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

Abstract

Long after Turing's seminal Reaction-Diffusion (RD) model, the elegance of his fundamental equations alleviated much of the skepticism surrounding pattern formation. Though Turing model is a simplification and an idealization, it is one of the best-known theoretical models to explain patterns as a reminiscent of those observed in nature. Over the years, concerted efforts have been made to align theoretical models to explain patterns in real systems. The apparent difficulty in identifying the specific dynamics of the RD system makes the problem particularly challenging. Interestingly, we observe Turing-like patterns in a system of neurons with adversarial interaction. In this study, we establish the involvement of Turing instability to create such patterns. By theoretical and empirical studies, we present a *pseudo-reaction-diffusion* model to explain the mechanism that may underlie these phenomena. While supervised learning attains homogeneous equilibrium, this paper suggests that the introduction of an adversary helps break this homogeneity to create non-homogeneous patterns at equilibrium. Further, we prove that randomly initialized gradient descent with over-parameterization can converge exponentially fast to an ϵ-stationary point even under adversarial interaction. In addition, different from sole supervision, we show that the solutions obtained under adversarial interaction are not limited to a tiny subspace around initialization.

Introduction

In this paper, we intend to demystify an interesting phenomenon: adversarial interaction between generator and discriminator creates non-homogeneous equilibrium by inducing Turing instability in a Pseudo-Reaction-Diffusion (PRD) model. This is in contrast to supervised learning where the identical model achieves homogeneous equilibrium while maintaining spatial symmetry over iterations.

Recent success of Generative Adversarial Networks (GANs) (Goodfellow et al. 2014; Arjovsky, Chintala, and Bottou 2017) has led to exciting applications in a wide variety of tasks (Luc et al. 2016; Zhu et al. 2017; Ledig et al. 2017; Engin, Genç, and Kemal Ekenel 2018; Rout et al. 2020). In adversarial learning paradigm, it is often required that a particular sample is generated subject to a conditional input. Typically, conditional GANs are employed to meet these demands (Mirza and Osindero 2014). Further, it has

been reported in copious literature that supervised learning with adversarial regularization performs better than sole supervision (Ledig et al. 2017; Rout 2020; Wang et al. 2018; Wang and Gupta 2016; Karacan et al. 2016; Sarmad, Lee, and Kim 2019). In all these prior works, one may notice several crucial properties of adversarial interaction. It is worth emphasizing that adversarial learning owes its benefits to the continuously evolving loss function which otherwise is extremely difficult to model. Motivated by these findings, we uncover another interesting property of adversarial training. We observe that adversarial interaction helps break the spatial symmetry and homogeneity to create non-homogeneous patterns in weight space.

The reason for studying these phenomena is multi-fold. The fact that adversarial interaction exhibits Turing-like patterns creates a dire need to investigate its connections to nature. In particular, these patterns often emerge in real world systems, such as butterfly wings, zebra, giraffe and leopard (Turing 1952; Meinhardt 1982; Rauch and Millonas 2004; Nakamasu et al. 2009; Kondo and Miura 2010). Interestingly, adversarial training captures some intricacies of this complex biological process that create evolutionary patterns in neural networks. Furthermore, it is important to understand neural synchronization in human brain to design better architectures (Budzynski et al. 2009). This paper is intended to shed light on some of these aspects.

It is widely believed that fully connected networks already capture certain important properties of deep learning (Saxe, McClelland, and Ganguli 2014; Li and Liang 2018). While one may wish to extend these analyses to more complex networks, it may not allow a comprehensive study of various fundamental aspects in the nascent state of understanding. Besides, the complexity involved in studying the Reaction-Diffusion (RD) dynamics of a large neural network is enormous. For this reason, we study two layer neural networks and focus more on the theory of Turing-like patterns.

While dynamical systems governed by different equations exhibit different patterns, it is crucial to study the dynamics through *reaction and diffusion* terms that laid the foundation of pattern formation (Turing 1952). Our key observation:

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.

To provide a thorough explanation to these empirical findings, we derive the governing dynamics of a PRD model.

From another perspective, the generator provides a short-range positive feedback as it tries to minimize the empirical risk directly. On the other hand, the discriminator provides a long-range negative feedback as it tries to maximize the generator cost. Since the adversary discriminates between real and fake samples, it indirectly optimizes the primary objective function. It is safe to assume that such signals from the discriminator to the generator form the basis of long-range negative feedback as studied by Rauch and Millonas.

Preliminaries

Notations   Bold upper-case letter $\boldsymbol{A}$ denotes a matrix. Bold lower-case letter $\boldsymbol{a}$ denotes a vector. Normal lower-case letter a denotes a scalar. $\|.\|_2$ represents Euclidean norm of a vector and spectral norm of a matrix. $\|.\|_F$ represents Frobenius norm of a matrix. $\lambda_{\min}(.)$ and $\lambda_{\max}(.)$ denote smallest and largest eigen value of a matrix. dx represents derivative of x and ∂x represents its partial derivative. For $g : \mathbb{R}^d \to \mathbb{R}$, ∇g and $\nabla^2 g$ denote gradient and Laplacian of g, respectively. $[m]$ denotes the set $\{1, 2, \ldots, m\}$.

Problem Setup   Consider that we are given n training samples $\{(\boldsymbol{x}_p, \boldsymbol{y}_p)\}_{p=1}^n \subset \mathbb{R}^{d_{in}} \times \mathbb{R}^{d_{out}}$. Formally, we use the following notations to represent two layer neural networks with rectified linear unit (ReLU) activation function ($\sigma(.)$).

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

Here, $\boldsymbol{U} \in \mathbb{R}^{m \times d_{in}}$ and $\boldsymbol{V} \in \mathbb{R}^{d_{out} \times m}$. Let us denote $\boldsymbol{u}_j = \boldsymbol{U}_{j,:}$ and $\boldsymbol{v}_j = \boldsymbol{V}_{:,j}$. The scaling factor $\frac{1}{\sqrt{d_{out} m}}$ is derived from Xavier initialization (Glorot and Bengio 2010). In supervised learning, the training is carried out by minimizing the l_2 loss over data as given by

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

The input data points are 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 $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n) \in \mathbb{R}^{d_{out} \times n}$. In regularized adversarial learning, the generator cost is augmented with an adversary:

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

The adversary, $g(\boldsymbol{W}, \boldsymbol{a}, y) = \frac{1}{\sqrt{m}} \boldsymbol{a}^T \sigma(\boldsymbol{W} \boldsymbol{y}) : \mathbb{R}^{d_{out}} \to \mathbb{R}$ is a two layer network with ReLU activation. Here, $\boldsymbol{W} \in \mathbb{R}^{m \times d_{out}}$ and $\boldsymbol{a} \in \mathbb{R}^m$. The discriminator cost is exactly identical to the critic of WGAN with gradient penalty (Gulrajani et al. 2017). We follow the common practice to train generator and discriminator alternatively using Wasserstein distance. In this study, $\mathcal{L}_{aug}$ is considered as the equivalent of a continuous field in a RD system (Turing 1952).

Learning Algorithm   We consider vanilla gradient descent with random initialization as our learning algorithm to minimize both supervised and augmented objective. For instance, we update each trainable parameter in augmented objective by the following Ordinary Differential Equations (ODE):

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

for $i \in [d_{out}]$, $j \in [m]$ and $k \in [d_{in}]$. In ideal condition, the system enters equilibrium when $\frac{du_{jk}}{dt} = \frac{dv_{ij}}{dt} = 0$. To circumvent tractability issues, we seek ϵ-approximate equilibrium, i.e. $\left| \frac{du_{jk}}{dt} \right| < \epsilon$ and $\left| \frac{dv_{ij}}{dt} \right| < \epsilon$ for a small ϵ.

Revisiting Reaction-Diffusion Model(Turing 1952)

We focus on two body morphogenesis though it may be applied generally to many bodies upon further investigation. Here, two bodies refer to two layers of generator network. There are $2m$ differential equations governing the reaction ($\mathfrak{R}$) and diffusion ($\mathfrak{D}$) dynamics of such a complex system:

$$\frac{d\boldsymbol{u}_j}{dt} = \mathfrak{R}_j^{\boldsymbol{u}}(\boldsymbol{u}_j, \boldsymbol{v}_j) + \mathfrak{D}_j^{\boldsymbol{u}}(\nabla^2 \boldsymbol{u}_j),$$
$$\frac{d\boldsymbol{v}_j}{dt} = \mathfrak{R}_j^{\boldsymbol{v}}(\boldsymbol{u}_j, \boldsymbol{v}_j) + \mathfrak{D}_j^{\boldsymbol{v}}(\nabla^2 \boldsymbol{v}_j), \qquad (5)$$

where $j = 1, 2, \ldots, m$. Here, m denotes the total number of neurons in the hidden layer. In the current setup, $\boldsymbol{u}_j = (u_{jk})_{k=1}^{d_{in}}, u_{jk} \in \mathbb{R}$ and $\boldsymbol{v}_j = (v_{ij})_{i=1}^{d_{out}}, v_{ij} \in \mathbb{R}$. Thus, $\frac{d\boldsymbol{u}_j}{dt} = \left(\frac{du_{jk}}{dt} \right)_{k=1}^{d_{in}}$ and $\frac{d\boldsymbol{v}_j}{dt} = \left(\frac{dv_{ij}}{dt} \right)_{i=1}^{d_{out}}$. In the current analogy, each neuron represents a morphogen as it fulfills the fundamental requirements of Turing pattern formation. For better understanding, we have grouped those in hidden layer to one entity ($\boldsymbol{u}_j$) and top layer to another entity ($\boldsymbol{v}_j$). Among several major advantages of RD systems, a few that are essential to the present body of analysis are separability, stability and strikingly rich spatio-temporal dynamics. Later parts of this paper will focus on deriving suitable expressions for the reaction and diffusion term.

Pseudo-Reaction-Diffusion Model

The analogy that has been made with RD systems in the foregoing analysis may be rather confusing to some readers. The succeeding analysis is intended to clarify some of these concerns. In the traditional setting, diffusion terms are limited to the Laplacian of the corresponding morphogens. In

the present account however, the diffusibility of one morphogen depends on the other morphogens, and hence the term *pseudo-reaction-diffusion*. Since later discoveries identified the root cause of pattern formation to be a short range positive feedback and a long range negative feedback (Meinhardt and Gierer 1974, 2000; Rauch and Millonas 2004), a system with adversarial interaction is fairly a pseudo-reaction-diffusion model.

Theoretical Analysis

First, we study symmetry and homogeneity in a simplified setup. In this regard, the separability property allows us to choose a scalar network, i.e., $d_{out} = 1$ and fix the second layer weights. There are $2m$ morphogens in the hidden layer itself making it a critically important analysis from mathematics perspective. Even with this simplification, the network is still non-convex and non-smooth[1]. The network architecture then becomes:

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

Our goal is to minimize

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

in a supervised setting and $\mathcal{L}_{aug}\left(\boldsymbol{U}, \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{a}\right)$

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

in an adversarial setting. The architecture of adversary is simplified to $g\left(w, a, y\right) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_j \sigma\left(w_j y\right)$. In the adversarial setting, this problem can be related to min-max optimization in non-convex-non-concave setting. We follow the definition of Gram matrix from Du et al.

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

Let us recall the following assumption which is crucial for the analysis in this paper.

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

The Gram matrix has several important properties (Tsuchida, Roosta, and Gallagher 2018; Xie, Liang, and Song 2017). One interesting property that justifies **Assumption 1** is given by Du et al.: *If no two inputs are parallel, then the Gram matrix is positive definite.* This is a valid assumption as very often we do not rely on a training dataset that contains too many parallel samples.

---

[1]We choose to fix the weights of the second layer because the network becomes convex and smooth if we fix the weights of the first layer. It motivates us to make this choice which is not far from practice and allows us to simplify the expressions.

Warm-Up: Reaction Without Diffusion

Before stating the main result, it is useful to get familiarized with the arguments of warm-up exercise.

Theorem 1. (Symmetry and Homogeneity) *Suppose Assumption 1 holds. Let us i.i.d. initialize* $u_j \sim \mathcal{N}\left(0, I\right)$ *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* $1 - \delta$:

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

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

Proof. We begin proof sketch with the following lemma.

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

Proof. Using the law of large numbers, it is trivial to prove symmetry and homogeneity since Gaussian distribution has a symmetric density function. We defer the proof of upper bound to Appendix.

Next, we prove how supervised cost helps maintain symmetry and homogeneity. Since $\boldsymbol{U}$ is initially symmetric and homogeneously distributed within ζ according to **Lemma 1**, the problem is now reduced to show that $\boldsymbol{U}(t)$ lies in the close proximity of $\boldsymbol{U}(0)$. We remark three crucial observations from Du et al. that are essential to our analysis.

