

788 A Proof

789 A.1 Lemmas

790 We use the following lemma in our derivation.

Lemma A.1 (Singh et al.).

$$\frac{\partial \mathbf{A} \mathbf{X} \mathbf{B}}{\partial \mathbf{X}} = \mathbf{A} \otimes \mathbf{B}^\top \quad (21)$$

791 **Lemma A.2.** Let $\mathbf{Y} \in \mathbb{R}^{S \times D}$ be an input matrix. Then, the Jacobian of Π with respect to \mathbf{Y} is given
792 by

$$\frac{\partial \Pi(\mathbf{Y})}{\partial \mathbf{Y}} = \text{blockdiag} \left(\left\{ \frac{1}{\|\mathbf{Y}_{[i,:]} \|} (\mathbf{I}_D - \frac{\mathbf{Y}_{[i,:]} \mathbf{Y}_{[i,:]}^\top}{\|\mathbf{Y}_{[i,:]} \|^2}) \right\}_{i=1}^S \right) \quad (22)$$

793 *Proof.* Since Π operates independently on each row of \mathbf{Y} , the Jacobian is block-diagonal, with each
794 block corresponding to the derivative of a single normalized row:

$$\frac{\partial \Pi(\mathbf{Y})}{\partial \mathbf{Y}} = \text{blockdiag}(\left\{ \frac{\partial \Pi(\mathbf{Y})_{[i,:]} }{\partial \mathbf{Y}_{[i,:]} } \right\}_{i=1}^S). \quad (23)$$

795 For each row, we compute the gradient of the normalized vector:

$$\frac{\partial \Pi(\mathbf{Y})_{[i,:]} }{\partial \mathbf{Y}_{[i,:]} } = \frac{\partial \mathbf{Y}_{[i,:]} / \|\mathbf{Y}_{[i,:]} \|}{\partial \mathbf{Y}_{[i,:]} } \quad (24)$$

$$= \frac{1}{\|\mathbf{Y}_{[i,:]} \|} \frac{\partial \mathbf{Y}_{[i,:]} }{\partial \mathbf{Y}_{[i,:]} } + \mathbf{Y}_{[i,:]} \frac{\partial 1/\|\mathbf{Y}_{[i,:]} \|}{\partial \mathbf{Y}_{[i,:]} } \quad (25)$$

$$= \frac{1}{\|\mathbf{Y}_{[i,:]} \|} \mathbf{I}_D - \mathbf{Y}_{[i,:]} \frac{\mathbf{Y}_{[i,:]}^\top}{\|\mathbf{Y}_{[i,:]} \|^3} \quad (26)$$

$$= \frac{1}{\|\mathbf{Y}_{[i,:]} \|} (\mathbf{I}_D - \frac{\mathbf{Y}_{[i,:]} \mathbf{Y}_{[i,:]}^\top}{\|\mathbf{Y}_{[i,:]} \|^2}). \quad (27)$$

796 \square

797 **Lemma A.3.** Let $\mathbf{Y} \in \mathbb{R}^{S \times D}$ be an input matrix. Then, the Jacobian of RMSNorm with respect to
798 \mathbf{Y} is given by

$$\frac{\partial \text{RMSNorm}(\mathbf{Y})}{\partial \mathbf{Y}} = \text{blockdiag} \left(\left\{ \frac{1}{\|\mathbf{Y}_{[i,:]} \|} \text{diag}(\boldsymbol{\gamma}) (\mathbf{I}_D - \frac{\mathbf{Y}_{[i,:]} \mathbf{Y}_{[i,:]}^\top}{\|\mathbf{Y}_{[i,:]} \|^2}) \right\}_{i=1}^S \right) \quad (28)$$

799 *Proof.* Since RMSNorm is expressed as

$$\text{RMSNorm}(\mathbf{Y})_{[i,:]} = \text{diag}(\boldsymbol{\gamma}) \Pi(\mathbf{Y})_{[i,:]}, \quad (29)$$

800 the result follows from lemma A.2 \square

801 A.2 Proof of Proposition 4.1

802 **Proposition 4.1** is restated.

803 Consider the continuous-time dynamics for single-head SA equipped with projection (3). The energy
804 function

$$E_{\text{single}}(\mathbf{X}) = - \sum_{i,j} \exp \left(\beta \mathbf{X}_{[i,:]}^\top \mathbf{W}^Q \mathbf{W}^{K^\top} \mathbf{X}_{[j,:]} \right) \quad (30)$$

805 is monotonically decreasing as $dE_{\text{single}}(\mathbf{X})/dt \leq 0$ under the condition:

$$\mathbf{W}^V = (\mathbf{W}^{Q^\top} \mathbf{W}^K + \mathbf{W}^Q \mathbf{W}^{K^\top})/2. \quad (31)$$

