Comparing the modeling powers of RNN and HMM

Achille Salaün*†, Yohan Petetin*, François Desbouvries*
*Samovar, CNRS, Telecom SudParis, Institut Polytechnique de Paris (Evry, France)
{yohan.petetin, francois.desbouvries}@telecom-sudparis.eu
†Nokia Bell Labs (Nozay, France)
achille.salaun@nokia.com

***Abstract*—Recurrent Neural Networks (RNN) and Hidden Markov Models (HMM) are popular models for processing sequential data and have found many applications such as speech recognition, time series prediction or machine translation. Although both models have been extended in several ways (eg. Long Short Term Memory and Gated Recurrent Unit architectures, Variational RNN, partially observed Markov models...), their theoretical understanding remains partially open. In this context, our approach consists in classifying both models from an information geometry point of view. More precisely, both models can be used for modeling the distribution of a sequence of random observations from a set of latent variables; however, in RNN, the latent variable is deterministically deduced from the current observation and the previous latent variable, while, in HMM, the set of (random) latent variables is a Markov chain. In this paper, we first embed these two generative models into a generative unified model (GUM). We next consider the subclass of GUM models which yield a stationary Gaussian observations probability distribution function (pdf). Such pdf are characterized by their covariance sequence; we show that the GUM model can produce any stationary Gaussian distribution with geometrical covariance structure. We finally discuss about the modeling power of the HMM and RNN submodels, via their associated observations pdf: some observations pdf can be modeled by a RNN, but not by an HMM, and vice versa; some can be produced by both structures, up to a re-parameterization.**

***Index Terms*—Recurrent neural networks, Hidden Markov Models**

I. INTRODUCTION

Let us consider the general time series prediction problem, which consists in predicting random future observations $\{X_{t+1}, \cdots, X_{t+j}\} = X_{t+1:j}$ from a realization of the past ones, $\{x_0, \cdots, x_t\} = x_{0:t} \in \mathbb{R}^{t+1}$.

This problem has many applications such as speech recognition [6] [12], time series prediction [8], or machine translation [3]. Various tools have been proposed to address this problem [14]. In particular, a solution to such a prediction problem is brought by generative models. Such models aim at modeling the observation sequence with a probability distribution $p_\theta(x_{0:t})$, where θ describes the set of parameters of the corresponding generative model. Once the model is known (i.e. θ has been estimated), the prediction of the future observations is computed from $p_\theta(x_{t+1:j}|x_{0:t})$.

In this paper, we focus on two popular generative models : Recurrent Neural Networks (RNN) on the one hand [10] [7], and Hidden Markov Models (HMM) [12] [5] on the other hand. Both models build a distribution $p_\theta(x_{0:T})$ via some latent and possibly random variables $H_{0:T}$:

- In the RNN, each latent variable is deduced deterministically from the past one and from the previous observation;
- By contrast, the distribution of the sequence of observations produced by an HMM is a marginal of a joint distribution which now involves random hidden variables. In the common vision, these hidden variables are discrete [12]; however, the general definition of such a model (see e.g. [5]) also encompasses the case where these variables are continuous. And indeed, such (continuous states) HMM have been widely used in many engineering applications such as econometric or tracking problems [5], [13]. Due to their proximity with RNN, we will only focus here on continuous states HMM.

So RNN and HMM share similarities since they both involve latent variables, but they differ from the way that those variables are built. Consequently, a natural question is to compare both models in terms of modeling power. More precisely, we wonder what kind of distributions $p_\theta(x_{0:T})$ can be modeled by each model, and what are the consequences induced by the different construction of the latent variables. As we shall see, there is no trivial inclusion of RNN into HMM or the converse but the modeling power of HMM is wider than that of RNN, even if there exists some distributions which can be only reached by the RNN structure.

II. GENERATIVE MODELS

In this section we first recall the principle of RNN and of HMM. We next show that both models can be seen as particular cases of a common probabilistic model called Generative Unified Model (GUM). Finally, we discuss the assumptions underlying our comparative study of these generative models.

