# Why Adversarial Interaction Creates Non-Homogeneous Patterns: A Pseudo-Reaction-Diffusion Model for Turing Instability

Litu Rout

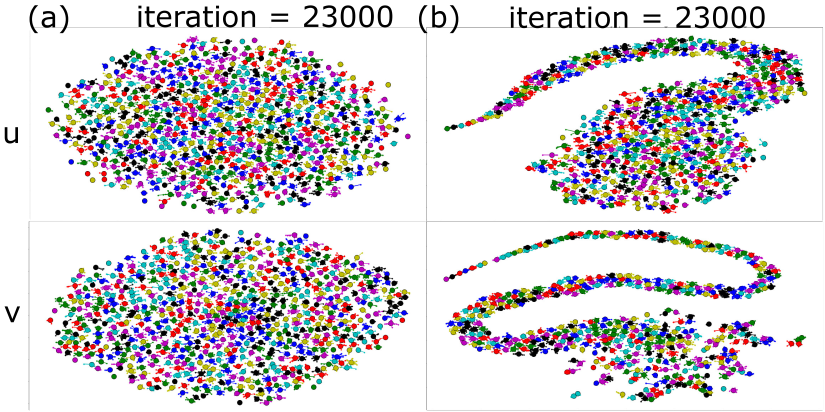
Space Applications Centre
Indian Space Research Organisation

*lr@sac.isro.gov.in*

January 6, 2021

# Overview

Supervised Learning vs Regularized Adversarial Learning



# Introduction

- Symmetry and homogeneity

# Introduction

- Symmetry and homogeneity
- Breakdown of symmetry and homogeneity

# Introduction

- Symmetry and homogeneity
- Breakdown of symmetry and homogeneity
- **Root cause: Turing instability**

# Introduction

- Symmetry and homogeneity
- Breakdown of symmetry and homogeneity
- **Root cause: Turing instability**

### Hypothesis

*A system in which a generator and a discriminator adversarially interact with each other exhibits Turing-like patterns in the hidden layer and top layer of a two layer generator network with ReLU activation.*

# Objectives

- Does it converge? If so, under what circumstance?

# Objectives

- Does it converge? If so, under what circumstance?
- Why do non-homogeneous patterns emerge?

# Objectives

- Does it converge? If so, under what circumstance?
- Why do non-homogeneous patterns emerge?
- Why is it important to study such patterns?

# Problem Setup

Consider $n$ i.i.d. training samples: $\{(\boldsymbol{x}_p, \boldsymbol{y}_p)\}_{p=1}^{n} \subset \mathbb{R}^{d_{in}} \times \mathbb{R}^{d_{out}}$.

Two layer network with ReLU activation ($\sigma(.)$):

$$f(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{x}) = \frac{1}{\sqrt{d_{out} m}} \boldsymbol{V} \sigma(\boldsymbol{U}\boldsymbol{x}). \tag{1}$$

Here, $\boldsymbol{U} \in \mathbb{R}^{m \times d_{in}}$ and $\boldsymbol{V} \in \mathbb{R}^{d_{out} \times m}$.

Let input data points be represented by $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) \in \mathbb{R}^{d_{in} \times n}$ and corresponding labels by $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n) \in \mathbb{R}^{d_{out} \times n}$.

# Problem Setup

Supervised learning:

$$\mathcal{L}_{sup}\left(\boldsymbol{U},\boldsymbol{V}\right) = \frac{1}{2}\sum_{p=1}^{n}\left\|\frac{1}{\sqrt{d_{out}m}}\boldsymbol{V}\sigma\left(\boldsymbol{U}\boldsymbol{x}_p\right) - \boldsymbol{y}_p\right\|_2^2 = \frac{1}{2}\left\|\frac{1}{\sqrt{d_{out}m}}\boldsymbol{V}\sigma\left(\boldsymbol{U}\boldsymbol{X}\right) - \boldsymbol{Y}\right\|_F^2. \quad (2)$$

Regularized adversarial learning:

$$\mathcal{L}_{aug}\left(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\boldsymbol{a}\right) = \underbrace{\frac{1}{2}\left\|\frac{1}{\sqrt{d_{out}m}}\boldsymbol{V}\sigma\left(\boldsymbol{U}\boldsymbol{X}\right) - \boldsymbol{Y}\right\|_F^2}_{\mathcal{L}_{sup}} \underbrace{- \frac{1}{m\sqrt{d_{out}}}\sum_{p=1}^{n}\boldsymbol{a}^T\sigma\left(\boldsymbol{W}\boldsymbol{V}\sigma\left(\boldsymbol{U}\boldsymbol{x}_p\right)\right)}_{\mathcal{L}_{adv}}. \quad (3)$$

# Learning Algorithm

Randomly initialized gradient descent:

$$\frac{du_{jk}}{dt} = -\frac{\partial \mathcal{L}_{aug}\left(\boldsymbol{U}(t), \boldsymbol{V}(t), \boldsymbol{W}(t), \boldsymbol{a}(t)\right)}{\partial u_{jk}(t)},$$
$$\frac{dv_{ij}}{dt} = -\frac{\partial \mathcal{L}_{aug}\left(\boldsymbol{U}(t), \boldsymbol{V}(t), \boldsymbol{W}(t), \boldsymbol{a}(t)\right)}{\partial v_{ij}(t)} \quad (4)$$

for $i \in [d_{out}]$, $j \in [m]$ and $k \in [d_{in}]$.

Equilibrium (ideal condition): $\frac{du_{jk}}{dt} = \frac{dv_{ij}}{dt} = 0$.

$\epsilon$-approximate equilibrium: $\left|\frac{du_{jk}}{dt}\right| < \epsilon$ and $\left|\frac{dv_{ij}}{dt}\right| < \epsilon$ for a small $\epsilon$.

