |A|^{-\beta}$ since $|A| \leqslant |X|$.

Since B is correct with probability at least $|A|^{-\varepsilon} = |A|^{-\beta/2}$ and these three cases in which T does not reach a significant vertex account for correctness at most $(r+1) \cdot |A|^{-\beta}$, which is much less than $\frac{1}{2} \cdot |A|^{-\beta/2}$, T must reach a significant vertex with probability at least $\frac{1}{2}|A|^{-\beta/2}$. □

The following lemma is the the key to the proof of the theorem.

Lemma 5.5. *Let s be any significant vertex of B. There is an $\eta = \delta\gamma/2 > 0$ such that for a uniformly random x chosen from X and a uniformly random computation path C, the probability that its truncation T ends at s is at most $|X|^{-\eta \log_r |A|}$.*

The proof of Lemma 5.5 requires a delicate progress argument and is deferred to the next subsection. We first show how Lemmas 5.4 and 5.5 immediately imply Theorem 4.1.

Proof of Theorem 4.1. By Lemma 5.4, for x chosen uniformly at random from X and T the truncation of a uniformly random computation path given concept x, T ends at a significant vertex with probability at least $|A|^{-\beta/2}/2$. On the other hand, by Lemma 5.5, for any significant vertex s, the probability that T ends at s is at most $|X|^{-\eta \log_r |A|}$. Therefore the number of significant vertices must be at least $2|X|^{\eta \log_r |A|}/|A|^{\beta/2}$ and since B has length at most $|A|^{\beta}$, there must be at least $2|X|^{\eta \log_r |A|}/|A|^{3\beta/2}$ significant vertices in some layer. Hence B requires space $\Omega(\delta\gamma \log_2 |X| \log_r |A|)$. □

5.1 Progress towards significance

In this section we prove Lemma 5.5 showing that for any particular significant vertex s a random truncated path reaches s only with probability $|X|^{-\Omega(\delta\gamma \ \log_r |A|)}$. For each vertex v in B let $\mathbf{Pr}[v]$ denote the probability over a random concept x, that the truncation of a random computation path in B given concept x visits v and for each edge e in B let $\mathbf{Pr}[e]$ denote the probability over a random concept x, that the truncation of a random computation path in B for x traverses e.

Since B is a leveled branching program, the vertices of B may be divided into disjoint sets V_t for $t = 0, 1, \ldots, T$ where T is the length of B and V_t is the set of vertices at distance t from the root, and disjoint sets of edges E_t for $t = 1, \ldots, T$ where E_t consists of the edges from V_{t-1} to V_t. For each vertex $v \in V_{t-1}$, note that by definition we only have

$$\mathbf{Pr}[v] \geqslant \sum_{(v,w) \in E_t} \mathbf{Pr}[(v,w)]$$

since some truncated paths may terminate at v.

For each t, since the truncated computation path visits at most one vertex and at most one edge at level t, we obtain a sub-distribution on V_t in which the probability of $v \in V_t$ is $\mathbf{Pr}[v]$ and a corresponding sub-distribution on E_t in which the probability of $e \in E_t$ is $\mathbf{Pr}[e]$. We write $v \sim V_t$ and $e \sim E_t$ to denote random selection from these sub-distributions, where the outcome $\perp$ corresponds to the case that no vertex (respectively no edge) is selected.

Fix some significant vertex s. We consider the progress that a truncated path makes as it moves from the start vertex to s. We measure the progress at a vertex v as

$$\rho(v) = \frac{\langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s} \rangle}{\langle \mathbb{P}_{x|s}, \mathbb{P}_{x|s} \rangle}.$$

Clearly $\rho(s) = 1$. We first see that ρ starts out at a tiny value.

Lemma 5.6. *If v_0 is the start vertex of B then $\rho(v_0) \leqslant |X|^{-\delta}$.*

Proof. By definition, $\mathbb{P}_{x|v_0}$ is the uniform distribution on X. Therefore

$$\langle \mathbb{P}_{x|v_0}, \mathbb{P}_{x|s} \rangle = \mathbf{E}_{x' \in X}[|X|^{-1} \cdot \mathbb{P}_{x|s}(x')] = |X|^{-2} \cdot \sum_{x' \in X} \mathbb{P}_{x|v_0}(x') = |X|^{-2}$$

since $\mathbb{P}_{x|s}$ is a probability distribution on X. On the other hand, since s is significant, $\langle \mathbb{P}_{x|s}, \mathbb{P}_{x|s} \rangle = \|\mathbb{P}_{x|s}\|_2^2 \geqslant |X|^{\delta} \cdot |X|^{-2}$. The lemma follows immediately. □

Since the truncated path is randomly chosen, the progress towards s after t steps is a random variable. We show that not only is the increase in this expected value of this random variable in each step very small, its higher moments also increase at a very small rate. Define

$$\Phi_t = \mathbf{E}_{v \sim V_t}[(\rho(v))^{\gamma \log_r |A|}]$$

where we extend ρ and define $\rho(\perp) = 0$. We will show that for $s \in V_t$, Φ_t is still $|X|^{-\Omega(\delta\gamma \, \log_r |A|)}$, which will be sufficient to prove Lemma 5.5.

Therefore, Lemma 5.5, and hence Theorem 4.1, will follow from the following lemma.

Lemma 5.7. *For every t with $1 \leqslant t \leqslant |A|^{\beta} - 1$,*

$$\Phi_t \leqslant \Phi_{t-1} \cdot (1 + |A|^{-2\beta}) + |X|^{-\gamma \log_r |A|}.$$

Proof of Lemma 5.5 from Lemma 5.7. By definition of Φ_t and Lemma 5.6 we have $\Phi_0 \leqslant |X|^{-\delta\gamma \log_r |A|}$. By Lemma 5.7, for every t with $1 \leqslant t \leqslant |A|^{\beta} - 1$,

$$\Phi_t \leqslant \sum_{j=0}^{t} (1 + |A|^{-2\beta})^j \cdot |X|^{-\delta\gamma \log_r |A|} < (t+1) \cdot (1 + |A|^{-2\beta})^t \cdot |X|^{-\delta\gamma \log_r |A|}.$$

In particular, for every $t \leqslant |A|^{\beta} - 1$,

$$\Phi_t \leqslant |A|^{\beta} \cdot (1 + |A|^{-2\beta})^{|A|^{\beta}} \cdot |X|^{-\delta\gamma \log_r |A|} \leqslant e^{1/|A|^{\beta}} \cdot |A|^{\beta} \cdot |X|^{-\delta\gamma \log_r |A|}.$$

Now fix t^* to be the level of the significant node s. Every truncated path that reaches s will have contribution to Φ_{t^*} of $(\rho(s))^{\gamma \log_r |A|} = 1$ times its probability of occurring. Therefore the truncation of a random computation path reaches s with probability at most $|X|^{-\eta \log_r |A|}$ for $\eta = \delta\gamma/2$ and $|A|, |X|$ sufficiently large, which proves the lemma. □

We now focus on the proof of Lemma 5.7. Because Φ_t depends on the sub-distribution over V_t and Φ_{t-1} depends on the sub-distribution over V_{t-1}, it is natural to consider the analogous quantities based on the sub-distribution over the set E_t of edges that join V_{t-1} and V_t. We can extend the definition of ρ to edges of B, where we write

$$\rho(e) = \frac{\langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s} \rangle}{\langle \mathbb{P}_{x|s}, \mathbb{P}_{x|s} \rangle}.$$

Then define

$$\Phi'_t = \mathbf{E}_{e \sim E_t}[(\rho(e))^{\gamma \log_r |A|}].$$

Intuitively, there is no gain of information in moving from elements E_t to elements of V_t. More precisely, we have the following lemma:

Lemma 5.8. *For all t, $\Phi_t \leqslant \Phi'_t$.*

Proof. Note that for $v \in V_t$, since the truncated paths that follow some edge $(u,v) \in E_t$ are precisely those that reach v, by definition, $\mathbf{Pr}[v] = \sum_{(u,v) \in E_t} \mathbf{Pr}[(u,v)]$. Since the same applies separately to the set of truncated paths for each concept $x' \in X$ that reach v, for each $x' \in X$ we have

$$\mathbf{Pr}[v] \cdot \mathbb{P}_{x|v}(x') = \sum_{(u,v) \in E_t} \mathbf{Pr}[(u,v)] \cdot \mathbb{P}_{x|(u,v)}(x').$$

Therefore,

$$\mathbf{Pr}[v] \cdot \frac{\langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s} \rangle}{\langle \mathbb{P}_{x|s}, \mathbb{P}_{x|s} \rangle} = \sum_{(u,v) \in E_t} \mathbf{Pr}[(u,v)] \cdot \frac{\langle \mathbb{P}_{x|(u,v)}, \mathbb{P}_{x|s} \rangle}{\langle \mathbb{P}_{x|s}, \mathbb{P}_{x|s} \rangle};$$

i.e., $\mathbf{Pr}[v] \cdot \rho(v) = \sum_{(u,v) \in E_t} \mathbf{Pr}[(u,v)] \cdot \rho((u,v))$. Since $\mathbf{Pr}[v] = \sum_{(u,v) \in E_t} \mathbf{Pr}[(u,v)]$, by the convexity of the map $s \mapsto s^{\gamma \log_r |A|}$ we have

$$\mathbf{Pr}[v] \cdot (\rho(v))^{\gamma \log_r |A|} \leqslant \sum_{(u,v) \in E_t} \mathbf{Pr}[(u,v)] \cdot (\rho((u,v))^{\gamma \log_r |A|}.$$