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* $\lambda_{min}\left(\mathcal{H}\right) \geq \frac{\lambda_0}{2}.$

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

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

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

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

Now, for $0 \leq s \leq t$,

$$\left\|\frac{d\boldsymbol{u}_j(s)}{ds}\right\|_2 = \left\|\frac{\partial \mathcal{L}_{sup}\left(\boldsymbol{U}, \boldsymbol{v}\right)}{\partial \boldsymbol{u}_j(s)}\right\|_2 = \left\|\sum_{p=1}^{n} \left(z_p(s) - y_p\right) \frac{\partial z_p(s)}{\partial \boldsymbol{u}_j(s)}\right\|_2$$

$$= \left\|\sum_{p=1}^{n} \left(z_p(s) - y_p\right) \frac{1}{\sqrt{m}} v_j 1_{\{\boldsymbol{u}_j(s)^T \boldsymbol{x}_p \geq 0\}} \boldsymbol{x}_p\right\|_2.$$
(9)

By triangle inequality,

$$\left\|\frac{d\boldsymbol{u}_j(s)}{ds}\right\|_2 \le \sum_{p=1}^n \left\|(z_p(s) - y_p)\frac{1}{\sqrt{m}}v_j 1_{\{u_j(s)^T x_p \ge 0\}} x_p\right\|_2. \tag{10}$$

Using the classical inequality of Cauchy-Schwarz, $\|x_p\|_2 = 1$, $|v_j| = 1$ and **Remark 3**, we get

$$\begin{aligned}
\left\|\frac{d\boldsymbol{u}_j(s)}{ds}\right\|_2 &\le \sum_{p=1}^n \frac{1}{\sqrt{m}} |(z_p(s) - y_p)| \, |v_j| \, \|x_p\|_2 \\
&= \frac{1}{\sqrt{m}} \sum_{p=1}^n |(z_p(s) - y_p)| \\
&\le \frac{\sqrt{n}}{\sqrt{m}} \|\boldsymbol{z}(s) - \boldsymbol{y}\|_2 \\
&\le \frac{\sqrt{n}}{\sqrt{m}} \exp\left(-\frac{\lambda_0}{2}s\right) \|\boldsymbol{z}(0) - \boldsymbol{y}\|_2.
\end{aligned} \tag{11}$$

By integral form of Jensen's inequality, the distance from initialization can be bounded by

$$\begin{aligned}
\|\boldsymbol{u}_j(t) - \boldsymbol{u}_j(0)\|_2 &= \left\|\int_0^t \frac{d\boldsymbol{u}_j(s)}{ds} ds\right\|_2 \le \int_0^t \left\|\frac{d\boldsymbol{u}_j(s)}{ds}\right\|_2 ds \\
&\le \frac{\sqrt{n}}{\sqrt{m}} \int_0^t \exp\left(-\frac{\lambda_0}{2}s\right) \|\boldsymbol{z}(0) - \boldsymbol{y}\|_2 \, ds \\
&\le \frac{2\sqrt{n}\|\boldsymbol{z}(0) - \boldsymbol{y}\|_2}{\sqrt{m}\lambda_0}\left(1 - \exp\left(-\frac{\lambda_0}{2}t\right)\right).
\end{aligned} \tag{12}$$

Since $\exp\left(-\frac{\lambda_0}{2}t\right)$ is a decreasing function of t, the above expression simplifies to

$$\|\boldsymbol{u}_j(t) - \boldsymbol{u}_j(0)\|_2 \le \frac{2\sqrt{n}\|\boldsymbol{z}(0) - \boldsymbol{y}\|_2}{\sqrt{m}\lambda_0}. \tag{13}$$

Using Markov's inequality, with probability at least $1 - \delta$, we get

$$\begin{aligned}
\|\boldsymbol{u}_j(t) - \boldsymbol{u}_j(0)\|_2 &\le \frac{2\sqrt{n}\mathbb{E}\left[\|\boldsymbol{z}(0) - \boldsymbol{y}\|_2\right]}{\sqrt{m}\lambda_0\delta} \\
&\le \mathcal{O}\left(\frac{n^{3/2}}{m^{1/2}\lambda_0\delta}\right).
\end{aligned} \tag{14}$$

Now, we can bound the distance from initialization.

$$\begin{aligned}
\|\boldsymbol{U}(t) - \boldsymbol{U}(0)\|_F &= \left(\sum_{j=1}^m \sum_{k=1}^{d_{in}} |u_{jk}(t) - u_{jk}(0)|^2\right)^{1/2} \\
&\le \left(\sum_{j=1}^m \|\boldsymbol{u}_j(t) - \boldsymbol{u}_j(0)\|_2^2\right)^{1/2} \\
&\le \left(\sum_{j=1}^m \frac{4n\left(\mathbb{E}\left[\|\boldsymbol{z}(0) - \boldsymbol{y}\|_2\right]\right)^2}{m\lambda_0^2\delta^2}\right)^{1/2} \\
&\le \frac{2\sqrt{n}\mathbb{E}\left[\|\boldsymbol{z}(0) - \boldsymbol{y}\|_2\right]}{\lambda_0\delta} \le \mathcal{O}\left(\frac{n^{3/2}}{\lambda_0\delta}\right),
\end{aligned} \tag{15}$$

which finishes the proof. $\qquad\square$

Main Result: Reaction With Diffusion

To limit the capacity of a discriminator, it is often suggested to enforce a Lipschitz constraint on its parameters. While gradient clipping has been quite effective in this regard (Arjovsky, Chintala, and Bottou 2017), recent success in adversarial training owes in part to gradient penalty (Gulrajani et al. 2017). We remark that min-max optimization under non-convexity and non-concavity is considered NP-hard to find a stationary point (Lei et al. 2019). Therefore, it is necessary to make certain assumptions about discriminator, such as Lipschitz constraint, regularization and structure of the network. Different from one layer generator and quadratic discriminator (Lei et al. 2019), we study two layer networks with ReLU activations and rely on gradient penalty to limit its expressive power. In the simplified theoretical analysis, we assume $\|\boldsymbol{w}\|_2 \le L$ for a small constant $L > 0$.

Theorem 2. (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*

$$\|\boldsymbol{w}\|_2 \le L \le \mathcal{O}\left(\frac{\epsilon\sqrt{m}}{\kappa n\sqrt{2\log(2/\delta)}}\right), \quad \kappa = \mathcal{O}(\kappa^\infty) \text{ where}$$

κ^∞ 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[2]:

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

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

Proof. We sketch the proof of the main result as following.

Reaction Term For $0 \le s \le t$ in augmented objective as given by equation (8), we get

$$\begin{aligned}
\left\|\frac{d\boldsymbol{u}_j(s)}{ds}\right\|_2 &= \left\|\frac{\partial \mathcal{L}_{aug}(\boldsymbol{U}, \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{a})}{\partial \boldsymbol{u}_j(s)}\right\|_2 \\
&= \left\|\frac{\partial \mathcal{L}_{sup}(\boldsymbol{U}, \boldsymbol{v})}{\partial \boldsymbol{u}_j(s)} - \frac{\partial}{\partial \boldsymbol{u}_j(s)}\sum_{p=1}^n g(\boldsymbol{w}, a, z_p)\right\|_2 \\
&\le \underbrace{\left\|\frac{\partial \mathcal{L}_{sup}(\boldsymbol{U}, \boldsymbol{v})}{\partial \boldsymbol{u}_j(s)}\right\|_2 + \left\|\frac{\partial}{\partial \boldsymbol{u}_j(s)}\sum_{p=1}^n g(\boldsymbol{w}, a, z_p)\right\|_2}_{\text{Triangle inequality}}.
\end{aligned} \tag{16}$$

We start our analysis by first deriving an asymptotic upper bound of the supervised part. Then, we shift our focus to the augmented part which essentially constitutes the adversary.

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

$$\frac{d\boldsymbol{z}(t)}{dt} = \mathcal{H}(t)(\boldsymbol{y} - \boldsymbol{z}(t)) + \mathcal{H}(t)\nabla_{\boldsymbol{z}(t)}g(\boldsymbol{w}(t), \boldsymbol{a}(t), \boldsymbol{z}(t)). \tag{17}$$

---

[2]Refer to Appendix for experimental evidence and further discussion on breakdown of symmetry and homogeneity.

Proof. The above ODE is obtained by analyzing the dynamics as following:

$$\frac{dz_p(t)}{dt} = \sum_{j=1}^{m} \left\langle \frac{\partial f\left(\boldsymbol{U}, \boldsymbol{v}, \boldsymbol{x}_p\right)}{\partial \boldsymbol{u}_j(t)}, \frac{d\boldsymbol{u}_j(t)}{dt} \right\rangle$$

$$= \underbrace{\sum_{j=1}^{m} \left\langle \frac{\partial f\left(\boldsymbol{U}, \boldsymbol{v}, \boldsymbol{x}_p\right)}{\partial \boldsymbol{u}_j(t)}, \frac{1}{\sqrt{m}} \sum_{q=1}^{n} (y_q - z_q) v_j \boldsymbol{x}_q \mathbf{1}_{\left\{\boldsymbol{u}_j^T \boldsymbol{x}_q \geq 0\right\}} \right\rangle}_{\mathcal{A}}$$

$$+ \underbrace{\sum_{j=1}^{m} \left\langle \frac{\partial f\left(\boldsymbol{U}, \boldsymbol{v}, \boldsymbol{x}_p\right)}{\partial \boldsymbol{u}_j(t)}, \frac{1}{m} \sum_{q=1}^{n} \sum_{r=1}^{m} a_r w_r v_j \boldsymbol{x}_q \mathbf{1}_{\left\{w_r z_q \geq 0, \boldsymbol{u}_j^T \boldsymbol{x}_q \geq 0\right\}} \right\rangle}_{\mathcal{B}}.$$

(18)

Following arguments of the warm-up exercise, the first part can be simplified as:

$$\mathcal{A} := \sum_{j=1}^{m} \left\langle \frac{1}{\sqrt{m}} v_j \boldsymbol{x}_p \mathbf{1}_{\left\{\boldsymbol{u}_j^T \boldsymbol{x}_p \geq 0\right\}}, \frac{1}{\sqrt{m}} \sum_{q=1}^{n} (y_q - z_q) v_j \boldsymbol{x}_q \mathbf{1}_{\left\{\boldsymbol{u}_j^T \boldsymbol{x}_q \geq 0\right\}} \right\rangle$$

$$= \sum_{q=1}^{n} (y_q - z_q) \boldsymbol{x}_p^T \boldsymbol{x}_q \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}_{\left\{\boldsymbol{u}_j^T \boldsymbol{x}_p \geq 0, \boldsymbol{u}_j^T \boldsymbol{x}_q \geq 0\right\}}$$

$$\triangleq \sum_{q=1}^{n} (y_q - z_q(t)) \mathcal{H}_{pq}(t),$$

(19)

where $\mathcal{H}_{pq}(t)$ denotes the elements of Gram matrix $\mathcal{H}(t)$ defined by

$$\mathcal{H}_{pq}(t) = \boldsymbol{x}_p^T \boldsymbol{x}_q \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}_{\left\{\boldsymbol{u}_j^T \boldsymbol{x}_p \geq 0, \boldsymbol{u}_j^T \boldsymbol{x}_q \geq 0\right\}}. \quad (20)$$

Using the predefined Gram matrix, the second part can be simplified as:

$$\mathcal{B} := \sum_{j=1}^{m} \left\langle \frac{1}{\sqrt{m}} v_j \boldsymbol{x}_p \mathbf{1}_{\left\{\boldsymbol{u}_j^T \boldsymbol{x}_p \geq 0\right\}}, \frac{1}{m} \sum_{q=1}^{n} \sum_{r=1}^{m} a_r w_r v_j \boldsymbol{x}_q \mathbf{1}_{\left\{w_r z_q \geq 0, \boldsymbol{u}_j^T \boldsymbol{x}_q \geq 0\right\}} \right\rangle$$

$$= \sum_{q=1}^{n} \underbrace{\left(\frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r w_r \mathbf{1}_{\left\{w_r z_q \geq 0\right\}} \right)}_{\nabla_z g} \boldsymbol{x}_p^T \boldsymbol{x}_q \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}_{\left\{\boldsymbol{u}_j^T \boldsymbol{x}_p \geq 0, \boldsymbol{u}_j^T \boldsymbol{x}_q \geq 0\right\}}$$

$$\triangleq \sum_{q=1}^{n} \frac{\partial g\left(\boldsymbol{w}, \boldsymbol{a}, z_q\right)}{\partial z_q} \mathcal{H}_{pq}(t)$$

(21)

Thus, the prediction dynamics are governed by

$$\frac{dz_p(t)}{dt} = \sum_{q=1}^{n} (y_q - z_q(t)) \mathcal{H}_{pq}(t)$$

$$+ \sum_{q=1}^{n} \frac{\partial g\left(\boldsymbol{w}(t), \boldsymbol{a}(t), z_q(t)\right)}{\partial z_q(t)} \mathcal{H}_{pq}(t).$$

(22)

Rearranging the above expression in matrix form, we get the statement of **Lemma 2**. $\square$

Lemma 3. (Hoeffding's inequality, two sided (Vershynin 2018)) *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*

$$\mathbb{P}\left\{ \left| \sum_{r=1}^{m} a_r w_r \right| \geq t \right\} \leq 2 \exp\left(-\frac{t^2}{2 \|\boldsymbol{w}\|_2^2} \right). \quad (23)$$

With probability at least $1 - \delta$, we get the following bound using two-sided Hoeffding's inequality:

$$\left| \sum_{r=1}^{m} a_r w_r \right| \leq \|\boldsymbol{w}\|_2 \sqrt{2 \log\left(\frac{2}{\delta} \right)}. \quad (24)$$

Now, the distance from true labels can be bounded by

$$\frac{d}{dt} \|\boldsymbol{z}(t) - \boldsymbol{y}\|_2^2$$

$$= 2 \left\langle \boldsymbol{z}(t) - \boldsymbol{y}, \frac{d\boldsymbol{z}(t)}{dt} \right\rangle \quad (25)$$

$$= 2 \langle \boldsymbol{z}(t) - \boldsymbol{y}, -\mathcal{H}(t)(\boldsymbol{z}(t) - \boldsymbol{y}) \rangle$$

$$+ 2 \langle \boldsymbol{z}(t) - \boldsymbol{y}, \mathcal{H}(t) \nabla_{\boldsymbol{z}(t)} g(\boldsymbol{w}(t), \boldsymbol{a}(t), \boldsymbol{z}(t)) \rangle$$

Lemma 4. *Suppose **Assumption 1** holds. If we denote $\lambda_{\max}(\mathcal{H}^\infty)$ by λ_1^∞, then $\lambda_{\max}(\mathcal{H}) \leq \frac{\lambda_1}{2} \triangleq \lambda_1^\infty + \frac{\lambda_0}{2}$.*
Proof. As the proof is relatively simpler, we defer it to appendix.

