Modelling conditional probabilities with Riemann-Theta Boltzmann Machines

Stefano Carrazza[1,2], **Daniel Krefl**[3], **Andrea Papaluca**[1]

[1] TIF Lab, Dipartimento di Fisica, Università degli Studi di Milano
[2] INFN Sezione di Milano, Via Celoria 16, 20133, Milano, Italy
[3] Department of Computational Biology, University of Lausanne, Switzerland

E-mail: stefano.carrazza@unimi.it, daniel.krefl@unil.ch, andrea.papaluca@studenti.unimi.it

Abstract. The probability density function for the visible sector of a Riemann-Theta Boltzmann machine can be taken conditional on a subset of the visible units. We derive that the corresponding conditional density function is given by a reparameterization of the Riemann-Theta Boltzmann machine modelling the original probability density function. Therefore the conditional densities can be directly inferred from the Riemann-Theta Boltzmann machine.

1. Introduction

Modelling the underlying probability density function of a dataset is a non-trivial problem, already in the low dimensional setting. In particular so in the non-normal and multi-modal case. To make things even more complicated, we often do not only want to model the probability density, but as well obtain related quantities like marginalizations, conditionals or the cumulative density function.

Several techniques to model probability densities of unknown functional form can be found in the literature. To mention a few: Kernel density estimation, mixture models, copulas, normalizing flows, and neural networks. However, each technique comes with its own advantages and drawbacks, and it is fair to say that so far no general use technique is at hand.

Inspired by Boltzmann machines [1], the authors of [2] introduced a novel kind of stochastic network, distinguished by an infinite hidden state space given by $\mathbb{Z}^{N_h}$, with N_h denoting the number of hidden units. Key quantities, like the visible sector probability density function, can be calculated in closed form involving Riemann-Theta functions [3]. Therefore, the network has been denoted as Riemann-Theta Boltzmann machine, for short RTBM. In particular, the visible sector density function is given by a novel parametric model, which can be made arbitrarily expressive via increasing the dimension of the hidden state space. The appealing property of this new kind of Boltzmann machine is that the normalization (summation over all states) is given in closed-form in terms of the Riemann-Theta function. The closed form solution allows to keep full analytic control, and in particular to derive related quantities, like for example the corresponding cumulative distribution function or conditional densities. The latter will be discussed in this note.

As conditional distributions are the essential ingredient of Bayes' theorem, modelling conditional distributions has wide applications in machine learning. For instance, in probabilistic

modelling a straight-forward application would be as a Bayes classifier.

1.1. RTBM

The visible sector probability density function of the Riemann-Theta Boltzmann machine is given by [2]

$$P(v) = \sqrt{\frac{\det T}{(2\pi)^{N_v}}} e^{-\frac{1}{2}\left((v+T^{-1}B_v)^t T (v+T^{-1}B_v)\right)} \frac{\tilde{\theta}(B_h^t + v^t W|Q)}{\tilde{\theta}(B_h^t - B_v^t T^{-1} W|Q - W^t T^{-1} W)}\,, \tag{1}$$

with

$$\tilde{\theta}(z^t|\Omega) := \sum_{n\in\mathbb{N}^{N_h}} e^{-\frac{1}{2}n^t\Omega n + n^t z}\,,$$

the Riemann-Theta function. The density $P(v)$ is parameterized by positive definite matrices Q and T of dimension $N_h \times N_h$, respectively $N_v \times N_v$, an arbitrary matrix W of dimension $N_v \times N_h$ and bias vectors B_h and B_v of dimensions $N_h \times 1$, respectively $N_v \times 1$.

As has been discussed for the first time in [2], the density $P(v)$ is a powerful density approximator due to its high intrinsic modelling capacity, determined by the number of hidden units N_h. For a given set of data samples, the underlying probability density function can be approximated by $P(v)$ via fixing the parameters in a maximum likelihood fashion. However, one should keep in mind that the modelling capacity one can reach in practice is limited by the rather high computational cost of evaluating the Riemann-Theta function, which at present limits N_h to be quite small. For details we refer to [2] and references therein.

It can be shown that $P(v)$ also possesses an interpretation in terms of a specific gaussian mixture model with an infinite number of gaussian constituents [4]. In particular, certain useful properties are inherited from the multi-variate gaussian density, like for example functional invariance under affine transformations of the datapoints v.

