Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Decoding Hidden Markov Models Faster than Viterbi Via Online Matrix-Vector (max, +)-Multiplication

Massimo Cairo
Department of Mathematics
University of Trento
Trento, Italy
massimo.cairo@unitn.it

Gabriele Farina
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano
I-20133, Milan, Italy
gabriele2.farina@mail.polimi.it

Romeo Rizzi
Department of Computer Science
University of Verona
Verona, Italy
romeo.rizzi@univr.it

Abstract

In this paper, we present a novel algorithm for the maximum *a posteriori* decoding (MAPD) of time-homogeneous Hidden Markov Models (HMM), improving the worst-case running time of the classical Viterbi algorithm by a logarithmic factor. In our approach, we interpret the Viterbi algorithm as a repeated computation of matrix-vector $(\max, +)$-multiplications. On time-homogeneous HMMs, this computation is *online*: a matrix, known in advance, has to be multiplied with several vectors revealed one at a time. Our main contribution is an algorithm solving this version of matrix-vector $(\max, +)$-multiplication in subquadratic time, by performing a polynomial preprocessing of the matrix. Employing this fast multiplication algorithm, we solve the MAPD problem in $\mathcal{O}(mn^2/\log n)$ time for any time-homogeneous HMM of size n and observation sequence of length m, with an extra polynomial preprocessing cost negligible for $m > n$. To the best of our knowledge, this is the first algorithm for the MAPD problem requiring subquadratic time per observation, under the assumption – usually verified in practice – that the transition probability matrix does not change with time.

Introduction

Hidden Markov Models (HMMs) are simple probabilistic models originally introduced (Viterbi 1967) to decode convolutional codes. Due to their universal and fundamental nature, these models have successfully been applied in several fields, with many important applications, such as gene prediction (Haussler and Eeckman 1996), speech, gesture and optical character recognition (Gales 1998; Huang, Ariki, and Jack 1990; Starner, Weaver, and Pentland 1998; Agazzi and Kuo 1993), and part-of-speech tagging (Kupiec 1992). Their applications to bioinformatics began in the early 1990 and soon exploded to the point that currently they hold a recognized place in that field (Yoon 2009; Mäkinen et al. 2015).

A HMM describes a stochastic process generating a sequence of observations $y_1, y_2, \ldots, y_n$. Internally, a sequence of hidden states $x_1, x_2, \ldots, x_n$ is generated according to a *Markov chain*. At each time instant $t = 1, 2, \ldots, n$, a symbol y_t is observed according to a probability distribution depending on x_t. We consider only time-homogeneous HMMs, i.e. models whose parameters do not depend on the time t. While this assumption covers the majority of applications, some notable exceptions involving time-inhomogeneous models are known (Lafferty, McCallum, and Pereira 2001).

Maximum *a posteriori* decoding (MAPD). Since the states of the model are hidden, i.e. only the generated symbols can be observed, a natural problem associated with HMMs is the MAPD problem: *given a HMM $\mathcal{M}$ and an observed sequence of symbols Y of length m, find any state path X through $\mathcal{M}$ maximizing the joint probability of X and Y. We call any such X a* most probable state path *explaining the observation Y*. Traditionally, the MAPD problem is solved by the Viterbi algorithm (Viterbi 1967), in $\mathcal{O}(mn^2)$ time and $\mathcal{O}(mn)$ memory for any model of size n and observation sequence of length m.

Over the years, much effort has been put into lowering the cost of the Viterbi algorithm, both in terms of memory and of running time. (Grice, Hughey, and Speck 1997) showed that a checkpointing technique can be employed to reduce the memory complexity to $\mathcal{O}(\sqrt{m} \cdot n)$; refinements of this idea (embedded checkpointing) deliver a family of time-memory tradeoffs, culminating into an $\mathcal{O}(n \log m)$ memory solution with a slightly increased running time $\mathcal{O}(mn^2 \log m)$.

At the same time, several works reducing the time complexity of the algorithm in the average-case were developed (Šrámek 2007; Churbanov and Winters-Hilt 2008; Felzenszwalb, Huttenlocher, and Kleinberg 2004; Esposito and Radicioni 2009; Kaji et al. 2010). Many of these works make assumptions on the structure of $\mathcal{T}$, and may lose the optimality or degenerate to the worst case $\Theta(mn^2)$ opera-

tions when these assumptions are not fulfilled. To the best of our knowledge no algorithm achieving a worst-case running time better than $\mathcal{O}(mn^2)$ is known under the only assumption of time-homogeneousness.

Approach. We give an algorithm solving the MAPD problem for time-homogeneous HMMs with time complexity asymptotically lower than $\mathcal{O}(mn^2)$, *in the worst case*. We regard the MAPD problem as an iterated computation of a matrix-vector multiplication. For time-homogeneous models, the matrix is known in advance and does not change with time. However, the sequence of vectors to be multiplied cannot be foreseen, as each vector depends on the result of the previous computation; this rules out the possibility to batch the vectors into a matrix and defer the computation. We call this version of the problem, in which a fixed matrix has to be multiplied with several vectors revealed one at a time, "the online matrix-vector multiplication (OMV MUL) problem".

Consider the problem of multiplying a $n \times n$ matrix with a column vector of size n. Without further assumptions, the trivial $\mathcal{O}(n^2)$ time algorithm is optimal, since all the n^2 elements of the matrix have to be read at least once. However, under the assumption that the matrix is known in advance and can be preprocessed, this trivial lower bound ceases to hold. Algorithms faster than the trivial quadratic one are known for the OMV MUL problem over *finite* semirings (Williams 2007), as well as over real numbers with standard $(+, \cdot)$-multiplication, if the matrix has only a *constant* number of distinct values (Liberty and Zucker 2009).