A. RNN

RNN are particular neural networks which take into account the temporal structure of the data and are described by a set of parameters $\overline{\theta} = (\overline{\theta}_0, \overline{\theta}_1, \overline{\theta}_2)$. The distribution of the observations is obtained by managing hidden units $\overline{h}_t \in \mathbb{R}$ that are sequentially and deterministically computed through a given activation function $f(.)$ and additional parameters $\overline{\theta}_1 = (W_{hh}, W_{xh}, k)$:

$$\overline{h}_t = f_{\overline{\theta}_1}(\overline{h}_{t-1}, x_t) = f(W_{hh}\overline{h}_{t-1} + W_{xh}x_t + k). \quad (1)$$

Next, the distribution of the observations is directly deduced from the hidden units,

$$p_{\overline{\theta}}(x_{0:T}) = p_{\overline{\theta}}(x_0)\prod_{t=1}^{T} p_{\overline{\theta}}(x_t|x_{0:t-1}) = p_{\overline{\theta}_0}(x_0)\prod_{t=1}^{T} p_{\overline{\theta}_2}(x_t|\overline{h}_{t-1}), \tag{2}$$

where $p_{\overline{\theta}_0}(x_0)$ and $p_{\overline{\theta}_2}(x_t|\overline{h}_{t-1})$ are given parametrized distributions. Since by construction the likelihood $p_{\overline{\theta}}(x_{0:T})$ is computable, the parameters $\overline{\theta}$ which define (2) can be estimated by applying a gradient ascent method. In particular, the popular backpropagation algorithm [11] provides a solution to compute this gradient.

B. HMM

Continuous states HMM models are graphical statistical models which have found many applications in signal processing, in particular in contexts where the objective is to estimate a sequence of hidden states from a sequence of observations (e.g., estimating the kinematics of a target from noisy radar measurements). However, such models can also be used to model a sequence of observations. Here, the distribution $p_{\tilde{\theta}}(x_{0:T})$ is the marginal of the joint distribution of the latent and observed variables $(H_{0:T}, X_{0:T})$,

$$p_{\tilde{\theta}}(h_{0:T}, x_{0:T}) = p_{\tilde{\theta}_0}(h_0)\prod_{t=1}^{T} p_{\tilde{\theta}_1}(h_t|h_{t-1})\prod_{t=0}^{T} p_{\tilde{\theta}_2}(x_t|h_t). \tag{3}$$

This factorization describes the fact that $H_{0:T}$ is a Markov chain characterized by an initial distribution $p_{\tilde{\theta}_0}(h_0)$ and a transition distribution $p_{\tilde{\theta}_1}(h_t|h_{t-1})$ and that given the latent variables $h_{0:t}$, the observations $X_{0:T}$ are independent, and X_t only depends on the hidden state at the same instant t via the likelihood $p_{\tilde{\theta}_2}(x_t|h_t)$. Here, the computation of the predictive likelihood $p_{\tilde{\theta}}(x_{t+1}|x_{0:t})$ relies on the Bayes filter and its associated approximations [2].

C. GUM

As we have seen, both HMM and RNN rely on a sequence of hidden variables to model a distribution $p_\theta(x_{0:t})$. Moreover, by translating the temporal indexes of the hidden units of the RNN, i.e. by setting $h_t = \overline{h}_{t-1}$, we observe that both models share a different but close structure in terms of conditional dependencies. Indeed, in both cases, the pair $\{H_t, X_t\}_{t\geq 0}$ is Markovian. In the HMM case, H_t is a random variable and its distribution is deduced from $p_{\tilde{\theta}_1}(h_t|h_{t-1})$,

$$p_{\tilde{\theta}}(h_t, x_t|h_{t-1}, x_{t-1}) = p_{\tilde{\theta}_1}(h_t|h_{t-1})p_{\tilde{\theta}_2}(x_t|h_t); \tag{4}$$