Therefore

$$\begin{aligned} \Phi_t = \sum_{v \in V_t} \mathbf{Pr}[v] \cdot (\rho(v))^{\gamma \log_r |A|} &\leqslant \sum_{v \in V_t} \sum_{(u,v) \in E_t} \mathbf{Pr}[(u,v)] \cdot (\rho((u,v)))^{\gamma \log_r |A|} \\ &= \sum_{e \in E_t} \mathbf{Pr}[e] \cdot (\rho(e))^{\gamma \log_r |A|} = \Phi'_t. \end{aligned}$$

□

Therefore, to prove Lemma 5.7 it suffices to prove that the same statement holds with Φ_t replaced by Φ'_t; that is,

$$\mathbf{E}_{e \in E_t}[(\rho(e))^{\gamma \log_r |A|}] \leqslant (1 + |A|^{-2\beta}) \cdot \mathbf{E}_{v \in V_{t-1}}[(\rho(v))^{\gamma \log_r |A|}] + |X|^{-\gamma \log_r |A|}$$

E_t is the disjoint union of the out-edges $\Gamma_{out}(v)$ for vertices $v \in V_{t-1}$, so it suffices to show that for each $v \in V_{t-1}$,

$$\sum_{e \in \Gamma_{out}(v)} \mathbf{Pr}[e] \cdot (\rho(e))^{\gamma \log_r |A|} \leqslant (1 + |A|^{-2\beta}) \cdot \mathbf{Pr}[v] \cdot (\rho(v))^{\gamma \log_r |A|} + |X|^{-\gamma \log_r |A|} \cdot \mathbf{Pr}[v]. \quad (1)$$

Since any truncated path that follows e must also visit v, we can write $\mathbf{Pr}[e|v] = \mathbf{Pr}[e]/\mathbf{Pr}[v]$. Moreover, both $\rho(v)$ and $\rho(e)$ have the same denominator $\langle \mathbb{P}_{x|s}, \mathbb{P}_{x|s} \rangle$ and therefore, by definition, inequality (1), and hence Lemma 5.7, follows from the following lemma.

Lemma 5.9. *For $v \in V_{t-1}$,*

$$\sum_{e \in \Gamma_{out}(v)} \mathbf{Pr}[e|v] \cdot \langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s} \rangle^{\gamma \log_r |A|} \leqslant (1 + |A|^{-2\beta}) \cdot \langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s} \rangle^{\gamma \log_r |A|} + |X|^{-\gamma \log_r |A|}.$$

Before we prove Lemma 5.9, we first prove some technical lemmas, the first relating the distributions for $v \in V_{t-1}$ and edges $e \in E_t$ and the last upper bounding $\|\mathbb{P}_{x|s}\|_2$.

Lemma 5.10. *Suppose that $v \in V_{t-1}$ is not significant and $e = (v, w) \in E_t$ has $\mathbf{Pr}[e] > 0$ and label (a, b). Then for $x' \in X$, $\mathbb{P}_{x|e}(x') > 0$ only if $x' \notin \mathrm{High}(v)$ and $L(a, x') = b$, in which case*

$$\mathbb{P}_{x|e}(x') = c_e^{-1} \cdot \mathbb{P}_{x|v}(x')$$

where $c_e \geqslant \frac{1}{r} - |A|^{-\gamma} - |X|^{-\delta}$.

Proof. If there exists $j \in \{1, \cdots, r-1\}$ such that $|(M^{(j)} \cdot \mathbb{P}_{x|v})(a)| \geqslant |A|^{-\gamma}$ then by definition of truncation we also will have $\mathbf{Pr}[e] = 0$. Therefore, since $\mathbf{Pr}[e] > 0$, e is not a high bias edge – that is, $\forall j \in [r-1]$, $|(M^{(j)} \cdot \mathbb{P}_{x|v})(a)| < |A|^{-\gamma}$. We now use Lemma 5.2 to derive that

$$\mathbf{Pr}_{x' \sim \mathbb{P}_{x|v}}[L(a, x') = b] \geqslant \frac{1}{r} - |A|^{-\gamma}.$$

Let $\mathcal{E}_e(x')$ be the event that both $L(a, x') = b$ and $x' \notin \mathrm{High}(v)$ and define

$$c_e = \mathbf{Pr}_{x' \sim \mathbb{P}_{x|v}}[\mathcal{E}_e(x')].$$

If $\mathcal{E}_e(x')$ fails to hold for all x', i.e., $x' \in \mathrm{High}(v)$ or $L(a, x') \neq b$, then any truncated path for concept x' that reaches v will not continue along e and hence $\mathbf{Pr}[e] = 0$. On the other hand, since $\mathbf{Pr}[e] > 0$, if $\mathcal{E}_e(x')$ holds for some concept x' then any truncated path for x' that reaches v will continue precisely if the input chosen at v is a, which happens with probability $|A|^{-1}$ for each such x'. The total probability over $x' \in X$, conditioned that the truncated path on x' reaches v and that the path continues along e is then $|A|^{-1} \cdot c_e$. Therefore, if $x' \in \mathcal{E}_e$ then $\mathbb{P}_{x|e}(x') = \frac{|A|^{-1} \cdot \mathbb{P}_{x|v}(x')}{|A|^{-1} \cdot c_e} = c_e^{-1} \cdot \mathbb{P}_{x|v}(x')$. Now by Lemma 5.3,

$$\mathbf{Pr}_{x' \sim \mathbb{P}_{x|v}}[x' \in \mathrm{High}(v)] \leqslant |X|^{-\delta}$$

and so

$$c_e = \mathbf{Pr}_{x' \sim \mathbb{P}_{x|v}}[L(a, x') = b \text{ and } x' \notin \mathrm{High}(v)] > \frac{1}{r} - |A|^{-\gamma} - |X|^{-\delta}$$

as required. □

We use this lemma together with an argument similar to that of Lemma 5.8 to upper bound $\|\mathbb{P}_{x|s}\|_2$ for our significant vertex s.

Lemma 5.11. $\|\mathbb{P}_{x|s}\|_2 \leqslant 2r \cdot |X|^{-(1-\delta/2)}$.

Proof. The main observation is that s is the first significant vertex of any truncated path that reaches it and so the probability distributions of each of the immediate predecessors v of s must have bounded expectation 2-norm and, by Lemma 5.10 and the proof idea from Lemma 5.8, the 2-norm of the distribution at s cannot grow too much larger than those at its immediate predecessors.

By Lemma 5.10, if $e = (v, s)$ and $\mathbf{Pr}[e] > 0$, then

$$\|\mathbb{P}_{x|e}\|_2 \leqslant c_e^{-1} \cdot \|\mathbb{P}_{x|v}\| \leqslant c_e^{-1} |X|^{-(1-\delta/2)} \leqslant 2r \cdot |X|^{-(1-\delta/2)}$$

since v is not significant and $c_e \geqslant \frac{1}{r} - |A|^{-\gamma} - |X|^{-\delta} > \frac{1}{2r}$ for $|A|$ and $|X|$ sufficiently large. Let $\Gamma_{in}(s)$ be the set of edges (v, s) in B. $\mathbf{Pr}[s] = \sum_{e=(v,s)\in\Gamma_{in}(s)} \mathbf{Pr}[e]$ and for each $x' \in X$,

$$\mathbf{Pr}[s] \cdot \mathbb{P}_{x|s}(x') = \sum_{e=(v,s)\in\Gamma_{in}(s)} \mathbf{Pr}[e] \cdot \mathbb{P}_{x|e}(x').$$

Since $\mathbf{Pr}[s] = \sum_{e=(v,s)\in\Gamma_{in}(s)} \mathbf{Pr}[e]$, by convexity of the map $r \mapsto r^2$, we have

$$\mathbf{Pr}[s] \cdot (\mathbb{P}_{x|s}(x'))^2 \leqslant \sum_{e=(v,s)\in\Gamma_{in}(s)} \mathbf{Pr}[e] \cdot (\mathbb{P}_{x|e}(x'))^2.$$

Summing over $x' \in X$ we have

$$\begin{aligned}\mathbf{Pr}[s] \cdot \|\mathbb{P}_{x|s}\|_2^2 &\leqslant \sum_{e=(v,s)\in\Gamma_{in}(s)} \mathbf{Pr}[e] \cdot \|\mathbb{P}_{x|e}\|_2^2 \\ &\leqslant \sum_{e=(v,s)\in\Gamma_{in}(s)} \mathbf{Pr}[e] \cdot (2r \cdot |X|^{-(1-\delta/2)})^2 = \mathbf{Pr}[s] \cdot (2r \cdot |X|^{-(1-\delta/2)})^2.\end{aligned}$$

Therefore $\|\mathbb{P}_{x|s}\| \leqslant 2r \cdot |X|^{-(1-\delta/2)}$ as required since $\mathbf{Pr}[s] > 0$. □

To complete the proof of Lemma 5.7, and hence Lemma 5.5, it only remains to prove Lemma 5.9.