However, none of the above algorithm can be applied to time-homogeneous HMMs, as their decoding relies on online real matrix-vector $(\max, +)$-multiplication (ORMV $(\max, +)$-MUL). In the specific case of real $(\max, +)$-multiplication, subcubic algorithms have been known for years (Dobosiewicz 1990; Chan 2008; 2015) for the *matrix-matrix* multiplication problem, with important applications to graph theory and boolean matrix multiplication, among others. However, we are not aware of any algorithm solving the ORMV $(\max, +)$-MUL problem in subquadratic time. Note that the ORMV $(\max, +)$-MUL can be used to compute the OMV MUL over the Boolean semiring: for this problem, it has been conjectured (Henzinger et al. 2015) that no "truly polynomially subquadratic" algorithm[1] exists for the ORMV $(\max, +)$-MUL problem.

We reduce the ORMV $(\max, +)$-MUL problem to a multi-dimensional geometric dominance problem, following an approach similar to that of (Bremner et al. 2006; Chan 2008). Then, the geometric problem is solved by a divide-and-conquer algorithm, which can be regarded as a transposition of the algorithm of (Chan 2008) to the online setting. Our technique yields a worst-case $\mathcal{O}(mn^2/\log n)$ algorithm, called GDFV, solving the MAPD problem after a polynomial preprocessing of the model.

Contributions. Our key contributions are as follows: (i) we extend the geometric dominance reporting problem introduced in (Chan 2008) to the online setting; (ii) we solve the ORMV $(\max, +)$-MUL problem in $\mathcal{O}(n^2/\log n)$ time after a polynomial preprocessing of the $n \times n$ matrix; (iii) we show an algorithm solving the MAPD problem on time-homogeneous HMMs in $\mathcal{O}(mn^2/\log n)$ time in the worst-case, after a polynomial preprocessing of the model.

Finally, we experimentally evaluate the performance of our algorithms, with encouraging results. Currently the problem sets in which we outperform Viterbi are limited, but we hope that the approach we propose will open the way to further improvements on this problem in future works.

[1] That is, running in time $\mathcal{O}(n^{2-\varepsilon})$ for some $\varepsilon > 0$ after a polynomial preprocessing or the matrix.

Preliminaries

Notation

The i-th component of a vector $\mathbf{v}$ is denoted by $\mathbf{v}[i]$; similarly, $\mathbf{M}[i, j]$ denotes the entry of row i and column j, in matrix $\mathbf{M}$. Indices will always be considered as starting from 1. Given two vectors $\mathbf{a}$ and $\mathbf{b}$ of dimension n, such that $\mathbf{a}[i] \leq \mathbf{b}[i]$ for every coordinate index $i = 1, \ldots, n$, we write $\mathbf{a} \preceq \mathbf{b}$ and say that $\mathbf{b}$ *dominates* $\mathbf{a}$, or, equivalently, that $(\mathbf{a}, \mathbf{b})$ is a *dominating pair*.

Given a matrix or vector $\mathbf{M}$ with non-negative entries, we write $\log \mathbf{M}$ to mean the matrix or vector that is obtained from $\mathbf{M}$ by applying the logarithm on every component. We will almost always work with the extended set $\mathbb{R}_* = \mathbb{R} \cup \{-\infty\}$, so that we can write $\log 0 = -\infty$. We assume that $-\infty + x = x + (-\infty) = -\infty$ and $x \geq -\infty$ for all $x \in \mathbb{R}_*$.

Hidden Markov Models (HMMs)

We formally introduce the concept of time-homogeneous Hidden Markov Models.

Definition 1. *A time-homogeneous HMM is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \Pi, \mathcal{T}, \mathcal{E})$, composed of:*

- *a set $\mathcal{S} = \{s_1, \ldots, s_n\}$ of n hidden states; n is called the* size *of the model,*
- *an output alphabet $\mathcal{A} = \{a_1, \ldots, a_{|\mathcal{A}|}\}$,*
- *a probability distribution vector $\Pi = \{\pi_1, \ldots, \pi_n\}$ over the initial states,*
- *a matrix $\mathcal{T} = \{t_s(s')\}_{s,s' \in \mathcal{S}}$ of transition probabilities between states,*
- *a matrix $\mathcal{E} = \{e_s(a)\}_{s \in \mathcal{S}}^{a \in \mathcal{A}}$ of emission probabilities.*

Matrices $\mathcal{T}$ and $\mathcal{E}$ are stochastic, i.e., the entries of every row sum up to 1.

For notational convenience, we relabel the states of a HMM with natural numbers, i.e. we let $\mathcal{S} = \{1, \ldots, n\}$.

As stated in the introduction, HMMs define generative processes over the alphabet $\mathcal{A}$. The initial state $x_0 \in \mathcal{S}$ is chosen according to the distribution Π; then, at each step, a symbol y is generated according to the probability distribution $e_x(y)$, where x is the current state; a new state x' is chosen according to the probability distribution induced by $t_x(x')$, and the process repeats. The probability of a state path $X = (x_1, \ldots, x_m)$ joint to an observation sequence $Y = (y_1, \ldots, y_m)$ is computed as:

$$\Pr(X, Y) = \pi_{x_1} \left(\prod_{i=1}^{m-1} t_{x_i}(x_{i+1}) \right) \left(\prod_{i=1}^{m} e_{x_i}(y_i) \right).$$

The Viterbi algorithm

The Viterbi algorithm consists of two phases: in the first phase, a simple dynamic programming approach is used to determine the probability of the most probable state path ending in each possible state. In the second phase, the data stored in the dynamic programming table is used to reconstruct a most probable state path.

Definition 2. *Assume given a HMM $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \Pi, \mathcal{T}, \mathcal{E})$ and an observed sequence $A = (a_1, \ldots, a_m)$. For every $s \in \mathcal{S}$ and $i = 1, \ldots, m$, denote by $q_i(s)$ the probability of any most probable path ending in state s explaining the observation $A_{i-1} = (a_1, \ldots, a_{i-1})$.*

By definition of $q_i(s)$, any most probable path explaining A has probability $\max_{s \in \mathcal{S}} \left\{ e_{s(a_m)} \cdot q_m(s) \right\}$. The $q_i(s)$ values can be computed inductively. Indeed, $q_1(s) = \pi_s$ for all $s \in \mathcal{S}$, while for every $i > 1$ and $s \in \mathcal{S}$ it holds:

$$q_i(s) = \max_{s' \in \mathcal{S}} \{ q_{i-1}(s') \cdot t_{s'}(s) \cdot e_{s'}(a_{i-1}) \} . \qquad (1)$$

In order to compute all the n values $q_i(s)$, for any fixed $i > 1$, $\Theta(n^2)$ comparisons have to be performed. This phase is in fact the bottleneck of the algorithm.