806 *Proof.* Let $\Delta = \text{softmax}(\beta \mathbf{X} \mathbf{W}^Q \mathbf{W}^{K\top} \mathbf{X}^\top) \mathbf{X} \mathbf{W}^V$ and let $\mathbf{A} = \mathbf{W}^Q \mathbf{W}^{K\top}$. The first-order
 807 derivative of $E_{\text{single}}(\mathbf{X})$ with respect to $\mathbf{X}_{[i,:]}$ is:

$$\frac{dE_{\text{single}}(\mathbf{X})}{d\mathbf{X}_{[i,:]}} \quad (32)$$

$$= - \sum_{j \neq i} \frac{d \exp(\beta \mathbf{X}_{[i,:]}^\top \mathbf{A} \mathbf{X}_{[j,:]})}{d\mathbf{X}_{[i,:]}} - \sum_{j \neq i} \frac{d \exp(\beta \mathbf{X}_{[j,:]}^\top \mathbf{A} \mathbf{X}_{[i,:]})}{d\mathbf{X}_{[i,:]}} - \frac{d \exp(\beta \mathbf{X}_{[i,:]}^\top \mathbf{A} \mathbf{X}_{[i,:]})}{d\mathbf{X}_{[i,:]}} \quad (33)$$

$$= - \sum_{j \neq i} \exp(\beta \mathbf{X}_{[i,:]}^\top \mathbf{A} \mathbf{X}_{[j,:]}) \mathbf{A}^\top \mathbf{X}_{[j,:]} - \sum_{j \neq i} \exp(\beta \mathbf{X}_{[j,:]}^\top \mathbf{A} \mathbf{X}_{[i,:]}) \mathbf{A} \mathbf{X}_{[j,:]} \quad (34)$$

$$- \exp(\beta \mathbf{X}_{[i,:]}^\top \mathbf{A} \mathbf{X}_{[i,:]}) (\mathbf{A}^\top + \mathbf{A}) \mathbf{X}_{[i,:]} \quad (35)$$

$$= - \sum_j \exp(\beta \mathbf{X}_{[i,:]}^\top \mathbf{A} \mathbf{X}_{[j,:]}) (\mathbf{A}^\top + \mathbf{A}) \mathbf{X}_{[j,:]} \quad (36)$$

808 Under the given condition on the weights, $\mathbf{A} = \mathbf{A}^\top$ we have:

$$\Delta_{[i,:]} = (\text{softmax}(\beta \mathbf{X} \mathbf{W}^V \mathbf{X}^\top)_{[i,:]} \mathbf{X} (\mathbf{A}^\top + \mathbf{A}))^\top / 2 \quad (37)$$

$$= \sum_j \frac{\exp(\beta \mathbf{X}_{[i,:]}^\top \mathbf{A} \mathbf{X}_{[j,:]})}{Z_i} (\mathbf{A}^\top + \mathbf{A}) \mathbf{X}_{[j,:]} / 2, \quad (38)$$

809 where $Z_i = \sum_{j'} \exp(\beta \mathbf{X}_{[i,:]}^\top \mathbf{A} \mathbf{X}_{[j',:]})$ is the normalization term.

810 Thus, we have,

$$\frac{dE_{\text{single}}(\mathbf{X})}{dt} \quad (39)$$

$$= \frac{dE_{\text{single}}(\mathbf{X})}{d\mathbf{X}} \cdot \frac{d\mathbf{X}}{dt} \quad (40)$$

$$= - \sum_i \left(\frac{dE_{\text{single}}(\mathbf{X})}{d\mathbf{X}_{[i,:]}} \cdot (\mathbf{I}_D - \mathbf{X}_{[i,:]} \mathbf{X}_{[i,:]}^\top) \Delta_{[i,:]} \right) \quad (41)$$

$$= - \sum_i \left(\sum_j \exp(\beta \mathbf{X}_{[i,:]}^\top \mathbf{A} \mathbf{X}_{[j,:]}) (\mathbf{A}^\top + \mathbf{A}) \mathbf{X}_{[j,:]} \cdot (\mathbf{I}_D - \mathbf{X}_{[i,:]} \mathbf{X}_{[i,:]}^\top) \Delta_{[i,:]} \right) \quad (42)$$

$$= -2 \sum_i \left(Z_i \Delta_{[i,:]} \cdot (\mathbf{I}_D - \mathbf{X}_{[i,:]} \mathbf{X}_{[i,:]}^\top) \Delta_{[i,:]} \right) \quad (43)$$

$$= -2 \sum_i \left(Z_i \Delta_{[i,:]}^\top (\mathbf{I}_D - \mathbf{X}_{[i,:]} \mathbf{X}_{[i,:]}^\top) \Delta_{[i,:]} \right) \quad (44)$$

$$\leq 0, \quad (45)$$

811 where, in the last inequality, we used the fact that the matrix $\mathbf{I}_D - \mathbf{X}_{[i,:]} \mathbf{X}_{[i,:]}^\top$ is positive semi-
 812 definite. \square