In this note we will discuss another useful property, which one may see as well to be inherited from the multi-variate gaussian. Namely, that the conditional density functions can again be expressed in terms of the original density $P(v)$, albeit under a different parameterization. Therefore, once we learned an approximation $P(v)$ of a multi-variate density via a RTBM, we obtain all the conditional densities automatically, as we will show in the following section.

2. Conditional probability

2.1. Derivation

We take $v = (y_1, \ldots, y_m, d_1, \ldots, d_n)$ with $m + n = N_v$ and consider the conditional density function

$$P(y|d) = \frac{P(v)}{P(d)}\,,$$

with $P(v)$ given by the density (1) and $P(d)$ its marginalization

$$P(d) = \int_{-\infty}^{\infty} [dy]\, P(y, d)\,. \tag{2}$$

It is usefull to decompose the parameter matrices T, W and B_v of the density (1) into the following block forms:

$$T = \left(\begin{array}{c|c} \bar{T}_0 & \bar{T}_1^t \\ \hline \bar{T}_1 & \tilde{T} \end{array}\right),$$

with $\bar{T}_0$ a $m \times m$ square matrix, $\bar{T}_1$ a $n \times m$ rectangular matrix and $\tilde{T}$ a $n \times n$ square matrix. Similarly,

$$W = \left(\begin{array}{c} W_0 \\ \hline \\ W_1 \end{array} \right) , \quad B_v = \left(\begin{array}{c} B_{v,0} \\ \hline B_{v,1} \end{array} \right) ,$$

with W_0 and W_1 rectangular matrices of dimension $m \times N_h$, respectively $n \times N_h$. $B_{v,0}$ and $B_{v,1}$ are column vectors of size m, respectively, n.

Imposing the above block structure onto the terms $v^t T v$, $B_v^t v$ and $v^t W$ occurring in the density (1) gives

$$\begin{aligned} v^t T v &= y^t \bar{T}_0 y + 2 d^t \bar{T}_1 y + d^t \tilde{T} d \,, \\ B_v^t v &= B_{v,0}^t y + B_{v,1}^t d \,, \\ v^t W &= y^t W_0 + d^t W_1 \,, \end{aligned} \tag{3}$$

and leads to the following expression for the joint density $P(v) = P(y,d)$:

$$P(y,d) = \sqrt{\frac{\det T}{(2\pi)^{Nv}}} \quad e^{-\frac{1}{2} y^t \bar{T}_0 y - d^t \bar{T}_1 y - \frac{1}{2} d^t \tilde{T} d - B_{v,0}^t y - B_{v,1}^t d - \frac{1}{2} B_v^t T^{-1} B_v} \\ \times \frac{\tilde{\theta}(B_h^t + y^t W_0 + d^t W_1 | Q)}{\tilde{\theta}(B_h^t - B_v^t T^{-1} W | Q - W^t T^{-1} W)} \,. \tag{4}$$

Note that the terms $B_v^t T^{-1} B_v$, $B_v^t T^{-1} W$ and $W^t T^{-1} W$ are not written in factorized form because they do not include y and therefore can be pulled out of the marginalization integration (2). After explicitation of the theta function the integral we have to solve to obtain $P(d)$ reads

$$I = \int_{-\infty}^{+\infty} [dy] \quad e^{-\frac{1}{2} y^t \bar{T}_0 y + (n^t W_0^t - d^t \bar{T}_1 - B_{v,0}^t) y} \,,$$

which is the well known generalized gaussian integral

$$\int_{-\infty}^{\infty} [dx] \quad e^{-\frac{1}{2} x^t A x + b^t x} = \sqrt{\frac{(2\pi)^n}{\det A}} \quad e^{\frac{1}{2} b^t A^{-1} b} .$$

Hence, we find the explicit expression

$$P(d) = \sqrt{\frac{\det T}{(2\pi)^{Nv}}} \quad e^{-\frac{1}{2} d^t \tilde{T} d - B_{v,1}^t d - \frac{1}{2} B_v^t T^{-1} B_v} \sqrt{\frac{(2\pi)^m}{\det \bar{T}_0}} \quad e^{(B_{v,0} + \bar{T}_1^t d)^t \bar{T}_0^{-1} (B_{v,0} + \bar{T}_1^t d)} \\ \times \frac{\tilde{\theta}(B_h^t + d^t W_1 - (B_{v,0} + \bar{T}_1^t d)^t \bar{T}_0^{-1} W_0 | Q - W_0^t \bar{T}_0^{-1} W_0)}{\tilde{\theta}(B_h^t - B_v^t T^{-1} W | Q - W^t T^{-1} W)} \quad , \tag{5}$$

