A. Sparse prior for recurrent layer: a control system perspective

To better understand our compression strategy for RNN, we stack all the weights together in a control systematic formulation [1]. Take an RNN with double hidden layers for example as shown in Fig.2. Our goal is to reduce the order of the system and the number of inputs, i.e., frozen some x and u. It should be noted that this is the same as goal of model reduction in control theory [1]. The yellow, purple and blue strips in Fig.2(b) illustrate the "reduction" idea. Similarly, it is relatively straightforward to apply the same idea to a single layer RNN as shown in Fig. 1.

B. Compute the Hessian of recurrent layer

As a RNN cell could be extended to fully connected layers along time sequence, the Hessian calculation at a certain time step can refer to the approach for fully-connected layer whose Hessian calculation methods has been expalined in Appendix.C.1 [2]. Therefore, for the Hessian calculation in a recurrent cell, the left main problem is how to conduct sequential Hessian calculations when considering the influence of time sequence and the backward propagation through time (BPTT) process. We denote W_i, W_x and W_o as the input weight, hidden weight and output weight of a RNN cell respectively. The sequence length for RNN cell is assumed to be T and τ is the backward propagation time horizon. u^t and x^t stands for the input vector and hidden vector. The Hessian for W_o could be computed as:

$$\mathbf{H}_o = \frac{1}{T}\sum_{t=1}^{T}\mathbf{H}_o^t \tag{1}$$

where $\mathbf{H}_o^t = (x^t)^2 \otimes H_o^t$ stands for the Hessian and H_o^t is the pre-activation Hessian of W_o at step t. When computing the Hessian for W_i and W_x, the BPTT process should also be evaluated besides the time sequence:

$$\mathbf{H}_x = \frac{1}{T\times\tau}\sum_{t=1}^{T}\sum_{bptt=T}^{T-\tau}\mathbf{H}_x^{t,bptt}, \quad \mathbf{H}_x^{t,bptt} = x_{t-1}^2 \otimes H_x^{t,bptt} \tag{2}$$

$$H_x^{t,bptt} = B^2 \circ (W_x^2 H_x^{t,\text{bptt+1}}) + D, \quad B = \sigma'(h^t), \quad D = \sigma''(h^t)\circ\frac{\partial L}{\partial x^t} \tag{3}$$

where *bptt* is the index for BPTT process and $H_x^{t,\text{bptt+1}}$ is supposed to be known before calculating $\mathbf{H}_x$. Similar to the calculation for output Hessian. The input Hessian could be calculated as:

$$\mathbf{H}_{in} = \frac{1}{T\times\tau}\sum_{t=1}^{T}\sum_{bptt=T}^{T-\tau}\mathbf{H}_{in}^{t,bptt} \tag{4}$$

where $\mathbf{H}_{in}^{t,bptt} = (u^{t-1})^2 \otimes H_{in}^{t,bptt}$ and $H_{in}^{t,bptt}$ stands for the pre-activation Hessian which can be computed as Eq.C.1.3 or Eq.C.1.5..

C. Control systematic representation for RNN with two hidden layers

$$\underbrace{\begin{bmatrix} x_t^1 \\ x_{t+1}^1 \\ x_t^2 \end{bmatrix}}_{\bar{x}_{t+1}} = \sigma_x\left(\underbrace{\begin{bmatrix} W_x^1 & 0 & 0 \\ 0 & W_x^1 & 0 \\ 0 & W_i^2 & W_x^2 \end{bmatrix}}_{A}\underbrace{\begin{bmatrix} x_{t-1}^1 \\ x_t^1 \\ x_{t-1}^2 \end{bmatrix}}_{\bar{x}_t} + \underbrace{\begin{bmatrix} W_i^1 & 0 \\ 0 & W_i^1 \\ 0 & 0 \end{bmatrix}}_{B}\underbrace{\begin{bmatrix} u_t \\ u_{t+1} \end{bmatrix}}_{\bar{u}_t}\right)$$

$$\bar{y}_t \triangleq y_{t-1} = \sigma_y\left(\underbrace{\begin{bmatrix} 0 & 0 & W_o \end{bmatrix}}_{C}\underbrace{\begin{bmatrix} x_{t-1}^1 \\ x_t^1 \\ x_{t-1}^2 \end{bmatrix}}_{\bar{x}_t}\right)$$

D. Experiment

We simulate a sparse RNN with Gaussian ($\mathcal{N}(0,1)$) weights (see upper left of Fig. 1(c) and Fig. 2(c)) and inputs to generate train and test dataset. We compare several compression strategies and find our method is likely to find the true network structure (see lower right of Fig. 1(c) and Fig. 2(c)).

