Notes on antisymmetric NeuralODEs

fnestaas

April 2022

1 Introduction

We introduce the general dynamics

$$\dot{x} = f_\theta(x) \quad (1)$$
$$\dot{\theta} = g(\theta). \quad (2)$$

By introducing different restrictions, we will see that this system can describe a wide class of stable residual recurrent neural network architectures. We start by enforcing that x is multiplied by an anti-symmetric matrix.

$$\dot{x} = f((A(t) - A(t)^T)x + b(t)) \quad (3)$$
$$\dot{A}(t) = g(A(t)) \quad (4)$$
$$\dot{b}(t) = h(b(t)) \quad (5)$$

Where $A \in \mathbb{R}^{d\times d}$ and $b \in \mathbb{R}^d$. Letting f act component wise as $f : \mathbb{R} \to \mathbb{R}, y \mapsto \sigma(y)$, we get a system studied in detail in [1];

$$\dot{x} = \sigma((A(t) - A(t)^T)x + b(t)) \quad (6)$$

where the authors do not comment too explicitly on the dynamics of A, but test the method for a constant matrix A, obtaining good results. Perhaps a simpler approach is to let f still act componentwise, but be the identity function instead. Additionally, we set $b = 0$;

$$\dot{x} = (A(t) - A(t)^T)x \quad (7)$$
$$\dot{A} = g(A). \quad (8)$$

This system has the interesting property that the norm of x is constant;

$$\frac{1}{2}\dot{\|x\|^2} = x^T\dot{x} = x^TAx = x^TA^Tx = -x^TAx \tag{9}$$
$$\Rightarrow x^TAx = 0. \tag{10}$$

While this property might seem desirable, it is so restrictive that the system becomes linear in its initial state. The reason for this is precisely that the norm of x remains constant, i.e.$\forall t \geq 0, \|x(t)\| = \|x(0)\|$. Then for any fixed t there exists an orthogonal matrix $Q_t \in R^{d\times d}$ such that $x(t) = Q_tx(0)$. In fact we readily find the dynamics that such a matrix exhibits;

$$\dot{x} = \dot{Q}_tx(0) = (A(t) - A(t)^T)Q_tx(0) \tag{11}$$
$$\dot{A} = g(A) \tag{12}$$
$$Q_0 := I_d \tag{13}$$

where I_d is the d-dimensional identity matrix. The dynamics above shows that for any system of the form 8 can be expressed as

$$\dot{Q}_t = (A(t) - A(t)^T)Q_t \tag{14}$$
$$\dot{A} = g(A) \tag{15}$$
$$Q_0 := I_d \tag{16}$$
$$x(t) = Q_tx(0) \tag{17}$$

which is unfortunately not terribly interesting due to the linear behavior in 17. This is problematic e.g. if we would like to classify images; the sum of a picture of a 0 and a 7 is not neccesarily a number.

Actually, any system of the form

$$\dot{x} = \alpha Ax \tag{18}$$
$$\dot{A} = g(A) \tag{19}$$

TODO: what is the effect of alpha? What is the effect of the initial state?

2 Ode to ODE

$$\dot{x} = Wx \tag{20}$$
$$\dot{W} = Wb_\theta(W, t) \tag{21}$$
$$W^TW = I \tag{22}$$

Then $\|x\|^2 = \|\dot{x}\|^2$, so if ever $x(t^*) = 0$, then $\forall t \geq t^*, x(t) = 0$.

3 My regularization idea

Adopting the notation from NeuralODE ([2]), we have $a \in \mathbb{R}^{1\times n}, f_\theta \in \mathbb{R}^n, z \in \mathbb{R}^n, \theta \in \mathbb{R}^d$, we could enforce $y_\theta := \left[a(t)\frac{df_\theta}{d\theta}\right]^T$ to have a constant magnitude, i.e. $y_\theta(z(t),t) = W_t y_\theta(z_{t_1}, t_1)$. Thus, y is linear in $y(t_1)$, but this is not problematic since it is always the case for neural ODEs;

$$\frac{dL}{d\theta} = -\int_{t_1}^{t_0} a(t)\frac{df_\theta(z(t),t)}{d\theta}\mathrm{d}t \tag{23}$$

$$= -\frac{dL}{dz(t_1)}\int_{t_1}^{t_0} \frac{dz(t_1)}{dz(t)}\frac{df_\theta(z(t),t)}{d\theta}\mathrm{d}t \tag{24}$$

We obtain

$$\frac{d}{dt}\left[a(t)\frac{df_\theta(z(t),t)}{d\theta}\right] = \dot{a}(t)\frac{df_\theta(z(t),t)}{d\theta} + a(t)\frac{d}{dt}\frac{df_\theta(z(t),t)}{d\theta} \tag{25}$$

$$= -a(t)\frac{d}{dz}f_\theta(z(t),t)\frac{df_\theta(z(t),t)}{d\theta} + a(t)\frac{d}{dt}\frac{df_\theta(z(t),t)}{d\theta} \tag{26}$$

$$= a(t)\left(\frac{d}{dt}\frac{df_\theta(z(t),t)}{d\theta} - \frac{d}{dz}f_\theta(z(t),t)\frac{df_\theta(z(t),t)}{d\theta}\right) \tag{27}$$