The second phase of the algorithm uses the $q_i(s)$ values to reconstruct an optimal path in $\mathcal{O}(mn)$ time. We will not deal with this second and faster part, and only mention that most of the previously developed solutions for it, including the memory saving ones (Šrámek 2007; Churbanov and Winters-Hilt 2008), are still applicable once the first part has been carried out based on our approach.

Online matrix-vector multiplication problem

We formally introduce the real online matrix-vector $(\max, +)$-multiplication problem briefly discussed in the introduction.

Definition 3 (Matrix-vector $(\max, +)$-multiplication)**.** *Given a $m \times n$ matrix $\mathbf{A}$ and a n-dimensional column vector $\mathbf{b}$ over $\mathbb{R}_*$, their $(\max, +)$-multiplication $\mathbf{A} * \mathbf{b}$ is a column vector of dimension m whose i-th component is:*

$$(\mathbf{A} * \mathbf{b})[i] = \max_{j=1}^{n} \{ \mathbf{A}[i, j] + \mathbf{b}[j] \} .$$

Problem 1 (ORMV $(\max, +)$-MUL problem)**.** *Given a $m \times n$ matrix $\mathbf{A}$ over $\mathbb{R}_*$, perform a polynomial-time preprocess of it, so that the product $\mathbf{A} * \mathbf{b}$ can be computed for any input vector $\mathbf{b}$ of size n.*

Lemma 1 is a simple result that bridges the gap between the MAPD problem and the ORMV $(\max, +)$-MUL problem, showing that any fast solution to Problem 1 can be turned into a fast solution for the MAPD problem.

Lemma 1. *Any algorithm for Problem 1 computing $\mathbf{A} * \mathbf{b}$ in $T(m, n)$ time after $U(m, n)$ preprocessing time can be used to solve the MAPD problem for any time-homogeneous HMM of size n and observation sequence of length m in $\mathcal{O}(m \cdot T(n, n))$ time after a $U(n, n)$ time preprocessing.*

Proof. Denote by ${}^t\mathcal{T}$ the transpose of the state transition matrix $\mathcal{T}$ of $\mathcal{M}$. Furthermore, for every $i = 1, \ldots, m$, introduce the pair of vectors

$$\mathbf{q}_i = (q_i(1), \ldots, q_i(n)), \quad \mathbf{e}_i = (e_1(a_i), \ldots, e_n(a_i)).$$

The value $q_i(s)$ corresponding to instant $i > 0$ and state $s \in \mathcal{S}$ can be computed as follows, in logaritmic scale:

$$\begin{aligned} \log q_i(s) &= \log \max_{s' \in \mathcal{S}} \{ q_{i-1}(s') \cdot t_{s'}(s) \cdot e_{s'}(a_{i-1}) \} \\ &= \max_{s' \in \mathcal{S}} \{ \log(t_{s'}(s)) + \log(q_{i-1}(s') \, e_{s'}(a_{i-1})) \} \\ &= ((\log {}^t\mathcal{T}) * (\log \mathbf{q}_{i-1} + \log \mathbf{e}_{i-1}))[s]. \end{aligned}$$

Notice that the $n \times n$ matrix $\log {}^t\mathcal{T}$ depends only on the model and is time-invariant; therefore, we can compute $q_i(s)$ for all $s = 1, \ldots, n$ in batch with an instance of ORMV $(\max, +)$-MUL:

$$\log \mathbf{q}_i = (\log {}^t\mathcal{T}) * (\log \mathbf{q}_{i-1} + \log \mathbf{e}_{i-1}).$$

Once the $q_i(s)$ values have been computed, we can use the second part of the Viterbi algorithm out of the box. The time required for the multiplication – the bottleneck of the algorithm – is $\mathcal{O}(T(n, n))$ by hypothesis, hence the final algorithm has time complexity $\mathcal{O}(m \cdot T(n, n))$. The time complexity of the preprocessing is $U(n, n)$. □

From multiplication to geometric dominance

Consider the following geometric problem, which apparently has no relation with the $(\max, +)$-multiplication, nor with the Viterbi algorithm.

Problem 2 (Online geometric dominance reporting)**.** *Let $\mathcal{B}$ be a set of d-dimensional vectors[2]. Given a vector $\mathbf{p} \in \mathcal{B}$, we define its* domination set *as*

$$\delta_{\mathcal{B}}(\mathbf{p}) = \{ \mathbf{b} \in \mathcal{B} : \mathbf{b} \preceq \mathbf{p} \}.$$

Preprocess $\mathcal{B}$ so that at a later time the set $\delta_{\mathcal{B}}(\mathbf{p})$ can be computed for any input vector $\mathbf{p}$.

Lemma 2. *Any algorithm solving Problem 2 in $\mathcal{O}(T(d, |\mathcal{B}|) + |\delta_{\mathcal{B}}(\mathbf{p})|)$ time after a preprocessing time $\mathcal{O}(U(d, |\mathcal{B}|))$ can be turned into an algorithm solving the ORMV $(\max, +)$-MUL problem for any $m \times t$ matrix in $\mathcal{O}(m + t \cdot T(t, m))$ time, with preprocessing time $\mathcal{O}(mt^2 + t \cdot U(t, m))$.*

Proof. This constructive proof comes in different stages. First, the result is obtained under the simplifying assumptions that (i) neither $\mathbf{A}$ nor $\mathbf{b}$ have any $-\infty$ entry, and (ii) the maximum sum on any row, $\mathbf{A}[\cdot, j] + \mathbf{b}[j]$, is achieved by exactly one value of the column index j. Later, these conditions will be dropped.