REFERENCES

[1] Zhou, Kemin, John Comstock Doyle, and Keith Glover. "Robust and optimal control." Vol. 40. New Jersey: Prentice hall, 1996.

[2] Botev, Aleksandar, Hippolyt Ritter, and David Barber. "Practical gauss-newton optimisation for deep learning." Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017.

[3] Sepp Hochreiter. "The vanishing gradient problem during learning recurrent neural nets and problem solutions". Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 6, 2 (April 1998), 107-116. DOI=http://dx.doi.org/10.1142/S0218488598000094

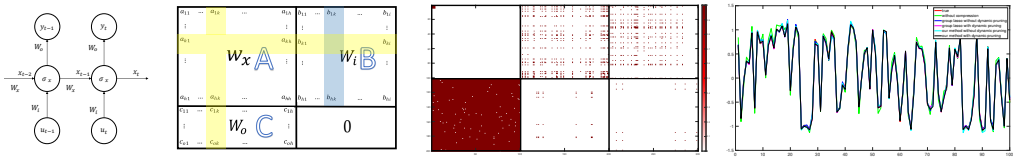

Fig. 1. RNN with single hidden layer. (a) Network Structure with single hidden layers formulated as $x_t = \sigma_x(W_x x_{t-1} + W_i u_t + b_h), y_t = \sigma_y(W_o x_t + b_y)$. (b) Control systematic representation of weight matrices formulated as $x_t = \sigma_x(Ax_{t-1} + Bu_t), y_t = \sigma_y(Cx_t)$. Yellow and blue strips marks the reduction for state x and input u respectively. (c) Estimated sparse structure of W_x using several compression strategies: ground truth (upper left); no compression (lower left); group Lasso without dynamic pruning (upper middle); group Lasso with dynamic pruning (lower middle); our method without dynamic pruning after 10 loops (upper right); our method with dynamic pruning after 10 loops (lower right). (d) Test using the weights in (c) with Gaussian input.

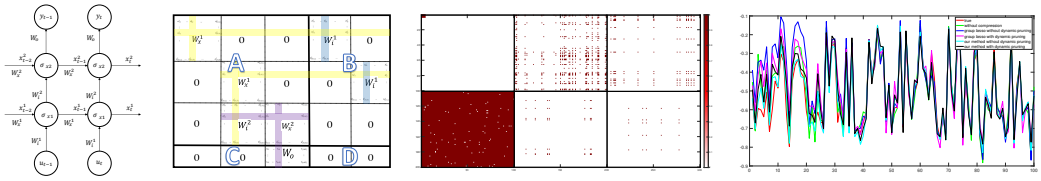

Fig. 2. RNN with double hidden layers. (a) Network structure formulated as $x_t^1 = \sigma_x^1(W_x^1 x_{t-1}^1 + W_i^1 u_t + b_x^1), x_t^2 = \sigma_x^2(W_x^2 x_{t-1}^2 + W_i^2 x_t^1 + b_x^2), y_t = \sigma_y(W_o x_t^2 + b_y)$. (b) Control systematic representation of weight matrices formulated in Section -C by introducing new augmented state variable $\bar{x}$ and input variable $\bar{u}$. Yellow, purple and blue strips marks the reduction for state x^1, state x^2 and input u respectively. (c) Estimated sparse structure of W_x^1 using several compression strategies: ground truth (upper left); no compression (lower left); group Lasso without dynamic pruning (upper middle); group Lasso with dynamic pruning (lower middle); our method without dynamic pruning after 10 loops (upper right); our method with dynamic pruning after 10 loops (lower right). (d) Test using the weights in (c) with Gaussian input.