5.1.1 Proof of Lemma 5.9

Since we know that if $v \in V_{t-1}$ is significant then any edge $e \in \Gamma_{out}(v)$ has $\mathbf{Pr}[e] = 0$, we can assume without loss of generality that v is not significant.

Define $g : X \to \mathbb{R}$ by

$$g(x') = \mathbb{P}_{x|v}(x') \cdot \mathbb{P}_{x|s}(x')$$

and note that $\langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s} \rangle = \mathbf{E}_{x' \in X}[g(x')]$. For $x' \in X$ define

$$f(x') = \begin{cases} g(x') & x' \notin \mathrm{High}(v) \\ 0 & \text{otherwise} \end{cases}$$

and let $F = \sum_{x' \in X} f(x')$. For every edge e where $\langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s} \rangle > 0$, we have $F > 0$.

The function f induces a new probability distribution on X, $\mathbb{P}_f$, given by $\mathbb{P}_f(x') = f(x') / \sum_{x \in X} f(x) = f(x')/F$ in which each point $x' \in X \setminus \mathrm{High}(v)$ is chosen with probability proportional to its contribution to $\langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s} \rangle$ and each $x' \in \mathrm{High}(v)$ has probability 0.

CLAIM: Let (a, b) be the label on an edge e, then

$$\langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s}\rangle \leqslant (rc_e)^{-1} \cdot (1 + \sum_{j=1}^{r-1} |(M^{(j)} \cdot \mathbb{P}_f)(a)|) \cdot F/|X| \leqslant (rc_e)^{-1} \cdot (1 + \sum_{j=1}^{r-1} |(M^{(j)} \cdot \mathbb{P}_f)(a)|) \cdot \langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s}\rangle$$

where c_e is given by Lemma 5.10.

We first prove the claim. By Lemma 5.10 and the definition of f,

$$\mathbb{P}_{x|e}(x') \cdot \mathbb{P}_{x|s}(x') = \begin{cases} c_e^{-1} \cdot f(x') & \text{if } L(a, x') = b \\ 0 & \text{otherwise.} \end{cases}$$

Therefore

$$\langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s}\rangle = \mathbf{E}_{x' \in_R X}[\mathbb{P}_{x|e}(x') \cdot \mathbb{P}_{x|s}(x')] = \mathbf{E}_{x' \in_R X}[c_e^{-1} f(x') \cdot \mathbf{1}_{L(a,x')=b}]$$

Let $z = M^{(1)}(a, x') \cdot \omega^{-b}$. Then $z \in \{1, \omega, \cdots, \omega^{r-1}\}$. The indicator function $\mathbf{1}_{L(a,x')=b}$ is 1 when $z = 1$, and $\mathbf{1}_{L(a,x')=b}$ is 0 when $z = \omega, \cdots, \omega^{r-1}$. By interpolation we have

$$\mathbf{1}_{L(a,x')=b} = \frac{1}{r} \sum_{j=0}^{r-1} z^j$$

Notice that $z^j = \omega^{-bj} \cdot M^{(j)}(a, x')$ for $j = 1, \cdots, r-1$, so we have

$$\begin{aligned}
\langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s}\rangle &= \mathbf{E}_{x' \in_R X}\left[c_e^{-1} f(x') \cdot (1 + \sum_{j=1}^{r-1} \omega^{-b \cdot j} \cdot M^{(j)}(a, x'))/r\right] \\
&= (rc_e)^{-1} \cdot \left(\mathbf{E}_{x' \in_R X}[f(x')] + \sum_{j=1}^{r-1} \omega^{-bj} \cdot \mathbf{E}_{x' \in_R X}[M^{(j)}(a, x') \cdot f(x')]\right) \\
&\leqslant (rc_e)^{-1} \cdot \left(\mathbf{E}_{x' \in_R X}[f(x')] + \sum_{j=1}^{r-1} \left|\mathbf{E}_{x' \in_R X}[M^{(j)}(a, x') \cdot f(x')]\right|\right) \\
&= (rc_e)^{-1} \cdot |X|^{-1} \cdot F \cdot \left(1 + \frac{\sum_{j=1}^{r-1} \left|\mathbf{E}_{x' \in_R X}[M^{(j)}(a, x') \cdot f(x')]\right|}{F}]\right) \\
&= (rc_e)^{-1} \cdot |X|^{-1} \cdot F \cdot (1 + \sum_{j=1}^{r-1} |(M^{(j)} \cdot \mathbb{P}_f)(a)|) \\
&\leqslant (rc_e)^{-1} \cdot (1 + \sum_{j=1}^{r-1} |(M^{(j)} \cdot \mathbb{P}_f)(a)|) \cdot \langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s}\rangle
\end{aligned}$$

since $|X|^{-1} \cdot F = \mathbf{E}_{x' \in_R X}[f(x')] \leqslant \mathbf{E}_{x' \in_R X}[g(x')] = \langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s}\rangle$, which proves the claim.

By Lemma 5.10, $rc_e \geqslant 1 - r \cdot |A|^{-\gamma} - r \cdot |X|^{-\delta}$ and so $(rc_e)^{-1} \leqslant 1 + |A|^{-\sigma} \leqslant 2$ for $\sigma = \min(\gamma, \delta)/2$ and sufficiently large $|A|$ since $|A| \leqslant |X|$. We consider two cases:

CASE $F \leqslant |X|^{-1}$: In this case, since $\mathbb{P}_f$ is a probability distribution, for every $a \in A$ and $j \in [r-1]$, we have $|(M^{(j)} \cdot \mathbb{P}_f)(a)| \leqslant \max_{x' \in X} |M^{(j)}(a, x')| = 1$ and from the claim we obtain for every edge $e \in \Gamma_{out}(v)$, $\langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s} \rangle \leqslant r \cdot (rc_e)^{-1} \cdot |X|^{-2}$. Therefore $\sum_{e \in \Gamma_{out}(v)} \mathbf{Pr}[e|v] \cdot \langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s} \rangle^{\gamma \log_r |A|}$ is at most $(2r \cdot |X|^{-2})^{\gamma \log_r |A|} \leqslant |X|^{-\gamma \log_r |A|}$ for $|X| \geqslant 2r$.

CASE $F \geqslant |X|^{-1}$: In this case we will show that $\|\mathbb{P}_f\|_2$ is not too large and use this together with the bound on the 2-norm amplification curve of each $M^{(j)}$ to show that $\|M^{(j)} \cdot \mathbb{P}_f\|_2$ is small for each $j = 1, \cdots, r-1$. This will be important because of the following connection:

By the Claim, we have

$$\sum_{e \in \Gamma_{out}(v)} \mathbf{Pr}[e|v] \cdot \langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s} \rangle^{\gamma \log_r |A|} \leqslant \sum_{e \in \Gamma_{out}(v)} \mathbf{Pr}[e|v] \cdot [(rc_e)^{-1} \cdot (1 + \sum_{j=1}^{r-1} |(M^{(j)} \cdot \mathbb{P}_f)(a_e)|)]^{\gamma \log_r |A|} \cdot \langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s} \rangle^{\gamma \log_r |A|} \tag{2}$$

where a_e is the input labelling edge e. By definition, for each $a \in A$ there are precisely r edges $e_0, \cdots, e_{r-1} \in \Gamma_{out}(v)$ with $a_{e_i} = a$ for $i = 0, \cdots, r-1$ and $\sum_{i=0}^{r-1} \mathbf{Pr}[e_i|v] \leqslant 1/|A|$ since the next input is chosen uniformly at random from A. (It would be equality but some inputs a have high bias and in that case $\mathbf{Pr}[e_i|v] = 0$ for all i.) Previously, we also observed that $(rc_e)^{-1} \leqslant 1 + |A|^{-\sigma}$ where $\sigma = \min(\gamma, \delta)/2$. Therefore,

$$\begin{aligned}
&\sum_{e \in \Gamma_{out}(v)} \mathbf{Pr}[e|v] \cdot \langle \mathbb{P}_{x|e}, \mathbb{P}_{x|s} \rangle^{\gamma \log_r |A|} \\
&\leqslant \sum_{a \in A} \frac{1}{|A|} [(1 + |A|^{-\sigma}) \cdot (1 + \sum_{j=1}^{r-1} |(M^{(j)} \cdot \mathbb{P}_f)(a_e)|)]^{\gamma \log_r |A|} \cdot \langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s} \rangle^{\gamma \log_r |A|} \\
&= (1 + |A|^{-\sigma})^{\gamma \log_r |A|} \cdot \mathbf{E}_{a \in_R A} [(1 + \sum_{j=1}^{r-1} |(M^{(j)} \cdot \mathbb{P}_f)(a_e)|)^{\gamma \log_r |A|}] \cdot \langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s} \rangle^{\gamma \log_r |A|}.
\end{aligned}$$

To prove the lemma we therefore need to bound $\mathbf{E}_{a \in_R A}[(1 + \sum_{j=1}^{r-1} |(M^{(j)} \cdot \mathbb{P}_f)(a_e)|)^{\gamma \log_r |A|}]$. We will bound this by first analyzing $\|M^{(j)} \cdot \mathbb{P}_f\|_2$ for each j.

