Supplementary Material for the RNN-RNADE model

Anonymous Author(s)
Affiliation
Address
email

Abstract

This document contains some of the mathematical derivations for the RNN-RNADE model.

1 Introduction

The RNADE fprop:

$$p(x) = \prod_{d=1}^{D} p(x_d|\mathbf{x}_{<\mathbf{d}}) \text{ with } p(x_d|\mathbf{x}_{<\mathbf{d}}) = p_{\mathcal{M}(x_d|\theta_d)}$$

$$\mathbf{a}_{d+1} = \mathbf{a}_d + x_d \mathbf{W}_{\cdot,d}$$

$$\mathbf{h}_\mathbf{d} = \text{sigm}(\rho_d \mathbf{a}_\mathbf{d})$$

$$\boldsymbol{\alpha_d} = \text{softmax}(\mathbf{V}^{\boldsymbol{\alpha}}_\mathbf{d} \mathbf{h}_\mathbf{d} + \mathbf{b}^{\boldsymbol{\alpha}}_\mathbf{d})$$

$$\mu_d = \mathbf{V}^\mu_d \mathbf{h}_d + \mathbf{b}^\mu_d$$

$$\boldsymbol{\sigma_d} = \exp(\mathbf{V}^{\boldsymbol{\sigma}}_{\boldsymbol{d}} \mathbf{h}_{\boldsymbol{d}} + \mathbf{b}^{\boldsymbol{\sigma}}_{\boldsymbol{d}})$$

Algorithm 1 RNADE fprop

```
procedure FPROP
    a ← c
    p ← 1
    for d from 1 to D do
        ψ_d = ρ_d a
        h_d = sigm(ψ_d)
        z_d^α = V_d^α^T h_d + b_d^α
        z_d^μ = V_d^μ^T h_d + b_d^μ
        z_d^σ = V_d^σ^T h_d + b_d^σ
        α_d = softmax(z_d^α)
        μ_d = z_d^μ
        σ_d = exp(z_d^σ)
        p(x) = p(x) p_M(x_d; α_d; μ_d; σ_d)
        a = a + x_d W_{·,d}
```

Gradients for the RNADE:

$$\phi_i(x_d|\mathbf{x}_{<d}) = \frac{1}{\sqrt{2\pi}\boldsymbol{\sigma}_{d,i}} \exp\left\{-\frac{(x_d - \boldsymbol{\mu}_{d,i})^2}{2\boldsymbol{\sigma}_{d,i}^2}\right\}$$

Posterior/Responsibility:

$$\pi_i(x_d|\mathbf{x}_{<d}) = \frac{\boldsymbol{\alpha}_{d,i}\phi_i(x_d|\mathbf{x}_{<d})}{\sum_{j=1}^{K}\boldsymbol{\alpha}_{d,j}\phi_j(x_d|\mathbf{x}_{<d})}$$

Gradients with respect to the gaussian parameters:

$$\frac{\partial p(\mathbf{x})}{\partial \mathbf{V}_d^{\alpha}} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\alpha}} \frac{\partial \mathbf{z}_{d,i}^{\alpha}}{\partial \mathbf{V}_d^{\alpha}} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\alpha}}\mathbf{h}$$

$$\frac{\partial p(\mathbf{x})}{\partial \mathbf{b}_d^{\alpha}} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\alpha}} \frac{\partial \mathbf{z}_{d,i}^{\alpha}}{\partial \mathbf{b}_d^{\alpha}} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\alpha}}$$

$$\frac{\partial p(\mathbf{x})}{\partial \mathbf{V}_d^{\mu}} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\mu}} \frac{\partial \mathbf{z}_{d,i}^{\mu}}{\partial \mathbf{V}_d^{\mu}} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\mu}}\mathbf{h}$$

$$\frac{\partial p(\mathbf{x})}{\partial \mathbf{b}_d^{\mu}} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\mu}} \frac{\partial \mathbf{z}_{d,i}^{\mu}}{\partial \mathbf{b}_d^{\mu}} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\mu}}$$