# Revisiting Turing's Reaction-Diffusion Model

Governing **Reaction** ($\mathfrak{R}$) and **Diffusion** ($\mathfrak{D}$) dynamics:

$$\begin{aligned} \frac{d\boldsymbol{u}_j}{dt} &= \mathfrak{R}_j^{\boldsymbol{u}}\left(\boldsymbol{u}_j, \boldsymbol{v}_j\right) + \mathfrak{D}_j^{\boldsymbol{u}}\left(\nabla^2 \boldsymbol{u}_j\right), \\ \frac{d\boldsymbol{v}_j}{dt} &= \mathfrak{R}_j^{\boldsymbol{v}}\left(\boldsymbol{u}_j, \boldsymbol{v}_j\right) + \mathfrak{D}_j^{\boldsymbol{v}}\left(\nabla^2 \boldsymbol{v}_j\right). \end{aligned} \tag{5}$$

- Turing, A. 1952. The Chemical Basis of Morphogenesis. Phil. Trans. of the Royal Society of London. Series B, Biological Sciences 237(641): 37–72.

# Simplified Setup: Scalar Network and Training One Layer

Simplified Generator network:

$$f\left(\boldsymbol{U}, \boldsymbol{v}, \boldsymbol{x}\right) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} v_j \sigma\left(u_j^T x\right) = \frac{1}{\sqrt{m}} \boldsymbol{v}^T \sigma\left(\boldsymbol{U}\boldsymbol{x}\right). \tag{6}$$

Supervised learning:

$$\mathcal{L}_{sup}\left(\boldsymbol{U}, \boldsymbol{v}\right) = \sum_{p=1}^{n} \frac{1}{2} \left(f\left(\boldsymbol{U}, \boldsymbol{v}, \boldsymbol{x}_p\right) - y_p\right)^2 \tag{7}$$

Regularized adversarial learning: $\mathcal{L}_{aug}\left(\boldsymbol{U}, \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{a}\right)$

$$= \sum_{p=1}^{n} \frac{1}{2} \left(f\left(\boldsymbol{U}, \boldsymbol{v}, \boldsymbol{x}_p\right) - y_p\right)^2 - \frac{1}{\sqrt{m}} \sum_{p=1}^{n} \boldsymbol{a}^T \sigma\left(\boldsymbol{w}\left(f\left(\boldsymbol{U}, \boldsymbol{v}, \boldsymbol{x}_p\right)\right)\right) \tag{8}$$

# Warm-up: Reaction Without Diffusion

**Definition 1.** (Du et al., 2018) *Define Gram matrix $\mathcal{H}^{\infty} \in \mathcal{R}^{n \times n}$. Each entry of $\mathcal{H}^{\infty}$ is computed by $\mathcal{H}_{ij}^{\infty} = \mathbb{E}_{u \sim \mathcal{N}(0,I)} \left[ x_i^T x_j 1_{\{u^T x_i \geq 0, u^T x_j \geq 0\}} \right]$.*

**Assumption 1.** (Du et al., 2018) *We assume $\lambda_0 \triangleq \lambda_{min}(\mathcal{H}^{\infty}) > 0$ which means that $\mathcal{H}^{\infty}$ is a positive definite matrix.*

**Lemma 1.** *If we i.i.d initialize $u_{jk} \sim \mathcal{N}(0,1)$ for $j \in [m]$ and $k \in [d_{in}]$, then with probability at least $(1-\delta)$, $u_{jk}$ induces a symmetric and homogeneously distributed matrix $U$ at initialization within a ball of radius $\zeta \triangleq \frac{2\sqrt{md_{in}}}{\sqrt{2\pi}\delta}$.*

# Warm-up: Reaction Without Diffusion

**Remark 1.** *Suppose* $\|\boldsymbol{u}_j - \boldsymbol{u}_j(0)\|_2 \leq \frac{c\delta\lambda_0}{n^2} \triangleq R$ *for some small positive constant* $c$*. In the current setup, the Gram matrix* $\mathcal{H} \in \mathbb{R}^{n \times n}$ *defined by*

$$\mathcal{H}_{ij} = \boldsymbol{x}_i^T \boldsymbol{x}_j \frac{1}{m} \sum_{r=1}^{m} 1_{\{\boldsymbol{u}_r^T \boldsymbol{x}_i \geq 0, \boldsymbol{u}_r^T \boldsymbol{x}_j \geq 0\}}$$

*satisfies* $\|\mathcal{H} - \mathcal{H}(0)\|_2 \leq \frac{\lambda_0}{4}$ *and* $\lambda_{min}(\mathcal{H}) \geq \frac{\lambda_0}{2}$.

**Remark 2.** *With Gram matrix* $\mathcal{H}(t)$*, the prediction dynamics,* $z(t) = f(\boldsymbol{U}(t), \boldsymbol{v}(t), \boldsymbol{x})$ *are governed by the following ODE:*

$$\frac{d\boldsymbol{z}(t)}{dt} = \mathcal{H}(t)(\boldsymbol{y} - \boldsymbol{z}(t)).$$

**Remark 3.** *For* $\lambda_{min}(\mathcal{H}(t)) \geq \frac{\lambda_0}{2}$*, we have*

$$\|\boldsymbol{z}(t) - \boldsymbol{y}\|_2 \leq \exp\left(-\frac{\lambda_0}{2} t\right) \|\boldsymbol{z}(0) - \boldsymbol{y}\|_2.$$

