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

Litu Rout

Association for the Advancement of Artificial Intelligence (AAAI-21)

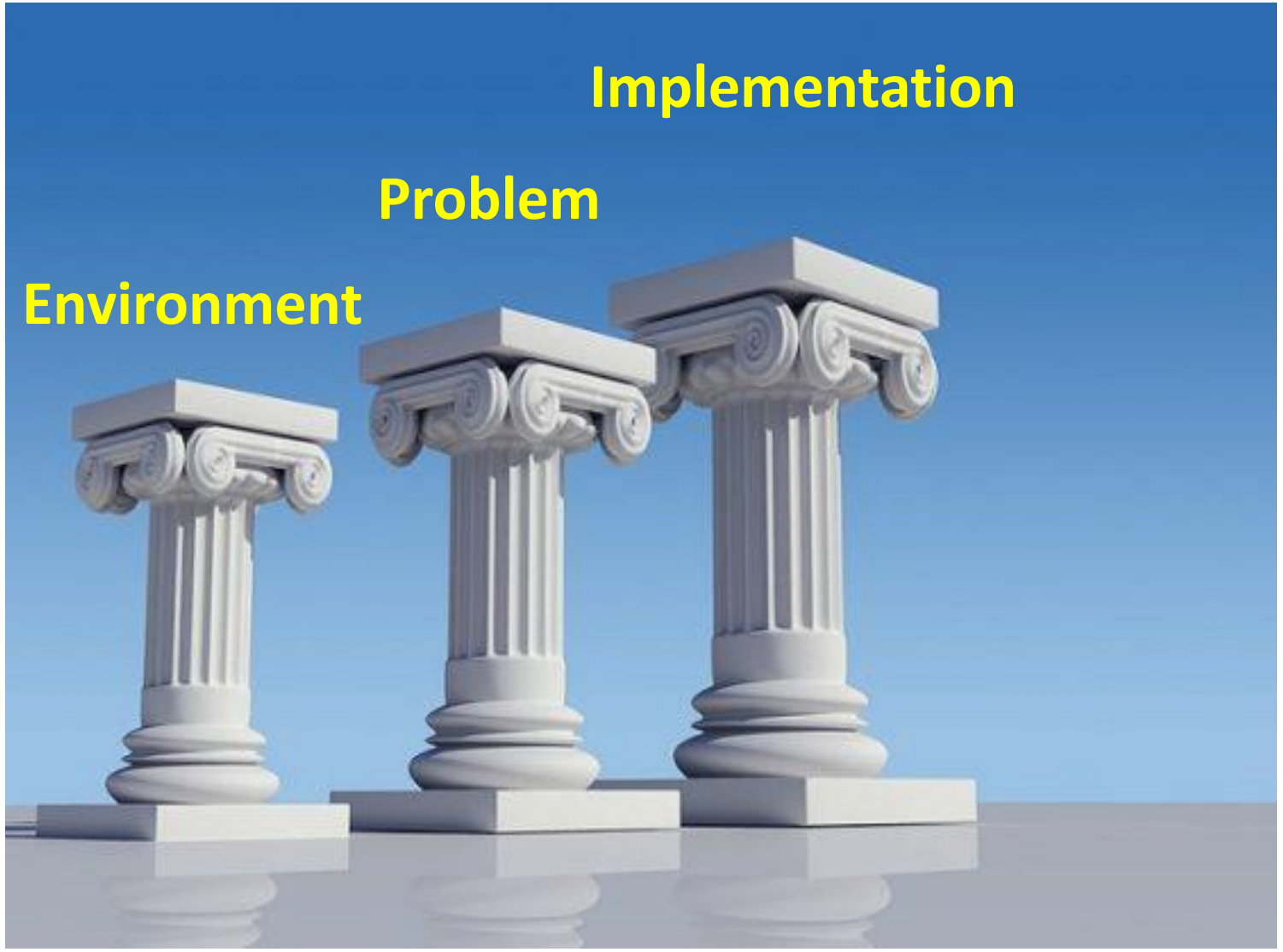
Preprint: https://liturout.github.io/data/aaai21_preprint.pdf

June 16, 2021

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

- Adversarial Interaction
  - Generative Adversarial Networks (GANs)
  - Application of conditional GANs
- Non-Homogeneous Patterns
  - Homogeneous patterns
  - Supervised learning
- Reaction-Diffusion
  - Turing's RD model (1952)
  - Gray-Scott RD model (1984)
- Turing Instability
  - Reaction dynamics
  - Diffusion dynamics

Three Pillars of Deep Learning



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

- Adversarial Interaction
  - Generative Adversarial Networks (GANs)
  - Application of conditional GANs
- Non-Homogeneous Patterns
  - Homogeneous patterns
  - Supervised learning
- Reaction-Diffusion
  - Turing's RD model (1952)
  - Gray-Scott RD model (1984)
- Turing Instability
  - Reaction dynamics
  - Diffusion dynamics

Outline

- Maximum Likelihood Estimation
- Reverse KL Divergence
- Introduction to GANs
- GANs in Remote Sensing
  - Problem Formulation

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

- Adversarial Interaction
  - Generative Adversarial Networks (GANs)
  - Application of conditional GANs
- Non-Homogeneous Patterns
  - Homogeneous patterns
  - Supervised learning
- Reaction-Diffusion
  - Turing's RD model (1952)
  - Gray-Scott RD model (1984)
- Turing Instability
  - Reaction dynamics
  - Diffusion dynamics

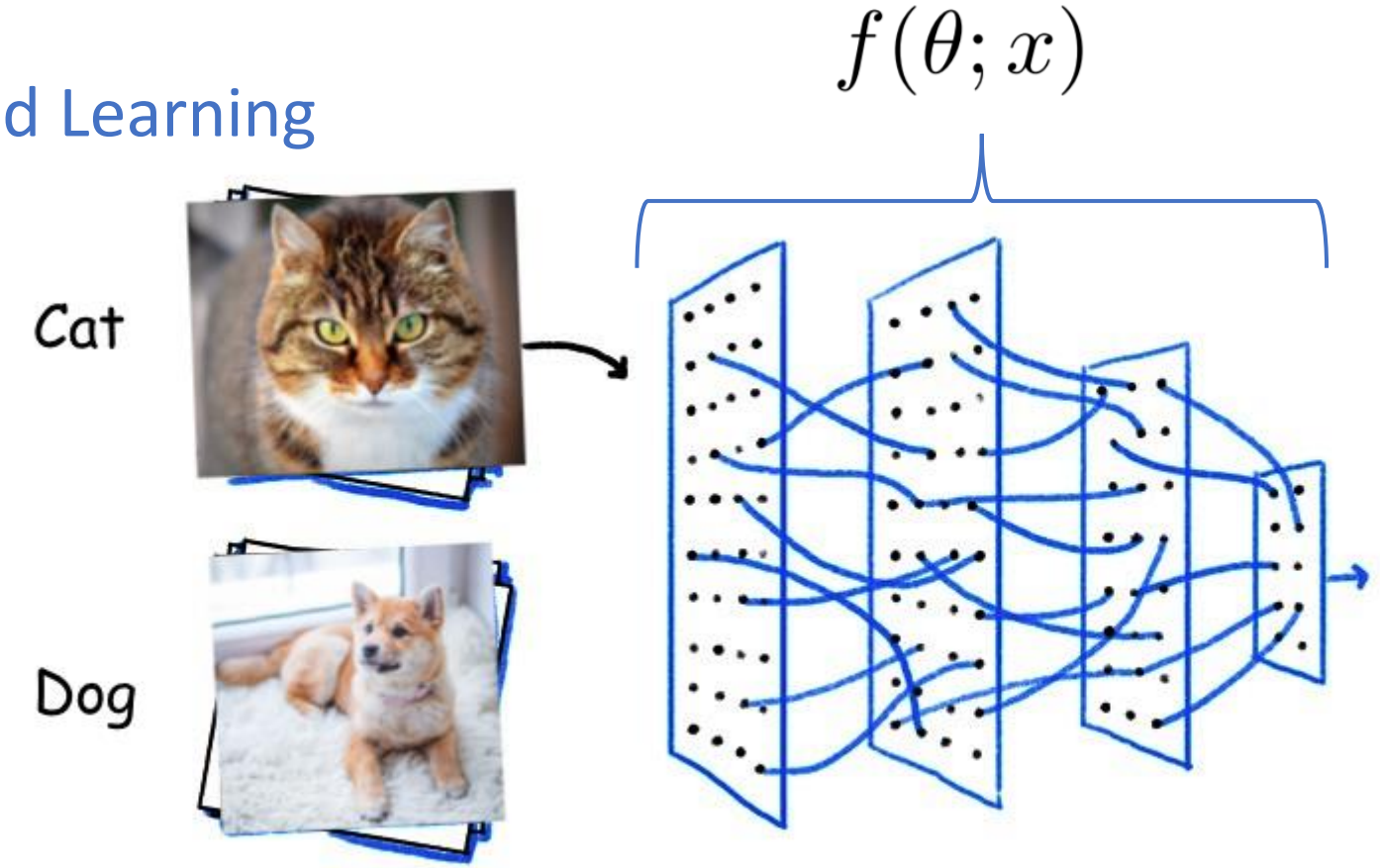
Understanding the Role of Adversarial Regularization in Supervised Learning

