Starting to Uncover the Relationship Between Stochastic Factorization and Hidden Markov Models

André M. S. Barreto
National Laboratory for Scientific Computing
Petrópolis, RJ, Brazil
amsb@lncc.br

Borja B. Pigem
McGill University
Montreal, QC, Canada
bballe@cs.mcgill.ca

Joelle Pineau
McGill University
Montreal, QC, Canada
jpineau@cs.mcgill.ca

Doina Precup
McGill University
Montreal, QC, Canada
dprecup@cs.mcgill.ca

1 Introduction

When a transition probability matrix is represented as the product of two stochastic matrices, one can swap the factors of the multiplication to obtain another transition matrix. Interestingly, the derived matrix retains some fundamental characteristics of its precursor, which makes it possible to substitute the former for the latter in some applications [1]. Since the new matrix can be much smaller than the original, such a replacement can lead to significant savings in terms of computational effort. This new strategy, dubbed the "stochastic-factorization trick," has been used with different objectives in mind: to compute the stationary distribution of a Markov chain [1], to determine the fundamental matrix of an absorbing chain [1], and to compute a decision policy in dynamic programming [2] and reinforcement learning [3, 4]. At a more conceptual level, the stochastic-factorization trick can also serve as a formalism for thinking about state aggregation [5, 2].

In this short paper we start to investigate the relationship between stochastic factorization and hidden Markov models (HMMs) [6]. Besides being interesting on its own, this relationship may also have practical applications. Here we exploit such a connection to show how well-established techniques to compute an HMM from data can also be used to compute a stochastic factorization.

2 Background

This section briefly reviews some concepts and introduces the notation adopted in the paper.

2.1 Stochastic Factorization

We start with a few basic definitions:

Definition 1. *A matrix* $\mathbf{P} \in \mathbb{R}^{n \times z}$ *is called* stochastic *if and only if* $p_{ij} \geq 0$ *for all* i, j *and* $\sum_{j=1}^{z} p_{ij} = 1$ *for all* i*. A square stochastic matrix is called a* transition matrix.

Definition 2. *Given a stochastic matrix* $\mathbf{P} \in \mathbb{R}^{n \times z}$*, the relation* $\mathbf{P} = \mathbf{DK}$ *is called a* stochastic factorization *of* $\mathbf{P}$ *if* $\mathbf{D} \in \mathbb{R}^{n \times m}$ *and* $\mathbf{K} \in \mathbb{R}^{m \times z}$ *are also stochastic matrices. The integer* $m > 0$ *is the* order *of the factorization.*

Definition 3. *The* stochastic rank *of a stochastic matrix* $\mathbf{P} \in \mathbb{R}^{n \times z}$*, denoted by* $\mathrm{srk}(\mathbf{P})$*, is the smallest possible order of the stochastic factorization* $\mathbf{P} = \mathbf{DK}$.

The stochastic factorization has appeared before in the literature, either as defined above [7, 8] or in slightly modified versions [9, 10]. This paper will focus on a useful property of this type of factorization that has only recently been noted, which we now proceed to describe [1].

Given a stochastic factorization of a transition matrix, $\mathbf{P} = \mathbf{DK}$, *swapping* the factors of the factorization yields another transition matrix $\bar{\mathbf{P}} = \mathbf{KD}$, potentially much smaller than the original, which retains the basic topology of $\mathbf{P}$—that is, the number of recurrent classes and their respective reducibilities and periodicities (see [1] for details and formal definitions). Since $\bar{\mathbf{P}}$ can be much smaller than $\mathbf{P}$, the idea of replacing the latter with the former comes almost inevitably: this is the "*stochastic-factorization trick*."

A question that naturally arises in the context of the stochastic-factorization trick is how to compute the factorization in reasonable time. Since this question depends on the particular application, it has been addressed in different ways [2, 3, 4]. In the current paper we make a connection between stochastic factorization and HMMs that may be useful in this context, for two reasons. First, it delineates a class of transition matrices $\mathbf{P} \in \mathbb{R}^{n \times n}$ that are "factorizable"—that is, whose stochastic rank is $m < n$. Second, it shows how techniques to compute an HMM can also be used to compute $\mathbf{D}$ and $\mathbf{K}$ based on transitions sampled from $\mathbf{P}$.