while in the RNN, h_t is deterministic given h_{t-1} and x_{t-1},

$$p_{\overline{\theta}}(h_t, x_t|h_{t-1}, x_{t-1}) = \delta_{f_{\overline{\theta}_1}(h_{t-1},x_{t-1})}(h_t)\times p_{\overline{\theta}_2}(x_t|h_t), \tag{5}$$

where $\delta_a(x)$ stands for the Dirac delta function at point a. Finally, the observation x_t is generated from the hidden state h_t whatever the considered model. From (4) and (5), both models can be seen as particular cases of a GUM parametrized by θ, in which the pair $\{H_T, X_T\}$ is Markovian and the associated transition distribution reads

$$p_\theta(h_t, x_t|h_{t-1}, x_{t-1}) = p_{\theta_1}(h_t|h_{t-1}, x_{t-1})p_{\theta_2}(x_t|h_t). \tag{6}$$

The graphical structures of the three models are summarized in Fig. 1.

D. Scope of the study

As we have explained above, RNN and HMM can now be seen as two particular cases of the GUM and only differ from the distribution of the latent variable H_t, given h_{t-1} and x_{t-1}. We now exploit this unified framework to compare the modeling power of each model by characterizing and comparing the distributions $p_{\overline{\theta}}(x_{0:T})$ and $p_{\tilde{\theta}}(x_{0:T})$ in function of $\overline{\theta}$ and of $\tilde{\theta}$. In order to address such a theoretical comparison between RNN and HMM with feasible computations, we focus on the case where the objective is to model a sequence of observations $X_{0:T}$ in which each observation X_t follows a known Gaussian distribution $p(x_t)$ which does not depend on t. Thus, the models can now be directly compared via the joint distribution $p_\theta(x_{0:T})$ in the GUM. We assume that

$$\begin{aligned} p_{\theta_0}(h_0) &= \mathcal{N}(h_0; m_0; \eta), & (7)\\ p_{\theta_1}(h_t|h_{t-1}, x_{t-1}) &= \mathcal{N}(h_t; ah_{t-1} + cx_{t-1}; \alpha), & (8)\\ p_{\theta_2}(x_t|h_t) &= \mathcal{N}(x_t; bh_t; \beta). & (9) \end{aligned}$$

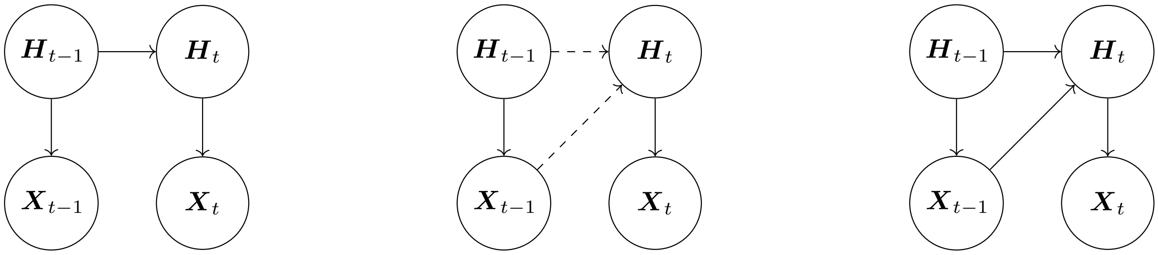
Note that the linear characteristic of the model (equations (8) and (9)) ensures that the distributions $p(x_t)$ of each observation X_t is Gaussian (and indeed that $p(x_{0:t})$ is Gaussian too). Let us now comment on these assumptions. Setting $c = 0$ we get linear and Gaussian HMM, which actually are ubiquitous in many applications such as navigation and tracking (see e.g. [9]); however setting $\alpha = 0$ we get RNN with linear activation functions, whereas activation functions are generally nonlinear and are a key of the modeling power of that model. Note however that we address a fair comparison in the sense that non linear activation function / transition distribution can be used in practice in both models.