- Mitigating Vanishing Gradient
- Asymptotic Iteration Complexity
- Sub-optimality Gap
- Provable Convergence
- Generalization Error (Open Problems)
- Neural Topology Analysis (Open Problems)

L.R., CVPR-AML-CV 2021: https://liturout.github.io/data/cvpr2021_aml_cv.pdf

Problem Setup

- Supervised Learning

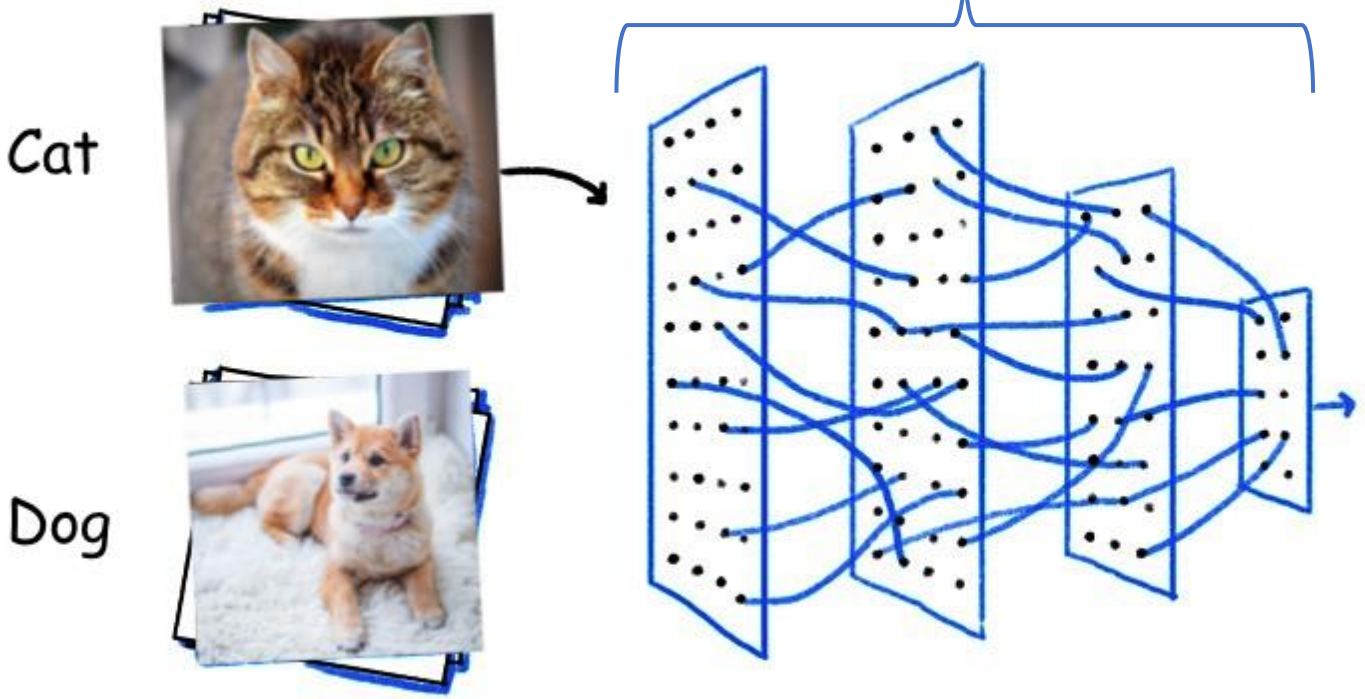


$$\arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{P}} \left[l\left(f(\theta; x); y \right) \right]$$

Problem Setup

- Adversarial Learning

Input: x



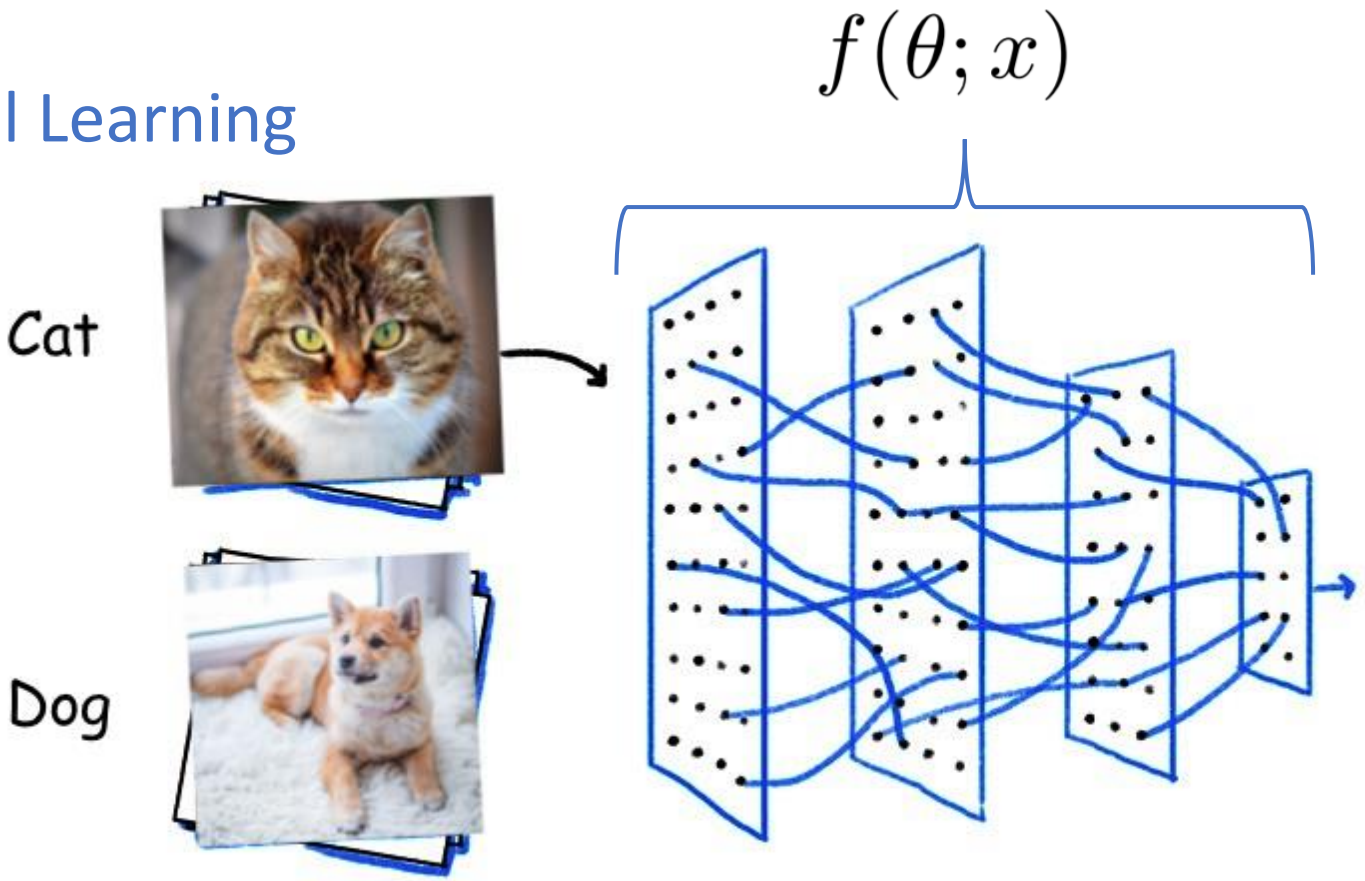
$f(\theta; x)$

Output: y

Problem Setup

Detective
(Discriminator: g(.))

