# Towards A Pseudo-Reaction-Diffusion Model for Turing Instability in Adversarial Learning

**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 justification, we present a *pseudo-reaction-diffusion* model to explain the mechanism that may underlie this phenomenon. While supervised learning attains homogeneous equilibrium, the introduction of an adversary helps break this homogeneity to create non-homogeneous patterns at equilibrium. In addition, different from sole supervision, we show that the solutions obtained under adversarial interaction are not limited to a tiny subspace around initialization.

## 1 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 finds homogeneous equilibrium while maintaining spatial symmetry over iterations.

The reason for studying this phenomenon is multifold. 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 physical systems, such as butterfly wings, zebra, giraffe and leopard [1, 2, 3, 4, 5]. 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 [6]. This paper is intended to shed light on some of these aspects.

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 [1]. Thus we state the 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 the two layer generator network.*

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

Third Workshop on Machine Learning and the Physical Sciences (NeurIPS 2020), Vancouver, Canada.

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 [3].

## 2 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 $\lambda_{\max}(.)$ denotes smallest and largest eigen value of a matrix. $dx$ represents derivative of $x$ and $\partial x$ represents its partial derivative. For $g : \mathbb{R}^d \to \mathbb{R}$, $\nabla g$ 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 denote two layer neural networks with rectified linear unit (ReLU) activation function $(\sigma(.))$ by $f(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{x}) = \frac{1}{\sqrt{d_{out}m}} \boldsymbol{V}\sigma(\boldsymbol{U}\boldsymbol{x})$. 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 [7]. In supervised learning, the training is carried out by minimizing the $l_2$ loss over data as given by

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

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:

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

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 [8]. 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 [1].

**Learning Algorithm.** We consider the 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):

$$\frac{du_{jk}}{dt} = -\frac{\partial \mathcal{L}_{aug}(\boldsymbol{U}(t), \boldsymbol{V}(t), \boldsymbol{W}(t), \boldsymbol{a}(t))}{\partial u_{jk}(t)}, \frac{dv_{ij}}{dt} = -\frac{\partial \mathcal{L}_{aug}(\boldsymbol{U}(t), \boldsymbol{V}(t), \boldsymbol{W}(t), \boldsymbol{a}(t))}{\partial v_{ij}(t)} \quad (3)$$

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 $\epsilon$-approximate equilibrium, i.e. $\frac{du_{jk}}{dt} < \epsilon$ and $\frac{dv_{ij}}{dt} < \epsilon$.

### 2.1 Revisiting Reaction-Diffusion Model[1]

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:

$$\frac{d\boldsymbol{u}_j}{dt} = \mathfrak{R}_j^{\boldsymbol{u}}(\boldsymbol{u}_j, \boldsymbol{v}_j) + \mathfrak{D}_j^{\boldsymbol{u}}\left(\nabla^2 \boldsymbol{u}_j\right), \frac{d\boldsymbol{v}_j}{dt} = \mathfrak{R}_j^{\boldsymbol{v}}(\boldsymbol{u}_j, \boldsymbol{v}_j) + \mathfrak{D}_j^{\boldsymbol{v}}\left(\nabla^2 \boldsymbol{v}_j\right), \quad (4)$$

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.

### 2.2 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 [9, 10, 3], a system with adversarial interaction is asserted to be a pseudo-reaction-diffusion model.

## 3 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. The network architecture then becomes:

$$f(\boldsymbol{U}, \boldsymbol{v}, \boldsymbol{x}) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} v_j \sigma\left(u_j^T x\right) = \frac{1}{\sqrt{m}} \boldsymbol{v}^T \sigma(\boldsymbol{U}\boldsymbol{x}). \tag{5}$$

Our goal is to minimize

$$\mathcal{L}_{sup}(\boldsymbol{U}, \boldsymbol{v}) = \sum_{p=1}^{n} \frac{1}{2} \left(f(\boldsymbol{U}, \boldsymbol{v}, \boldsymbol{x}_p) - y_p\right)^2 \tag{6}$$

in supervised setting and

$$\mathcal{L}_{aug}(\boldsymbol{U}, \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{a}) = \sum_{p=1}^{n} \frac{1}{2} \left(f(\boldsymbol{U}, \boldsymbol{v}, \boldsymbol{x}_p) - y_p\right)^2 - \frac{1}{\sqrt{m}} \sum_{p=1}^{n} \boldsymbol{a}^T \, \sigma\left(\boldsymbol{w}\left(f(\boldsymbol{U}, \boldsymbol{v}, \boldsymbol{x}_p)\right)\right) \tag{7}$$

in adversarial setting. The architecture of adversary is simplified to $g(w, a, y) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_j \sigma(w_j y)$. We follow the definition of Gram matrix from [11]

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

The Gram matrix has several important properties [12, 13]. One interesting property that justifies **Assumption 1** is given by Du et al. [11]: *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.

### 3.1 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 $u_j \sim \mathcal{N}(0, I)$ 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 $1 - \delta$:*

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

*Proof.* Refer to Appendix B.1.

### 3.2 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 [14], recent success in adversarial training owes in part to gradient penealty [8]. We remark that min-max optimization under non-convexity and non-concavity is considered NP-hard to find a stationary point [15]. 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 [15], we study two layer networks with ReLU activations and rely on gradient penalty to limit its capacity. In the simplified theoretical analysis, we assume $\|\boldsymbol{w}\|_2 \leq L$ for a small positive 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 $L \leq \mathcal{O}\left(\frac{\epsilon\sqrt{m}}{\kappa n\sqrt{2\log(2/\delta)}}\right)$, $\kappa = \mathcal{O}(\kappa^\infty)$ 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[1]:*

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

*Proof.* Refer to Appendix B.2.

## 4 Discussion of Insights from Analysis

A profound implication of this finding is that adversarial learning allows gradient descent-ascent to explore a large subspace in contrast to supervised learning where a tiny subspace around initialization is merely explored [16]. 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 [1, 2, 17, 3].

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 per the statement of **Theorem 2**. 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 [4]. More importantly, it fits the description of a short and a long range feedback which indicates a similar mechanism must be involved in adversarial learning. This analogy essentially provides positive support to the developed PRD theory.

