

---

# Supplementary material for RNN-RNADE: A Tractable Model for High-Dimensional Real Valued Sequences

---

Anonymous Author(s)

Affiliation

Address

email

## 1 Introduction

This document contains supplementary material for the paper “RNN-RNADE: A Tractable Model for High-Dimensional Real Valued Sequences”. Section 2 describes an algorithm for calculating the gradients of the joint RNN-RNADE model and updating the parameters of the model based on these gradients. The last 2 sections describe videos that are included in the supplementary material, which were generated as outputs of the experiments.

## 2 Training Algorithm

This section describes an algorithm for training the RNN-RNADE using the back-propagation. Further details on the derivation of the gradients for the RNADE can be found in the supplementary for the original RNADE paper [1]. We first present the algorithm for the calculation of the gradients for the RNADE. Our gradients differ slightly from the gradients in [1], because we use sigmoid activation functions for the RNADE instead of rectified linear activations as prescribed in [1]. In our initial experiments with *mocap* dataset, we found that the training algorithm for the joint model was more stable and less prone to exploding gradients, when sigmoid activations were used for the RNADE.

Algorithm 1 shows the how the gradients for the RNADE are computed. Algorithm 2 then shows how the gradients from the RNADE can be used to train the joint model.

054  
055  
056  
057  
058  
059  
060  
061  
062  
063  
064  
065  
066  
067  
068  
069  
070  
071  
072  
073  
074  
075  
076  
077  
078  
079  
080  
081  
082  
083  
084  
085  
086  
087  
088  
089  
090  
091  
092  
093  
094  
095  
096  
097  
098  
099  
100  
101  
102  
103  
104  
105  
106  
107

---

**Algorithm 1** RNADE gradients

---

```

a  $\leftarrow$  c
for  $d$  from 1 to  $D$  do
    a  $\leftarrow$  a +  $x_d \mathbf{W}_{:,d}$ 
end for
for  $d$  from  $D$  to 1 do
     $\psi \leftarrow \rho_d \mathbf{a}$ 
     $\mathbf{h} \leftarrow \sigma(\psi)$ 
     $\mathbf{z}^\alpha \leftarrow \mathbf{V}_d^{\alpha T} \mathbf{h} + \mathbf{b}_d^\alpha$ 
     $\mathbf{z}^\mu \leftarrow \mathbf{V}_d^{\mu T} \mathbf{h} + \mathbf{b}_d^\mu$ 
     $\mathbf{z}^\sigma \leftarrow \mathbf{V}_d^{\sigma T} \mathbf{h} + \mathbf{b}_d^\sigma$ 
     $\alpha \leftarrow \text{softmax}(\mathbf{z}^\alpha)$ 
     $\mu = \mathbf{z}^\mu$ 
     $\sigma \leftarrow \exp(\mathbf{z}^\sigma)$ 
     $\phi \leftarrow \frac{1}{2} \frac{(\mu - \mathbf{x}_d)^2}{\sigma^2} - \log \sigma - \frac{1}{2} \log(2\pi)$ 
     $\pi \leftarrow \frac{\alpha \phi}{\sum_{j=1}^K \alpha_j \phi_j}$ 
     $\partial z^\alpha \leftarrow \pi - \alpha$ 
     $\partial \mathbf{V}_d^\alpha \leftarrow \partial z^\alpha \mathbf{h}$ 
     $\partial \mathbf{b}_d^\alpha \leftarrow \partial z^\alpha$ 
     $\partial z^\mu \leftarrow \pi (x_d - \mu) / \sigma^2$ 
     $\partial \mathbf{V}_d^\mu \leftarrow \partial z^\mu \mathbf{h}$ 
     $\partial \mathbf{b}_d^\mu \leftarrow \partial z^\mu$ 
     $\partial z^\sigma \leftarrow \pi \{ (x_d - \mu) / \sigma^2 - 1 \}$ 
     $\partial \mathbf{V}_d^\sigma \leftarrow \partial z^\sigma \mathbf{h}$ 
     $\partial \mathbf{b}_d^\sigma \leftarrow \partial z^\sigma$ 
     $\partial \mathbf{h} \leftarrow z^\alpha \mathbf{V}_d^\alpha + z^\mu \mathbf{V}_d^\mu + z^\sigma \mathbf{V}_d^\sigma$ 
     $\partial \phi = \partial \mathbf{h} \sigma(\psi) (1 - \sigma(\psi))$ 
     $\partial \rho_d \leftarrow \sum_j \partial \psi_j a_j$ 
     $\partial \mathbf{a} \leftarrow \partial \mathbf{a} + \partial \psi_\rho$ 
     $\partial \mathbf{W}_{:,d} \leftarrow \partial \mathbf{a} x_d$ 
    if  $d = 1$  then
         $\partial \mathbf{c} \leftarrow \partial \mathbf{a}$ 
    else
         $\mathbf{a} \leftarrow \mathbf{a} - x_d \mathbf{W}_{:,d}$ 
    end if
end for

