A Maximum Matching Algorithm for Basis Selection in Spectral Learning

Ariadna Quattoni and **Xavier Carreras** and **Matthias Gallé**
Xerox Research Centre Europe (XRCE)
Meylan, France
{ariadna.quattoni,xavier.carreras,matthias.galle}@xrce.xerox.com

Abstract

We present a solution to scale spectral algorithms for learning sequence functions. We are interested in the case where these functions are sparse (that is, for most sequences they return 0). Spectral algorithms reduce the learning problem to the task of computing an SVD decomposition over a special type of matrix called the Hankel matrix. This matrix is designed to capture the relevant statistics of the training sequences. What is crucial is that to capture long range dependencies we must consider very large Hankel matrices. Thus the computation of the SVD becomes a critical bottleneck. Our solution finds a subset of rows and columns of the Hankel that realizes a compact and informative Hankel submatrix. The novelty lies in the way that this subset is selected: we exploit a maximal bipartite matching combinatorial algorithm to look for a sub-block with full *structural* rank, and show how computation of this sub-block can be further improved by exploiting the specific structure of Hankel matrices.

1 INTRODUCTION

Our goal is to model functions whose domain are discrete sequences over some finite alphabet. Our focus is on sparse functions, by which we mean functions that have the property that only a very small proportion of the sequences in the domain map to a non-zero value. We call those sequences the support of the function. The main motivation lies in solving problems arising in Natural Language Processing (NLP) applications, where sparse sequence functions are of special interest. For example, think of all possible sequences of T letters that constitute valid English words of length T. If Σ is the set of English letters, is clear that out of the Σ^T possible letter sequences only a very small fraction are valid words (i.e. should have non-zero probability).

One interesting function class over $\Sigma^\star$ is that of functions computed by Non-Deterministic Weighted Automata (WA), since this class properly includes classes such as ngram models and hidden Markov models. In recent years several approaches for estimating WAs have been proposed that are based on representing the function computed by a WA using a Hankel matrix [Beimel et al., 2000, Jaeger, 2000, Hsu et al., 2009, Anandkumar et al., 2012, Balle et al., 2013].

As an illustration of the method, consider the following problem: Assume we are given a set of pairs $(x, f(x))$, where x is a sequence in the support of some target function f over $\Sigma^\star$ and we wish to learn a WA that approximates f. The spectral method provides a solution to this problem and it would work in four steps:

1. Basis Selection: Choose a set of prefixes $\mathcal{P}$ and suffixes $\mathcal{S}$.
2. Build a Hankel matrix: $\mathbf{H} \in \mathbb{R}^{|\mathcal{P}|\times|\mathcal{S}|}$ where the entry $\mathbf{H}(p, s)$ is the value of the target function on the sequence obtained by concatenating prefix p with suffix s.
3. Perform SVD on $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$.
4. Use the factorization $\mathbf{F} = \mathbf{U}\mathbf{\Sigma}$ and $\mathbf{B} = \mathbf{V}^\top$ and $\mathbf{H}$ to recover the parameters of the minimal WA, following Hsu et al. [2009] (see §2.3 for details).

The computational cost of the algorithm will be dominated by the SVD step $\mathcal{O}\big(\min(|\mathcal{P}|, |\mathcal{S}|)^3\big)$, thus to control the computational complexity, it is critical to choose a small and yet informative basis.

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

The theory of spectral learning tells us that if the target function has a minimal WA representation of size n, there will be a complete basis where $|\mathcal{P}| = |\mathcal{S}| = n$, where complete means that the rank of the corresponding Hankel defined over that basis is the same as the size of the minimal WA. But the theory does not give a practical answer to how to choose such a basis. The design of efficient algorithms for choosing an informative and yet small sample-dependent basis is still an open problem, which is the focus of our paper.

We propose an efficient combinatorial algorithm for sample-dependent basis selection. At its core, our strategy computes a maximum matching of the bipartite graph associated with the sparsity pattern of a Hankel matrix. The main idea is quite simple, we find a subset of prefixes and suffixes in the given sample, such that the corresponding Hankel matrix defined over that basis has full structural rank. The key insight is that for sparse matrices it is easy to remove symbolic dependencies (i.e. dependencies at the level of the sparsity pattern of the matrix). Similar ideas have a long history in the numerical optimization literature, where combinatorial algorithms are used for computing preconditioners for solving large sparse linear systems. However, to the best of our knowledge we are the first ones in applying this idea in the context of spectral learning.

We show that when the Hankel matrix of a function satisfies some non-degeneracy assumptions, our basis selection algorithm is optimal, in the sense that it computes the smallest complete basis. While the non-degeneracy assumption will not be always satisfied, our experiments suggest that it is always almost satisfied for sparse sequence functions.

Our experiments on a real sequence modeling tasks show that the proposed algorithm can select a basis that is at least an order of magnitude smaller than the best alternative methods for basis selection, resulting in an SVD step which is at least two orders of magnitude faster.