[1]Refer to Appendix for further discussion on breakdown of symmetry and homogeneity.

## 5 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


[1] A. Turing, "The chemical basis of morphogenesis," *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, vol. 237, no. 641, pp. 37–72, 1952.

[2] H. Meinhardt, "Models of biological pattern formation," *New York*, p. 118, 1982.

[3] E. M. Rauch and M. M. Millonas, "The role of trans-membrane signal transduction in turing-type cellular pattern formation," *Journal of theoretical biology*, vol. 226, no. 4, pp. 401–407, 2004.

[4] A. Nakamasu, G. Takahashi, A. Kanbe, and S. Kondo, "Interactions between zebrafish pigment cells responsible for the generation of turing patterns," *Proceedings of the National Academy of Sciences*, vol. 106, no. 21, pp. 8429–8434, 2009.

[5] S. Kondo and T. Miura, "Reaction-diffusion model as a framework for understanding biological pattern formation," *science*, vol. 329, no. 5999, pp. 1616–1620, 2010.

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

[7] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, 2010.

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

[9] H. Meinhardt and A. Gierer, "Applications of a theory of biological pattern formation based on lateral inhibition," *Journal of cell science*, vol. 15, no. 2, pp. 321–346, 1974.

[10] H. Meinhardt and A. Gierer, "Pattern formation by local self-activation and lateral inhibition," *Bioessays*, vol. 22, no. 8, pp. 753–760, 2000.

[11] S. S. Du, X. Zhai, B. Poczos, and A. Singh, "Gradient descent provably optimizes over-parameterized neural networks," in *International Conference on Learning Representations*, 2018.

[12] R. Tsuchida, F. Roosta, and M. Gallagher, "Invariance of weight distributions in rectified mlps," in *International Conference on Machine Learning*, pp. 4995–5004, 2018.

[13] B. Xie, Y. Liang, and L. Song, "Diverse neural network learns true target functions," in *Artificial Intelligence and Statistics*, pp. 1216–1224, 2017.

[14] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, 2017.

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

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

[17] P. Gray and S. Scott, "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*, vol. 39, no. 6, pp. 1087–1097, 1984.

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

[19] J. Bernoulli, "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*, 1695.

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

[21] T. Gregor, D. W. Tank, E. F. Wieschaus, and W. Bialek, "Probing the limits to positional information," *Cell*, vol. 130, no. 1, pp. 153–164, 2007.

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

[23] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," in *NIPS Workshop on Adversarial Training*, 2016.

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

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

[26] D. Engin, A. Genç, and H. Kemal Ekenel, "Cycle-dehaze: Enhanced cyclegan for single image dehazing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 825–833, 2018.

[27] L. Rout, I. Misra, S. Manthira Moorthi, and D. Dhar, "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*, pp. 188–189, 2020.

[28] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[29] L. Rout, "Alert: Adversarial learning with expert regularization using tikhonov operator for missing band reconstruction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 6, pp. 4395–4405, 2020.

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

[31] X. Wang and A. Gupta, "Generative image modeling using style and structure adversarial networks," in *European conference on computer vision*, pp. 318–335, Springer, 2016.

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

[33] M. Sarmad, H. J. Lee, and Y. M. Kim, "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*, pp. 5898–5907, 2019.

[34] P.-F. Verhulst, "Notice sur la loi que la population suit dans son accroissement," *Corresp. Math. Phys.*, vol. 10, pp. 113–126, 1838.

[35] E. Rogers, "Diffusion of innovations . delran," *NJ: Simon & Schuster. Schneider, L.(1971). Dialectic in sociology. American Sociological Review*, vol. 36, p. 667678, 2003.

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

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

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

## Appendix A 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 $\frac{1}{\sqrt{d_{out}m}}\boldsymbol{V}\sigma\left(\boldsymbol{U}\boldsymbol{x}_p\right)$.

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

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

$$\begin{aligned}
\mathfrak{R}_j^{\boldsymbol{u}}\left(\boldsymbol{u}_j,\boldsymbol{v}_j\right)&=\mathcal{O}\left(nd_{in}\sqrt{\frac{d_{out}}{m}}\right),\\
\mathfrak{D}_j^{\boldsymbol{u}}\left(\nabla^2\boldsymbol{u}_j\right)&=\mathcal{O}\left(nm^2d_{in}d_{out}^{3/2}\right),\\
\mathfrak{R}_j^{\boldsymbol{v}}\left(\boldsymbol{u}_j,\boldsymbol{v}_j\right)&=\mathcal{O}\left(nd_{in}\sqrt{\frac{d_{out}}{m}}\right),\\
\mathfrak{D}_j^{\boldsymbol{v}}\left(\nabla^2\boldsymbol{v}_j\right)&=\mathcal{O}\left(nm^2d_{in}d_{out}^{1/2}\right).
\end{aligned}$$

*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.

## Appendix B Technical Proofs

### B.1 Proof of Theorem 1

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 $(1-\delta)$, $u_{jk}$ induces a symmetric and homogeneously distributed matrix $U$ at initialization within a ball of radius $\zeta\triangleq\frac{2\sqrt{md_{in}}}{\sqrt{2\pi}\delta}$.*

Using law of large numbers, it is trivial to prove symmetry and homogeneity since Gaussian distribution has a symmetric density function. Now, we derive the upper bound.

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

$$\left|u_{jk}\left(0\right)\right|\leq\frac{\mathbb{E}\left[\left|u_{jk}\left(0\right)\right|\right]}{\delta}=\frac{2}{\sqrt{2\pi}\delta}. \tag{8}$$

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

$$\begin{aligned}
\left\|U(0)\right\|_F&=\left(\sum_{j=1}^{m}\sum_{k=1}^{d_{in}}\left|u_{jk}(0)\right|^2\right)^{1/2}\\
&\leq\left(\sum_{j=1}^{m}\sum_{k=1}^{d_{in}}\frac{4}{2\pi\delta^2}\right)^{1/2}\\
&\leq\frac{2\sqrt{md_{in}}}{\sqrt{2\pi}\delta}\triangleq\zeta.
\end{aligned} \tag{9}$$

This finishes the proof of **Lemma 1**. □

Next, we prove how supervised cost helps maintain symmetry and homogeneity. Since $\boldsymbol{U}$ is initially symmetric and homogeneously distributed within $\zeta$, 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 [11] 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* $\lambda_{min}(\mathcal{H}) \geq \frac{\lambda_0}{2}$.

