Role of Parametrization in Learning Dynamics of Recurrent Neural Networks

Adwait Datar[1] ADWAIT.DATAR@TUHH.DE

Chinmay Datar[2,3] CHINMAY.DATAR@TUM.DE

Zahra Monfared[2,4,5] ZAHRA.MONFARED@IWR.UNI-HEIDELBERG.DE

Felix Dietrich[2] FELIX.DIETRICH@TUM.DE

[1] *Institute for Data Science Foundations, Hamburg University of Technology, Germany*
[2] *School of Computation, Information and Technology, Technical University of Munich, Germany*
[3] *Institute for Advanced Study, Technical University of Munich, Germany*
[4] *Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, Germany*
[5] *Faculty of Mathematics and Computer Science, Heidelberg University, Germany*

Abstract

The characteristics of the loss landscape are vital for ensuring efficient gradient-based optimization of recurrent neural networks (RNNs). Learning dynamics in continuous-time RNNs are prone to plateauing effects, with recent studies focusing on this issue by analyzing loss landscapes, particularly in the setting of linear time-invariant (LTI) systems. Building on this work, we explore a fairly simplified setting and study the loss landscape under modal and canonical parametrizations, derived from their respective state-space realizations. We find that canonical parametrization offers improved quasi-convexity properties and faster learning compared to modal forms. Theoretical results are corroborated by numerical experiments. We also show that autonomous ReLU-based RNNs in a modal structure generate trajectories which can be produced by an LTI system while those with a canonical structure produce complex trajectories beyond the scope of LTI systems.

1. Introduction

In training recurrent neural networks (RNNs), the loss landscape plays a critical role in determining the convergence and stability of optimization dynamics. Some parametrizations can lead to favorable landscapes mitigating issues such as vanishing and exploding gradients, allowing for better training dynamics and improved convergence properties while others could lead to the existence of flat regions, saddle points and local minima. We investigate the role of parametrization on convexity properties of the loss landscape for continuous-time linear RNNs and follow closely along the work of [9] and [12]. A canonical parametrization based on the controller canonical form is considered in [9], whereas [12] considers a modal parametrization[1]. These parametrizations are taken up in [22] where a trained neural network with an auto-encoder like architecture is mapped to a linear time-invariant (LTI) system but numerical results are presented only for the canonical parametrization. State-space parametrizations of such LTI systems have been well-studied in the system-identification literature [4–6, 13, 14]. Different parametrizations have been proposed in

[1] These parametrizations will be introduced in the next section.

© A. Datar^1, C. Datar^{2,3}, Z. Monfared^{2,4,5} & F. Dietrich^2.

[15–19, 30, 31] and it is shown that these parametrizations show better properties under Gauss-Newton type optimization methods which are common in the system identification. We contribute to this line of work while focusing on gradient-based optimization techniques and at the two specific parametrizations studied in [12] and [9].

The main contributions of this paper can be summarized as follows. We extend the results on weak quasi-convexity analysis of the loss landscape with the canonical parametrization from [9] to the continuous-time setting and derive a linear matrix inequality to estimate the region of weak quasi-convexity. Furthermore, we show that loss function under the modal parametrization shows poor convexity properties. A discussion on characterizing stationary points of the loss reconciling relevant results [12, 20] is given in Appendix B. Our work together with the results of [9, 12] suggests that the canonical parametrization shows superior theoretical properties compared to the modal parametrization. Preliminary numerical studies support this finding. Finally, we derive an elementary result on non-linear recurrent neural networks with rectified linear unit (ReLU) activation function that proves the superior richness properties of the canonical[2] form over the modal form. These results offer a deeper understanding of the sensitive dependence of the loss landscape on specific parametrizations and introduce systematic tools to compare and analyze different candidate parametrizations. Additionally, they motivate the search for novel parametrizations that are specifically designed to improve training efficiency and stability in RNN training. Theoretical results are presented in Sections 2 and 3 with the proofs being deferred to Appendix D.1. Numerical results are presented in Section 4 followed by conclusions in Section 5.

2. Main Results

We consider a continuous-time RNN with input u and output y described by

$$\frac{d}{dt}x(t) = \sigma(Ax(t) + bu(t)), \qquad x(0) = 0 \quad \text{and} \quad y(t) = cx(t),$$

where $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^{n\times 1}$ and $c \in \mathbb{R}^{1\times n}$ are composed of the trainable weights of RNN and σ is an activation function. We call an RNN *linear* when σ is the identity operator and we will make this assumption throughout this paper, except for Section 3, where we set σ to be the ReLU function[3]. We refer to Appendix F for a discussion on the applicability of the results to the discrete-time setting. Following [9, 12], we choose the infinite-horizon loss function[4]

$$\mathcal{L}(A, b, c) = \int_0^\infty |c_* \cdot \exp(A_*\tau) \cdot b_* - c \cdot \exp(A\tau) \cdot b|^2 d\tau,$$

where matrices (A_*, b_*, c_*) are the model matrices of a true underlying system. Note that for any non-singular matrix T, $\mathcal{L}(A, b, c) = \mathcal{L}(T^{-1}AT, T^{-1}b, cT)$. This illustrates the non-uniqueness of state-space realizations and shows that the loss indeed merely depends on the input-output behavior. [12] uses a state-space realization $(\hat{A}, \hat{b}, \hat{c})$ in a modal form whereas [9] uses a state-space realization

[2] The use of the words "canonical" or "modal" is not well-founded in the non-linear setting but we nevertheless use them to represent the sparsity pattern.

[3] The ReLU activation function is defined as $\sigma(x) = \max\{0, x\}$.

[4] The loss function for the general setting is given by $\mathcal{L} = \int_0^T |y(\tau) - y_*(\tau)|^2 d\tau$, where y and y_* represent the predicted output and the true output for a given input u. We restrict attention to infinite-horizon loss ($T \to \infty$).

$(\tilde{A}, \tilde{b}, \tilde{c})$ in the controller canonical form to parametrize the respective models as

$$\left[\begin{array}{c|c} \hat{A} & \hat{b} \\ \hline \hat{c} & 0 \end{array}\right] = \left[\begin{array}{cccc|c} \hat{a}_1 & 0 & \dots & 0 & 1 \\ 0 & \hat{a}_2 & \ddots & \vdots & 1 \\ \vdots & \ddots & \ddots & 0 & 1 \\ 0 & \dots & 0 & \hat{a}_n & 1 \\ \hline \hat{c}_1 & \hat{c}_2 & \dots & \hat{c}_n & 0 \end{array}\right], \quad \left[\begin{array}{c|c} \tilde{A} & \tilde{b} \\ \hline \tilde{c} & 0 \end{array}\right] = \left[\begin{array}{cccc|c} 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & \ddots & \vdots & 0 \\ \vdots & \ddots & \ddots & 1 & 0 \\ \tilde{a}_1 & \tilde{a}_2 & \dots & \tilde{a}_n & 1 \\ \hline \tilde{c}_1 & \tilde{c}_2 & \dots & \tilde{c}_n & 0 \end{array}\right],$$

where we collect the modal parameters in a vector $\hat{\theta} = (\hat{a}_1, \cdots, \hat{a}_n, \hat{c}_1, \cdots, \hat{c}_n)$ and the canonical parameters in a vector $\tilde{\theta} = (\tilde{a}_1, \cdots, \tilde{a}_n, \tilde{c}_1, \cdots, \tilde{c}_n)$. The mapping between parametrizations is given in Appendix A. We use the notation $\hat{a} = (\hat{a}_1, \cdots, \hat{a}_n)$, $\hat{c} = (\hat{c}_1, \cdots, \hat{c}_n)$ and similarly for $\tilde{a}$ and $\tilde{c}$. Following [9, 12], we make the following assumption to ensure that the true system can be represented in both parametrizations of order n.

Assumption 1 *A_* has distinct, negative, real eigenvalues and (A_*, b_*, c_*) is a minimal realization.*

2.1. Convexity Analysis

The notion of weak quasi-convexity is central in proving convergence with a rate $\mathcal{O}(\frac{1}{t})$ [9] (see Appendix D for details). Translating ideas from [9, Lemma 3] to the continuous-time setting, we obtain the following result that gives a sufficient condition for weak quasi-convexity of the loss.