2.2 Hidden Markov Models

We redirect the reader to one of the many introductions to HMMs available in the literature, such as Rabiner's tutorial, for a detailed description of these models [6]. Here we restrict ourselves to presenting the notation adopted.

We denote the sequence of hidden states by $H_t \in \{1, 2, ..., m\}$ and the sequence of observations by $S_t \in \{1, 2, ..., n\}$. Then:

- $\mathbf{T} \in \mathbb{R}^{m \times m}$ is the transition matrix, where $t_{ij} = \Pr(H_{t+1} = j | H_t = i)$;
- $\mathbf{K} \in \mathbb{R}^{m \times n}$ is the observation matrix, where $k_{ij} = \Pr(S_t = j | H_t = i)$;
- $\mu \in \mathbb{R}^{1 \times m}$ is the initial distribution, where $u_i = \Pr(H_1 = i)$.

Therefore, an HMM is defined as $H \equiv (\mathbf{T}, \mathbf{K}, \mu)$. We assume that $m < n$ throughout the paper.

3 Computing a Stochastic Factorization from an HMM

We start by presenting the following result:

Proposition 1. *Let $H \equiv (\hat{\mathbf{P}}, \mathbf{K}, \mu)$ and let $\mathbf{P}^{(t)}$ be the transition matrix in which $p_{ij}^{(t)} = \Pr(S_{t+1} = j | S_t = i)$. Then, $\mathrm{srk}(\mathbf{P}^{(t)}) \leq m$, where m is the number of hidden states in H.*

Proof. In order to show that $\mathrm{srk}(\mathbf{P}^{(t)}) \leq m$, it suffices to show that there exists an order m stochastic factorization of $\mathbf{P}^{(t)}$. Given an HMM H, it is always true that

$$\Pr(S_{t+1} = j | S_t = i) = \sum_w \Pr(S_{t+1} = j | S_t = i, H_{t+1} = w) \Pr(H_{t+1} = w | S_t = i).$$

Since in an HMM the probabilities of an observation are completely defined by the hidden state, $\Pr(S_{t+1} = j | S_t = i, H_{t+1} = w) = \Pr(S_{t+1} = j | H_{t+1} = w)$, and thus:

$$\Pr(S_{t+1} = j | S_t = i) = \sum_w \Pr(S_{t+1} = j | H_{t+1} = w) \Pr(H_{t+1} = w | S_t = i). \quad (1)$$

Therefore, if we define matrix $\mathbf{D}^{(t)} \in \mathbb{R}^{n \times m}$ as $d_{iw}^{(t)} = \Pr(H_{t+1} = w | S_t = i)$, we have $\mathbf{P}^{(t)} = \mathbf{D}^{(t)}\mathbf{K}$. □

Proposition 1 allows us to enunciate the following result:

Proposition 2. *Let $H \equiv (\hat{\mathbf{P}}, \mathbf{K}, \mu)$. If $\mu = \hat{\rho}$, where $\hat{\rho}$ is the stationary distribution of $\hat{\mathbf{P}}$, then $\Pr(S_{t+1} | S_t)$ is the same for all t.*

Proof. From Proposition 2 we know that $\mathbf{P}^{(t)} = \mathbf{D}^{(t)}\mathbf{K}$. Thus, if we show that $\mathbf{D}^{(t)} = \mathbf{D}$ for all t, the result follows.