**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:*

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

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

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

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

$$\begin{aligned}\left\|\frac{d\boldsymbol{u}_j(s)}{ds}\right\|_2 &= \left\|\frac{\partial \mathcal{L}_{sup}(\boldsymbol{U}, \boldsymbol{v})}{\partial \boldsymbol{u}_j(s)}\right\|_2 = \left\|\sum_{p=1}^{n}(z_p(s) - y_p)\frac{\partial z_p(s)}{\partial \boldsymbol{u}_j(s)}\right\|_2 \\ &= \left\|\sum_{p=1}^{n}(z_p(s) - y_p)\frac{1}{\sqrt{m}} v_j 1_{\{\boldsymbol{u}_j(s)^T \boldsymbol{x}_p \geq 0\}} \boldsymbol{x}_p\right\|_2.\end{aligned} \tag{10}$$

By triangle inequality,

$$\left\|\frac{d\boldsymbol{u}_j(s)}{ds}\right\|_2 \leq \sum_{p=1}^{n}\left\|(z_p(s) - y_p)\frac{1}{\sqrt{m}} v_j 1_{\{u_j(s)^T x_p \geq 0\}} x_p\right\|_2. \tag{11}$$

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

$$\begin{aligned}\left\|\frac{d\boldsymbol{u}_j(s)}{ds}\right\|_2 &\leq \sum_{p=1}^{n}\frac{1}{\sqrt{m}}\left|(z_p(s) - y_p)\right| |v_j| \|x_p\|_2 \\ &= \frac{1}{\sqrt{m}}\sum_{p=1}^{n}\left|(z_p(s) - y_p)\right| \\ &\leq \frac{\sqrt{n}}{\sqrt{m}}\|\boldsymbol{z}(s) - \boldsymbol{y}\|_2 \\ &\leq \frac{\sqrt{n}}{\sqrt{m}}\exp\left(-\frac{\lambda_0}{2}s\right)\|\boldsymbol{z}(0) - \boldsymbol{y}\|_2.\end{aligned} \tag{12}$$

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

$$\begin{aligned}\|\boldsymbol{u}_j(t) - \boldsymbol{u}_j(0)\|_2 &= \left\|\int_0^t \frac{d\boldsymbol{u}_j(s)}{ds} ds\right\|_2 \leq \int_0^t \left\|\frac{d\boldsymbol{u}_j(s)}{ds}\right\|_2 ds \\ &\leq \frac{\sqrt{n}}{\sqrt{m}}\int_0^t \exp\left(-\frac{\lambda_0}{2}s\right)\|\boldsymbol{z}(0) - \boldsymbol{y}\|_2 ds \\ &\leq \frac{2\sqrt{n}\,\|\boldsymbol{z}(0) - \boldsymbol{y}\|_2}{\sqrt{m}\lambda_0}\left(1 - \exp\left(-\frac{\lambda_0}{2}t\right)\right).\end{aligned} \tag{13}$$

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

$$\|\boldsymbol{u}_j(t) - \boldsymbol{u}_j(0)\|_2 \leq \frac{2\sqrt{n}\,\|\boldsymbol{z}(0) - \boldsymbol{y}\|_2}{\sqrt{m}\lambda_0}. \tag{14}$$

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

$$\|\boldsymbol{u}_j(t) - \boldsymbol{u}_j(0)\|_2 \leq \frac{2\sqrt{n}\mathbb{E}\left[\|z(0) - y\|_2\right]}{\sqrt{m}\lambda_0\delta} \leq \mathcal{O}\left(\frac{n^{3/2}}{m^{1/2}\lambda_0\delta}\right). \tag{15}$$

Now, we can bound the distance from initialization.

$$\begin{aligned}
\|\boldsymbol{U}(t) - \boldsymbol{U}(0)\|_F &= \left(\sum_{j=1}^{m}\sum_{k=1}^{d_{in}} |u_{jk}(t) - u_{jk}(0)|^2\right)^{1/2} \\
&\leq \left(\sum_{j=1}^{m} \|\boldsymbol{u}_j(t) - \boldsymbol{u}_j(0)\|_2^2\right)^{1/2} \\
&\leq \left(\sum_{j=1}^{m} \frac{4n\left(\mathbb{E}\left[\|\boldsymbol{z}(0) - \boldsymbol{y}\|_2\right]\right)^2}{m\lambda_0^2\delta^2}\right)^{1/2} \\
&\leq \frac{2\sqrt{n}\mathbb{E}\left[\|z(0) - y\|_2\right]}{\lambda_0\delta} \leq \mathcal{O}\left(\frac{n^{3/2}}{\lambda_0\delta}\right),
\end{aligned} \tag{16}$$

which finishes the proof. □

### B.2 Proof of Theorem 2

We sketch the proof of the main result in two parts: **Reaction Term** and **Diffusion Term**.

#### B.2.1 Reaction Term

For $0 \leq s \leq t$ in augmented objective as given by equation (7), we get

$$\begin{aligned}
\left\|\frac{d\boldsymbol{u}_j(s)}{ds}\right\|_2 &= \left\|\frac{\partial\mathcal{L}_{aug}\left(\boldsymbol{U}, \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{a}\right)}{\partial\boldsymbol{u}_j(s)}\right\|_2 \\
&= \left\|\frac{\partial\mathcal{L}_{sup}\left(\boldsymbol{U}, \boldsymbol{v}\right)}{\partial\boldsymbol{u}_j(s)} - \frac{\partial}{\partial\boldsymbol{u}_j(s)}\sum_{p=1}^{n} g\left(\boldsymbol{w}, a, z_p\right)\right\|_2 \\
&\leq \underbrace{\left\|\frac{\partial\mathcal{L}_{sup}\left(\boldsymbol{U}, \boldsymbol{v}\right)}{\partial\boldsymbol{u}_j(s)}\right\|_2 + \left\|\frac{\partial}{\partial\boldsymbol{u}_j(s)}\sum_{p=1}^{n} g\left(\boldsymbol{w}, a, z_p\right)\right\|_2}_{\text{Triangle inequality}}.
\end{aligned} \tag{17}$$

