A New Algebra for the Treatment of Markov Models

featuring

A Novel Inverse

T D Barfoot and G M T D'Eleuterio

University of Toronto Institute for Aerospace Studies
4925 Dufferin Street, Toronto, Ontario, Canada, M3H 5T6
<tim.barfoot@utoronto.ca, gabriele.deleuterio@utoronto.ca>

Keywords

Stochastic Matrix, Stochastic Algebra, Hidden Markov Model

Abstract

In this paper we question the appropriateness of using conventional matrix algebra in the analyses of such systems as Markov chains and Hidden Markov Models. *Stochastic matrices*, whose entries are probabilities and whose columns sum to unity, are central to many of these systems yet the set of such matrices under the usual matrix addition and scalar multiplication does not constitute a vector space. To solve this dilemma, we describe a new algebra which does allow the addition and scalar multiplication of stochastic matrices. We show several examples of how the new operators required to construct this *stochastic algebra* are useful in the analyses of Markov models. In particular, we cast the development of the classic Baum-Welch algorithm in this new algebra. We also propose a gradient-ascent algorithm that is compared to Baum-Welch on a simple example inspired by Markov's original paper.

1 Introduction

Markov models are simple examples of stochastic processes that have connections to statistical physics. Since their inception circa 1913 by Andrei Andreevich Markov [10], the most basic *chains* have been expanded in several ways to include Hidden Markov Models (HMMs), Markov Decision Processes, Partially Observable Markov Decision Processes, Decentralized Partially Observable Markov Decision Processes, Products of Hidden Markov Models, and Markov Networks. This partial list shows the fervour with which Markov's original ideas have proliferated.

The primary objective of this paper is to suggest a rigourous way of treating Markov models. It is our conjecture that conventional matrix algebra with which are all familiar is perhaps not the ideal tool for working with matrices whose entries are probabilities. If we represent the probability density of a random variable over m discrete states[1] as a column vector then the axiom of total probability dictates that the sum of the entries down this column must be unity. More generally, matrices whose entries are probabilities (i.e., real numbers between 0 and 1) and whose columns sum to unity have been referred to as *stochastic matrices* in the context of Markov chains for quite some time [5]. If $\boldsymbol{x}[t] = \left[p(X_i[t])\right]$ is a stochastic column and $\boldsymbol{A} = \left[p(X_i[t+1]|X_j[t])\right]$ is an appropriately sized square stochastic matrix, a Markov chain is represented by the simple difference equation

$$\boldsymbol{x}[t+1] = \boldsymbol{A}\boldsymbol{x}[t]$$

The matrix, $\boldsymbol{A}$, is often called the *transition matrix*. The constraint on the columns of stochastic matrices is also known as the *simplex constraint*. Unfortunately, there is a problem when applying conventional matrix algebraic operators, such as addition and scalar multiplication, to stochastic matrices.

The root of the problem mathematically is that although the stochastic matrices are a well defined *subset* of the real matrices, they do not form a *subspace* of the real matrices. This is immediately clear when one tries to

[1] In this paper we discuss only discrete random variables.

add two stochastic matrices or multiply by a scalar using conventional operations. Alternatively, consider the zero vector from the vector space formed by the real matrices. It is a matrix with all entries equal to zero. The zero vector is not a stochastic matrix but its inclusion is a requirement of the establishment of a subspace.

Although we have made the situation to sound quite dire, we are prepared to offer a resolution. In what follows, we will show that the set of stochastic matrices can still be thought of as a vector space but not under the conventional matrix operations. We must redefine addition and scalar multiplication to be more suited to stochastic matrices [2, 4]. As a point of comparision, the zero vector in our new vector space is the *uniform probability density.* Our approach is really no more than a straightforward application of the general theory of vector spaces [12]. Once this is done we no longer have to worry about the constraints associated with probability densities, they are handled implicitly. In particular, we do not require awkward projections or the more elegant Lagrange multipliers often called upon to re-establish said constraint. It furthermore becomes possible to establish an inner product space, an associative algebra, and a vector calculus including a gradient operator for our new vector space. To distinguish our new methodology from conventional matrix algebra, it will be referred to as *stochastic algebra*, as it is based on stochastic matrices.

Once this stochastic algebra has been established it is relatively easy to handle the equations of Markov models. We might stress from the onset, however, that the equation of even a basic Markov chain is considered to be *nonlinear.* Thus we have traded an equation that appears linear (although it is not) in conventional algebra for a nonlinear one in stochastic algebra. In this paper we will examine both Markov chains and Hidden Markov Models to show the applicability of this new approach. These Markov models have been used to model population dynamics, human speech, and noisy robot sensors, for example. When fitting an HMM to data, the classic Baum-Welch algorithm is a popular choice. We show how it may be developed in our new algebra. In the course of this development, the gradient of a function of a stochastic matrix is computed. As such, a simple gradient-ascent approach to fitting an HMM to data follows quite naturally. We compare this new algorithm to Baum-Welch on a simple example inspired by Markov's original paper, in which he fits a Markov chain to data reflecting the alternation of vowels and consonants in Aleksandr Pushkin's classic poem, *Eugene Onegin.*

We begin with a presentation of the new stochastic algebra and stochastic calculus, followed by analyses of Markov chains and Hidden Markov Models, and conclude with the example.