Observe that $(\mathbf{A} * \mathbf{b})[i] = \mathbf{A}[i, j^*] + \mathbf{b}[j^*]$ iff

$$\mathbf{A}[i, j^*] + \mathbf{b}[j^*] \geq \mathbf{A}[i, j] + \mathbf{b}[j] \qquad (2)$$

for every column index j. Under assumption (i), this inequality can be rewritten as

$$\mathbf{A}[i, j] - \mathbf{A}[i, j^*] \leq \mathbf{b}[j^*] - \mathbf{b}[j].$$

Defining the values

$$a_{i,j^*}(j) = \mathbf{A}[i, j] - \mathbf{A}[i, j^*], \quad b_{j^*}(j) = \mathbf{b}[j^*] - \mathbf{b}[j]$$

[2]The coordinates of the vectors can range over any chosen totally-ordered set, as long as any two coordinate values can be compared in constant time.

for all the feasible values of i, j, j^*, we obtain

$$(\mathbf{A} * \mathbf{b})[i] = \mathbf{A}[i, j^*] + \mathbf{b}[j^*] \iff a_{i,j^*}(j) \leq b_{j^*}(j) \quad \forall j. \tag{3}$$

Notice that the last expression is actually a statement of geometric dominance, between the two t-dimensional vectors $\tilde{\mathbf{A}}_{i,j^*} = (a_{i,j^*}(1), \ldots, a_{i,j^*}(t))$ and $\tilde{\mathbf{b}}_{j^*} = (b_{j^*}(1), \ldots, b_{j^*}(t))$. This immediately leads to the algorithm streamlined in Algorithm 1.

Algorithm 1 $m \times t$ matrix-vector $(\max, +)$-multiplication

```
1: procedure PREPROCESS(A)
2:     for j* = 1, ..., t do
3:         for i = 1, ..., m do
4:             Ã_{i,j*} ← (a_{i,j*}(1), ..., a_{i,j*}(t))
5:         B_{j*} ← {Ã_{1,j*}, ..., Ã_{m,j*}}
6:         Preprocess B_{j*}

1: procedure MULTIPLY(b)                          ▷ Returns A * b
2:     for j* = 1, ..., t do
3:         b̃_j ← (b_{j*}(1), ..., b_{j*}(t))
4:         δ ← δ_{B_{j*}}(b̃_{j*}), as defined in Problem 2
5:         for all Ã_{i,j*} ∈ δ do
6:             m[i] ← A[i, j*] + b[j*]
7:     return m
```

Line 6 in procedure PREPROCESS refers to the preprocessing routine for the ORMV $(\max, +)$-MUL problem given by assumption. It requires $U(t, m)$ time, making the total preprocessing cost of MULTIPLY $\mathcal{O}(mt^2 + t \cdot U(t, m))$. As for the running time, Line 4 of MULTIPLY takes $\mathcal{O}(T(t, m) + |\delta|)$ time, and is executed t times. Under assumption (ii), there is only one column j^* that satisfies Equation 3 for each row i, hence the total number of elements appearing in δ is exactly m. As a consequence, Line 6 of MULTIPLY is executed m times and the total time complexity of MULTIPLY is $\mathcal{O}(m + t \cdot T(m, t))$.

We now relax assumptions (i) and (ii) by applying a transformation of the input. Instead of working on $\mathbb{R}_*$, we work on triples over $\mathbb{N} \times \mathbb{R} \times \mathbb{N}$. The matrix $\mathbf{A}$ is transformed as follows: each element $x > -\infty$ is replaced by $\langle 0, x, 0 \rangle$, while each occurence of $-\infty$ is replaced by $\langle -1, 0, 0 \rangle$. The input vector $\mathbf{b}$ is transformed similary, but the third coordinate is used to hold the index of the replaced element. Namely, element $\mathbf{b}[j] = x$ is replaced by $\langle 0, x, j \rangle$ if $x > -\infty$ and by $\langle -1, 0, j \rangle$ otherwise. One can informally regard the first two entries of each triple $\langle a, b, \cdot \rangle$ as a shorthand for the value $a \cdot \infty + b$. Any two triples are compared according to lexicographical order, while addition and subtraction are performed element-wise.

The crucial observation is that, if $\mathbf{A}[i, j^*] + \mathbf{b}[j^*] > \mathbf{A}[i, j] + \mathbf{b}[j]$ before the transformation, then the same holds also after the transformation. Hence, we can solve the transformed problem to obtain the solution of the original problem. Our algorithm can be applied as it is to the transformed problem: indeed, the inequality in Equation 2 can be rearranged without any further assumption; moreover, there can be no two distinct columns j and j' achieving the maximum, as the two triples $\mathbf{A}[i, j] + \mathbf{b}[j]$ and $\mathbf{A}[i, j'] + \mathbf{b}[j']$ differ at least on the third element. Once the output vector is obtained, replace each triple $\langle k, x, j \rangle$ with x if $k = 0$ and with $-\infty$ otherwise. Conveniently, the third element j holds the index of the column achieving the maximum. □

Lemma 3 shows that every fast algorithm for the OMV $(\max, +)$-MUL of narrow rectangular matrices can be turned into a OMV $(\max, +)$-MUL algorithm for square matrices.

Lemma 3. *Any algorithm computing the* OMV $(\max, +)$-MUL *of a $m \times t$ matrix in $T(m, t)$ time and $U(m, t)$ preprocessing time, can be used to multiply any $m \times n$ matrix, $n \geq t$, in $\mathcal{O}(n/t \cdot (T(m, t) + m))$ time and $\mathcal{O}(n/t \cdot U(m, t))$ preprocessing time.*

Proof. Assume without loss of generality that n is an integer multiple of t (otherwise, add columns to $\mathbf{A}$ and elements to $\mathbf{b}$ with value $-\infty$ until the condition is met). The idea is to split $\mathbf{A}$ and $\mathbf{b}$ into n/t blocks, each of size $m \times t$ and $t \times 1$ respectively:

$$\mathbf{A} = (\mathbf{A}_1 | \cdots | \mathbf{A}_{n/t}), \quad \mathbf{b} = (\mathbf{b}_1 | \cdots | \mathbf{b}_{n/t}).$$

Observe that $(\mathbf{A} * \mathbf{b})[i] = \max_{\ell=1}^{n/t} \{(\mathbf{A}_\ell * \mathbf{b}_\ell)[i]\}$. This immediately leads to the following algorithm. First, we preprocess each block $\mathbf{A}_\ell$ in $U(m, t)$ time with the given algorithm, so that the product $\mathbf{A}_\ell * \mathbf{b}_\ell$ can be later computed in $T(m, t)$ time. As soon as the vector $\mathbf{b}$ is received, compute $\mathbf{m}_\ell = \mathbf{A}_\ell * \mathbf{b}_\ell$ for all $\ell = 1, \ldots, n/t$, and finally the output vector by $(\mathbf{A} * \mathbf{b})[i] = \max_{\ell=1}^{n/t} \{\mathbf{m}_\ell[i]\}$.