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:*

$$\frac{d\boldsymbol{z}(t)}{dt} = \mathcal{H}(t)\left(\boldsymbol{y} - \boldsymbol{z}(t)\right) + \mathcal{H}(t)\nabla_{\boldsymbol{z}(t)}g(\boldsymbol{w}(t), \boldsymbol{a}(t), \boldsymbol{z}(t)). \tag{18}$$

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

$$\begin{aligned}
\frac{dz_p(t)}{dt} &= \sum_{j=1}^{m} \left\langle \frac{\partial f\left(\boldsymbol{U}, \boldsymbol{v}, \boldsymbol{x}_p\right)}{\partial \boldsymbol{u}_j(t)}, \frac{d\boldsymbol{u}_j(t)}{dt} \right\rangle \\
&= \sum_{j=1}^{m} \left\langle \frac{\partial f\left(\boldsymbol{U}, \boldsymbol{v}, \boldsymbol{x}_p\right)}{\partial \boldsymbol{u}_j(t)}, \frac{1}{\sqrt{m}} \sum_{q=1}^{n} \left(y_q - z_q\right) v_j \boldsymbol{x}_q 1_{\left\{\boldsymbol{u}_j^T \boldsymbol{x}_q \geq 0\right\}} + \frac{1}{m} \sum_{q=1}^{n} \sum_{r=1}^{m} a_r w_r v_j \boldsymbol{x}_q 1_{\left\{w_r z_q \geq 0, \boldsymbol{u}_j^T \boldsymbol{x}_q \geq 0\right\}} \right\rangle \\
&= \underbrace{\sum_{j=1}^{m} \left\langle \frac{\partial f\left(\boldsymbol{U}, \boldsymbol{v}, \boldsymbol{x}_p\right)}{\partial \boldsymbol{u}_j(t)}, \frac{1}{\sqrt{m}} \sum_{q=1}^{n} \left(y_q - z_q\right) v_j \boldsymbol{x}_q 1_{\left\{\boldsymbol{u}_j^T \boldsymbol{x}_q \geq 0\right\}} \right\rangle}_{\mathcal{A}} \\
&\quad + \underbrace{\sum_{j=1}^{m} \left\langle \frac{\partial f\left(\boldsymbol{U}, \boldsymbol{v}, \boldsymbol{x}_p\right)}{\partial \boldsymbol{u}_j(t)}, \frac{1}{m} \sum_{q=1}^{n} \sum_{r=1}^{m} a_r w_r v_j \boldsymbol{x}_q 1_{\left\{w_r z_q \geq 0, \boldsymbol{u}_j^T \boldsymbol{x}_q \geq 0\right\}} \right\rangle}_{\mathcal{B}}.
\end{aligned} \tag{19}$$

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

$$\begin{aligned}
\mathcal{A} &\coloneqq \sum_{j=1}^{m} \left\langle \frac{1}{\sqrt{m}} v_j \boldsymbol{x}_p 1_{\left\{\boldsymbol{u}_j^T \boldsymbol{x}_p \geq 0\right\}}, \frac{1}{\sqrt{m}} \sum_{q=1}^{n} \left(y_q - z_q\right) v_j \boldsymbol{x}_q 1_{\left\{\boldsymbol{u}_j^T \boldsymbol{x}_q \geq 0\right\}} \right\rangle \\
&= \sum_{q=1}^{n} \left(y_q - z_q\right) v_j^2 \boldsymbol{x}_p^T \boldsymbol{x}_q \frac{1}{m} \sum_{j=1}^{m} 1_{\left\{\boldsymbol{u}_j^T \boldsymbol{x}_p \geq 0, \boldsymbol{u}_j^T \boldsymbol{x}_q \geq 0\right\}} \\
&\triangleq \sum_{q=1}^{n} \left(y_q - z_q(t)\right) \mathcal{H}_{pq}(t),
\end{aligned} \tag{20}$$

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

$$\mathcal{H}_{pq}(t) = \boldsymbol{x}_p^T \boldsymbol{x}_q \frac{1}{m} \sum_{j=1}^{m} 1_{\left\{\boldsymbol{u}_j^T \boldsymbol{x}_p \geq 0, \boldsymbol{u}_j^T \boldsymbol{x}_q \geq 0\right\}}. \tag{21}$$

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

$$\begin{aligned}
\mathcal{B} &\coloneqq \sum_{j=1}^{m} \left\langle \frac{1}{\sqrt{m}} v_j \boldsymbol{x}_p 1_{\left\{\boldsymbol{u}_j^T \boldsymbol{x}_p \geq 0\right\}}, \frac{1}{m} \sum_{q=1}^{n} \sum_{r=1}^{m} a_r w_r v_j \boldsymbol{x}_q 1_{\left\{w_r z_q \geq 0, \boldsymbol{u}_j^T \boldsymbol{x}_q \geq 0\right\}} \right\rangle \\
&= \sum_{q=1}^{n} v_j^2 \underbrace{\left( \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r w_r 1_{\{w_r z_q \geq 0\}} \right)}_{\nabla_{\boldsymbol{z}} g} \boldsymbol{x}_p^T \boldsymbol{x}_q \frac{1}{m} \sum_{j=1}^{m} 1_{\left\{\boldsymbol{u}_j^T \boldsymbol{x}_p \geq 0, \boldsymbol{u}_j^T \boldsymbol{x}_q \geq 0\right\}} \\
&\triangleq \sum_{q=1}^{n} \frac{\partial g\left(\boldsymbol{w}, \boldsymbol{a}, z_q\right)}{\partial z_q} \mathcal{H}_{pq}(t)
\end{aligned} \tag{22}$$

Thus, the prediction dynamics are governed by

$$\frac{dz_p(t)}{dt} = \sum_{q=1}^{n} \left(y_q - z_q(t)\right) \mathcal{H}_{pq}(t) + \sum_{q=1}^{n} \frac{\partial g\left(\boldsymbol{w}(t), \boldsymbol{a}(t), z_q(t)\right)}{\partial z_q(t)} \mathcal{H}_{pq}(t). \tag{23}$$