Theorem 1 (Convexity analysis) *Let $\tilde{\theta}_* = (a_1, \cdots, a_n, c_1, \cdots, c_n)$ be the optimal parameters corresponding to the true system in canonical form. If there exists a positive constant τ and a neighborhood $\mathcal{S}$ of $\tilde{\theta}_*$ such that for all $\omega \in \mathbb{R}$ and for all $\tilde{\theta} = (\tilde{a}_1, \cdots, \tilde{a}_n, \tilde{c}_1, \cdots, \tilde{c}_n) \in \mathcal{S}$*

$$2\mathfrak{Re}\left[\frac{(i\omega)^n - a_n(i\omega)^{n-1} \cdots - a_2(i\omega) - a_1}{(i\omega)^n - \tilde{a}_n(i\omega)^{n-1} - \cdots - \tilde{a}_2(i\omega) - \tilde{a}_1}\right] \geq \tau, \tag{1}$$

then $\mathcal{L}$ is τ-wqc over $\mathcal{S}$ with respect to $\tilde{\theta}_*$.

Theorem 1 provides a useful way of analyzing the domain of weak quasi-convexity of the loss function. In particular, owing to the well-known positive-real lemma (see [2, Section 2.7.2]), we obtain Corollary 2 as an immediate consequence.

Corollary 2 (Region of convergence estimation) *Let $\tilde{\theta}_* = (a_1, \cdots, a_n, c_1, \cdots, c_n)$ be the optimal parameters corresponding to the true system in canonical form. $\mathcal{L}$ is τ-wqc over $\mathcal{S}$ with respect to θ_* where* [5]

$$\mathcal{S} = \left\{\tilde{\theta} \in \mathbb{R}^{2n} \mid \exists P \succ 0, \begin{bmatrix} \tilde{A}^T P + P\tilde{A} & P\tilde{b} - \tilde{a}^T + \tilde{a}_*^T \\ \tilde{b}^T P - \tilde{a} + \tilde{a}_* & 0 \end{bmatrix} \preceq 0\right\}.$$

In order to avoid evaluating the set $\mathcal{S}$ via a linear matrix inequality, a local result is given next.

Corollary 3 *(Local weak quasi-convexity) For every $\tau \in (0, 1)$, there exists a positive constant R such that $\mathcal{L}$ is τ-wqc over the domain $\mathcal{B}_R(\tilde{\theta}_*) := \{\tilde{\theta}_* + \delta \in \mathbb{R}^{2n} \mid \|\delta\| \leq R\}$ with respect to $\tilde{\theta}_*$.*

[5] The notation $P \succ 0$ means P is symmetric positive definite whereas $\preceq$ is used to mean symmetric negative semi-definiteness.

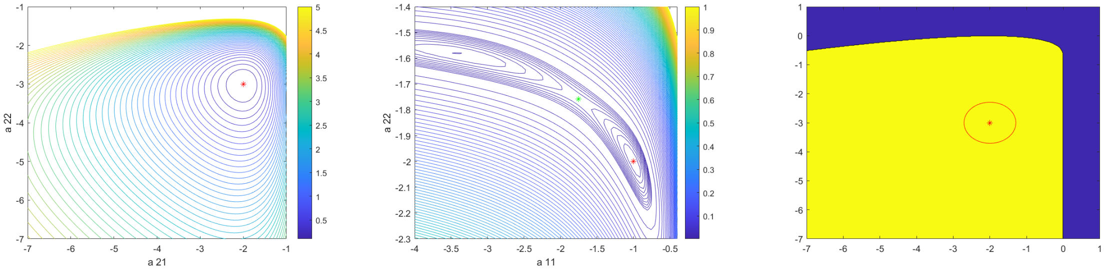

Figure 1: Illustration on a low dimensional example with $n = 2$ where the c matrix for both parametrizations is set to their respective optimal values. Contour plots of the loss under canonical parametrization is shown on the left and under modal parametrization in the middle. Right figure shows the domain of weak quasi-convexity where the yellow region denotes $\mathcal{S}$ from Corollary 2 whereas the red circle denotes $\mathcal{B}_R(\tilde{\theta}_*)$ from Corollary 3. Red star $*$ denotes the global optimum and green star $*$ denotes a saddle point.

We now move our attention to the modal parametrization and obtain the following lemma where $\mathbb{1}$ is used to denote the vector of all ones.

Lemma 4 *Let $\hat{\theta}_* = (a_1, \cdots, a_n, c_1, \cdots, c_n)$ be the optimal parameters corresponding to the true system in modal form. Then $(\hat{\theta} - \hat{\theta}_*)^T \nabla \mathcal{L}(\hat{\theta}) = \int_0^\infty E(i\omega)^* \left(2\mathfrak{Re}\left[M(i\omega)\right]\right) E(i\omega) d\omega$ where $M(s) = \mathbb{1}\left[\frac{s-a_1}{s-\hat{a}_1} \quad \dots \quad \frac{s-a_n}{s-\hat{a}_n}\right]$ and $E(s) = \left[\left(\frac{\hat{c}_1}{s-\hat{a}_1} - \frac{c_1}{s-a_1}\right) \quad \dots \quad \left(\frac{\hat{c}_n}{s-\hat{a}_n} - \frac{c_n}{s-a_n}\right)\right]^T$.*

Following the strategy used in the canonical parametrization, we can try to use Lemma 4 to verify $\tau-$weak quasi-convexity of $\mathcal{L}$ by checking if the integrand can be uniformly lower bounded as $E(i\omega)^* \left(2\mathfrak{Re}\left[M(i\omega) - \tau\mathbb{1}\mathbb{1}^T\right]\right) E(i\omega) \geq 0 \quad \forall \omega \in \mathbb{R}$. This, however, turns out to be impossible, because one can construct $\hat{\theta}$ arbitrarily close to $\hat{\theta}_*$ such that $E(0)^* \left(2\mathfrak{Re}\left[M(0)\right]\right) E(0) < 0$ as illustrated in Appendix E.

Figure 2.1 illustrates the application of these results on a low-dimensional example with $n = 2$. It can be seen that the canonical parametrization leads to convex sub-level sets and a unique global minimizer as expected whereas the modal parametrization shows poor convexity-properties with the presence of a saddle point and two minimizers.

3. Autonomous RNNs with ReLU Activations

In contrast to the results on LTI systems, defining a canonical form for general non-linear systems is fairly involved. Furthermore, similarity transformations $(A, b, c) \to (T^{-1}AT, T^{-1}b, cT)$ do not keep the input-output response invariant. Acknowledging these difficulties and subtleties involved in the analysis of non-linear systems, we investigate the richness of the model classes obtained by imposing the sparsity structure on the model matrices motivated from the modal and canonical parametrizations. To simplify the analysis, we focus first on autonomous systems (i.e., $u \equiv 0$) with non-zero initial conditions. We observe that the set of output trajectories generated when imposing a modal structure is a strict subset of the trajectories that can be generated by an LTI system of potentially smaller order.

Proposition 5 *Let σ be the element-wise ReLU activation function and let x and y be the state and output trajectories generated by*

$$\frac{d}{dt}x(t) = \sigma(\hat{A}x(t)) \quad \text{and} \quad y(t) = \hat{c}x(t)$$

for some parameters $\hat{\theta} = (\hat{a}_1, \cdots, \hat{a}_n, \hat{c}_1, \cdots, \hat{c}_n)$ and initial condition $x(0)$. There exist $A \in \mathbb{R}^{m \times m}$ (diagonal) and $c \in \mathbb{R}^{1 \times m}$ with $m \leq n$ such that $y(t) = c\exp(At)\mathbb{1}$.

Conversely, we show that there exist trajectories generated by an RNN with a canonical structure on A that cannot be generated by any LTI system. Consider

$$\frac{d}{dt}x(t) = \sigma\left(\begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} x(t)\right) \quad \text{and} \quad y(t) = \begin{bmatrix} 0 & 1 \end{bmatrix} x(t),$$

with initial condition $\tilde{x}(0) = [-1 \quad 0]^T$. This system can be solved analytically to obtain the output trajectory $y(t) = 2e^{-2t}(e^t - 1)$ for $t \in [0, \ln(2))$. For $t \geq \ln(2)$, $y(t) = \frac{1}{2}$. Note that since $y(t)$ is not twice continuously differentiable, there does not exist A, c and x_0 such that $y(t) = ce^{At}x_0$, which illustrates that the canonical structure produces richer trajectories.

4. Numerical Experiments

We conduct numerical experiments to explore the loss landscapes and parameter trajectories, systematically comparing the modal and controller canonical (cc) forms in terms of contour plots, loss curves, and the number of iterations needed to reach a specific loss. We also study the performance of three optimizers, GD, Nesterov, and Adam, for both parametrizations. In order to study the local and global features of the loss landscapes and the performance of various optimizers for the two parametrizations, we initialize the learnable parameters both "near" and "far" from the true parameter values, focusing on hidden dimensions 2 and 8. For details on initialization, comparison, and the problem setup, see Appendix G, and for case studies on learning systems with hidden dimensions 2 and 8, please refer to Appendix G.1, Appendix G.2, respectively. The code to reproduce the experimental results will be published as an open-source repository upon acceptance.