Fix j. By definition,

$$\|f\|_2^2 = \mathbf{E}_{x' \in_R X} \mathbf{1}_{x' \notin \mathrm{High}(v)} \cdot \mathbb{P}_{x|v}^2(x') \cdot \mathbb{P}_{x|s}^2(x') \leqslant |X|^{-2\alpha} \cdot \mathbf{E}_{x' \in_R X} \mathbb{P}_{x|s}^2(x') = |X|^{-2\alpha} \cdot \|\mathbb{P}_{x|s}\|_2^2.$$

Therefore, by Lemma 5.11, and the fact that $F \geqslant |X|^{-1}$,

$$\|\mathbb{P}_f\|_2 = \frac{\|f\|_2}{F} \leqslant \frac{|X|^{-\alpha} \cdot \|\mathbb{P}_{x|s}\|_2}{|X|^{-1}} \leqslant |X|^{(1-\alpha)} \cdot 2r \cdot |X|^{-(1-\delta/2)} = |X|^{1-\alpha+\delta/2+\log_{|X|} 2r} \cdot |X|^{-1}.$$

Since, for sufficiently large $|X|$,

$$1 - \alpha + \delta/2 + \log_{|X|} 2r = 2\delta + \delta/2 + \log_{|X|} 2r \leqslant 3\delta = \delta'/2,$$

we have $\|\mathbb{P}_f\|_2 \leqslant |X|^{-(1-\delta'/2)}$. So, definition of τ we have $\|M^{(j)} \cdot \mathbb{P}_f\|_2 \leqslant |A|^{\tau_{M^{(j)}}(\delta')} \leqslant |A|^{-2\gamma}$. Thus $\mathbf{E}_{a \in_R A}[|(M^{(j)} \cdot \mathbb{P}_f)(a)|^2] = \|M^{(j)} \cdot \mathbb{P}_f\|_2^2 \leqslant |A|^{-4\gamma}$. So, by Markov's inequality,

$$\mathbf{Pr}_{a \in_R A}[|(M^{(j)} \cdot \mathbb{P}_f)(a)| \geqslant |A|^{-\gamma}] = \mathbf{Pr}_{a \in_R A}[|(M^{(j)} \cdot \mathbb{P}_f)(a)|^2 \geqslant |A|^{-2\gamma}] \leqslant |A|^{-2\gamma}.$$

By a union bound,

$$\mathbf{Pr}_{a\in_R A}\left[\exists j \in [r-1],\ |(M^{(j)}\cdot \mathbb{P}_f)(a)| \geqslant |A|^{-\gamma}\right] \leqslant (r-1)\cdot |A|^{-2\gamma}.$$

Therefore, since we always have $|(M^{(j)}\cdot \mathbb{P}_f)(a)| \leqslant 1$,

$$\begin{aligned}
&\mathbf{E}_{a\in_R A}[(1+\sum_{j=1}^{r-1}|(M^{(j)}\cdot \mathbb{P}_f)(a)|^{\gamma \log_r |A|}]\\
\leqslant&\mathbf{E}_{a\in_R A}\left[\mathbf{1}_{\forall j\in[r-1],\ |(M^{(j)}\cdot\mathbb{P}_f)(a)|\leqslant |A|^{-\gamma}}\cdot(1+(r-1)\cdot|A|^{-\gamma})^{\gamma\log_r|A|}\right]\\
&+\mathbf{E}_{a\in_R A}\left[\mathbf{1}_{\exists j\in[r-1],\ |(M^{(j)}\cdot\mathbb{P}_f)(a)|> |A|^{-\gamma}}\cdot r^{\gamma\log_r|A|}\right]\\
\leqslant&(1+(r-1)\cdot|A|^{-\gamma})^{\gamma\log_r|A|}+(r-1)\cdot|A|^{-2\gamma}\cdot|A|^{\gamma}\\
=&(1+(r-1)\cdot|A|^{-\gamma})^{\gamma\log_r|A|}+(r-1)\cdot|A|^{-\gamma}\leqslant 1+|A|^{-\gamma/2}
\end{aligned}$$

for $\gamma \log_r |A|$ sufficiently large. Therefore, the total factor increase over $\langle \mathbb{P}_{x|v}, \mathbb{P}_{x|s}\rangle^{\gamma \log_r |A|}$ is at most $(1+|A|^{-\sigma})^{\gamma\log_r|A|}\cdot(1+|A|^{-\gamma/2})$ where $\sigma = \min(\gamma,\delta)/2$. Therefore, for sufficiently large $|A|$ this is at most $1+|A|^{-\min(\gamma,\delta)/4}$. Since $\beta \leqslant \min(\gamma,\delta)/8$ this is at most $1+|A|^{-2\beta}$ as required to prove Lemma 5.9.

6 An SDP Relaxation for Norm Amplification on the Positive Orthant

For a matrix $M \in \mathbb{C}^{A\times X}$,

$$\tau_M(\delta) = \sup_{\substack{\mathbb{P}\in\Delta_X\\ \|\mathbb{P}\|_2\leqslant 1/|X|^{1-\delta/2}}} \log_{|A|}(\|M\cdot\mathbb{P}\|_2).$$

That is, $\tau_M(\delta) = \frac{1}{2}\log_{|A|} OPT_{M,\delta}$ where $OPT_{M,\delta}$ is the optimum of the following quadratic program:

$$\begin{aligned}
&\text{Maximize} && \|M\cdot\mathbb{P}\|_2^2 = \langle M\cdot\mathbb{P}, M\cdot\mathbb{P}\rangle,\\
&\text{subject to:} && \\
& && \sum_{i\in X}\mathbb{P}_i = 1,\\
& && \sum_{i\in X}\mathbb{P}_i^2 \leqslant |X|^{\delta-1},\\
& && \mathbb{P}_i \geqslant 0 \qquad\qquad \text{for all } i\in X.
\end{aligned} \tag{3}$$

Instead of attempting to solve (3), presumably a difficult quadratic program, we consider the following semidefinite program (SDP):

$$\begin{aligned} &\text{Maximize} \quad \langle M^*M, U\rangle \cdot |X|^2/|A| \\ &\text{subject to:} \\ &\qquad [V] \quad U \succeq 0, \\ &\qquad [w] \quad \sum_{i,j\in X} U_{ij} = 1, \\ &\qquad [z] \quad \sum_{i\in X} U_{ii} \leqslant |X|^{\delta-1}, \\ &\qquad\qquad U_{ij} \in \mathbb{R}, U_{ij} \geqslant 0 \qquad \text{for all } i,j \in X. \end{aligned} \tag{4}$$

Recall that M^* is the conjugate transpose of M. Note that for any $\mathbb{P} \in \Delta_X$ achieving the optimum value of (3) the positive semidefinite matrix $U = \mathbb{P} \cdot \mathbb{P}^T$ has the same value in (4) (where the $|X|^2/|A|$ factor accounts for the difference in scaling factors based on the dimensions for the two expectation inner products), and hence (4) is an SDP relaxation of (3).

But this is not a standard SDP, since M is over $\mathbb{C}$ and M^*M might contain complex entries. In order to apply techniques on real matrices, we define $N : X \times X \to \mathbb{R}$ as $N(x,x') = Re(M^*M(x,x'))$, that is, N is the real part of M^*M. Then we have the following (real) program:

$$\begin{aligned} &\text{Maximize} \quad \langle N, U\rangle \cdot |X|^2/|A| \\ &\text{subject to:} \\ &\qquad [V] \quad U \succeq 0, \\ &\qquad [w] \quad \sum_{i,j\in X} U_{ij} = 1, \\ &\qquad [z] \quad \sum_{i\in X} U_{ii} \leqslant |X|^{\delta-1}, \\ &\qquad\qquad U_{ij} \geqslant 0 \qquad \text{for all } i,j \in X. \end{aligned} \tag{5}$$

The key observation is that (4) and (5) have the same optimal value. This is because for any $U \in \mathbb{R}^{X\times X}$,

$$|X|^2\langle M^*M, U\rangle = \sum_{i,j}(M^*M)_{ij} \cdot U_{ij} = \sum_{ij} Re((M^*M)_{ij}) \cdot U_{ij} + i \cdot \sum_{x,x'} Im((M^*M)_{ij}) \cdot U_{ij}$$

Since M^*M is a Hermitian matrix, we have $(M^*M)_{ij} = \overline{(M^*M)_{ji}}$. But U is real symmetric, so we have $\sum_{i,j} Im((M^*M)_{ij}) \cdot U_{ij} = 0$, namely

$$|X|^2\langle M^*M, U\rangle = \sum_{i,j} Re((M^*M)_{ij}) \cdot U_{ij} = |X|^2\langle N, U\rangle$$

and we only need to consider the real parts. In order to upper bound the value of (5), we consider its dual program:

$$\begin{array}{rll}
\text{Minimize} & w + z \cdot |X|^{\delta-1} & \\
\text{subject to:} & & \\
[U] & V \succeq 0, & \\
[U_{ii}] & w + z \geqslant V_{ii} + N_{ii}/|A|, & \text{for all } i \in X \\
[U_{ij}] & w \geqslant V_{ij} + N_{ij}/|A|, & \text{for all } i \neq j \in X \\
& z \geqslant 0 &
\end{array} \tag{6}$$

or equivalently,

$$\begin{array}{rl}
\text{Minimize} & w + z \cdot |X|^{\delta} \cdot |X|^{-1} \\
\text{subject to:} & \\
& V \succeq 0, \\
& zI + wJ \geqslant V + N/|A|, \\
& z \geqslant 0.
\end{array} \tag{7}$$

where I is the identity matrix and J is the all 1's matrix over $X \times X$.