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

$$\frac{d}{dt} \|\boldsymbol{z}(t) - \boldsymbol{y}\|_2^2$$

$$\leq -\lambda_0 \|\boldsymbol{z}(t) - \boldsymbol{y}\|_2^2$$

$$+ \lambda_1 \langle \boldsymbol{z}(t) - \boldsymbol{y}, \nabla_{\boldsymbol{z}(t)} g(\boldsymbol{w}(t), \boldsymbol{a}(t), \boldsymbol{z}(t)) \rangle$$

$$\leq -\lambda_0 \|\boldsymbol{z}(t) - \boldsymbol{y}\|_2^2$$

$$+ \lambda_1 \underbrace{\|\boldsymbol{z}(t) - \boldsymbol{y}\|_2 \|\nabla_{\boldsymbol{z}(t)} g(\boldsymbol{w}(t), \boldsymbol{a}(t), \boldsymbol{z}(t))\|_2}_{\text{Cauchy-Schwarz inequality}}$$

$$\leq -\lambda_0 \|\boldsymbol{z}(t) - \boldsymbol{y}\|_2^2$$

$$+ \lambda_1 \|\boldsymbol{z}(t) - \boldsymbol{y}\|_2 \|\nabla_{\boldsymbol{z}(t)} g(\boldsymbol{w}(t), \boldsymbol{a}(t), \boldsymbol{z}(t))\|_1$$

$$\leq -\lambda_0 \|\boldsymbol{z}(t) - \boldsymbol{y}\|_2^2$$

$$+ \lambda_1 \|\boldsymbol{z}(t) - \boldsymbol{y}\|_2 \sum_{q=1}^{n} \left| \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r w_r \mathbf{1}_{\left\{w_r z_q \geq 0\right\}} \right|$$

$$\leq -\lambda_0 \|\boldsymbol{z}(t) - \boldsymbol{y}\|_2^2$$

$$+ \lambda_1 \|\boldsymbol{z}(t) - \boldsymbol{y}\|_2 \frac{n}{\sqrt{m}} \left| \sum_{r=1}^{m} a_r w_r \right|$$

(26)

Substituting equation (24) in equation (26), we get

$$\frac{d}{dt} \|\boldsymbol{z}(t) - \boldsymbol{y}\|_2^2$$

$$\leq -\lambda_0 \|\boldsymbol{z}(t) - \boldsymbol{y}\|_2^2 + \lambda_1 \|\boldsymbol{z}(t) - \boldsymbol{y}\|_2 \frac{n}{\sqrt{m}} \|\boldsymbol{w}\|_2 \sqrt{2 \log\left(\frac{2}{\delta} \right)}$$

$$\leq -\lambda_0 \|\boldsymbol{z}(t) - \boldsymbol{y}\|_2^2 + \frac{\lambda_1 L n \sqrt{2 \log\left(\frac{2}{\delta} \right)}}{\sqrt{m}} \|\boldsymbol{z}(t) - \boldsymbol{y}\|_2.$$

(27)

Let us define $\mu \triangleq \frac{Ln\sqrt{2\log\left(\frac{2}{\delta}\right)}}{\sqrt{m}}$. Then,

$$\frac{d}{dt}\|\boldsymbol{z}(t)-\boldsymbol{y}\|_2^2 \leq -\lambda_0\|\boldsymbol{z}(t)-\boldsymbol{y}\|_2^2 + \lambda_1\mu\|\boldsymbol{z}(t)-\boldsymbol{y}\|_2 \tag{28}$$

The above non-linear ODE is a special Bernoulli Differential Equation (BDE)[3] which has known exact solutions (Bernoulli 1695). For simplicity, let us suppose $\psi = \|\boldsymbol{z}(t)-\boldsymbol{y}\|_2^2$. Now,

$$\frac{d\psi}{dt} \leq -\lambda_0\psi + \lambda_1\mu\psi^{1/2} \tag{29}$$

Substituting $\psi = \varphi^2$, the BDE is reduced to an Initial Value Problem (IVP): $\frac{d\varphi}{dt} + \frac{\lambda_0}{2}\varphi \leq \frac{\lambda_1}{2}\mu$. By substituting $\varphi = \nu\zeta$, the IVP is decomposed into two linear ODEs of the form $\frac{d\nu}{dt} + \frac{\lambda_0}{2}\nu = 0$ and $\nu\frac{d\zeta}{dt} - \frac{\lambda_1}{2}\mu = 0$. Since these ODEs have separable forms, for arbitrary constants C_1 and C_2, we get

$$\nu = C_1\exp\left(-\frac{\lambda_0 t}{2}\right), \quad \zeta = C_2 + \frac{\kappa\mu}{C_1}\exp\left(\frac{\lambda_0 t}{2}\right), \tag{30}$$

where $\kappa = \frac{\lambda_1}{\lambda_0} = \frac{2\left(\lambda_1^\infty + \frac{\lambda_0}{2}\right)}{\lambda_0} = \mathcal{O}(\kappa^\infty)$. Here, κ^∞ is the condition number of $\mathcal{H}^\infty$. Thus, the solution of the BDE is given by $\psi = \varphi^2 = \left(C\exp\left(-\frac{\lambda_0 t}{2}\right) + \kappa\mu\right)^2$ for another constant C. Using initial value of ψ, we get the exact solution:

$$\|\boldsymbol{z}(t)-\boldsymbol{y}\|_2 \leq (\|\boldsymbol{z}(0)-\boldsymbol{y}\|_2 - \kappa\mu)\exp\left(-\frac{\lambda_0}{2}t\right) + \kappa\mu. \tag{31}$$

From equation (11) in the warm-up exercise, we know for $0 \leq s \leq t$,

$$\left\|\frac{\partial\mathcal{L}_{sup}(\boldsymbol{U},\boldsymbol{v})}{\partial\boldsymbol{u}_j(s)}\right\|_2 \leq \frac{\sqrt{n}}{\sqrt{m}}\|\boldsymbol{z}(s)-\boldsymbol{y}\|_2. \tag{32}$$

Now, substituting equation (31), we get

$$\left\|\frac{\partial\mathcal{L}_{sup}(\boldsymbol{U},\boldsymbol{v})}{\partial\boldsymbol{u}_j(s)}\right\|_2$$
$$\leq \frac{\sqrt{n}}{\sqrt{m}}(\|\boldsymbol{z}(0)-\boldsymbol{y}\|_2 - \kappa\mu)\exp\left(-\frac{\lambda_0}{2}s\right) + \frac{\sqrt{n}}{\sqrt{m}}\kappa\mu. \tag{33}$$

Therefore, the reaction dynamics are given by

$$\mathfrak{R}_j^u(\boldsymbol{u}_j(t)) \leq \frac{\sqrt{n}}{\sqrt{m}}(\|\boldsymbol{z}(0)-\boldsymbol{y}\|_2 - \kappa\mu)\exp\left(-\frac{\lambda_0}{2}t\right) + \frac{\sqrt{n}}{\sqrt{m}}\kappa\mu. \tag{34}$$

Diffusion Term The augmented part on the other hand becomes:

$$\left\|\frac{\partial}{\partial\boldsymbol{u}_j(s)}\sum_{p=1}^n g(\boldsymbol{w},a,z_p)\right\|_2$$
$$= \left\|\sum_{p=1}^n\sum_{r=1}^m\frac{1}{\sqrt{m}}a_r 1_{\{w_r z_p \geq 0\}}w_r\frac{1}{\sqrt{m}}v_j 1_{\{\boldsymbol{v}_j^T\boldsymbol{x}_p \geq 0\}}\boldsymbol{x}_p\right\|_2. \tag{35}$$

---
[3]A Bernoulli differential equation is an ODE of the form $\frac{dx(t)}{dt} + P(t)x(t) = Q(t)x^n(t)$ for $n \in \mathbb{R}\setminus\{0,1\}$.

By Triangle and Cauchy-Schwarz inequality, we get

$$\left\|\frac{\partial}{\partial\boldsymbol{u}_j(s)}\sum_{p=1}^n g(\boldsymbol{w},a,z_p)\right\|_2$$
$$\leq \frac{1}{m}\sum_{p=1}^n\left\|v_j 1_{\{\boldsymbol{v}_j^T\boldsymbol{x}_p \geq 0\}}\boldsymbol{x}_p\sum_{r=1}^m a_r w_r 1_{\{w_r z_p \geq 0\}}\right\|_2$$
$$\leq \frac{1}{m}\sum_{p=1}^n|v_j|\|\boldsymbol{x}_p\|_2\left|\sum_{r=1}^m a_r w_r\right| \tag{36}$$
$$\leq \frac{1}{m}\sum_{p=1}^n\left|\sum_{r=1}^m a_r w_r\right|$$

Substituting equation (24) in equation (36), we arrive at the following inequality:

$$\left\|\frac{\partial}{\partial\boldsymbol{u}_j(s)}\sum_{p=1}^n g(\boldsymbol{w},a,z_p)\right\|_2 \leq \frac{1}{m}\sum_{p=1}^n\|\boldsymbol{w}\|_2\sqrt{2\log\left(\frac{2}{\delta}\right)}$$
$$\leq \frac{Ln\sqrt{2\log\left(\frac{2}{\delta}\right)}}{m} = \mathcal{O}\left(\frac{\mu}{\sqrt{m}}\right). \tag{37}$$

Thus, the diffusion dynamics are given by

$$\mathfrak{D}_j^u(\boldsymbol{u}_j(t)) \leq \frac{Ln\sqrt{2\log\left(\frac{2}{\delta}\right)}}{m}. \tag{38}$$

Now integrating the gradients over $0 \leq s \leq t$,

$$\|\boldsymbol{u}_j(t)-\boldsymbol{u}_j(0)\|_2 \leq \int_0^t\left\|\frac{d\boldsymbol{u}_j(s)}{ds}\right\|_2 ds$$
$$\leq \int_0^t\frac{\sqrt{n}}{\sqrt{m}}(\|\boldsymbol{z}(0)-\boldsymbol{y}\|_2 - \kappa\mu)\exp\left(-\frac{\lambda_0}{2}s\right)ds$$
$$+ \int_0^t\frac{\mu(1+\kappa\sqrt{n})}{\sqrt{m}}ds \tag{39}$$
$$\leq \frac{2\sqrt{n}(\|\boldsymbol{z}(0)-\boldsymbol{y}\|_2 - \kappa\mu)}{\sqrt{m}\lambda_0}\left(1 - \exp\left(-\frac{\lambda_0}{2}t\right)\right)$$
$$+ \left(\frac{\mu(1+\kappa\sqrt{n})}{\sqrt{m}}\right)t.$$

Using Markov's inequality, $\|\boldsymbol{z}(0)-\boldsymbol{y}\|_2 \leq \frac{\mathbb{E}\left[\|\boldsymbol{z}(0)-\boldsymbol{y}\|_2\right]}{\delta} = \mathcal{O}\left(\frac{n}{\delta}\right)$ with probability at least $1-\delta$. Thus,

$$\|\boldsymbol{u}_j(t)-\boldsymbol{u}_j(0)\|_2$$
$$\leq \mathcal{O}\left(\frac{n^{3/2}}{m^{1/2}\lambda_0\delta} + \left(\frac{\mu(1+\kappa\sqrt{n})}{m^{1/2}}\right)t\right). \tag{40}$$

Furthermore, the spatial grid of neurons satisfies:

$$\|\boldsymbol{U}(t)-\boldsymbol{U}(0)\|_F \leq \sqrt{m}\|\boldsymbol{u}_j(t)-\boldsymbol{u}_j(0)\|_2$$
$$\leq \mathcal{O}\left(\frac{n^{3/2}}{\lambda_0\delta} + \mu(1+\kappa\sqrt{n})t\right). \tag{41}$$

To circumvent tractability issues, it is common to seek an ϵ-stationary point. As given by equation (31), $\boldsymbol{z}(t)$ in adversarial learning converges uniformly to an ϵ-neighborhood of $\boldsymbol{y}$ for any $t \geq T_0 \triangleq \frac{2}{\lambda_0} \log \left(\frac{\|\boldsymbol{z}(0) - \boldsymbol{y}\|_2 - \kappa\mu}{\epsilon - \kappa\mu} \right)$. For finite time convergence, we need $\kappa\mu < \epsilon < \|\boldsymbol{z}(0) - \boldsymbol{y}\|_2$. The second inequality holds because we usually look for a solution where the error is better than what obtained during initialization. The first inequality gives the upper bound on gradient penalty, i.e., $L \leq \mathcal{O}\left(\frac{\epsilon \sqrt{m}}{\kappa n \sqrt{2 \log(2/\delta)}} \right)$ by substituting the value of μ. It is an important result in a sense that over-parameterized networks can still enjoy linear rate of convergence even under adversarial interaction. $\qquad \square$

In a general configuration, **Remark 1** asserts that the induced Gram matrix is stable and satisfies our assumptions on eigen values as long as $\|\boldsymbol{u}_j - \boldsymbol{u}_j(0)\| \leq R$. Intuitively, this is satisfied when the points visited by gradient descent in adversarial learning lie within this R-ball. Formally, we need the following condition to be satisfied for finding the least expensive ϵ-stationary point:

$$\mathcal{O}\left(\frac{n^{3/2}}{m^{1/2} \lambda_0 \delta} + \left(\frac{\mu \left(1 + \kappa \sqrt{n}\right)}{m^{1/2}} \right) T_0 \right) \leq R. \qquad (42)$$

Substituting $R = \frac{c\delta\lambda_0}{n^2}$ in the above expression, we get

$$m = \Omega\left(\left(\frac{n^{7/2}}{\lambda_0^2 \delta^2} + \frac{n^2 \mu \left(1 + \kappa \sqrt{n}\right) T_0}{\lambda_0 \delta} \right)^2 \right). \qquad (43)$$

It is worth mentioning that the polynomial node complexity, $m = poly\left(n, \frac{1}{\lambda_0}, \frac{1}{\delta} \right)$ is also essential for finding an ϵ-stationary point in sole supervision. By ignoring the diffusible factors, i.e., setting $\mu = 0$, we recover the lower bound, $m = \Omega\left(\frac{n^7}{\lambda_0^4 \delta^4} \right)$ in supervised learning.