The time analysis is straightforward. The computation of each $\mathbf{m}_\ell$ takes $T(m, t)$ time. There are n/t such computations and merging the results takes $\mathcal{O}(m \cdot n/t)$ time, yielding a total time of $\mathcal{O}(n/t \cdot (T(m, t) + m))$. The total preprocessing time is $\mathcal{O}(n/t \cdot U(m, t))$. □

An $\mathcal{O}(n^2/\log n)$ algorithm (GDFV)

Lemma 4. *There exists an algorithm solving Problem 2 in $\mathcal{O}(d\, c_\varepsilon^d\, |\mathcal{B}|^\varepsilon + |\delta_{\mathcal{B}}(\mathbf{p})|)$ time for every $\varepsilon \in (0, 1]$, where $c_\varepsilon := 1/(2^\varepsilon - 1)$. The preprocessing requires $\mathcal{O}({c'_\varepsilon}^d\, |\mathcal{B}|^{1+\varepsilon})$ time and memory for every $\varepsilon \in (0, \log_2 3/2]$, where $c'_\varepsilon := 1/(2^{1+\varepsilon} - 2)$.*

Proof. We assume without loss of generality that $|\mathcal{B}|$ is a power of two, and show a simple divide-and-conquer algorithm.

Overview. If $d = 0$, return $\delta_{\mathcal{B}}(\mathbf{p}) = \mathcal{B}$. If $\mathcal{B}$ contains only one vector $\mathbf{b}$, check if $\mathbf{b} \preceq \mathbf{p}$ and return either $\{\mathbf{b}\}$ or the empty set accordingly. In all the other cases, split $\mathcal{B}$ into two sets $\mathcal{B}^-$ and $\mathcal{B}^+$ of size $|\mathcal{B}|/2$, according to the median d-th coordinate γ of the vectors in $\mathcal{B}$, so that $\mathbf{b}^-[d] \leq \gamma \leq \mathbf{b}^+[d]$ for all $\mathbf{b}^- \in \mathcal{B}^-$ and $\mathbf{b}^+ \in \mathcal{B}^+$. Now consider the d-th coordinate of the query vector $\mathbf{p}$: if it is strictly less than γ, then all the vectors in $\mathcal{B}^+$ do not occur in the solution. Hence, solve the problem recursively on $\mathcal{B}^-$. Otherwise, both the sets $\mathcal{B}^+$ and $\mathcal{B}^-$ need to be considered; however, the d-th coordinate for the vectors in $\mathcal{B}^-$ is known to be $\leq \mathbf{p}[d]$ and can be dropped. Hence, solve the problem recursively on

$\mathcal{B}^+$ and $\mathbf{p}$ and on $(\mathcal{B}^-)'$ and $\mathbf{p}'$, where the apostrophe denotes the discard of the last coordinate, and merge the solutions. The recursive step is summarized by the following recurrence:

$$\delta_{\mathcal{B}}(\mathbf{p}) = \begin{cases} \delta_{\mathcal{B}^-}(\mathbf{p}) & \text{if } \mathbf{p}[d] < \gamma, \\ \delta_{\mathcal{B}^+}(\mathbf{p}) \cup \delta_{(\mathcal{B}^-)'}(\mathbf{p}') & \text{otherwise.} \end{cases}$$

In order to make the algorithm faster, we exploit the fact that $\mathcal{B}$ is known in advance. At preprocessing time, we build a tree that guides the execution of the algorithm, where each node u corresponds to a subproblem $\mathcal{B}_u$ over a d_u-dimensional space. The root corresponds to the original set $\mathcal{B}$. If $|\mathcal{B}_u| \geq 2$ and $d_u \geq 1$, then the node u stores the median value γ and has three children, corresponding to the subproblems $\mathcal{B}_u^+$, $\mathcal{B}_u^-$ and $(\mathcal{B}_u^-)'$. Otherwise, u is a leaf storing the content of $\mathcal{B}_u$. We analyze the cost of building the tree later on. For now, notice that the size of the tree is at most polynomial in $|\mathcal{B}|$: the height is at most $\log |\mathcal{B}|$, as the value $|\mathcal{B}_u|$ halves at each level, so the nodes are at most $\mathcal{O}(3^{\log_2 |\mathcal{B}|}) = \mathcal{O}(|\mathcal{B}|^{\log_2 3}) = \mathcal{O}(|\mathcal{B}|^{1.59})$.

Time analysis. Our algorithm starts from the root node, and visits recursively the nodes in the tree that are needed to solve the problem. When we reach a leaf u with $d_u = 0$, we output $\mathcal{B}_u$ (which is not empty) in $\mathcal{O}(|\mathcal{B}_u|)$ time. If instead $d_u > 0$ and $|\mathcal{B}_u| = 1$, we pay $\mathcal{O}(d)$ time to check if $\mathbf{b} \preceq \mathbf{p}$, and $\mathcal{O}(1)$ to output $\mathbf{b}$ if needed. On internal nodes, we only pay constant extra time as the median coordinate γ is known from the tree. The cost of producing the output is $\mathcal{O}(|\delta_{\mathcal{B}}(\mathbf{p})|)$, and is measured separately. Hence, the running time is $\mathcal{O}(T_d(|\mathcal{B}|)+|\delta_{\mathcal{B}}(\mathbf{p})|)$ where $T_d(n)$ satisfies the linear recurrence relation