Any dual solution of (7) yields an upper bound on the optimum of (4) and hence $OPT_{M,\delta}$ and $\tau_M(\delta)$. To simplify the complexity of analysis we restrict ourselves to considering semidefinite matrices V that are suitably chosen Laplacian matrices. For any set S in $X \times X$ and any $\alpha : S \to \mathbb{R}_+$ the Laplacian matrix associated with S and α is defined by $L_{(S,\alpha)} := \sum_{(i,j)\in S} \alpha(i,j) L_{ij}$ where $L_{ij} = (e_i - e_j)(e_i - e_j)^T$ for the standard basis $\{e_i\}_{i\in X}$. Intuitively, in the dual SDP (7), by adding matrix $V = L_{S,\alpha}$ for suitable S and α depending on M we can shift weight from the off-diagonal entries of N to the diagonal where they can be covered by the $z + w$ entries on the diagonal rather than being covered by the w values in the off-diagonal entries. This will be advantageous for us since the objective function has much smaller coefficient for z which helps cover the diagonal entries than coefficient for w, which is all that covers the off-diagonal entries.

Definition 6.1. *Suppose that $N \in \mathbb{R}^{X\times X}$ is a symmetric matrix. For $\kappa \in \mathbb{R}_+$, define*

$$W_\kappa(N) = \max_{i\in X} \sum_{j\in X:\ N_{i,j}>\kappa} (N_{i,j} - \kappa).$$

The following lemma is the basis for our bounds on $\tau_M(\delta)$.

Lemma 6.2. *Let $\kappa \in \mathbb{R}_+$. Then*

$$OPT_{M,\delta} \leqslant (\kappa + W_\kappa(N) \cdot |X|^{\delta-1})/|A|.$$

Proof. For each off-diagonal entry of N with $N(i,j) > \kappa$, include matrix L_{ij} with coefficient $(N(i,j) - \kappa)/|A|$ in the sum for the Laplacian V. By construction, the matrix $V + N/|A|$ has off-diagonal entries at most $\kappa/|A|$ and diagonal entries at most $(\kappa + W_\kappa(N))/|A|$. The solution to (7) with $w = \kappa/|A|$ and $z = W_\kappa(N)/|A|$ is therefore feasible, which yields the bound as required. □

It may not be easy to bound $W_\kappa(N)$ directly, since the real part of M^*M may not have good structure. Fortunately, we have the following measure:

Definition 6.3. *Let $M \in \mathbb{C}^{A \times X}$ be a complex matrix. For $\kappa \in \mathbb{R}_+$, define*

$$\tilde{W}_\kappa(M) = \max_{i \in X} \sum_{j \in X\colon |(M^*M)_{i,j}| > \kappa} (|(M^*M)_{i,j}| - \kappa)$$

.

Proposition 6.4. *Let $\kappa \in \mathbb{R}_+$. Then $W_\kappa(N) \leqslant \tilde{W}_\kappa(M)$*

Proof. Whenever $N_{i,j} > \kappa$, we have $|(M^*M)_{i,j}| \geqslant N_{i,j} > \kappa$. Moreover, this gives $|(M^*M)_{i,j}| - \kappa \geqslant N_{i,j} - \kappa$. Then the statement follows the two definitions. □

For specific matrices M, we obtain the required bounds on $\tau_M(\delta) < 0$ for some $0 < \delta < 1$ by showing that we can set $\kappa = |A|^\gamma$ for some $\gamma < 1$ and obtain that $W_\kappa(N)$ or $\tilde{W}_\kappa(M)$ is at most $\kappa \cdot |X|^{\gamma'}$ for some $\gamma' < 1$.

7 Applications to Learning Polynomial Functions over Finite Fields

In this section, we prove all the bounds on the norm amplification curves needed to obtain the lower bounds on learning polynomials discussed in Section 4. We use the strategy in Section 6 by studying the W_κ function for matrices associated with learning polynomials over finite fields. We show that the values of this function are determined by the weight distribution and expected bias of these polynomials.

7.1 The Bias of $\mathbb{F}_2$ Polynomials and the Weight Distribution of Reed-Muller Codes

Let $d \geqslant 2$ be an integer. For any integer $n \geqslant d$, consider the learning problem for $\mathbb{F}_2$ polynomials in n variables of degree at most d, That is $A = \mathbb{F}_2^n$ and, expressing polynomials by their coefficients, we have $X = \mathbb{F}_2^m$ where $m = \sum_{i=0}^{d} \binom{n}{d}$ and for $a \in A$ and $x \in X$, $L(a, x) = x(a) = \sum_{S: 0 \leqslant |S| \leqslant d} x_S \prod_{i \in S} a_i$ over $\mathbb{F}_2$.

Recall that since the range of L is $\{0, 1\}$, we have a $N = M^T \cdot M$ where $M(a, x) = (-1)^{L(a,x)} = (-1)^{x(a)}$. Let M_x denote the x-th column of M where $x \in \{0, 1\}^n$. Then $N_{xy} = 2^n \cdot \langle M_x, M_y \rangle$.

Proposition 7.1. *Let $\mathbf{0} = 0^m$. Then $\langle M_x, M_y \rangle = \langle M_\mathbf{0}, M_{x+y} \rangle$.*

Proof.

$$\begin{aligned}\langle M_x, M_y \rangle &= \mathbf{E}_{a \in \mathbb{F}_2^n} M_x(a) M_y(a) = \mathbf{E}_{a \in \mathbb{F}_2^n} (-1)^{x(a)} (-1)^{y(a)} = \mathbf{E}_{a \in \mathbb{F}_2^n} (-1)^{x(a)+y(a)} \\ &= \mathbf{E}_{a \in \mathbb{F}_2^n} (-1)^{(x+y)(a)} = \mathbf{E}_{a \in \mathbb{F}_2^n} M_\mathbf{0}(a) M_{x+y}(a) = \langle M_\mathbf{0}, M_{x+y} \rangle\end{aligned}$$

□

Since the mapping $y \mapsto x + y$ for $x \in \mathbb{F}_2^m$ is 1-1 on $\mathbb{F}_2^m$, every row of N_x for $x \in X$ contains the same multi-set of values. Therefore, in order to analyze the function $W_\kappa(N)$, we only need to examine the fixed row $N_\mathbf{0}$ of N, where each entry

$$N_{\mathbf{0}x} = \sum_{a \in \mathbb{F}_2^n} M(a, x) = \sum_{a \in \mathbb{F}_2^n} (-1)^{x(a)}.$$

For $x \in X$, define $\mathrm{weight}(x) = |\{a \in \mathbb{F}_2^n : x(a) = 1\}|$. By definition, for $x \in \mathbb{F}_2^m$, $N_{\mathbf{0}x} = \sum_{a \in \mathbb{F}_2^n} (-1)^{x(a)} = 2^n - 2 \cdot \mathrm{weight}(x)$. Thus, understanding the function $W_\kappa(N)$ that we use to derive our bounds via Theorem 4.3 reduces to understanding the distribution of $\mathrm{weight}(x)$ for $x \in X$. In particular, our goal of showing that for some κ for which $(\kappa + W_\kappa(N))/2^n$ is at most $2^{2\tau n}$ for some $\tau < 0$ follows by showing that the distribution of $\mathrm{weight}(x)$ is tightly concentrated around $2^n/2$.

We can express this question in terms of the Reed-Muller error-correcting code $RM(d,n)$ over $\mathbb{F}_2$ (see, e.g. [4]).

Definition 7.2. *The Reed-Muller code $RM(d,n)$ over $\mathbb{F}_2$ is the set of vectors $\{G \cdot x \mid x \in \{0,1\}^m\}$ where G is the $2^n \times m$ matrix for $m = \sum_{t=0}^{d} \binom{n}{t}$ over $\mathbb{F}_2$ with rows indexed by vectors $a \in \{0,1\}^n$ and columns indexed by subsets $S \subseteq [n]$ with $|S| \leqslant d$ given by $G(a,S) = \prod_{i \in S} a_i$.*

Evaluating $\mathrm{weight}(x)$ for all $x \in \{0,1\}^m$ is that of understanding the distribution of Hamming weights of the vectors in $RM(d,n)$, a question with a long history.

Quadratic polynomials over $\mathbb{F}_2$ For the special case that $d = 2$, Sloane and Berlekamp [16] derived an exact enumeration of the number of vectors of each weight in $RM(2,n)$.

Proposition 7.3 ([16])**.** *The weight of every codeword of $RM(2,n)$ is of the form $2^{n-1} \pm 2^{n-i}$ for some integer i with $1 \leqslant i \leqslant \lceil n/2 \rceil$ or precisely 2^{n-1} and the number of codewords of weight $2^{n-1} + 2^{n-i}$ or $2^{n-1} - 2^{n-i}$ is precisely*

$$2^{i(i+1)} \prod_{j=0}^{i-1} \frac{2^{n-2j}(2^{n-2j-1}-1)}{2^{2(j+1)}-1}.$$

(Though the original proof used other methods, a simpler alternative proof by McEliece [10] follows from a lemma of Dickson [5] giving a normal form theorem for quadratic polynomials over $\mathbb{F}_{2^t}$. We will use a similar approach when we analyze quadratic polynomials over $\mathbb{F}_{p^t}$.)