2 The Algebra

Stochastic matrices are often used to represent probability densities and conditional probability tables (CPTs) when random variables are discrete. The earliest reference we have found to the use of the term *stochastic matrix* is Dmitriev [5] but they have also been called Markov matrices [10]. The set of *stochastic matrices* ${}^m\mathbb{S}^n$ is

$$ {}^m\mathbb{S}^n = \left\{ \boldsymbol{A} = [a_{ij}] \in {}^m\mathbb{R}^n \;\middle|\; \sum_{i=1}^{m} a_{ij} = 1, \;\; a_{ij} > 0 \right\} $$

Each column of a stochastic matrix may be thought of as probability density over m discrete "states". In the limiting case that only one state is occupied with probability 1, the density is called *deterministic* and must be treated carefully. In the event that no one state is occupied with probability 1 but at least one state is occupied with probability 0, the density is called *partially stochastic* and again must be treated with care. When all states are equally probable we have a uniform probability density. Thus we introduce the *uniform matrix*, $\boldsymbol{\Omega} \in {}^m\mathbb{S}^n$, which is

$$ \boldsymbol{\Omega} = [u_{ij}], \quad u_{ij} = \frac{1}{m} $$

This will be referred to simply as $\boldsymbol{\omega}$ in the case of a single column. As discussed above, some new operators are now introduced for stochastic matrices. These definitions are critical in establishing a vector space. The *normalization* operator denoted $\downharpoonleft\!\boldsymbol{R}$, where $\boldsymbol{R} = [r_{ij}] \in {}^m\mathbb{R}^n$ with $r_{ij} > 0$ is

$$ \downharpoonleft\!\boldsymbol{R} = \left[\frac{r_{ij}}{\sum_{k=1}^{m} r_{kj}} \right] $$

This operation renders any positive real matrix a stochastic matrix ($\downharpoonleft\!\boldsymbol{R} \in {}^m\mathbb{S}^n$).

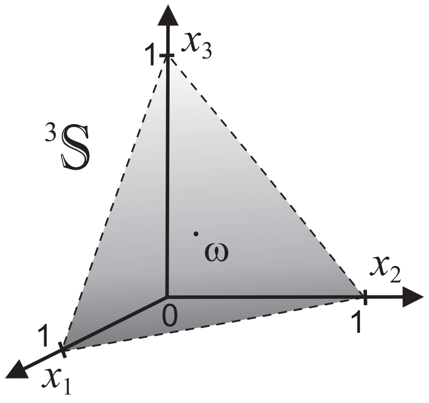

Figure 1: Graphical depiction of the vector space, $^3\mathbb{S}$, in relation to the usual Cartesian space, $^3\mathbb{R}$. The new vector space is the two-dimensional shaded triangular surface shown with the zero vector, $\boldsymbol{\omega}$, marked at the centroid of the triangle.

We redefine the addition operator for stochastic matrices as follows. Let $\boldsymbol{A} = [a_{ij}],\ \ \boldsymbol{B} = [b_{ij}] \in {}^m\mathbb{S}^n$. The *vector addition* of $\boldsymbol{A}$ and $\boldsymbol{B}$, denoted $\boldsymbol{A} \oplus \boldsymbol{B}$, is

$$\boldsymbol{A} \oplus \boldsymbol{B} = \downarrow[a_{ij}b_{ij}]$$

In the case that the operands are deterministic, vector addition must be computed in the limit. Also, when addition is negative, the symbol, $\ominus$, is used. In words, vector addition is accomplished by taking the direct product of the entries and then renormalizing each column. This vector addition has also been called *logarithmic opinion pooling* [6] and can be traced back to the *Nash product* [11]. It is also the operator used to combine experts in recent *products-of-experts* models [8].

Scalar multiplication must be redefined as well to be compatible with stochastic matrices. It is carried out by taking the exponent of each entry with the scalar and then renormalizing the columns. Let $\boldsymbol{A} = [a_{ij}] \in {}^m\mathbb{S}^n$ and $\lambda \in \mathbb{R}$. The scalar multiplication of λ with vector $\boldsymbol{A}$, denoted $\lambda{\cdot}\boldsymbol{A}$, is

$$\lambda{\cdot}\boldsymbol{A} = \downarrow[a_{ij}^{\lambda}]$$

In the case that $\boldsymbol{A}$ is deterministic, scalar multiplication must be computed in the limit. With these definitions in hand it is possible to prove that the set ${}^m\mathbb{S}^n$ is a vector space over the field $\mathbb{R}$ under the vector addition and scalar multiplication defined above. The uniform matrix, $\boldsymbol{\Omega}$, is the *zero vector* of ${}^m\mathbb{S}^n$.

We also define an *inner product* associated with this space. Let $\boldsymbol{x} = [x_i],\ \boldsymbol{y} = [y_i] \in {}^m\mathbb{S}$. Then

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \frac{1}{2m} \sum_{i=1}^{m} \sum_{j=1}^{m} \ln\left(\frac{x_i}{x_j}\right) \ln\left(\frac{y_i}{y_j}\right)$$

The inner product will be necessary in establishing the gradient of a scalar function with a stochastic matrix as a parameter. Note the general properties of the inner product, namely, $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle$, $\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0$, and $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0$ only when $\boldsymbol{x} = \boldsymbol{\omega}$ (i.e., the zero vector). The inner product is also linear such that

$$\langle \alpha{\cdot}\boldsymbol{u} \oplus \beta{\cdot}\boldsymbol{v}, \gamma{\cdot}\boldsymbol{x} \oplus \delta{\cdot}\boldsymbol{y} \rangle = \alpha\gamma\langle \boldsymbol{u}, \boldsymbol{x} \rangle + \alpha\delta\langle \boldsymbol{u}, \boldsymbol{y} \rangle + \beta\gamma\langle \boldsymbol{v}, \boldsymbol{x} \rangle + \beta\delta\langle \boldsymbol{v}, \boldsymbol{y} \rangle$$

We furthermore make the claim that we have an *associative algebra* for stochastic matrices, or a *stochastic algebra.* To justify this claim, we require a vector product (in addition to the already established vector space) but in the interests of brevity, we elect not to present it here. This algebra allows one to consider all the usual algebraic concepts including: the adjoint, an outer product, bases, subspaces, projections, determinant, rank, the eigenproblem, the Cayley-Hamilton theorem, and an isomorphism to the familiar matrix algebra [2, 4, 3]. Naturally, it provides *a novel inverse* for a stochastic matrix.