$$\frac{\partial p(\mathbf{x})}{\partial \mathbf{V}_d^{\sigma}} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\sigma}} \frac{\partial \mathbf{z}_{d,i}^{\sigma}}{\partial \mathbf{V}_d^{\sigma}} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\sigma}}\mathbf{h}$$

$$\frac{\partial p(\mathbf{x})}{\partial \mathbf{b}_d^{\sigma}} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\sigma}} \frac{\partial \mathbf{z}_{d,i}^{\sigma}}{\partial \mathbf{b}_d^{\sigma}} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\sigma}}$$

Gradient with respect to the hidden activations:

$$\frac{\partial p(\mathbf{x})}{\partial \mathbf{h}_d} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\alpha}} \frac{\partial \mathbf{z}_{d,i}^{\alpha}}{\partial \mathbf{h}_d} + \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\mu}} \frac{\partial \mathbf{z}_{d,i}^{\mu}}{\partial \mathbf{h}_d} + \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\sigma}} \frac{\partial \mathbf{z}_{d,i}^{\sigma}}{\partial \mathbf{h}_d}$$

$$\frac{\partial p(\mathbf{x})}{\partial \mathbf{h}_d} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\alpha}}\mathbf{V}_d^{\alpha} + \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\mu}}\mathbf{V}_d^{\mu} + \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\sigma}}\mathbf{V}_d^{\sigma}$$

2 RNN-RNADE fprop

$$p(x_1^T) = \prod_1^T p(x^t|\mathcal{A}^t) \text{ where } \mathcal{A}^t \equiv \{x^\tau|\tau < t\}$$

Maximum-likelihood Training:

$$\log p(x_1^T) = \sum_1^T \log p(x^t|\mathcal{A}^t)$$
$$\log p(x_1^T) = \sum_{t=1}^T \sum_{d=1}^D \log p(x_d^t|\mathbf{x}_{<d}^t)$$

W' is the weight matrix from the reccurrent layer to the mixing coefficients.
W'' is the weight matrix from the reccurrent layer to the means.
W''' is the weight matrix from the reccurrent layer to the sigmas.
The rest of the architecture is similar to the RNN-RBM paper.

$$\hat{h}^{(t)} = \sigma(W_2\mathbf{x}^t + W_3\hat{h}^{(t-1)} + b_{\hat{h}})$$
$$r_\alpha = W^\alpha \hat{h}^{(t-1)}$$
$$r_\mu = W^\mu \hat{h}^{(t-1)}$$
$$r_\sigma = W^\sigma \hat{h}^{(t-1)}$$

Algorithm 2 RNN-RNADE fprop

procedure FPROP
- $a \leftarrow c$
- $p \leftarrow 1$
- **for** t from 1 to T **do**
  - **for** d from 1 to D **do**
    - $\psi_\mathbf{d} = \rho_\mathbf{d}\mathbf{a}$
    - $\mathbf{h_d} = \text{sigm}(\psi_\mathbf{d})$
    - $\mathbf{z}_d^\alpha = \mathbf{V_d^\alpha}^\mathbf{T}\mathbf{h_d} + \mathbf{b_d^\alpha} + \mathbf{r_d^\alpha} = \mathbf{V_d^\alpha}^\mathbf{T}\mathbf{h_d} + \mathbf{b_d^\alpha} + \mathbf{r_d^\alpha}$
    - $\mathbf{z}_d^\mu = \mathbf{V_d^\mu}^\mathbf{T}\mathbf{h_d} + \mathbf{b_d^\mu} + \mathbf{r_d^\mu}$
    - $\mathbf{z}_d^\sigma = \mathbf{V_d^\sigma}^\mathbf{T}\mathbf{h_d} + \mathbf{b_d^\sigma} + \mathbf{r_d^\sigma}$
    - $\boldsymbol{\alpha_d} = \text{softmax}(\mathbf{z}_{\boldsymbol{d}}^{\boldsymbol{\alpha}})$
    - $\mu_d = \mathbf{z}_d^\mu$
    - $\boldsymbol{\sigma_d} = \exp(\mathbf{z}_{\boldsymbol{d}}^{\boldsymbol{\sigma}})$
    - $p(\mathbf{x}) = p(\mathbf{x})p_\mathcal{M}(x_d; \alpha_d; \mu_d; \sigma_d)$
    - $\mathbf{a} = \mathbf{a} + x_d^t\mathbf{W}_{\cdot,d}$