Our comparison study is next led in three steps. First, in the linear and Gaussian GUM framework, we identify the class of models parametrized by $\theta = (a, b, c, \eta, \alpha, \beta)$ which satisfy $p(x_t) = \mathcal{N}(x_t; 0; 1)$ for all t; we thus obtain a class of joint Gaussian distributions $p_\theta(x_{0:t})$ which only differ by their associated covariance matrix. We next study the modeling power of this family of distributions and we show that the linear and Gaussian GUM can model any multivariate stationary Gaussian distribution in which, for all τ, the covariance function has the form $\text{cov}(X_t, X_{t+\tau}) = A^{\tau-1}B$, for appropriate constants A and B. Finally, we discuss the modeling power of the RNN (2) and of the HMM (3) w.r.t. the GUM, by drawing a cartography of the three models.

III. Modeling power of the GUM

A. Structure of the covariance sequence

From now on, we study the modeling power of the GUM with assumptions (7)-(9) and $m_0 = 0$. By using the Markovianity of (H_t, X_t), we build $p_\theta(h_{0:T}, x_{0:T})$, and next $p_\theta(x_{0:T})$ by marginalizing out the hidden states $h_{0:T}$. Due to the linear and Gaussian assumption (see section II-D), it is easy to check that $p_\theta(x_{0:T})$ is Gaussian, and is thus described by a mean



(a) HMM (b) RNN (c) GUM

Fig. 1: Conditional dependencies in RNN, HMM, and GUM. Dashed arrows represent deterministic dependencies. Plain ones are probabilistic dependencies. The HMM and the RNN are particular instances of the GUM. In the first case, H_t is conditionally independent of X_{t-1}; in the second one, H_t is no longer random given H_{t-1} and X_{t-1}.

vector and a covariance matrix. In this paper, we focus on the dependency structure induced by the three models. For that reason, we assume without loss of generality that $p(x_t)$ does not depend on t, and indeed without loss of generality that

$$(\star) : \forall t \in \mathbb{N}, p_\theta(x_t) = \mathcal{N}(x_t; 0; 1)$$

(In practice, choosing standard marginals can be seen as a renormalization of the data; releasing $(\star)$ would introduce translations and dilatations in the computations). The fact that $\mathrm{var} X_t = 1$ for all t implies that:

$$\beta = 1 - b^2\eta \tag{10}$$

$$\alpha = (1 - a^2 - 2abc)\eta - c^2 \tag{11}$$

The first equation is obtained from the computation of $p(x_0)$; the second one comes from the computation of $p(x_1)$. As α and β are functions of a, b, c, η, any linear Gaussian GUM under constraint $(\star)$ is fully described by these last four parameters. Nonetheless, they cannot be chosen freely since α and β have to be positive. Finally, for all $T \in \mathbb{N}^*, p(x_{0:T}) = \mathcal{N}(x_{0:T}; \mathbf{0}_T; \mathbf{\Sigma}_T)$, where the covariance matrix $\mathbf{\Sigma}_T$ has ones on its diagonal and is defined elsewhere by the covariances:

$$\forall t \in \mathbb{N}, \forall \tau \in \mathbb{N}^*, \ \mathrm{cov}(X_t, X_{t+\tau}) = (a + bc)^{\tau-1}(ab^2\eta + bc) \tag{12}$$

B. Positivity constraints on the covariance parameters

As we have just seen, for any linear and Gaussian GUM under constraint $(\star)$, $\mathrm{cov}(X_t, X_{t+\tau})$ is geometrical *ie.* $\mathrm{cov}(X_t, X_{t+\tau}) = A^{\tau-1}B$ for some A and B. Conversely, for any A and B, the Toeplitz symmetric matrix $\boldsymbol{R}_T(A, B)$ with first row $[1, B, AB, ..., A^{T-2}B]$ is not necessarily a covariance matrix, since at this point we do not know whether $\boldsymbol{R}_T(A, B)$ is indeed positive semi-definite. We thus characterize the (A, B) domain for which $\boldsymbol{R}_T(A, B)$ is a covariance matrix for all $T \in \mathbb{N}^*$. We have the following result (the proof is omitted due to lack of space but relies on the Carathéodory-Toeplitz theorem [1]):