- Adversarial Learning



Input: x

Output: y

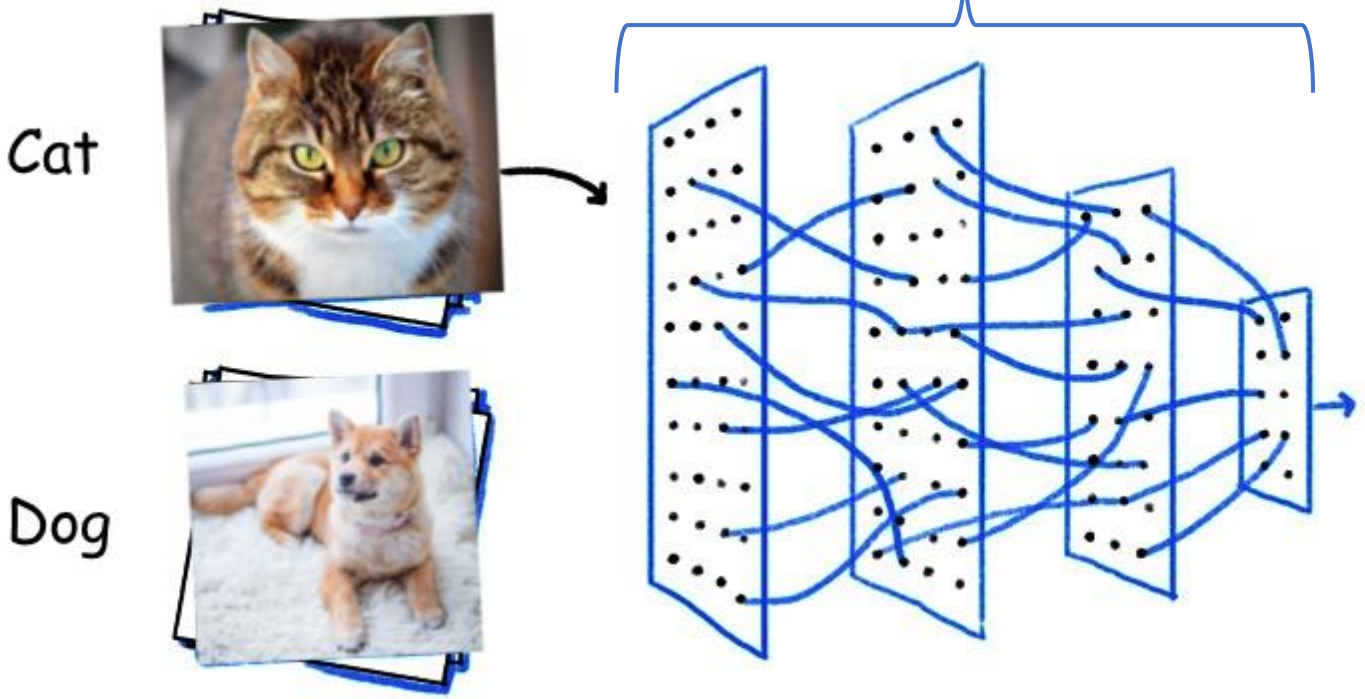
$$\arg\min_{\theta} -\mathbb{E}_{x\sim\mathcal{P}_X}\left[g\left(\psi; f\left(\theta; x\right)\right)\right]$$

Problem Setup

Detective
(Discriminator: g(.))

- Adversarial Regularization

Input: x



$f(\theta; x)$

Output: y

$$\arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{P}} \left[l\left(f(\theta; x); y \right) - g\left(\psi; f\left(\theta; x \right) \right) \right]$$

Problem Setup

- Supervised Learning

$$\arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{P}} \left[l\left(f(\theta; x); y \right) \right]$$

Problem Setup

- Supervised Learning

$$\arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{P}} \left[l\left(f(\theta; x); y \right) \right]$$

- Adversarial Learning

$$\arg\min_{\theta} -\mathbb{E}_{x\sim\mathcal{P}_X} \left[g\left(\psi; f\left(\theta; x \right) \right) \right]$$

Problem Setup

- Supervised Learning

$$\arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{P}} [l\left(f(\theta; x); y\right)]$$

- Adversarial Learning

$$\arg\min_{\theta} -\mathbb{E}_{x\sim\mathcal{P}_X} [g\left(\psi; f\left(\theta; x\right)\right)]$$

- Adversarial Regularization

$$\arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{P}} [l\left(f(\theta; x); y\right) - g\left(\psi; f\left(\theta; x\right)\right)]$$

Problem Setup

- Supervised Learning

$$\arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{P}} [l\left(f(\theta; x); y\right)]$$

- Adversarial Learning

$$\arg\min_{\theta} -\mathbb{E}_{x\sim\mathcal{P}_X} [g\left(\psi; f\left(\theta; x\right)\right)]$$

- Adversarial Regularization

$$\arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{P}} [l\left(f(\theta; x); y\right) - g\left(\psi; f\left(\theta; x\right)\right)]$$

Understanding the Role of Adversarial Regularization in Supervised Learning

- Mitigating Vanishing Gradient
- Asymptotic Iteration Complexity
- Sub-optimality Gap
- Provable Convergence
- Generalization Error (Open Problems)
- Neural Topology Analysis (Open Problems)

Mitigating Vanishing Gradient

$$\theta(t+1) = \theta(t) - \eta \frac{\partial L(\theta)}{\partial \theta(t)}$$

Supervised Objective

Assumption 1. *The function $f(\theta; x)$ is L-Lipschitz in θ.*

Assumption 2. *The loss function $l(p; y)$, where $p = f(\theta; x)$, is β-smooth in p.*

Lemma 1. *Let **Assumption 1** and **Assumption 2** hold. If $\|\theta - \theta^*\| \leq \epsilon$, then $\left\|\nabla_\theta \mathbb{E}_{(x,y)\sim\mathcal{P}} \left[l\left(f(\theta; x); y\right)\right]\right\| \leq L^2\beta\epsilon$.*

Mitigating Vanishing Gradient

$$\theta(t+1) = \theta(t) - \eta \frac{\partial L(\theta)}{\partial \theta(t)}$$

Augmented Objective

Theorem 1 (Informal).

Under valid assumptions, the adversarial regularization mitigates vanishing gradient in the near optimal region.

Theorem 1. *Let us suppose* ***Assumption 1*** *and* ***Assumption 2*** *hold. If* $\|\theta - \theta^*\| \leq \epsilon$ *and* $\|g - g^*\| \leq \delta$, *then* $\left\|\nabla_\theta \mathbb{E}_{(x,y)\sim\mathcal{P}} \left[l\left(f(\theta; x); y\right) - g\left(\psi; f\left(\theta; x\right)\right)\right]\right\| \leq \left(L^2\beta\epsilon + L\delta\right)$.

Mitigating Vanishing Gradient

Experimental Result

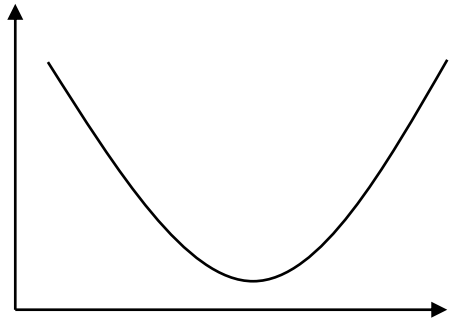
- Gradient norm in hidden layer
- Gradient norm in top layer
- Application of adversarial regularization