1.1 Related Work

Although choosing a basis is in practice an important task for having a robust spectral learning algorithm, not much research has focused on this problem. One popular approach is to choose a basis by selecting all observed prefixes and suffixes of length less than T, for some $T > 0$ [Hsu et al., 2009, Siddiqi et al., 2010]. In practice, this strategy only works if there are no long-range dependencies in the target function. Wiewiora [2005] presented a greedy heuristic where for each prefix added to the basis a computation taking exponential time in the number of states n is required. Bailly et al. [2009] suggest to include all observed prefixes and suffixes (observed in the sample) in the basis. There are some theoretical results [Denis et al., 2016] that suggest that under certain assumptions this is the optimal strategy, in the sense that there is no *statistical harm* in considering all prefixes and suffixes. However, this approach is in practice unfeasible: to give a concrete example if one considers modeling the distribution of n-grams up to length 10 in a standard NLP benchmark, the unique number of observed prefixes and suffixes is at least tens of millions. Finally, Balle et al. [2012] gave the first theoretical results for the problem of basis selection. They show that by sampling prefixes and suffixes proportional to their frequency in a large enough sample, with high probability, a complete basis will be found. They also provide experimental results [Balle et al., 2013].

2 PRELIMINARIES

2.1 Non Deterministic Weighted Finite State Automata

We start by defining a class of functions over discrete sequences. More specifically, let $x = x_1 \cdots x_t$ be a sequence of length t over some finite alphabet Σ. We use $\Sigma^\star$ to denote the set of all finite sequences with elements in Σ, and we use ϵ to denote the empty sequence. The domain of our functions is $\Sigma^\star$.

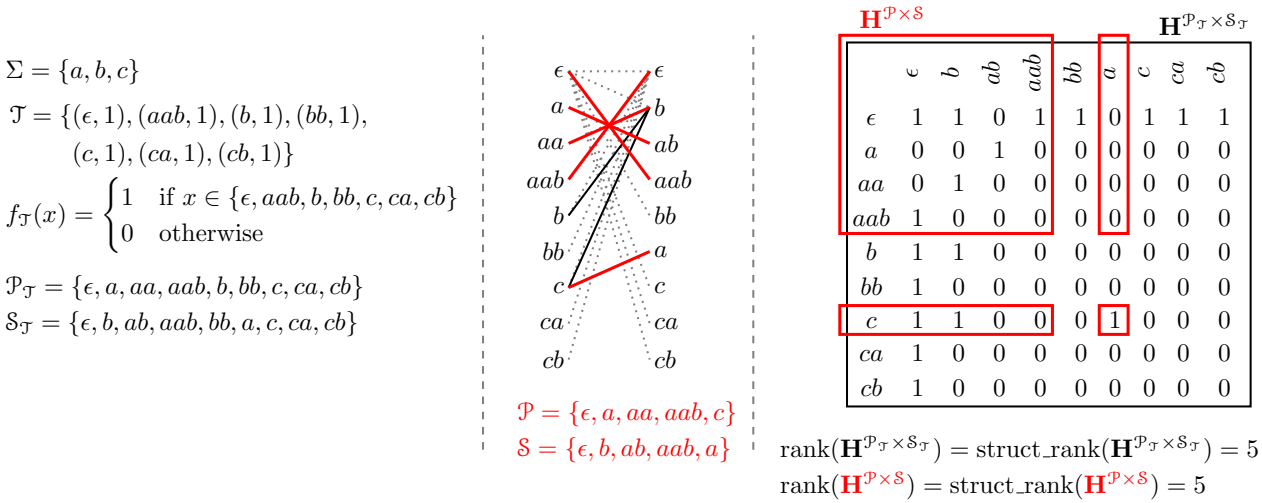
An Non-Deterministic Weighted Automaton (WA) with n states is defined as a tuple: $A = \langle \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_\infty, \{\mathbf{A}_\sigma\}_{\sigma\in\Sigma}\rangle$ where $\boldsymbol{\alpha}_0$, $\boldsymbol{\alpha}_\infty \in \mathbb{R}^n$ are the initial and final weight vectors and $A_\sigma \in \mathbb{R}^{n\times n}$ are the transition matrices associated to each symbol $\sigma \in \Sigma$. The function $f_A : \Sigma^\star \to \mathbb{R}$ realized by an WA A is defined as:

$$f_A(x) = \boldsymbol{\alpha}_0^\top \mathbf{A}_{x_1} \cdots \mathbf{A}_{x_t} \boldsymbol{\alpha}_\infty \quad . \tag{1}$$

The above equation is an algebraic representation of the computation performed by an WA on a sequence x. To see this consider a state vector $\mathbf{s}_i \in \mathbb{R}^n$ where the jth entry represents the sum of the weights of all the state paths that generate the prefix $x_{1:i}$ and end in state j. Initially, $\mathbf{s}_0 = \boldsymbol{\alpha}_0$, and then $\mathbf{s}_i^\top = \mathbf{s}_{i-1}^\top \mathbf{A}_{x_i}$ updates the state distribution by simultaneously emiting the symbol x_i and transitioning to generate the next state vector. WAs constitute a rich function class which properly includes popular sequence models such as HMMs.

2.2 Hankel Matrices

We now introduce the concept of Hankel matrices for WA, which are central to the spectral learning algorithm, and to the result in this paper.



