On the Stochastic Stability of Deep Markov Models

Ján Drgoňa[1], **Sayak Mukherjee**[1], **Jiaxin Zhang**[2], **Frank Liu**[2], **Mahantesh Halappanavar**[1]
[1] Pacific Northwest National Laboratory
Richland, Washington, USA
[2] Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
{jan.drgona, sayak.mukherjee, mahantesh.halappanavar}@pnnl.gov
{zhangj, liufy}@ornl.gov

1 Supplementary Material

1.1 Background

In this section, we recall few important mathematical definitions and theorems used in the paper.

Definition 1. *Induced operator norm of a matrix* $\mathbf{A} \in \mathbb{R}^{n \times m}$ *is defined as:*

$$||\mathbf{A}||_p = \max_{\mathbf{x} \neq 0} \frac{||\mathbf{A}\mathbf{x}||_p}{||\mathbf{x}||_p} = \max_{||\mathbf{x}||_p = 1} ||\mathbf{A}\mathbf{x}||_p, \ \ \forall \mathbf{x} \in \mathcal{X}, \tag{1}$$

where $\mathcal{X}$ *is a compact normed vector space, and* $|| \cdot ||_p : \mathbb{R}^n \to \mathbb{R}$ *represents vector* p*-norm inducing the matrix norm* $||\mathbf{A}||_p : \mathbb{R}^{n \times m} \to \mathbb{R}$*. In case of Euclidean norm* $|| \cdot ||_2$*, the operator norm corresponds to the largest singular value* $\sigma_{max}(\mathbf{A})$*:*

$$||\mathbf{A}||_2 = \max_{||\mathbf{x}||_2 = 1} ||\mathbf{A}\mathbf{x}||_2 = \sigma_{max}(\mathbf{A}). \tag{2}$$

The matrix norm is sub-additive:

$$||\mathbf{A} + \mathbf{B}||_p \leq ||\mathbf{A}||_p + ||\mathbf{B}||_p. \tag{3}$$

Induced p*-norm* $|| \cdot ||_p : \mathbb{R}^{n \times m} \to \mathbb{R}$ *is called submultiplicative if it satisfies [Malek-Shahmirzadi, 1983]:*

$$||\mathbf{A}\mathbf{B}||_p \leq ||\mathbf{A}||_p ||\mathbf{B}||_p. \tag{4}$$

Remark 2. *For commonly used norms* $|| \cdot ||_p$ *where* $p \in \{1, 2, \infty\}$ *the submultiplicativity* (4) *holds.*

Definition 3. *Asymptotic stability of the dynamical system for a bounded initial condition* $\mathbf{x}_0$ *implies that its states converge to the equilibrium point* $\bar{\mathbf{x}}_e$*:*

$$||\mathbf{x}_0 - \bar{\mathbf{x}}_e|| < \delta \implies \lim_{t \to \infty} ||\bar{\mathbf{x}}_t|| = \bar{\mathbf{x}}_e \tag{5}$$

Definition 4. *Given a metric space* $(\mathcal{X}, d)$*, a mapping* $T : \mathcal{X} \to \mathcal{X}$ *is called contractive if there exist a constant* $c \in [0, 1)$ *and a metric* d *such that following holds:*

$$d(T(\mathbf{x_1}), T(\mathbf{x_2})) \leq c d(\mathbf{x_1}, \mathbf{x_2}), \ \forall \mathbf{x_1}, \mathbf{x_2} \in \mathcal{X} \tag{6}$$

Contraction generalizes to maps $T : \mathcal{X} \to \mathcal{Y}$ *between two metric spaces* $(\mathcal{X}, d)$*, and* $(\mathcal{Y}, d')$ *as:*

$$d'(T(\mathbf{x_1}), T(\mathbf{x_2})) \leq c d(\mathbf{x_1}, \mathbf{x_2}), \ \forall \mathbf{x_1}, \mathbf{x_2} \in \mathcal{X} \tag{7}$$