# Warm-up: Reaction Without Diffusion

**Theorem (Symmetry and Homogeneity)**

*Suppose* **Assumption 1** *holds. Let us i.i.d. initialize* $u_j \sim \mathcal{N}(0, I)$ *and sample* $v_j$ *uniformly from* $\{+1, -1\}$ *for all* $j \in [m]$*. If we choose* $\|x_p\|_2 = 1$ *for* $p \in [n]$*, then we obtain the following with probability at least* $1 - \delta$*:*

$$\|\boldsymbol{u}_j(t) - \boldsymbol{u}_j(0)\|_2 \leq \mathcal{O}\left(\frac{n^{3/2}}{\sqrt{m}\lambda_0\delta}\right),$$

$$\|\boldsymbol{U}(t) - \boldsymbol{U}(0)\|_F \leq \mathcal{O}\left(\frac{n^{3/2}}{\lambda_0\delta}\right).$$

# Warm-up: Reaction Without Diffusion

- Symmetry and homogeneity
- Breakdown of symmetry and homogeneity
- **Root cause: Turing instability**

# Main Result: Reaction With Diffusion

**Theorem (Breakdown of Symmetry and Homogeneity)**

*Suppose* **Assumption 1** *holds. Let us i.i.d. initialize* $u_j, w_r \sim \mathcal{N}(0, I)$ *and sample* $v_j, a_r$ *uniformly from* $\{+1, -1\}$ *for* $j, r \in [m]$. *Let* $\|x_p\|_2 = 1$ *for all* $p \in [n]$. *If we choose* $\|\boldsymbol{w}\|_2 \leq L \leq \mathcal{O}\left(\frac{\epsilon\sqrt{m}}{\kappa n\sqrt{2\log(2/\delta)}}\right)$, $\kappa = \mathcal{O}(\kappa^\infty)$ *where* $\kappa^\infty$ *denotes the condition number of* $\mathcal{H}^\infty$, *and define* $\mu \triangleq \frac{Ln\sqrt{2\log(2/\delta)}}{\sqrt{m}}$, *then with probability at least* $1-\delta$, *we obtain the following:*

$$\|\boldsymbol{u}_j(t) - \boldsymbol{u}_j(0)\|_2 \leq \mathcal{O}\left(\frac{n^{3/2}}{\sqrt{m}\lambda_0\delta} + \left(\frac{\mu\left(1+\kappa\sqrt{n}\right)}{\sqrt{m}}\right)t\right),$$

$$\|\boldsymbol{U}(t) - \boldsymbol{U}(0)\|_F \leq \mathcal{O}\left(\frac{n^{3/2}}{\lambda_0\delta} + \mu\left(1+\kappa\sqrt{n}\right)t\right).$$

# Reaction With Diffusion: Proof Sketch

Gradient Flow:

$$\begin{aligned}\left\|\frac{d\boldsymbol{u}_j(s)}{ds}\right\|_2 &= \left\|\frac{\partial\mathcal{L}_{aug}\left(\boldsymbol{U},\boldsymbol{v},\boldsymbol{w},\boldsymbol{a}\right)}{\partial\boldsymbol{u}_j(s)}\right\|_2 \\ &= \left\|\frac{\partial\mathcal{L}_{sup}\left(\boldsymbol{U},\boldsymbol{v}\right)}{\partial\boldsymbol{u}_j(s)} - \frac{\partial}{\partial\boldsymbol{u}_j(s)}\sum_{p=1}^{n} g\left(\boldsymbol{w},a,z_p\right)\right\|_2 \\ &\leq \underbrace{\left\|\frac{\partial\mathcal{L}_{sup}\left(\boldsymbol{U},\boldsymbol{v}\right)}{\partial\boldsymbol{u}_j(s)}\right\|_2 + \left\|\frac{\partial}{\partial\boldsymbol{u}_j(s)}\sum_{p=1}^{n} g\left(\boldsymbol{w},a,z_p\right)\right\|_2}_{\text{Triangle inequality}}.\end{aligned} \tag{9}$$

# Reaction With Diffusion: Reaction Term

Gradient Flow:

$$\begin{aligned}\left\| \frac{d\boldsymbol{u}_j(s)}{ds} \right\|_2 &= \left\| \frac{\partial \mathcal{L}_{aug}\left(\boldsymbol{U}, \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{a}\right)}{\partial \boldsymbol{u}_j(s)} \right\|_2 \\ &= \left\| \frac{\partial \mathcal{L}_{sup}\left(\boldsymbol{U}, \boldsymbol{v}\right)}{\partial \boldsymbol{u}_j(s)} - \frac{\partial}{\partial \boldsymbol{u}_j(s)} \sum_{p=1}^{n} g\left(\boldsymbol{w}, a, z_p\right) \right\|_2 \\ &\leq \underbrace{\left\| \frac{\partial \mathcal{L}_{sup}\left(\boldsymbol{U}, \boldsymbol{v}\right)}{\partial \boldsymbol{u}_j(s)} \right\|_2 + \left\| \frac{\partial}{\partial \boldsymbol{u}_j(s)} \sum_{p=1}^{n} g\left(\boldsymbol{w}, a, z_p\right) \right\|_2}_{\text{Triangle inequality}}.\end{aligned} \tag{10}$$

# Reaction With Diffusion: Reaction Term

**Lemma 2.** *In contrast to* **Remark 2**, *the prediction dynamics in adversarial regularization are governed by the following ODE:*

$$\frac{d\boldsymbol{z}(t)}{dt} = \mathcal{H}(t)\left(\boldsymbol{y} - \boldsymbol{z}(t)\right) + \mathcal{H}(t)\nabla_{\boldsymbol{z}(t)} g(\boldsymbol{w}(t), \boldsymbol{a}(t), \boldsymbol{z}(t)). \tag{11}$$