813 A.3 Proof of Proposition 4.2

814 **Proposition 4.2 is restated.** Consider the continuous-time dynamics for multi-head SA without
 815 projection: $d\mathbf{X}/dt = \sum_{h=1}^H S\mathbf{A}_h(\mathbf{X})$. An energy function

$$E_{\text{multi}}(\mathbf{X}) = - \sum_h \sum_{i,j} \exp(\beta \mathbf{X}_{[i,:]}^\top \mathbf{W}_h^Q \mathbf{W}_h^{K\top} \mathbf{X}_{[j,:]}) \quad (46)$$

816 is monotonically decreasing as $dE_{\text{multi}}(\mathbf{X})/dt \leq 0$ under the condition

$$\mathbf{W}_h^V = (\mathbf{W}_h^{Q\top} \mathbf{W}_h^K + \mathbf{W}_h^Q \mathbf{W}_h^{K\top})/2, \quad \mathbf{W}_h^Q \mathbf{W}_h^{K\top} = \mathbf{U}_{1,h} \mathbf{U}_{2,h}^\top, \quad (47)$$

817 where $\mathbf{U}_{1(2),h} \in \mathbb{R}^{D \times D/(2H)}$ ($h \in [1, H]$) satisfies the orthogonality condition $\mathbf{U}_{k,h}^\top \mathbf{U}_{k',h'} =$
 818 $\delta_{hh'} \delta_{kk'} \mathbf{I}_{D/(2H)}$.

819 *Proof.* Let $\Delta_h = \text{softmax}(\beta \mathbf{X} \mathbf{W}_h^Q \mathbf{W}_h^{K\top} \mathbf{X}^\top) \mathbf{X} \mathbf{W}_h^V$ and let $\mathbf{A}_h = \mathbf{W}_h^Q \mathbf{W}_h^{K\top}$. The first-order
 820 derivative of $E_{\text{multi}}(\mathbf{X})$ with respect to $\mathbf{X}_{[i,:]}$ is, similarly to the single-head case, given by:

$$\frac{\partial E_{\text{multi}}(\mathbf{X})}{\partial \mathbf{X}_{[i,:]}} = - \sum_h \sum_j \exp(\beta \mathbf{X}_{[i,:]}^\top \mathbf{A}_h \mathbf{X}_{[j,:]}) (\mathbf{A}_h^\top + \mathbf{A}_h) \mathbf{X}_{[j,:]} \quad (48)$$

821 Under the given condition on the weights, similar to the single-head case, we have:

$$\Delta_{h[i,:]} = \sum_j \frac{\exp(\beta \mathbf{X}_{[i,:]}^\top \mathbf{A}_h \mathbf{X}_{[j,:]})}{Z_{h,i}} (\mathbf{A}_h^\top + \mathbf{A}_h) \mathbf{X}_{[j,:]} / 2 \quad (49)$$

822 where $Z_{h,i} = \sum_{j'} \exp(\beta \mathbf{X}_{[i,:]}^\top \mathbf{A}_h \mathbf{X}_{[j',:]})$ is the normalization term. Thus, we have,

$$\frac{dE_{\text{multi}}(\mathbf{X})}{dt} \quad (50)$$

$$= \frac{dE_{\text{multi}}(\mathbf{X})}{d\mathbf{X}} \cdot \frac{d\mathbf{X}}{dt} \quad (51)$$

$$= - \sum_i \left(\frac{\partial E_{\text{multi}}(\mathbf{X})}{\partial \mathbf{X}_{[i,:]}} \cdot \sum_h \Delta_{h[i,:]} \right) \quad (52)$$

$$= - \sum_i \left(\sum_h \sum_j \exp(\beta \mathbf{X}_{[i,:]}^\top \mathbf{A}_h \mathbf{X}_{[j,:]}) (\mathbf{A}_h^\top + \mathbf{A}_h) \mathbf{X}_{[j,:]} \cdot \sum_h \Delta_{h[i,:]} \right) \quad (53)$$

$$= -2 \sum_i \left(\sum_h Z_{h,i} \Delta_{h[i,:]} \cdot \sum_h \Delta_{h[i,:]} \right) \quad (54)$$

$$= -2 \sum_i \sum_h (Z_{h,i} \Delta_{h[i,:]} \cdot \Delta_{h[i,:]}) \quad (55)$$

$$= -2 \sum_i \sum_h Z_{h,i} \|\Delta_{h[i,:]\|}^2 \quad (56)$$

$$\leq 0. \quad (57)$$

823 In the third-to-last line, we use the fact that for $h \neq h'$,

$$\mathbf{A}_h^\top \mathbf{A}_{h'} = \mathbf{U}_{2,h} \mathbf{U}_{1,h}^\top \mathbf{U}_{1,h'} \mathbf{U}_{2,h'}^\top = \mathbf{O} \quad (58)$$

$$\mathbf{A}_h \mathbf{A}_{h'} = \mathbf{U}_{1,h} \mathbf{U}_{2,h}^\top \mathbf{U}_{1,h'} \mathbf{U}_{2,h'}^\top = \mathbf{O}, \quad (59)$$

