RBM cumulants

Tor Erlend Fjelde

January 19, 2019

Contents

1 Notation

- $K(t)$ denotes the cumulant-generating function of a probability distribution
- $\kappa^{(n)} = \partial_t^n|_{t=0} K(t)$ denotes the n-th cumulant of a probability distribution
- $E(\mathbf{v}, \mathbf{h}) = - \sum_j a_j(v_j) - \sum_\mu b_\mu(h_\mu - \sum_j \sum_\mu v_j W_{j\mu} h_\mu$ is the energy
- $a_j(v_j)$ denotes the bias of v_j in the energy function $E(\mathbf{v}, \mathbf{h})$
- $b_\mu(h_\mu)$ denotes the bias of h_μ in the energy function $E(\mathbf{v}, \mathbf{h})$
- $q_\mu(h_\mu) = e^{b_\mu(h_\mu)} / Z_\mu$, where
- $K_\mu(t) = \log \sum_{h_\mu} q_\mu(h_\mu) e^{t h_\mu}$ denotes cumulative-generating function for q_μ
- $\kappa_\mu^{(n)}$ denote the n-th cumulant of the distribution $q_\mu(h_\mu)$
- $A_{n,m}$ denotes the *Eulerian number* corresponding to n and m

2 Cumulants

Cumulants of a distribution are useful since

- they provide an alternative to moments of the distribution,
- the moments determine cumulants in the sense that *any two probability distributions whose moments are identical will have identical cumulants*, and the converse is also true.

Definition 2.1. The cumulants of a random variable X can be defined using the **cumulant-generating-funcion** $K(t)$, which is

$$K(t) = \log M(t) = \log \mathbb{E}[e^{tX}]$$

where $M(t) = \mathbb{E}[e^{tX}]$ is the *moment-generating function.*

Definition 2.2. The **cumulants** $\kappa^{(n)}$ are obtained from a power series expansion of the 2.1:

$$K(t) = \sum_{n=1}^{\infty} \kappa^{(n)} \frac{t^n}{n!} = \mu t + \sigma^2 \frac{t^2}{2} + \dots$$

This is a Mclauring series, hence

$$\kappa^{(n)} = \frac{\partial^n}{\partial t^n}\Big|_{t=0} K(t) = K^{(n)}(0)$$

3 RBMs

As seen from Eqs. 203-207 in [1], we can define a energy function which only depends on $\mathbf{v}$, by letting

$$p(\mathbf{v}) := \frac{1}{Z} e^{-E(\mathbf{v})} = \frac{1}{Z} \sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}$$

Therefore,

$$\begin{aligned}
E(\mathbf{v}) &= -\log \sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} \\
&= -\log \sum_{\mathbf{h}} \exp\Big(a_j(v_j) + b_\mu(h_\mu) + v_j W_{j\mu} h_\mu\Big) \\
&= -\log \exp\Big(\sum_j a_j(v_j)\Big) \\
&\quad - \log \sum_{\mathbf{h}} \exp\Big(\sum_\mu b_\mu(h_\mu) + \sum_j v_j W_{j\mu} h_\mu\Big) \\
&= -\log \exp\Big(\sum_j a_j(v_j)\Big) \\
&\quad - \log \prod_{\mu=1}^{|\mathcal{H}|} \sum_{h_\mu} \exp\Big(b_\mu(h_\mu)\Big) \exp\Big(\sum_j v_j W_{j\mu} h_\mu\Big) \\
&= -\sum_j a_j(v_j)\Big) \\
&\quad - \sum_\mu \log \sum_{h_\mu} \exp\Big(b_\mu(h_\mu)\Big) \exp\Big(\sum_j v_j W_{j\mu} h_\mu\Big)
\end{aligned}$$

If we then introduce the distributions

$$q_\mu(h_\mu) = \frac{1}{Z_\mu} e^{b_\mu(h_\mu)}$$

for each h_μ, then the corresponding 2.1 is given by

$$K_\mu(t) = \log M_\mu(t) = \log \mathbb{E}[e^{th_\mu}] = \log \sum_{h_\mu} q_\mu(h_\mu) e^{th_\mu}$$