which in turn leads to the desired conditional probability

$$P(y|d) = \frac{P(v)}{P(d)} = \sqrt{\frac{\det \bar{T}_0}{(2\pi)^m}} \quad e^{-\frac{1}{2} y^t \bar{T}_0 y - (B_{v,0} + \bar{T}_1^t d)^t y - \frac{1}{2} (B_{v,0} + \bar{T}_1^t d)^t \bar{T}_o^{-1} (B_{v,0} + \bar{T}_1^t d)} \\ \times \frac{\tilde{\theta}(B_h^t + d^t W_1 + y^t W_0 | Q)}{\tilde{\theta}(B_h^t + d^t W_1 - (B_{v,0} + \bar{T}_1^t d)^t \bar{T}_0^{-1} W_0 | Q - W_0^t \bar{T}_0^{-1} W_0)} \,. \tag{6}$$

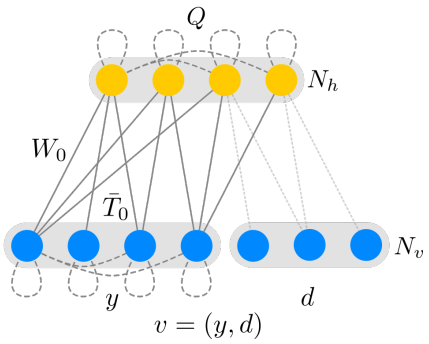


Figure 1: Graphical representation of a conditional probability RTBM architecture.

Note that $P(y|d)$ as given above is easily seen to correspond to a probability density function $P(v)$ of a $N_v = m$ RTBM with the following reparameterization:

$$\begin{aligned} T &\to \bar{T}_0\,, \\ W &\to W_0\,, \\ B_v &\to B_{v,0} + \bar{T}_1^t d\,, \\ B_h &\to B_h + W_1^t d\,. \end{aligned} \tag{7}$$

We conclude that starting from a "parent" RTBM modelling a multidimensional density, we can generate "child" RTBMs modelling its conditional probabilities simply by choosing the parameters accordingly. For illustration, we sketched the surviving parameters of the corresponding network architecture in figure 1.

2.2. Examples

As an example of what we have discussed above, let us consider the multivariate Student's t-distribution:

$$f(x) = \frac{\Gamma((v+p)/2)}{\Gamma(v/2)\,(v\pi)^{\frac{p}{2}}\,|\Sigma|^{\frac{1}{2}}}\left[1 + \frac{1}{v}\,(x-\mu)^t\Sigma^{-1}(x-\mu)\right]^{-\frac{v+p}{2}}. \tag{8}$$

As derived in [5], [6], it possesses an analytic expression for its conditional. For $x = (x_1, x_2)$ one finds that

$$x_2\,|\,x_1 \;\sim\; t_{p_2}\left(\mu_{2|1}, \frac{v+d_1}{v+p_1}\,\Sigma_{22|1}, v+p_1\right). \tag{9}$$

Note that the conditional is again a t-distribution. This allows us to easily compare the analytical conditional with the one obtained from an RTBM trained to fit the corresponding t-distribution.

We consider a t-distribution with the following parameters:

$$\mu = (0,0), \quad \Sigma = \begin{pmatrix} 2 & -1 \\ -1 & 4 \end{pmatrix}, \quad v = 6, \tag{10}$$

A RTBM with $N_v = 2$ and $N_h = 2$ is trained with the CMA-ES algorithm on 5000 samples thereof. The best solution was found at a log-likelihood loss of $\sim 1.3 \cdot 10^4$. The found RTBM parameters fitting the above t-distribution read:

$$\begin{aligned} W &= \begin{pmatrix} -1.11 & 1.02 \\ -0.66 & 0.60 \end{pmatrix}, \quad T = \begin{pmatrix} 0.56 & 0.18 \\ 0.18 & 0.30 \end{pmatrix}, \quad B_v = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \\ B_h &= \begin{pmatrix} 8.22 \\ 17.40 \end{pmatrix}, \quad Q = \begin{pmatrix} 24.15 & -0.44 \\ -0.44 & 41.57 \end{pmatrix}. \end{aligned} \tag{11}$$