824 and thus

$$\Delta_{h[i,:]} \cdot \Delta_{h'[i,:]} \quad (60)$$

$$= \Delta_{h[i,:]}^\top \Delta_{h'[i,:]} \quad (61)$$

$$= \left((\mathbf{A}_h + \mathbf{A}_h^\top) \sum_j \frac{\exp(\beta \mathbf{X}_{[i,:]}^\top \mathbf{A}_h \mathbf{X}_{[j,:]})}{Z_{h,i}} \mathbf{X}_{[j,:]} \right)^\top \quad (62)$$

$$(\mathbf{A}_{h'} + \mathbf{A}_{h'}^\top) \sum_j \frac{\exp(\beta \mathbf{X}_{[i,:]}^\top \mathbf{A}_{h'} \mathbf{X}_{[j,:]})}{Z_{h',i}} \mathbf{X}_{[j,:]} / 4 \quad (63)$$

$$= \left(\sum_j \frac{\exp(\beta \mathbf{X}_{[i,:]}^\top \mathbf{A}_h \mathbf{X}_{[j,:]})}{Z_{h,i}} \mathbf{X}_{[j,:]} \right)^\top \quad (64)$$

$$(\mathbf{A}_h^\top + \mathbf{A}_h) (\mathbf{A}_{h'} + \mathbf{A}_{h'}^\top) \sum_j \frac{\exp(\beta \mathbf{X}_{[i,:]}^\top \mathbf{A}_{h'} \mathbf{X}_{[j,:]})}{Z_{h',i}} \mathbf{X}_{[j,:]} / 4 \quad (65)$$

$$= 0. \quad (66)$$

825 \square

826 A.4 Proof of Proposition 5.1

827 **Proposition 5.1 is restated.** Suppose that, in the update of ItrSA (9), the input to the normalization
 828 layer satisfies $\|\mathbf{X}_{[i,:]} + \eta\Delta\mathbf{X}_{[i,:]}\| \geq R$ for all $i \in [1, S]$. Then, the spectral norm of the Jacobian
 829 satisfies the upper bound

$$\left\| \frac{\partial \text{RMSNorm}(\mathbf{X} + \eta\Delta\mathbf{X})}{\partial \mathbf{X}} \right\|_2 \leq \frac{\max_j(|\gamma_j|)}{R} (1 + |\eta| \|\mathbf{J}_{\text{MSA}}(\mathbf{X})\|_2), \quad (67)$$

830 where $\mathbf{J}_{\text{MSA}}(\mathbf{X}) := \partial \text{MSA}(\mathbf{X}) / \partial \mathbf{X}$ denotes the Jacobian of MSA.

831 *Proof.* First, for any vector $a \in \mathbb{R}^D$, the eigenvalues of the matrix $I_D - \frac{aa^\top}{\|a\|^2}$ are 1 (with multiplicity
 832 $D - 1$) and 0 (with multiplicity 1). Hence,

$$\left\| I_D - \frac{aa^\top}{\|a\|^2} \right\|_2 \leq 1. \quad (68)$$

833 Using Lemma A.3 we have

$$\left\| \frac{\partial \text{RMSNorm}^{(\text{osc})}(\mathbf{Y})}{\partial \mathbf{Y}} \right\|_2 = \left\| \text{blockdiag} \left(\left\{ \frac{1}{\|\tilde{\mathbf{Y}}_{i,j}\|} \text{diag}(\gamma) \left(I_D - \frac{\tilde{\mathbf{Y}}_{i,j} \tilde{\mathbf{Y}}_{i,j}^\top}{\|\tilde{\mathbf{Y}}_{i,j}\|^2} \right) \right\}_{i,j} \right) \right\|_2 \quad (69)$$

$$= \max_i \left\| \frac{1}{\|\mathbf{Y}_{[i,:]}\|} \text{diag}(\gamma) \left(I_D - \frac{\mathbf{Y}_{[i,:]} \mathbf{Y}_{[i,:]}^\top}{\|\mathbf{Y}_{[i,:]\|^2} \right) \right\|_2 \quad (70)$$

$$\leq \frac{\max_j |\gamma_j|}{R}. \quad (71)$$

834 Therefore, setting $\mathbf{Y} = \mathbf{X} + \eta\Delta\mathbf{X} = \mathbf{X} + \eta(\mathbf{C} + \text{MSA}(\mathbf{X}))$, we have

$$\left\| \frac{\partial \text{RMSNorm}(\mathbf{X} + \eta\Delta\mathbf{X})}{\partial \mathbf{X}} \right\|_2 \quad (72)$$

$$= \left\| \frac{\partial \text{RMSNorm}(\mathbf{Y})}{\partial \mathbf{Y}} \frac{\partial \mathbf{Y}}{\partial \mathbf{X}} \right\|_2 \quad (73)$$

$$\leq \left\| \frac{\partial \text{RMSNorm}(\mathbf{Y})}{\partial \mathbf{Y}} \right\|_2 \left\| \frac{\partial \mathbf{Y}}{\partial \mathbf{X}} \right\|_2 \quad (74)$$

$$\leq \frac{\max_j |\gamma_j|}{R} \|\mathbf{I}_{SD} + \eta \mathbf{J}_{\text{MSA}}\|_2 \quad (75)$$

$$\leq \frac{\max_j (|\gamma_j|)}{R} (1 + |\eta| \|\mathbf{J}_{\text{MSA}}(\mathbf{X})\|_2). \quad (76)$$