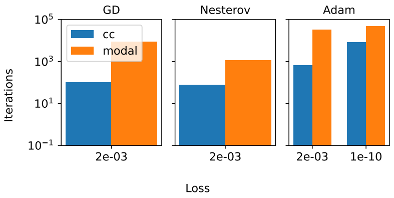

Figure 2: Iterations needed for reaching a prescribed loss for $n = 2$.

When the learnable parameters are initialized close to the global minimum, the results for both parametrizations and for $n = 2$ and $n = 8$ show no notable difference in the number of iterations required to reach a given loss, implying that neither parametrization holds an advantage in terms of faster convergence to the true parameters. While GD and Nesterov reach a loss value of at least 10^{-5}, they need more than 50k steps for convergence.In contrast, the Adam optimizer consistently reaches the true parameters in fewer iterations.

When initialized far from the true parameters, we observe that for $n = 2$ and $n = 8$, the canonical form consistently requires fewer iterations to reach a specific loss compared to the modal form, indicating favorable characteristics (See Figure 2 for $n = 2$). All optimizers, Adam, GD, and Nesterov, attain a loss of around 2 to 8 orders of magnitude higher than when the parameters are initialized near the true parameters. The Adam optimizer usually results in a loss of around 2 to 6 orders of magnitude lower than GD and Nesterov. Even though this aligns with the theory, the comparison favors the canonical form as the loss at the initialized parameters in the cc form ($= 4.3$) is slightly lower than the loss in the modal form ($= 19.85$) for $n = 2$. We intend to compare the two parametrizations in different ways in the future to make the comparison more fair (cf. Appendix G).

Figure 3 shows the contour plot of the learnable parameters of the state matrix and their respective trajectories in the two parametrizations. The $\hat{c}$ and $\tilde{c}$ parameters are eliminated by plugging in their optimal values for each $\hat{a}$ and $\tilde{a}$, respectively. Note that this is different from the plots in Figure 2.1 where $\hat{c}$ and $\tilde{c}$ are set to the parameters of the true system. As the canonical form has a unique global minimum, all trajectories slowly converge towards the global minimum. The modal form has multiple global minima, and we observe that different initializations may lead to trajectories converging towards different global minima. In both cases, the Adam optimizer finds globally optimal parameters, whereas GD and Nesterov optimizers reach the shallow valley in which the decay of the loss function is very slow, and do not find the true parameters within 50000 iterations. The Adam optimizer's use of adaptive learning rates for each parameter (unlike GD and Nesterov) and momentum (absent in GD) enables faster convergence to the global minimum by taking advantage of momentum and allows increasing learning rates in the shallow valley regions.

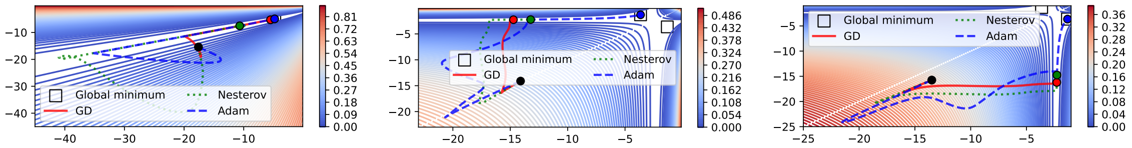

Figure 3: Trajectories of learnable parameters of the state matrix A on a contour plot for 2×2 system and parameters initialized away from the true parameters. The black circle indicates the starting point, and the respective colored circles indicate the parameter values at the last (50000^{th}) iteration. (Left): CC - trajectory 1, (Middle): Modal - trajectory - 1, (Right): Modal trajectory 2.

5. Conclusions and Future Work

This work investigates the loss landscape properties of continuous-time RNNs under canonical and modal parametrizations in a fairly simplified setting. We demonstrate that, under the canonical parametrization, the loss function exhibits weak quasi-convexity within a region that can be estimated. In contrast, the modal parametrization displays poor convexity properties, as evidenced by a counterexample and contour plots from a low-dimensional case. Preliminary numerical studies support these findings, with more extensive systematic studies currently underway. Lastly, we establish that autonomous RNNs with ReLU activation possess richer dynamical properties under the canonical structure than the modal structure.

Several open research directions show promise. For example, it would be interesting to see how the different parametrizations proposed in [15–19, 30, 31] affect the convexity properties of the loss and, therefore, the learning dynamics. Following the discussion on the geometric aspects of the manifold presented in Appendix C, one can exploit the Riemannian structure of the set of stable transfer functions to investigate an algorithm based on the Riemannian gradient descent.

Acknowledgments

F.D. is thankful to the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - for funding through project No. 468830823, and also acknowledge the association to DFG-SPP-229. Z.M. is also grateful to the Bundesministerium für Bildung und Forschung (BMBF, Federal Ministry of Education and Research) for funding through project OIDLITDSM, No. 01IS24061. C.D. is grateful for the funding from the Institute of Advanced Study at the Technical University of Munich (TUM-IAS).

References

[1] Michel Antsaklis. *A Linear Systems Primer*. Birkhäuser, Boston, MA, 2007. ISBN 978-0-8176-4460-4. doi: 10.1007/978-0-8176-4661-5.

[2] Stephen P. Boyd, editor. *Linear Matrix Inequalities in System and Control Theory*. Number vol. 15 in SIAM Studies in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, 1994. ISBN 978-0-89871-334-3.

[3] R. Brockett. Some geometric questions in the theory of linear systems. *IEEE Transactions on Automatic Control*, 21(4):449–455, August 1976. ISSN 0018-9286. doi: 10.1109/TAC.1976.1101301.

[4] M. Gevers and V. Wertz. Parametrization Issues in System Identification. *IFAC Proceedings Volumes*, 20(5):287–292, July 1987. ISSN 14746670. doi: 10.1016/S1474-6670(17)55514-7.

[5] K. Glover and J. Willems. Parametrizations of linear dynamical systems: Canonical forms and identifiability. *IEEE Transactions on Automatic Control*, 19(6):640–646, December 1974. ISSN 0018-9286. doi: 10.1109/TAC.1974.1100711.

[6] Keith Glover. *Structural aspects of system identification.* PhD thesis, Massachusetts Institute of Technology, 1973.

[7] B. Hanzon. A Geometric Approach to System Identification Using Model-reduction Techniques. *IFAC Proceedings Volumes*, 21(9):695–700, August 1988. ISSN 14746670. doi: 10.1016/S1474-6670(17)54809-0.

[8] Bernard Hanzon and Raimund J. Ober. Overlapping Block-Balanced Canonical Forms and Parametrizations: The Stable SISO Case. *SIAM Journal on Control and Optimization*, 35(1): 228–242, January 1997. ISSN 0363-0129, 1095-7138. doi: 10.1137/S0363012993260549.

[9] Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018.

[10] Michiel Hazewinkel and Rudolf E Kalman. On invariants, canonical forms and moduli for linear, constant, finite dimensional, dynamical systems. In *Mathematical Systems Theory: Proceedings of the International Symposium Udine, Italy, June 16–27, 1975*, pages 48–60. Springer, 1976.

[11] Hassan K. Khalil. *Nonlinear Systems*. Prentice Hall, 2002. ISBN 978-0-13-067389-3.

[12] Zhong Li, Jiequn Han, Qianxiao Li, et al. On the curse of memory in recurrent neural networks: Approximation and optimization analysis. *arXiv preprint arXiv:2009.07799*, 2020.

[13] L. Ljung. On Convexification of System Identification Criteria. *Automation and Remote Control*, 80(9):1591–1606, September 2019. ISSN 0005-1179, 1608-3032. doi: 10.1134/S0005117919090030.

[14] Lennart Ljung and Tianshi Chen. Convexity issues in system identification. In *2013 10th IEEE International Conference on Control and Automation (ICCA)*, pages 1–9, June 2013. doi: 10.1109/ICCA.2013.6565206.

[15] T. McKelvey and A. Helmersson. System identification using an over-parametrized model class-improving the optimization algorithm. In *Proceedings of the 36th IEEE Conference on Decision and Control*, volume 3, pages 2984–2989, San Diego, CA, USA, 1997. IEEE. ISBN 978-0-7803-4187-6. doi: 10.1109/CDC.1997.657905.

[16] Tomas McKelvey. *Identification of State-Space Models from Time and Frequency Data*. Univ, Linköping, 1995. ISBN 978-91-7871-531-2.

[17] Tomas McKelvey. A new minimal local parametrization for multivariable linear systems. *IFAC Proceedings Volumes*, 35(1):247–252, 2002.

[18] Tomas McKelvey, Anders Helmersson, and Thomas Ribarits. Data driven local coordinates for multivariable linear systems and their application to system identification. *Automatica*, 40 (9):1629–1635, September 2004. ISSN 00051098. doi: 10.1016/j.automatica.2004.04.015.