**Lemma 3.** (Hoeffding's inequality, two sided (vershynin et al.)) *Suppose* $\boldsymbol{a} = (a_1, a_2, \ldots, a_m) \in \{\pm 1\}^m$ *be a collection of independent symmetric Bernoulli random variables, and* $\boldsymbol{w} = (w_1, w_2, \ldots, w_m) \in \mathbb{R}^m$. *Then, for any* $t > 0$, *we have*

$$\mathbb{P}\left\{\left|\sum_{r=1}^{m} a_r w_r\right| \geq t\right\} \leq 2\exp\left(-\frac{t^2}{2\left\|\boldsymbol{w}\right\|_2^2}\right). \tag{12}$$

# Reaction With Diffusion: Reaction Term

**Lemma 4.** *Suppose* **Assumption 1** *holds. If we denote* $\lambda_{\max}(\mathcal{H}^{\infty})$ *by* $\lambda_1^{\infty}$*, then* $\lambda_{\max}(\mathcal{H}) \leq \frac{\lambda_1}{2} \triangleq \lambda_1^{\infty} + \frac{\lambda_0}{2}$.

The distance from true labels can be bounded by

$$\begin{aligned}\frac{d}{dt}\left\|\boldsymbol{z}(t)-\boldsymbol{y}\right\|_2^2 &= 2\left\langle \boldsymbol{z}(t)-\boldsymbol{y}, \frac{d\boldsymbol{z}(t)}{dt}\right\rangle \\ &= 2\left\langle \boldsymbol{z}(t)-\boldsymbol{y}, -\mathcal{H}(t)\left(\boldsymbol{z}(t)-\boldsymbol{y}\right)\right\rangle \\ &\quad + 2\left\langle \boldsymbol{z}(t)-\boldsymbol{y}, \mathcal{H}(t)\nabla_{\boldsymbol{z}(t)} g(\boldsymbol{w}(t), \boldsymbol{a}(t), \boldsymbol{z}(t))\right\rangle\end{aligned} \tag{13}$$

# Reaction With Diffusion: Reaction Term

**Lemma 4.** *Suppose* **Assumption 1** *holds. If we denote* $\lambda_{\max}(\mathcal{H}^\infty)$ *by* $\lambda_1^\infty$, *then* $\lambda_{\max}(\mathcal{H}) \le \frac{\lambda_1}{2} \triangleq \lambda_1^\infty + \frac{\lambda_0}{2}$.

The distance from true labels can be bounded by

$$\begin{aligned}\frac{d}{dt}\left\|\boldsymbol{z}(t)-\boldsymbol{y}\right\|_2^2 &= 2\left\langle \boldsymbol{z}(t)-\boldsymbol{y}, \frac{d\boldsymbol{z}(t)}{dt}\right\rangle \\ &= 2\left\langle \boldsymbol{z}(t)-\boldsymbol{y}, -\mathcal{H}(t)\left(\boldsymbol{z}(t)-\boldsymbol{y}\right)\right\rangle \\ &\quad + 2\left\langle \boldsymbol{z}(t)-\boldsymbol{y}, \mathcal{H}(t)\nabla_{\boldsymbol{z}(t)} g(\boldsymbol{w}(t), \boldsymbol{a}(t), \boldsymbol{z}(t))\right\rangle\end{aligned} \tag{13}$$

Since $\lambda_{\min}(\mathcal{H}) \ge \frac{\lambda_0}{2}$ (**Remark 1**) and $\lambda_{\max}(\mathcal{H}) \le \frac{\lambda_1}{2}$ (**Lemma 4**), we get

$$\begin{aligned}\frac{d}{dt}\left\|\boldsymbol{z}(t)-\boldsymbol{y}\right\|_2^2 \le -\lambda_0\left\|\boldsymbol{z}(t)-\boldsymbol{y}\right\|_2^2 \\ + \lambda_1\left\langle \boldsymbol{z}(t)-\boldsymbol{y}, \nabla_{\boldsymbol{z}(t)} g(\boldsymbol{w}(t), \boldsymbol{a}(t), \boldsymbol{z}(t))\right\rangle\end{aligned} \tag{14}$$

# Reaction With Diffusion: Reaction Term

Upon simplification using **Lemma 3**,

$$\frac{d}{dt}\left\|\boldsymbol{z}(t)-\boldsymbol{y}\right\|_2^2 \leq -\lambda_0\left\|\boldsymbol{z}(t)-\boldsymbol{y}\right\|_2^2 + \lambda_1\mu\left\|\boldsymbol{z}(t)-\boldsymbol{y}\right\|_2 \tag{15}$$

For simplicity, let us suppose $\psi = \left\|\boldsymbol{z}(t)-\boldsymbol{y}\right\|_2^2$. Now,

$$\frac{d\psi}{dt} \leq -\lambda_0\psi + \lambda_1\mu\psi^{1/2} \tag{16}$$

# Reaction With Diffusion: Reaction Term

Upon simplification using **Lemma 3**,

$$\frac{d}{dt}\left\|\boldsymbol{z}(t)-\boldsymbol{y}\right\|_2^2 \leq -\lambda_0\left\|\boldsymbol{z}(t)-\boldsymbol{y}\right\|_2^2 + \lambda_1\mu\left\|\boldsymbol{z}(t)-\boldsymbol{y}\right\|_2 \tag{15}$$

For simplicity, let us suppose $\psi = \left\|\boldsymbol{z}(t)-\boldsymbol{y}\right\|_2^2$. Now,

$$\frac{d\psi}{dt} \leq -\lambda_0\psi + \lambda_1\mu\psi^{1/2} \tag{16}$$

**Bernoulli Differential Equation (BDE) (Bernoulli, 1695)**

$$\frac{dx(t)}{dt} = -P(t)x(t) + Q(t)x^n(t) \text{ for } n \in \mathbb{R}\setminus\{0,1\}$$