	ϵ	b	ab	aab	bb	a	c	ca	cb
ϵ	1	1	0	1	1	0	1	1	1
a	0	0	1	0	0	0	0	0	0
aa	0	1	0	0	0	0	0	0	0
aab	1	0	0	0	0	0	0	0	0
b	1	1	0	0	0	0	0	0	0
bb	1	0	0	0	0	0	0	0	0
c	1	1	0	0	0	1	0	0	0
ca	1	0	0	0	0	0	0	0	0
cb	1	0	0	0	0	0	0	0	0

Figure 1: Illustration of the Maximum Bipartite Matching sub-block. Left: a training set and the associated target function. Middle: a prefix-suffix graph with a corresponding maximum matching in red. Right: the full Hankel matrix for the training set, and the submatrix given by the matching.

Let $f : \Sigma^\star \to \mathbb{R}$ be an arbitrary function from sequences to reals (not necessarily computed by a WA). Let $\mathcal{P}, \mathcal{S} \subseteq \Sigma^\star$ be sets of sequences. We call prefixes the elements $p \in \mathcal{P}$, and suffixes the elements $s \in \mathcal{S}$. The Hankel matrix $\mathbf{H}_f \in \mathbb{R}^{\mathcal{P}\times\mathcal{S}}$ for f over the block $(\mathcal{P}, \mathcal{S})$ is defined by entries $\mathbf{H}(p, s) = f(ps)$, where ps is the concatenation of prefix $p \in \mathcal{P}$ and suffix $s \in \mathcal{S}$. The following theorem gives a bijection between the class of functions computed by WA and Hankel matrices:

Theorem 1. *[Schützenberger, 1961, Carlyle and Paz, 1971, Fliess, 1974] A function $f : \Sigma^\star \to \mathbb{R}$ can be realized by a WA with n states if and only if, for every possible block $(\mathcal{P}, \mathcal{S})$, the corresponding Hankel matrix $\mathbf{H}_f$ has rank at most n.*

2.3 The Spectral Method

We now give a brief description of the spectral method for estimating a minimal WA representation for a target function. The algorithm is a constructive version of the theorem above: it builds a Hankel matrix of rank n and computes the associated n state WA from it. We only provide a higher-level description of the method; for a complete derivation and the theory justifying the algorithm we refer the reader to the works by Hsu et al. [2009] and Balle et al. [2013].

Assume a training set $\mathcal{T}$ in the form of a collection of sequences, each paired with a target real value. We will denote as $f_{\mathcal{T}}$ the function obtained from the training set, i.e. if $x \in \mathcal{T}$, $f_{\mathcal{T}}(x)$ is the target value. For example, $\mathcal{T}$ could be a corpus of English sentences, and $f_{\mathcal{T}}(x)$ the probability with which x appears in $\mathcal{T}$.

Given a training set $\mathcal{T}$, the spectral algorithm computes a WA A with n states, where n is a parameter of the algorithm, such that f_A is a good approximation of $f_{\mathcal{T}}$. See Hsu et al. [2009] for the generalization theory of the algorithm. The method is described by the following steps:

(1) Select a Hankel block. Let $\mathcal{P}_{\mathcal{T}}$ and $\mathcal{S}_{\mathcal{T}}$ be respectively the sets of all unique prefixes and suffixes of sequences in $\mathcal{T}$. Select a block out of them, namely, a subset of prefixes $\mathcal{P} \subseteq \mathcal{P}_{\mathcal{T}}$ and a subset of suffixes $\mathcal{S} \in \mathcal{S}_{\mathcal{T}}$.

(2) Compute Hankel matrices for $(\mathcal{P}, \mathcal{S})$.

  (a) Compute $\mathbf{H} \in \mathbb{R}^{\mathcal{P}\times\mathcal{S}}$, with entries $\mathbf{H}(p, s) = f_{\mathcal{T}}(ps)$.

  (b) Compute $\mathbf{h}_{\mathcal{P}} \in \mathbb{R}^{\mathcal{P}}$ with $\mathbf{h}_{\mathcal{P}}(p) = f_{\mathcal{T}}(p)$ and $\mathbf{h}_{\mathcal{S}} \in \mathbb{R}^{\mathcal{S}}$ with $\mathbf{h}_{\mathcal{S}}(s) = f_{\mathcal{T}}(s)$.

  (c) For each $\sigma \in \Sigma$, compute $\mathbf{H}_\sigma \in \mathbb{R}^{\mathcal{P}\times\mathcal{S}}$ with entries $\mathbf{H}_\sigma(p, s) = f_{\mathcal{T}}(p\sigma s)$.

(3) Compute an n-rank factorization of $\mathbf{H}$. Compute the truncated SVD of $\mathbf{H}$, i.e. $\mathbf{H} \approx \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ resulting in a matrix $\mathbf{F} = \mathbf{U}\mathbf{\Sigma} \in \mathbb{R}^{\mathcal{P}\times n}$ and a matrix $\mathbf{B} = \mathbf{V} \in \mathbb{R}^{\mathcal{S}\times n}$.

(4) Recover the WA A of n states. Let $\mathbf{M}^+$ denote the Moore-Penrose pseudo-inverse of a matrix $\mathbf{M}$. The elements of A are recovered as follows. Initial vector: $\boldsymbol{\alpha}_0^\top = \mathbf{h}_{\mathcal{S}}^\top \mathbf{B}$. Final vector: $\boldsymbol{\alpha}_\infty = \mathbf{F}^+ \mathbf{h}_{\mathcal{P}}$. Transition Matrices: $\mathbf{A}_\sigma = \mathbf{F}^+ \mathbf{H}_\sigma \mathbf{B}$, for $\sigma \in \Sigma$.