[19] Tomas McKelvey, Anders Helmersson, and Thomas Ribarits. Data driven local coordinates for multivariable linear systems and their application to system identification. *Automatica*, 40 (9):1629–1635, September 2004. ISSN 00051098. doi: 10.1016/j.automatica.2004.04.015.

[20] L. Meier and D. Luenberger. Approximation of linear constant systems. *IEEE Transactions on Automatic Control*, 12(5):585–588, October 1967. ISSN 0018-9286, 1558-2523. doi: 10.1109/TAC.1967.1098680.

[21] Petar Mlinarić, Christopher Beattie, Zlatko Drmač, and Serkan Gugercin. Irka is a riemannian gradient descent method. *arXiv preprint arXiv:2311.02031*, 2023.

[22] Tobias Nagel and Marco F. Huber. Autoencoder-Inspired Identification of LTI Systems. In *2021 European Control Conference (ECC)*, pages 2352–2357, Delft, Netherlands, June 2021. IEEE. ISBN 978-94-6384-236-5. doi: 10.23919/ECC54610.2021.9655185.

[23] Yurii Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer International Publishing, Cham, 2018. ISBN 978-3-319-91577-7 978-3-319-91578-4. doi: 10.1007/978-3-319-91578-4.

[24] Mitsuaki Obara, Kazuhiro Sato, Hiroki Sakamoto, Takayuki Okuno, and Akiko Takeda. Stable linear system identification with prior knowledge by riemannian sequential quadratic optimization. *IEEE Transactions on Automatic Control*, 2023.

[25] Ralf Peeters. System identification based on riemannian geometry: theory and algorithms. *(No Title)*, 1994.

[26] Hiroyuki Sato and Kazuhiro Sato. Riemannian optimal system identification algorithm for linear mimo systems. *IEEE control systems letters*, 1(2):376–381, 2017.

[27] Kazuhiro Sato, Hiroyuki Sato, and Tobias Damm. Riemannian optimal identification method for linear systems with symmetric positive-definite matrix. *IEEE Transactions on Automatic Control*, 65(11):4493–4508, 2019.

[28] Eduardo D. Sontag. *Mathematical Control Theory: Deterministic Finite Dimensional Systems*. Springer Science & Business Media, November 2013. ISBN 978-1-4612-0577-7.

[29] Konstantin Usevich and Ivan Markovsky. Optimization on a grassmann manifold with application to system identification. *Automatica*, 50(6):1656–1662, 2014.

[30] J. Vayssettes and G. Mercere. A new parametrisation of matrix fraction descriptions to improve gradient-based optimisation methods. In *53rd IEEE Conference on Decision and Control*, pages 1011–1016, Los Angeles, CA, USA, December 2014. IEEE. ISBN 978-1-4673-6090-6 978-1-4799-7746-8 978-1-4799-7745-1. doi: 10.1109/CDC.2014.7039514.

[31] J. Vayssettes, G. Mercere, Y. Bury, and V. Pommier-Budinger. Structured model identification algorithm based on constrained optimisation. In *2015 European Control Conference (ECC)*, pages 1285–1290, Linz, Austria, July 2015. IEEE. ISBN 978-3-9524269-3-7. doi: 10.1109/ECC.2015.7330715.

[32] Patrick M. Wensing and Jean-Jacques Slotine. Beyond convexity—Contraction and global convergence of gradient descent. *PLOS ONE*, 15(8):e0236661, August 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0236661.

[33] Kemin Zhou and John Comstock Doyle. *Essentials of robust control*, volume 104. Prentice hall Upper Saddle River, NJ, 1998.

Appendix A. Relationship between the parametrizations

We briefly comment on the non-linear mapping between the two parametrization. Let us first consider the low-dimensional example with $n = 2$ before giving a general formula for the mapping. By equating the transfer functions, observe that

$$\frac{\hat{c}_1}{(s-\hat{a}_1)} + \frac{\hat{c}_2}{(s-\hat{a}_2)} = \frac{\hat{c}_1(s-\hat{a}_2)+\hat{c}_2(s-\hat{a}_1)}{(s-\hat{a}_1)(s-\hat{a}_2)} = \frac{\overbrace{(\hat{c}_1+\hat{c}_2)}^{\tilde{c}_2}\, s + \overbrace{(-\hat{c}_1\hat{a}_2-\hat{c}_2\hat{a}_1)}^{\tilde{c}_1}}{s^2 - \underbrace{(\hat{a}_1+\hat{a}_2)}_{\tilde{a}_2}\, s - \underbrace{(-\hat{a}_1\hat{a}_2)}_{\tilde{a}_1}}.$$

Thus, the relationship between the modal parametrization and the canonical parametrization for $n = 2$ can be described via the map

$$\mathfrak{f}_{\text{mod-can}} : \mathbb{R}^4 \ni (\hat{a}_1, \hat{a}_2, \hat{c}_1, \hat{c}_2) \mapsto (-\hat{a}_1\hat{a}_2, \hat{a}_1+\hat{a}_2, -\hat{c}_1\hat{a}_2-\hat{c}_2\hat{a}_1, \hat{c}_1+\hat{c}_2) \in \mathbb{R}^4.$$

Note that $\mathfrak{f}_{\text{mod-can}}$ is not injective since $\mathfrak{f}_{\text{mod-can}}(\hat{a}_1, \hat{a}_2, \hat{c}_1, \hat{c}_2) = \mathfrak{f}_{\text{mod-can}}(\hat{a}_2, \hat{a}_1, \hat{c}_2, \hat{c}_1)$. Furthermore, there does not exist a point $(\hat{a}_1, \hat{a}_2, \hat{c}_1, \hat{c}_2) \in \mathbb{R}^4$ such that $\mathfrak{f}_{\text{mod-can}}(\hat{a}_1, \hat{a}_2, \hat{c}_1, \hat{c}_2) = (-1, 0, 1, 1)$ showing that $\mathfrak{f}_{\text{mod-can}}$ is not surjective either. Therefore, corresponding to a particular input-output relationship, there exist many state-space realizations in modal form that are mapped to a single state-space realization in canonical form. For higher order systems, this mapping can be described via a transformation matrix T that maps a state-space realization in modal form to one in the canonical form as

$$\left[\begin{array}{c|c} \tilde{A} & \tilde{b} \\ \hline \tilde{c} & 0 \end{array}\right] = \left[\begin{array}{c|c} T^{-1}\hat{A}T & T^{-1}\hat{b} \\ \hline \hat{c}T & 0 \end{array}\right] \text{ where } T = \begin{bmatrix} e_n^T \mathcal{C}^{-1} \\ e_n^T \mathcal{C}^{-1}\hat{A} \\ \vdots \\ e_n^T \mathcal{C}^{-1}\hat{A}^{n-1} \end{bmatrix} \tag{2}$$

with e_n being the n^{th} column of the n dimensional identity matrix and and $\mathcal{C} = \begin{bmatrix} \hat{b} & \hat{A}\hat{b} & \cdots & \hat{A}^{n-1}\hat{b} \end{bmatrix}$ is the controllability matrix[6]. For more details on this transformation, see [1, Section 6.4.1].

Appendix B. Characterization of stationary points and global minimizers

A number of results illuminating the difficulties in learning of RNNs in the modal form have been derived in [12]. We now review some of these results that are relevant for our setting and investigate their analogues for the canonical form. For example, [12, Theorem D.1 and Theorem D.2] together show that there exist $m!$ global minimizers while the number of $d-$coincided critical affine spaces[7], each containing infinitely many stationary points, is at least polynomial times larger than the number of global minimizers. The $m!$ global minimizers are characterized in the proof of [12, Theorem D.1] and these are obtained from a simple permutation. Furthermore, [12, Theorem D.3] shows that the Hessian $\nabla^2\mathcal{L}(\theta_{\text{mod}})$ on these $d-$coincided critical affine spaces has at least $n - d$ zero eigenvalues. See [12, Section D.1.3] for details on the application of this theory on the low-dimensional example.

[6] Note that the controllability matrix obtained with the modal form is a Vandermonde matrix and one can use formulae for it's inverse to compute the transformation matrix T.

[7] For $1 \leq d \leq n$, a $d-$coincided critical affine spaces are defined to be the set of points in the space of parameters θ_{mod} where $\nabla\mathcal{L}(\theta_{\text{mod}}) = 0$ and there are at least d distinct entries in the parameter vector $\hat{a} = (\hat{a}_1, \cdots, \hat{a}_n)$.

It is shown that there exist 2 global minimizers of the form (a_1, a_2, c_1, c_2) and (a_2, a_1, c_2, c_1) and a $1-$coincided critical affine space $\{(\hat{a}_1, a - \hat{a}, w, w \mid a \in \mathbb{R})\}$ which contains saddle points and degenerate stable-points that are not global minimizes.