$$\begin{aligned} T_d(n) &= 1 + \max \begin{cases} T_d(n/2) \\ T_{d-1}(n/2) + T_d(n/2) \end{cases} \\ &= T_{d-1}(n/2) + T_d(n/2) + 1, \end{aligned} \tag{4}$$

with base cases $T_d(1) = d$ and $T_0(n) = 0$. (The time required to handle the latter case is included in $\mathcal{O}(|\delta_{\mathcal{B}}(\mathbf{p})|)$). We show by induction that

$$T_d(n) \leq \overline{T}_d(n) \coloneqq d\, c_\varepsilon^d\, n^\varepsilon,$$

for any chosen $\varepsilon \in (0, 1]$, where $c_\varepsilon \coloneqq 1/(2^\varepsilon - 1)$. Notice that $c_\varepsilon \geq 1$ for $\varepsilon \in (0, 1]$, thus the statement is true for the base cases as $\overline{T}_d(n) \geq d$. Assuming the inductive hypothesis, we obtain for $n \geq 2$ and $d \geq 1$:

$$\begin{aligned} T_d(n) &\leq \overline{T}_d(n/2) + \overline{T}_{d-1}(n/2) + 1 \\ &= d\, c_\varepsilon^d\, (n/2)^\varepsilon + (d-1)\, c_\varepsilon^{d-1}\, (n/2)^\varepsilon + 1 \\ &\leq d\, c_\varepsilon^d\, (n/2)^\varepsilon + d\, c_\varepsilon^{d-1}\, (n/2)^\varepsilon \\ &= d\, c_\varepsilon^d\, n^\varepsilon\, \frac{1 + c_\varepsilon^{-1}}{2^\varepsilon} = \overline{T}_d(n)\, \frac{1 + 2^\varepsilon - 1}{2^\varepsilon} = \overline{T}_d(n), \end{aligned}$$

completing the induction. Thus, the time complexity of the algorithm is $\mathcal{O}(d\, c_\varepsilon^d\, |\mathcal{B}|^\varepsilon + |\delta_{\mathcal{B}}(\mathbf{p})|)$.

Preprocessing. The tree is built starting from the root. Finding the median d-th coordinate γ, computing $\mathcal{B}_u^+$, $\mathcal{B}_u^-$ and $(\mathcal{B}_u^-)'$, and storing the data in the node u, all require $\mathcal{O}(|\mathcal{B}_u|)$ time and memory. Hence, the time and memory cost to build the tree is $\mathcal{O}(U_d(|\mathcal{B}|))$, where $U_d(n)$ satisfies the recurrence

$$U_d(n) = 2U_d(n/2) + U_{d-1}(n/2) + n$$

with base cases $U_0(n) = n$ and $U_d(1) = 1$. We show by induction that

$$U_d(n) \leq \overline{U}_d(n) \coloneqq 3{c'_\varepsilon}^d\, n^{1+\varepsilon} - 2n$$

for any chosen $\varepsilon \in (0, \log_2 3/2]$, where $c'_\varepsilon \coloneqq 1/(2^{1+\varepsilon} - 2)$. Notice that $c'_\varepsilon \geq 1$ for $\varepsilon \in (0, \log_2 3/2]$. Hence, the statement is true for the base cases, as $\overline{U}_d(n) \geq 3{c'_\varepsilon}^d\, n^{1+\varepsilon} - 2n \geq 3n - 2n \geq n$. Assuming the inductive hypothesis, we obtain for $n \geq 2$ and $d \geq 1$:

$$\begin{aligned} U_d(n) &\leq 2\,\overline{U}_d(n/2) + \overline{U}_{d-1}(n/2) + n \\ &= 2 \cdot (3\, {c'_\varepsilon}^d\, (n/2)^{1+\varepsilon} - n) \\ &\quad + (3\, {c'_\varepsilon}^{d-1}\, (n/2)^{1+\varepsilon} - n) + n \\ &= 2 \cdot 3\, {c'_\varepsilon}^d\, (n/2)^{1+\varepsilon} + 3\, {c'_\varepsilon}^{d-1}\, (n/2)^{1+\varepsilon} - 2n \\ &= 3\, {c'_\varepsilon}^d\, n^{1+\varepsilon} \cdot \frac{2 + {c'_\varepsilon}^{-1}}{2^{1+\varepsilon}} - 2n \\ &= 3\, {c'_\varepsilon}^d\, n^{1+\varepsilon} - 2n = \overline{U}_d(n). \end{aligned}$$

completing the induction. Hence, the time and memory cost of the preprocessing phase is $\mathcal{O}(U_d(|\mathcal{B}|)) = \mathcal{O}({c'_\varepsilon}^d\, n^\varepsilon)$. □

We remark that the time complexity and the recurrence of the simple algorithm given in the above proof differ from those of (Chan 2008) by necessity, as the online setting requires each vector to be treated separately. On the other hand, his result follows directly from ours:

Theorem 1 ((Chan 2008), Lemma 2.1)**.** *Given n red/blue points in $\mathbb{R}_*^d$ we can report all K dominating pairs in $\mathcal{O}(k_\varepsilon^d\, n^{1+\varepsilon} + K)$ time for any $\varepsilon \in (0, 1)$, where $k_\varepsilon \coloneqq 2^\varepsilon/(2^\varepsilon - 1)$.*

Proof. After reading the set $\mathcal{B}$ of blue points, preprocess them as described in the proof of Lemma 4. Then, for each red point $\mathbf{p}$, perform a query to find all the dominators of $\mathbf{p}$, i.e. $\delta_{\mathcal{B}}(\mathbf{p})$; simply flush out the union of all the dominating pairs obtained. By Lemma 4, the cost of the preprocessing is $\mathcal{O}((2^{1+\varepsilon} - 2)^{-d}\, n^{1+\varepsilon}) = \mathcal{O}(k_\varepsilon^d\, n^{1+\varepsilon})$. On the other hand, each of the n queries takes time $\mathcal{O}(d(2^\varepsilon - 1)^{-d}\, n^\varepsilon)$, that is[3] $\mathcal{O}(k_\varepsilon^d\, n^\varepsilon)$, excluding the output; the overhead due to the actual output of the pairs is $\mathcal{O}(K)$. The final cost of the algorithm is therefore $\mathcal{O}(c_\varepsilon^d\, n^{1+\varepsilon} + K)$ as desired. □

Finally, Lemmas 1, 4, and 3 combine into the following.