TABLE I

HYPER-PARAMETER UPDATE RULE FOR RNN WITH DOUBLE HIDDEN LAYERS

Category	Sparse prior	$R(\omega \circ \boldsymbol{W})$	ω	γ
$\boldsymbol{W}^1$ (Yellow)	$\prod_{n=1}^{N^1} \mathcal{N}(\mathbf{0}, \gamma_o^1 \mathbf{I}_{n,:})$	$\boldsymbol{W}^1 = [2\boldsymbol{W}_x^1, 2(\boldsymbol{W}_x^1)^\top, 2\boldsymbol{W}_i^1, (\boldsymbol{W}_i^2)^\top]$ $R_1 = \sum_{n=1}^{N^1} \lVert \omega_{n,:}^l \circ \boldsymbol{W}_{n,:}^l \rVert_2$	$\omega_o^1 = \sqrt{\sum_n \lvert \alpha_{n,:}^1 \rvert}$ $\omega_{n,:}^1 = \omega_o^1 \cdot \boldsymbol{I}_{n,:}^1$	$\gamma_o^1 = \frac{\lVert \boldsymbol{W}_{n,:}^1 \rVert_2}{(\omega_{n,:}^1)^{k-1}}$ $\gamma_{n,:}^1 = \gamma_o^1 \cdot \boldsymbol{I}_{n,:}^1$
$\boldsymbol{W}^2$ (Purple)	$\prod_{n=1}^{N^2} \mathcal{N}(\mathbf{0}, \gamma_o^2 \mathbf{I}_{n,:})$	$\boldsymbol{W}^2 = [\boldsymbol{W}_x^2, (\boldsymbol{W}_x^2)^\top, \boldsymbol{W}_i^2, \boldsymbol{W}_o^\top]$ $R_2 = \sum_{n=1}^{N^2} \lVert \omega_{n,:}^2 \circ \boldsymbol{W}_{n,:}^2 \rVert_2$	$\omega_o^2 = \sqrt{\sum_n \lvert \alpha_{n,:}^2 \rvert}$ $\omega_{n,:}^2 = \omega_o^2 \cdot \boldsymbol{I}_{n,:}^2$	$\gamma_o^2 = \frac{\lVert \boldsymbol{W}_{n,:}^2 \rVert_2}{(\omega_{n,:}^2)^{k-1}}$ $\gamma_{n,:}^2 = \gamma_o^2 \cdot \boldsymbol{I}_{n,:}^2$
$\boldsymbol{W}^3$ (Blue)	$\prod_{n=1}^{N^3} \mathcal{N}(\mathbf{0}, \gamma_o^3 \mathbf{I}_{n,:})$	$\boldsymbol{W}^3 = (\boldsymbol{W}_i^1)\top$ $R_3 = \sum_{n=1}^{N^3} \lVert \omega_{n,:}^3 \circ \boldsymbol{W}_{n,:}^3 \rVert_2$	$\omega_o^3 = \sqrt{\sum_n \lvert \alpha_{n,:}^3 \rvert}$ $\omega_{n,:}^3 = \omega_o^3 \cdot \boldsymbol{I}_{n,:}^3$	$\gamma_o^3 = \frac{\lVert \boldsymbol{W}_{n,:}^3 \rVert_2}{(\omega_{n,:}^3)^{k-1}}$ $\gamma_{n,:}^3 = \gamma_o^3 \cdot \boldsymbol{I}_{n,:}^3$

TABLE II

HYPER-PARAMETER UPDATE RULE FOR RNN WITH SINGLE HIDDEN LAYER

Category	Sparse prior	$R(\omega \circ \boldsymbol{W})$	ω	γ
$\boldsymbol{W}^1$ (Yellow)	$\prod_{n=1}^{N^1} \mathcal{N}(\mathbf{0}, \gamma_o^1 \mathbf{I}_{n,:})$	$\boldsymbol{W}^1 = [\boldsymbol{W}_h, \boldsymbol{W}_h^\top, \boldsymbol{W}_i, \boldsymbol{W}_o^T]$ $\sum_{n=1}^{N^1} \lVert \omega_{n,:}^l \circ \boldsymbol{W}_{n,:}^l \rVert_2$	$\omega_o^1 = \sqrt{\sum_n \lvert \alpha_{n,:}^1 \rvert}$ $\omega_{n,:}^1 = \omega_o^1 \cdot \boldsymbol{I}_{n,:}^1$	$\gamma_o^1 = \frac{\lVert \boldsymbol{W}_{n,:}^1 \rVert_2}{(\omega_{n,:}^1)^{k-1}}$ $\gamma_{n,:}^1 = \gamma_o^1 \cdot \boldsymbol{I}_{n,:}^1$
$\boldsymbol{W}^2$ (Blue)	$\prod_{n=1}^{N^2} \mathcal{N}(\mathbf{0}, \gamma_o^2 \mathbf{I}_{n,:})$	$\boldsymbol{W}^2 = \boldsymbol{W}_i^T$ $\sum_{n=1}^{N^2} \lVert \omega_{n,:}^2 \circ \boldsymbol{W}_{n,:}^2 \rVert_2$	$\omega_o^2 = \sqrt{\sum_n \lvert \alpha_{n,:}^2 \rvert}$ $\omega_{n,:}^2 = \omega_o^2 \cdot \boldsymbol{I}_{n,:}^2$	$\gamma_o^2 = \frac{\lVert \boldsymbol{W}_{n,:}^2 \rVert_2}{(\omega_{n,:}^2)^{k-1}}$ $\gamma_{n,:}^2 = \gamma_o^2 \cdot \boldsymbol{I}_{n,:}^2$