Discussion of Insights from Analysis

A profound implication of this finding is that adversarial learning allows gradient descent to explore a large subspace in contrast to supervised learning where a tiny subspace around initialization is merely explored (Gur-Ari, Roberts, and Dyer 2018). As a result, it offers the provision to exploit full capacity of network architectures by encouraging local interaction. In other words, the neurons in supervised learning do not interact with each other as much as they do in adversarial learning. By introducing the diffusible factors, it helps break the spatial symmetry and homogeneity in this tiny subspace. Due to more local interaction and diffusion, it exhibits patterns as a reminiscent of those observed in nature. More importantly, this is consistent with the well-studied theory of pattern formation (Turing 1952; Meinhardt 1982; Gray and Scott 1984; Rauch and Millonas 2004).

The system of neurons is initially in a stable homogeneous condition due to non-diffusive elements in sole supervision. It is perturbed by irregularities introduced under the influence of an adversary. For a RD system, it is necessary that these irregularities are small enough, which otherwise would destabilize the whole system, and it may never converge to a reasonable solution. This is easily satisfied in over-parameterized networks as given by equation (38). Thus, it is not unreasonable to suppose that adversarial interaction in augmented objective is the only one in which conditions are such to break the spatial symmetry. Different from strict RD systems, the diffusibility here does not directly depend on Laplacian of each morphogen. This is not uncommon because bell-like pattern formation in the skin of a zebrafish is a typical example where patterns emerge even when the system is different from the original Turing model (Nakamasu et al. 2009). More importantly, it fits the description of short and long range feedback which indicates a similar mechanism must be involved in adversarial learning. This analogy provides positive support to the developed PRD theory.

It is well known that randomly initialized gradient descent with over-parameterization finds solutions close to its initialization (Du et al. 2018; Li and Liang 2018; Neyshabur et al. 2018; Nagarajan and Kolter 2019). The distance from initialization has helped unveil several mysteries of deep learning in part including the generalization puzzle and ϵ-stationarity. We ask whether such implicit restriction to a tiny search space is a *necessary condition* to achieve similar performance. The expressive power of a large network is not fully exploited by limiting the search space. This argument is supported by Gulrajani et al. who show that the generator in WGAN with weight clipping (Arjovsky, Chintala, and Bottou 2017) fails to capture higher order moments. One reason for such behavior is the implicit restriction of discriminator weights to a tiny subspace around extremas due to weight clipping. It is resolved however by incorporating gradient penalty which allows exploration in a larger search space within clipping boundaries. In this regard, we provide both theoretical and empirical evidence that imposing such restriction is not a necessary condition. With over-parameterization, randomly initialized gradient descent can still find a global minimizer relatively farther from its initialization. It is possible because of adversarial interaction that helps introduce diffusible factors into the system.

Conclusion & Future Work

In this paper, we studied the evolutionary patterns formed in a system of neurons with adversarial interaction. We provided a theoretical justification and empirical evidence of Turing instability in a pseudo-reaction-diffusion model that underlie these phenomena. Furthermore, it was shown that randomly initialized gradient descent with over-parameterization could still enjoy exponentially fast convergence to an ϵ-stationary point even under adversarial interaction. However, unlike sole supervision, it was found that the obtained solutions were not limited to a tiny subspace around initialization. It was observed that adversarial interaction helped in the breakdown of spatial symmetry and homogeneity which allowed exploration in a larger subspace.

While this work takes a step towards explaining non-homogeneous pattern formation due to adversarial interaction, it is far from being conclusive. Though diffusibility ensures more local interaction, it will certainly be interesting to synchronize neurons based on this observation in future.

Broader Impact

This paper investigates the underlying phenomena that may cause evolutionary patterns to emerge with the advent of adversarial interaction. By theoretical and empirical evidence, it tries to corroborate the developed pseudo-reaction-diffusion system. We believe this work does not present any forseeable societal consequence.

References

Allen-Zhu, Z.; and Li, Y. 2020. Backward feature correction: How deep learning performs deep learning. *arXiv preprint arXiv:2001.04413* .

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning*.

Bernoulli, J. 1695. Explicationes, Annotationes & Additiones ad ea, quae in Actis sup. de Curva Elastica, Isochrona Paracentrica, & Velaria, hinc inde memorata, & paratim controversa legundur; ubi de Linea mediarum directionum, alliisque novis. *Acta Eruditorum* .

Budzynski, T. H.; Budzynski, H. K.; Evans, J. R.; and Abarbanel, A. 2009. *Introduction to quantitative EEG and neurofeedback: Advanced theory and applications*. Academic Press.

Du, S. S.; Zhai, X.; Poczos, B.; and Singh, A. 2018. Gradient Descent Provably Optimizes Over-parameterized Neural Networks. In *International Conference on Learning Representations*.

Engin, D.; Genç, A.; and Kemal Ekenel, H. 2018. Cycledehaze: Enhanced cyclegan for single image dehazing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 825–833.

Frey, B. J.; and Dueck, D. 2006. Mixture modeling by affinity propagation. In *Advances in neural information processing systems*, 379–386.

Frey, B. J.; and Dueck, D. 2007. Clustering by passing messages between data points. *science* 315(5814): 972–976.

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 249–256.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Gray, P.; and Scott, S. 1984. Autocatalytic reactions in the isothermal, continuous stirred tank reactor: Oscillations and instabilities in the system A+ 2B to 3B; B to C. *Chemical Engineering Science* 39(6): 1087–1097.

Gregor, T.; Tank, D. W.; Wieschaus, E. F.; and Bialek, W. 2007. Probing the limits to positional information. *Cell* 130(1): 153–164.

Gulrajani, I.; Ahmed, F.; vsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved training of wasserstein gans. In *Advances in neural information processing systems*, 5767–5777.

Gur-Ari, G.; Roberts, D. A.; and Dyer, E. 2018. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754* .

Karacan, L.; Akata, Z.; Erdem, A.; and Erdem, E. 2016. Learning to generate images of outdoor scenes from attributes and semantic layouts. *arXiv preprint arXiv:1612.00215* .

Kondo, S.; and Miura, T. 2010. Reaction-diffusion model as a framework for understanding biological pattern formation. *science* 329(5999): 1616–1620.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4681–4690.

Lei, Q.; Lee, J. D.; Dimakis, A. G.; and Daskalakis, C. 2019. Sgd learns one-layer networks in wgans. *arXiv preprint arXiv:1910.07030* .

Li, Y.; and Liang, Y. 2018. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, 8157–8166.

Luc, P.; Couprie, C.; Chintala, S.; and Verbeek, J. 2016. Semantic Segmentation using Adversarial Networks. In *NIPS Workshop on Adversarial Training*.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9(Nov): 2579–2605.

Meinhardt, H. 1982. Models of biological pattern formation. *New York* 118.

Meinhardt, H.; and Gierer, A. 1974. Applications of a theory of biological pattern formation based on lateral inhibition. *Journal of cell science* 15(2): 321–346.

Meinhardt, H.; and Gierer, A. 2000. Pattern formation by local self-activation and lateral inhibition. *Bioessays* 22(8): 753–760.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* .

Nagarajan, V.; and Kolter, J. Z. 2019. Generalization in deep networks: The role of distance from initialization. *arXiv preprint arXiv:1901.01672* .

Nakamasu, A.; Takahashi, G.; Kanbe, A.; and Kondo, S. 2009. Interactions between zebrafish pigment cells responsible for the generation of Turing patterns. *Proceedings of the National Academy of Sciences* 106(21): 8429–8434.

Neyshabur, B.; Li, Z.; Bhojanapalli, S.; LeCun, Y.; and Srebro, N. 2018. The role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*.

Rauch, E. M.; and Millonas, M. M. 2004. The role of trans-membrane signal transduction in Turing-type cellular pattern formation. *Journal of theoretical biology* 226(4): 401–407.

Rogers, E. 2003. Diffusion of innovations . Delran. *NJ: Simon & Schuster. Schneider, L.(1971). Dialectic in sociology. American Sociological Review* 36: 667678.

Rout, L. 2020. Alert: Adversarial learning with expert regularization using tikhonov operator for missing band reconstruction. *IEEE Transactions on Geoscience and Remote Sensing* 58(6): 4395–4405.

Rout, L.; Misra, I.; Manthira Moorthi, S.; and Dhar, D. 2020. S2A: Wasserstein GAN with Spatio-Spectral Laplacian Attention for Multi-Spectral Band Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 188–189.

Sarmad, M.; Lee, H. J.; and Kim, Y. M. 2019. Rl-gan-net: A reinforcement learning agent controlled gan network for real-time point cloud shape completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5898–5907.

Saxe, A.; McClelland, J.; and Ganguli, S. 2014. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. International Conference on Learning Represenatations 2014.

Sayama, H. 2015. *Introduction to the modeling and analysis of complex systems*. Open SUNY Textbooks.

Tsuchida, R.; Roosta, F.; and Gallagher, M. 2018. Invariance of weight distributions in rectified MLPs. In *International Conference on Machine Learning*, 4995–5004.

Turing, A. 1952. THE CHEMICAL BASIS OF MORPHO-GENESIS. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 237(641): 37–72.

Verhulst, P.-F. 1838. Notice sur la loi que la population suit dans son accroissement. *Corresp. Math. Phys.* 10: 113–126.

Vershynin, R. 2018. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wang, X.; and Gupta, A. 2016. Generative image modeling using style and structure adversarial networks. In *European conference on computer vision*, 318–335. Springer.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; and Change Loy, C. 2018. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 0–0.

Wolpert, L.; Tickle, C.; and Arias, A. M. 2015. *Principles of development*. Oxford University Press, USA.

Xie, B.; Liang, Y.; and Song, L. 2017. Diverse neural network learns true target functions. In *Artificial Intelligence and Statistics*, 1216–1224.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

Appendix
More Related Works

Reaction-Diffusion Systems

The original RD model is a simplification and an idealization of practical systems whose complexity makes it hard to understand the phenomena. With slight modification to the theory, one may easily extend this mathematical analysis to explain pattern formation in real world systems. In addition, it can generate limitless variety of patterns depending on the parameters in the reaction and diffusion terms.

Numerous methods seek to explain pattern formation in complex systems. Among many reasonable attempts, one that experimental biologists may recall is gradient model (Wolpert, Tickle, and Arias 2015). Different from RD model, it assumes a fixed source of morphogens that provides positional information. In other words, it can be designed as a special case of RD model by carefully choosing the boundary conditions. Experiments have shown the necessity of molecular interaction and boundary condition to create more realistic patterns (Meinhardt 1982). To model interactions of molecular elements in gradient analysis, (Gregor et al. 2007) developed a framework that is essentially similar to RD model.

Concerted efforts have been made towards extension and identification of root causes to explain pattern formation. The fact that a short range positive feedback and a long range negative feedback are enough to generate Turing patterns is indeed a big revelation in this direction (Meinhardt and Gierer 1974, 2000). This refinement helps envision a wide variety of patterns in more complex systems.

Particularly intriguing is the fact that these interacting elements need not be limited to molecules. The interaction between cellular signals also generates Turing patterns (Nakamasu et al. 2009). Further, there is no restriction on how the system diffuses to break spatial symmetry. A relayed series of cell to cell signal transmission may induce diffusible factors in a system (Rauch and Millonas 2004). All these scenarios have a common ground in a sense that these systems exhibit a short range positive feedback and a long range negative feedback similar to adversarial framework.

Bernoulli Differential Equation

Bernoulli differential equation was discussed by Bernoulli in 1695. This fundamental equation arises naturally in a wide variety fields, such as modelling of population growth (Verhulst 1838), modelling of a pandemic, modelling of growth of tumors, Fermi-Dirac statistics, modelling of crop response, and modelling of diffusion of innovations in economics and sociology (Rogers 2003). In Verhulst model of population growth, the rate of reproduction is proportional to current population and available resources. Formally,

$$\frac{dP}{dt} = rP\left(1 - \frac{P}{K}\right), \tag{44}$$

where P, r, and K denote population size, rate of growth, and carrying capacity, respectively. In ecology, N is often used in place of P to represent population. An interesting

theory, namely r/K selection theory builds on simplified Verhulst model by drawing r and K from ecological algebra. In the similar spirit, the presented PRD model arrives at a special Bernoulli differential equation where error being the population size in the modelling of population growth. Different from traditional settings, the rate of change here is proportional to the square root of current error. To put more succinctly,

$$\frac{d\psi}{dt} \leq r\psi^{1/2}\left(1 - \frac{\psi^{1/2}}{K}\right), \tag{45}$$

where $r = \lambda_1\mu$ and $K = \kappa\mu$. The interpretation of this equation is reversed in the present analysis as we are interested in the decay of total error. Nevertheless, the compact representation captures the essence of reaction and diffusion dynamics.

Jointly Training Both Layers

In this section, we extend theoretical analyses from a single layer scalar network architecture to jointly training both layers with multiple classes. For simplicity, let $\boldsymbol{z}_p$ denotes $\frac{1}{\sqrt{d_{out}m}}\boldsymbol{V}\sigma(\boldsymbol{U}\boldsymbol{x}_p)$.

Definition 2. *Let us define* $\mathfrak{R}_j^{\boldsymbol{u}}(\boldsymbol{u}_j, \boldsymbol{v}_j)$ *and* $\mathfrak{R}_j^{\boldsymbol{v}}(\boldsymbol{u}_j, \boldsymbol{v}_j)$ *as the reaction terms in hidden and top layer, respectively. Also, let* $\mathfrak{D}_j^{\boldsymbol{u}}(\nabla^2\boldsymbol{u}_j)$ *and* $\mathfrak{D}_j^{\boldsymbol{v}}(\nabla^2\boldsymbol{v}_j)$ *denote the diffusion terms in hidden and top layer, respectively.*

Theorem 3. (Reaction-Diffusion Dynamics) *If we absorb constants in* $\mathcal{O}(.)$ *and set* $(\boldsymbol{y}_p - \boldsymbol{z}_p)_i\, v_{ij}\mathbb{1}_{\left\{\boldsymbol{u}_j^T\boldsymbol{x}_p \geq 0\right\}}x_{p,k} = \mathcal{O}(1)$ *for* $i \in [d_{out}]$ *and* $p \in [n]$*, then for all* $j \in [m]$ *the RD dynamics satisfy:*

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

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

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

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

Proof. See next section. The diffusion terms are greatly affected by other morphogens in the system, suggesting a special case scenario of Turing's RD model. To put more succinctly, $\mathfrak{D}_j^{\boldsymbol{u}}$ and $\mathfrak{D}_j^{\boldsymbol{v}}$ are dominated by $\boldsymbol{v}_j$ and $\boldsymbol{u}_j$, respectively. While the asymptotic reaction terms are bounded by similar norms, the apparent difference between diffusion terms explains why we observe different patterns in the hidden and top layer.