# Reaction With Diffusion: Reaction Term

Exact solution of the BDE:

$$\|\boldsymbol{z}(t) - \boldsymbol{y}\|_2 \leq (\|\boldsymbol{z}(0) - \boldsymbol{y}\|_2 - \kappa\mu) \exp\left(-\frac{\lambda_0}{2} t\right) + \kappa\mu. \tag{17}$$

From warm-up exercise, we know for $0 \leq s \leq t$,

$$\begin{aligned}\left\|\frac{\partial \mathcal{L}_{sup}\left(\boldsymbol{U}, \boldsymbol{v}\right)}{\partial \boldsymbol{u}_j(s)}\right\|_2 &\leq \frac{\sqrt{n}}{\sqrt{m}} \|\boldsymbol{z}(s) - \boldsymbol{y}\|_2 \\ &\leq \frac{\sqrt{n}}{\sqrt{m}} (\|\boldsymbol{z}(0) - \boldsymbol{y}\|_2 - \kappa\mu) \exp\left(-\frac{\lambda_0}{2} t\right) + \frac{\sqrt{n}}{\sqrt{m}} \kappa\mu.\end{aligned} \tag{18}$$

# Pseudo-Reaction-Diffusion Model

Governing Dynamics:

$$\frac{d\boldsymbol{u}_j}{dt} = \mathfrak{R}_j^{\boldsymbol{u}}\left(\boldsymbol{u}_j, \boldsymbol{v}_j\right) + \mathfrak{D}_j^{\boldsymbol{u}}\left(\boldsymbol{u}_j\right) \tag{19}$$

Reaction Dynamics:

$$\mathfrak{R}_j^{\boldsymbol{u}}\left(\boldsymbol{u}_j(t), \boldsymbol{v}_j(t)\right) \leq \frac{\sqrt{n}}{\sqrt{m}}\left(\|\boldsymbol{z}(0) - \boldsymbol{y}\|_2 - \kappa\mu\right)\exp\left(-\frac{\lambda_0}{2}t\right) + \frac{\sqrt{n}}{\sqrt{m}}\kappa\mu. \tag{20}$$

# Pseudo-Reaction-Diffusion Model

Governing Dynamics:

$$\frac{d\boldsymbol{u}_j}{dt} = \mathfrak{R}_j^{\boldsymbol{u}}\left(\boldsymbol{u}_j, \boldsymbol{v}_j\right) + \mathfrak{D}_j^{\boldsymbol{u}}\left(\boldsymbol{u}_j\right) \tag{19}$$

Reaction Dynamics:

$$\mathfrak{R}_j^{\boldsymbol{u}}\left(\boldsymbol{u}_j(t), \boldsymbol{v}_j(t)\right) \leq \frac{\sqrt{n}}{\sqrt{m}}\left(\|\boldsymbol{z}(0) - \boldsymbol{y}\|_2 - \kappa\mu\right)\exp\left(-\frac{\lambda_0}{2}t\right) + \frac{\sqrt{n}}{\sqrt{m}}\kappa\mu. \tag{20}$$

Diffusion Dynamics:

$$\mathfrak{D}_j^{\boldsymbol{u}}\left(\boldsymbol{u}_j\right) \leq ?$$

# Reaction With Diffusion: Diffusion Term

Augmented part:

$$\left\| \frac{d\boldsymbol{u}_j(s)}{ds} \right\|_2 \leq \left\| \frac{\partial \mathcal{L}_{sup}\left(\boldsymbol{U}, \boldsymbol{v}\right)}{\partial \boldsymbol{u}_j(s)} \right\|_2 + \left\| \frac{\partial}{\partial \boldsymbol{u}_j(s)} \sum_{p=1}^{n} g\left(\boldsymbol{w}, a, z_p\right) \right\|_2 \tag{21}$$

Upon expansion,

$$\begin{aligned} &\left\| \frac{\partial}{\partial \boldsymbol{u}_j(s)} \sum_{p=1}^{n} g\left(\boldsymbol{w}, a, z_p\right) \right\|_2 \\ &= \left\| \sum_{p=1}^{n} \sum_{r=1}^{m} \frac{1}{\sqrt{m}} a_r 1_{\{w_r z_p \geq 0\}} w_r \frac{1}{\sqrt{m}} v_j 1_{\left\{\boldsymbol{v}_j^T \boldsymbol{x}_p \geq 0\right\}} \boldsymbol{x}_p \right\|_2 . \end{aligned} \tag{22}$$

# Reaction With Diffusion: Diffusion Term

By triangle inequality, Cauchy-Schwarz inequality, and **Lemma 3**, we get

$$\begin{aligned}
\left\| \frac{\partial}{\partial \boldsymbol{u}_j(s)} \sum_{p=1}^{n} g\left(\boldsymbol{w}, a, z_p\right) \right\|_2 &\leq \frac{1}{m} \sum_{p=1}^{n} \left\| v_j 1_{\left\{\boldsymbol{v}_j^T \boldsymbol{x}_p \geq 0\right\}} \boldsymbol{x}_p \sum_{r=1}^{m} a_r w_r 1_{\{w_r z_p \geq 0\}} \right\|_2 \\
&\leq \frac{1}{m} \sum_{p=1}^{n} \left| \sum_{r=1}^{m} a_r w_r \right| \\
&\leq \frac{1}{m} \sum_{p=1}^{n} \|\boldsymbol{w}\|_2 \sqrt{2 \log\left(\frac{2}{\delta}\right)} \\
&\leq \frac{Ln\sqrt{2\log\left(\frac{2}{\delta}\right)}}{m} = \mathcal{O}\left(\frac{\mu}{\sqrt{m}}\right)
\end{aligned} \tag{23}$$