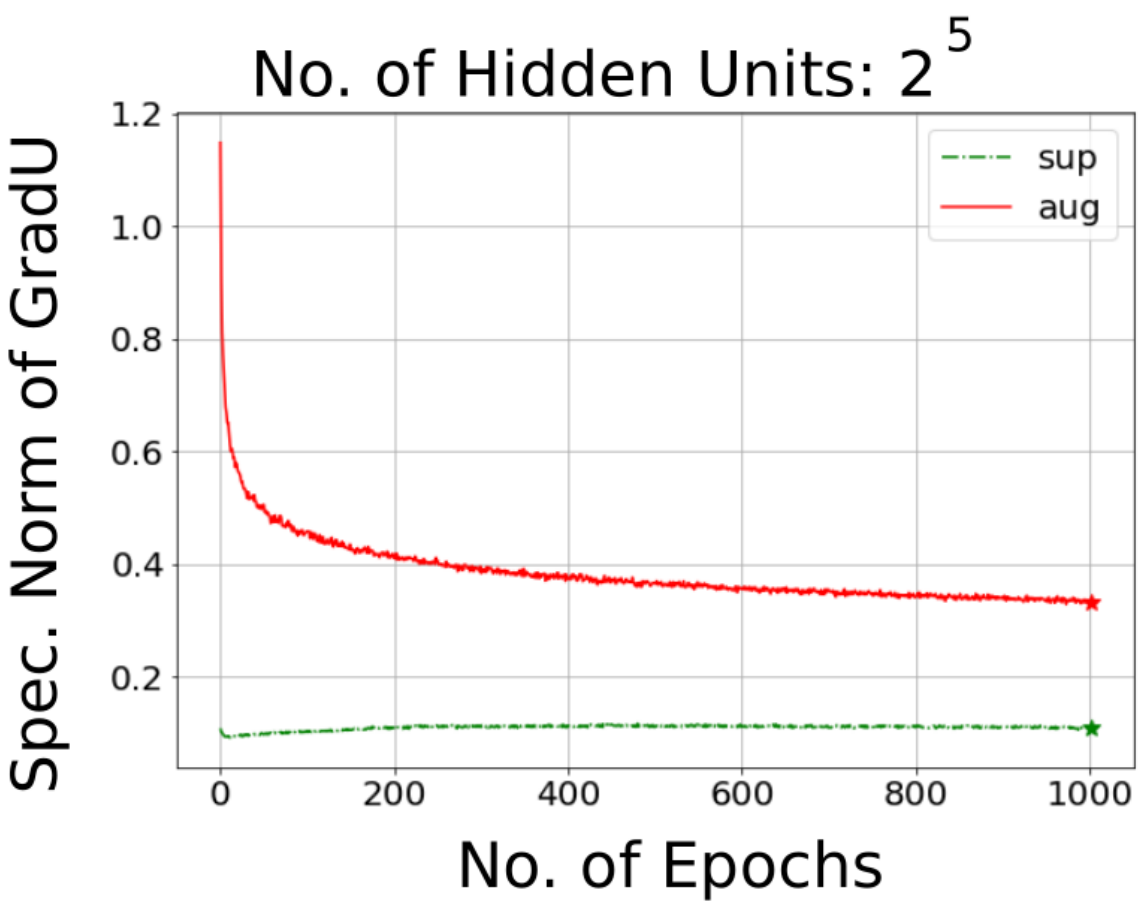
Mitigating Vanishing Gradient

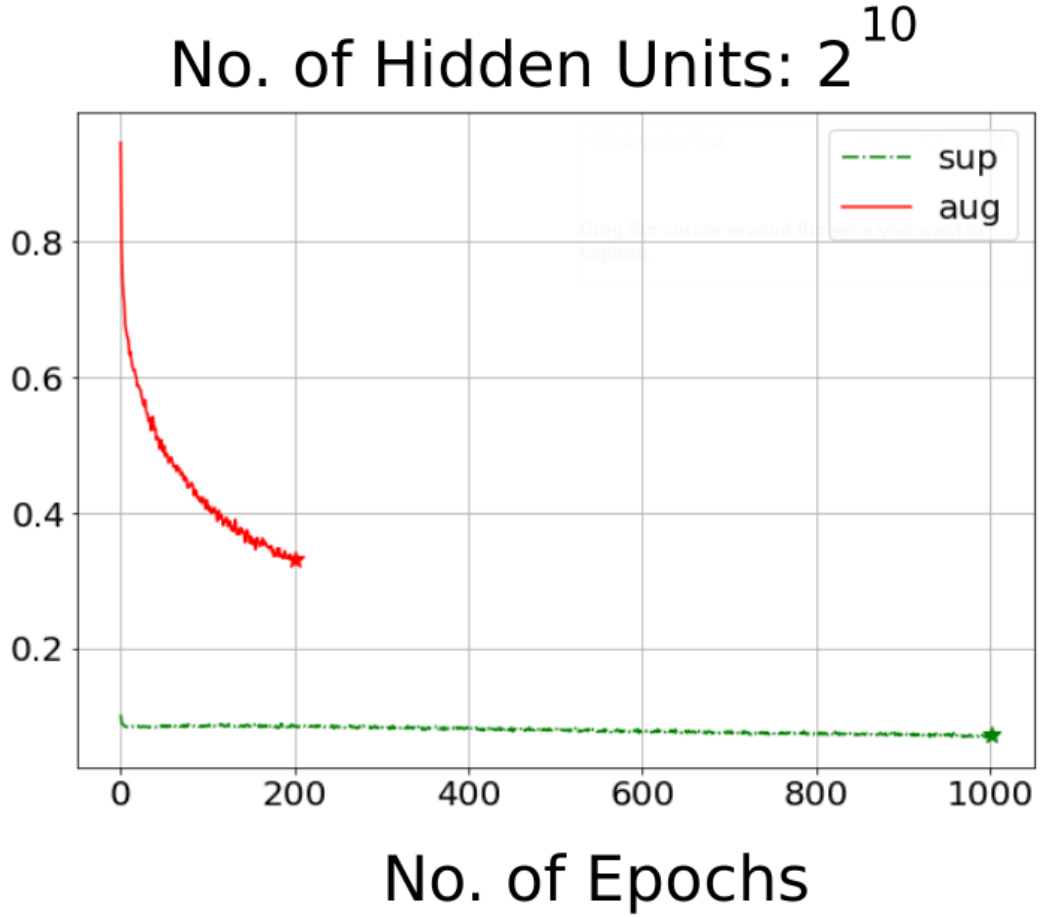
Experimental Result

- Gradient norm in hidden layer
- Gradient norm in top layer
- Application of adversarial regularization