There is another operator that is not required in the establishment of the stochastic algebra but which is particularly useful for Markov models. The *stochastic transpose* operator, denoted $\boldsymbol{A}^{\dagger}$, where $\boldsymbol{A} = [a_{ij}]$, is

$$\boldsymbol{A}^{\dagger} = \downarrow[a_{ji}]$$

Note, in the case that $\sum_k a_{jk} = 0$ then the j^{th} column of $\boldsymbol{A}^\dagger$ is defined to be the uniform column (this is a limiting case). This operator takes the transpose of a stochastic matrix and then renormalizes the columns.

3 Vector Calculus

It is not surprising that we can associate with our stochastic algebra a corresponding stochastic calculus. As ${}^m\mathbb{S}^n$ is a vector space, all the typical results from vector calculus may be obtained including derivatives of stochastic matrices, partial derivatives, and Jacobians [2, 4]. In much of the work involving Markov models to follow, we will be trying to minimize a scalar error function of a stochastic parameter. To this end we must be able to compute the gradient of a scalar function with respect to a stochastic vector.

The *stochastic gradient* of a real scalar function, $F(\boldsymbol{x})$, with respect to stochastic vector, $\boldsymbol{x} \in {}^n\mathbb{S}$, is denoted $\boldsymbol{\nabla}_{\boldsymbol{x}} F \in {}^n\mathbb{S}$ and is the unique vector that satisfies

$$F(\boldsymbol{x} \oplus \Delta\boldsymbol{x}) - F(\boldsymbol{x}) = \langle \boldsymbol{\nabla}_{\boldsymbol{x}} F,\ \Delta\boldsymbol{x} \rangle + O(\Delta\boldsymbol{x}^2)$$

where $\Delta\boldsymbol{x} \in {}^n\mathbb{S}$. When $\Delta\boldsymbol{x}$ is sufficiently small we may neglect the $O(\Delta\boldsymbol{x}^2)$ terms. Upon doing so it is possible to establish that the gradient may be expressed as

$$\boldsymbol{\nabla}_{\boldsymbol{x}} F = \downarrow\!\left[\exp \frac{\eth F(\boldsymbol{x})}{\eth x_j}\right]$$

The stochastic partial derivative, denoted $\eth F(\boldsymbol{x})/\eth x_j$, is given by

$$\frac{\eth F(\boldsymbol{x})}{\eth x_j} \triangleq \lim_{\lambda \to 0} \frac{1}{\lambda}\big(F(\boldsymbol{x} \oplus \lambda \cdot \boldsymbol{\xi}_j) - F(\boldsymbol{x})\big) = x_j \left(\frac{\partial F}{\partial x_j} - \sum_{k=1}^{n} x_k \frac{\partial F}{\partial x_k} \right)$$

where $\partial F/\partial x_j$ is the usual partial derivative of a scalar function with respect to a scalar variable, $\boldsymbol{\xi}_j$ is the j^{th} column of $\boldsymbol{\Xi} = \downarrow[\exp \delta_{ij}]$, and δ_{ij} is the Kronecker delta. Note that $\sum_{j=1}^{n} \eth F/\eth x_j = 0$.

If we are looking for a density, $\boldsymbol{x}$, to minimize (or maximize) a scalar function, $F(\boldsymbol{x})$, the usual approach is to compute the gradient, set it equal to the zero vector, and solve for the critical points. In conventional algebra one must include a Lagrange multiplier to maintain the constraint that probabilities sum to unity. We should mention that the Lagrange multiplier approach assumes the critical points are on the interior of the probability simplex (as opposed to the boundary).

We will see that in stochastic algebra we do not need to include a Lagrange multiplier explicitly when solving for the critical points as the simplex constraint has been maintained implicitly. Setting the stochastic gradient to the stochastic zero vector we have

$$\boldsymbol{\nabla}_{\boldsymbol{x}} F = \boldsymbol{\omega}$$

which implies that $\eth F/\eth x_j = \phi$ where ϕ is an unknown constant. It turns out that $\phi = 0$ in general which may be found by summing over all j and using $\sum_{j=1}^{n} \eth F/\eth x_j = 0$. Thus, the critical points are found by solving

$$(\forall j)\ \ \frac{\eth F}{\eth x_j} = x_j \left(\frac{\partial F}{\partial x_j} - \sum_{k=1}^{n} x_k \frac{\partial F}{\partial x_k} \right) = 0$$

If we furthermore assume that the critical points are not on the boundary of the probability simplex[2] (i.e., $(\forall j)\ x_j > 0$), which is what the Lagrange multiplier approach assumes, we may divide out the factor x_j above. It is not hard to show that an equivalent set of equations is

$$(\forall j)\ \ \frac{\partial F}{\partial x_j} - \frac{1}{n} \sum_{k=1}^{n} \frac{\partial F}{\partial x_k} = 0$$

which are precisely the $n-1$ independent equations one ends up with by using conventional algebra with a Lagrange multiplier. Thus, our algebra (and calculus) effectively has the Lagrange multiplier built in and thus we need not worry about including it explicitly. These are all the basic results we need to begin our analyses of Markov models.

[2] This will always be so given our definition of the stochastic vector space as it does not allow for zero entries except in the limit.

4 Markov Chains

We begin with some thoughts on the most basic form of Markov chains. A Markov chain can be expressed by the following stochastic equation

$$\boldsymbol{x}[t+1] = \boldsymbol{A}\boldsymbol{x}[t] \qquad \boldsymbol{x}[0] = \boldsymbol{x}_0$$

where $\boldsymbol{x} = [x_i]$, $\boldsymbol{x}_0 = [x_{0,i}] \in {}^n\mathbb{S}$, and $\boldsymbol{A} = [a_{ij}] \in {}^n\mathbb{S}^n$. A Markov chain is characterized by $\Theta = \{\boldsymbol{A}, \boldsymbol{x}_0\}$. The state of such a system is random variable, X, which takes on values from an alphabet of size n at each time-step, drawn from density, $\boldsymbol{x} = [x_i] \in {}^n\mathbb{S}$. The ordered sequence for X up to time τ is thus

$$H = (H_0, H_1, H_2, \ldots, H_\tau)$$

where $H_t \in \{1, \ldots, n\}$ are independently drawn and τ is the size of the sequence.