Definition 5. *An affine map* $T(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ *with* $\mathbf{x} \in \mathbb{R}^n$ *and metric* $d = || \cdot ||_p$ *is a contraction mapping if the spectral norm of its linear component* $\mathbf{A}$ *is less than one, i.e.* $||\mathbf{A}||_2 < 1$*.*

Theorem 6. *Banach fixed-point theorem. Lets have non-empty complete metric space* $(\mathcal{X}, d)$ *then every contraction* (6) *is converging towards a unique fixed point* $T(\mathbf{x}_{ss}) = \mathbf{x}_{ss}$*.*

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

1.2 Stable Deep Markov Models with Bounded Equilibria

Corollary 7. *An equilibrium $\bar{\mathbf{x}}_e$ of a mean-square stable deep Markov model is bounded by $\underline{\mathbf{x}} \leq ||\bar{\mathbf{x}}_e||_p \leq \overline{\mathbf{x}}$ if either of the conditions in Theorem 3 or Corollalry 5 are satisfied, and the following holds $\forall \mathbf{x} \in dom(\mathbf{f}_{\theta_\mathbf{f}}(\mathbf{x}))$.*

$$\overline{\mathbf{x}} \leq \frac{\|\mathbf{b}_\mathbf{f}(\mathbf{x})\|_p}{1 - \|\mathbf{A}_\mathbf{f}(\mathbf{x})\|_p}, \quad \frac{\|\mathbf{b}_\mathbf{f}(\mathbf{x})\|_p}{1 + \|\mathbf{A}_\mathbf{f}(\mathbf{x})\|_p} \leq \underline{\mathbf{x}} \tag{8}$$

Proof. The conditions in Theorem 3 and Corollary 5 imply mean square stability of the DMM with a mean dynamics converging towards a stable equilibrium $\bar{\mathbf{x}}_e$:

$$\bar{\mathbf{x}}_e = \mathbf{f}_{\theta_\mathbf{f}}(\bar{\mathbf{x}}_e) = \lim_{t \to \infty} \mathbf{f}_{\theta_\mathbf{f}}(\bar{\mathbf{x}}_t) \tag{9}$$

Now by substitution of the PWA form of the DNN into (9) we get:

$$\bar{\mathbf{x}}_e = \mathbf{A}_\mathbf{f}(\bar{\mathbf{x}}_e)\bar{\mathbf{x}}_e + \mathbf{b}_\mathbf{f}(\bar{\mathbf{x}}_e) \tag{10}$$

Where the matrix $\mathbf{A}_\mathbf{f}(\bar{\mathbf{x}}_e)$ and bias vector $\mathbf{b}_\mathbf{f}(\bar{\mathbf{x}}_e)$ uniquely define the affine equilibrium dynamics.

For the upper bound in (8), we apply operator norm (1) to the equation (10) to obtain:

$$||\bar{\mathbf{x}}_e||_p = ||\mathbf{A}_\mathbf{f}(\bar{\mathbf{x}}_e)\bar{\mathbf{x}}_e + \mathbf{b}_\mathbf{f}(\bar{\mathbf{x}}_e)||_p \tag{11}$$

Then applying triangle inequality (3) and operator upper bound $||\mathbf{A}\mathbf{x}||_p \leq ||\mathbf{A}||_p||\mathbf{x}||_p$ we get:

$$||\bar{\mathbf{x}}_e||_p \leq ||\mathbf{A}_\mathbf{f}(\bar{\mathbf{x}}_e)||_p||\bar{\mathbf{x}}_e||_p + ||\mathbf{b}_\mathbf{f}(\bar{\mathbf{x}}_e)||_p \tag{12}$$

And by applying straightforward algebra we have:

$$(1 - ||\mathbf{A}_\mathbf{f}(\bar{\mathbf{x}}_e)||_p)||\bar{\mathbf{x}}_e||_p \leq ||\mathbf{b}_\mathbf{f}(\bar{\mathbf{x}}_e)||_p \tag{13}$$