# Main Result: Reaction With Diffusion

Reaction Dynamics:

$$\mathfrak{R}_j^u(\boldsymbol{u}_j(t)) \leq \frac{\sqrt{n}}{\sqrt{m}} \left(\|\boldsymbol{z}(0) - \boldsymbol{y}\|_2 - \kappa\mu\right) \exp\left(-\frac{\lambda_0}{2} t\right) + \frac{\sqrt{n}}{\sqrt{m}} \kappa\mu. \tag{24}$$

Diffusion Dynamics:

$$\mathfrak{D}_j^u(\boldsymbol{u}_j(t)) \leq \frac{Ln\sqrt{2\log\left(\frac{2}{\delta}\right)}}{m}. \tag{25}$$

Integrating over $0 \leq s \leq t$,

$$\begin{aligned} \|\boldsymbol{u}_j(t) - \boldsymbol{u}_j(0)\|_2 &\leq \int_0^t \left\| \frac{d\boldsymbol{u}_j(s)}{ds} \right\|_2 ds \\ &\leq \int_0^t \mathfrak{R}_j^u(\boldsymbol{u}_j(s)) + \mathfrak{D}_j^u(\boldsymbol{u}_j(s))\ ds. \end{aligned} \tag{26}$$

# Main Result: Reaction With Diffusion

## Individual Neuron

$$\|\boldsymbol{u}_j(t) - \boldsymbol{u}_j(0)\|_2 \leq \mathcal{O}\left(\frac{n^{3/2}}{m^{1/2}\lambda_0\delta} + \left(\frac{\mu\left(1+\kappa\sqrt{n}\right)}{m^{1/2}}\right) t\right)$$

# Main Result: Reaction With Diffusion

## Individual Neuron

$$\left\|\boldsymbol{u}_j(t) - \boldsymbol{u}_j(0)\right\|_2 \leq \mathcal{O}\left(\frac{n^{3/2}}{m^{1/2}\lambda_0\delta} + \left(\frac{\mu\left(1+\kappa\sqrt{n}\right)}{m^{1/2}}\right) t\right)$$

## Spatial Grid of Neurons

$$\left\|\boldsymbol{U}(t) - \boldsymbol{U}(0)\right\|_F \leq \mathcal{O}\left(\frac{n^{3/2}}{\lambda_0\delta} + \mu\left(1 + \kappa\sqrt{n}\right) t\right)$$

# Main Result: Reaction With Diffusion

### Individual Neuron

$$\|\boldsymbol{u}_j(t) - \boldsymbol{u}_j(0)\|_2 \leq \mathcal{O}\left(\frac{n^{3/2}}{m^{1/2}\lambda_0\delta} + \left(\frac{\mu\left(1+\kappa\sqrt{n}\right)}{m^{1/2}}\right) t\right)$$

### Spatial Grid of Neurons

$$\|\boldsymbol{U}(t) - \boldsymbol{U}(0)\|_F \leq \mathcal{O}\left(\frac{n^{3/2}}{\lambda_0\delta} + \mu\left(1 + \kappa\sqrt{n}\right) t\right)$$

### Breakdown Threshold

$$m = \Omega\left(\left(\frac{n^{7/2}}{\lambda_0^2\delta^2} + \frac{n^2\mu\left(1+\kappa\sqrt{n}\right)T_0}{\lambda_0\delta}\right)^2\right)$$

# Jointly Training Both Layers

**Theorem (Reaction-Diffusion Dynamics)**

*If we absorb constants in $\mathcal{O}(.)$ and set $(\boldsymbol{y}_p - \boldsymbol{z}_p)_i \, v_{ij} \mathbb{1}_{\{\boldsymbol{u}_j^T \boldsymbol{x}_p \geq 0\}} x_{p,k} = \mathcal{O}(1)$ for $i \in [d_{out}]$ and $p \in [n]$, then for all $j \in [m]$ the RD dynamics satisfy:*

$$\mathfrak{R}_j^{\boldsymbol{u}}(\boldsymbol{u}_j, \boldsymbol{v}_j) = \mathcal{O}\left(n d_{in} \sqrt{\frac{d_{out}}{m}}\right),$$

$$\mathfrak{D}_j^{\boldsymbol{u}}(\nabla^2 \boldsymbol{u}_j) = \mathcal{O}\left(n m^2 d_{in} d_{out}^{3/2}\right),$$

$$\mathfrak{R}_j^{\boldsymbol{v}}(\boldsymbol{u}_j, \boldsymbol{v}_j) = \mathcal{O}\left(n d_{in} \sqrt{\frac{d_{out}}{m}}\right),$$

$$\mathfrak{D}_j^{\boldsymbol{v}}(\nabla^2 \boldsymbol{v}_j) = \mathcal{O}\left(n m^2 d_{in} d_{out}^{1/2}\right).$$

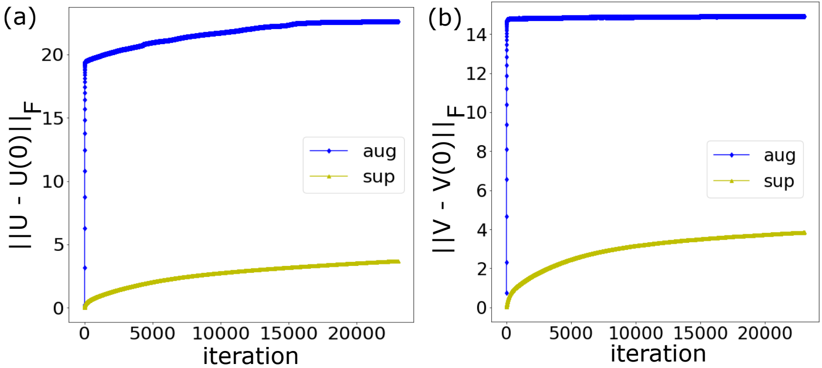
# Experiments

- **Linear Rate:** Solution in a larger subspace around initialization.
- **Theorem 1:** Maintaining symmetry and homogeneity.
- **Theorem 2:** Breakdown of symmetry and homogeneity.