Note that $d_{iw}^{(t)} = \Pr(H_{t+1} = w|S_t = i) = \sum_l \Pr(H_{t+1} = w|H_t = l, S_t = i)\Pr(H_t = l|S_t = i)$. Since

$$\Pr(H_t = l|S_t = i) = \frac{\Pr(S_t = i|H_t = l)\Pr(H_t = l)}{\sum_z \Pr(S_t = i|H_t = z)\Pr(H_t = z)},$$

we can write

$$d_{iw}^{(t)} = \frac{\sum_l \Pr(H_{t+1} = w|H_t = l)\Pr(S_t = i|H_t = l)\Pr(H_t = l)}{\sum_z \Pr(S_t = i|H_t = z)\Pr(H_t = z)}. \tag{2}$$

Now, since $\mu = \hat{\rho}$, $\Pr(H_t = z) = \hat{\rho}_z$ for all t, which implies that $\mathbf{D}^{(t)}$ does not depend on t. □

Propositions 1 and 2 are interesting for two reasons. First, they show how the structural assumptions necessary for the application of the stochastic-factorization trick naturally emerges from HMMs. Second, they suggest a simple algorithm to compute a stochastic factorization from data sampled from $\mathbf{P}$:

1. Use transitions sampled from $\mathbf{P}$ to compute $\tilde{\mathbf{T}}$ and $\tilde{\mathbf{K}}$—estimates of $\mathbf{T}$ and $\mathbf{K}$—using one of the many algorithms available to compute an HMM based on observations [11, 12, 13];
2. Compute $\tilde{\pi}$, the stationary distribution of $\tilde{\mathbf{T}}$;
3. Compute $\tilde{\mathbf{D}}$ as in (2);
4. Compute $\bar{\mathbf{P}} = \tilde{\mathbf{K}}\tilde{\mathbf{D}}$.

Note that in an actual application the approximation $\tilde{\mathbf{P}} = \tilde{\mathbf{D}}\tilde{\mathbf{K}}$ is never explicitly computed. Besides the obvious benefit of allowing one to replace $\mathbf{P}$ with $\bar{\mathbf{P}}$, the algorithm above may also be advantageous in terms of sample complexity, since we are estimating $O(m^2 + mn)$ parameters instead of $O(n^2)$.[1] We have run preliminary experiments that corroborate this hypothesis. See Figure 1 for an illustrative example.

4 Future Research

This is still an incipient research, and as such it leaves many open questions. Here we briefly discuss a few questions that we find particularly important.

At a conceptual level, it is interesting to ask whether there is a useful connection between $\mathbf{T}$ and $\bar{\mathbf{P}}$, since in general $\mathbf{T} \neq \bar{\mathbf{P}}$. Also, there should be a class of HMMs for which $\mathbf{T} = \bar{\mathbf{P}}$; one may ask whether it is possible to exploit the fact that $\mathbf{T} = \mathbf{KD}$ in this specific case.

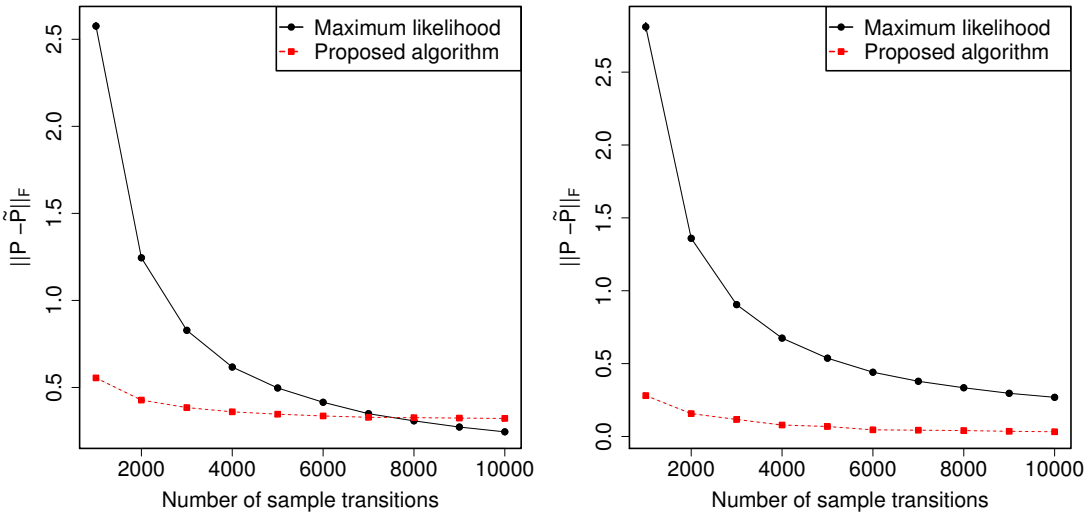
From a theoretical perspective, an important question is whether the converse of Proposition 2 is also true, that is, if the fact that $\mathrm{srk}(\mathbf{P}) \leq m$ implies that there exist an HMM $H = (\mathbf{T}, \mathbf{K}, \pi)$ that gives rise to $\mathbf{P}$, with $\mathbf{T} \in \mathbb{R}^{m\times m}$. If this is not the case, then we should ask why the proposed algorithm works when $\mathbf{P}$ is factorizable, as Figure 1b seems to indicate (note that in the referred experiment matrix $\mathbf{P}$ was generated from $\mathbf{D}$ and $\mathbf{K}$, without any explicit connection to an HMM).