Gradients for the RNN-RNADE:

$$\phi_i(x_d|\mathbf{x}_{<d}) = \frac{1}{\sqrt{2\pi}\boldsymbol{\sigma}_{d,i}} \exp\left\{-\frac{(x_d - \boldsymbol{\mu}_{d,i})^2}{2\boldsymbol{\sigma}_{d,i}^2}\right\}$$

Posterior/Responsibility:

$$\pi_i(x_d|\mathbf{x}_{<d}) = \frac{\boldsymbol{\alpha}_{d,i}\phi_i(x_d|\mathbf{x}_{<d})}{\sum_{j=1}^K \boldsymbol{\alpha}_{d,j}\phi_j(x_d|\mathbf{x}_{<d})}$$

Gradients with respect to the gaussian parameters:

$$\frac{\partial p(\mathbf{x})}{\partial \mathbf{V}_d^\alpha} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^\alpha}\frac{\partial \mathbf{z}_{d,i}^\alpha}{\partial \mathbf{V}_d^\alpha} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^\alpha}\mathbf{h}$$

$$\frac{\partial p(\mathbf{x})}{\partial \mathbf{b}_d^{\alpha}} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\alpha}} \frac{\partial \mathbf{z}_{d,i}^{\alpha}}{\partial \mathbf{b}_d^{\alpha}} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\alpha}}$$

$$\frac{\partial p(\mathbf{x})}{\partial \mathbf{V}_d^{\mu}} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\mu}} \frac{\partial \mathbf{z}_{d,i}^{\mu}}{\partial \mathbf{V}_d^{\mu}} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\mu}} \mathbf{h}$$

$$\frac{\partial p(\mathbf{x})}{\partial \mathbf{b}_d^{\mu}} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\mu}} \frac{\partial \mathbf{z}_{d,i}^{\mu}}{\partial \mathbf{b}_d^{\mu}} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\mu}}$$

$$\frac{\partial p(\mathbf{x})}{\partial \mathbf{V}_d^{\sigma}} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\sigma}} \frac{\partial \mathbf{z}_{d,i}^{\sigma}}{\partial \mathbf{V}_d^{\sigma}} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\sigma}} \mathbf{h}$$

$$\frac{\partial p(\mathbf{x})}{\partial \mathbf{b}_d^{\sigma}} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\sigma}} \frac{\partial \mathbf{z}_{d,i}^{\sigma}}{\partial \mathbf{b}_d^{\sigma}} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\sigma}}$$

Gradient with respect to the hidden activations:

$$\frac{\partial p(\mathbf{x})}{\partial \mathbf{h}_d} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\alpha}} \frac{\partial \mathbf{z}_{d,i}^{\alpha}}{\partial \mathbf{h}_d} + \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\mu}} \frac{\partial \mathbf{z}_{d,i}^{\mu}}{\partial \mathbf{h}_d} + \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\sigma}} \frac{\partial \mathbf{z}_{d,i}^{\sigma}}{\partial \mathbf{h}_d}$$

$$\frac{\partial p(\mathbf{x})}{\partial \mathbf{h}_d} = \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\alpha}} \mathbf{V}_d^{\alpha} + \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\mu}} \mathbf{V}_d^{\mu} + \frac{\partial p(\mathbf{x})}{\partial \mathbf{z}_{d,i}^{\sigma}} \mathbf{V}_d^{\sigma}$$