With resulting equilibrium upper bound given as:

$$||\bar{\mathbf{x}}_e||_p \leq \frac{||\mathbf{b}_\mathbf{f}(\bar{\mathbf{x}}_e)||_p}{1 - ||\mathbf{A}_\mathbf{f}(\bar{\mathbf{x}}_e)||_p} \tag{14}$$

For deriving the lower bound in (8), we start with straightforward algebraic operations on (10) to obtain equilibrium dynamics in a form:

$$(\mathbf{I} - \mathbf{A}_\mathbf{f}(\bar{\mathbf{x}}_e))\bar{\mathbf{x}}_e = \mathbf{b}_\mathbf{f}(\bar{\mathbf{x}}_e) \tag{15}$$

Again for two equivalent vectors their norms must be equal:

$$||(\mathbf{I} - \mathbf{A}_\mathbf{f}(\bar{\mathbf{x}}_e))\bar{\mathbf{x}}_e||_p = ||\mathbf{b}_\mathbf{f}(\bar{\mathbf{x}}_e)||_p \tag{16}$$

Now applying operator norm upper bound inequality $||\mathbf{A}\mathbf{x}||_p \leq ||\mathbf{A}||_p||\mathbf{x}||_p$ to (16) we have:

$$||(\mathbf{I} - \mathbf{A}_\mathbf{f}(\bar{\mathbf{x}}_e))||_p||\bar{\mathbf{x}}_e||_p \geq ||\mathbf{b}_\mathbf{f}(\bar{\mathbf{x}}_e)||_p \tag{17}$$

$$||\bar{\mathbf{x}}_e||_p \geq \frac{||\mathbf{b}_\mathbf{f}(\bar{\mathbf{x}}_e)||_p}{||\mathbf{I} - \mathbf{A}_\mathbf{f}(\bar{\mathbf{x}}_e)||_p} \tag{18}$$

Then from triangle inequality $||\mathbf{I} - \mathbf{A}_\mathbf{f}(\bar{\mathbf{x}}_e)||_p \leq ||\mathbf{I}|| + ||\mathbf{A}_\mathbf{f}(\bar{\mathbf{x}}_e)||_p$ we obtain a lower bound on the equilibrium norm as follows:

$$||\bar{\mathbf{x}}_e||_p \geq \frac{||\mathbf{b}_\mathbf{f}(\bar{\mathbf{x}}_e)||_p}{1 + ||\mathbf{A}_\mathbf{f}(\bar{\mathbf{x}}_e)||_p} \tag{19}$$

Now clearly if the conditions of Corollary 5 are satisfied then the conditions (14) and (19) hold. □

1.3 Stability Regularizations for Deep Markov Models

This section proposes additional regularization methods for learning stable deep Markov models. The most direct approach is to include the stability conditions as extra penalties in the DMM loss function.

$$\mathcal{L}_{\texttt{stable}} = \max\big(1, \|\mathbf{A}_\mathbf{f}(\mathbf{x})\|_p\big) + \max\big(K, ||\mathbf{A}_\mathbf{g}(\mathbf{x})||_p + \frac{||\mathbf{b}_\mathbf{g}(\mathbf{x})||_p}{\|\mathbf{x}\|_p}\big) \tag{20}$$

Then the stability will be enforced by assigning large penalty weights to (20). Additionally, we can bound the norms of the equilibria of DMMs by enforcing constraints (8) via following penalties.

$$\mathcal{L}_{\texttt{bounded}} = \max\big(0, \underline{\mathbf{x}} - \frac{\|\mathbf{b}_\mathbf{f}(\mathbf{x})\|_p}{1 - \|\mathbf{A}_\mathbf{f}(\mathbf{x})\|_p}\big) + \max\big(0, \frac{\|\mathbf{b}_\mathbf{f}(\mathbf{x})\|_p}{1 - \|\mathbf{A}_\mathbf{f}(\mathbf{x})\|_p} - \overline{\mathbf{x}}\big) \tag{21}$$

We leave the empirical validation of the proposed regularizations (20) and (21) for future work.