Technical Proofs

Proof of Lemma 1

With probability at least $(1 - \delta)$, by Markov's inequality, we get

$$|u_{jk}(0)| \leq \frac{\mathbb{E}\left[|u_{jk}(0)|\right]}{\delta} = \frac{2}{\sqrt{2\pi}\delta}. \tag{46}$$

We use matrix norm properties to bound the Frobenius norm of $U(0)$:

$$\|U(0)\|_F = \left(\sum_{j=1}^{m} \sum_{k=1}^{d_{in}} |u_{jk}(0)|^2 \right)^{1/2}$$

$$\leq \left(\sum_{j=1}^{m} \sum_{k=1}^{d_{in}} \frac{4}{2\pi\delta^2} \right)^{1/2} \quad (47)$$

$$\leq \frac{2\sqrt{md_{in}}}{\sqrt{2\pi}\delta} \triangleq \zeta.$$

This finishes the proof of the statement. $\square$

Proof of Lemma 4

For clarity, let us recall **Lemma 3.1** of Du et al.: *If $m = \Omega\left\{ \frac{n^2}{\lambda_0^2} \log\left(\frac{n}{\delta}\right) \right\}$, then we have with high probability $1 - \delta$, $\|\mathcal{H}(0) - \mathcal{H}^\infty\|_2 \leq \frac{\lambda_0}{4}$ and $\lambda_{\min}\left(\mathcal{H}(0)\right) \geq \frac{3}{4}\lambda_0$.* From **Remark 1**, we know

$$\|\mathcal{H}\|_2 - \|\mathcal{H}(0)\|_2 \leq \|\mathcal{H} - \mathcal{H}(0)\|_2 \leq \frac{\lambda_0}{4}. \quad (48)$$

Using similar arguments, we get

$$\|\mathcal{H}(0)\|_2 - \|\mathcal{H}^\infty\|_2 \leq \|\mathcal{H}(0) - \mathcal{H}^\infty\|_2 \leq \frac{\lambda_0}{4}, \quad (49)$$

which implies $\lambda_{\max}(\mathcal{H}(0)) \leq \lambda_{\max}(\mathcal{H}^\infty) + \frac{\lambda_0}{4}$. By plugging this, the expression gets simplified to

$$\lambda_{\max}\left(\mathcal{H}\right) \leq \lambda_{\max}\left(\mathcal{H}^\infty\right) + \frac{\lambda_0}{4} + \frac{\lambda_0}{4}$$

$$\leq \lambda_1^\infty + \frac{\lambda_0}{2} \triangleq \frac{\lambda_1}{2}. \quad (50)$$

This justifies the upper bound assumption of the largest eigen value over iterations. $\square$

Proof of Theorem 3

Our first step is to derive dynamics of weights due to supervised cost. In the hidden layer, the weights are updated by the following PDE.

$$\frac{\partial \mathcal{L}_{sup}(U,V)}{\partial u_{jk}}$$

$$= \frac{1}{2} \sum_{p=1}^{n} \frac{\partial}{\partial u_{jk}} \sum_{i=1}^{d_{out}} (z_p - y_p)_i^2$$

$$= \sum_{p=1}^{n} \sum_{i=1}^{d_{out}} (z_p - y_p)_i \frac{\partial z_{p,i}}{\partial u_{jk}} \quad (51)$$

$$= \sum_{p=1}^{n} \sum_{i=1}^{d_{out}} (z_p - y_p)_i \frac{1}{\sqrt{d_{out}m}} v_{ij} \frac{\partial}{\partial u_{jk}} \sigma\left(u_j^T x_p\right)$$

$$= \frac{1}{\sqrt{d_{out}m}} \sum_{p=1}^{n} \sum_{i=1}^{d_{out}} (z_p - y_p)_i \, v_{ij} 1_{\left\{ u_j^T x_p \geq 0 \right\}} x_{p,k}.$$

Next, we calculate dynamics of weights in the top layer.

$$\frac{\partial \mathcal{L}_{sup}(U,V)}{\partial v_{ij}}$$

$$= \frac{1}{2} \sum_{p=1}^{n} \frac{\partial}{\partial v_{ij}} \sum_{i=1}^{d_{out}} (z_p - y_p)_i^2$$

$$= \sum_{p=1}^{n} (z_p - y_p)_i \frac{\partial z_{p,i}}{\partial v_{ij}} \quad (52)$$

$$= \sum_{p=1}^{n} (z_p - y_p)_i \frac{1}{\sqrt{d_{out}m}} \frac{\partial}{\partial v_{ij}} \sum_{j=1}^{m} v_{ij} \sigma\left(u_j^T x_p\right)$$

$$= \frac{1}{\sqrt{d_{out}m}} \sum_{p=1}^{n} (z_p - y_p)_i \, 1_{\left\{ u_j^T x_p \geq 0 \right\}} u_j^T x_p.$$

Now, we proceed to compute dynamics of weights due to adversarial regularization. In the hidden layer, the weights obey the following dynamics:

$$\frac{\partial \mathcal{L}_{adv}(U,V,W,a)}{\partial u_{jk}}$$

$$= \frac{1}{m\sqrt{d_{out}}} \sum_{p=1}^{n} a^T \text{diag}\left(1_{\{WV\sigma(Ux_p) \geq 0\}}\right) Wv_j 1_{\left\{ u_j^T x_p \geq 0 \right\}} x_{p,k}. \quad (53)$$

The weights in the top layer are governed by:

$$\frac{\partial \mathcal{L}_{adv}(U,V,W,a)}{\partial v_{ij}}$$

$$= \frac{1}{m\sqrt{d_{out}}} \sum_{p=1}^{n} a^T \text{diag}\left(1_{\{WV\sigma(Ux_p) \geq 0\}}\right) W_{:,i} 1_{\left\{ u_j^T x_p \geq 0 \right\}} u_j^T x_p. \quad (54)$$

Analogous to equation (5), the reaction and diffusion terms in augmented objective are defined as:

$$\mathfrak{R}_j^u(u_j,v_j)$$

$$\triangleq \left\{ \frac{1}{\sqrt{d_{out}m}} \sum_{p=1}^{n} \sum_{i=1}^{d_{out}} (y_p - z_p)_i \, v_{ij} 1_{\left\{ u_j^T x_p \geq 0 \right\}} x_{p,k} \right\}_{k=1}^{d_{in}}, \quad (55)$$

$$\mathfrak{D}_j^u(\nabla^2 u_j)$$

$$\triangleq \left\{ \frac{1}{m\sqrt{d_{out}}} \sum_{p=1}^{n} a^T \text{diag}\left(1_{\{WV\sigma(Ux_p) \geq 0\}}\right) Wv_j 1_{\left\{ u_j^T x_p \geq 0 \right\}} x_{p,k} \right\}_{k=1}^{d_{in}}, \quad (56)$$

$$\mathfrak{R}_j^v(u_j,v_j)$$

$$\triangleq \left\{ \frac{1}{\sqrt{d_{out}m}} \sum_{p=1}^{n} (y_p - z_p)_i \, 1_{\left\{ u_j^T x_p \geq 0 \right\}} u_j^T x_p \right\}_{i=1}^{d_{out}}, \quad (57)$$

$$\mathfrak{D}_j^v(\nabla^2 v_j)$$

$$\triangleq \left\{ \frac{1}{m\sqrt{d_{out}}} \sum_{p=1}^{n} a^T \text{diag}\left(1_{\{WV\sigma(Ux_p) \geq 0\}}\right) W_{:,i} 1_{\left\{ u_j^T x_p \geq 0 \right\}} u_j^T x_p \right\}_{i=1}^{d_{out}}. \quad (58)$$

Ignoring constants and assuming $(\boldsymbol{y}_p - \boldsymbol{z}_p)_i \, v_{ij} \mathbb{1}_{\{\boldsymbol{u}_j^T \boldsymbol{x}_p \geq 0\}} x_{p,k} = \mathcal{O}(1)$, we get asymptotic bounds on the norm of reaction and diffusion terms[4]:

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

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

$$\mathfrak{R}_j^{\boldsymbol{v}}(\boldsymbol{u}_j, \boldsymbol{v}_j) = \mathcal{O}\left(n d_{in} \sqrt{\frac{d_{out}}{m}}\right),$$ (59)

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

which completes the proof. □

Experiments

Our experiments aim to answer the following questions:

- (**Linear Rate**) *How does randomly initialized gradient descent converge exponentially fast to a solution in a larger subspace around initialization?*

- (**Theorem 1**) *How does purely supervised cost maintain symmetry and homogeneity over iterations?*

- (**Theorem 2**) *How does adversarial interaction help break this symmetry and homogeneity?*

Training Details

Architecture We use two layer neural network with ReLU activation in all the experiments. Refer to section **Preliminaries** for mathematical details on the network architecture. The discriminator is updated once per each generator update. In addition to gradient penalty, we set $L = 0.01$ for the discriminator. The primary results are reported with the number of hidden units, $m = 2^{13}$ to capture over-parameterization. Also, we experiment with 13 different hidden units ranging from $m = 2^3$ to 2^{15} so as to study the generalization of our theory. The input to the network is $d_{in} = 784$ on synthetic dataset, MNIST and FashionMNIST. We also study the effect of $d_{in} = 1024$, which is like gray scale CIFAR10 dataset, on pattern formation. The output of the network is set to $d_{out} = 10$ in all experiments. The specific choice of scalar network, i.e., $d_{out} = 1$ is made for mathematical convenience though we demonstrate empirically that the theory can be successfully applied to analyze more complex networks in less restrictive setting. A system with 64GB RAM, one V100 gpu and PyTorch library is used for all the experiments reported in this paper.

Synthetic Datasets It has been established in multitude of tasks that natural images often lie on a low dimensional manifold. To emulate this structure, we have created a synthetic dataset: $\{(\boldsymbol{x}_p, \boldsymbol{y}_p)\}_{p=1}^n \subset \mathbb{R}^{784} \times \mathbb{R}^{10}$. To enforce data

---

[4]More precisely, one may choose a generator to have different number of hidden units than discriminator. In that case, the asymptotic bounds may contain m_{dis} and m_{gen}. To simplify the expression and focus more on analysis, we assume equal number of hidden units in generator and discriminator.
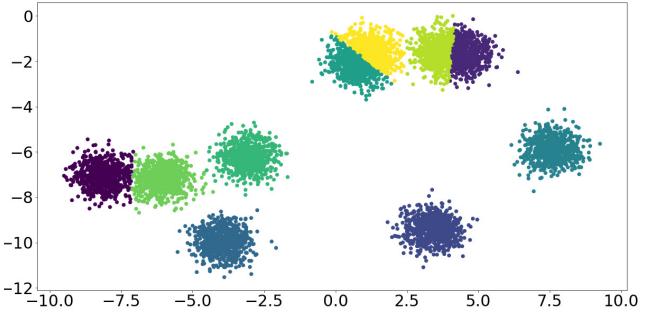
Figure 1: Synthetic dataset on 2-dimensional manifold.

points to lie on a low dimensional manifold (here, 2), the initial two elements of $\boldsymbol{x}_p$ are sampled from a mixture of Gaussian distribution having $d_{in} = 10$ different modes similar to MNIST, FashionMNIST and CIFAR10. The rest of the elements of $\boldsymbol{x}_p$ are set to 1. In another dataset, we set $d_{in} = 1024$ with initial two elements sampled from the same mixture of Gaussian distribution. In both these datasets, we have $n = 7000$ samples with 6000 used for training and remaining 1000 used for testing. We use a batch size of 256, a learning rate of $1e-2$, and train for 1000 epochs using SGD optimizer with momentum 0.9. The random seed is set to 1 in the mixture of Gaussian distribution (ref. `make_blobs` from `sklearn`). Figure 1 illustrates different classes of this toy dataset.

MNIST and FashionMNIST Datasets Here, out of $n = 70000$ samples of size $[28 \times 28]$, we select 60000 for training and 10000 for testing purpose. We use a batch size of 512, a learning rate of $1e-2$, and train for 1000 epochs or until convergence using SGD optimizer with momentum 0.9. In this setting, $d_{in} = 784$. We assume convergence when the training error reaches at least 0.001.

CIFAR10 Dataset Out of 60000 samples of size $[32 \times 32 \times 3]$, we select 50000 for training and remaining 10000 for testing. Here, d_{in} is set to 3072. Other than the convergence criterion which is set to 0.02, all the hyperparameters are identical to MNIST and FashionMNIST.

Experimental Results

The empirical evidence is outlined as following. First, we demonstrate that the spatial grid of neurons follows an exponential upper bound as predicted by our theory. Second, we show that the system of neurons in sole supervision preserves symmetry and homogeneity over iterations as per **Theorem 1**. Finally, we verify **Theorem 2** by showcasing Turing-like patterns under the influence of adversarial interaction. More importantly, these computer simulations on toy datasets resemble Turing-like patterns formed while training on real datasets, suggesting both share similar underlying principles. An artistic view of the breakdown of symmetry and how it leads to pattern formation as a result of local interaction is shown in Figure 2.