There are some observations to make that motivate the contribution of this paper. Consider the complete

training block $(\mathcal{P}_\mathcal{T}, \mathcal{S}_\mathcal{T})$, and let $\mathbf{H}^\mathcal{T}$ denote the Hankel matrix for this complete block. If we want to fully reconstruct the function $f_\mathcal{T}$, we need an automata A that has as many states as the rank of $\mathbf{H}^\mathcal{T}$. By using less states, we will be learning a low-rank approximation of $f_\mathcal{T}$ in the form of a WA.

The second observation is that any sub-block $(\mathcal{P}, \mathcal{S})$ whose Hankel submatrix has full rank (with respect to the rank of $\mathbf{H}^\mathcal{T}$) can be used to fully recover $f_\mathcal{T}$.

Thus, in the ideal case, step (1) of the algorithm would select a compact submatrix of $\mathbf{H}^\mathcal{T}$ that preserves the rank. By doing so, the cost of steps (4) and (5) would only depend on the size of the submatrix. Even if we can not get the ideal block, it would be good to have a method for step (1) that produces a small and informative block. Unfortunately, in the general case (i.e. for any real matrix) finding the submatrix of fixed size that has maximal rank is known to be NP-complete [Peeters, 2003]. In this paper we propose an algorithm to find a small submatrix of $\mathbf{H}^\mathcal{T}$ of high rank.

As a final note, spectral methods can be used to learn a language model, that is, a probability distribution over all sentences of a language. A straightforward way to learn a language model is to regard the training collection $\mathcal{T}$ as an empirical distribution over sequences of words, where the probability of a sequence is proportional to the number of times it appears, i.e. $f_\mathcal{T}(x) = \Pr_\mathcal{T}(x)$. Another choice, sometimes referred to as moment matching, is to set the function $f_\mathcal{T}(x)$ to be the expected number of times that the sequence x appears as a subsequence of a random sequence sampled from an empirical distribution. In this case, the spectral algorithm will learn a WA that computes expectations of subsequence frequencies. One useful result is that this WA can be converted to another WA that corresponds to the underlying language model, i.e. a distribution over sequences; see Balle et al. [2013] for details. In practice this second method is preferred, since subsequence frequency expectations are statistics that are more stable to estimate from a training set.

3 SUB-BLOCK SELECTION VIA BEST BIPARTITE MATCHING

We start this section by defining the structural rank of a matrix. Our proposed algorithm will search for a submatrix of $\mathbf{H}$ with full structural rank. The structural rank of a matrix is the maximum rank of all numerical matrices with the same non-zero pattern. In the context of WA and Hankel matrices this has a nice interpretation as a notion of *complexity* of the support of a function. This is because the structural rank of a Hankel matrix corresponds to the number of states of the minimal WA for the *hardest* function defined over that support.

Notice that by definition, the numerical rank of a matrix is always less or equal than its structural rank, thus the structural rank of the Hankel matrix $\mathbf{H}$ of a function f_A will be always greater or equal than the number of states of the minimal WA computing f_A. Our algorithm is based on finding a submatrix of $\mathbf{H}$ of full structural rank.

The problem of finding a full structural rank sub-block of $\mathbf{H}$ can be casted as an instance of maximum bipartite matching [Edmonds, 1967]. Given a bipartite graph (V, G) where V are the set of vertices and G the set of edges, the maximum bipartite matching is defined as the largest set of non-intersecting edges, where *non-intersecting* means that no two edges in the set share a common vertex.

In the case of the Hankel matrix for a function f_A we would have a bipartite graph (V, G) where on one side we have vertices corresponding to all unique prefixes in the support of f_A and on the other side we have all unique suffixes, thus: $|V| = |\mathcal{P}| + |\mathcal{S}|$. There will be an edge connecting node i an j if the corresponding sequence made by the concatenation of prefix i and suffix j is in the support of f_A. For every sequence s of length T in the support of f_A and every possible cut of s into a prefix and a suffix, there will be $T + 1$ corresponding edges in G, thus $|G| = O(T|f_A|)$ where we use $|f_A|$ to refer to the number of sequences in the support of f_A.

The maximum bipartite matching of a set of sequences is a subset of the sequences such that no two sequences share a common prefix or suffix and there is no larger subset that satisfies that property. Figure 1 shows an example of a function f_A and its corresponding graph, and a maximum bipartite matching for that graph.

We define the *maximum bipartite matching sub-block* as the block consisting of all vertices (prefixes and suffixes) in a maximum matching. Figure 1 shows an example of a function, a maximum bipartite matching, and the corresponding sub-block and Hankel submatrix.

To find a maximum bipartite matching there are several classical algorithms. The Augmented Paths algorithm runs in $\mathcal{O}\big(|V||E|\big)$, but in practice it has a much lower average case complexity. The Hopcroft-Karp algorithm runs in $\mathcal{O}\big(|E|\sqrt{|V|}\big)$, removing the linear dependence on V (however, in our experiments the Augmented Paths algorithm was alredy very fast). In the next section we propose an algoritm that takes advantage of the structure of the Hankel matrix to obtain further speed ups.