And using the expansion of K_μ seen in the Definition 2.2, we get

$$K_\mu(t) = \sum_n \kappa_\mu^{(n)} \frac{t^n}{n!}$$

Observe that if $t = \sum_j v_j W_{j\mu}$, then we have

$$K_\mu\Big(\sum_j v_j W_{j\mu}\Big) = \log \sum_{h_\mu} \exp\Big(b_\mu(h_\mu)\Big) \exp\Big(\sum_j v_j W_{j\mu} h_\mu\Big)$$

Subsituting into Eq. 3, we get

$$\begin{aligned} E(\mathbf{v}) &= -\sum_j a_j(v_j) - \sum_\mu K_\mu\Big(\sum_j v_j W_{j\mu}\Big) \\ &= -\sum_j a_j(v_j) - \sum_\mu \sum_n \kappa_\mu^{(n)} \frac{\Big(\sum_j v_j W_{j\mu}\Big)^n}{n!} \\ &= -\sum_j a_j(v_j) - \sum_\mu \kappa_\mu^{(1)} \sum_i W_{i\mu} v_i \\ &\quad - \frac{1}{2} \sum_\mu \kappa_\mu^{(2)} \sum_{i,j} W_{i\mu} W_{j\mu} v_i v_j - \dots \end{aligned}$$

From this expression, we can see that an RBM can capture *moments* of arbitary order for the random variables V_j.

It's important to remember that in this expansion from Definition 2.2 we're expanding around the point $t = 0$, which in this case would be

$$\sum_j W_{j\mu} v_j = 0, \quad \forall \mu = 1, \dots, |\mathcal{H}|$$

4 Bernoulli RBM

4.1 The cumulants

Now suppose we are using a Bernoulli RBM, that is,

$$a_j(v_j) = a_j v_j, \qquad b_\mu(h_\mu) = b_\mu h_\mu$$

Then

$$q_\mu(h_\mu) = \frac{e^{b_\mu h_\mu}}{1 + e^{b_\mu}}$$

and therefore,

$$\begin{aligned} K_\mu(t) &:= \log \sum_{h_\mu} q_\mu(h_\mu) e^{t h_\mu} \\ &= \log\Big(q_\mu(0) + q_\mu(1) e^t\Big) \\ &= \log\Big(1 + e^{b_\mu + t}\Big) - \log\Big(1 + e^{b_\mu}\Big) \end{aligned}$$

Thus,

$$\begin{aligned} \frac{\partial K_\mu}{\partial t} &= \frac{\partial}{\partial t} \log\Big(1 + e^{b_\mu + t}\Big) \\ &= \frac{e^{b_\mu + t}}{1 + e^{b_\mu + t}} \\ &= \sigma\big(b_\mu + t\big) \end{aligned} \tag{1}$$

Letting $z = b_\mu + t$, we observe that we have the relation

$$K_\mu^{(n+1)}(z) = \sigma^{(n)}(z) \tag{2}$$

This is useful, since one can obtain a general expression for the n-th derivative of the sigmoid function $\sigma(z)$. Remark 5 in [2] tells us

$$\sigma^{(n)}(z) = \sum_{k=1}^{n} \big(-1\big)^{k-1} A_{n,k-1} \sigma(z)^k \big(1 - \sigma(z)\big)^{n+1-k} \tag{3}$$

where $A_{n,k-1}$ are known as the *Eulerian numbers*, which can easily be computed using the recursion

$$A_{n,m} = (n-m)A_{n-1,m-1} + (m+1)A_{n-1,m} \tag{4}$$

or in explicit form,

$$A_{n,m} = \sum_{k=0}^{m} \big(-1\big)^k \binom{n+1}{k} \big(m+1-k\big)^n$$

See Table 1 in [2] for an example of some Eulerian numbers. Substituting Eq. 2 into Eq. 3, we get