It is useful to examine the process of fitting a Markov chain to some sequence, H, as described above. Namely, given H, what is the best model, $\Theta = \{\boldsymbol{A}, \boldsymbol{x}_0\}$? This is, which model will maximize the probability of producing the data? In turns out to be easier to construct our objective function, $F(\Theta)$, to be the logarithm[3] of the probability of the data

$$F(\Theta) \triangleq \ln p(H|\Theta) = \ln\left(\prod_{t=1}^{\tau} a(H_t, H_{t-1})\right) x_0(H_0) = \ln x_0(H_0) + \sum_{t=1}^{\tau} \ln a(H_t, H_{t-1})$$

where $a(H_t, H_{t-1})$ is the element from $\boldsymbol{A}$ for transitioning from the state at time $t-1$ to that at time t and $x_0(H_0)$ is the probability of starting in state, H_0. Note that we have introduced the notation $a(\alpha, \beta)$, in place of $a_{\alpha\beta}$ to avoid the use of subscripts on subscripts; these two notations will be used interchangeably throughout the rest of the paper.

We would like to find a $\Theta^* = \{\boldsymbol{A}^*, \boldsymbol{x}_0^*\}$ to maximize $F(\Theta)$. To do this we compute the gradients of $F(\Theta)$ with respect to each of $\boldsymbol{A}$ and $\boldsymbol{x}_0$ which are both stochastic matrices. Now

$$\frac{\partial F}{\partial a_{ij}} = \frac{1}{a_{ij}} \sum_{t=1}^{\tau} \delta(H_t, i)\delta(H_{t-1}, j)$$

where $\delta(H_t, i)$ is a Kronecker delta which is 1 when $H_t = i$ and 0 otherwise. We thus have for the partial derivative with respect to stochastic element, a_{ij}, that

$$\begin{aligned}
\frac{\eth F}{\eth a_{ij}} &= a_{ij}\left(\frac{\partial F}{\partial a_{ij}} - \sum_{k=1}^{n} a_{kj}\frac{\partial F}{\partial a_{kj}}\right) \\
&= \sum_{t=1}^{\tau} \left(\delta(H_t, i)\delta(H_{t-1}, j) - a_{ij}\delta(H_{t-1}, j)\right) \\
\boldsymbol{\nabla}_{\boldsymbol{A}} F &= \downarrow\left[\exp \frac{\eth F}{\eth a_{ij}}\right] \\
&= \downarrow\left[\exp \sum_{t=1}^{\tau} \left(\delta(H_t, i)\delta(H_{t-1}, j) - a_{ij}\delta(H_{t-1}, j)\right)\right]
\end{aligned}$$

where we have noticed $\sum_{k=1}^{n} \delta(H_t, k) = 1$. Setting the gradient[4] equal to the zero vector, $\boldsymbol{\nabla}_{\boldsymbol{A}} F = \boldsymbol{\Omega}$, it is straightforward to solve analytically for

$$\boldsymbol{A}^* = \downarrow\left[\sum_{t=1}^{\tau} \delta(H_t, i)\delta(H_{t-1}, j)\right]$$

[3] The logarithm is a monotonically increasing function so to maximize this is equivalent to maximizing its argument.

[4] Technically speaking, we only defined the stochastic partial derivative and stochastic gradient for stochastic columns, but we hope their extensions to stochastic matrices are clear from usage.

which is simply the normalized frequency count of transitioning from state j to i, as expected. Using the same process for $\boldsymbol{x}_0$ we arrive at

$$\boldsymbol{x}_0^* = \downarrow[\delta(H_0, i)]$$

which has a 1 in the row corresponding to H_0 and 0 in all other rows. With this basic result in hand, we now turn to a more complicated Markov model for which it is not possible to solve for the optimal model so easily.

5 Hidden Markov Models

A Hidden Markov Model (HMM) can be expressed by the following stochastic equations

$$\begin{aligned} \boldsymbol{x}[t+1] &= \boldsymbol{A}\boldsymbol{x}[t] \qquad\qquad \boldsymbol{x}[0] = \boldsymbol{x}_0 \\ \boldsymbol{y}[t] &= \boldsymbol{C}\boldsymbol{x}[t] \end{aligned}$$

where $\boldsymbol{x} = [x_i]$, $\boldsymbol{x}_0 = [x_{0,i}] \in {}^n\mathbb{S}$, $\boldsymbol{A} = [a_{ij}] \in {}^n\mathbb{S}^n$, $\boldsymbol{y} = [y_i] \in {}^m\mathbb{S}$, and $\boldsymbol{C} = [c_{ij}] \in {}^m\mathbb{S}^n$. an HMM is a Markov chain with an associated output equation producing symbols from some alphabet of size, m, and is characterized by $\Theta = \{\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{x}_0\}$. The random variable, X, is called the *hidden variable*, because its identity at each time-step is unknown. Let Y be the random variable representing the output, also called the *visible variable.* The probability density associated with Y is $\boldsymbol{y} = [P(Y_i)] = [y_i] \in {}^m\mathbb{S}$. Our visible sequence, V, drawn from $\boldsymbol{y}$ is,

$$V = (V_0, V_1, \ldots, V_\tau)$$

where $V_t \in \{1, \ldots, m\}$ are independently drawn and τ is the size of the sequence. We assume Y has taken on value V_t when X has taken on H_t. To be clear, the usual assumption with this model is an observer has access only to the visible sequence, V, and not the hidden sequence, H. With this in mind, we will examine two key problems associated with HMMs:

- *Compute the probability of some particular visible sequence, V, given the model, Θ.*
- *Fit a model, Θ, given a visible sequence, V.*

The latter problem may really be seen to be the inverse of the former. As we will see, our new framework may be used to develop both the classic Baum-Welch solution and a new gradient-ascent approach to this inverse problem. Hence, we claim to have *a novel inverse* algorithm.