In contrast, it has been shown that the canonical form for single-input single-output system is an identifiable model structure, i.e., there exists a unique global minimizer (see [6]). It can be shown that all $m!$ global minimizers in the modal form are mapped via $\mathfrak{f}_{\text{mod-can}}$ to a unique minimizer in the canonical form. Furthermore, since the loss landscape satisfies the local weak quasi-convexity property (Definition 6), it can be shown that there are no other stationary points in the neighborhood of the global minimum. For a global picture, the stationary points can be characterized by a set of $2n$ algebraic equations [20, see equation (33)]

$$G(-\hat{p}) = G_\theta(-p),$$
$$\frac{dG}{ds}(-\hat{p}) = \frac{dG_\theta}{ds}(-\hat{p}),$$

where $\hat{p}$ are the poles of the RNN $G_\theta(s)$ and $G(-\hat{p})$ is the true transfer system. These equations can be numerically solved to obtain the set of stationary points. Characterizing the set of stationary points using these algebraic equations and comparing them with the number of global minimizers in a way similar to the one adopted in [12] is an interesting question for future work.

Appendix C. Geometric viewpoint

The study of geometric structures that can be imposed on the set linear system has a long history and we cite some of the most relevant works next. [10] develops a differentiable manifold structure on the set of state-space models of order at most n by considering equivalence classes defined via similarity transformations. [3] investigates the geometric aspects of the set of linear single-input single-output systems and shows that this set consists of multiple connected components. [7] defines a so-called Finsler metric on the manifold of systems and presents a parametrization [8] that is especially suitable for the implementation of a parametrization independent Riemannian gradient descent. [21] use a Riemannian manifold structure to interpret existing $\mathcal{H}_2$ model reduction algorithm as a Riemannian gradient descent. Application of such geometric ideas for system identification has been considered in [24–27, 29].

Appendix D. Convergence analysis under weak quasi-convexity

Let us first recall the notion of weak quasi-convexity as defined in [9].

Definition 6 (Weak quasi-convexity) *Let τ be a positive constant. A differentiable $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}$ is said to be τ-weakly quasi-convex ($\tau-$wqc) over a domain $\mathcal{S}$ with respect to θ_* if for all $\theta \in \mathcal{S}$,*

$$(\theta - \theta_*)^T \nabla \mathcal{L}(\theta) \geq \tau(\mathcal{L}(\theta) - \mathcal{L}(\theta_*)).$$

Convergence analysis of negative gradient flows has been well studied and convexity plays a central role in obtaining linear and sub-linear convergence rate guarantees (see for example [32] and [23]). For the sake of completeness, we provide a standard convergence analysis based on the notion of weak quasi-convexity which acts as a continuous-time analogue of the global convergence result from [9].

Theorem 7 (Convergence Analysis) *Suppose $\mathcal{L}$ is τ-weakly quasi-convex ($\tau-$wqc) over a domain $\mathcal{S}$ with respect to θ_* and let c be a positive constant such that $\mathcal{B}_c(\theta_*) \subset \mathcal{S}$. If θ is a solution of the gradient flow dynamics*

$$\dot{\theta}(t) = -\nabla\mathcal{L}(\theta(t))$$

with $\theta(0) \in \mathcal{B}_c(\theta_*)$, then

$$(\mathcal{L}(\theta(t)) - \mathcal{L}(\theta_*)) \leq \frac{1}{2\tau t}||\theta(0) - \theta_*||^2 \qquad \forall t \geq 0.$$

Proof The proof follows a standard argument based on a Lyapunov function [11]. If the trajectory $\theta(t)$ stays in $\mathcal{S}$ for all $t \geq 0$, consider the evolution of the energy function $V(t) := \frac{1}{2}||\theta(t) - \theta_*||^2 + t\tau\left(\mathcal{L}(\theta(t)) - \mathcal{L}(\theta_*)\right)$. The chain rule gives us

$$\dot{V}(t) = -(\theta(t) - \theta_*)^T\nabla\mathcal{L}(\theta(t)) + \tau\left(\mathcal{L}(\theta(t)) - \mathcal{L}(\theta_*)\right) - t\tau||\nabla\mathcal{L}(\theta(t))||^2 \leq 0.$$

This implies that for all $t \geq 0$,

$$V(t) = \frac{1}{2}||\theta(t) - \theta_*||^2 + t\tau\left(\mathcal{L}(\theta(t)) - \mathcal{L}(\theta_*)\right) \leq V(0) = \frac{1}{2}||\theta(0) - \theta_*||^2$$

which gives us the desired convergence bound. The only thing left to be proven then is if $\theta(0) \in \mathcal{B}_c(\theta_*) \subset \mathcal{S}$, then $\theta(t) \in \mathcal{S}$ for all $t \geq 0$. This follows from a standard invariance argument [11] completing the proof. ■

D.1. Proofs of theorems

Proof (Theorem 1) The proof proceeds exactly analogous to the proof of [9, Lemma 3] except that the Parseval relation [33, Problem 4.2] in continuous-time evaluates the integral along the imaginary axis in the complex plane instead of the unit circle. We thus obtain

$$\begin{aligned}&(\tilde{\theta} - \theta_*)^T\nabla\mathcal{L}(\tilde{\theta})\\ &\quad= \int_0^\infty 2\mathfrak{Re}\left[\frac{(i\omega)^n - a_n(i\omega)^{n-1}\cdots - a_2(i\omega) - a_1}{(i\omega)^n - \tilde{a}_n(i\omega)^{n-1} - \cdots - \tilde{a}_2(i\omega) - \tilde{a}_1}\right]||G(i\omega) - \hat{G}(iw)||^2 d\omega.\end{aligned}$$

Using the uniform lower bound hypothesis and the definition of weak quasi-convexity, we get the desired result. ■

Proof (Corollary 2) This is a straight-forward application of the positive real lemma [2, Section 2.7.2]. The main idea is that positive realness of a transfer function over all frequencies can be verified by solving an algebraic system of equations. ■

Proof (Corollary 3) We use a linear fractional representation [33, Chapter 9] and a small-gain argument [33, Section 8.2] common in the robust control literature and we refer the reader to these references for more details. With the deviation variables $\delta = (\delta_{a_1}, \cdots, \delta_{a_n}, \delta_{c_1}, \cdots, \delta_{c_n}) = \tilde{\theta} - \theta_*$, we can show that

$$\frac{s^n - a_n s^{n-1} \cdots - a_2 s - a_1}{s^n - \tilde{a}_n s^{n-1} - \cdots - \tilde{a}_2 s - \tilde{a}_1} = 1 + \frac{\delta_{a_n} s^{n-1} \cdots + \delta_{a_2} s + \delta_{a_1}}{s^n - (a_n + \delta_{a_n}) s^{n-1} \cdots - ((a_2 + \delta_{a_2}))s - (a_1 + \delta_{a_1})}$$

$$= 1 + \mathcal{F}_u \left(\underbrace{\begin{bmatrix} P_{11}(s) & P_{12}(s) \\ P_{21}(s) & 0 \end{bmatrix}}_{P(s)}, \underbrace{\begin{bmatrix} \delta_{a_1} & & \\ & \ddots & \\ & & \delta_{a_n} \end{bmatrix}}_{\Delta} \right),$$

where $\mathcal{F}_u\left(P(s), \Delta\right) := P_{21}(s)\Delta(I - P_{11}(s)\Delta)^{-1}P_{12}(s)$ and $P(s)$ does not depend on Δ. Finally, it can be shown that if[8]

$$\|\Delta\| < \frac{1-\tau}{\|(1-\tau)P_{11}(s)\|_{\mathcal{H}_\infty} + \|P_{12}(s)\|_{\mathcal{H}_\infty} \cdot \|P_{21}(s)\|_{\mathcal{H}_\infty}} =: R,$$

then $\|\mathcal{F}_u\left(P(s), \Delta\right)\|_{\mathcal{H}_\infty} < 1$ implying $\frac{s^n - a_n s^{n-1} \cdots - a_2 s - a_1}{s^n - \tilde{a}_n s^{n-1} - \cdots - \tilde{a}_2 s - \tilde{a}_1} \geq \tau$ completing the proof. ■

Proof (Lemma 4) Proceeding in the same way as in the proof of Theorem 1, we use Parseval's relation and reorganize the integrand using the definitions of

$$M(s) = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} \frac{s-a_1}{s-\hat{a}_1} & \cdots & \frac{s-a_n}{s-\hat{a}_n} \end{bmatrix} \quad \text{and} \quad E(s) = \begin{bmatrix} \left(\frac{\hat{c}_1}{s-\hat{a}_1} - \frac{c_1}{s-a_1} \right) \\ \vdots \\ \left(\frac{\hat{c}_n}{s-\hat{n}_2} - \frac{c_n}{s-a_n} \right) \end{bmatrix}$$

to get the desired result.