3.1 On the Optimality of the Maximum Matching Sub-block

We will use a weak version the *matching property*, an assumption used by Hoffman and McCormick [1982]. Let $\mathbf{M}$ be a matrix of structural rank s. $\mathbf{M}$ has the weak matching property (WMP) if for any submatrix $\mathbf{M}'$ of at least s rows and s columns, the rank of $\mathbf{M}'$ is equal to the structural rank of $\mathbf{M}'$.

Lemma 1. *Let $\mathbf{H}$ be a Hankel matrix that satisfies the weak matching property. Let $\mathcal{B}$ be a maximum bipartite matching of $\mathbf{H}$ and let $\mathbf{H}^{\mathcal{B}}$ be the corresponding submatrix. $\mathcal{B}$ is a basis of $\mathbf{H}$, i.e. the rank of $\mathbf{H}^{\mathcal{B}}$ is equal to the rank of $\mathbf{H}$.*

Proof. Let $s(\mathbf{M})$ be the structural rank of a matrix $\mathbf{M}$. Let n be the rank of $\mathbf{H}$, and note that $s(\mathbf{H})$ is n because $\mathbf{H}$ has WMP. Now note that $s(\mathbf{H}^{\mathcal{B}})$ is also n, because the maximum bipartite matching of $\mathbf{H}$ is included in $\mathbf{H}^{\mathcal{B}}$, thus $s(\mathbf{H}^{\mathcal{B}})$ is at least n; and it is at most n, otherwise $s(\mathbf{H}) \geq s(\mathbf{H}^{\mathcal{B}}) > n$. Since $\mathbf{H}$ has WMP, the rank of $\mathbf{H}^{\mathcal{B}}$ is n. □

Ideally, we would not have to assume the *matching property* and instead we could provide theoretical guarantees for the maximum gap between the structural and numeric rank of a matrix. Unfortunately, because of the discrete nature of the structural rank, deriving useful bounds for this gap has been shown to be a hard theoretical challenge [Hoffman and McCormick, 1982]. Thus to provide validation for our assumption we resorted to an empirical evaluation of the gap on a wide range of sequence modeling datasets, where we observe that the weak matching property is a reasonable assumption. The complete results are in section B of the supplementary materials.

4 FASTER BIPARTITE MATCHING FOR HANKEL MATRICES

As said in the previous section, finding the structural rank can be reduced to the maximum bipartite matching problem. In this section, we propose a simple heuristic to speed-up the maximum bipartite matching for the specific case where the underlying matrix is a Hankel. We do this by exploiting structural properties of these matrices for an underlying subroutine, the *augmenting path* algorithm. Each basic application of the augmenting path increases the matching by one, and a matching is maximum if and only if there is no further augmenting path. The straightforward solution of applying it on each node is equivalent to the maximal flow algorithm, and while more sophisticated algorithms where proposed [Hopcroft and Karp, 1973] which find several paths per iteration, benchmarks [Setubal, 1996] have shown that the simple algorithm works in general faster.

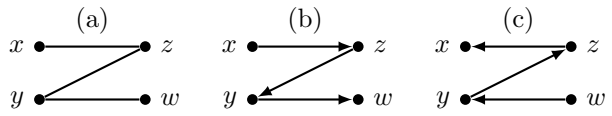


Figure 2: Illustrations of the augmenting path algorithm.

We first describe the basic procedure: assume the graph depicted in Figure 2 (a), and furthermore assume that the current matching (not maximal) is as follows: $M = \{(y, z)\}$. This is clearly not maximal, as a better (and maximal) matching would be $\{(x, z), (y, w)\}$. The augmenting path procedure maps the previous matching to the *directed* graph G depicted in Figure 2 (b).

Unselected edges will be directed from left to right, while selected edges will be directed from right to left. An augmenting path is then defined as a path $x_1, \ldots, x_m$ over G, such that x_1 belongs to the left partition, x_m to the right one, and both x_1 and x_m are unmatched (this is, do not belong yet to a matching). No restrictions are put on the intermediate nodes, but it becomes clear that the path alternates between unmatched pairs (left to right edges) and matched pairs (right to left). Note that such paths can now easily be retrieved with a standard graph traversal (in our implementation we use a depth-first search, which we assumed was faster on sparse graphs although this was not verified). Starting from node x, the following path can then be retrieved: x, z, y, w, and the graph will then be rewired to the graph depicted in Figure 2 (c). No further augmenting paths exist here, and the maximum matching algorithm therefore finishes with the following matching: $\{(x, z), (y, w)\}$.

The specific case where the left part of the bipartite are prefixes and the right part are suffixes creates some strong structural constraints. Notably:

Property 1. *$(p\sigma, s)$ is an edge in the graph iff $(p, \sigma s)$ is an edge*

This is, the edges of the bipartite graph denoting a Hankel matrix come by (possibly overlapping) groups of edges, each group originating in one of the support sequences.

We propose to take advantage of that structural knowledge to speed-up the maximum matching algorithm. First, we sort the prefixes by their lengths, and start applying the augmenting path procedure from the longest prefix node. Each augmenting path procedure returns a set of edges R to be removed from the matching, and a set of edges A to be added to the matching.

For each edge $(\sigma_1 \dots \sigma_k, s) \in A$ we consider all *shifted* pairs $(\sigma_1 \dots \sigma_i, \sigma_{i+1} \dots \sigma_k s)$. Due to Property 1, each one of these pairs is an edge in the bipartite graph. We check each such pair, and if both nodes are unmatched we simply add them to the matching.