At a more practical level, an obvious question is whether the proposed algorithm provides benefits from a computational point of view. Since the proposed method involves more operations than estimating $\mathbf{P}$ through directly maximizing the likelihood of $P(S_{t+1}|S_t)$ (which comes down to counting transitions), this extra cost must be compensated by the fact that $\bar{\mathbf{P}}$ will be manipulated in place of $\mathbf{P}$.

References

[1] André M. S. Barreto and Marcelo D. Fragoso. Computing the stationary distribution of a finite Markov chain through stochastic factorization. *SIAM Journal on Matrix Analysis and App.*, 32:1513–1523, 2011.

[1] As the recently introduced spectral methods to compute an HMM start from an estimate of $\mathbf{P}$, it seems unlikely that they provide benefits in terms of sample complexity [12, 13].



(a) Matrix $\mathbf{P}$ generated by sampling its elements from a uniform distribution and then normalizing them to enforce each row to have sum 1.

(b) $\mathbf{P} = \mathbf{DK}$, with $\mathbf{D}$ and $\mathbf{K}$ generated by sampling their elements from a uniform distribution and then normalizing them.

Figure 1: $\| \mathbf{P} - \tilde{\mathbf{P}} \|_{\mathrm{F}}$, where $\| \cdot \|_{\mathrm{F}}$ is the Frobenius norm and $\tilde{\mathbf{P}}$ is an estimate of $\mathbf{P}$, computed either by directly maximizing the likelihood of $P(S_{t+1}|S_t)$ (counting) or by the proposed method combined with the Baum-Welch algorithm [11]. Results generated with $n = 50$ and $m = 5$ using a single trajectory whose length is shown in the x axis. Shadows represent one standard error over 30 runs.

[2] André M. S. Barreto, Joelle Pineau, and Doina Precup. Policy iteration based on stochastic factorization. *Journal of Artificial Intelligence Research*, 50:763–803, 2014.

[3] André M. S. Barreto, Doina Precup, and Joelle Pineau. Reinforcement learning using kernel-based stochastic factorization. In *Adv. in Neural Information Processing Systems (NIPS)*, pages 720–728, 2011.

[4] André M. S. Barreto. Tree-based on-line reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2417–2423, 2014.

[5] André M. S. Barreto and Marcelo D. Fragoso. Lumping the states of a finite Markov chain through stochastic factorization. In *Proceedings of the World Congress of the International Federation of Automatic Control (IFAC)*, pages 4206–4211, 2011.

[6] Lawrence R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[7] Joel E. Cohen and Uriel G. Rothblum. Nonnegative ranks, decompositions and factorizations of nonnegative matrices. *Linear Algebra and its Applications*, 190:149–168, 1991.

[8] Ngoc-Diep Ho and Paul van Dooren. Non-negative matrix factorization with fixed row and column sums. *Linear Algebra and Its Applications*, 429(5–6):1020–1025, 2007.

[9] Adele Cutler and Leo Breiman. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994.

[10] Chris H. Q. Ding, Tao Li, and Michael I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(1):45–55, 2010.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38, 1977.

[12] Daniel Hsu, Sham M. Kakade, and Tong Zhang. A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

[13] Anima Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *CoRR*, abs/1210.7559, 2012.