835 \square

836 A.5 Derivation of the eigenvalue bound in oscillatory cases (Section 5.2)

837 We show that all eigenvalues λ_j of the Jacobian

$$\mathbf{J}(x) = \left(I_D - \frac{\mathbf{y}\mathbf{y}^\top}{\|\mathbf{y}\|^2} \right) \frac{1}{\|\mathbf{y}\|} (I_D + \eta\Omega), \quad (77)$$

838 satisfy $|\lambda_j| \leq 1$, where $\mathbf{y} = (I_D + \eta\Omega)x$.

839 We begin by computing the norm of \mathbf{y} :

$$\|\mathbf{y}\|^2 = \|(I_D + \eta\Omega)x\|^2 \quad (78)$$

$$= x^\top (I_D + \eta\Omega^\top) (I_D + \eta\Omega) x \quad (79)$$

$$= x^\top x + \eta x^\top \Omega x + \eta x^\top \Omega^\top x + \eta^2 x^\top \Omega^\top \Omega x \quad (80)$$

$$= 1 + \eta^2 \|\Omega x\|^2, \quad (81)$$

where we used the fact that for an antisymmetric matrix Ω , $\mathbf{x}^\top \Omega \mathbf{x} = 0$. Note also that an antisymmetric matrix has eigenvalues of the form $\pm i\omega_j$, where $\omega_j \geq 0$ ($j = 1, 2, \dots$). For simplicity, assume all eigenvalues have identical magnitude $\omega_j = \omega$. Then, we have

$$\|\mathbf{y}\|^2 = 1 + \eta^2 \omega^2. \quad (82)$$

We also use the facts that

$$\left\| \mathbf{I}_D - \frac{\mathbf{y}\mathbf{y}^\top}{\|\mathbf{y}\|^2} \right\|_2 \leq 1 \quad (83)$$

and

$$\|\mathbf{I}_D + \eta\Omega\|_2 = |1 \pm i\eta\omega| = \sqrt{1 + \eta^2 \omega^2}. \quad (84)$$

Combining these, we obtain the following bound on the spectral norm of $\mathbf{J}(\mathbf{x})$:

$$\|\mathbf{J}(\mathbf{x})\|_2 = \left\| \left(\mathbf{I}_D - \frac{\mathbf{y}\mathbf{y}^\top}{\|\mathbf{y}\|^2} \right) \frac{1}{\|\mathbf{y}\|} (\mathbf{I}_D + \eta\Omega) \right\|_2 \quad (85)$$

$$\leq \frac{1}{\|\mathbf{y}\|} \|\mathbf{I}_D + \eta\Omega\|_2 \quad (86)$$

$$\leq \frac{\sqrt{1 + \eta^2 \omega^2}}{\sqrt{1 + \eta^2 \omega^2}} = 1. \quad (87)$$

This implies that all eigenvalues of $\mathbf{J}(\mathbf{x})$ satisfy $|\lambda_j| \leq 1$.

B Experimental details

B.1 Experimental setup

We solved Sudoku task, which is a puzzle played on a 9×9 grid, where some of the cells are pre-filled with digits from 1 to 9, and the remaining cells are left blank. The objective is to fill in the blank cells such that each 1) row, 2) column, and 3) 3×3 subgrid contains each digit exactly once.

In our experiments, we used two Sudoku datasets: the SATNet [Wang et al., 2019] and RRN dataset [Palm et al., 2018]. The key differences between the two are that the RRN dataset is more difficult (with only 17–34 given digits compared to 31–42 in SATNet) and larger in size (198k samples vs. 10k samples). Following Miyato et al. [2025], we used the SATNet dataset for training as in-distribution (ID) data and the RRN dataset as out-of-distribution (OOD) data. This setup allows us to evaluate the ability of models to generalize to more challenging settings.

We primarily followed Miyato et al. [2025] and used their official implementation¹ using $N = 4$ as the dimension of oscillators. We used the Adam optimizer [Kingma and Ba, 2015] and trained for 100 epochs with batch size 100. We tuned the learning rate across $\{1 \times 10^{-6}, 5 \times 10^{-6}, \dots, 1 \times 10^{-3}\}$ for all settings based on the OOD accuracy at the iteration $T = 16$. All experiments were conducted on NVIDIA H200 GPUs, and we run experiments with 5 different random seeds.

We also conducted experiments on the CIFAR-10 dataset [Krizhevsky et al.]. See table S.2 for training and model configurations.