Results on Synthetic Datasets While the experiments suggest approximately 2300 iterations are enough to achieve
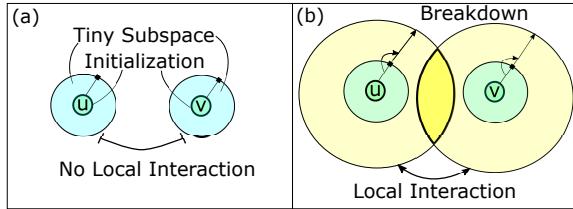
Figure 2: Breakdown of symmetry and homogeneity leads to local interaction. (a) Without Diffusion. (b) With Diffusion.
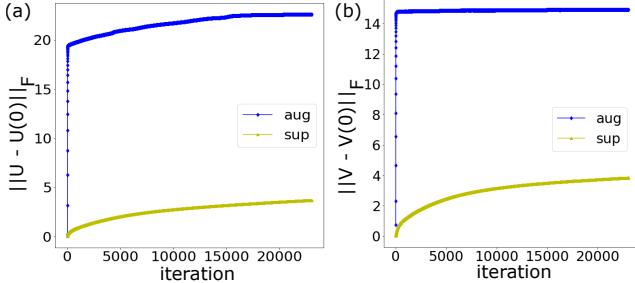
Figure 3: Distance from initialization in the (a) hidden layer (b) top layer over iterations on MNIST-like ($d_{in} = 784$) toy dataset.

Figure 4: Distance from initialization in the *hidden layer* on (a) MNIST-like ($d_{in} = 784$) and (b) CIFAR10-like ($d_{in} = 1024$) toy datasets.

Figure 5: Distance from initialization in the *top layer* on (a) MNIST-like ($d_{in} = 784$) and (b) CIFAR10-like ($d_{in} = 1024$) toy datasets.

an acceptable test error (here, $\epsilon = 0.016$) on this toy dataset, we continued training up to 23000 iterations to study the impact of μt on distance from initialization. As shown in Figure 3, the upper bound is dominated by the magnitude of exponential term, i.e., $\mathcal{O}\left(\frac{n^{3/2}}{\lambda_0 \delta}\right)$. Observe that μ is an infinitesimally small quantity in over-parameterized network which justifies tiny linear slope. In fact, diffusion coefficients are often orders of magnitude smaller than reaction coefficents (Turing 1952; Gray and Scott 1984)[5]. Unless the model is trained for infinite iterations, it is fair to assume the settling point of the upper bound to be $\mathcal{O}\left(\frac{n^{3/2}}{\lambda_0 \delta}\right)$ with a very small tolerance of $\mathcal{O}(\mu T_0)$. To further our understanding of exponential rate, we analyze initial few iterations on MNIST- and CIFAR10-like toy datasets. Figure 4 and 5 support our theoretical analysis as we observe that $\|\boldsymbol{U}(t) - \boldsymbol{U}(0)\|_F$ and $\|\boldsymbol{V}(t) - \boldsymbol{V}(0)\|_F$ are of the form $\alpha\left(1 - \exp(-\beta t)\right)$ for $\alpha > 0$ and $\beta > 0$.

Results on MNIST As shown in Figure 6, we observe breakdown of symmetry to explore larger subspace on MNIST. In addition, we study the dependency of these phenomena on width in Figure 7 and random initialization in Figure 8 to gain a better understanding.

Dependence on Width. While the settling point is different in $m = 2^{14}$ as compared to $m = 2^{13}$, it still provides empirical support to exponential breakdown due to adversarial interaction.

Dependence on Initialization. To investigate robustness,

---

[5]The reader is referred to the simulated patterns by RD model and Gray-Scott model in later parts of this paper where the impact of tiny diffusibility is discernible.

it is essential to run the experiments multiple times with random seeds. For this reason, we run the experiments 5 times each for supervised and adversarial learning, i.e., without and with diffusible factors. In adversarial setting, the PRD model is shown to break the symmetric and homogeneous barrier exponentially fast. Further, the experiments are consistent over multiple initializations. Though the variance appears to be quite large in hidden layer, it is relatively small ($< 2.5\%$) compared to the settling point in each trajectory. Nevertheless, it agrees with exponential rate in all the experiments.

Error Analysis. It is well known that two layer neural networks with over-parameterization can find an ϵ-stationary point on MNIST in a finite time (Neyshabur et al. 2018; Nagarajan and Kolter 2019). For clarity purpose, we conduct experiments in this problem setting. Figure 9 provides further evidence of this phenomena and admits the theoretical results.

Weight Analysis. In Figure 10, we plot hidden layer filters, $\boldsymbol{u}_j$ with top-9 L_2 norm. An interesting observation is the organization of weights in the spatial grid of size $[28 \times 28]$. Since the breakdown of symmetry allows more local interaction, the neurons coalesce in the decision making process. As shown in Figure 10, neurons in the hidden layer learn to focus on relevant parts of the input image in adversarial setting. This is in contrast to supervised setting where neurons pay attention to all spatial locations. To quantitatively analyze the non-homogeneity of a single node, the variance is shown alongside the hidden unit. The high vari-
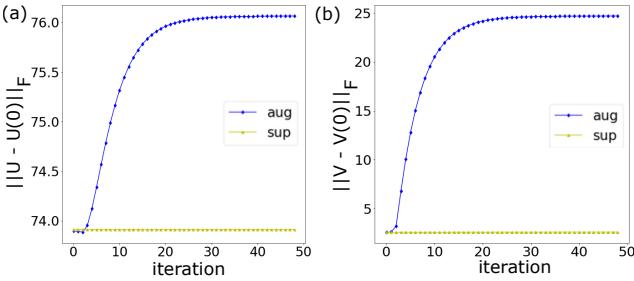
Figure 6: Distance from initialization in the (a) hidden layer (b) top layer on MNIST.
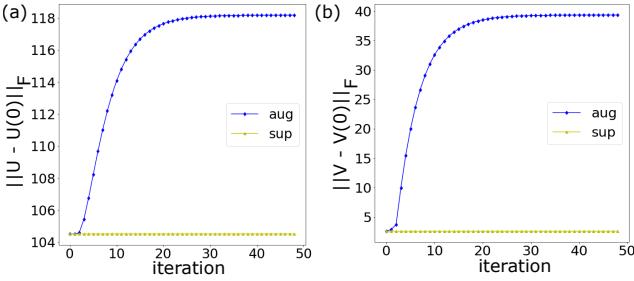
Figure 7: Distance from initialization in the (a) hidden layer (b) top layer on MNIST with 2^{14} hidden units.

Figure 8: Distance from multiple initialization in the (a) hidden layer and (b) top layer on MNIST.
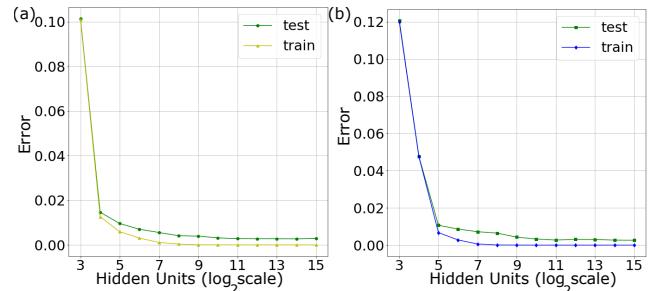
Figure 9: Training and testing error on MNIST. (a) Without Diffusion. (b) With Diffusion.

ance in advesarial setting captures non-homogeneity up to some extent.

Feature Visualization. Here, we try to understand the sensitivity of neurons to perturbations in the input space. For this reason, a common random input image is fed as input to both the systems without and with diffusion. The output of a particular neuron in the hidden layer is maximized subject to perturbations in the input image (Allen-Zhu and Li 2020). Formally, for all $j \in [2^{13}]$,

$$\delta_j = \arg\max_{\delta \in \Delta} \boldsymbol{u}_j^T (x + \delta), \tag{60}$$

where x denotes a random vector in $\mathbb{R}^{d_{in}}$, δ denotes the perturbation, and Δ is the allowed set of perturbation. While one can explore different choices of allowed pertubation, a common choice, which we use in feature visualization, is l_∞ ball:

$$\Delta = \{\delta : \|\delta\|_\infty \le \epsilon\} \tag{61}$$

The l_∞ norm of a vector in $\mathbb{R}^{d_{in}}$ is defined as:

$$\|\delta\|_\infty = \max_k |\delta_k|, \tag{62}$$

where $k \in [d_{in}]$. Basically, the perturbation in each component is restricted to $[-\epsilon, \epsilon]$. Here, we choose $\epsilon = 0.007$, and run gradient descent with learning rate $1e - 1$ for 100 iterations. The extracted features are reshaped to $[28 \times 28]$ for visualization purpose.

As shown in Figure 11, maximization of excitation without diffusion leads to perturbation in every spatial location of the input image. In contrast to that, the system with diffusion leads to perturbations in locations where essential information is present. From another perspective, this indicates

excitation when the input images are pushed towards actual manifold of MNIST data. The fact that both start from the same initial point asserts this intriguing property of adversarial interaction. The input image is shown in Figure 12.

Results on FashionMNIST As shown in Figure 13, the distance from initialization is larger than synthetic datasets in both supervised and adversarial learning. One reason for such large deviation is the increased complexity on FashionMNIST. Moreover, it follows exponential trend in augmented objective and explores a larger subspace around initialization as predicted by our theory.

Weight Analysis. Similar to MNIST, we observe the self-organization tendency even on FashionMNIST. As shown in Figure 14, local interaction with the advent of diffusion allows neurons in the hidden layer to cooperate while operating on the input samples.

Feature Visualization. Figure 15 provides further evidence of non-homogeneity in learned features. Recall that the task here is not to generate realistic looking images, but classify them into different categories. The discriminator does not have access to real images in this setting. It only sees the predicted and true labels in $\mathbb{R}^{10}$. Interestingly, adversarial interaction allows to learn the manifold of actual data while performing the desired classification task. Contrary to that, supervised learning seems to take a random walk in the feature space, suggesting misunderstanding of the true characteristics of natural images.
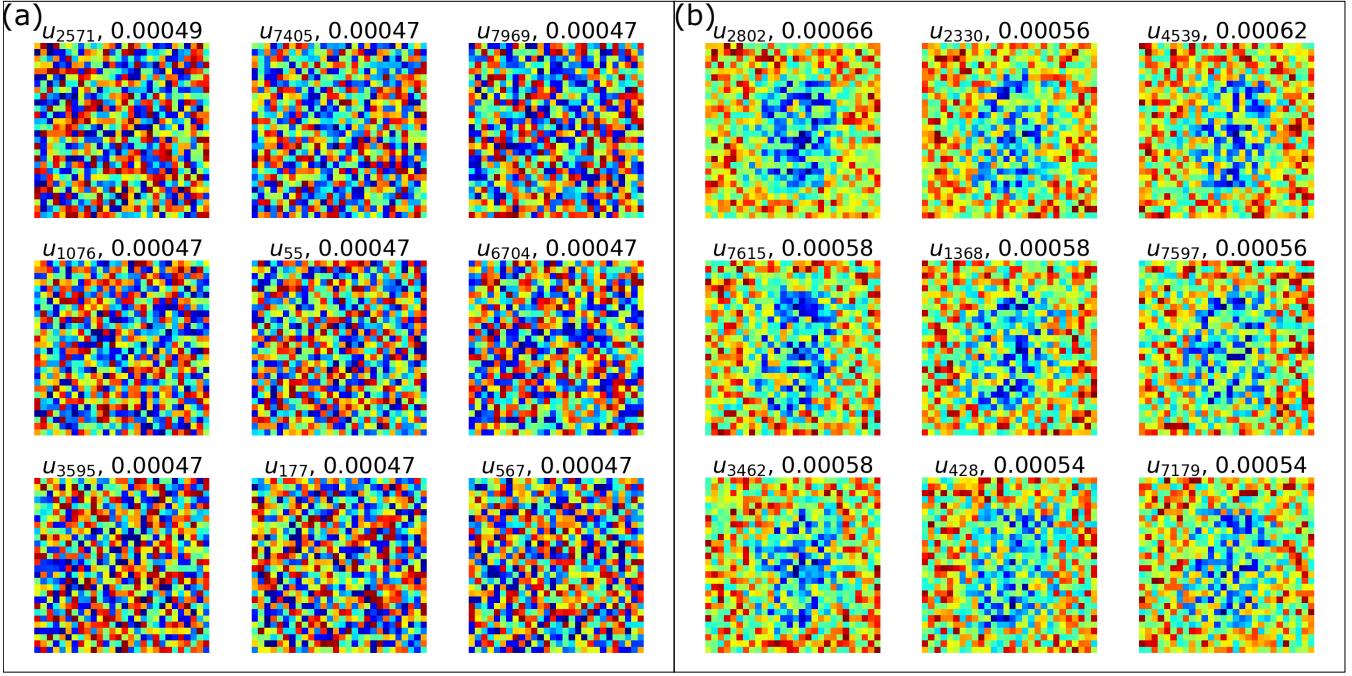
Figure 10: Comparing hidden layer filters on MNIST. For $j \in [2^{13}]$, $\|\boldsymbol{u}_j\|_2$ is in descending order from left to right and top to bottom. (a) Without Diffusion. (b) With Diffusion. Filters with diffusion have high variances compared to without diffusion. Since digits are centred in MNIST, adversarial interaction learns to discriminate background and foreground information.
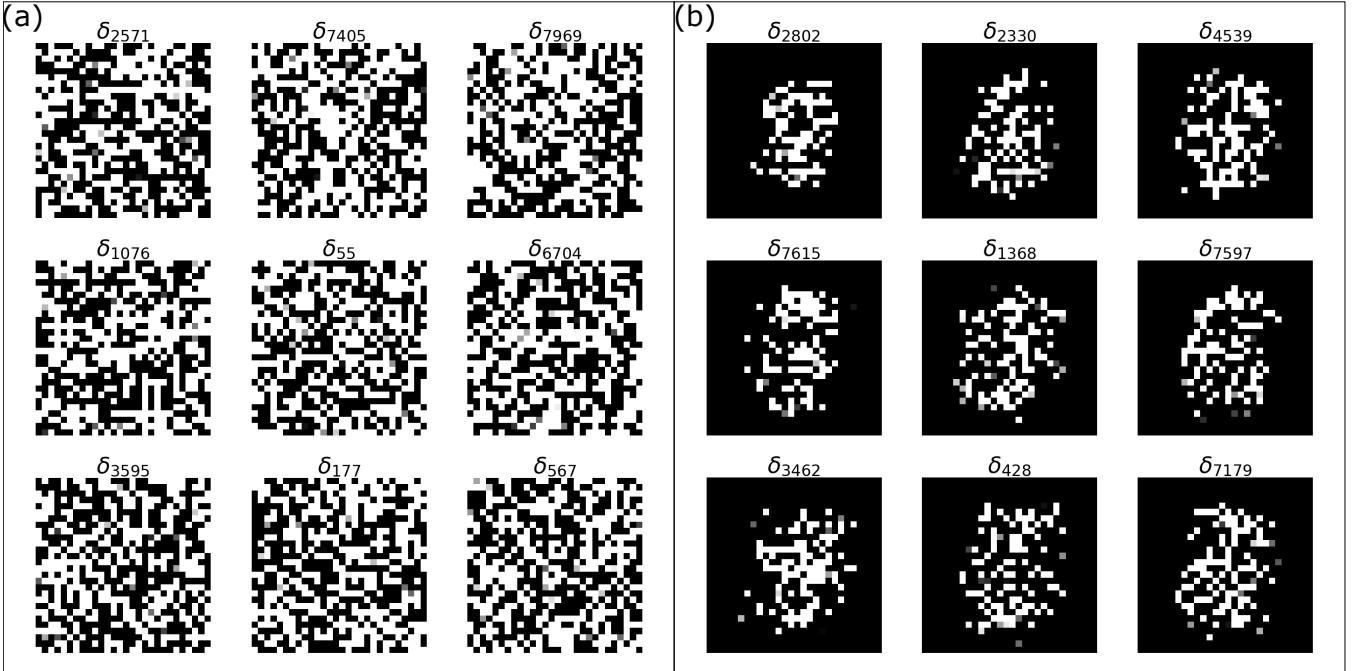
Figure 11: Visualization of features in the hidden layer on MNIST. For $j \in [2^{13}]$, $\|\boldsymbol{u}_j\|_2$ is in descending order from left to right and top to bottom. (a) Without Diffusion. (b) With Diffusion. While the system without diffusion takes a random walk, the system with diffusion moves in the direction of real manifold of natural images in order to maximize excitation.