$$= a(t)\frac{df_\theta(z(t),t)}{d\theta}(A - A^T) \tag{28}$$

where the last equality follows from the fact that if $\dot{q} = qS$ where $q \in \mathbb{R}^{1\times n}$ and S is anti-symmetric in $\mathbb{R}^{n\times n}$, then the norm is constant in time. As such, the equation allows us to look for dynamics where $\|y(t)\|$ is constant, or to check if certain dynamics satisfy the equation. One approach is to solve

$$\frac{d}{dt}\frac{df_\theta(z(t),t)}{d\theta} - \frac{d}{dz}f_\theta(z(t),t)\frac{df_\theta(z(t),t)}{d\theta} = \frac{df_\theta(z(t),t)}{d\theta}(A - A^T) \tag{29}$$

where A is a matrix that can depend on θ, t and z. In fact, if W_t obeys to the dynamics $\dot{W}_t = (A^T - A)W_t$, then $\dot{y(t)}^T = y(t)^T(A - A^T)$.

This approach satisfies that the gradients do not explode since

$$\left\|\frac{dL}{d\theta}\right\| \leq \int_{t_0}^{t_1} \|y(t)\|\,\mathrm{d}t = (t_1 - t_0)\,\|y_{t_1}\|\,. \tag{30}$$

However, they might still vanish, e.g. if W_t is any matrix satisfying $W_{t+T/2} = -W_t, \forall t \in [0, t_0 + T/2]$ and $T = t_1 - t_0$. Then the integral evaluates to 0. We can try to remedy this, e.g. by putting some further constraints on y. One such constraint could be to enforce that the inner product with a certain direction is always positive. E.g. using $y(t^*)$ and some $\gamma \in [-1, 1]$;

$$y(t)^T y_{t^*} \geq \gamma \left\|y(t_1)\right\|^2, \forall t \tag{31}$$

$$\Rightarrow \frac{dL}{d\theta} y_{t^*} = -\int_{t_1}^{t_0} y(t)^T y_{t^*} \mathrm{d}t \geq \gamma \left\|y_T\right\|^2 (t_1 - t_0). \tag{32}$$

While this might seem somewhat arbitrary, intuitively, we can understand it as restricting the general direction of the loss not to change drastically in the network. In this case, let $Q := W_t^T W_{t^*}$. Q is orthogonal and

$$y(t)^T y(t^*) = y(t_1)^T Q y(t_1) = \sum_i \lambda_i |k_i^T y(t_1)|^2 \tag{33}$$

since $Q = K \Lambda K^T$ where $\Lambda_{i,j} = 1(i = j)\lambda_i$, λ_i are eigenvalues of Q and $K = (k_1, \ldots, k_n)$ are eigenvectors such that $Qk_i = \lambda_i k_i$. Then $\{k_i\}_{i=1}^n$ forms an orthonormal basis and we have $\left\|y(t_1)\right\|^2 = \sum_i |k_i^T y(t_1)|^2$. According to Wikipedia, normal matrices (B s.t. $B^* B = BB^*$, where $(.)^*$ is the conjugate transpose) are never defective, and hence also not orthogonal matrices. `https://en.wikipedia.org/wiki/Defective_matrix`. Then we can write

$$\sum_i \lambda_i |k_i^T y(t_1)|^2 = \sum_i \mathrm{Re}(\lambda_i) |k_i^T y(t_1)|^2 \geq \min_i \mathrm{Re}(\lambda_i) \left\|y(t_1)\right\|^2 \tag{34}$$

and e.g. enforce that $\forall i, \mathrm{Re}(\lambda_i) \geq \gamma$.

If we enforce $\forall i, \lambda_i = 1$, then we always satisfy 31, and we can parameterize W using the Cayley transform (see [4]). But then $Q = I$ (see diagonalization).

Another approach is to bound $\|a(t)\| = \left\|\frac{dL}{dz(t)}\right\|$, as is done in numerous papers (e.g. [1] [3]). We could again root for an approach where the norm remains constant, and obtain

$$a(t) = a(t_1) W_t \tag{35}$$

where, again, W_t is orthogonal.

References

[1] Bo Chang, Minmin Chen, Eldad Haber, and Ed H. Chi. Antisymmetricrnn: A dynamical system view on recurrent neural networks, 2019. URL: `https://arxiv.org/abs/1902.09689`, `doi:10.48550/ARXIV.1902.09689`.

[2] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations, 2018. URL: `https://arxiv.org/abs/1806.07366`, `doi:10.48550/ARXIV.1806.07366`.

[3] Krzysztof Choromanski, Jared Quincy Davis, Valerii Likhosherstov, Xingyou Song, Jean-Jacques Slotine, Jacob Varley, Honglak Lee, Adrian Weller, and Vikas Sindhwani. An ode to an ode, 2020. URL: `https://arxiv.org/abs/2006.11421`, `doi:10.48550/ARXIV.2006.11421`.

[4] Kyle Helfrich, Devin Willmott, and Qiang Ye. Orthogonal recurrent neural networks with scaled cayley transform, 2017. URL: https://arxiv.org/abs/1707.09520, doi:10.48550/ARXIV.1707.09520.