B.2 Single-head generalized symmetric SA

For single-head generalized symmetric SA, we define

$$R_{\text{E-single}} := \|\mathbf{W}_1^V \mathbf{W}_1^O - (\mathbf{W}_1^V \mathbf{W}_1^O)^\top\|_F^2, \quad (88)$$

under the condition that $H = 1$. If $R_{\text{E-single}} = 0$, setting $\mathbf{W}^V = \mathbf{W}_1^V \mathbf{W}_1^O$ satisfies the condition on \mathbf{W}^V described in Proposition 4.1

¹<https://github.com/autonomousvision/akorn>

869 B.3 Lyapunov exponent

870 The Lyapunov exponent quantifies the exponential rate at which nearby trajectories in a dynamical
 871 system diverge. For a discrete-time system $\mathbf{x}^{(t+1)} = \mathbf{f}(\mathbf{x}^{(t)})$, the Lyapunov spectrum $\{\lambda_i\}$ is defined
 872 as:

$$\lambda_i = \lim_{T \rightarrow \infty} \frac{1}{2T} \log |\alpha_i^{(T)}|, \quad (89)$$

873 where $\alpha_i^{(T)}$ is the i -th eigenvalue of the positive semi-definite matrix

$$\Lambda^{(T)} = \left(\frac{d\mathbf{f}^T(\mathbf{x}^{(0)})}{d\mathbf{x}^{(0)}} \right)^\top \frac{d\mathbf{f}^T(\mathbf{x}^{(0)})}{d\mathbf{x}^{(0)}}, \quad (90)$$

874 and \mathbf{f}^T denotes the T -fold composition of the function \mathbf{f} . The *maximum Lyapunov exponent* is then
 875 defined as

$$\lambda_{\max} := \max_i \lambda_i. \quad (91)$$

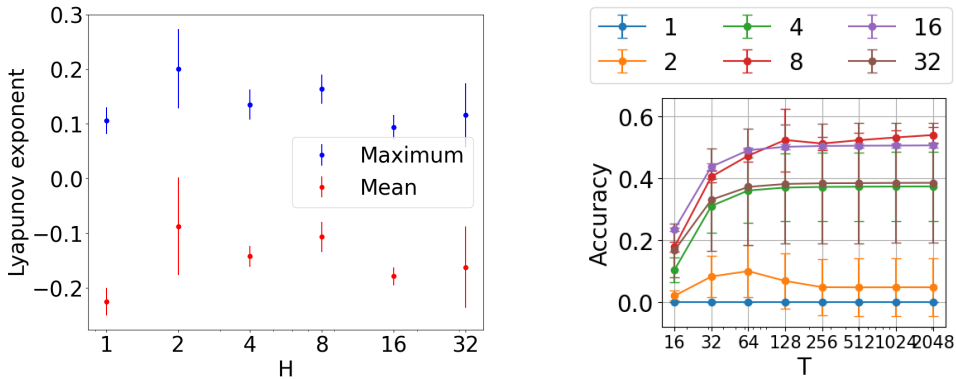
876 In our experiments, we approximated the Lyapunov spectrum using a finite time horizon of $T = 16$
 877 on a randomly selected sample. For models without normalization and symmetric SA, we trained
 878 them for only one epoch, as full training was not feasible due to instability. To evaluate how the
 879 Lyapunov exponent varies, we adjusted the input scaling of \mathbf{X} , the step size η , and the norms of the
 880 value projection weights, $\|\mathbf{W}_h^V\|$ and $\|\mathbf{W}_h^O\|$, in the SA update of \mathbf{X} .

881 B.4 Details of other figures

882 For Figure 2c, we computed the eigenvalues of the Jacobian matrix at $T = 16$ on a randomly selected
 883 sample from the Sudoku dataset. For the model with normalization, we used the fully trained model.
 884 For models without normalization, we followed the same setup as in the Lyapunov experiments and
 885 used models trained for only one epoch.

886 For the computation of the SA’s Jacobian in Figure 1, we used the CCDV arXiv summarization
 887 dataset [Cohan et al., 2018], as it provides text data suitable for varying the number of tokens. We used
 888 an initialized SA and computed the Jacobian and SA followed by normalization over 500 randomly
 889 selected samples. The norm of tokens was set to $R = 100$ and their dimensions to $D = 256$.

890 B.5 Additional results



(a) Number of attention heads H vs. Lyapunov exponent in ItrSA with normalization.

(b) Number of attention heads H and OOD accuracy on the Sudoku dataset.

Figure S.1: Effect of the number of attention heads H on Lyapunov exponent and OOD accuracy in the Sudoku dataset.

891 **The number of attention heads.** To further investigate the effect of normalization, we calculated the
 892 Lyapunov exponents while varying the number of attention heads. The results in Figure S.1a indicate

that the number of heads has little to no impact on the Lyapunov exponents. Figure S.1b also shows the OOD test accuracy for models with different numbers of attention heads. The results indicate that models with a small number of heads ($H = 1, 2$) exhibit poor performance, while models with $H = 8$ or 16 achieve the highest accuracy.