Dissection of Diffusion

To understand the contribution of diffusion term as given by equation (38), we experiment with less number of hidden units on MNIST and CIFAR10 datasets. Recall that **Theorem 2** suggests a larger linear slope with smaller number
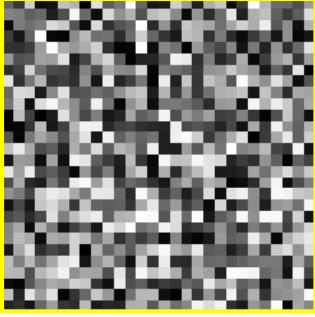
Figure 12: Input image used in the visualization of features.
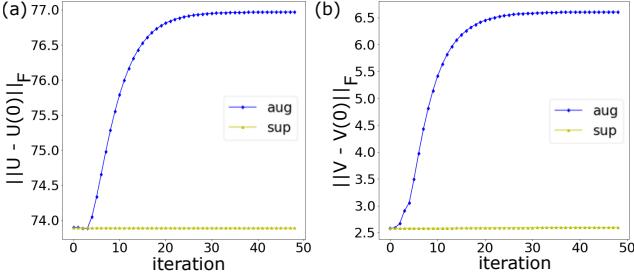
Figure 13: Distance from initialization in the (a) hidden layer (b) top layer on FashionMNIST.

of hidden units. Since reaction term settles down exponentially fast, a larger linear slope is expected to dominate the distance from initialization beyond the settling point. This hypothesis is supported by the experimental results in Figure 16 and 17 where the dominance of linear slope is discernible. To this end, we have verified the assumptions and theorems empirically on two synthetic datasets and three benchmark datasets. In the succeeding discussion, we put emphasis on pattern formation using Turing's RD model, the developed PRD model, and Gray-Scott model.

Turing Patterns by RD Model

In this section, we provide numerical simulations to better understand the specific dynamics in RD model. Turing's fundamental equations capture the dynamics of patterns as a reminiscent of those seen in nature (Turing 1952):

$$
\begin{aligned}
\frac{\partial u}{\partial t} &= a(u - h) + b(v - k) + \mu'\nabla^2 u \\
\frac{\partial v}{\partial t} &= c(u - h) + d(v - k) + \nu'\nabla^2 v.
\end{aligned}
\tag{63}
$$

Here, u and v denote two morphogens which attain equilibrium at h and k respectively. The reaction term is controlled by a, b, c and d. On the other hand, the diffusion term is controlled by μ' and ν'.

Implementation Details Figure 18 illustrates Turing patterns with parameters, $a = 1, b = -1, c = 3, d = -1.5, h = 1, k = 1, \mu' = 0.0001, \nu' = 0.0006$. The grid size is

$[100 \times 100]$ and temporal resolution, $0.02s$.[6]

Analysis on Simulated Dataset Observe that by discarding diffusion terms, i.e., $\mu' = 0$ and $\nu' = 0$, the system attains homogenous equilibrium with $u_{eq} = h$ and $v_{eq} = k$. For this reason, the morphogens, u and v are initialized with $u_{eq} + \mathcal{U}(-0.03, 0.03)$ and $v_{eq} + \mathcal{U}(-0.03, 0.03)$. The uniformly distributed perturbation, $\mathcal{U}(.)$ helps break homogeneity and spatial symmetry when diffusible factors are introduced in the system. Figure 18 supports this hypothesis as we observe emergence of evolutionary patterns as a result of diffusion. Notice that the diffusible factors are orders of magnitude smaller than reaction coefficients. This mild diffusion is enough to induce Turing instability that results in evolutionary patterns.

Turing-Like Patterns by PRD Model

Implementation Details Here, Turing-like patterns contain m neurons each in the hidden layer (u) and top layer (v). For visualization purpose, the neurons in $\mathbb{R}^{d_{in}}$ are mapped to $\mathbb{R}^{10}$ using Principal Component Analysis (PCA). Further, t-Distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton 2008) is employed to map each neuron to $\mathbb{R}^2$. t-SNE is particularly well-suited for visualizing high dimensional data. However, sometime it requires a first stage dimensionality reduction due to its high computational cost. In the top layer, t-SNE is directly employed to map from $\mathbb{R}^{10}$ to $\mathbb{R}^2$. To reduce computational bottleneck, we have selected initial 2048 neurons out of 2^{13} in the hidden layer for visualization. The absolute scales of these plots are irrelevant to the present body of analysis. In each pattern, different colors are used to represent different clusters of neurons for the sole purpose of visual assimilation. For clustering purpose, Affinity Propagation (AP) is used to capture the essence of node connectivity during information propagation among hidden units (Frey and Dueck 2006, 2007).

Analysis on Synthetic Datasets Similar to Turing's RD model, we observe spatially symmetric and homogeneous equilibrium in the absence of adversarial regularization. The system retains symmetry over iterations as long as it does not receive any signal that may induce diffusibility. As shown in Figure 19(a) and Figure 20(a), the neurons in the hidden layer (u) and top layer (v) lie close to their initial values. While gradient descent finds an ϵ-stationary point, the neurons in both layers do not deviate much from their initial topology. Consequently, the spatial symmetry and homogeneity is preserved throughout training.

On the other hand, with the introduction of an adversary into the system, the exchange of signal between generator and discriminator constitutes the basis of diffusible factors. Evident from Figure 19(b) and Figure 20(b), the adversarial interaction helps break the inherent spatial symmetry and homogeneity while randomly initialized gradient de-

---

[6]One can select another set of parameters to create a different pattern. The present discussion is least affected by the choice of these parameters. However, it is essential to draw insights from the fundamental equations.
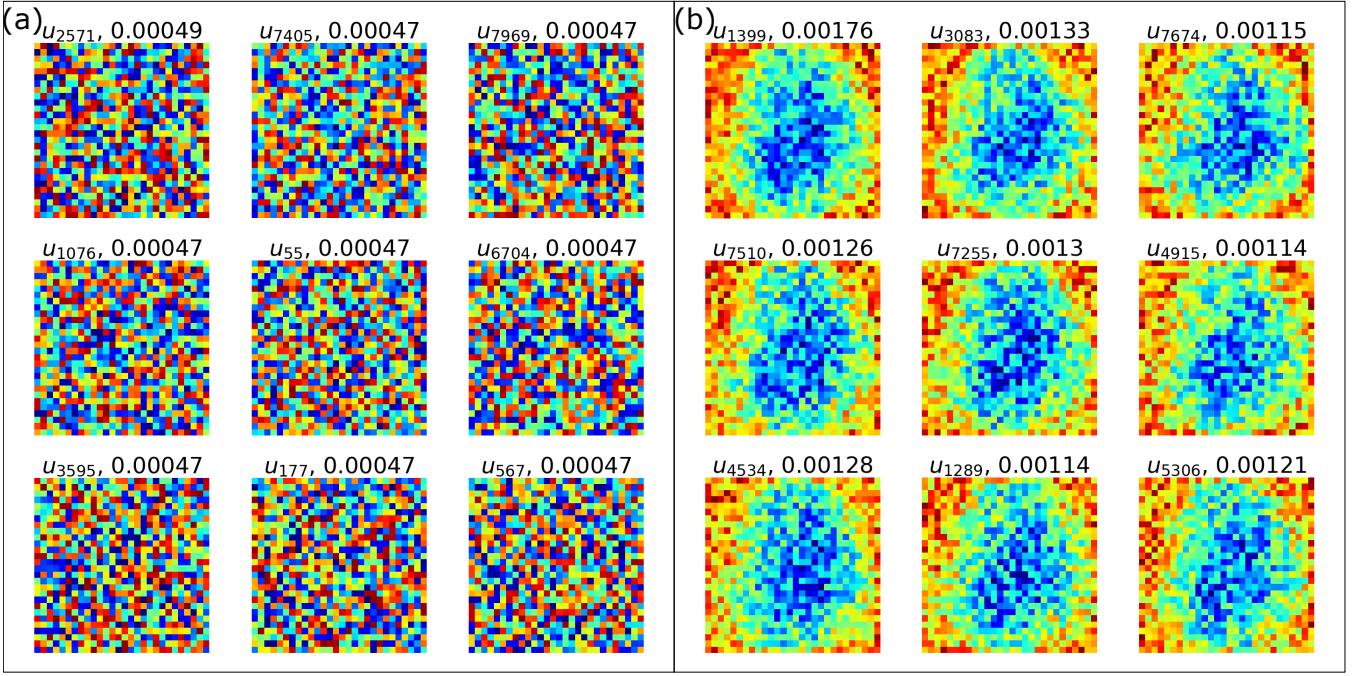
Figure 14: Comparing hidden layer filters on FashionMNIST. For $j \in [2^{13}]$, $\|\boldsymbol{u}_j\|_2$ is in descending order from left to right and top to bottom. (a) Without Diffusion. (b) With Diffusion. High variance captures non-homogeneity in weight space. Adversarial interaction allows neurons in the hidden layer to focus on central region where useful information is accessible.
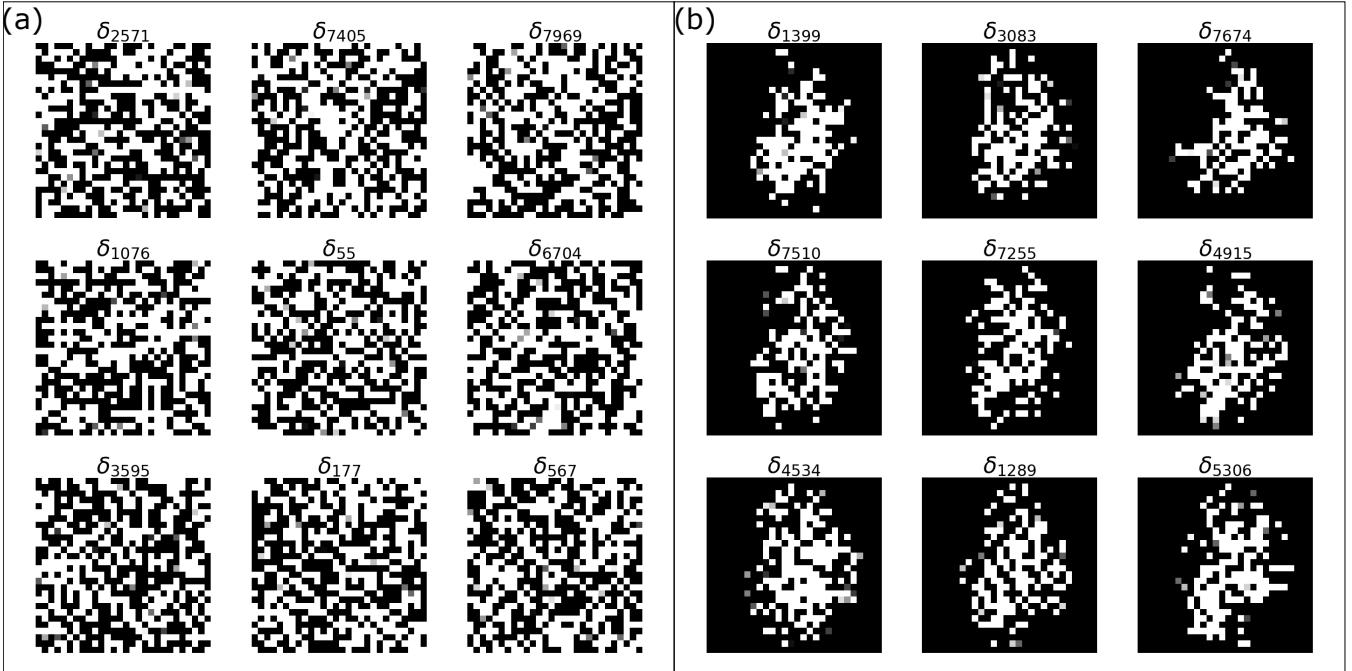
Figure 15: Visualization of features in the hidden layer on FashionMNIST. For $j \in [2^{13}]$, $\|\boldsymbol{u}_j\|_2$ is in descending order from left to right and top to bottom. (a) Without Diffusion. (b) With Diffusion. The system with diffusion gets excited the most when input images are pushed towards actual manifold of real images.