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

**Lemma 3.** (Hoeffding's inequality, two sided [18]) *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*

$$\mathbb{P}\left\{ \left| \sum_{r=1}^{m} a_r w_r \right| \geq t \right\} \leq 2 \exp\left( -\frac{t^2}{2 \left\| \boldsymbol{w} \right\|_2^2} \right). \tag{24}$$

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

$$\left|\sum_{r=1}^{m} a_r w_r\right| \leq \|\boldsymbol{w}\|_2 \sqrt{2\log\left(\frac{2}{\delta}\right)}. \tag{25}$$

Now, the distance from true labels can be bounded by

$$\begin{aligned}
\frac{d}{dt}\|\boldsymbol{z}(t)-\boldsymbol{y}\|_2^2 &= \left\langle \boldsymbol{z}(t)-\boldsymbol{y}, \frac{d\boldsymbol{z}(t)}{dt}\right\rangle \\
&= 2\left\langle \boldsymbol{z}(t)-\boldsymbol{y}, \mathcal{H}(t)\left(\boldsymbol{y}-\boldsymbol{z}(t)\right) + \mathcal{H}(t)\nabla_{\boldsymbol{z}(t)} g(\boldsymbol{w}(t),\boldsymbol{a}(t),\boldsymbol{z}(t))\right\rangle \\
&= 2\left\langle \boldsymbol{z}(t)-\boldsymbol{y}, -\mathcal{H}(t)\left(\boldsymbol{z}(t)-\boldsymbol{y}\right)\right\rangle + 2\left\langle \boldsymbol{z}(t)-\boldsymbol{y}, \mathcal{H}(t)\nabla_{\boldsymbol{z}(t)} g(\boldsymbol{w}(t),\boldsymbol{a}(t),\boldsymbol{z}(t))\right\rangle
\end{aligned} \tag{26}$$

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

*Proof.* For clarity, let us recall **Lemma 3.1** of Du et al.[11]: *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 $\lambda_{\min}(\mathcal{H}(0)) \geq \frac{3}{4}\lambda_0$.* From **Remark 1**, we know

$$\|\mathcal{H}\|_2 - \|\mathcal{H}(0)\|_2 \leq \|\mathcal{H}-\mathcal{H}(0)\|_2 \leq \frac{\lambda_0}{4}. \tag{27}$$

Using similar arguments, we get

$$\|\mathcal{H}(0)\|_2 - \|\mathcal{H}^\infty\|_2 \leq \|\mathcal{H}(0)-\mathcal{H}^\infty\|_2 \leq \frac{\lambda_0}{4}, \tag{28}$$

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

$$\begin{aligned}
\lambda_{\max}(\mathcal{H}) &\leq \lambda_{\max}(\mathcal{H}^\infty) + \frac{\lambda_0}{4} + \frac{\lambda_0}{4} \\
&\leq \lambda_1^\infty + \frac{\lambda_0}{2} \triangleq \frac{\lambda_1}{2}.
\end{aligned} \tag{29}$$

This justifies the upper bound assumption of the largest eigen value over iterations. □

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

$$\begin{aligned}
\frac{d}{dt}\|\boldsymbol{z}(t)-\boldsymbol{y}\|_2^2 &\leq -\lambda_0\|\boldsymbol{z}(t)-\boldsymbol{y}\|_2^2 + \lambda_1\left\langle \boldsymbol{z}(t)-\boldsymbol{y}, \nabla_{\boldsymbol{z}(t)} g(\boldsymbol{w}(t),\boldsymbol{a}(t),\boldsymbol{z}(t))\right\rangle \\
&\leq -\lambda_0\|\boldsymbol{z}(t)-\boldsymbol{y}\|_2^2 + \lambda_1 \underbrace{\|\boldsymbol{z}(t)-\boldsymbol{y}\|_2 \left\|\nabla_{\boldsymbol{z}(t)} g(\boldsymbol{w}(t),\boldsymbol{a}(t),\boldsymbol{z}(t))\right\|_2}_{\text{Cauchy-Schwarz inequality}} \\
&\leq -\lambda_0\|\boldsymbol{z}(t)-\boldsymbol{y}\|_2^2 + \lambda_1\|\boldsymbol{z}(t)-\boldsymbol{y}\|_2 \left\|\nabla_{\boldsymbol{z}(t)} g(\boldsymbol{w}(t),\boldsymbol{a}(t),\boldsymbol{z}(t))\right\|_1 \\
&\leq -\lambda_0\|\boldsymbol{z}(t)-\boldsymbol{y}\|_2^2 \lambda_1\|\boldsymbol{z}(t)-\boldsymbol{y}\|_2 \sum_{q=1}^{n}\left|\frac{1}{\sqrt{m}}\sum_{r=1}^{m} a_r w_r 1_{\{w_r z_q \geq 0\}}\right| \\
&\leq -\lambda_0\|\boldsymbol{z}(t)-\boldsymbol{y}\|_2^2 \lambda_1\|\boldsymbol{z}(t)-\boldsymbol{y}\|_2 \frac{n}{\sqrt{m}}\left|\sum_{r=1}^{m} a_r w_r\right|
\end{aligned} \tag{30}$$

Substituting equation (25) in equation (30), we get

$$\begin{aligned}
\frac{d}{dt}\|\boldsymbol{z}(t)-\boldsymbol{y}\|_2^2 &\leq -\lambda_0\|\boldsymbol{z}(t)-\boldsymbol{y}\|_2^2 + \lambda_1\|\boldsymbol{z}(t)-\boldsymbol{y}\|_2 \frac{n}{\sqrt{m}}\|\boldsymbol{w}\|_2\sqrt{2\log\left(\frac{2}{\delta}\right)} \\
&\leq -\lambda_0\|\boldsymbol{z}(t)-\boldsymbol{y}\|_2^2 + \frac{\lambda_1 L n\sqrt{2\log\left(\frac{2}{\delta}\right)}}{\sqrt{m}}\|\boldsymbol{z}(t)-\boldsymbol{y}\|_2.
\end{aligned} \tag{31}$$