$$K_\mu^{(n+1)}(z) = \sum_{k=1}^{n} \big(-1\big)^{k-1} A_{n,k-1} \sigma(z)^k \big(1 - \sigma(z)\big)^{n+1-k}$$

as a general expression for the (n + 1)-th derivative of the cumulant-generating function. Substituting back in $z = b_\mu + t$, and letting $t = 0$, we get the general expression for the (n + 1)-th cumulant

$$\kappa_\mu^{(n+1)} = \sum_{k=1}^{n} \big(-1\big)^{k-1} A_{n,k-1} \sigma(b_\mu)^k \big(1 - \sigma(b_\mu)\big)^{n+1-k} \tag{5}$$

4.2 Second order interactions between visible units

This section is due to [3], which is the group I tagged along with for my summer project.

With Eq. 5 we can rewrite the expansion of $E(\mathbf{v})$ in Eq. 3 as

$$E(\mathbf{v}) = -\sum_j a_j v_j - \sum_\mu \sum_{n=1}^{\infty} \frac{\kappa_\mu^{(n)}}{n!} \Big(\sum_j W_{j\mu} v_j\Big)^n$$

Say we're interested in the *second order interactions* between two visible units v_{j_1} and v_{j_2}. We then observe that every n-th term with $n \geq 2$ in the above sum will contribute to the second order interactions between v_{j_1} and v_{j_2} by summing over all possible

$$W_{i_1\mu} W_{i_2\mu} \cdots W_{i_n\mu} v_{i_1} v_{i_2} \cdots v_{i_n}$$

with $i_m \in \{j_1, j_2\}$, $m = 1, \ldots, n$ and excluding the case where $i_m = j_1, \forall m$ and $i_m = j_2, \forall m$, as these cases will only contribute to the first order moment.

Suppose we want to assign $2 \leq k < n$ of these $i_1, \ldots, i_n$ to j_1 and the rest to j_2, then we have $\binom{n}{k}$ ways of making these assignments. Therefore the sum over all possible combinations above can be compactly written

$$\begin{aligned}
&\Bigg(\sum_{i_1 \in \{j_1, j_2\}} \cdots \sum_{i_n \in \{j_1, j_2\}} W_{i_1\mu} \cdots W_{i_n}\Bigg) - \big(W_{j_1\mu}\big)^n - \big(W_{j_2\mu}\big)^n \\
=&\Bigg(\sum_{m=0}^{n} \binom{n}{m} \big(W_{j_1\mu}\big)^m \big(W_{j_2\mu}\big)^{n-m}\Bigg) - \big(W_{j_1\mu}\big)^n - \big(W_{j_2\mu}\big)^n \\
=&\Big(W_{j_1\mu} + W_{j_2\mu}\Big)^n - \big(W_{j_1\mu}\big)^n - \big(W_{j_2\mu}\big)^n
\end{aligned}$$

where we have made use of the Binomial theorem in the last equality. Hence, all terms involving second order interactions between v_{j_1} and v_{j_2} is given by

$$\sum_{\mu} \sum_{n=2}^{\infty} \frac{\kappa_\mu^{(n)}}{n!} \Big[\Big(W_{j_1\mu} + W_{j_2\mu} \Big)^n - \Big(W_{j_1\mu} \Big)^n - \Big(W_{j_2\mu} \Big)^n \Big]$$

Furthermore, some clever people noticied that this can be expressed using the shift-operator