Theorem 1: Let $\boldsymbol{R}_T(A, B)$ be a Toeplitz symmetric matrix with first row $[1, B, AB, ..., A^{T-2}B]$. $\boldsymbol{R}_T(A, B)$ is a covariance matrix for all $T \in \mathbb{N}^*$ if and only if (A, B) belongs to the parallelogram $\mathcal{P}$ defined by $A \in [-1, 1]$ and $\frac{A-1}{2} \leq B \leq \frac{A+1}{2}$; or to the line $\mathcal{D}$ defined by $B = 0$. We set $\mathcal{S} \overset{\text{def}}{=} \mathcal{P} \cup \mathcal{D}$.

IV. GUM EQUIVALENCE CLASSES

At this point we know that the observations distribution of any linear and Gaussian GUM under constraint $(\star)$ is Gaussian stationary with geometrical covariance structure $A^{\tau-1}B$, with $(A, B) \in \mathcal{S}$. Conversely, we now wonder whether any such pdf can be modeled by some GUM. In other words, we study the inverse mapping of:

$$\phi : (a, b, c, \eta) \mapsto (A = a + bc, B = ab^2\eta + bc) \tag{13}$$

One can easily show that ϕ is surjective, i.e. for any $(A, B) \in \mathcal{S}$, there exists at least one GUM providing $p_{A,B}(x_{0:t})$ and we can characterize this GUM. However it is not injective since two different GUM can model a same observation distribution.

A. Modeling power of RNN and HMM

We now study whether some distribution $p_{A,B}(x_{0:T})$ for $(A, B) \in \mathcal{S}$, can be produced either by an RNN, or by an HMM, or both, or neither of them.

- *HMM:* In the HMM case, $c = 0$, which implies $A = a$ and $B = ab^2\eta$. Taking into account this constraint yields the exact set of HMM models $\phi^{-1}(A, B)$ described by $|B| \leq |A|$ and $AB \geq 0$.
- *RNN:* In the standard RNN (2) with constraint $(\star)$, $\alpha = 0$ and $c^2 = \eta$. This second constraint comes from the way we usually initialize the RNN which is opposite to that of the GUM (the standard RNN starts by modeling the initial distribution of x_0 from which is computed the distribution of $\tilde{h}_0$ via the deterministic transition). These contraints yield to the exact set of standard RNN models $\phi^{-1}(A, B)$ described by $(B = A(2A^2 - 1)$ and $-1 \leq A \leq 1) \cup (A = B$ and $-1 \leq A \leq 1) \cup (A \in \mathbb{R}, B = 0)$.

B. Discussion

Fig. 2 shows whether a distribution can be modeled by an RNN (orange), an HMM (blue), a GUM (light blue) or none of them. For example, there is at least one instance of GUM modeling the distribution described by $(A = \frac{1}{2}, B = \frac{1}{4})$. Among all the possible instances of GUM reaching that distribution, some are actually HMM but none are RNN. In that example, the set of parameters $(a = \frac{1}{3}, b = \frac{1}{2}, c = \frac{1}{3}, \eta = 1)$ provides a proper solution (among others). As we see from the figure, any pdf in $\mathcal{S}$ can be obtained by at least one GUM; furthermore some can be produced by an HMM, but not by a RNN, and vice versa; finally some points can be reached by both.