Mitigating Vanishing Gradient





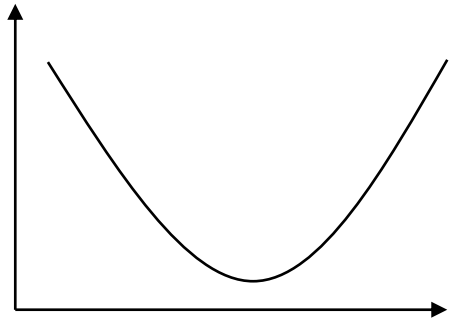


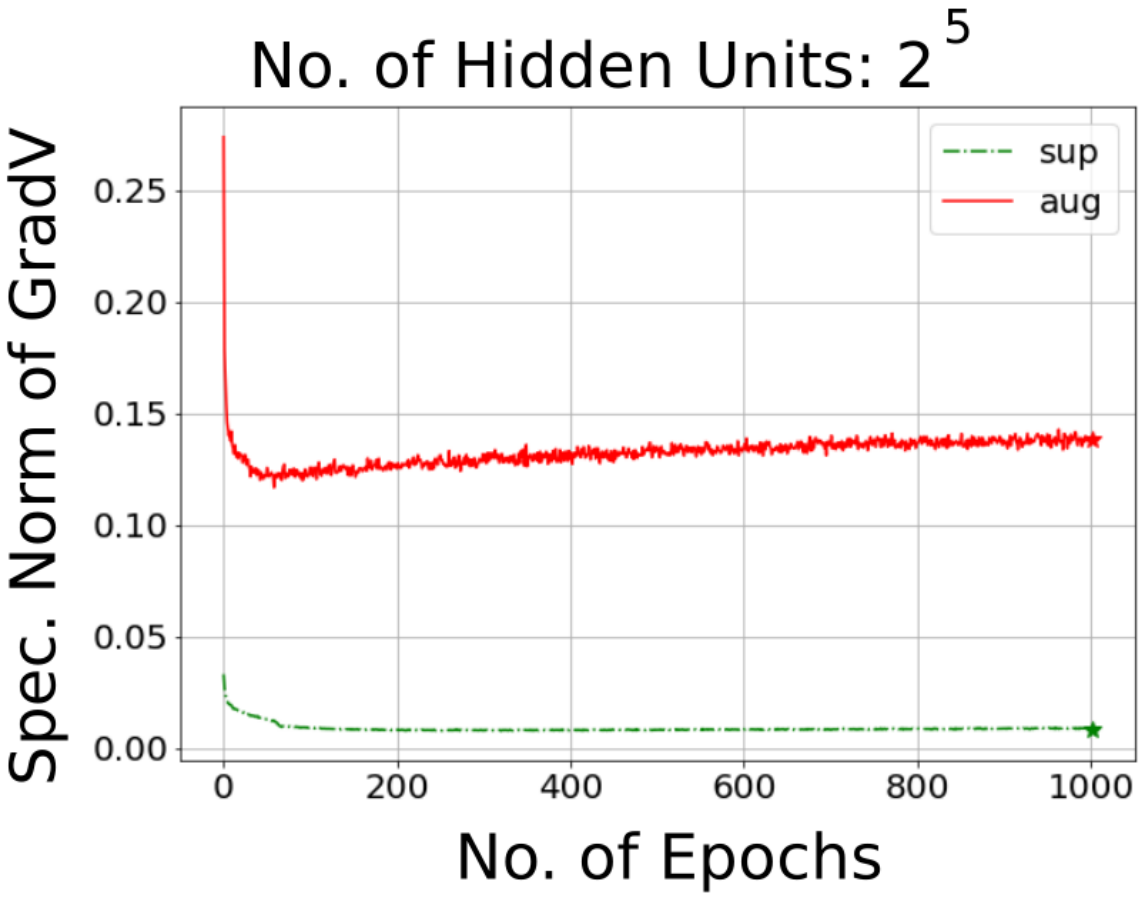
Mitigating Vanishing Gradient

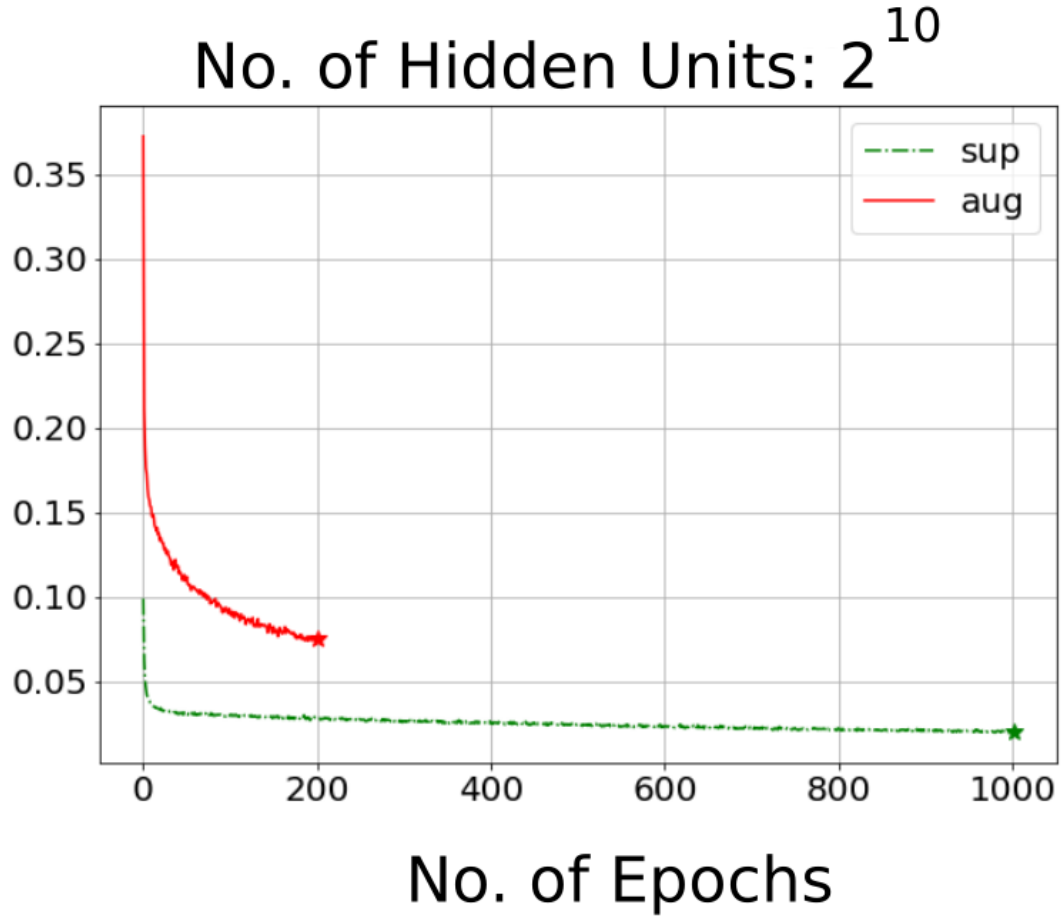
Experimental Result

- Gradient norm in hidden layer
- Gradient norm in top layer
- Application of adversarial regularization

Mitigating Vanishing Gradient







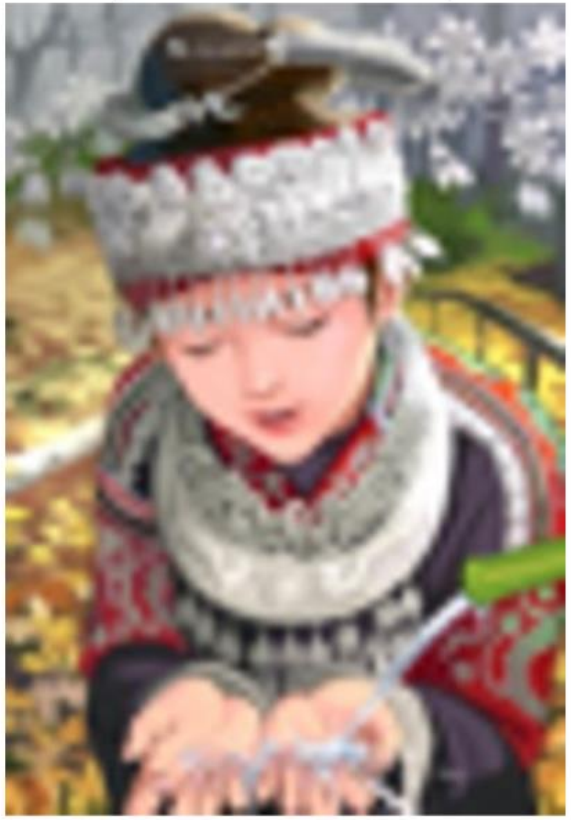
Mitigating Vanishing Gradient

Experimental Result

- Gradient norm in hidden layer
- Gradient norm in top layer
- Application of adversarial regularization

Mitigating Vanishing Gradient

bicubic
(21.59dB/0.6423)