Theorem 2. *There exists an algorithm solving the* MAPD *problem in $\mathcal{O}(mn^2/\log n)$ time after a polynomial preprocessing of the model, for any HMM of size n and observation sequence of length m.*

[3] Indeed, using the binomial expansion formula we have:

$$k_\varepsilon^d = \left(1 + \frac{1}{2^\varepsilon - 1}\right)^d \geq d \left(\frac{1}{2^\varepsilon - 1}\right)^{d-1} = \Omega(d\, c_\varepsilon^d).$$

Proof. It is enough to show how to solve the ORMV $(\max,+)$-MUL problem for any $n \times n$ matrix in $\mathcal{O}(n^2/\log n)$ time. To this end, apply Lemma 3 with $m = n$ and $t = \alpha \log_2 n$, where $\alpha \in (0, 1/2) \subseteq \mathbb{R}$. This gives a running time of $\mathcal{O}(n \cdot T(\alpha \log_2 n, n) + n^2/\log n)$, where $T(d, n)$ is the cost of computing the domination set of a d-dimensional vector over a fixed set of n points (see Problem 2). Substituting the bounds of Lemma 4 into the time complexity, yields a polynomial preprocessing cost, and the following time bound for each multiplication:

$$\mathcal{O}((2^\varepsilon - 1)^{-\alpha \log_2 n} \alpha n^{1+\varepsilon} \log^2 n + n^2/\log n) = \\ = \mathcal{O}(n^{1+\varepsilon-\alpha \log_2(2^\varepsilon-1)} \log^2 n + n^2/\log n)$$

for all $\varepsilon \in (0, 1)$. Setting $\varepsilon = 2\alpha$, the exponent of the first term becomes $1 + \alpha(2 - \log_2(4^\alpha - 1)) < 2$ for all $0 < \alpha < 1/2$. Therefore, the time complexity of the algorithm is $\mathcal{O}(n^2/\log n)$. □

We call the resulting algorithm *geometric dominance faster Viterbi* (GDFV).

Experimental evaluation

Methodology. All the algorithms are implemented in the C++11 language, compiled using the `clang` compiler and run on the OSX 10.10.3 operating system. The main memory is a 8GB 1600MHz DDR3 RAM, and the processor is an Intel Core i7-4850HQ CPU, with 6MB shared L3 cache. All the matrices and vectors used for the experiments have entries sampled from a uniform distribution over $(0, 1] \subseteq \mathbb{R}$.

Results. *How does our proposed* ORMV $(\max,+)$-MUL *algorithm for narrow matrices compare to the trivial one?* We analyze the throughput of Algorithm 1, based on the geometric subroutines exposed in the proof of Lemma 4, comparing it with the trivial multiplication approach. For every chosen pair (n, t), we run 25 tests, each of which consists of an online multiplication of a $n \times t$ matrix with 10 000 vectors. The results of our tests are summarized in Figure 1, where we see that our algorithm can be up to 4 times faster than the trivial one. This is mainly due to the fact that the number of accesses to the tree is much less than $n \cdot t$, and to the lower number of comparisons needed to find the answer.

How does the complete GDFV *algorithm compare with the Viterbi algorithm?* We experimentally evaluate the first phase of the GDFV algorithm, i.e. the computation of the $q_i(s)$ values defined in Equation 1. This is the most expensive task in the decoding of HMMs. We implement the algorithm as described in the proof of Theorem 2, using $\alpha = 0.25$, that is splitting the $n \times n$ transition probability matrix $\mathcal{T}$ of the model in approximately $n/2$ blocks when $n \leq 4000$. We summarize the results in Figure 2, where we see that our algorithm is roughly twice as fast as the Viterbi algorithm, in line with expectations. However, we note that the amount of memory required by our algorithm makes it impractical for larger values of α. Indeed, we have verified that when the memory pressure becomes high other factors slow down the implemented algorithm, such as cache and page misses, or, for bigger allocations, the hard drive latency.

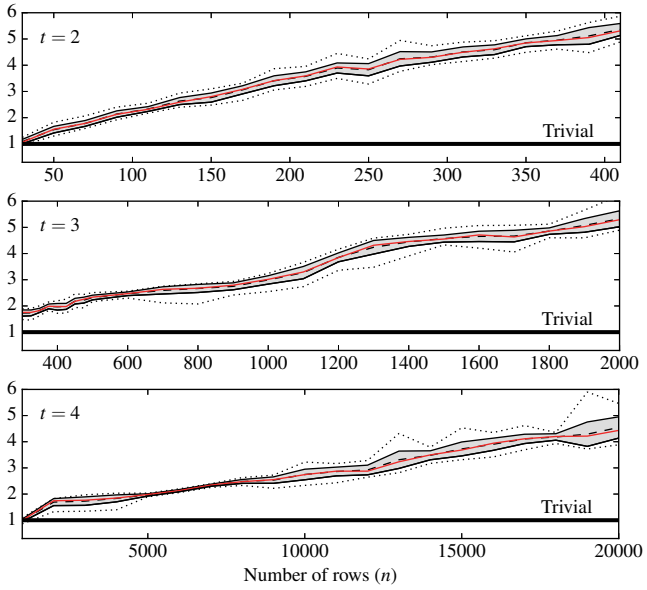

Figure 1: Relative throughput of Algorithm 1 when compared to the trivial quadratic algorithm on matrices of size $n \times t$ for $t = 2, 3, 4$. A higher throughput implies faster computation. Legend: **dotted black** lines denote min and max values, **red solid** lines denote median values, **black dashed** lines denote mean values (μ), and **gray shading** denotes the range $\mu \pm \sigma$, where σ is the standard deviation.