Proof of Theorem 4.2 (a). Let the threshold $\kappa = 2^{n-k}$ for some integer k to be determined later. By Lemma 6.2 with $X = \{0,1\}^m$, for (3), we have $OPT_{M,\delta} \leqslant (\kappa + W_\kappa(N) 2^{(\delta-1)m})/2^n$ where $N = M^T \cdot M$. By definition for all $x \in X$ we have $N_{\mathbf{0}x} = 2^n - 2 \cdot \mathrm{weight}(x)$ and by Proposition 7.3, we know that if $2^n - 2 \cdot \mathrm{weight}(x) > 0$ then it is 2^{n-i+1} for some $1 \leqslant i \leqslant \lceil n/2 \rceil$. Also by Proposition 7.3, the number, c_i, of $x \in X$ such that $N_{\mathbf{0}x} = 2^{n-i+1}$ is at most

$$2^{i(i+1)} \prod_{j=0}^{i-1} \frac{2^{n-2j}(2^{n-2j-1}-1)}{2^{2(j+1)}-1} \leqslant 2^{2(i-1)n}.$$

Therefore, by definition of W_κ and Proposition 7.1, for any $x \in X$ we have

$$W_\kappa(N) \leqslant \sum_{y \in X : N_{xy} > 2^{n-k}} N_{xy} = \sum_{i=1}^{k} c_i \cdot 2^{n-i+1} \leqslant \sum_{i=1}^{k} 2^{2(i-1)n} \cdot 2^{n-i+1} = \sum_{i=1}^{k} 2^{(2n-1)(i-1)+n} < 2^{(2n-1)k}.$$

Thus for any k,

$$OPT_{M,\delta} \leqslant (2^{n-k} + 2^{(2n-1)k + (\delta-1)m})/2^n \leqslant 2^{-k} + 2^{(2n-1)k - (1-\delta)n(n+1)/2 - n}.$$

The first term is larger for $k \leqslant (1-\delta)n/4 + (3-\delta)/4$ so to balance them as much as possible we choose $k = \lfloor (1-\delta)n/4 + (3-\delta)/4 \rfloor \geqslant (1-\delta)n/4 - (1+\delta)/4$. Hence $OPT_{M,\delta} \leqslant 2 \cdot 2^{-k} \leqslant 2^{-\frac{1-\delta}{4}n+\frac{5+\delta}{4}}$ Therefore, $\tau_M(\delta) = \frac{1}{2}\log_{2^n} OPT_{M,\delta} \leqslant -\frac{(1-\delta)}{8} + \frac{(5+\delta)}{8n}$ as required. □

Polynomials of degree $d > 2$ over $\mathbb{F}_2$ For the case that $d > 2$, the minimum distance, the smallest weight of a non-zero codeword, in $RM(d,n)$ is known to be 2^{n-d} but for $2 < d < n-2$, no exact enumeration of the weight distribution of the code $RM(d,n)$ is known. It was a longstanding problem even to approximate the number of codewords of different weights in $RM(d,n)$. Relatively recently, bounds on these weights (or more precisely the associated biases) that are good enough for our purposes were shown by Ben-Eliezer, Hod, and Lovett [3].

Proposition 7.4. *For $\varepsilon > 0$ there are constants c_1, c_2 with $0 < c_1, c_2 < 1$ such that if p is a uniformly random degree d polynomial over $\mathbb{F}_2^n$ and $d \leqslant (1-\varepsilon)n$ then*

$$\mathbf{Pr}[|\mathbf{E}_{a\in\{0,1\}^n}(-1)^{p(a)}| > 2^{-c_1 n/d}] \leqslant 2^{-c_2\sum_{i=0}^{d}\binom{n}{i}}.$$

From this form we can obtain the bound on the norm amplification curve of the associated matrix fairly directly.

Proof of Theorem 4.2 (b). Fix $\varepsilon > 0$ and let $0 < c_1, c_2 < 1$ be the constants depending on ε from Proposition 7.4. Let $\delta = c_2/2$ so $0 < \delta < 1/2$. Let M be the $2^n \times 2^m$ matrix associated with learning polynomials of degree at most d over $\mathbb{F}_2$, let $N = M^T \cdot M$ and Setting $\kappa = 2^{(1-c_1/d)n}$, by Proposition 7.4 at most $2^{(1-c_2)m}$ polynomials p have entries N_{0p} larger than κ. Each such entry has value at most 2^n so $W_\kappa(N) \leqslant 2^n \cdot 2^{(1-c_2)m}$. by Lemma 6.2 with $X = \{0,1\}^m$ we have

$$OPT_{M,\delta} \leqslant (\kappa + W_\kappa(N)\cdot 2^{(\delta-1)m})/2^n \leqslant 2^{-c_1 n/d} + 2^{(\delta-c_2)m+1} \leqslant 2^{-c_1 n/d} + 2^{1-\delta m}$$

which is at most $2^{-c' n/d}$ for some constant $c' > 0$. Hence $\tau_M(\delta) \leqslant -c'/d$. □

7.2 The Bias of $\mathbb{F}_p$ Polynomials for Odd Prime p

Let $d \geqslant 1$ be an integer and p be an odd prime. For any integer $n \geqslant d$, consider the learning problem for $\mathbb{F}_p$ polynomials in n variables of degree at most d. Unlike the case over $\mathbb{F}_2$, the monomials are not necessarily multilinear but can have degree at most $p-1$ in each variable. Let $\mathcal{M}_p(d,n)$ be the set of monomials in n variables of total degree at most d and degree at most $p-1$ in each variable. That is $A = \mathbb{F}_p^n$ and, expressing polynomials by their coefficients, we have $X = \mathbb{F}_p^m$ where $m = |\mathcal{M}_p(d,n)|$ is the number of monomials of total degree at most d and degree at most $p-1$ in each variable. As in the case of $\mathbb{F}_2$, m is the dimension of a Reed-Muller code $RM_p(d,n)$ over $\mathbb{F}_p$, and for $a \in A$ and $x \in X$, $L(a,x) = x(a) \in \mathbb{F}_p$. For $d \geqslant p$ there is no convenient closed form known for $|\mathcal{M}_p(d,n)|$ but the following is known:

Proposition 7.5. *For $d < p$, $|\mathcal{M}_p(d,n)| = \binom{n+d}{d}$ and for $2 < p \leqslant d \leqslant n$, $\sum_{i=0}^{d}\binom{n}{d} \leqslant |\mathcal{M}_p(d,n)| \leqslant \binom{n+d}{d}$.*

Since $p > 2$, the learning problem for $\mathbb{F}_p$ polynomials is governed by $p-1$ complex matrices $M^{(1)}, \ldots, M^{(p-1)}$ where $M^{(j)}(a,x) = \omega^{j \cdot x(a)}$ and $\omega = e^{2\pi i/p}$. We need to bound the norm amplification curves of all these matrices. We will relate these curves to the values of $\text{bias}_j(x)$ for $j \in \{1, \ldots, p-1\}$ and $x \in X$, where

$$\text{bias}_j(x) = \mathbf{E}_{a \in_R A} \omega^{j \cdot x(a)}.$$

Fix an arbitrary $j^* \in \{1, \ldots, p-1\}$, For $N = (M^{(j^*)})^* \cdot M^{(j^*)}$, the (x,y) entry of N is $p^n \langle M_x^{(j^*)}, M_y^{(j^*)} \rangle$ where $\langle \cdot, \cdot \rangle$ is the complex inner product.

Proposition 7.6. *Let $\mathbf{0} = 0^m$. Then for $x, y \in X$, $\langle M_x^{(j^*)}, M_y^{(j^*)} \rangle = \langle M_{\mathbf{0}}^{(j^*)}, M_{y-x}^{(j^*)} \rangle$.*

Proof.

$$\begin{aligned}\langle M_x^{(j^*)}, M_y^{(j^*)} \rangle &= \mathbf{E}_{a \in \mathbb{F}_p^n} \overline{M_x^{(j^*)}(a)} M_y^{(j^*)}(a) = \mathbf{E}_{a \in \mathbb{F}_p^n} \omega^{-j^* \cdot x(a)} \omega^{j^* \cdot y(a)} = \mathbf{E}_{a \in \mathbb{F}_p^n} \omega^{-j^* \cdot x(a) + j^* \cdot y(a)} \\ &= \mathbf{E}_{a \in \mathbb{F}_p^n} \omega^{j^*(y-x)(a)} = \mathbf{E}_{a \in \mathbb{F}_p^n} M_{\mathbf{0}}^{(j^*)}(a) M_{y-x}^{(j^*)}(a) = \langle M_{\mathbf{0}}^{(j^*)}, M_{y-x}^{(j^*)} \rangle\end{aligned}$$

□

Since the mapping $y \mapsto y - x$ for $x \in \mathbb{F}_p^m$ is 1-1 on $\mathbb{F}_p^m$, every row of N_x for $x \in X$ contains the same multi-set of values. Therefore, in order to analyze the function $\tilde{W}_\kappa(N)$, we only need to examine the fixed row $N_{\mathbf{0}}$ of N. where each entry

$$N_{\mathbf{0}x} = \sum_{a \in \mathbb{F}_p^n} \omega^{j^* \cdot x(a)} = p^n \cdot \text{bias}_{j^*}(x).$$

Therefore we have shown the following:

Lemma 7.7. *Let $j^* \in \{1, \ldots, p-1\}$. For every $v \in \mathbb{C}$, the number of entries in each row of $N = (M^{(j^*)})^* \cdot M^{(j^*)}$ equal to v is precisely the number of polynomials $x \in X$ such that $p^n \cdot \text{bias}_{j^*}(x) = v$.*

Therefore, to bound $\tilde{W}_\kappa(N)$ it suffices to bound the numbers of polynomials $x \in X$ such that $|\text{bias}_{j^*}(x)|$ is large.