scent still finds ϵ-stationary point. In supervised and adversarial setting, the qualitative observation of Turing-like pat-

terns favors the quantitative analyses made in **Experimental Results**. Furthermore, **Theorem 3** provides a justification to
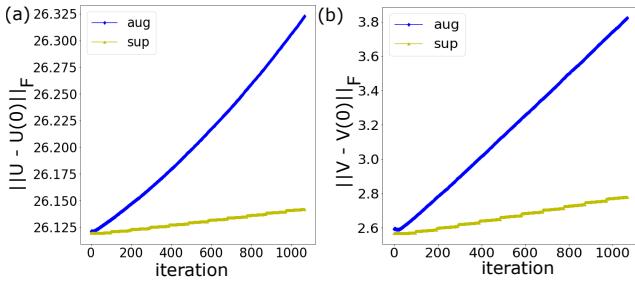
Figure 16: Distance from initialization in the (a) hidden layer (b) top layer on CIFAR10 with 2^{10} hidden units.
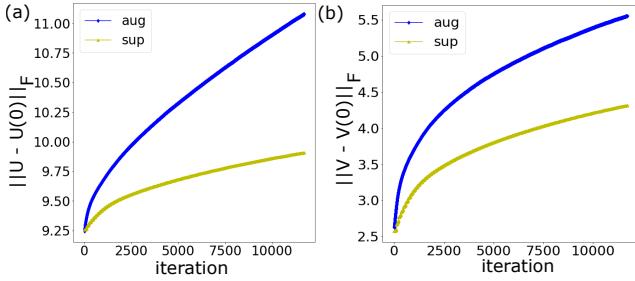
Figure 17: Distance from initialization in the (a) hidden layer (b) top layer on MNIST with 2^7 hidden units.
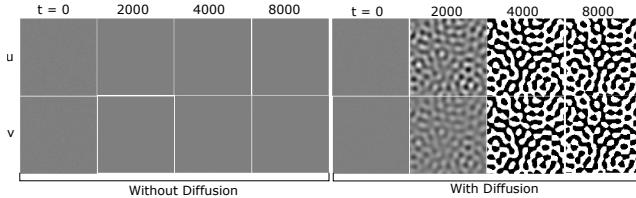
Figure 18: Turing pattern formation. The diffusible factors help break the symmetry and homogeneity.
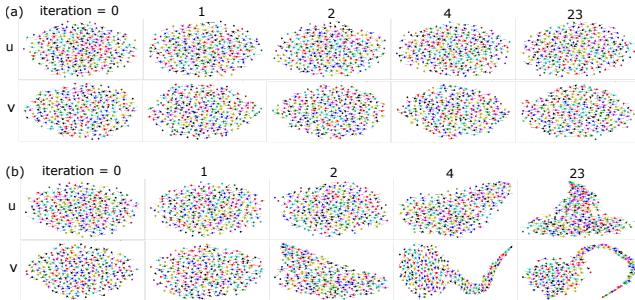
Figure 19: Pattern formation on synthetic data, $d_{in} = 784$. (a) Without Diffusion. (b) With Diffusion.

late breakdown of symmetry in the hidden layer compared to top layer. Since $\mathfrak{D}_j^{\boldsymbol{u}} \left(\nabla^2 \boldsymbol{u}_j \right)$ has a tighter asymptotic bound compared to $\mathfrak{D}_j^{\boldsymbol{v}} \left(\nabla^2 \boldsymbol{v}_j \right)$, it has a lower rate of diffusion on the experimented datasets.

Analysis on MNIST and FashionMNIST   Interestingly, the simulated patterns resemble those observed in the hidden
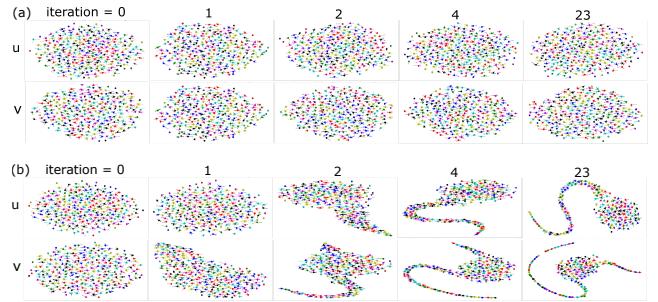
Figure 20: Pattern formation on synthetic data, $d_{in} = 1024$. (a) Without Diffusion. (b) With Diffusion.

layer and top layer when trained on actual datasets: MNIST (Figure 21) and FashionMNIST (Figure 22). Here, similarity is measured in a topological sense that reflects breakdown of symmetry and homogeneity under adversarial interaction. Figure 23 illustrates Turing-like patterns formed by all neurons after fully trained on MNIST.
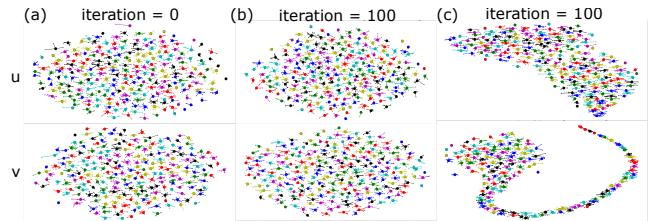
Figure 21: Pattern formation on MNIST. (a) Initialization. (b) Without Diffusion. (c) With Diffusion.
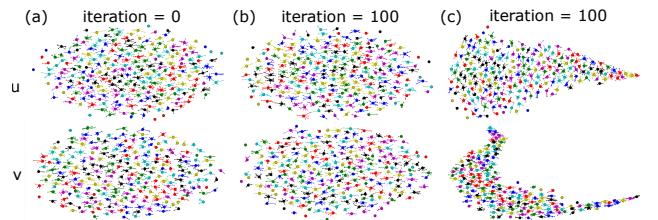
Figure 22: Pattern formation on FashionMNIST. (a) Initialization. (b) Without Diffusion. (c) With Diffusion.

Turing-Like Patterns by Gray-Scott Model

For better understanding of RD systems, we have also simulated Gray-Scott patterns (Gray and Scott 1984)[7]:

$$\begin{aligned}
\frac{\partial u}{\partial t} &= F \left(1 - u \right) - uv^2 + \mu' \nabla^2 u \\
\frac{\partial v}{\partial t} &= - \left(F + k \right) v + uv^2 + \nu' \nabla^2 v
\end{aligned} \tag{64}$$

---

[7](Sayama 2015) provides a comprehensive overview of Gray-Scott pattern formation subject to different parameters.
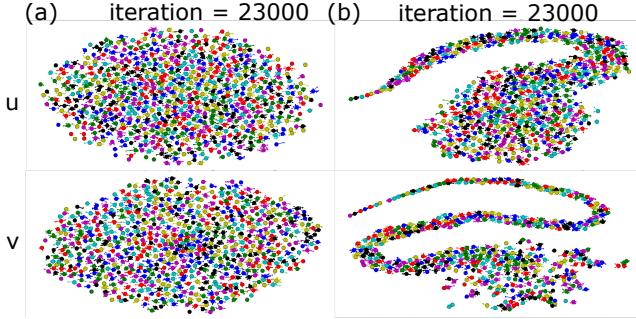
Figure 23: Pattern formation on MNIST. Visualization of all 2^{13} neurons. (a) Without Diffusion. (b) With Diffusion.

Implementation Details The morphogens lie on a two dimensional grid of size $[100 \times 100]$. The initial conditions are such that there exist certain concentration difference in each morphogen which gives rise to diffusion. The grid therefore is initialized by all ones (u) and all zeros (v) with corresponding central $[5 \times 5]$ patch inverted as shown in Figure 24, 25, 26 and 27 at $t = 0$. With very high rate of diffusion, the morphogens get destroyed too quickly to have a positive influence on the reactions, which are the most important parts of the assumptions in RD model. Therefore, the diffusion coefficients are set to $\mu' = 2e-5$ and $\nu' = 1e-5$ in the present analysis of Gray-Scott patterns.

Analysis on Simulated Dataset By changing the values of F and k, a wide variety of patterns can be created. In all these figures, different irregularities induce different patterns despite exactly identical intitial condition. This is a crucial observation as it justifies non-stationary Turing-like patterns in the intermediate iterations of adversarial training as shown in Figure 19(b) and Figure 20(b). In this case, different irregularities get introduced at each iteration due to stochasticity in the finite sum optimization problem.
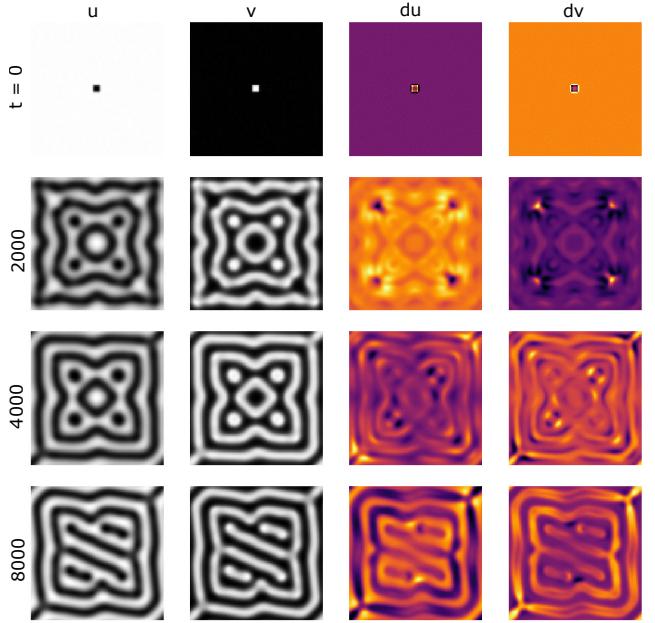
Figure 24: Gray-Scott pattern formation. $F = 0.025$ and $k = 0.055$.
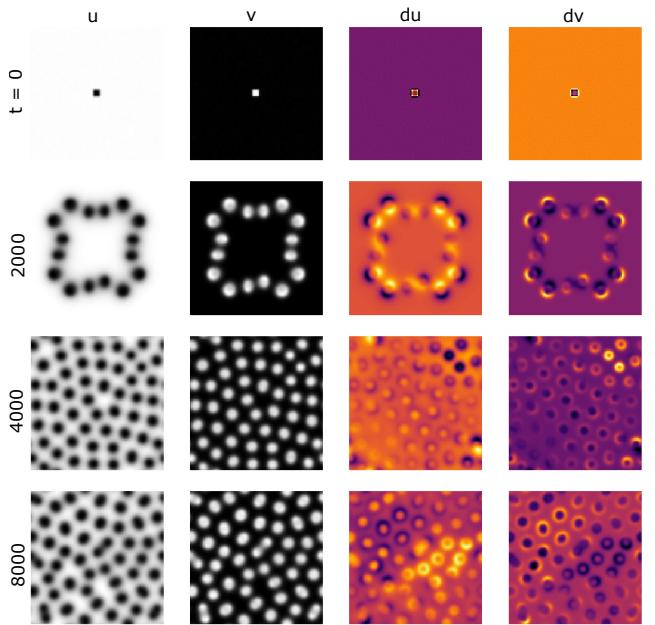
Figure 25: Gray-Scott pattern formation. $F = 0.025$ and $k = 0.060$.
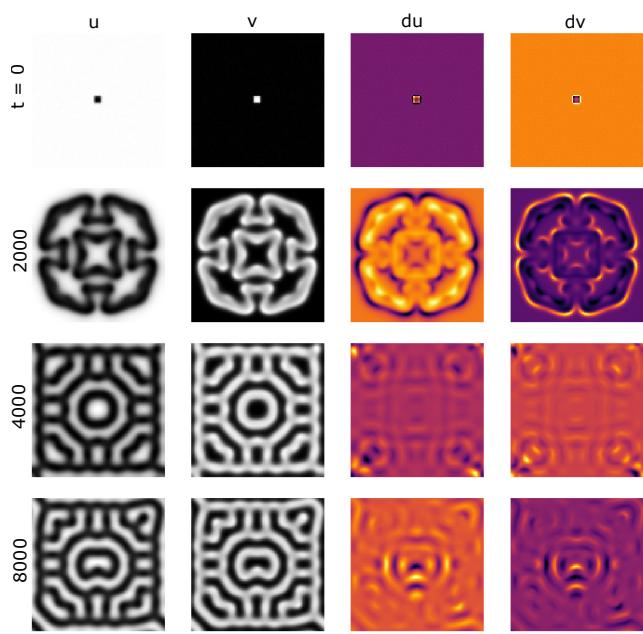
Figure 26: Gray-Scott pattern formation. $F = 0.040$ and $k = 0.060$.
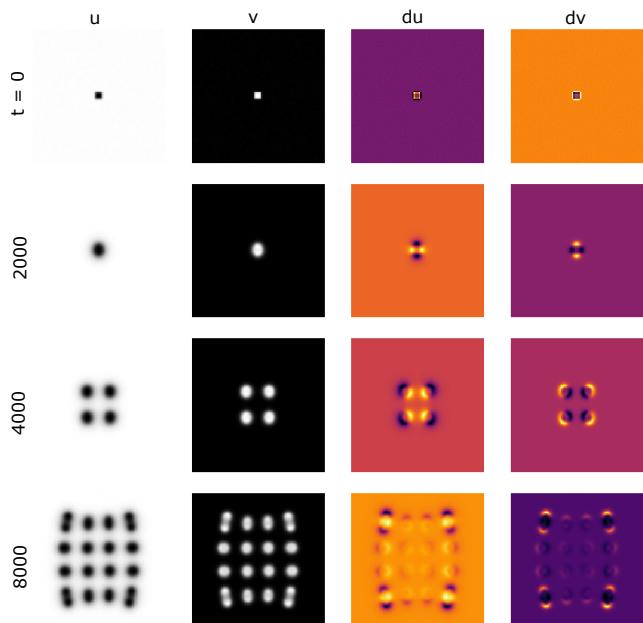
Figure 27: Gray-Scott pattern formation. $F = 0.035$ and $k = 0.065$.