5.1 Probability of Visible Sequence

Traditionally this problem is solved using either the forward or backward procedure which make use of the recursively defined sequences

$$\begin{array}{rclclcrcl} p(V_0, \ldots, V_t, H_t = i|\Theta) &=& \alpha_i[t+1] &=& c(V_{t+1}, i) \sum_{j=1}^n a_{ij}\alpha_j[t] & & \alpha_i[0] &=& c(V_0, i)\, x_{0,i} \\ p(V_{t+1}, \ldots, V_\tau|H_t = i, \Theta) &=& \beta_i[t] &=& \sum_{j=1}^n a_{ji}\beta_j[t+1]\, c(V_{t+1}, j) & & \beta_i[\tau] &=& 1 \end{array}$$

where the α_i are called the *forward variables* and the β_i are the *backwards variables.* The probability of the visible sequence, V, given the model, can be efficiently computed as

$$p(V|\Theta) = \sum_{i=1}^n \alpha_i[\tau] = \sum_{i=1}^n \beta_i[0]\, c(V_0, i)\, x_{0,i}$$

Incidently, we may use the forward variables to construct a recursive forward state estimator in matrix form as follows:

$$\big[p(H_t = i|V_0, \ldots, V_t, \Theta)\big] \quad=\quad \hat{\boldsymbol{x}}_{\text{for}}[t+1] \quad=\quad \boldsymbol{A}\hat{\boldsymbol{x}}_{\text{for}}[t] \oplus \boldsymbol{C}^\dagger \boldsymbol{y}[t+1] \qquad \hat{\boldsymbol{x}}_{\text{for}}[0] \quad=\quad \boldsymbol{x}_0$$

where $\boldsymbol{y}[t+1] = \big[\delta(V_{t+1}, i)\big] \in {}^m\mathbb{S}$. This is related to the forward variables through

$$\hat{\boldsymbol{x}}_{\text{for}}[t] = \downarrow\big[\alpha_i[t]\big]$$

This may be used to estimate the state of random variable, X, as visible symbols are generated online. It was originally proposed by [1] and now gets used frequently as a *belief estimator* when solving POMDPs [7], for example. In stochastic algebra, it can be seen to take the general form of nonlinear observer.

5.2 Fitting an HMM to Data

We now examine the problem of determining the best model, $\Theta^* = \{\boldsymbol{A}^*, \boldsymbol{C}^*, \boldsymbol{x}_0^*\}$, given a particular visible sequence, V. We assume the corresponding hidden sequence, H, is unknown. As in the case of the Markov chain we construct our objective function as the logarithm of the probability of the data, given the model,

$$F(\Theta) \triangleq \ln p(V|\Theta) = \ln \sum_{\nu} p(V, H^{\nu}|\Theta)$$

where ν is an index over all possible hidden sequences (there are n^{τ} of them). Also note that

$$p(V, H^{\nu}|\Theta) = \left(\prod_{t=1}^{\tau} c(V_t, H_t^{\nu}) a(H_t^{\nu}, H_{t-1}^{\nu}) \right) x_0(H_0^{\nu})$$

As before we must compute the gradient of F with respect to our stochastic parameters. For simplicity we will assume we are only going to fit the transition matrix, $\boldsymbol{A}$, for the remainder of this section. In general one would do the same thing for $\boldsymbol{C}$ and $\boldsymbol{x}_0$ as well. After a little manipulation we find for $\boldsymbol{A}$ that

$$\begin{aligned}
\frac{\partial F}{\partial a_{ij}} &= \sum_{\nu} p(H^{\nu}|V,\Theta) \frac{\partial}{\partial a_{ij}} \ln p(V, H^{\nu}|\Theta) \\
&= \sum_{\nu} p(H^{\nu}|V,\Theta) \sum_{t=1}^{\tau} \frac{\partial}{\partial a_{ij}} \ln a(H_t^{\nu}, H_{t-1}^{\nu}) \\
&= \sum_{\nu} p(H^{\nu}|V,\Theta) \sum_{t=1}^{\tau} a_{ij} \left(\frac{\partial}{\partial a_{ij}} \ln a(H_t^{\nu}, H_{t-1}^{\nu}) - \sum_{k=1}^{n} a_{kj} \frac{\partial}{\partial a_{kj}} \ln a(H_t^{\nu}, H_{t-1}^{\nu}) \right) \\
&= \sum_{\nu} p(H^{\nu}|V,\Theta) \sum_{t=1}^{\tau} \Big(\delta(H_t^{\nu}, i)\delta(H_{t-1}^{\nu}, j) - a_{ij}\delta(H_{t-1}^{\nu}, j) \Big) \\
&= \mu_j \left(\bar{a}_{ij} - a_{ij} \right)
\end{aligned}$$

where we have defined

$$\begin{aligned}
\mu_j &= \sum_{\nu} p(H^{\nu}|V,\Theta) \sum_{t=1}^{\tau} \delta(H_{t-1}^{\nu}, j) \\
\bar{a}_{ij} &= \frac{1}{\mu_j} \sum_{\nu} p(H^{\nu}|V,\Theta) \sum_{t=1}^{\tau} \delta(H_t^{\nu}, i)\delta(H_{t-1}^{\nu}, j)
\end{aligned}$$

It is important to note that $\bar{\boldsymbol{A}} = \left[\bar{a}_{ij}\right] \in {}^{n}\mathbb{S}^{n}$ is the *expected* normalized frequency count and still depends on $\boldsymbol{A}$. The gradient may now be written as

$$\begin{aligned}
\boldsymbol{\nabla}_{\boldsymbol{A}} F &= \downarrow\left[\exp \frac{\partial F}{\partial a_{ij}}\right] \\
&= \downarrow\left[\exp \mu_j \left(\bar{a}_{ij} - a_{ij} \right)\right]
\end{aligned}$$