```

---

---

**Algorithm 2** RNN-RNADE gradients
 

---

```

112 for  $t$  from  $T$  to 1 do
113    $\mathbf{a} \leftarrow \mathbf{c}$ 
114   for  $d$  from 1 to  $D$  do
115      $\mathbf{a} \leftarrow \mathbf{a} + x_d \mathbf{W}_{:,d}$ 
116   end for
117   for  $d$  from  $D$  to 1 do
118      $\psi_t \leftarrow \rho_d \mathbf{a}$ 
119      $\mathbf{h}_t \leftarrow \sigma(\psi_t)$ 
120      $\mathbf{z}_t^\alpha \leftarrow \mathbf{V}_d^{\alpha T} \mathbf{h}_t + \mathbf{b}_{d(t)}^\alpha$ 
121      $\mathbf{z}_t^\mu \leftarrow \mathbf{V}_d^{\mu T} \mathbf{h}_t + \mathbf{b}_{d(t)}^\mu$ 
122      $\mathbf{z}_t^\sigma \leftarrow \mathbf{V}_d^{\sigma T} \mathbf{h}_t + \mathbf{b}_{d(t)}^\sigma$ 
123      $\boldsymbol{\alpha}_t \leftarrow \text{softmax}(\mathbf{z}_t^\alpha)$ 
124      $\boldsymbol{\mu}_t = \mathbf{z}_t^\mu$ 
125      $\boldsymbol{\sigma}_t \leftarrow \exp(\mathbf{z}_t^\sigma)$ 
126      $\phi_t \leftarrow \frac{1}{2} \frac{(\boldsymbol{\mu}_t - \mathbf{x}_d^t)^2}{\boldsymbol{\sigma}_t^2} - \log \boldsymbol{\sigma}_t - \frac{1}{2} \log(2\pi)$ 
127      $\boldsymbol{\pi}_t \leftarrow \frac{\boldsymbol{\alpha}_t \phi_t}{\sum_{j=1}^K \alpha_j \phi_j}$ 
128      $\partial z_t^\alpha \leftarrow \boldsymbol{\pi}_t - \boldsymbol{\alpha}_t$ 
129      $\partial \mathbf{V}_d^\alpha \leftarrow \partial z_t^\alpha \mathbf{h}_t$ 
130      $\partial \mathbf{b}_{d(t)}^\alpha \leftarrow \partial z_t^\alpha$ 
131      $\partial z_t^\mu \leftarrow \boldsymbol{\pi}_t (x_d - \boldsymbol{\mu}_t) / \boldsymbol{\sigma}_t^2$ 
132      $\partial \mathbf{V}_d^\mu \leftarrow \partial z_t^\mu \mathbf{h}_t$ 
133      $\partial \mathbf{b}_{d(t)}^\mu \leftarrow \partial z_t^\mu$ 
134      $\partial z_t^\sigma \leftarrow \boldsymbol{\pi}_t \{ (x_d - \boldsymbol{\mu}_t) / \boldsymbol{\sigma}_t^2 - 1 \}$ 
135      $\partial \mathbf{V}_d^\sigma \leftarrow \partial z_t^\sigma \mathbf{h}_t$ 
136      $\partial \mathbf{b}_{d(t)}^\sigma \leftarrow \partial z_t^\sigma$ 
137      $\partial \mathbf{h} \leftarrow z_t^\alpha \mathbf{V}_d^\alpha + z_t^\mu \mathbf{V}_d^\mu + z_t^\sigma \mathbf{V}_d^\sigma$ 
138      $\partial \phi_t = \partial \mathbf{h}_t \sigma(\psi_t) (1 - \sigma(\psi_t))$ 
139      $\partial \rho_d(t) \leftarrow \sum_j \partial \psi_j a_j$ 
140      $\partial \mathbf{a} \leftarrow \partial \mathbf{a} + \partial \psi_\rho$ 
141      $\partial \mathbf{W}_{:,d} \leftarrow \partial \mathbf{a} x_d$ 
142     if  $d = 1$  then
143        $\partial \mathbf{c} \leftarrow \partial \mathbf{a}$ 
144     else
145        $\mathbf{a} \leftarrow \mathbf{a} - x_d^t \mathbf{W}_{:,d}$ 
146     end if
147   end for
148    $\partial W_\alpha \leftarrow \partial W_\alpha + \partial \mathbf{b}_t^\alpha \mathbf{h}^{t-1 T}$ 
149    $\partial W_\mu \leftarrow \partial W_\mu + \partial \mathbf{b}_t^\mu \mathbf{h}^{t-1 T}$ 
150    $\partial W_\sigma \leftarrow \partial W_\sigma + \partial \mathbf{b}_t^\sigma \mathbf{h}^{t-1 T}$ 
151   if  $t = T$  then
152      $\partial \mathbf{h}^t \leftarrow W_\alpha \partial \mathbf{b}_{t+1}^\alpha + W_\mu \partial \mathbf{b}_{t+1}^\mu + W_\sigma \partial \mathbf{b}_{t+1}^\sigma$ 
153   else
154      $\partial \mathbf{h}^t \leftarrow W_{rec} \partial \mathbf{h}^{t+1} \mathbf{h}^{t+1} (1 - \mathbf{h}^{t+1}) + W_\alpha \partial \mathbf{b}_{t+1}^\alpha + W_\mu \partial \mathbf{b}_{t+1}^\mu + W_\sigma \partial \mathbf{b}_{t+1}^\sigma$ 
155   end if
156    $\partial \mathbf{b}_h \leftarrow \partial \mathbf{b}_h + \partial \mathbf{h}^t \mathbf{h}^t (1 - \mathbf{h}^t)$ 
157    $\partial W_{rec} \leftarrow \partial W_{rec} + \partial \mathbf{h}^t \mathbf{h}^t (1 - \mathbf{h}^t) \mathbf{h}^{t-1 T}$ 
158    $\partial W_{in} \leftarrow \partial W_{in} + \partial \mathbf{h}^t \mathbf{h}^t (1 - \mathbf{h}^t) \mathbf{x}_t^T$ 
159 end for