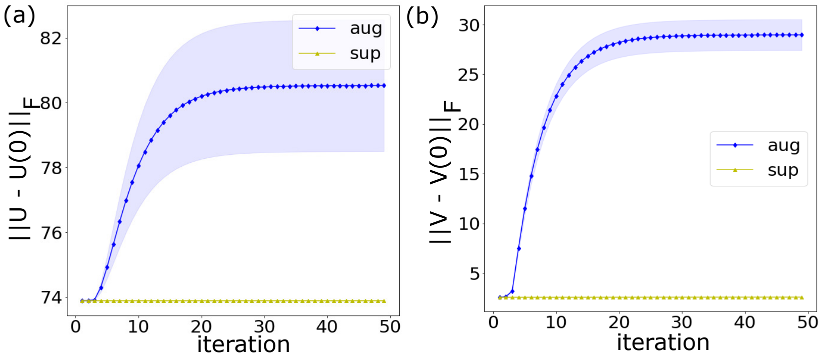
# Experiments

- **Linear Rate:** Solution in a larger subspace around initialization.
- **Theorem 1:** Maintaining symmetry and homogeneity.
- **Theorem 2:** Breakdown of symmetry and homogeneity.

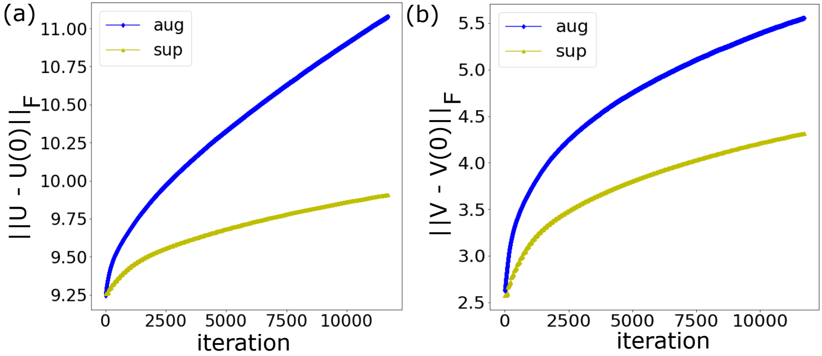
# Experimental Results: Linear Rate



# Experimental Results: Linear Rate



# Experimental Results: Dissecting Diffusion



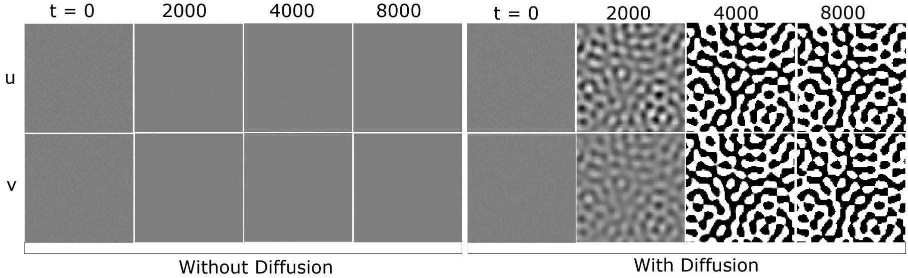
# Experiments

- **Linear Rate:** Solution in a larger subspace around initialization.
- **Theorem 1:** Maintaining symmetry and homogeneity.
- **Theorem 2:** Breakdown of symmetry and homogeneity.

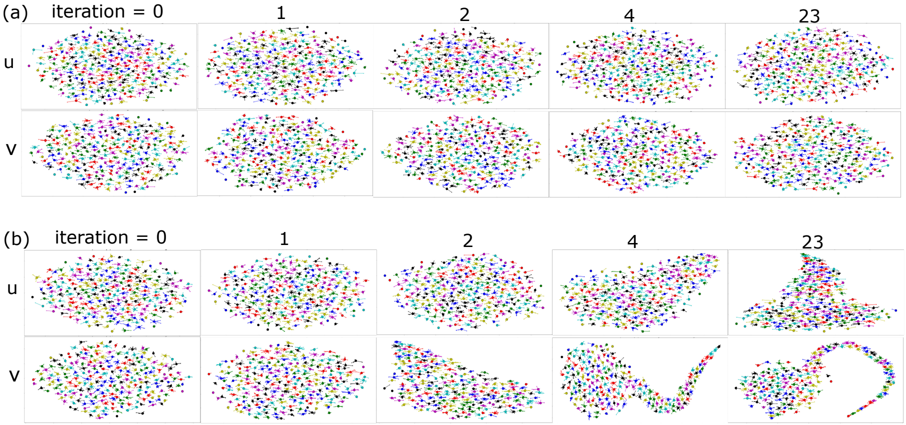
# Turing Patterns by RD Model

## Reaction-Diffusion Model

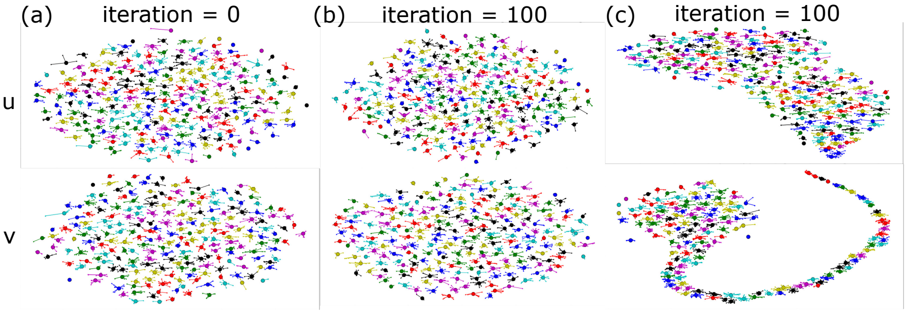
$$\frac{d\boldsymbol{u}_j}{dt} = \mathfrak{R}_j^{\boldsymbol{u}}\left(\boldsymbol{u}_j, \boldsymbol{v}_j\right) + \mathfrak{D}_j^{\boldsymbol{u}}\left(\nabla^2 \boldsymbol{u}_j\right),$$
$$\frac{d\boldsymbol{v}_j}{dt} = \mathfrak{R}_j^{\boldsymbol{v}}\left(\boldsymbol{u}_j, \boldsymbol{v}_j\right) + \mathfrak{D}_j^{\boldsymbol{v}}\left(\nabla^2 \boldsymbol{v}_j\right).$$