Assuming a bitset implementation of sets, the checks can be done in $\mathcal{O}(|E|)$, but in the worst-case scenario, it may well be that none of the shifted pairs are free, and therefore only add computation without improving the matching. In § B of the supplementary material we report synthetic experiments that show the speed-ups of this strategy compared to the standard method.

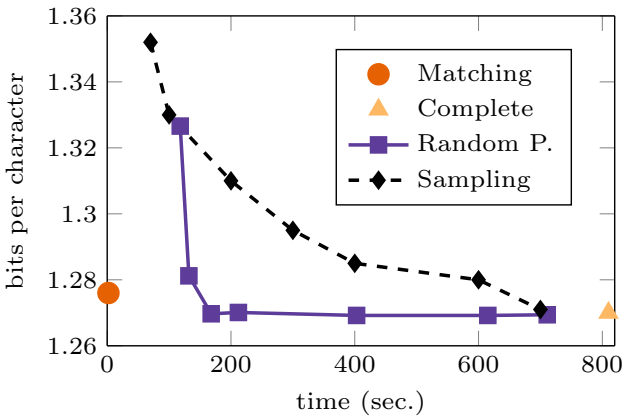


Figure 3: Comparison of Strategies for Scaling Spectral Learning.

5 EXPERIMENTS

To validate our sub-block selection strategy, we present comparisons to methods for scaling up spectral learning. We first compare to general methods to scale SVD, and then to sub-block selection strategies for Hankel matrices. We end this section with a comparison to state-of-the-art methods on the SPiCE benchmark.

In all experiments we use natural language data for the task of language modeling. The goal is to learn a language model that predicts the next symbol for a sentence prefix (including ending the sentence). As evaluation metric we use *Bits per Character (BpC)*, the average log-2 probability that the model gives to each symbol in the evaluation senquences, including sequence ends. As datasets we use the English Penn Treebank [Marcus et al., 1994] using standard splits[1], the War and Peace dataset [Karpathy et al., 2016][2], and the NLP datasets of the SPiCe benchmark [Balle et al., 2016].

5.1 Scalable SVD Methods

We conducted experiments comparing our method with two other strategies for scaling SVD. The first uses *Randomized Projections* to perform SVD [N. Halko and Tropp., 2009]. This idea was previously used to scale spectral learning [Hamilton et al., 2013]. The second strategy is based on *Sampling*, and selects the k top rows and columns that have the highest norm [Deshpande and Vempala., 2006].

For this comparison we used the Penn TreeBank dataset with simplified part-of-speech tags (12 symbols). We chose this dataset because it results in a relatively small Hankel matrix where we can run sparse SVD. In particular, we used moment size of $T = 5$, which results in a square Hankel matrix of size 52,450, numeric rank of 312, and structural rank of 313. Thus, the *Complete* method will use run sparse SVD on this matrix.

We present a trade-off between performance (in terms of bits-per-character) and training time of a method. When appropriate, we generate solutions that utilize different amounts of time. For Sampling, since it selects k rows and columns proportional to their norm, a natural way of generating different solutions is to vary k. For Randomized Projections we do not select a sub-block, instead we project the Hankel matrix to a lower ℓ-dimensional space and then the SVD is performed on the projected matrix. Thus to get performance as a function of training cost we can change the size of the projection.

The training time[3] of a method consists of: (1) time spent in selecting the Hankel sub-block (for algorithms that start by sub-block selection (e.g. best matching); (2) time spent on computing the singular value decomposition; and (3) time spent computing inverses, i.e. recovering operators. Notice that all spectral methods will perform SVD of a Hankel sub-block. Whenever we compute SVD we take the cost of the most efficient (i.e. sparse or full SVD) to be the cost of the algorithm. Another important observation is that the sparse SVD algorithm takes as a parameter the number of singular values to compute. We take this to be the optimal number of states found using the validation data.

Figure 3 shows the trade-off, for the four methods. The first observation is that with sufficient amount of computational time both Random Projections and Sampling achieve the same performance as using the

[1] 49 characters; 5017k / 393k / 442k characters in the train / dev / test portions.

[2] 84 symbols; 2658k / 300k / 300k characters in the train / dev / test portions.

[3] All experiments were run on a 2.2 GHz Intel Core processor.

Complete Hankel. This is expected since by setting k and ℓ sufficiently large we should always obtain the same result as using the complete Hankel. Random Projections seems to be significantly better than Sampling in terms of speed up, and it can obtain the same solution as Complete in less than 1/4 of the time. Best Bipartite Matching obtains a slightly higher bits-per-character than Random Projections, but is significantly more efficient. More precisely, to achieve the same performance as with Matching, Random Projections requires about 50 times more time.

5.2 Sub-block Selection Strategies for Spectral Methods

We now present an empirical comparison between the most prevalent sub-block selection strategies for spectral learning.

We train spectral language models at character level that use a fixed window of T characters both at training and test time. At training, we collect all substrings x of length up to T. Following Balle et al. [2013], we set a target function $f_{\mathcal{T}}(x)$ to be the expected number of times that x appears as a subsequence of a random sentence sampled from training. We run the spectral algorithm with $f_{\mathcal{T}}$ and obtain a WA. At test, we run the WA to compute the probability of the next character given a sliding prefix of length $T-1$.