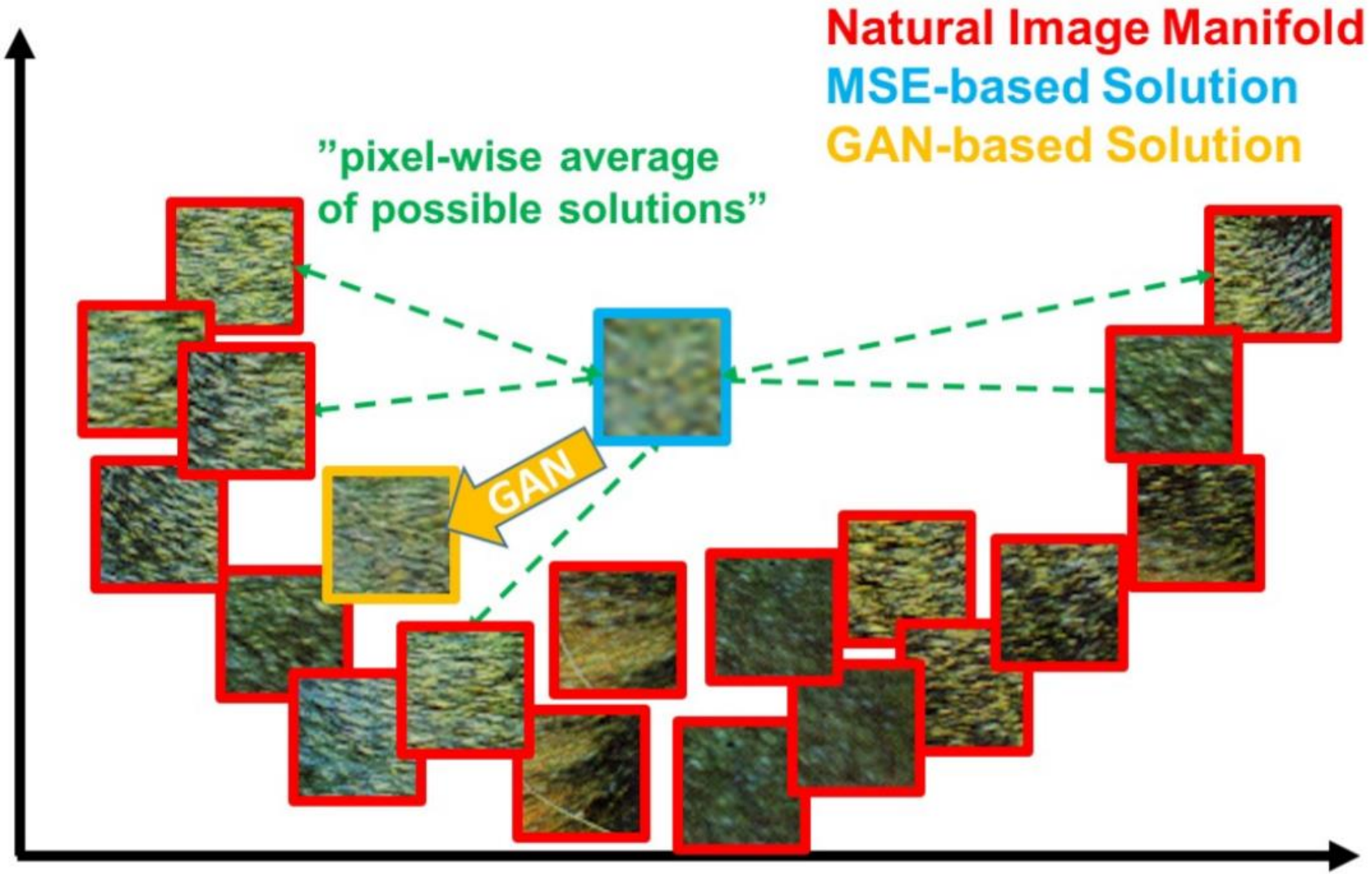
SRResNet
(23.53dB/0.7832)

SRGAN
(21.15dB/0.6868)

original

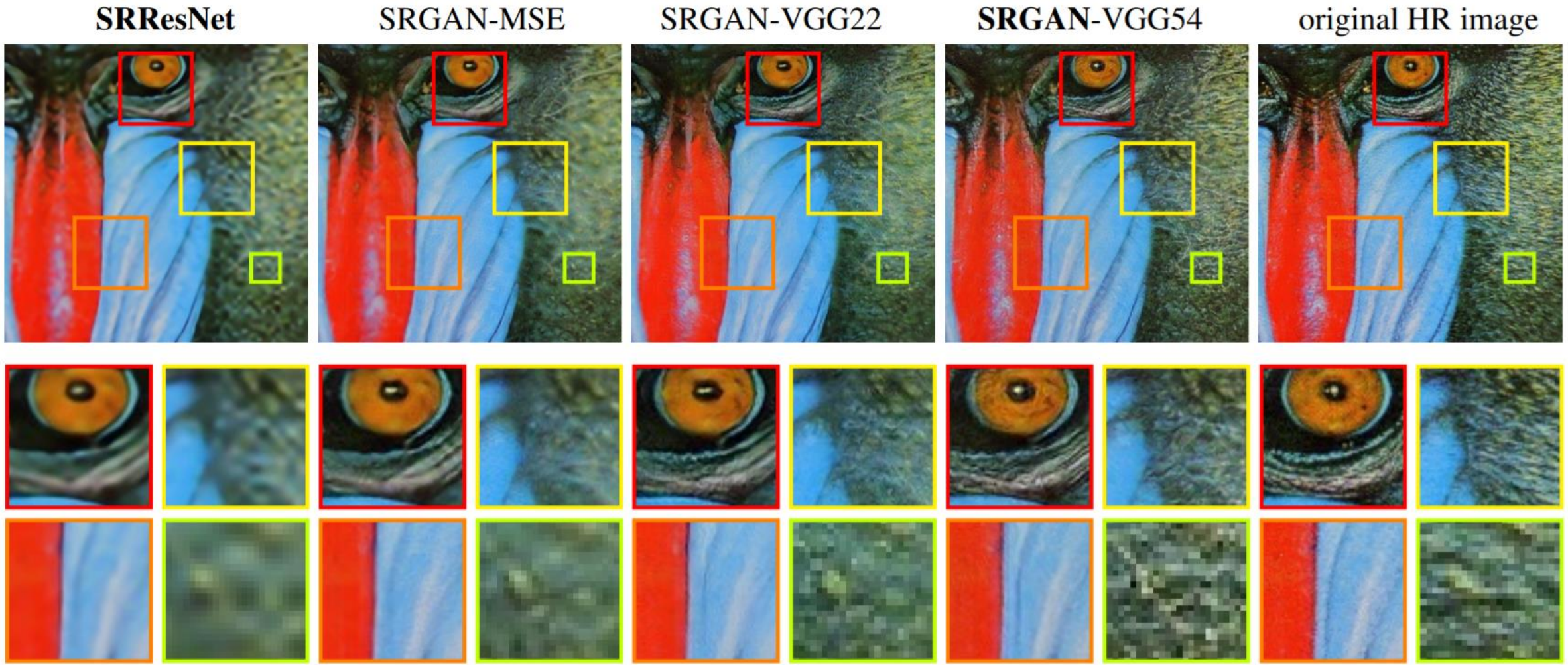
Ledig et al. CVPR, 2017

Mitigating Vanishing Gradient



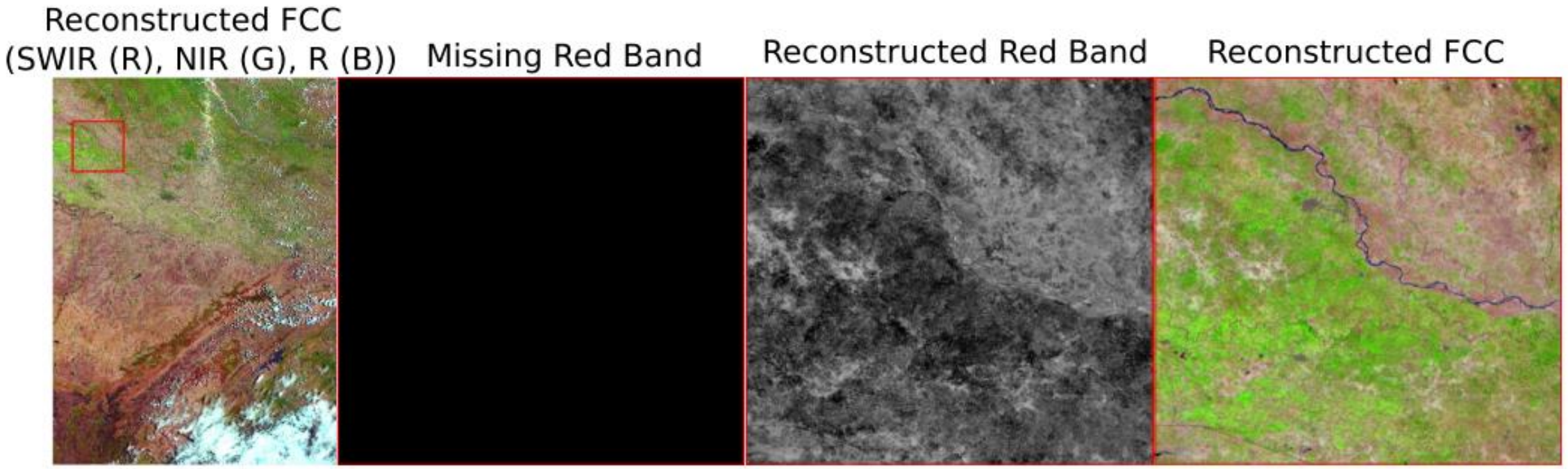
Ledig et al. CVPR, 2017

Mitigating Vanishing Gradient



Ledig et al. CVPR, 2017

Mitigating Vanishing Gradient



Metrics	Parameters	Training epochs	Training RMSE	Inference Time
AeroGAN	2 M	100	41.57	64 s
DeepAWiFS	4.7 M	1000	18.2	495 s
DSen2	4.7 M	1000	21.2	456 s
ALERT	2.8 M	50	10.3	70 s
% gain	40.42	95.00	51.41	84.64

Understanding the Role of Adversarial Regularization in Supervised Learning

- Mitigating Vanishing Gradient
- Asymptotic Iteration Complexity
- Sub-optimality Gap
- Provable Convergence
- Generalization Error (Open Problems)
- Neural Topology Analysis (Open Problems)

Asymptotic Iteration Complexity

Supervised Objective

Assumption 5 ((L_0, L_1)-Smoothness). *The function is (L_0, L_1)-smooth, i.e., there exist positive constants L_0 and L_1 such that* $\left\|\nabla^2 l(\theta)\right\| \leq L_0 + L_1 \left\|\nabla l(\theta)\right\|$.