It is not possible to equate the gradient to the zero vector, $\boldsymbol{\Omega}$, and solve for the best $\boldsymbol{A}^*$. We instead must turn our attention to approaches that iteratively improve the model. We will examine two such procedures, the classic Baum-Welch algorithm and a direct gradient-ascent algorithm. It is important to point out that the most efficient way to compute $\bar{\boldsymbol{A}}$ is using the forward/backward variables as follows

$$\bar{\boldsymbol{A}} = \downarrow\left[\sum_{t=1}^{\tau-1} \frac{c(V_{t+1}, i)\beta_i[t+1]a_{ij}\alpha_j[t]}{\sum_{k,l=1}^{n} c(V_{t+1}, k)\beta_k[t+1]a_{kl}\alpha_l[t]} \right]$$

which avoids the highly expensive proposition of actually considering all possible hidden sequences.

5.3 Baum-Welch Algorithm

The Baum-Welch algorithm falls into the category of *expectation-maximization* wherein one uses the current model to estimate the expected normalized frequency count, $\bar{\boldsymbol{A}}$. This is performed iteratively until the model has converged. Thus, if we let $\bar{\boldsymbol{A}}^{(s)} = \left[\bar{a}_{ij}^{(s)}\right]$ represent this quantity, evaluated using the model, $\boldsymbol{A}^{(s)} = \left[a_{ij}^{(s)}\right]$, from iteration, s, we may make the following approximation to the gradient

$$\boldsymbol{\nabla_A} F \approx \downarrow\left[\exp \mu_j \left(\bar{a}_{ij}^{(s)} - a_{ij}^{(s+1)}\right)\right]$$

so that upon equating to zero, $\boldsymbol{\Omega}$, we find $\boldsymbol{A}^{(s+1)} = \bar{\boldsymbol{A}}^{(s)}$. Thus, the optimal model at iteration $(s+1)$ is the expected normalized frequency count as computed using the model from iteration (s). This is the same as the result for the Markov chain but it is the *expected* normalized frequency count. The Baum-Welch iterative update for $\boldsymbol{A}$ is thus

$$\boldsymbol{A}^{(s+1)} \leftarrow \bar{\boldsymbol{A}}^{(s)}$$

We can show that this update is guaranteed increase the objective function. To first order, the change in the objective function according to our definition of the stochastic gradient[5] will be

$$\begin{aligned} F(\boldsymbol{A}^{(s+1)}) - F(\boldsymbol{A}^{(s)}) &= \langle \boldsymbol{\nabla_A} F\big|_{\boldsymbol{A}^{(s)}}, \boldsymbol{A}^{(s+1)} \ominus \boldsymbol{A}^{(s)} \rangle \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \mu_j \left(\bar{a}_{ij}^{(s)} - a_{ij}^{(s)}\right) \left(\ln \bar{a}_{ij}^{(s)} - \ln a_{ij}^{(s)}\right) \end{aligned}$$

which is positive definite[6]. It is in fact the symmetrical version of the Kullback-Leibler information theoretic measure [9] which is often used as the objective function from the beginning when solving this problem. We elected to show it comes out quite naturally when using stochastic algebra. One hopes that after a great many iterations, the process converges such that

$$\boldsymbol{A}^* = \lim_{s\to\infty} \boldsymbol{A}^{(s)}$$

the final solution is the optimal model. It should be pointed out, however, that the Baum-Welch algorithm finds a local maximum. The update equations are similar for $\boldsymbol{C}$ and $\boldsymbol{x}_0$.

5.4 Gradient-Ascent Algorithm

We now suggest using simple gradient-ascent in our stochastic algebra. Rather than selectively choosing to evaluate part of the gradient using the model from the previous iteration, we evaluate the entire gradient using the previous model. We then add this gradient to the stochastic parameter (using stochastic addition) to arrive at an update rule. For $\boldsymbol{A}$ the update is thus

$$\boldsymbol{A}^{(s+1)} \leftarrow \boldsymbol{A}^{(s)} \oplus \eta \cdot \boldsymbol{\nabla_A} F\big|_{\boldsymbol{A}^{(s)}}$$

where $\eta > 0 \in \mathbb{R}$ is a small constant. This algorithm is guaranteed to increase the objective function when η is small which may easily be seen using the definition of the gradient

$$\begin{aligned} F(\boldsymbol{A}^{(s+1)}) - F(\boldsymbol{A}^{(s)}) &= \langle \boldsymbol{\nabla_A} F\big|_{\boldsymbol{A}^{(s)}}, \eta \cdot \boldsymbol{\nabla_A} F\big|_{\boldsymbol{A}^{(s)}} \rangle \\ &= \eta \sum_{i=1}^{n} \sum_{j=1}^{n} \mu_j^2 \left(\bar{a}_{ij}^{(s)} - a_{ij}^{(s)}\right)^2 \end{aligned}$$

which is positive definite given $\eta > 0$. This algorithm uses all the same quantities as the Baum-Welch algorithm so it will be interesting to compare the two. One of the main differences is that we have some control over the convergence rate whereas classic Baum-Welch does not.

[5] In computing the change in objective function we use the gradient evaluated entirely at iteration (s), rather than the approximation employed in developing the Baum-Welch update.

[6] Technically speaking, we only defined the inner product for stochastic columns. To arrive at the expression for the change in objective function, we computed the sum over all j of the inner product between column j of the stochastic gradient of $\boldsymbol{A}$ and column j of the change, $\delta\boldsymbol{A} = \boldsymbol{A}^{(s+1)} \ominus \boldsymbol{A}^{(s)}$.