We compare maximum matching sub-block selection to three strategies: full block, random cuts, and length up-to. *Full block* uses all substrings of the support of $f_{\mathcal{T}}$ as prefixes and suffixes. *Random Cuts* follows Balle et al. [2012]: it samples a string x of the support, and chooses a random cut of x into a prefix and suffix, which are added to the sub-block. This process is repeated until the sub-block reaches size k (a parameter). *Length* $\leq \ell$ selects all substrings up to length ℓ.

Table 1 compares sub-block selection methods in terms of the numeric rank of the sub-matrix, the time it takes to compute an n-rank factorization, and the quality of the resulting n-state WA in terms of bits per character (BpC). n is a parameter that we tune on validation data with a range of values up to the rank of the sub-matrix. The matching sub-block obtains results that are very close to using the full matrix. However, it is much faster: the time to compute a matching is negligible, and the time to factorize the matrix is three orders of magnitude faster. Compared to other strategies, the matching sub-block is the most accurate and the most compact, and thus it is drastically faster. This improvement is achieved because it selects a very compact sub-block (of size 1,661) that has approximately full numeric rank (rank is 1,612). In contrast, using Random Cuts, a block of the same size as the matching sub-block (1,661) has only a rank of 739, which results in lower quality predictions. When increasing the block of Random Cuts up to 7 times the size of the matching, we obtain a rank of 1,114 and very close results to the matching and full sub-blocks; however, factorizing the sub-matrix is 100 times more costly. Sub-block selection by maximum length also performs poorly. This last result is evidence that long-range statistical dependencies exist in this data, and these are not captured by small moments. On the other hand, a brute-force approach to capture such long range dependencies is prohibitive. Our method clearly offers a very competitive solution.

Table 1: Comparison Between Sub-block Selection Methods for Support Strings up to Size $T = 5$

method	**size**	**rank**	**sec.**	**BpC**
Full	144,378	-	18,000	1.735
Matching	1,661	1,612	8	1.741
Random Cuts 1×	1,661	739	10	2.011
Random Cuts 2×	3,322	807	74	1.828
Random Cuts 3×	4,983	902	163	1.812
Random Cuts 4×	6,664	989	271	1.791
Random Cuts 5×	8,305	1,010	302	1.769
Random Cuts 6×	9,966	1,086	411	1.761
Random Cuts 7×	11,627	1,114	825	1.752
Length ≤ 2	861	92	2	3.105
Length ≤ 3	7,455	417	290	2.662
Length ≤ 4	38,314	907	3,500	1.856

Table 2: Results of spectral models for increasing length of strings in the support

	Penn Treebank		War and Peace (Lik)		
T	Full	Match.	Full	Match.	KJL16
5	1.735	1.741	1.377	1.405	1.451
6	1.623	1.653	1.326	1.393	1.339
7	1.597	1.622	1.323	1.369	1.321

Next we present results of models trained on larger substrings, of up to size 7, for the Penn Treebank and War and Peace datasets. Table 2 compares the performance of the matching sub-block to using the full block.[4] As we increase the size of the substrings (T) the models get better. There is always a performance gap between using the full or the matching blocks, however the matching sub-block scales much better:

[4]For the War and Peace data, we measure performance in terms of test negative log-likelihood, such that we can compare to published results.

Table 3: Results on NLP Datasets of SPiCe Sequence Prediction Competition.

	Verbs	LM (words)	LM (characters)	POS	Normalization	All
RNN-P	***0.6078***	***0.5434***	***0.8101***	***0.6573***	***0.5882***	***0.6414***
COMBO-NN-1	0.5794	**0.5014**	0.7632	0.6331	0.5181	0.5990
COMBO-B	0.5514	0.4264	**0.7978**	0.5890	0.3843	0.5498
LSTM	0.5123	0.4034	0.7630	0.5941	0.4187	0.5383
COMBO-Sp	0.5273	0.4148	0.6142	0.6235	0.4990	0.5358
Sp-BM	**0.5928**	0.4998	0.7820	**0.6356**	**0.5441**	**0.6109**

the cost of computing a matching is negligible (less than 15 seconds), and the cost of the factorization is at three orders of magnitude faster. Table 2 also compares to the results by Karpathy et al. [2016] (noted KJL16), in terms of negative log-likelihood on test characters (noted Lik). We report their results, corresponding to non-recurrent feed-forward neural models, which condition each prediction on the $T-1$ latest characters (see Table 2 of their paper). The results are fairly comparable, exhibiting the same trend.

5.3 Comparison with State-of-the-Art

In order to compare the performance of our proposed method to other state of the art methods for sparse sequence modeling, we run experiments on the five NLP datasets of the SPiCe sequence prediction competition [Balle et al., 2016]. The task of the competition was the following: given a string (prefix) of symbols in a finite alphabet the goal is to predict a ranking of possible next symbols to be the next element of the sequence. The metric used for evaluation measures the average ranking that the model gives to the correct next symbol.[5] Both validation and test sets are available from the challenge website.

There were a total of 26 teams implementing a wide range of methods, including: many different types of neural network models, boosting, spectral and classical state-merging algorithms for learning weighted automaton, and ensemble methods that combined several techniques.