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

$$\frac{d}{dt}\left\|\boldsymbol{z}(t)-\boldsymbol{y}\right\|_2^2 \leq -\lambda_0\left\|\boldsymbol{z}(t)-\boldsymbol{y}\right\|_2^2 + \lambda_1\mu\left\|\boldsymbol{z}(t)-\boldsymbol{y}\right\|_2 \tag{32}$$

The above non-linear ODE is a special Bernoulli Differential Equation (BDE)[2] which has known exact solutions [19]. For simplicity, let us suppose $\psi = \left\|\boldsymbol{z}(t)-\boldsymbol{y}\right\|_2^2$. Now,

$$\frac{d\psi}{dt} \leq -\lambda_0\psi + \lambda_1\mu\psi^{1/2} \tag{33}$$

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

$$\nu = C_1\exp\left(-\frac{\lambda_0 t}{2}\right),\ \zeta = C_2 + \frac{\kappa\mu}{C_1}\exp\left(\frac{\lambda_0 t}{2}\right), \tag{34}$$

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, $\kappa^\infty$ 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 $\psi$, we get the exact solution:

$$\left\|\boldsymbol{z}(t)-\boldsymbol{y}\right\|_2 \leq \left(\left\|\boldsymbol{z}(0)-\boldsymbol{y}\right\|_2 - \kappa\mu\right)\exp\left(-\frac{\lambda_0}{2}t\right) + \kappa\mu. \tag{35}$$

From equation (12) in the warm-up exercise, we know for $0 \leq s \leq t$,

$$\left\|\frac{\partial\mathcal{L}_{sup}\left(\boldsymbol{U},\boldsymbol{v}\right)}{\partial\boldsymbol{u}_j(s)}\right\|_2 \leq \frac{\sqrt{n}}{\sqrt{m}}\left\|\boldsymbol{z}(s)-\boldsymbol{y}\right\|_2. \tag{36}$$

Now, substituting equation (35), we get

$$\begin{aligned}&\left\|\frac{\partial\mathcal{L}_{sup}\left(\boldsymbol{U},\boldsymbol{v}\right)}{\partial\boldsymbol{u}_j(s)}\right\|_2\\ &\leq \frac{\sqrt{n}}{\sqrt{m}}\left(\left\|\boldsymbol{z}(0)-\boldsymbol{y}\right\|_2 - \kappa\mu\right)\exp\left(-\frac{\lambda_0}{2}s\right) + \frac{\sqrt{n}}{\sqrt{m}}\kappa\mu.\end{aligned} \tag{37}$$

Therefore, the reaction dynamics are governed by

$$\mathfrak{R}_j^u(\boldsymbol{u}_j(t)) \leq \frac{\sqrt{n}}{\sqrt{m}}\left(\left\|\boldsymbol{z}(0)-\boldsymbol{y}\right\|_2 - \kappa\mu\right)\exp\left(-\frac{\lambda_0}{2}t\right) + \frac{\sqrt{n}}{\sqrt{m}}\kappa\mu. \tag{38}$$

### B.2.2 Diffusion Term

The augmented part on the other hand becomes:

$$\begin{aligned}&\left\|\frac{\partial}{\partial\boldsymbol{u}_j(s)}\sum_{p=1}^{n} g\left(\boldsymbol{w},a,z_p\right)\right\|_2\\ &= \left\|\sum_{p=1}^{n}\sum_{r=1}^{m}\frac{1}{\sqrt{m}}a_r 1_{\{w_r z_p \geq 0\}} w_r \frac{1}{\sqrt{m}} v_j 1_{\left\{\boldsymbol{v}_j^T\boldsymbol{x}_p \geq 0\right\}}\boldsymbol{x}_p\right\|_2.\end{aligned} \tag{39}$$

By Triangle and Cauchy-Schwarz inequality, we get

$$\begin{aligned}\left\|\frac{\partial}{\partial\boldsymbol{u}_j(s)}\sum_{p=1}^{n} g\left(\boldsymbol{w},a,z_p\right)\right\|_2 &\leq \frac{1}{m}\sum_{p=1}^{n}\left\|v_j 1_{\left\{\boldsymbol{v}_j^T\boldsymbol{x}_p \geq 0\right\}}\boldsymbol{x}_p\sum_{r=1}^{m} a_r w_r 1_{\{w_r z_p \geq 0\}}\right\|_2\\ &\leq \frac{1}{m}\sum_{p=1}^{n}\left|v_j\right|\left\|\boldsymbol{x}_p\right\|_2\left|\sum_{r=1}^{m} a_r w_r\right|\\ &\leq \frac{1}{m}\sum_{p=1}^{n}\left|\sum_{r=1}^{m} a_r w_r\right|\end{aligned} \tag{40}$$

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

Substituting equation (25) in equation (40), we arrive at the following inequality:

$$\begin{aligned}\left\|\frac{\partial}{\partial \boldsymbol{u}_j(s)}\sum_{p=1}^{n} g\left(\boldsymbol{w}, a, z_p\right)\right\|_2 &\leq \frac{1}{m}\sum_{p=1}^{n}\|\boldsymbol{w}\|_2\sqrt{2\log\left(\frac{2}{\delta}\right)} \\ &\leq \frac{Ln\sqrt{2\log\left(\frac{2}{\delta}\right)}}{m} = \mathcal{O}\left(\frac{\mu}{\sqrt{m}}\right).\end{aligned} \tag{41}$$

Thus, the diffusion dynamics are given by

$$\mathfrak{D}_j^u(\boldsymbol{u}_j(t)) \leq \frac{Ln\sqrt{2\log\left(\frac{2}{\delta}\right)}}{m}. \tag{42}$$