1.4 Effect of Activation Functions on Stability of Deep Markov Models

As demonstrated by [Massaroli et al., 2021], different activations generate different phase fields in the context of neural ODEs. In a similar spirit we explore the phase space behaviors of DMMs using different activations and weight regularizations. We kept the weights to be marginally stable $||\mathbf{A}_i|| \approx 1$, and thereafter vary the activation functions and weight regularization type. In the top row of Fig. 1, we consider SVD based regularization for a $2-$layer DNN $\mathbf{f}_{\theta_\mathbf{f}}(\mathbf{x})$ for mean, and $3-$layer DNN $\mathbf{g}_{\theta_\mathbf{g}}(\mathbf{x})$ for diagonal covariances, and test with `ReLU, SELU, Softplus`. We can see that both `ReLU` and `SELU` produce sufficiently stable behaviors with few state trajectories for `SELU` remain oscillating near the equilibrium. Similar to `SELU`, state excursion generated by `Softplus` networks remain bounded but with higher uncertainty. In the bottom row of Fig. 1 we consider tight weight eigenvalue constraints based on Greshgorin discs. For the `ReLU`, and `Softplus`, along with the origin, both the axes become attractors, and `Softplus` produces higher uncertain trajectory oscillations. On the other hand, due to bounded (and contractive) tails `tanh` produces much more stable behavior with Greshgorin factorization. The behavior of different activations can be explained by the contractivity conditions given as $||\mathbf{\Lambda}^{\mathbf{f}}_{\mathbf{z}_i}||_p \leq 1$. The contractivity of activation functions is uniquely defined by their Lipschitz constants. For instance, we know that functions with trivial nullspace and Lipschitz constant $\mathcal{K} < 1$, such as `ReLU, tanh`, are asymptotically stable and hence show less uncertain state excursions. However, activations such as `SELU`, or Softplus are not contractive over the entire domain, and therefore, they can generate unstable dynamics even in the case of contractive weights.

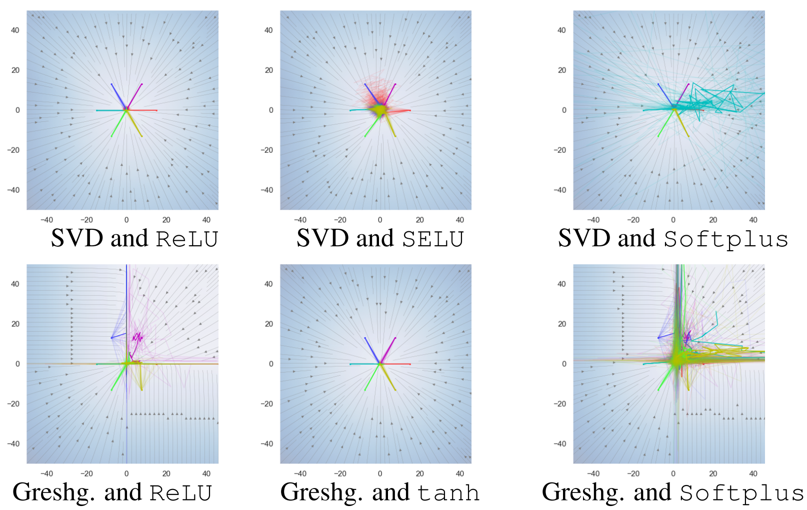

Figure 1: Experiment with different activations for marginally stable weight regularizations using SVD-based and Greshgorin disc-based factorizations. Thin lines are different realizations of the stochastic dynamics with bold lines being their mean trajectory.

References

M. Malek-Shahmirzadi. A characterization of certain classes of matrix norms. *Linear and Multilinear Algebra*, 13(2):97–99, 1983. doi:10.1080/03081088308817508. URL https://doi.org/10.1080/03081088308817508.

S. Massaroli, M. Poli, J. Park, A. Yamashita, and H. Asama. Dissecting neural ODEs, 2021. URL https://arxiv.org/abs/2002.08071.