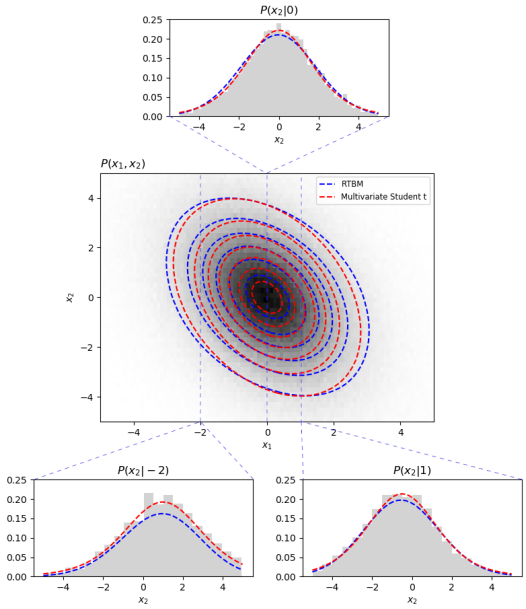


Figure 2: In the center the two dimensional histogram sampled by a multivariate Student t-distribution (8) with parameters (10) is shown. The contour plot of the trained RTBM (11) is shown in blue, the analytic distribution in red. On the bottom the comparison between the analytic conditionals given by (9) with the ones generated using (7) are shown.

Multivariate Student t		2D Example		3D Example	
Conditional	MSE	Conditional	MSE	Conditional	MSE
$P(x_2\|-2)$	$3.255 \cdot 10^{-4}$	$P(y\|2)$	$1.176 \cdot 10^{-4}$	$P(y_1, y_2\|-0.4)$	$4.953 \cdot 10^{-5}$
$P(x_2\|0)$	$4.083 \cdot 10^{-5}$	$P(y\|1.3)$	$6.888 \cdot 10^{-5}$	$P(y_1, y_2\|-0.6)$	$6.775 \cdot 10^{-5}$
$P(x_2\|1)$	$4.433 \cdot 10^{-5}$	$P(y\|0.4)$	$2.538 \cdot 10^{-4}$	$P(y_1, y_2\|-0.8)$	$5.304 \cdot 10^{-5}$

Table 1: MSE calculated for each of the conditional distributions shown in figure 2, 3a, 3b. Note that for the 2D and 3D examples, the MSE is calculated with respect to the empirical conditional derived from the relative histogram.

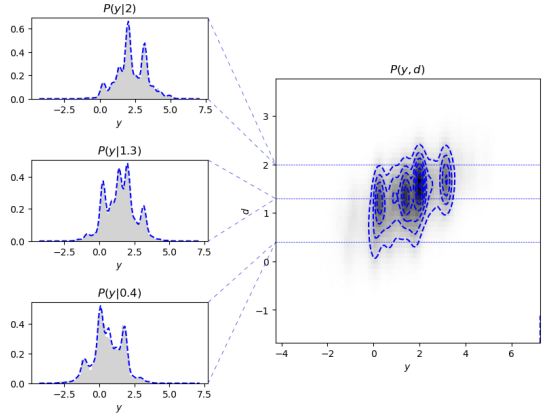
The analytic contribution and its RTBM based fit are shown in figure 2. The figure also shows three examples of conditionals derived following section 2.1 and the corresponding analytic solution obtained from (9).

In order to quantify the error in modelling the conditionals with the RTBM, we calculated the mean squared error (MSE) between the analytic distribution and the RTBM fit at the sample points used for training. The results are shown in table 1.

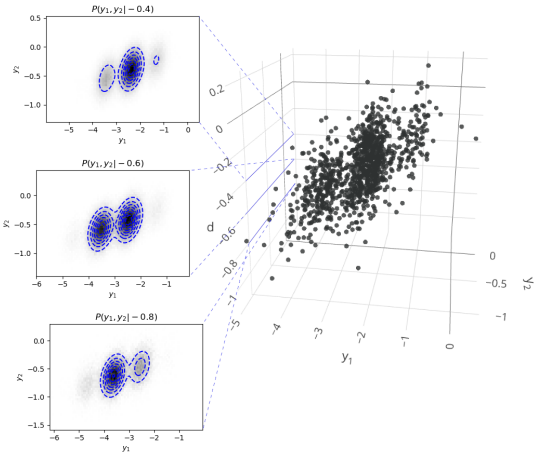
For illustration purposes, we consider two further examples for the conditional densities obtainable through relation (7). We only show results with $N_v = 2$ and $N_v = 3$ in order to simplify the visualization of results. However one should note that the current methodology is valid for higher dimensional examples as well.