■

Proof (Proposition 5) Consider the system of differential equations in coordinate form

$$\frac{d}{dt} x^{(i)}(t) = \sigma\left(\hat{a}_i x^{(i)}(t) \right),$$

with initial condition $x^{(i)}(0)$. Observe that if $\hat{a}_i x^{(i)}(0) > 0$, then $x^{(i)}(t) = e^{\hat{a}_i \cdot t} x^{(i)}(0)$ solves the differential equation. On the other hand, if $\hat{a}_i x^{(i)}(0) \leq 0$, then $x^{(i)}(t) = x^{(i)}(0)$ solves the differential equation. Since σ is Lipschitz-continuous, we also have uniqueness of the solution [28, Theorem 54]. Without loss of generality, let m be such that $\hat{a}_i x^{(i)}(0) > 0$ for all $i \in \{1, 2, \cdots, m\}$ and $\hat{a}_i x^{(i)}(0) \leq 0$ for all $i \in \{m+1, m+2, \cdots, n\}$. The trajectory y can now be described as the impulse response of the following LTI system which completes the proof.

$$\frac{d}{dt}\bar{x}(t) = \begin{bmatrix} \hat{a}_1 & & & \\ & \ddots & & \\ & & \hat{a}_m & \\ & & & 0 \end{bmatrix} \bar{x}(t), \qquad \bar{x}(0) = \mathbb{1},$$

$$y(t) = \begin{bmatrix} \left(x^{(1)}(0)\hat{c}_1\right) & \left(x^{(2)}(0)\hat{c}_2\right) & \cdots & \left(x^{(m)}(0)\hat{c}_m\right) & \left(\sum_{i=m+1}^{n} x^{(i)}(0)\hat{c}_i\right) \end{bmatrix} \bar{x}(t).$$

■

[8] We use the $\mathcal{H}_\infty$ norm defined as $\|P(s)\|_{\mathcal{H}_\infty} = \sup_{\omega \in \mathbb{R}} \|P(i\omega)\|$.

Appendix E. Counter example for weak quasi-convexity in modal form

For $\alpha \in \mathbb{R}$, define $\mu(\alpha) = \frac{c_2 - a_2\alpha}{c_2 + a_2\alpha}$ and let

$$\hat{a}_i = \begin{cases} a_i & \text{if } i \in \{1, \cdots, n\} \setminus \{2\}, \\ \frac{a_2}{\mu(\alpha)} & \text{if } i = 2 \end{cases}$$
$$\hat{c}_i = \begin{cases} c_1 - a_1\alpha \left(1 - \mu(\alpha) - \sqrt{2 + 2\mu(\alpha)^2}\right)\left(\frac{1}{\mu(\alpha)} - 1\right) & \text{if } i = 1 \\ c_i & \text{if } i \in \{2, \cdots, n\}. \end{cases}$$

Note that since $\lim_{\alpha\to 0} \mu(\alpha) = 1$, we have that $\lim_{\alpha\to 0} \hat{a}_i = a_i$ and $\lim_{\alpha\to 0} \hat{c}_i = c_i$. It can further be shown that for any $\alpha \in \mathbb{R}$, $E(0)^* \left(2\mathfrak{Re}\left[M(0)\right]\right) E(0) < 0$. This implies that the wqc condition is violated at the 0 frequency no matter how small we choose the neighborhood $\mathcal{S}$ of the optimum and how small we choose τ. This example thus shows that the proof technique we used for the canonical parametrization breaks down for the modal parametrization. We note however that since the sum of non-convex functions can potentially be convex, this example does not conclusively establish that the loss under the modal parametrization is not weakly quasi-convex.

Appendix F. From continuous-time to the discrete-time

In this section, we shortly comment on the analogous discrete-time results when studying discrete-time RNNs which is the more common setting in the literature. As already stated in the main text, Theorem 1 is motivated from its discrete-time analogue which appeared in [9, Lemma 3]. Corollaries 2 and 3 can be extended in a rather straight-forward manner to the discrete-time setting since the robust control tools such as the positive-real lemma and the small-gain theorem already have their discrete-time counterparts. Similarly, Lemma 4 can be extended to the discrete-time setting using a discrete-time version of the Parseval relation. Proposition 5 has an exact analogue in the discrete-time setting which is stated next.

Proposition 8 *Let σ be the element-wise ReLU activation function and let x and y be the state and output trajectories generated by*

$$x(k+1) = \sigma(\hat{A}x(k)) \quad \text{and} \quad y(k) = \hat{c}x(k).$$

for some parameters $\hat{\theta} = (\hat{a}_1, \cdots, \hat{a}_n, \hat{c}_1, \cdots, \hat{c}_n)$ and initial condition $x(0)$. There exist $A \in \mathbb{R}^{m\times m}$ (diagonal) and $c \in \mathbb{R}^{1\times m}$ with $m \leq n$ such that $y(k) = cA^k\mathbb{1}$ for $k \geq 2$.

Proof The proof follows arguments similar to the proof of Proposition 5. Observe that if $\hat{a}_i > 0$ and $x^{(i)}(0) > 0$, then $x^{(i)}(k) = \hat{a}_i^k x^{(i)}(0)$ solves the difference equation. In all other cases, note that $x^{(i)}(k) = 0$ for all $k \geq 2$. Without loss of generality, let m be such that $\hat{a}_i > 0$ and $x^{(i)}(0) > 0$ for all $i \in \{1, 2, \cdots, m\}$ and either $\hat{a}_i \leq 0$ or $x^{(i)}(0) \leq 0$ or both for all $i \in \{m+1, m+2, \cdots, n\}$. The trajectory y can now be described as the impulse response of the following LTI system which completes the proof.

$$\bar{x}(k+1) = \begin{bmatrix} \hat{a}_1 & & \\ & \ddots & \\ & & \hat{a}_m \end{bmatrix} \bar{x}(k), \qquad \bar{x}(0) = \mathbb{1},$$
$$y(k) = \begin{bmatrix} \left(x^{(1)}(0)\hat{c}_1\right) & \left(x^{(2)}(0)\hat{c}_2\right) & \ldots & \left(x^{(m)}(0)\hat{c}_m\right) \end{bmatrix} \bar{x}(k).$$

■

Finally, extension of the results from [12] such as Theorem D.2 and Theorem D.7 is ongoing work and we believe that these results can be extended to their natural counterparts in the discrete-time setting.

Appendix G. Numerical experiments

LTI system: We choose the LTI system proposed in [12] as a starting point. In particular, we populate the matrix $Z \in \mathbb{R}^{n \times n}$ with randomly sampled entries from the Gaussian distribution $\mathcal{N}(0,\, n^{-0.5})$. The state matrix is then computed as $-\mathcal{I} - Z^T Z$, where, $\mathcal{I} \in \mathbb{R}^{n \times n}$ is an identity matrix. The entries of vectors b and c are sampled from the normal distribution $\mathcal{N}(0,\, 1)$ and scalar d is set to 0. We first convert the systems into modal and canonical (cc) parametrizations, respectively, using similarity transformations. In the canonical form, entries of b are not learnable, and in modal form, we fix all entries of b to 1 and absorb the parameters in c without loss of generality.

Setup for numerical experiments: We stack the learnable parameters of the state matrix $A \in \mathbb{R}^{n \times n}$ and the output vector $c \in \mathbb{R}^{1 \times n}$ in a vector denoted by $w \in \mathbb{R}^{2n}$. Let $d_1, d_2 \in \mathbb{R}^{2n}$ be vectors with entries sampled from a uniform distributions $\mathbb{U}(-1, 0)$, $\mathbb{U}(-15, -10)$, respectively. We denote the initial and true parameters by $w_{init}, w_{true} \in \mathbb{R}^{2n}$. We focus on the following cases to understand the local and global aspects of the loss landscape and parameter trajectories:

- Case (A): Parameters initialized "near" the global minimum $w_{init} = w_{true} + d_1$.
- Case (B): Parameters initialized "away" from the global minimum $w_{init} = w_{true} + d_2$.

Note that we first compute true parameters in the respective parametrizations and then initialize the learnable parameters away from the true parameters by perturbations d_1 and d_2. We perform all experiments with three optimizers - Gradient Descent (GD), Nesterov, and Adam. We perform all experiments with three random seeds, thereby changing the initial values for learnable parameters for each seed. The hyper-parameter details are included in Table 1 and Table 2. In Appendix G.1 and Appendix G.2, we perform case studies with LTI systems having state dimensions 2 and 8, respectively. We choose the learning rate that results in the lowest mean of the losses over three different initializations for each parametrization. We select optimal hyper-parameters to plot the loss curves and contour plots. In all the loss curves, we plot the mean of the losses with a solid line, and the shaded region is used to indicate the minimum and maximum of the losses at a given iteration. We check that the eigenvalues of the learned state matrix are in the negative half-plane after every 500 iterations to ensure the stability of the learned LTI system during training and to ensure that the loss is meaningful.