Affine Functions over $\mathbb{F}_p$ For $d = 1$, an $x \in X = \mathbb{F}_p^{n+1}$ yields the function $x(a) = x_0 + \sum_{i=1}^n x_i a_i$. Unless $x_1 = \cdots = x_n = 0$, for every $k \in \mathbb{F}_p$ we have exactly p^{n-1} values $a \in \mathbb{F}_p^n$ for which $x(a) = k$ and hence $\text{bias}_{j^*}(x) = 0$. For each of the remaining p inputs with $x_1 = \cdots = x_n = 0$ and different values for x_0, we get $\text{bias}_{j^*}(x) = \omega^{j^* \cdot x_0}$ and hence $|\text{bias}_{j^*}(x)| = 1$. In this case we choose $\kappa = 0$ and observe that $\tilde{W}_0(N) = p^{n+1}$. Therefore for any δ with $0 \leqslant \delta \leqslant 1$, we have

$$OPT_{M^{(j^*)},\delta} \leqslant p^{n+1}|X|^{\delta-1}/|A| = p^{1+(\delta-1)(n+1)} = (p^n)^{-(1-\delta)+\delta/n},$$

so $\tau_{M^{(j^*)}}(\delta) = \frac{1}{2}\log_{|A|} OPT_{M^{(j^*)},\delta} \leqslant -\frac{1-\delta}{2} + \frac{\delta}{2n}$. This proves Theorem 4.4 (a). If we only took linear functions instead of affine functions, all non-zero x would be balanced and the term $\frac{\delta}{2n}$ would not appear. (This is the analog of the parity learning bound for higher moduli.) This proves Theorem 4.4 (b).

Quadratic Polynomials over $\mathbb{F}_p$

Lemma 7.8. *Let p be an odd prime and $n \geqslant 2$ be an integer. Let X be the set of quadratic polynomials over $A = \mathbb{F}_p^n$. Then for $j^* \in \{1, \ldots, p-1\}$,*

1. *For any $x \in X$, $\mathrm{bias}_{j^*}(x) = 0$ or $|\mathrm{bias}_{j^*}(x)| \in \{p^{-n/2}, p^{(n-1)/2}, \cdots, p^{-1/2}, 1\}$.*
2. *For $0 \leqslant k \leqslant n$ the number of $x \in X$ such that $|\mathrm{bias}_{j^*}(x)| = p^{-k/2}$ is less than $p^{kn+2k+1}$.*

To prove Lemma 7.8 we start with the following structure lemma for quadratic polynomials over fields of odd characteristic. This lemma is an easier analog of Dickson's Lemma for characteristic 2 [5] and is well known but we include a proof for completeness.

Lemma 7.9. *Let p be an odd prime and integer $t \geqslant 1$. For every quadratic polynomial q over $\mathbb{F}_{p^t}$ in variables $z = (z_1, \ldots, z_n)$, there is an invertible affine transformation T over $\mathbb{F}_{p^t}$ such that for $z' = T(z)$, there is a unique $k \leqslant n$, and $(c_1, \ldots, c_k) \in \{1, \cdots, p-1\}^k$, and an affine form ℓ over $\mathbb{F}_{p^t}$ in $n-k$ variables such that:*

$$q(z) = \sum_{i=1}^{k} c_i z_i'^2 + \ell(z'_{k+1}, \cdots, z'_n)$$

Proof. We show this by induction on n. The statement is clearly true when $n = 0$. Assume that this is true for any polynomial in $n-1$ variables. We have several cases when q has n variables:
CASE 1: q is affine: Then the statement is true with $k = 0$.
CASE 2: q contains some square term $b_i \cdot z_i^2$: In this case we can write q as $b_i \cdot z_i^2 + \ell_i \cdot z_i + q'$, where ℓ_i is affine, q' is a quadratic polynomial, and neither of them involves z_i. Then we can define

$$z_i' = z_i + 2^{-1} b_i^{-1} \cdot \ell_i$$

since b_i^{-1} and 2^{-1} are defined in field $\mathbb{F}_{p^t}$ because $b_i \neq 0$ and the characteristic p is odd. Also define $q'' = q' - 2^{-2} b_i^{-1} \ell_i^2$. Thus

$$\begin{aligned}
&b_i(z_i')^2 + q'' \\
&= b_i(z_i')^2 + q' - 2^{-2} b_i^{-1} \ell_i^2 \\
&= b_i(z_i + 2^{-1} b_i^{-1} \cdot \ell_i)^2 + q' - 2^{-2} b_i^{-1} \ell_i^2 \\
&= b_i(z_i^2 + b_i^{-1} \ell_i \cdot z_i + 2^{-2} b_i^{-2} \ell_i^2) + q' - 2^{-2} b_i^{-1} \ell_i^2 \\
&= b_i \cdot z_i^2 + \ell_i \cdot z_i + q' = q.
\end{aligned}$$

Define T_i to be the map which sets $z_j' = z_j$ for $j \neq i$ and replaces z_i with z_i' according to the above formula. Clearly by the definition of z_i', T_i is an affine map; moreover, it is invertible, with T_i^{-1} setting $z_i = z_i' - 2^{-1} b_i^{-1} \cdot \ell_i$ and leaving all other z_j for $j \neq i$ unchanged. By definition, q'' is a quadratic form defined only on the $m-1$ variables $z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_m$, a property inherited from q' and ℓ_i. Let P_{in} be the permutation that swaps positions i and n and leaves the rest alone and define $q''' = P_{in}(q'')$.

We now can apply the inductive hypothesis to q''' and derive that there is an invertible affine mapping T' on the $n-1$ variables (excluding z_i) and some k' together with constants $a'_1, \dots, a'_{k'} \in \mathbb{F}_p^*$, yielding variables $z''_1, \dots, z''_{n-1}$ as affine functions of the previous values such that

$$q''' = \sum_{j=1}^{k'} c_i {z''_j}^2 + \ell''(z''_{k'+1}, \dots, z''_{n-1}).$$

We can extend T' to an affine transformation T'' on n variables by keeping the n-th variable unchanged.

Finally, define $k = k'+1$, $c_k = b_i$ and the invertible affine transformation, $T = P_{nk} \circ T'' \circ P_{in} \circ T_i$ where P_{nk} is the permutation that swaps positions k and n. Then $T(z) = (z''_1, \dots, z''_{k-1}, z'_i, z''_{k+1}, \dots, z''_{n-1}, z''_k)$.

$$T(q) = \sum_{j=1}^{k-1} c_i {z''_j}^2 + c_k (z'_i)^2 + \ell''(z''_{k+1}, \dots, z''_{n-1}, z''_k)$$

which is of the required form.

CASE 3: q has no squared terms and is not affine. Then q must contain some cross term $b_{ij} \cdot z_i z_j$ for $i \neq j$. Here we can use the identity

$$z_i z_j = 2^{-2} \cdot ((z_i + z_j)^2 - (z_i - z_j)^2)$$

and let S_{ij} be the affine mapping that leaves all other variables unchanged and assigns $z'_i = 2^{-1}(z_i + z_j)$ and $z'_j = 2^{-1}(z_i - z_j)$ which exists since 2 is invertible over $\mathbb{F}_{p^t}$. S_{ij} is clearly invertible since $z_i = z'_i + z'_j$ and $z_j = z'_i - z'_j$. Hency, for $z' = S_{ij}(z)$, we have $q(z) = q_{ij}(z')$ for some quadratic q_{ij} that has two squared terms $(z'_i)^2$ and $(z'_j)^2$ and hence is covered by Case 2 above. Let T_2 be the resulting affine transformation derived for q_{ij}. It follows that $T = T_2 \circ S_{ij}$ is the required transformation for q. □

Lemma 7.9 provides a clean way of studying the bias of quadratic polynomials. For any invertible affine mapping T on $\mathbb{F}_p^n$, for $y \in X$ and $x(z) = y(T(z))$, we have $x \in X$ and

$$\text{bias}_{j^*}(x) = \mathbf{E}_{a \in \mathbb{F}_p^n} \omega^{j^* \cdot x(a)} = \mathbf{E}_{a \in \mathbb{F}_p^n} \omega^{j^* \cdot y(T(a))} = \mathbf{E}_{b \in \mathbb{F}_p^n} \omega^{j^* y(b)} = \text{bias}_{j^*}(y)$$

since T is a bijection on $\mathbb{F}_p^n$.

