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

Anonymous Author(s)
Affiliation
Address
email

1 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.

Algorithm 1 RNADE gradients

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

Algorithm 2 RNN-RNADE gradients

for t from T to 1 **do**
    $\mathbf{a} \leftarrow \mathbf{c}$
    for d from 1 to D **do**
        $\mathbf{a} \leftarrow \mathbf{a} + x_d \mathbf{W}_{.,d}$
    end for
    for d from D to 1 **do**
        $\boldsymbol{\psi}_t \leftarrow \rho_d \mathbf{a}$
        $\mathbf{h}_t \leftarrow \sigma(\boldsymbol{\psi}_t)$
        $\mathbf{z}_t^{\alpha} \leftarrow \mathbf{V}_d^{\alpha T} \mathbf{h}_t + \mathbf{b}_{d(t)}^{\alpha}$
        $\mathbf{z}_t^{\mu} \leftarrow \mathbf{V}_d^{\mu T} \mathbf{h}_t + \mathbf{b}_{d(t)}^{\mu}$
        $\mathbf{z}_t^{\sigma} \leftarrow \mathbf{V}_d^{\sigma T} \mathbf{h}_t + \mathbf{b}_{d(t)}^{\sigma}$
        $\boldsymbol{\alpha}_t \leftarrow \mathrm{softmax}(\mathbf{z}_t^{\alpha})$
        $\boldsymbol{\mu}_t = \mathbf{z}_t^{\mu}$
        $\boldsymbol{\sigma}_t \leftarrow \exp(\mathbf{z}_t^{\sigma})$
        $\boldsymbol{\phi}_t \leftarrow \frac{1}{2} \frac{(\boldsymbol{\mu}_t - \mathbf{x}_d^t)^2}{\boldsymbol{\sigma}^2} - \log \boldsymbol{\sigma}_t - \frac{1}{2} \log(2\pi)$
        $\boldsymbol{\pi}_t \leftarrow \frac{\boldsymbol{\alpha}_t \phi_t}{\sum_{j=1}^{K} \alpha_j \phi_j}$
        $\partial z_t^{\alpha} \leftarrow \boldsymbol{\pi}_t - \boldsymbol{\alpha}_t$
        $\partial \boldsymbol{V}_d^{\alpha} \leftarrow \partial z_t^{\alpha} \mathbf{h}_t$
        $\partial \mathbf{b}_{d(t)}^{\alpha} \leftarrow \partial z_t^{\alpha}$
        $\partial z_t^{\mu} \leftarrow \boldsymbol{\pi_t}(x_d - \boldsymbol{\mu}_t)/\boldsymbol{\sigma}_t^2$
        $\partial \boldsymbol{V}_d^{\mu} \leftarrow \partial z_t^{\mu} \mathbf{h}_t$
        $\partial \mathbf{b}_{d(t)}^{\mu} \leftarrow \partial z_t^{\mu}$
        $\partial z_t^{\sigma} \leftarrow \boldsymbol{\pi}_t \left\{ (x_d - \boldsymbol{\mu}_t)/\boldsymbol{\sigma}_t^2 - 1 \right\}$
        $\partial \boldsymbol{V}_d^{\sigma} \leftarrow \partial z_t^{\sigma} \mathbf{h}_t$
        $\partial \mathbf{b}_{d(t)}^{\sigma} \leftarrow \partial z_t^{\sigma}$
        $\partial \mathbf{h} \leftarrow z_t^{\alpha} \boldsymbol{V}_d^{\alpha} + z_t^{\mu} \boldsymbol{V}_d^{\mu} + z_t^{\sigma} \boldsymbol{V}_d^{\sigma}$
        $\partial \boldsymbol{\phi}_t = \partial \mathbf{h}_t \sigma(\boldsymbol{\psi}_t)(1 - \sigma(\boldsymbol{\psi}_t))$
        $\partial \rho_d(t) \leftarrow \sum_j \partial \boldsymbol{\psi}_j a_j$
        $\partial \mathbf{a} \leftarrow \partial \mathbf{a} + \partial \boldsymbol{\psi}_{\rho}$
        $\partial \mathbf{W}_{.,d} \leftarrow \partial \mathbf{a} x_d$
        if $d = 1$ **then**
            $\partial \mathbf{c} \leftarrow \partial \mathbf{a}$
        else
            $\mathbf{a} \leftarrow \mathbf{a} - x_d^t \mathbf{W}_{.,d}$
        end if
    end for
    $\partial W_{\alpha} \leftarrow \partial W_{\alpha} + \partial \mathbf{b}_t^{\alpha} \mathbf{h}^{t-1^T}$
    $\partial W_{\mu} \leftarrow \partial W_{\mu} + \partial \mathbf{b}_t^{\mu} \mathbf{h}^{t-1^T}$
    $\partial W_{\sigma} \leftarrow \partial W_{\sigma} + \partial \mathbf{b}_t^{\sigma} \mathbf{h}^{t-1^T}$
    if $t = T$ **then**
        $\partial \mathbf{h}^t \leftarrow W_{\alpha} \partial \mathbf{b}_{t+1}^{\alpha} + W_{\mu} \partial \mathbf{b}_{t+1}^{\mu} + W_{\sigma} \partial \mathbf{b}_{t+1}^{\sigma}$
    else
        $\partial \mathbf{h}^t \leftarrow W_{rec} \partial \mathbf{h}^{t+1} \mathbf{h}^{t+1} (1 - \mathbf{h}^{t+1}) + W_{\alpha} \partial \mathbf{b}_{t+1}^{\alpha} + W_{\mu} \partial \mathbf{b}_{t+1}^{\mu} + W_{\sigma} \partial \mathbf{b}_{t+1}^{\sigma}$
    end if
    $\partial \mathbf{b}_h \leftarrow \partial \mathbf{b}_h + \partial \mathbf{h}^t \mathbf{h}^t (1 - \mathbf{h}^t)$
    $\partial W_{rec} \leftarrow \partial W_{rec} + \partial \mathbf{h}^t \mathbf{h}^t (1 - \mathbf{h}^t) \mathbf{h}^{t-1^T}$
    $\partial W_{in} \leftarrow \partial W_{in} + \partial \mathbf{h}^t \mathbf{h}^t (1 - \mathbf{h}^t) \mathbf{x}_t^T$
end for

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.