We also show contour plots of the learnable parameters of the state matrix and their respective trajectories in the two parametrizations. In both cases, parameter values of c are set to the optimal values for the given parameter values of the matrix A at any iteration in the computation of the loss for the contour plot. Since we have more parameters in the true optimization problem, the parameter trajectories do not always appear along the gradient direction in the contour plot. We observe a unique global minimum in the canonical form and $n!$ global minima in the modal form.

Note on fair empirical comparison of the two forms: It is not trivial to define a problem setup for a fair comparison of the two approaches, and we wish to investigate more and better ways of comparing them in the future. For our setting, the parameter values in the canonical form usually

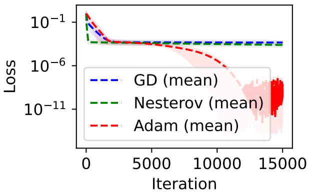

Figure 4: Loss curve: cc

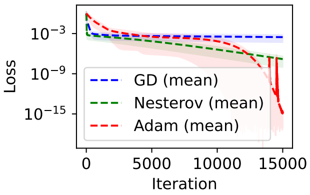

Figure 5: Loss curve: modal

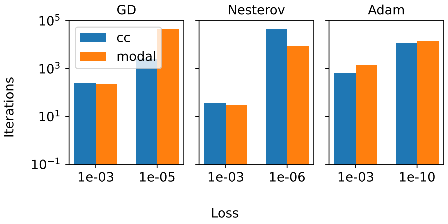

Figure 6: Iterations Vs Loss

Figure 7: **Loss trajectories:** 2×2 **system and Case (A)**

have a larger magnitude in comparison to the parameter values in the modal form. Though we initialize parameters with the same perturbation from the true parameters in both parametrizations, in our examples, the canonical form usually has a lower loss than the modal form for the initial guess and thus has an advantage. Investigating different initialization schemes for the respective forms is ongoing work and this will allow us to remove this bias in the comparison.

G.1. Case study - 1: 2×2 systems

G.1.1. CASE (A): PARAMETERS INITIALIZED "NEAR" THE GLOBAL MINIMUM.

Loss curves: Figure 4 and Figure 5 show that for canonical (cc) form and modal forms, the loss drops quickly within the first few iterations for GD and Nesterov, compared to the Adam optimizer. However, with enough iterations, Adam reduces the loss down to almost machine precision in both parametrizations, whereas Nesterov and GD exhibit extremely slow loss decay after the first drop. Figure 6 shows that the iteration count to attain a certain loss is roughly the same regardless of the parametrization if the weights are initialized close to the global minimum.

Contour plots: Figure 10 shows the contour plot of the learnable parameters of the state matrix and their respective trajectories in the two parametrizations. In both cases, parameter values of the matrix C are set to the optimal values for the given parameter values of the matrix A at any iteration in the computation of the loss. We observe that the Adam optimizer reaches the global minimum, whereas the SGD and Nesterov need more iterations to reach the global minimum.

Parameter trajectories: Figure 11 and Figure 12 show the parameter trajectories for canonical (cc) and modal forms, respectively. The learnable parameters in canonical form are the entries in the last row of the state matrix, viz., $A(2,1)$ and $A(2,2)$ (denoted by $\tilde{a}_1$ and $\tilde{a}_2$ in Section 2), whereas, in the modal form, they are the diagonal entries $A(1,1)$ and $A(2,2)$ (denoted by $\hat{a}_1$ and $\hat{a}_2$ in Section 2). We observe that Adam finds the global minimum in both forms and is very efficient compared to GD and Nesterov. Nesterov finds the global minimum in the modal form but needs more iterations in the canonical form to converge to true parameters, whereas GD performs relatively poorly in both forms.

G.1.2. CASE (B): PARAMETERS INITIALIZED AWAY FROM THE GLOBAL MINIMUM.

Loss curves: Figure 14 and Figure 15 show that for canonical, the loss drops quickly for all optimizers, whereas for the modal form, the loss relatively quickly with GD and Nesterov optimizers,

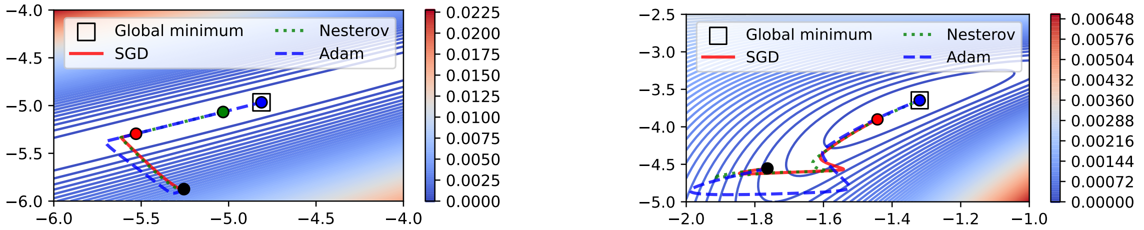

Figure 8: Contour plot: cc

Figure 9: Contour plot: modal

Figure 10: **Contour plots:** 2×2 **system and Case (A)** Trajectories of learnable parameters of the state matrix A on a contour plot. The black circle indicates the starting point, and the respective colored circles indicate the parameter values at the last (50000^{th}) iteration for GD, Nesterov, and Adam optimizers.

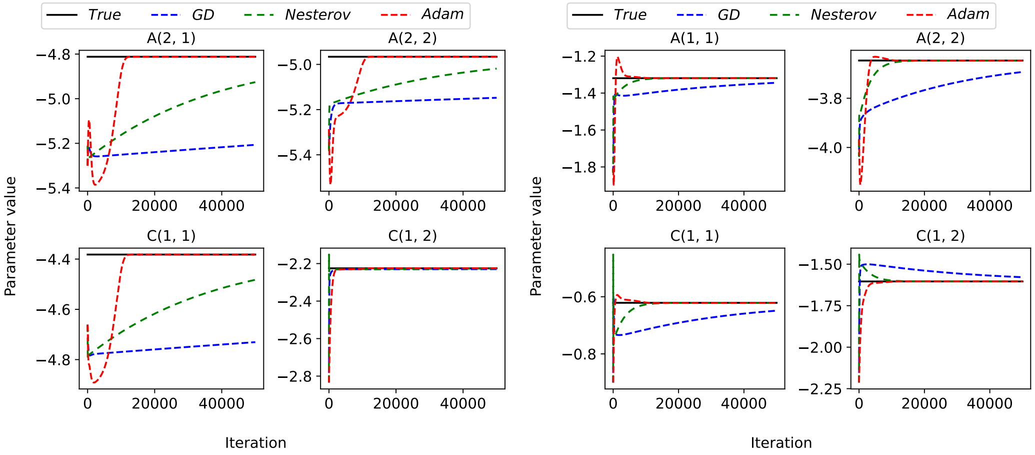

Figure 11: Parameter trajectory: cc

Figure 12: Parameter trajectory: modal

Figure 13: **Parameter trajectories:** 2×2 **system and Case (A)**

compared to the Adam optimizer. However, with enough iterations, Adam reduces the loss down to almost machine precision with both parametrizations, whereas Nesterov and GD exhibit extremely slow loss decay after the first drop.

Parameter trajectories: Figure 17 and Figure 18 show the parameter trajectories for canonical and modal forms, respectively. We observe that Adam finds the global minimum in both forms and is very efficient compared to GD and Nesterov. Interestingly, GD performs better than Nesterov in the canonical form. It seems like momentum does not necessarily help Nesterov in the valley region as it cannot adjust the learning rate, unlike Adam.