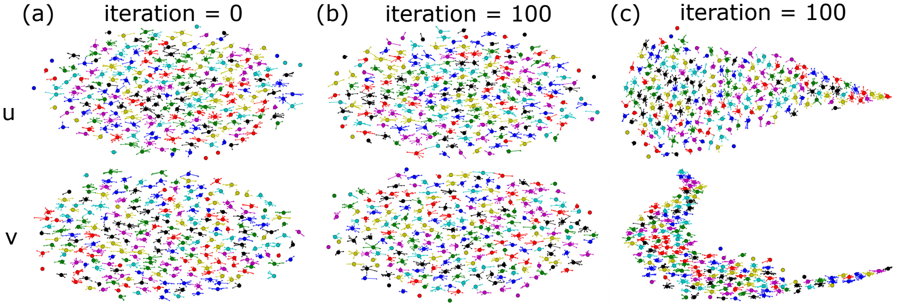
# Turing-like Patterns by PRD Model: Synthetic Dataset



# Turing-like Patterns by PRD Model: MNIST



# Turing-like Patterns by PRD Model: FashionMNIST



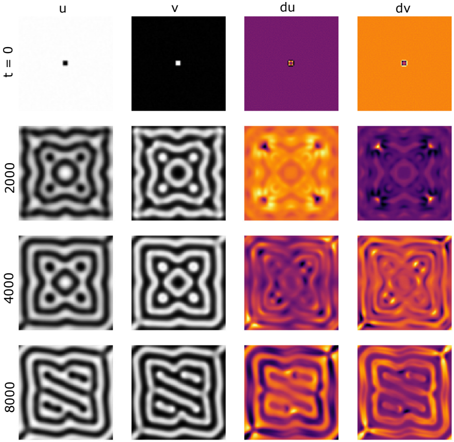
# Turing-like Patterns by Gray-Scott Model

## Gray-Scott Model

$$\frac{\partial u}{\partial t} = F(1-u) - uv^2 + \mu'\nabla^2 u$$
$$\frac{\partial v}{\partial t} = -(F+k)v + uv^2 + \nu'\nabla^2 v$$

## Parameters

$F = 0.025, K = 0.055, \mu' = 2e-5, \nu' = 1e-5$



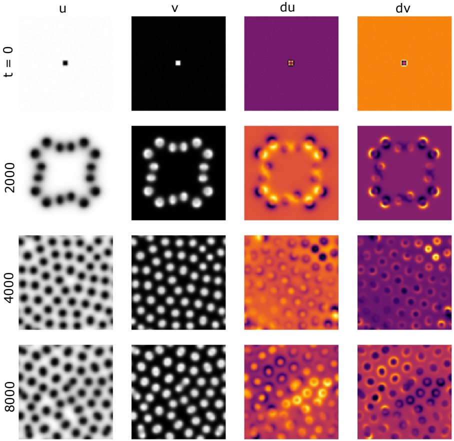
# Turing-like Patterns by Gray-Scott Model

## Gray-Scott Model

$$\frac{\partial u}{\partial t} = F(1-u) - uv^2 + \mu' \nabla^2 u$$
$$\frac{\partial v}{\partial t} = -(F+k)v + uv^2 + \nu' \nabla^2 v$$

## Parameters

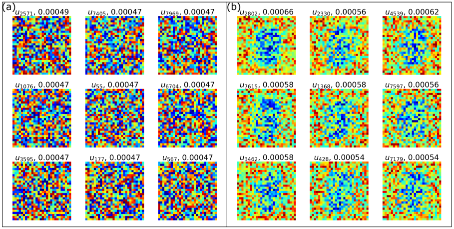
$F = 0.025, K = 0.060, \mu' = 2e-5, \nu' = 1e-5$

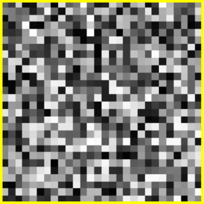


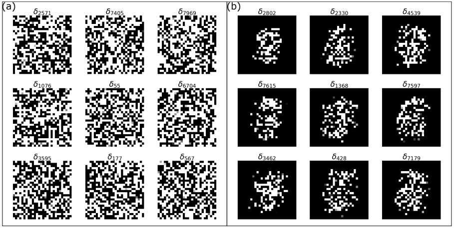
# Importance of Diffusion in PRD Model

- Reminiscent of patterns observed in nature
- Interpretable kernel weights
- Feature visualization:

$$\delta_j = \arg\max_{\delta \in \Delta} \boldsymbol{u}_j^T (x + \delta)$$





# Summary

- Exponentially fast convergence of over-parameterized networks under adversarial interaction.
- Theoretical justification of symmetry and homogeneity.
- Exploration of larger subspace around initialization beyond breakdown of symmetry and homogeneity.
- Interpretable kernels in regularized adversarial learning.
- Turing-like pattern formation under mild diffusion.
- Resemblance with naturally occurring Bernoulli differential equation.