We therefore first analyze the polynomials of the normal form in Lemma 7.9. Let $y(z) = \sum_{i=1}^k c_i z_i^2 + \ell(z_{k+1}, \dots, z_n)$ where each $c_1, \dots, c_k \neq 0$. Write $\ell = c_0 + \sum_{i=k+1}^n c_i z_i$. If there is any j with $k+1 \leqslant j \leqslant n$ such that $c_j \neq 0$ then then $\text{bias}_{j^*}(y) = 0$ just as in the affine case. Therefore it remains to consider

$$y(z) = \sum_{i=1}^k c_i z_i^2 + c_0 \quad \text{for } c_1, \dots, c_k \in \mathbb{F}_p^*,\ c_0 \in \mathbb{F}_p. \tag{8}$$

Observe that the number of such y is $(p-1)^k p < p^{k+1}$. Furthermore,

$$p^n \cdot |\text{bias}_{j^*}(y)| = |\sum_{a \in \mathbb{F}_p^n} \omega^{j^* \cdot \sum_{i=1}^k c_i a_i^2 + c_0}| = p^{n-k} \cdot \prod_{i=1}^k |\sum_{a_i=0}^{p-1} \omega^{j^* \cdot c_i a_i^2}|$$

The term $\sum_{a_i=0}^{p-1} \omega^{j^* \cdot c_i a_i^2}$ in the product is called a *quadratic Gauss sum* and has been studied previously. For our purpose, we need the following result:

Proposition 7.10 (Proposition 6.3.2 in [8]). *Let p be an odd prime. For $c \in \{1, \cdots, p-1\}$,*

$$|\sum_{j=0}^{p-1} \omega^{cj^2}| = \sqrt{p}.$$

Therefore setting $c = c_i \cdot j^*$ for the i-th term, we have $|\text{bias}_{j^*}(y)| = p^{-k/2}$. We now put things together to prove Lemma 7.8.

Proof of Lemma 7.8. By Lemma 7.9, since bias_{j^*} is preserved under invertible linear transformations T of the inputs, it follows that every polynomial x such that $\text{bias}_{j^*}(x) \neq 0$ must have $|\text{bias}_{j^*}(x)| = p^{-k/2}$ for some non-negative integer $k \leqslant n$. Moreover, the number of polynomials x whose normal form y of the form (8) is at most the number of affine transformations that define $z'_1, \ldots, z'_k$ in terms of $z_1, \ldots, z_k$ which is $(p^{n+1})^k$ since there are precisely p^{n+1} affine functions on $\mathbb{F}_p^n$. Therefore the total number of x such that $\text{bias}_{j^*}(x) = p^{-k/2}$ is less than $p^{(n+1))k} \cdot p^{k+1} = p^{nk+2k+1}$. □

Now we can use Proposition 7.8 to prove part (c) of Theorem 4.4.

Proof of Theorem 4.4 (c). Proposition 7.8 implies that for $N = (M^{(j^*)}) * \cdot M^{(j^*)}$,

$$W_{p^{n-k/2}}(N) \leqslant \tilde{W}_{p^{n-k/2}}(N) \leqslant \sum_{t=0}^{k-1} p^{n-t/2} \cdot p^{tn+2t+1} = p^{n+1} \cdot \frac{p^{kn+3k/2}-1}{p^{n+3/2}-1} \leqslant p^{kn+3k/2}$$

Therefore if we set $k = \lfloor \frac{1-\delta}{2} n \rfloor \geqslant \frac{1-\delta}{2} n - 1$, then by Lemma 6.2, since $|X| = p^{\binom{n+2}{2}}$ we have

$$OPT_{M^{(j^*)},\delta} \leqslant p^{-k} + p^{-n} \cdot p^{kn+3k/2+(\delta-1)\binom{n+2}{2}} \leqslant 2p^{-k} \leqslant p^{-\frac{1-\delta}{2}n+2}$$

Therefore,

$$\tau_{M^{(j^*)}}(\delta) = \frac{1}{2} \log_{p^n} OPT_{M^{(j^*)},\delta} \leqslant \frac{1-\delta}{4} + \frac{1}{n}$$

Since j^* was an arbitrary fixed element of $\{1, \ldots, p-1\}$, the theorem follows. □

Polynomials of degree $d > 2$ over $\mathbb{F}_p$ Similar to the $\mathbb{F}_2$ case, we need to understand the weight distribution of Reed-Muller codes over $\mathbb{F}_p$. In our companion paper [2] we give the following estimate which is the analogue for odd prime fields of the bounds of Ben-Eliezer, Hod, and Lovett [3].

Proposition 7.11 ([2]). *For $0 < \varepsilon < 1/2$, for all $j \in \mathbb{F}_p^*$, there are constants c_1, c_2 depending on ε with $0 < c_1, c_2 < 1$ such that if f is a uniformly random degree d polynomial over $\mathbb{F}_p^n$ and $d \leqslant \varepsilon n$ then*

$$\mathbf{Pr}[|\text{bias}_{j^*}(f)| > p^{-c_1 n/d}] \leqslant p^{-c_2 m}.$$

From this form we can obtain the bound on the norm amplification curve of the associated matrix fairly directly, and complete our proof of Theorem 4.4.

Proof of Theorem 4.4 (d). Fix $\varepsilon > 0$, $j \in \mathbb{F}_p^*$, and let $0 < c_1, c_2 < 1$ be the constants depending on ε from Proposition 7.4. Let $\delta = c_2/2$ so $0 < \delta < 1/2$. For $N = (M^{(j^*)})^* \cdot M^{(j^*)}$, when we set $\kappa = p^{(1-c_1/d)n}$, Proposition 7.11 implies that at most $p^{(1-c_2)m}$ polynomials f satisfy $|N_{0f}| > \kappa$. The norm of each entry is at most p^n so $\tilde{W}_\kappa(N) \leqslant p^n \cdot p^{(1-c_2)m}$. by Lemma 6.2 with $X = \mathbb{F}_p^m$ we have

$$OPT_{M,\delta} \leqslant (\kappa + W_\kappa(N) \cdot p^{(\delta-1)m})/p^n \leqslant p^{-c_1 n/d} + p^{(\delta-c_2)m+1} \leqslant p^{-c_1 n/d} + p^{1-\delta m}$$

which is at most $p^{-c' n/d}$ for some constant $c' > 0$. Hence $\tau_{M^{(j^*)}}(\delta) \leqslant -c'/d$. □

References

[1] Paul Beame, Shayan Oveis Gharan, and Xin Yang. Time-space tradeoffs for learning from small test spaces: Learning low degree polynomial functions. Technical Report TR17-120, Electronic Colloquium on Computational Complexity (ECCC), 2017.

[2] Paul Beame, Shayan Oveis Gharan, and Xin Yang. On the bias of Reed-Muller codes over odd prime fields. *ArXiv*, 2018.

[3] Ido Ben-Eliezer, Rani Hod, and Shachar Lovett. Random low-degree polynomials are hard to approximate. *Computational Complexity*, 21(1):63–81, 2012.

[4] Elwyn R Berlekamp. *Algebraic Coding Theory*. McGraw-Hill, 1968.

[5] Leonard E. Dickson. *Linear Groups with an Exposition of the Galois Field Theory*. B.G. Trubner, Leipzig, 1901.

[6] Sumegha Garg, Ran Raz, and Avishay Tal. Extractor-based time-space tradeoffs for learning. Technical Report TR17-121, Electronic Colloquium on Computational Complexity (ECCC), 2017.

[7] Sumegha Garg, Ran Raz, and Avishay Tal. Extractor-based time-space lower bounds for learning. In *Proceedings of the Fiftieth Annual ACM on Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA*, 2018. To appear.

[8] Kenneth Ireland and Michael Rosen. *A classical introduction to modern number theory*, volume 84. Springer Science & Business Media, 2013.

[9] Gillat Kol, Ran Raz, and Avishay Tal. 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*, pages 1067–1080, 2017.

[10] Robert James McEliece. *Linear recurring sequences over finite fields*. PhD thesis, California Institute of Technology, 1967.

[11] Dana Moshkovitz and Michal Moshkovitz. Mixing implies lower bounds for space bounded learning. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, pages 1516–1566, 2017.

[12] Dana Moshkovitz and Michal Moshkovitz. 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*, pages 28:1–28:20, 2018.

[13] Ran Raz. Fast learning requires good memory: A time-space lower bound for parity learning. In *Proceedings, 57th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2016, New Brunswick, New Jersey, USA*, pages 266–275, October 2016.

[14] Ran Raz. A time-space lower bound for a large class of learning problems. In *Proceedings, 58th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, California, USA*, pages 732–742, October 2017.

[15] Ohad Shamir. 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*, pages 163–171, Montreal, Quebec, Canada, 2014.

[16] Neil J. A. Sloane and Elwyn R. Berlekamp. Weight enumerator for second-order Reed-Muller codes. *IEEE Trans. Information Theory*, 16(6):745–751, 1970.

[17] Jacob Steinhardt, Gregory Valiant, and Stefan Wager. Memory, communication, and statistical queries. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA*, pages 1490–1516, 2016.