Note that the expression of $\text{cov}(X_t, X_{t+\tau})$ given in (12) displays a behavior of the GUM already known in HMM and RNN; unless $|a + bc| = 1$, the current output is geometrically uncorrelated from the past outputs; in the case where $|a+bc| = 1$, (10) and (11) yield $\alpha = 0$ and $\beta c^2 = 0$. In other words, to have long term dependencies, determinism through time is required. This phenomenon has already been observed in [4].

Let us now comment on the dimensionality of the models, which is related to the associated set of points in Figure 2. The RNN model is parameterized only by two parameters, η and the product bc (this product bc comes from the initialization constraint $c^2 = \eta$), whence the curve. By contrast the HMM is parameterized by three parameters a, b and η, whence the two triangles.

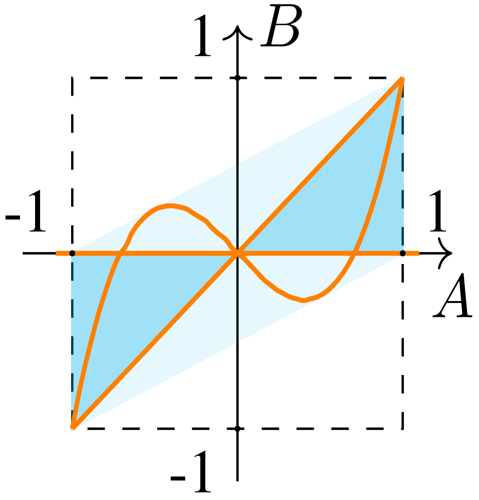


Fig. 2: Modeling powers of RNN, HMM and GUM with regards to A and B. The parallelogram (light blue) coincides with all the multivariate centered Gaussian distributions with a covariance matrix which satisfy $\text{Cov}(X_t, X_{t+\tau}) = A^{\tau-1}B$. Such distributions can be modeled by a GUM. The blue (resp. orange) areas (resp. curves) coincide with the value of A and B which can be taken by the HMM (resp. the RNN). This results shows that the modeling power of the GUM is larger than that of the HMM, which is larger than that of the RNN. However, let us note that in the context of this study, the RNN is finally defined by only 2 free parameters, the HMM relies on 3 parameters and the GUM on 4 parameters.

V. Concluding remarks

In this paper we compared two popular models for processing time series, the HMM and the RNN. First, we have encompassed both models in a common probabilistic model: both models can be seen as a particular instance of a GUM (whatever the activation function for the RNN, or the transition and likelihood distributions for the HMM). In order to address an exact comparison of the expressivity power of both models, we have focussed on the linear and Gaussian case. We have thus shown that the linear and Gaussian GUM can model a large class of stationary multivariate Gaussian distributions with geometrical covariance sequence. We also showed that none of the RNN or HMM sets is included into the other one, but that the modeling power of each model could easily be extended to the GUM framework, at the price of an augmentation of the number of parameters which characterize the model. However by considering deterministic transitions this price augmentation can be overcome. This highlights a more general trade-off between expressivity and practicability.

References

[1] N. I. Akhiezer and N. Kemmer. *The classical moment problem and some related questions in analysis*, volume 5. Oliver & Boyd Edinburgh, 1965.

[2] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear / non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, February 2002.

[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015.

[4] Y. Bengio and P. Frasconi. Diffusion of credit in markovian models. In *Advances in Neural Information Processing Systems*, pages 553–560, 1995.

[5] O. Cappé, É. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer-Verlag, 2005.

[6] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, et al. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778. IEEE, 2018.

[7] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *NeurIPS*, 2014.

[8] J. T. Connor, R. D. Martin, and L. E. Atlas. Recurrent neural networks and robust time series prediction. *IEEE transactions on neural networks*, 5(2):240–254, 1994.

[9] A. C. Harvey. *Forecasting, structural time series models and the Kalman filter*. Cambridge university press, 1990.

[10] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[11] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318, 2013.

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

[13] B. Ristic, S. Arulampalam, and N. Gordon. *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artec House, 2004.

[14] R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications (Springer Texts in Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2005.