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

$$\begin{aligned}\|\boldsymbol{u}_j(t) - \boldsymbol{u}_j(0)\|_2 &\leq \int_0^t \left\|\frac{d\boldsymbol{u}_j(s)}{ds}\right\|_2 ds \\ &\leq \int_0^t \frac{\sqrt{n}}{\sqrt{m}}\left(\|\boldsymbol{z}(0)-\boldsymbol{y}\|_2 - \kappa\mu\right)\exp\left(-\frac{\lambda_0}{2}s\right)ds + \int_0^t \frac{\mu\left(1+\kappa\sqrt{n}\right)}{\sqrt{m}}ds \\ &\leq \frac{2\sqrt{n}\left(\|\boldsymbol{z}(0)-\boldsymbol{y}\|_2 - \kappa\mu\right)}{\sqrt{m}\lambda_0}\left(1-\exp\left(-\frac{\lambda_0}{2}t\right)\right) + \left(\frac{\mu\left(1+\kappa\sqrt{n}\right)}{\sqrt{m}}\right)t.\end{aligned} \tag{43}$$

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,

$$\|\boldsymbol{u}_j(t) - \boldsymbol{u}_j(0)\|_2 \leq \mathcal{O}\left(\frac{n^{3/2}}{m^{1/2}\lambda_0\delta} + \left(\frac{\mu\left(1+\kappa\sqrt{n}\right)}{m^{1/2}}\right)t\right). \tag{44}$$

Furthermore, the spatial grid of neurons satisfies:

$$\begin{aligned}\|\boldsymbol{U}(t) - \boldsymbol{U}(0)\|_F &\leq \sqrt{m}\,\|\boldsymbol{u}_j(t) - \boldsymbol{u}_j(0)\|_2 \\ &\leq \mathcal{O}\left(\frac{n^{3/2}}{\lambda_0\delta} + \mu\left(1+\kappa\sqrt{n}\right)t\right).\end{aligned} \tag{45}$$

To circumvent tractability issues, it is common to seek an $\epsilon$-stationary point. As given by equation (35), $\boldsymbol{z}(t)$ in adversarial learning converges uniformly to an $\epsilon$-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 $\mu$. It is an important result in a sense that over-parameterized networks can still enjoy linear rate of convergence even under adversarial interaction.

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 $\epsilon$-stationary point:

$$\mathcal{O}\left(\frac{n^{3/2}}{m^{1/2}\lambda_0\delta} + \left(\frac{\mu\left(1+\kappa\sqrt{n}\right)}{m^{1/2}}\right)T_0\right) \leq R. \tag{46}$$

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

$$m = \Omega\left(\left(\frac{n^{7/2}}{\lambda_0^2\delta^2} + \frac{n^2\mu\left(1+\kappa\sqrt{n}\right)T_0}{\lambda_0\delta}\right)^2\right). \tag{47}$$

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 $\epsilon$-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. □

### B.3 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.

$$\begin{aligned}\frac{\partial \mathcal{L}_{sup}\left(\boldsymbol{U},\boldsymbol{V}\right)}{\partial u_{jk}} &= \frac{1}{2}\sum_{p=1}^{n}\frac{\partial}{\partial u_{jk}}\sum_{i=1}^{d_{out}}\left(\boldsymbol{z}_p-\boldsymbol{y}_p\right)_i^2 \\ &= \sum_{p=1}^{n}\sum_{i=1}^{d_{out}}\left(\boldsymbol{z}_p-\boldsymbol{y}_p\right)_i\frac{\partial z_{p,i}}{\partial u_{jk}} \\ &= \sum_{p=1}^{n}\sum_{i=1}^{d_{out}}\left(\boldsymbol{z}_p-\boldsymbol{y}_p\right)_i\frac{1}{\sqrt{d_{out}m}}v_{ij}\frac{\partial}{\partial u_{jk}}\sigma\left(\boldsymbol{u}_j^T\boldsymbol{x}_p\right) \\ &= \frac{1}{\sqrt{d_{out}m}}\sum_{p=1}^{n}\sum_{i=1}^{d_{out}}\left(\boldsymbol{z}_p-\boldsymbol{y}_p\right)_i v_{ij}1_{\left\{\boldsymbol{u}_j^T\boldsymbol{x}_p\geq 0\right\}}x_{p,k}.\end{aligned} \tag{48}$$

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

$$\begin{aligned}\frac{\partial \mathcal{L}_{sup}\left(\boldsymbol{U},\boldsymbol{V}\right)}{\partial v_{ij}} &= \frac{1}{2}\sum_{p=1}^{n}\frac{\partial}{\partial v_{ij}}\sum_{i=1}^{d_{out}}\left(\boldsymbol{z}_p-\boldsymbol{y}_p\right)_i^2 \\ &= \sum_{p=1}^{n}\left(\boldsymbol{z}_p-\boldsymbol{y}_p\right)_i\frac{\partial z_{p,i}}{\partial v_{ij}} \\ &= \sum_{p=1}^{n}\left(\boldsymbol{z}_p-\boldsymbol{y}_p\right)_i\frac{1}{\sqrt{d_{out}m}}\frac{\partial}{\partial v_{ij}}\sum_{j=1}^{m}v_{ij}\sigma\left(\boldsymbol{u}_j^T\boldsymbol{x}_p\right) \\ &= \frac{1}{\sqrt{d_{out}m}}\sum_{p=1}^{n}\left(\boldsymbol{z}_p-\boldsymbol{y}_p\right)_i 1_{\left\{\boldsymbol{u}_j^T\boldsymbol{x}_p\geq 0\right\}}\boldsymbol{u}_j^T\boldsymbol{x}_p.\end{aligned} \tag{49}$$

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

$$\frac{\partial \mathcal{L}_{adv}\left(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\boldsymbol{a}\right)}{\partial u_{jk}} = \frac{1}{m\sqrt{d_{out}}}\sum_{p=1}^{n}\boldsymbol{a}^T\mathrm{diag}\left(1_{\{\boldsymbol{W}\boldsymbol{V}\sigma(\boldsymbol{U}\boldsymbol{x}_p)\geq 0\}}\right)\boldsymbol{W}\boldsymbol{v}_j 1_{\left\{\boldsymbol{u}_j^T\boldsymbol{x}_p\geq 0\right\}}x_{p,k}. \tag{50}$$

The weights in the top layer are governed by:

$$\frac{\partial \mathcal{L}_{adv}\left(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\boldsymbol{a}\right)}{\partial v_{ij}} = \frac{1}{m\sqrt{d_{out}}}\sum_{p=1}^{n}\boldsymbol{a}^T\mathrm{diag}\left(1_{\{\boldsymbol{W}\boldsymbol{V}\sigma(\boldsymbol{U}\boldsymbol{x}_p)\geq 0\}}\right)\boldsymbol{W}_{:,i}1_{\left\{\boldsymbol{u}_j^T\boldsymbol{x}_p\geq 0\right\}}\boldsymbol{u}_j^T\boldsymbol{x}_p. \tag{51}$$