γ in RMSNorm. Table S.1 presents the values of the γ parameter learned by ItrSA. The results indicate that the trained models exhibit small γ values, with $\max_j |\gamma_j| < 1$.

Table S.1: γ in RMSNorm with different N .

N	$\ \gamma\ $	$\max_j \gamma_j $
4	0.229 ± 0.0004	0.229 ± 0.0004
8	0.031 ± 0.0017	0.052 ± 0.0000
16	0.098 ± 0.0062	0.098 ± 0.0062
32	0.348 ± 0.0004	0.348 ± 0.0004
64	0.489 ± 0.0006	0.489 ± 0.0006
128	0.841 ± 0.0000	0.841 ± 0.0000
256	0.738 ± 0.0004	0.738 ± 0.0004
512	0.811 ± 0.0002	0.811 ± 0.0002

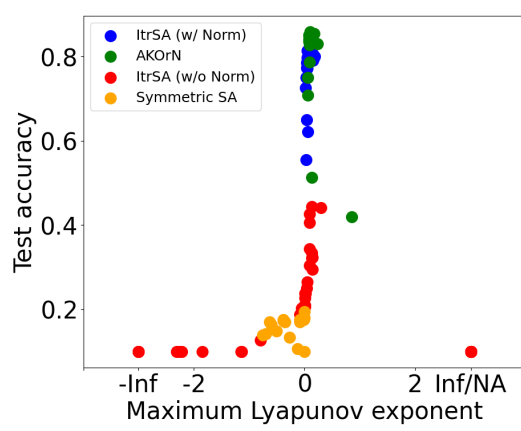
Table S.2: Training and model configurations.

Parameter	Sudoku	CIFAR-10
Hidden dimension D	512	384
Number of heads H	8	8
Initial value of η	1.0	1.0
Batch size	100	128
Number of epochs	100	200

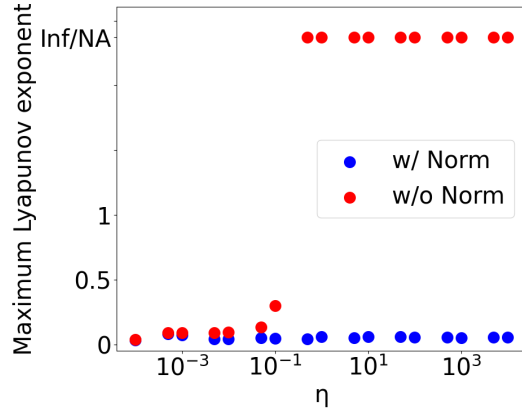
B.6 CIFAR-10

For the experiments on the CIFAR-10 dataset [Krizhevsky et al.], we used the same architecture and setup as in the Sudoku experiments. We used the Adam optimizer and tuned the learning rate across $\{1 \times 10^{-6}, 5 \times 10^{-6}, \dots, 1 \times 10^{-3}\}$ based on the test accuracy at the iteration $T = 16$. We trained models for 200 epochs and set the batch size 128.

Figure S.2 shows the Lyapunov exponent on the CIFAR-10 dataset. This result is in the same trend with that on the Sudoku dataset.



(a) Maximum Lyapunov exponent vs. test accuracy



(b) Maximum Lyapunov exponent with and without normalization

Figure S.2: Lyapunov exponent on the CIFAR-10 dataset.

C Other details

C.1 Extended related work

Energy-based understanding. The Transformer architecture has been a focus of efforts to provide theoretical grounding. [Geshkovski et al. \[2023a,b, 2024\]](#) formulated recurrent SA dynamics as interactions among tokens (“particles”), enabling theoretical analysis of phenomena such as meta-stable clustering and rank-one collapse. Their continuous-time dynamics monotonically decrease an energy (Lyapunov) function given by a summation over exponential functions, commonly requiring constraints such as single-head attention, hyperspherical token states, and symmetric weights. [Karagodin et al. \[2024\]](#) extended this framework to the case of causal attention masking. [Bruno et al. \[2025\]](#) succeeded in mathematically characterizing the meta-stable clustering as a Wasserstein gradient flow of mean-field token dynamics, with the energy serving as its potential function, although they replaced the softmax function with an unnormalized exponential function and restricted their analysis to identity weight matrices. [Yang et al. \[2022\]](#) considered an exponential energy function similar to that of [Geshkovski et al. \[2023a\]](#), describing the Transformer as performing alternating majorization-minimization updates on distinct energy functions. Their approach also accommodates discrete state updates and MLP layers, although it entails complex conditions, including constraints on step sizes and proximity to fixed points. [Ramsauer et al. \[2021\]](#) formalized the cross-attention mechanism as modern Hopfield networks. [Hoover et al. \[2023\]](#), [Hu et al. \[2025\]](#) further developed energy functions for Transformers including self-attentions. We do not address approaches based on Hopfield networks in this work, as they require architectural modifications, such as adding auxiliary signal paths that are absent in standard Transformers, which are beyond our scope.