Theorem 2. *Suppose the functions in $\mathscr{L}$ satisfy* ***Assumption 3, 4*** *and* ***5****. Given $\epsilon > 0$, the iteration complexity in sole supervision is upper bounded by*

$$\mathcal{O}\left(\frac{(l(\theta_0)-l^*)\left(L_0+L_1L^2\beta\epsilon\right)}{\epsilon^2}\right).$$

Asymptotic Iteration Complexity

Augmented Objective

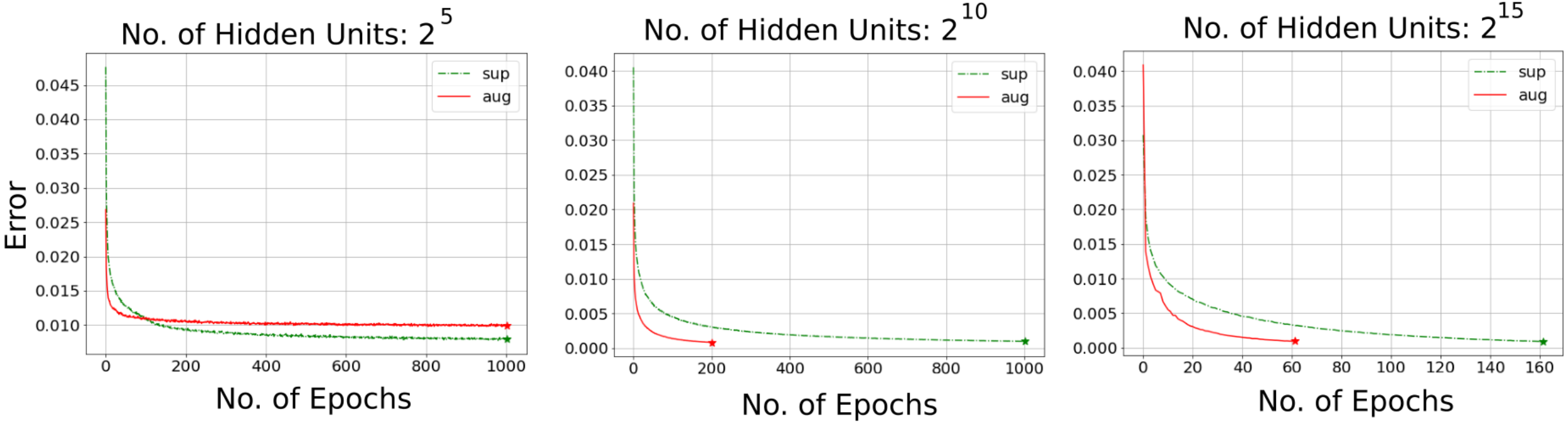
Theorem 2 and 3 (Informal).

Under valid assumptions, the number of iterations required to find a good solution in the augmented objective is equal to or less than supervised objective.

Theorem 3. *Suppose the functions in* $\mathscr{L}$ *satisfy* ***Assumption 3, 4*** *and* ***5****. Given* ***Assumption 6*** *holds,* $\epsilon > 0$ *and* $\delta \leq \frac{\sqrt{2\epsilon\zeta}}{L}$*, the iteration complexity in adversarial regularization is upper bounded by* $\mathcal{O}\left(\frac{(l(\theta_0)-l^*)\left(L_0+L_1L^2\beta\epsilon\right)}{\epsilon^2+2\epsilon\zeta-L^2\delta^2}\right)$.

Asymptotic Iteration Complexity

Experimental Result



Asymptotic Iteration Complexity

Experimental Result: Generator Network Configurations

Architecture	No. Layer	Activation	No. ResBlock	No. DenseBlock	No. Epoch Sup	No. Epoch Aug	Hypothesis
MLP-Deep	6	ELU	2	0	391	55	✓
CNN-ResNet	6	ReLU	2	0	215	41	✓
CNN-DenseNet	6	ReLU	2	1	163	39	✓
CNN-DenseNet-L1	6	ReLU	2	1	1000	39	✓
CNN-DenseNet-L2	6	ReLU	2	1	155	39	✓
CNN-ResNet-AvgPool	6	ReLU	2	0	109	29	✓

Adversarial regularization is at least as good as sole supervision.

Understanding the Role of Adversarial Regularization in Supervised Learning

- Mitigating Vanishing Gradient
- Asymptotic Iteration Complexity
- Sub-optimality Gap
- Provable Convergence
- Generalization Error (Open Problems)
- Neural Topology Analysis (Open Problems)

Sub-optimality Gap

Supervised Objective

Theorem 4. *In purely supervised learning, the sub-optimality gap at the average over all iterates in a trajectory of T time steps is upper bounded by* $\mathcal{O}\left(\frac{\|\theta(0)-\theta^*\|^2}{2T}\right)$.

Generator: $\kappa(t) = \kappa(\theta(t)) := l(\theta(t)) - l(\theta^*)$

Sub-optimality Gap

Augmented Objective

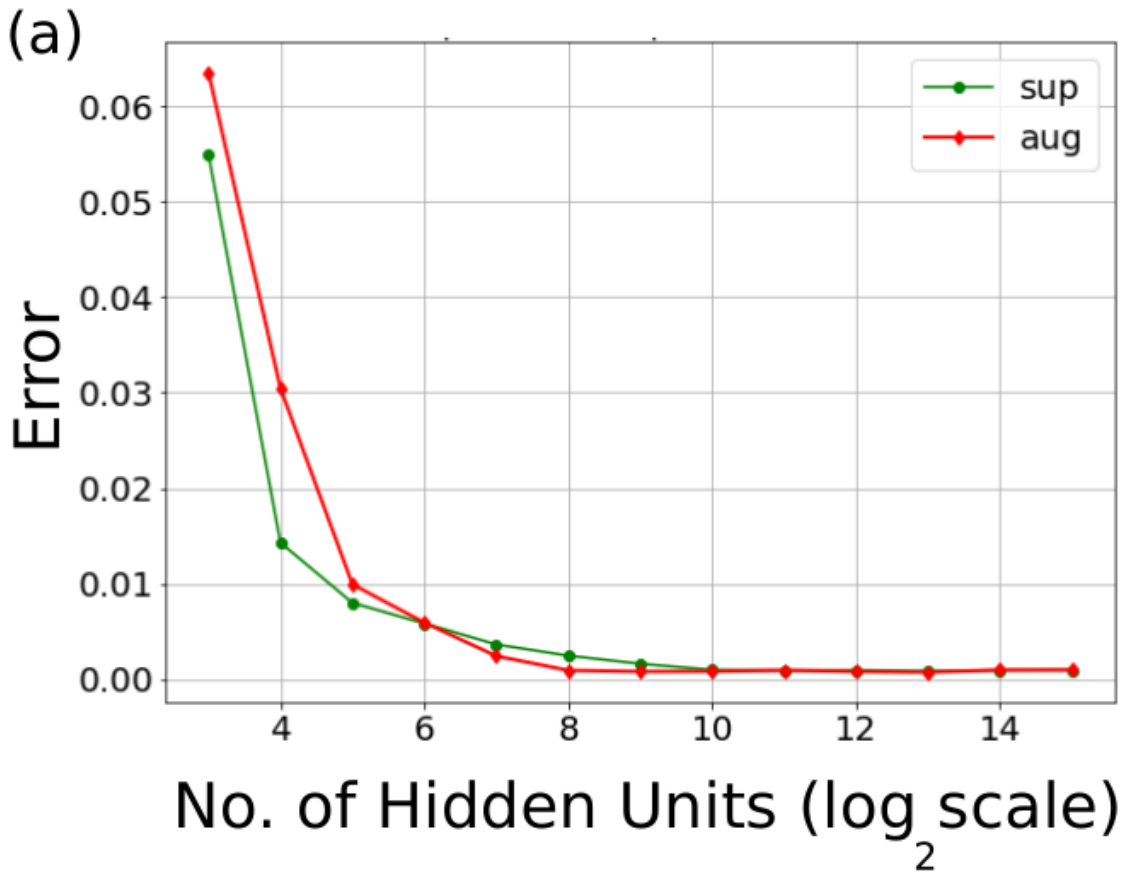
Theorem 5. *In supervised learning with adversarial regularization, the sub-optimality gap at the average over all iterates in a trajectory of T time steps is upper bounded by*