We should mention that our proposed algorithm is equivalent to reparametrizing a stochastic column, $\boldsymbol{x} \in {}^{n}\mathbb{S}$, using the following

$$\boldsymbol{x} = \downarrow\!\left[\exp y_i\right]$$

where the y_i are real numbers, not constrained in any way. We then perform exact gradient-ascent using the y_i parameters rather than the x_i (which are constrained). Computing the gradient using conventional matrix algebra with the chain rule reveals

$$\begin{aligned}\nabla_y F &= \left[\tfrac{\partial x_i}{\partial y_j}\right]^T \nabla_x F \\ &= \left[\textstyle\sum_{j=1}^{n} \tfrac{\partial F}{\partial x_j}\tfrac{\partial x_j}{\partial y_i}\right] \\ &= \left[x_i \left(\tfrac{\partial F}{\partial x_i} - \textstyle\sum_{j=1}^{n} x_j \tfrac{\partial F}{\partial x_j}\right)\right] \\ &= \left[\tfrac{\eth F}{\eth x_i}\right]\end{aligned}$$

whereupon the update rule for the y_i is

$$y_i \leftarrow y_i + \eta \frac{\eth F}{\eth x_i}$$

or in terms of $\boldsymbol{x}$ we have

$$\boldsymbol{x} \leftarrow \boldsymbol{x} \oplus \eta \cdot \left[\exp \tfrac{\eth F}{\eth x_i}\right]$$

which is precisely gradient-ascent ($\eta > 0$) in stochastic algebra.

6 A Simple Example

We now consider an example inspired by Markov's original analysis of the alternation of vowels and consonants in Aleksandr Pushkin's classic poem, *Eugene Onegin*[7]. We use only Chapter 1, Sonnet 6 which in the original Russian text is

ЛАТЫНЪ НЗ МОДЫ ВЫШЛА НЫНЕ:	*De gustibus non disputandum*
ТАК, ЕСЛИ ПРАВДУ ВАМ СКАЗАТЬ,	Has lost cachet, for Latin's dead;
ОН ЗНАЛ ДОВОЛЪНО ПО-ЛАТЫНИ,	Yet shown a Latin phrase at random,
ЧТОБ ЭПИГРАФЫ РАЗБИРАТЪ,	Eugene could tell you what it said;
ПОТОЛКОВАТЪ ОБ ЮВЕПАЛЕ,	He'd carve the meat from Juvenal's gristle,
В КОНЦЕ ПИСЪМА ПОСТАВИТЪ *VALE*,	Conclude with *Vale* an epistle,
ДА ПОМНИЛ, ХОТЬ НЕ БЕЗ ГРЕХА,	And knew by heart, though slightly skew,
ИЗ ЭНЕИДЫ ДВА СТИХА.	*Aeneid* verses – one or two.
ОН РЫТЪСЯ НЕ ИМЕЛ ОХОТЫ	He lacked the yen to go out poking
В ХРОНОЛОГИЧЕСКОЙ ПЫЛИ	Into the dusty lives of yore –
БЫТОПИСАНИЯ ЗЕМЛИ:	Historic details made him snore;
НО ДНЕЙ МИНУВШИХ АНЕКДОТЫ	But as for anecdotes and joking –
ОТ РОМУЛА ДО НАШИХ ДНЕЙ	Droll tales from Romulus till now –
ХРАНИЛ ОН В ПАМЯТИ СВОЕЙ.	He'd stocked a pile behind his brow.

where the English translation to the right may be found in Hofstadter [13]. The canonical Cyrillic alphabet consists of 33 symbols, 11 of which are vowels:

А, Е, Ё, И, Й, О, У, Ы, Э, Ю, Я

and the rest we consider to be consonants. Removing all punctuation, spaces, and the word 'vale' we create the sequence

$$\begin{aligned}H = (&1,2,1,2,1,1,1,1,1,2,1,2,1,2,1,1,2,1,2,1,2,1,2,1,2,1,1,2,1,1,2,1,1,2,1,2,1,1,1,2,1,2,\\ &1,1,2,1,1,1,2,1,1,2,1,2,1,1,1,2,1,2,1,2,1,2,1,2,1,1,2,1,2,1,2,1,1,2,1,2,1,2,1,1,2,1,2,\\ &1,1,1,2,1,2,1,1,2,1,2,1,1,2,1,2,1,2,1,2,1,2,1,1,2,1,1,2,1,2,1,1,1,2,1,2,1,1,2,1,2,1,1,\\ &1,2,1,2,1,1,2,1,1,2,1,1,1,2,1,2,1,1,1,2,1,2,2,1,2,1,2,2,1,2,1,1,2,1,1,2,1,2,2,1,1,2,1,\\ &1,1,2,1,2,2,1,2,1,2,1,2,1,2,1,1,1,2,1,2,1,2,1,2,1,2,1,1,2,2,1,2,1,2,1,2,1,2,1,2,1,2,1,\\ &2,2,1,2,1,1,2,1,2,1,1,2,2,1,2,1,2,1,1,2,1,2,1,2,1,1,2,1,2,2,1,1,2,1,2,1,2,1,2,1,2,1,2,\\ &1,1,1,2,2,1,1,2,1,2,1,2,1,1,1,2,1,2,1,2,1,1,2,2,2)\end{aligned}$$

[7] It is somewhat ironic that since Markov's original paper many English translations of Pushkin's poem have been attempted, differing greatly in their details. It would be amusing to consider Pushkin's poem as a Hidden Markov Model, itself, with the translations as visible sequences.