$$\begin{aligned}
& \sum_{\mu} \sum_{n=2}^{\infty} \Big[\Big(W_{i_1\mu} + W_{j_2\mu} \Big)^n - \Big(W_{j_1\mu} \Big)^n - \Big(W_{j_2\mu} \Big)^n \Big] \frac{1}{n!} \partial_t^n K_\mu(t) \big|_{t=0} \\
= \;& \sum_{\mu} \sum_{n=0}^{\infty} \Big[\Big(W_{i_1\mu} + W_{j_2\mu} \Big)^n - \Big(W_{j_1\mu} \Big)^n - \Big(W_{j_2\mu} \Big)^n \Big] \frac{1}{n!} \partial_t^n K_\mu(t) \big|_{t=0} \\
& - \sum_{\mu} \Big[\Big(W_{i_1\mu} + W_{j_2\mu} \Big) - \Big(W_{j_1\mu} \Big) - \Big(W_{j_2\mu} \Big) \Big] \partial_t K_\mu(t) \big|_{t=0} \\
& - \sum_{\mu} \Big[1 - 1 - 1 \Big] K_\mu(0) \\
= \;& \sum_{\mu} \Big[\exp\Big((W_{j_1\mu} + W_{j_2\mu}) \partial_t \Big) - \exp\Big(W_{j_1\mu} \partial_t \Big) - \exp\Big(W_{j_2\mu} \partial t \Big) + 1 \Big] K_\mu(t) \big|_{t=0}
\end{aligned}$$

where the negative term involving $\partial_t K_\mu(t)$ vanishes since the coefficient is zero. The shift-operator has the property $\exp\big(a \; \partial x\big) f(x) = f(a + x)$, hence the above expression simply becomes

$$\sum_{\mu} \Big[K_\mu \Big(W_{j_1\mu} + W_{j_2\mu} \Big) - K_\mu \Big(W_{j_1\mu} \Big) - K_\mu \Big(W_{j_2\mu} \Big) + K_\mu(0) \Big]$$

Providing us with a closed form expression for the second order interaction between visible units for a Bernoulli RBM.

4.3 Issues with numerical approximation to series

Before the insanely clever members of the Edinburgh Lattice QCD team obtained the closed form expression for the second order interactions seen in Eq. 4.2, we attempted to compute truncated series using Eq. 4.2. Attempting with orders up to $n \approx 50$, the terms would blow up, giving us unreasonable results. In this section we investigate why the numerical computations turned out to be insufficient.

4.3.1 Bounding the coefficients

First we observed that one can obtain an upper-bound for the magintude of the coefficients in the expansion of Eq. 3. As noted in Eq. 5, we can write the (n + 1)-th cumulant as

$$\kappa_\mu^{(n+1)} = \sum_{k=1}^{n} \big(-1 \big)^{k-1} A_{n,k-1} \sigma(b_\mu)^k \big(1 - \sigma(b_\mu) \big)^{n+1-k}$$

Or equivalently, by shifting k by -1,

$$\kappa_\mu^{(n+1)} = \sum_{k=0}^{n-1} \big(-1 \big)^{k} A_{n,k} \sigma(b_\mu)^{k+1} \big(1 - \sigma(b_\mu) \big)^{n-k}$$

Let

$$\alpha_\mu = \max \big\{ \sigma(b_\mu), \big(1 - \sigma(b_\mu) \big) \big\}$$

we observe

$$\begin{aligned}\left|\kappa_\mu^{(n+1)}\right| &= \left|\sum_{k=0}^{n-1}\big(-1\big)^k A_{n,k}\sigma(b_\mu)^{k+1}\big(1-\sigma(b_\mu)\big)^{n-k}\right| \\ &\leq \sum_{k=0}^{n-1}\left|A_{n,k}\right|\left|\sigma(b_\mu)^{k+1}\big(1-\sigma(b_\mu)\big)^{n-k}\right| \\ &\leq \sum_{k=0}^{n-1}\left|A_{n,k}\right|\alpha_\mu^{k+1}\alpha_\mu^{n-k} \\ &= \alpha_\mu^{n+1}\sum_{k=0}^{n-1}A_{n,k} \\ &= \alpha_\mu^{n+1}n!\end{aligned}$$

where we've used the fact that

$$A_{n,k} \geq 0, \quad \forall n \in \mathbb{N} \quad \text{and} \quad \sum_{k=0}^{n-1}A_{n,k} = n!, \quad n \geq 1$$

Relating to the coefficient of the Taylor expansion for the energy in Eq. 3, we instead consider $\kappa_\mu^{(n+1)}/(n+1)!$:

$$\frac{\left|\kappa_\mu^{(n+1)}\right|}{(n+1)!} \leq \frac{\alpha_\mu^{n+1}}{n+1}$$

Clearly $\alpha_\mu \in \left(\frac{1}{2},1\right)$ for b_μ in any bounded interval, hence this upper-bound decreases rapidly wrt. n.