Analogous to equation (4), the reaction and diffusion terms in augmented objective are defined as:

$$\mathfrak{R}_j^{\boldsymbol{u}}\left(\boldsymbol{u}_j,\boldsymbol{v}_j\right) \triangleq \left\{\frac{1}{\sqrt{d_{out}m}}\sum_{p=1}^{n}\sum_{i=1}^{d_{out}}\left(\boldsymbol{y}_p-\boldsymbol{z}_p\right)_i v_{ij}1_{\left\{\boldsymbol{u}_j^T\boldsymbol{x}_p\geq 0\right\}}x_{p,k}\right\}_{k=1}^{d_{in}}, \tag{52}$$

$$\mathfrak{D}_j^{\boldsymbol{u}}\left(\nabla^2\boldsymbol{u}_j\right) \triangleq \left\{\frac{1}{m\sqrt{d_{out}}}\sum_{p=1}^{n}\boldsymbol{a}^T\mathrm{diag}\left(1_{\{\boldsymbol{W}\boldsymbol{V}\sigma(\boldsymbol{U}\boldsymbol{x}_p)\geq 0\}}\right)\boldsymbol{W}\boldsymbol{v}_j 1_{\left\{\boldsymbol{u}_j^T\boldsymbol{x}_p\geq 0\right\}}x_{p,k}\right\}_{k=1}^{d_{in}}, \tag{53}$$

$$\mathfrak{R}_j^{\boldsymbol{v}}\left(\boldsymbol{u}_j,\boldsymbol{v}_j\right) \triangleq \left\{\frac{1}{\sqrt{d_{out}m}}\sum_{p=1}^{n}\left(\boldsymbol{y}_p-\boldsymbol{z}_p\right)_i 1_{\left\{\boldsymbol{u}_j^T\boldsymbol{x}_p\geq 0\right\}}\boldsymbol{u}_j^T\boldsymbol{x}_p\right\}_{i=1}^{d_{out}}, \tag{54}$$

$$\mathfrak{D}_j^{\boldsymbol{v}}\left(\nabla^2 \boldsymbol{v}_j\right) \triangleq \left\{ \frac{1}{m\sqrt{d_{out}}} \sum_{p=1}^{n} \boldsymbol{a}^T \text{diag}\left(1_{\{\boldsymbol{W}\boldsymbol{V}\sigma(\boldsymbol{U}\boldsymbol{x}_p)\geq 0\}}\right) \boldsymbol{W}_{:,i} 1_{\left\{\boldsymbol{u}_j^T \boldsymbol{x}_p \geq 0\right\}} \boldsymbol{u}_j^T \boldsymbol{x}_p \right\}_{i=1}^{d_{out}}. \tag{55}$$

Ignoring constants and assuming $\left(\boldsymbol{y}_p - \boldsymbol{z}_p\right)_i v_{ij} 1_{\left\{\boldsymbol{u}_j^T \boldsymbol{x}_p \geq 0\right\}} x_{p,k} = \mathcal{O}\left(1\right)$, we get asymptotic bounds on the norm of reaction and diffusion terms[3]:

$$\begin{aligned}
\mathfrak{R}_j^{\boldsymbol{u}}\left(\boldsymbol{u}_j, \boldsymbol{v}_j\right) &= \mathcal{O}\left(nd_{in}\sqrt{\frac{d_{out}}{m}}\right), \\
\mathfrak{D}_j^{\boldsymbol{u}}\left(\nabla^2 \boldsymbol{u}_j\right) &= \mathcal{O}\left(nm^2 d_{in} d_{out}^{3/2}\right), \\
\mathfrak{R}_j^{\boldsymbol{v}}\left(\boldsymbol{u}_j, \boldsymbol{v}_j\right) &= \mathcal{O}\left(nd_{in}\sqrt{\frac{d_{out}}{m}}\right), \\
\mathfrak{D}_j^{\boldsymbol{v}}\left(\nabla^2 \boldsymbol{v}_j\right) &= \mathcal{O}\left(nm^2 d_{in} d_{out}^{1/2}\right),
\end{aligned} \tag{56}$$

which completes the proof. □

# Appendix C More Related Works

## C.1 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 of 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 [20]. 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 [2]. To model interactions of molecular elements in gradient analysis, Gregor et al.[21] 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 [9, 10]. 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 [4]. 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 [3]. 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.

## C.2 Adversarial Learning

Recent success of Generative Adversarial Networks (GANs) [22, 14] has led to exciting applications in a wide variety of tasks [23, 24, 25, 26, 27]. 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 [28]. Further, it has been reported in copious literature that supervised learning with adversarial regularization performs better than sole supervision [25, 29,

[3] 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.

30, 31, 32, 33]. 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 symmetry and homogeneity to create non-homogeneous patterns in weight space.

### C.3 Bernoulli Differential Equation

Bernoulli differential equation was discussed in 1695 [19]. This fundamental equation arises naturally in a wide variety fields, such as modelling of population growth [34], 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 [35]. In Verhulst model [34] of population growth, the rate of reproduction is proportional to current population and available resources. Formally,

$$\frac{dP}{dt} = rP\left(1 - \frac{P}{K}\right), \tag{57}$$

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 [34] 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,

$$\frac{d\psi}{dt} \leq r\psi^{1/2}\left(1 - \frac{\psi^{1/2}}{K}\right), \tag{58}$$

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.

## Appendix D Further Discussion of Insights from Analysis

It is well known that randomly initialized gradient descent with over-parameterization finds solutions close to its initialization [11, 36, 37, 38]. The distance from initialization has helped unveil several mysteries of deep learning in part including the generalization puzzle and $\epsilon$-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. [8] who show that the generator in WGAN with weight clipping [14] 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 such implicit restriction to a tiny subspace is not a necessary condition. With over-parameterization, randomly initialized gradient descent-ascent 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.

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.