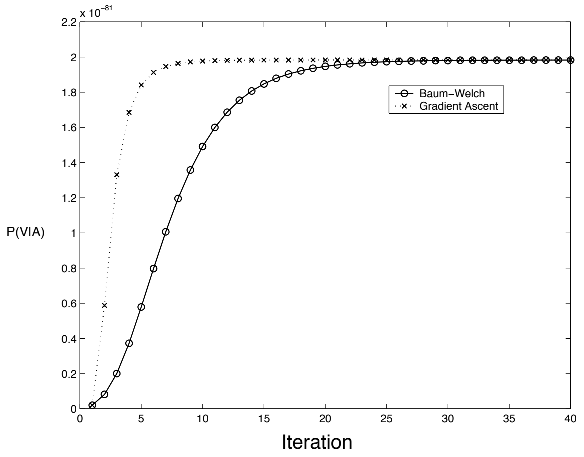

Figure 2: Convergence history for training a transition matrix representing the alternation of consonants and vowels in Pushkin's Eugene Onegin. Plot shows a typical convergence history for both the Baum-Welch algorithm and direct gradient-ascent with $\eta = 0.05$. In this limited example gradient-ascent is superior.

where state 1 indicates a consonant and state 2 indicates a vowel. Fitting a two-state Markov chain to this sequence reveals

$$\boldsymbol{A} = \downarrow\begin{bmatrix} 55 & 107 \\ 108 & 11 \end{bmatrix} \qquad \boldsymbol{x}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

We also consider an HMM wherein a visible sequence is produced by viewing the hidden sequence through a noisy channel[8]. We pick

$$\boldsymbol{C} = \begin{bmatrix} p & 1-p \\ 1-p & p \end{bmatrix}$$

with $p = 0.8$ and generate a visible sequence, V, from the above hidden sequence. We attempt to come up with a transition matrix, $\boldsymbol{A}$, to maximize the probability of this visible sequence, $p(V|\boldsymbol{A})$, while keeping $\boldsymbol{C}$ and $\boldsymbol{x}_0$ fixed.

We compared the Baum-Welch algorithm with basic gradient-ascent (with $\eta = 0.05$) as discussed above for 100 different visible sequences. Both algorithms started with $\boldsymbol{A} = \boldsymbol{\Omega}$ for all 100 test cases. Figure 2 shows the convergence of $p(V|\boldsymbol{A})$, our objective function, for a single test case. In every case both algorithms converged to identical transition matrices. We found gradient descent to convergence approximately twice as fast as Baum-Welch in 86 out of 100 cases.

In the other 14 cases both algorithms still converged to the same solution but Baum-Welch was faster (by at least an order of magnitude). In all of these 14 cases both algorithms converged to a matrix of the form

$$\boldsymbol{A} = \begin{bmatrix} q & 1 \\ 1-q & 0 \end{bmatrix}$$

That is, the second column was deterministic. The given visible sequence led to a transition matrix that transitions from a vowel to a consonant with probability 1 and transitions from a vowel to another vowel with probability 0. It is not difficult to understand why gradient-ascent is slower in this case. In stochastic algebra, deterministic columns may be viewed as being infinitely far away from the zero column, $\boldsymbol{\omega}$. Thus, when the gradient step is fixed in size it will take a very long time to get to "infinity". Baum-Welch is less susceptible to this problem as the step size it takes is effectively variable as can be seen in the change of the objective function. It will therefore get to "infinity" more quickly. In the 86 cases that gradient-ascent was faster than Baum-Welch, the ability to vary step size was far less important.

[8] This is fairly representative of the authors' lack of proficiency in Russian.

7 Conclusion

We have shown that a new algebra may be constructed wherein the set of stochastic matrices forms a vector space. Various new operators were introduced and used in the analyses of Markov chains and Hidden Markov Models. This rigourous framework allows an elegant treatment of both systems. Using the stochastic gradient operator, a straightforward gradient-ascent algorithm was tested against the classic Baum-Welch algorithm for Hidden Markov Models on a simple example. It was found that some tuning of the convergence rate was necessary, but that gradient-ascent could be made to converge faster than Baum-Welch except when the converged solution involved a deterministic column. In such an event Baum-Welch was much faster than gradient-ascent. The robustness of the Baum-Welch algorithm is superior but these results suggest further study of gradient-ascent for training HMMs is warranted.

Acknowledgements

The authors would like to thank the Natural Sciences and Engineering Research Council of Canada for support as well as Douglas R Hofstadter for allowing us to include his novel versification of Pushkin.

References

[1] K J Åstrom. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10:174–205, 1965.

[2] T D Barfoot. *Stochastic Decentralized Systems.* PhD thesis, Institute for Aerospace Studies, University of Toronto, 2002.

[3] T D Barfoot and G M T D'Eleuterio. An algebra for the analysis of stochastic systems. Submitted to *Linear Algebra and Its Applications*, 2002.

[4] T D Barfoot and G M T D'Eleuterio. An algebra for the control of stochastic systems: Exercises in linear algebra. Cambridge, UK, July 14-18 2002. The Fifth International Conference on Dynamics and Control of Structures in Space.

[5] N Dmitriev and E Dynkin. Eigenvalues of a stochastic matrix. *Izv. Akad. Nauk. SSSR Ser. Math.*, 10:167–184, 1946.

[6] Christian Genest and James V Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–148, 1986.

[7] Eric A Hansen. *Finite-Memory Control of Partially Observable Systems.* PhD thesis, Dept. of Computer Science, University of Massachusetts Amherst, 1998.

[8] Geoffrey E Hinton. Products of experts. In *Proceedings of the 9th International Conference on Artificial Neural Networks (ICANN 99)*, volume 1, pages 13–18, 1999.

[9] S Kullback. *Information Theory and Statistics.* Wiley, New York, 1959.

[10] Andrei Andreevich Markov. Essai d'une recherche statistique sur le texte du roman 'eugene onegin' illustrant la liaison des epreuve en chain. *Izvestia Imperatorskoi Akademii Nauk (Bulletin de l'Academie Imperiale des Sciences de St-Petersbourg)*, 7:153–162, 1913.

[11] John F Nash. The bargaining problem. *Econometrica*, 18:155–162, 1950.

[12] Giuseppe Peano. *Geometrical Calculus.* 1888.

[13] Alexander Sergeevich Pushkin. *Eugene Onegin, A Novel in Verse.* Basic Books, 1999. A Novel Versification by Douglas Hofstadter.