Table 3 shows results for the top 5 teams of the competition. The top team (RNN-P) is a novel RNN architecture where the state vector is augmented with an indicator vector representing the previous ngram in the history. The second best team (COMBO-NN) is an ensemble of MLP, CNN, LSTM and ngram models. The third team (COMBO-B) is also an ensemble method of ngram, spectral, RNN and tree boosting. The fourth team (LSTM) is an RNN with LSTM cells and the fifth team (COMBO-Sp) is another ensemble method that combines a spectral model with ngram models.

The performance of our spectral method with the proposed sub-block selection using best bipartite matching (Sp-BM) is given in the last row. We indicate with bold and stars the top performing method for each dataset and with bold the second best. Running the proposed algorithm out-of-the-box and without any model combination we get a very competitive performance: second best overall (0.6414 vs 0.6109) and second in 3 out of 5 datasets. One of the most attractive properties of our method is that the most costly training times (those corresponding to datasets with Hankel matrices of higher structural rank) were less than 5 minutes.

6 CONCLUSIONS

We presented a novel strategy for scaling spectral learning algorithms that is specifically designed for modeling long range dependencies in sparse sequence functions. The main idea is to use maximal bipartite matching to find a Hankel sub-block of maximal structural rank. Our experiments on a real sequence modeling task show that: (1) Exploiting large Hankel matrices is essential for the success of spectral learning algorithms; and that: (2) Our proposed sub-block selection strategy to handle large Hankel matrices can be much faster than using sparse SVD over the complete Hankel matrix without a significant loss in performance. Our algorithm leads to a very appealing trade-off between computational complexity and model performance.

References

Animashree Anandkumar, Daniel J. Hsu, and Sham M. Kakade. A method of moments for mixture models and hidden markov models. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, volume 23 of *JMLR Proceedings*, pages 33.1–33.34. JMLR.org, 2012.

[5] We refer the reader to the SPiCe benchmark website: `http://spice.lif.univ-mrs.fr`.

R. Bailly, F. Denis, and L. Ralaivola. Grammatical inference as a principal component analysis problem. In *Proc. ICML*, 2009.

B. Balle, X. Carreras, F.M. Luque, and A. Quattoni. Spectral learning of weighted automata: A forward-backward perspective. *Machine Learning*, 2013.

Borja Balle, Ariadna Quattoni, and Xavier Carreras. Local loss optimization in operator models: A new insight into spectral learning. In *ICML '12*, 2012.

Borja Balle, Rémi Eyraud, Franco M. Luque, Ariadna Quattoni, and Sicco Verwer. Results of the sequence prediction challenge (spice): a competition on learning the next symbol in a sequence. In *Proceedings of the 13th International Conference on Grammatical Inference*, 2016.

A. Beimel, F. Bergadano, N.H. Bshouty, E. Kushilevitz, and S. Varricchio. Learning functions represented as multiplicity automata. *JACM*, 2000.

J. W. Carlyle and M A. Paz. Realizations by stochastic finite automata. *Journal of Computer Systems Science*, 1971.

François Denis, Mattias Gybels, and Amaury Habrard. Dimension-free concentration bounds on hankel matrices for spectral learning. *Journal of Machine Learning Research*, 17(31):1–32, 2016. URL `http://jmlr.org/papers/v17/14-501.html`.

A. Deshpande and S. Vempala. Adaptive sampling and fast low-rank matrix approximation. *RANDOM 06*, 2006.

J. Edmonds. Systems of distinct representatives and linear algebra. *Journal of Research of the National Bureau of Standards*, 71B(4):241–245, 1967.

M. Fliess. Matrices de Hankel. *Journal de Mathématiques Pures et Appliquées*, 1974.

William L. Hamilton, Mahdi M. Fard, and Joelle Pineau. Modelling sparse dynamical systems with compressed predictive state representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 178–186. JMLR Workshop and Conference Proceedings, 2013.

A.J. Hoffman and S. T. McCormick. A fast algorithm that makes matrices optimally sparse. Technical Report 13, Stanford University Systems Optimization Laboratory Report, 1982.

John E. Hopcroft and Richard M. Karp. An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. *SIAM Journal on Computing*, 2(4):225–231, 1973.

D. Hsu, S. M. Kakade, and T. Zhang. A spectral algorithm for learning hidden markov models. In *Proc. of COLT*, 2009.

H. Jaeger. Observable operator models for discrete stochastic time series. *Neural Computation*, 12: 1371–1398, 2000.

Andrej Karpathy, Justin Johnson, and Fei-Fei Li. Visualizing and understanding recurrent networks. In *ICLR Workshop Track*, 2016.

Mitchell P. Marcus, Beatrice Santorini, and Mary A. Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19, 1994.

P.G. Martisson N. Halko and J. A. Tropp. Finding structure with randomness: Stochastic algorithms for constructing aproximate matrix decompositions. *arXiv: 0909.4061*, 2009.

R. Peeters. The maximum edge biclique problem is np-complete. *Discrete Applied Mathematics*, 2003.

M. P. Schützenberger. On the definition of a family of automata. *Information and Control*, 1961.

Joao C. Setubal. Sequential and parallel experimental results with bipartite matching algorithms, 1996.

Sajid Siddiqi, Byron Boots, and Geoffrey J. Gordon. Reduced-rank hidden Markov models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS-2010)*, 2010.

Eric Wiewiora. Learning predictive representations from a history. In *Proceedings of the 22nd international conference on Machine learning*, pages 964–971. ACM, 2005.