Jacobian-based analysis. The Jacobian of state updates is fundamental for characterizing neural network dynamics. For example, it has been used to analyze edge-of-chaos behavior for stable signal propagation and gradient control [\[Boedecker et al. 2012, Poole et al. 2016, Pennington et al. 2017\]](#). [Haber and Ruthotto \[2017\]](#) interpreted forward propagation in neural networks as continuous-time dynamical systems and analyzed their Jacobians to prevent exploding and vanishing gradients. [Chang et al. \[2019\]](#) extended the ODE-based perspective to recurrent neural networks and proposed using anti-symmetric weight matrices to satisfy discrete-time stability conditions. Several studies have explored Jacobian-based regularization techniques. [Yoshida and Miyato \[2017\]](#) proposed spectral norm regularization to reduce sensitivity to input perturbations and improve generalization. [Miyato et al. \[2018\]](#) applied spectral normalization to stabilize the training of generative adversarial networks. [Lewandowski et al. \[2025\]](#) introduced spectral regularization for continual learning, aiming to prevent the loss of plasticity and maintain trainability across tasks by keeping the maximum singular value of each layer close to one. Regarding SA specifically, [Noci et al. \[2022\]](#) analyzed Jacobians to explain rank collapse, while [Castin et al. \[2024\]](#) evaluated their spectral properties mathematically. In this work, we use Jacobian analysis to understand inference dynamics in realistic SAs and also employ them as regularizers and performance indicators.

Looped architectures. Looped architectures in Transformers have been explored since their introduction by [Dehghani et al. \[2018\]](#). One example is weight tying, as seen in the ALBERT model [\[Lan et al. 2020\]](#). Equilibrium models [\[Bai et al. 2019\]](#) use fixed-point solutions, which can be interpreted as infinitely looped computations. [Yang et al. \[2024\]](#), [Giannou et al. \[2023\]](#) showed that Transformers with looped structures are capable of learning algorithmic tasks. [Saunshi et al. \[2025\]](#) further showed that looped architectures enhance reasoning ability through strong inductive bias. As the number of recurrent updates (i.e., loops) increases, performance scales efficiently, a phenomenon we refer to as test-time scaling. [Geiping et al. \[2025\]](#) successfully applied test-time scaling to reasoning benchmarks, and [Bansal et al. \[2022\]](#) showed that it enables models to solve problems at test time that are more difficult than those seen during training. [Miyato et al. \[2025\]](#) proposed artificial Kuramoto oscillatory neurons (AKOrN), a looped architecture that successfully solves tasks in a neuroscience-inspired manner, demonstrating strong empirical results in unsupervised object discovery, adversarial robustness, calibrated uncertainty quantification, and reasoning.

956 **C.2 Details of preliminaries**

957 **Energy-based analysis by** [Yang et al. \[2022\]](#) [Yang et al. \[2022\]](#) formalized updates of SA using
 958 alternating inexact minimization algorithm as:

$$\mathbf{X}^{(t+1)} = \text{softmax}_\beta(\mathbf{X}^{(t)} \mathbf{W}^s \mathbf{X}^{(t)\top}) \mathbf{X}^{(t)} \mathbf{W}^s, \quad (92)$$

959 where $\mathbf{W}^s \in \mathbb{R}^{D \times D}$ is a symmetric matrix and softmax_β is a function reweighted with coefficient
 960 vector β .

961 **Operation on oscillators** We use $\widetilde{\mathbf{X}}_{i,j}$ to refer to the j -th oscillator of the i -th token of X , which
 962 is defined as $\widetilde{\mathbf{X}}_{i,j} = \mathbf{X}_{[i,(j-1)N+1:jN]} \in \mathbb{R}^N$. They are defined as:

$$\widetilde{\text{Omg}}^{(\text{osc})}(\mathbf{X}^{(t)})_{i,j} = \Omega_j \widetilde{\mathbf{X}}_{i,j}, \quad \widetilde{\text{Proj}}_X^{(\text{osc})}(\mathbf{Y})_{i,j} = \left(I_N - \widetilde{\mathbf{X}}_{i,j} \widetilde{\mathbf{X}}_{i,j}^\top \right) \widetilde{\mathbf{Y}}_{i,j}, \quad \widetilde{\Pi}^{(\text{osc})}(\mathbf{Y})_{i,j} = \frac{\widetilde{\mathbf{Y}}_{i,j}}{\|\widetilde{\mathbf{Y}}_{i,j}\|}, \quad (93)$$

963 AKOrN then uses a readout module to read out patterns independent of the phase.

$$\mathbf{C}' = \mathbf{g}(\mathbf{m}) \in \mathbb{R}^{D \times N}, m_k = \|\mathbf{z}_k\|, \mathbf{z}_k = \sum_i U_{kij} \widetilde{\mathbf{X}}_{i,j} \in \mathbb{R}^{N'}, \quad (94)$$

964 where $U_{kij} \in \mathbb{R}^{N' \times N}$ is a learned weight matrix, \mathbf{g} is a learned function and $k = 1 \cdots DN$.