4.3.2 Initial results

Unfortunately, the upper bound provided by α_μ cannot necessarily tell us anything about whether or not higher-order terms will have non-vanishing contributions, unless $\big(W_{j_1\mu}\big)^{n_1}\big(W_{j_2\mu}\big)^{n_2} \leq \frac{1}{\alpha^n}$, but it can serve as verifaction of the numerical procedure used by ensuring that the bounds hold for all n.

Using the weights of a RBM trained on a 16×16 Ising model with $T = 1.8$, we computed the second order interactions using Eq. 4.2 and then we computed the corresponding $\{\alpha_\mu\}$. In this particular case, we observed that the bounds established in Eq. 4.3.1 were preserved for $n \leq 39$, but not for $n > 39$.[1]

As a response, we computed the identity in Eq. 4.3.1 for these values of n, and observed that the identity failes for large n, and for $n = 40$ in particular the resulting value was *negative.*

Hence the numerical approximation fails due to numerical errors, especially errors accumulated by the intermediate computations of the Eulerian numbers.

4.3.3 Results after improvement

In these experiments, we were using weights obtained from RBMs trained on different 2D Ising models.

A member of the research group suggested we have a look at the magnitudes of the weights in comparison with upper bounds on the coefficients of the series; if the magnitudes of the weights were greater than the upper bounds on the coefficients, then the $W_{j_1\mu}$ seen earlier will dominate. In our trained RBMs, several $W_{j_1\mu}$ had magnitudes greater than 1. Therefore we would expect higher-order terms to have non-negligible contributions to the series, that is, truncated series are not necessarily expected to provide a good approximation.

We could also observe that computing $\kappa_\mu^{(n)}/n!$ for $n \leq 40$ numerically, the upper-bounds where satisfied, but when consdering $n = 50$, for multiple values of b_μ, $\kappa^{(n)}/n!$ were larger than the respective upper-bounds. This made us suspicious of numerical errors in the intermediate computations leading to significant errors in

[1] 128-bit floating points were used for all computations.

the computation of $\kappa^{(n)}/n!$. Suspecting that this had something to do with the computation of the Eulerian numbers, since this computation involves large binomial series, we decided to check if the following was satisfied

$$\sum_{m=0}^{n-1} A_{n,m} = n!$$

For $n = 50$ we observed $\sum_{m=0}^{n-1} A_{n,m}$ to be several orders of magnitude larger than $n!$. Hence, numerical errors also play a part in the failure of approximating the infinite series of Eq. 4.2 using numerically computed truncated series.[2]

4.3.4 Note on absolute convergence and the upper-bound

We know

$$\left|\sum_{\mu} \frac{\kappa_{\mu}^{(n)}}{n!}(W_{j_1\mu})^{n_1}(W_{j_2\mu})^{n_2}\right| \leq \sum_{\mu} \frac{\alpha_{\mu}^{n}}{n}\left|(W_{j_1\mu})^{n_1}(W_{j_2\mu})^{n_2}\right|$$

In the following we consider a specific μ, and therefore drop μ form the notation. We observe that the terms in the series defines a convergent sequence in the case where

$$\frac{\alpha^n}{n}(W_{j_1})^{n_1}(W_{j_2})^{n_2} < 1$$

where $n_1 + n_2 = n$, or rather,

$$(W_{j_1})^{n_1}(W_{j_2})^{n_2} < \frac{n}{\alpha^n}$$

which, in the case

$$(W_{j_1})^{n_1}(W_{j_2})^{n_2} > 1$$

For every $x \in \mathbb{R}$ such that $x > 1$, we know

$$\exists N \in \mathbb{N} : \quad x^n > n, \quad \forall n > N$$

Hence, the above is true if and only if

$$\alpha^n(W_{j_1})^{n_1}(W_{j_2})^{n_2} < 1$$

Therefore, for this to define an *absolutely* convergent sequence, we α^n to be smaller than the inverse of the absolute value of the weights.