```

---

### 3 2-D stochastic trajectory supplementary material

The experiment described in Section 5.1 in the main text applied an RNN-RNADE to the problem of predicting the next point in a stochastic sequence. Enclosed with this document is the video, *trajectory.m4v*, which shows an example unseen trajectory and predictions made by the RNN-RNADE. At each time-step  $T = t$ , the RNN-RNADE produces 10 samples from the predicted distribution at  $T = t + 1$  conditioned on  $T < t + 1$ . The ground truth (actual value at  $t + 1$ ) is depicted with a circle marker while the predictions are displayed with a cross. The means of the distribution clearly follow the path of the trajectory.

### 4 Videos of bouncing balls supplementary material

The experiment described in Section 5.3 in the main text applied an RNN-RNADE to the problem of predicting the pixels in the next frame of a video depicting 3 bouncing balls. The video was randomly generated and has not been previously seen by the RNN-RNADE. At each time-step  $T = t$ , the RNN-RNADE produces 1 samples from the predicted distribution at  $T = t + 1$  conditioned on  $T < t + 1$ . Enclosed with this document is the video, *bouncing.m4v*, which shows (top row left) the ground truth (actual frame at  $t + 1$ ), (top row middle) the frame sampled by the RNN-RNADE from the predicted distribution at  $t + 1$ , (top row right) the absolute difference per pixel between the ground truth and the sampled value and (bottom) the sum of the squared errors between the ground truth and sampled pixels per frame. The video shows that the sampled frame is similar to the ground truth but there is a small amount of noise in space where there are no balls. There is also noise around the outline of the balls. Apart from the first frame, the highest error occurs where two of the balls overlap at  $t = 69$  and  $t = 110$ , implying that the dynamics of this interaction have not been entirely captured by the model. Elsewhere, the sample error is low. The mean of the sum of the squared errors per frame on this sequence is 2.08.

216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269

## References

- [1] Benigno Uria, Iain Murray, and Hugo Larochelle. Rnade: The real-valued neural autoregressive density-estimator. In *Advances in Neural Information Processing Systems 26*, pages 2175–2183. 2013.