On both examples the RTBM parameters are initialized manually in order to achieve a more complicated distribution $P(v)$ than the t-distribution discussed above. The conditional distributions are obtained as before following the transformations presented in section 2.1. On both examples we compare the resulting conditional probabilities with the empirical distributions obtained using the RTBM sampling algorithm presented in [4].

The RTBM in the two dimensional example with $N_v = 2$ and $N_h = 4$ is defined via the



(a) On the right, the two dimensional histogram of a RTBM with $N_h = 4$, $N_v = 2$ and initialized with parameters (12) is shown. The contour plot is represented in blue. On the left the conditional probabilities $P(y|d)$ for three different values $d = 2, 1.3, 0.4$ are shown. The relative one dimensional histogram is shown in grey.



(b) On the right, a sampling of the three dimensional distribution given by a RTBM with $N_v = 3$, $N_h = 1$ and parameters (13). On the left, the contour plots of the conditional probabilities $P(y_1, y_2|d)$, represented by the dashed blue curves, evaluated at $d = -0.4, -0.6, -0.8$ are shown respectively. The corresponding two dimensional empirical histograms are shown in gray.

Figure 3: Examples of conditional probabilities.

following parameter choice:

$$W = \begin{pmatrix} 18.54 & 3.02 & -12.89 & -5.45 \\ 0.46 & 1.01 & -1.32 & -5.54 \end{pmatrix}, \quad T = \begin{pmatrix} 28.77 & 0 \\ 0 & 6.3 \end{pmatrix}, \quad B_v = \begin{pmatrix} -1.76 \\ -2.69 \end{pmatrix},$$
$$B_h = \begin{pmatrix} -0.31 \\ 2.29 \\ 1.65 \\ -2.73 \end{pmatrix}, \quad Q = \begin{pmatrix} 15.48 & 8.82 & -3.19 & -3.67 \\ 8.82 & 17.99 & 8.94 & -4.04 \\ -3.19 & 8.94 & 15.74 & 4.14 \\ -3.67 & -4.04 & 4.14 & -5.54 \end{pmatrix}. \tag{12}$$

The corresponding distribution and derived conditional distributions are shown in figure 3a.

For the three dimensional example, we define a RTBM with $N_v = 3$ and $N_h = 1$. The chosen parameters are

$$W = \begin{pmatrix} -15.76 \\ 2.29 \\ 2.09 \end{pmatrix}, \quad T = \begin{pmatrix} 16.02 & -6.52 & -6.76 \\ -6.52 & 29.04 & -2.56 \\ -6.76 & -2.56 & 42.16 \end{pmatrix}, \quad B_v = \begin{pmatrix} 1.08 \\ -0.67 \\ 4.86 \end{pmatrix},$$
$$B_h = \begin{pmatrix} 3.17 \end{pmatrix}, \quad Q = \begin{pmatrix} 19.18 \end{pmatrix}. \tag{13}$$

The corresponding distribution and examples of obtained two dimensional conditional densities are show in figure 3b.

The MSE between the conditional distributions derived from the RTBM and the empirical conditionals derived from histograms are listed in table 1. However, one should keep in mind that the purpose of the MSE calculation in these two examples is soley to illustrate that the relation (7) is correct. By construction, the RTBM constitute the true (analytic) underlying distribution for these examples.

Acknowledgments
S.C. is supported by the European Research Council under the European Unions Horizon 2020 research and innovation Programme (grant agreement number 740006).

References

[1] Hinton G E and Sejnowski T J May 1983 Analyzing cooperative computation *In Proceedings of the 5th Annual Congress of the Cognitive Science Society* New York
[2] Krefl D, Carrazza S, Haghighat B and Kahlen J 2017 Riemann-theta boltzmann machine *Preprint* stat-ml/1712.07581
[3] Mumford D. 1983 Tata lectures on theta I *Progress in Mathematics* book series **28** Boston
[4] Carrazza S and Krefl D 2018 Sampling the riemann-theta boltzmann machine *Preprint* stat-ml/1804.07768
[5] Nadarajah S and Kotz S 2005 Mathematical properties of the multivariate t distribution *Acta Applicandae Mathematicae* **89** 53-84 Springer
[6] Ding P 2016 On the conditional distribution of the multivariate t distribution *Preprint* math-st/1604.00561