Therefore the question is; **can we improve this bound**, and then show that a series which ought to have convergent coefficients does *not* have this when using numerical approximation?

5 TODO Further work: partial sums

Suppose $n \in$ Odds, then we will have a an odd number of terms. Let

$$\begin{aligned} S_1 &= \sum_{k=0}^{(n/2)-2} \big(-1\big)^k A_{n,k}\sigma^k\big(1-\sigma^k\big)^{n-1-k} \\ S_2 &= \sum_{k=(n/2)}^{n-1} \big(-1\big)^k A_{n,k}\sigma^k\big(1-\sigma^k\big)^{n-1-k} \end{aligned}$$

Then we can write

$$\begin{aligned} \kappa_{\mu}^{(n)} &= S_1 + S_2 + \big(-1\big)^{n/2} A_{n,(n/2)}\sigma^{(n/2)-1}\big(1-\sigma\big)^{(n/2)-1} \\ &= S_1 + S_2 + A_{n,(n/2)}\sigma^{(n/2)-1}\big(1-\sigma\big)^{(n/2)-1} \end{aligned}$$

[2] 128-bit floating points were used for all computations.

$A_{n,(n/2)}$ is the largest Eulerian number of any $n-1 \in$ Odds, hence this term is dominated by σ.
Further, due to the $n \in$ Odds:

$$\sum_{k=1}^{n} \big(-1\big)^{k-1} A_{n,k-1} \sigma^{k+1} \big(1-\sigma\big)^{n-k}$$

Then

$$S_1 = \sum_{k=1}^{(n-1)/2} \big(-1\big)^{k-1} A_{n,k-1} \sigma^{k+1} \big(1-\sigma\big)^{n-k}$$

$$S_2 = \sum_{k=(n+3)/2}^{n} \big(-1\big)^{k-1} A_{n,k-1} \sigma^{k+1} \big(1-\sigma\big)^{n-k}$$

Letting $k' = k - \frac{n+1}{2}$, we have $k = k' + \frac{n+1}{2}$, thus

$$\sum_{k=1}^{(n-1)/2} \big(-1\big)^{n-k-1} A_{n,n-1-k} \sigma^{n-k} \big(1-\sigma\big)^{k+1}$$

which is just

$$\sum_{k=1}^{(n-1)/2} \big(-1\big)^{k-1} A_{n,k-1} \sigma^{n-k} \big(1-\sigma\big)^{k+1}$$

Hence,

$$S_1 + S_2 = \sum_{k=1}^{(n-1)/2} \big(-1\big)^{k-1} A_{n,k-1} \Big[\sigma^{k+1} \big(1-\sigma\big)^{n-k} + \sigma^{n-k} \big(1-\sigma\big)^{k+1}\Big]$$

5.1 Ideas

1. Separate into two sums S_1 and S_2 plus an extra term corresponding to the middle Eulerian number, which also is the *dominant* Eulerian number

2. Observe that S_1 and S_2 can be combined to produce a single sum with

$$\sigma^a \big(1-\sigma\big)^{b-a} - \sigma^{b-a} \big(1-\sigma\big)^a$$

   for some $a, b \in \mathbb{N}$.

3. Obtain stricter inequality

4. Do same for $n \in$ Evens, but ignore middle term

References

[1] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G. R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab. A high-bias, low-variance introduction to machine learning for physicists. *CoRR*, 2018.

[2] Ali A Minai and Ronald D Williams. On the derivatives of the sigmoid. *Neural Networks*, 6(6):845–853, 1993.

[3] Guido Cossu, Luigi Del Debbio, Tommaso Giani, Ava Khamseh, and Michael Wilson. Machine learning determination of dynamical parameters: the ising model case. *CoRR*, 2018.