Conclusion and future works

In this paper, we gave the first algorithm for the maximum *a posteriori* decoding (MAPD) of time-homogeneous Hidden Markov Models requiring asymptotically less than $\mathcal{O}(mn^2)$ operations in the worst case. To this end, we first introduced an *online* geometric dominance reporting problem, and proposed a simple divide-and-conquer solution generalizing the result by (Chan 2008). At an intermediate step, we also gave the first algorithm solving the *online* matrix-vector $(\max,+)$-multiplication problem over $\mathbb{R}_*$ in subquadratic time after a polynomial preprocessing of the matrix. Finally, we applied the faster multiplication to obtain an algorithm for the MAPD problem. By extending to the online setting the results about multi-dimensional geometric dominance reporting, we effectively bridged the gap between matrix-matrix $(\max,+)$-multiplication and the online matrix-vector counterpart.

We think that our proposal leads to several questions which we intend to explore in future works:

- cut larger polylogarithmic factors, by splitting cases in Equation 4 in a different manner, as in (Chan 2015);
- study and implement a more succinct version of the decision tree, in order to mitigate the memory footprint;
- analyze the relationship of our work with other existing heuristics, such as (Lifshits et al. 2009) and (Esposito and Radicioni 2009);
- explore the possibility of implementing our algorithm at a hardware level, resulting in specialized chips performing

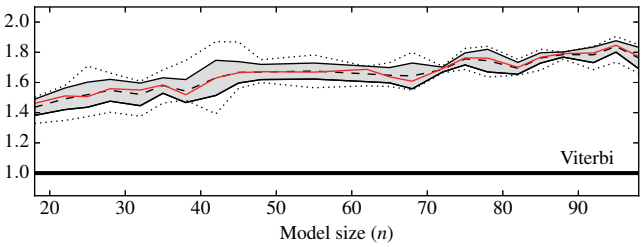

Figure 2: Relative throughput of the GDFV algorithm when compared to the Viterbi algorithm. A higher throughput implies faster computation. Legend: as in Figure 1.

asymptotically less that $\mathcal{O}(n^2)$ operations per observed symbol, in the worst case.

Acknowledgements

Massimo Cairo was supported by the Department of Computer Science, University of Verona under PhD grant "Computational Mathematics and Biology".

We would like to thank Marco Elver, Nicola Gatti, Zu Kim, and Luigi Laura for their valuable suggestions.

References

Agazzi, O., and Kuo, S. 1993. Hidden markov model based optical character recognition in the presence of deterministic transformations. *Pattern recognition* 26(12):1813–1826.

Bremner, D.; Chan, T. M.; Demaine, E. D.; Erickson, J.; Hurtado, F.; Iacono, J.; Langerman, S.; and Taslakian, P. 2006. Necklaces, convolutions, and $x + y$. In *Algorithms–ESA 2006*. Springer. 160–171.

Chan, T. M. 2008. All-pairs shortest paths with real weights in $o(n^3/\log n)$ time. *Algorithmica* 50(2):236–243.

Chan, T. M. 2015. Speeding up the four russians algorithm by about one more logarithmic factor. In *SODA*, 212–217. SIAM.

Churbanov, A., and Winters-Hilt, S. 2008. Implementing em and viterbi algorithms for hidden markov model in linear memory. *BMC bioinformatics* 9(1):224.

Dobosiewicz, W. 1990. A more efficient algorithm for the min-plus multiplication. *INT J COMPUT MATH* 32(1-2):49–60.

Esposito, R., and Radicioni, D. P. 2009. Carpediem: Optimizing the viterbi algorithm and applications to supervised sequential learning. *J MACH LEARN RES* 10:1851–1880.

Felzenszwalb, P. F.; Huttenlocher, D. P.; and Kleinberg, J. M. 2004. Fast algorithms for large-state-space hmms with applications to web usage analysis. *Advances in NIPS* 16:409–416.

Gales, M. J. 1998. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer speech & language* 12(2):75–98.

Grice, J.; Hughey, R.; and Speck, D. 1997. Reduced space sequence alignment. *Computer applications in the biosciences: CABIOS* 13(1):45–53.

Haussler, D. K. D., and Eeckman, M. G. R. F. H. 1996. A generalized hidden markov model for the recognition of human genes in dna. In *Proc. Int. Conf. on Intelligent Systems for Molecular Biology, St. Louis*, 134–142.

Henzinger, M.; Krinninger, S.; Nanongkai, D.; and Saranurak, T. 2015. Unifying and strengthening hardness for dynamic problems via the online matrix-vector multiplication conjecture. In *STOC*, 21–30. New York, NY, USA: ACM.

Huang, X. D.; Ariki, Y.; and Jack, M. A. 1990. *Hidden Markov models for speech recognition*, volume 2004. Edinburgh university press Edinburgh.

Kaji, N.; Fujiwara, Y.; Yoshinaga, N.; and Kitsuregawa, M. 2010. Efficient staggered decoding for sequence labeling. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 485–494. Association for Computational Linguistics.

Kupiec, J. 1992. Robust part-of-speech tagging using a hidden markov model. *Computer Speech & Language* 6(3):225–242.

Lafferty, J. D.; McCallum, A.; and Pereira, F. C. N. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 282–289.

Liberty, E., and Zucker, S. W. 2009. The mailman algorithm: A note on matrix–vector multiplication. *Information Processing Letters* 109(3):179–182.

Lifshits, Y.; Mozes, S.; Weimann, O.; and Ziv-Ukelson, M. 2009. Speeding up hmm decoding and training by exploiting sequence repetitions. *Algorithmica* 54(3):379–399.

Mäkinen, V.; Belazzougui, D.; Cunial, F.; and Tomescu, A. I. 2015. *Genome-Scale Algorithm Design*. Cambridge University Press.

Šrámek, R. 2007. The on-line viterbi algorithm. *KAI FMFI UK, Bratislava, máj*.

Starner, T.; Weaver, J.; and Pentland, A. 1998. Real-time american sign language recognition using desk and wearable computer based video. *IEEE T PATTERN ANAL* 20(12):1371–1375.

Viterbi, A. J. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE T INFORM THEORY* 13(2):260–269.

Williams, R. 2007. Matrix-vector multiplication in sub-quadratic time:(some preprocessing required). In *SODA*, volume 7, 995–1001.

Yoon, B.-J. 2009. Hidden markov models and their applications in biological sequence analysis. *Current genomics* 10(6):402.