G.2. Case study - 2: 8×8 systems

We now study loss curves for the 8×8 systems. Figure 20 and Figure 21 show that all optimizers can attain a loss lower than $1e - 5$ and Adam often results in a loss lower than the other optimizers

Table 1: Details of the hyper-parameters: 2×2 system

Form	Optimization algorithm	Learning rate values	Minimum Loss
CC (Case A)	Gradient Descent	$[10^2, 10^1, \mathbf{10^{-1}}, 10^{-3}, 10^{-5}, 10^{-7}]$	7.77×10^{-6}
	Nesterov	$[10^2, 10^1, \mathbf{10^{-1}}, 10^{-3}, 10^{-5}, 10^{-7}]$	7.54×10^{-7}
	Adam	$[10^2, 10^1, 10^{-1}, \mathbf{10^{-3}}, 10^{-5}, 10^{-7}]$	6.15×10^{-15}
CC (Case B)	Gradient Descent	$[10^2, \mathbf{10^1}, 10^{-1}, 10^{-3}, 10^{-5}, 10^{-7}]$	5.47×10^{-5}
	Nesterov	$[10^2, \mathbf{10^1}, 10^{-1}, 10^{-3}, 10^{-5}, 10^{-7}]$	8.24×10^{-4}
	Adam	$[10^2, 10^1, \mathbf{10^{-1}}, 10^{-3}, 10^{-5}, 10^{-7}]$	9.80×10^{-12}
Modal (Case A)	Gradient Descent	$[10^2, 10^1, \mathbf{10^{-1}}, 10^{-3}, 10^{-5}, 10^{-7}]$	7.05×10^{-6}
	Nesterov	$[10^2, 10^1, \mathbf{10^{-1}}, 10^{-3}, 10^{-5}, 10^{-7}]$	4.13×10^{-18}
	Adam	$[10^2, 10^1, 10^{-1}, \mathbf{10^{-3}}, 10^{-5}, 10^{-7}]$	3.76×10^{-18}
Modal (Case B)	Gradient Descent	$[10^2, 10^1, \mathbf{10^{-1}}, 10^{-3}, 10^{-5}, 10^{-7}]$	1.74×10^{-4}
	Nesterov	$[10^2, 10^1, \mathbf{10^{-1}}, 10^{-3}, 10^{-5}, 10^{-7}]$	1.64×10^{-3}
	Adam	$[10^2, 10^1, 10^{-1}, \mathbf{10^{-3}}, 10^{-5}, 10^{-7}]$	1.41×10^{-17}

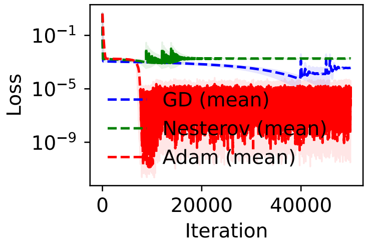

Figure 14: Loss curve: cc

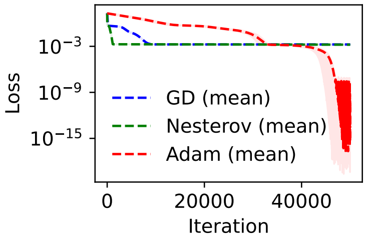

Figure 15: Loss curve: modal

Figure 16: **Loss trajectories:** 2×2 **system and Case (B)**

Table 2: Details of the hyper-parameters: 8×8 system

Form	Optimization algorithm	Learning rate values	Minimum Loss
CC (Case A)	Gradient Descent	$[10^2, \mathbf{10^1}, 10^{-1}, 10^{-3}, 10^{-5}, 10^{-7}]$	3.14×10^{-7}
	Nesterov	$[10^2, \mathbf{10^1}, 10^{-1}, 10^{-3}, 10^{-5}, 10^{-7}]$	3.43×10^{-9}
	Adam	$[10^2, 10^1, 10^{-1}, \mathbf{10^{-3}}, 10^{-5}, 10^{-7}]$	2.12×10^{-12}
CC (Case B)	Gradient Descent	$[10^2, \mathbf{10^1}, 10^{-1}, 10^{-3}, 10^{-5}, 10^{-7}]$	1.92×10^{-4}
	Nesterov	$[10^2, \mathbf{10^1}, 10^{-1}, 10^{-3}, 10^{-5}, 10^{-7}]$	2.3×10^{-6}
	Adam	$[10^2, 10^1, \mathbf{10^{-1}}, 10^{-3}, 10^{-5}, 10^{-7}]$	6.96×10^{-10}
Modal (Case A)	Gradient Descent	$[10^2, 10^1, \mathbf{10^{-1}}, 10^{-3}, 10^{-5}, 10^{-7}]$	2.96×10^{-6}
	Nesterov	$[10^2, 10^1, \mathbf{10^{-1}}, 10^{-3}, 10^{-5}, 10^{-7}]$	2.69×10^{-7}
	Adam	$[10^2, 10^1, 10^{-1}, \mathbf{10^{-3}}, 10^{-5}, 10^{-7}]$	2.84×10^{-9}
Modal (Case B)	Gradient Descent	$[10^2, 10^1, \mathbf{10^{-1}}, 10^{-3}, 10^{-5}, 10^{-7}]$	1.49×10^{-4}
	Nesterov	$[10^2, 10^1, \mathbf{10^{-1}}, 10^{-3}, 10^{-5}, 10^{-7}]$	2.59×10^{-4}
	Adam	$[10^2, 10^1, \mathbf{10^{-1}}, 10^{-3}, 10^{-5}, 10^{-7}]$	7.09×10^{-7}

by up to five orders of magnitude. The final loss attained by the canonical form is lower than the one obtained by the modal form. However, importantly, Figure 22 shows that the canonical form

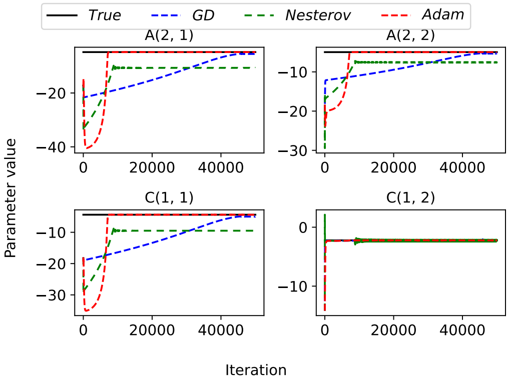

Figure 17: Parameter trajectory: cc

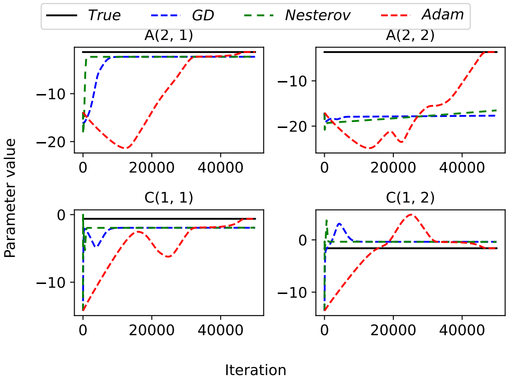

Figure 18: Parameter trajectory: modal

Figure 19: **Parameter trajectories:** 2×2 **system and Case (B)**

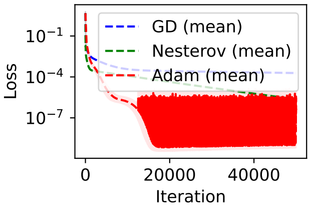

Figure 20: Loss curve: cc

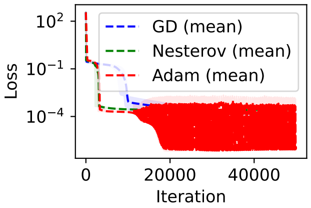

Figure 21: Loss curve: modal

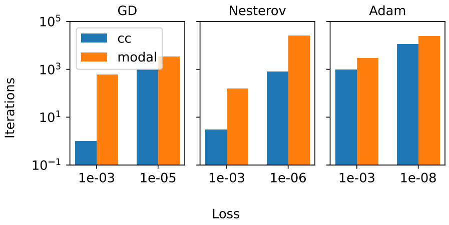

Figure 22: Iterations Vs Loss

Figure 23: **Loss trajectories:** 8×8 **system and Case (A)**

requires much fewer iterations than the modal form, to attain a particular loss. The observations in case (B), when we start away from the global minimizer are similar (cf Figure 24, Figure 25, Figure 26).

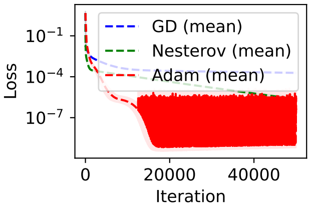

Figure 24: Loss curve: cc

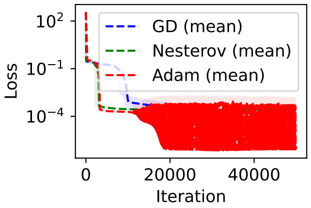

Figure 25: Loss curve: modal

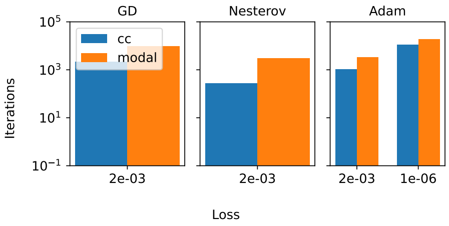

Figure 26: Iterations Vs Loss

Figure 27: **Loss trajectories:** 8×8 **system and Case (B)**