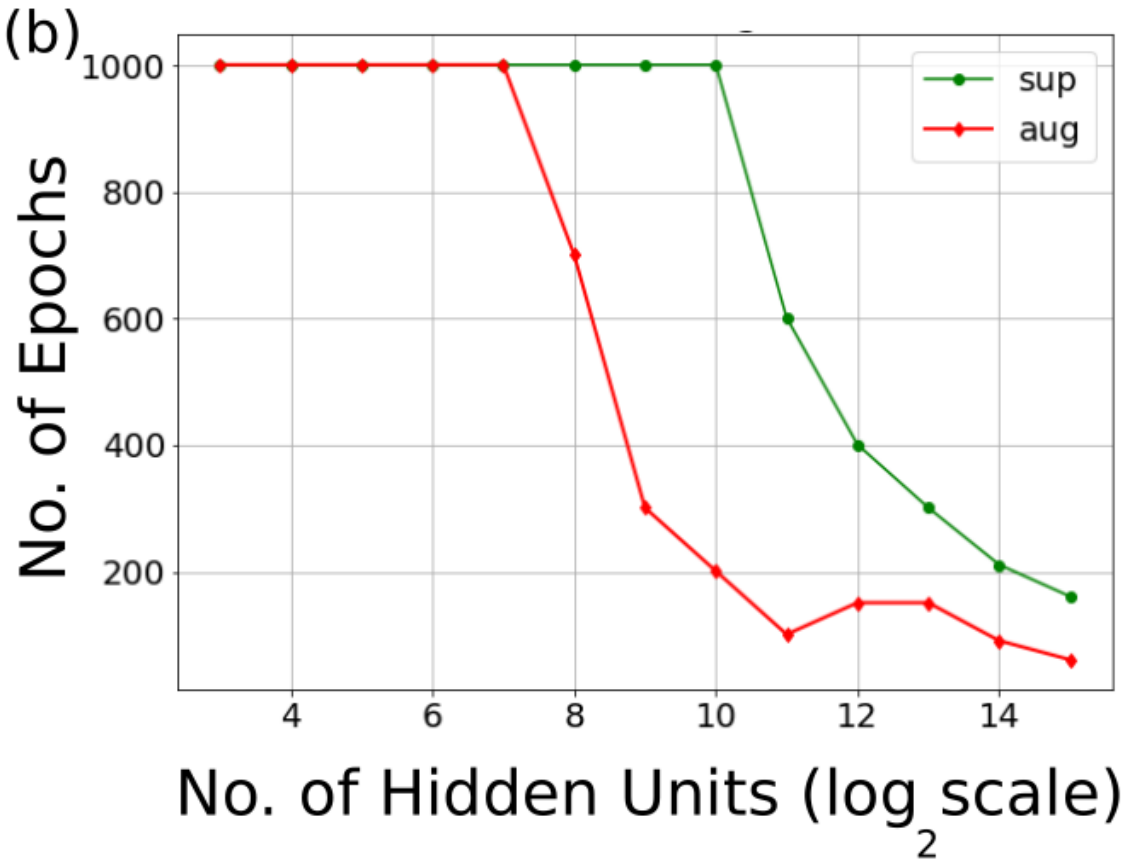
$$\mathcal{O}\left(\frac{\|\theta(0)-\theta^*\|^2}{2T}-\pi\left(\frac{1}{T}\int_0^T\theta(t)dt\right)\right).$$

Discriminator: $\pi(t) = \pi(\theta(t)) := g(\theta^*) - g(\theta(t))$

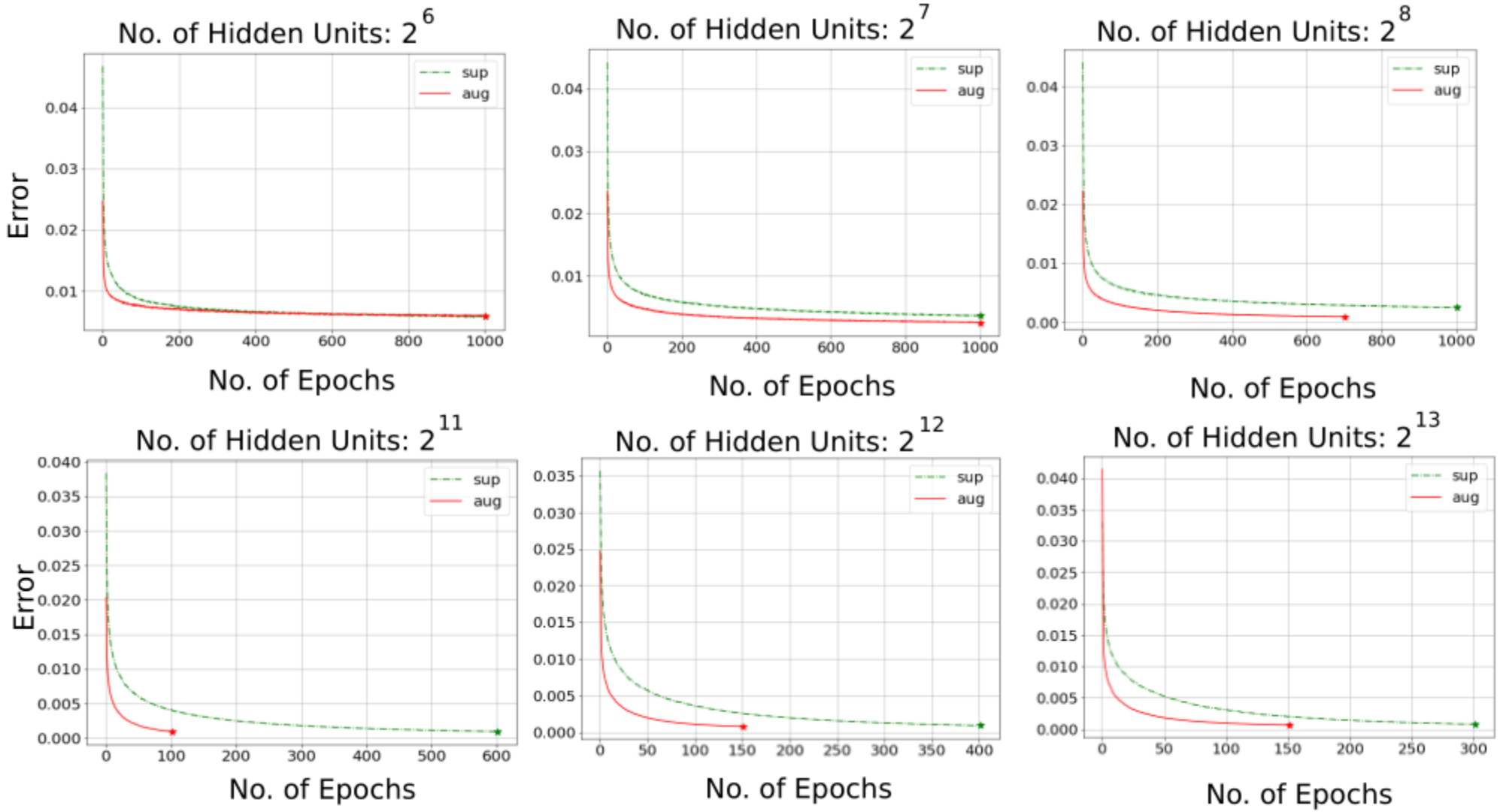
Sub-optimality Gap

Experimental Result

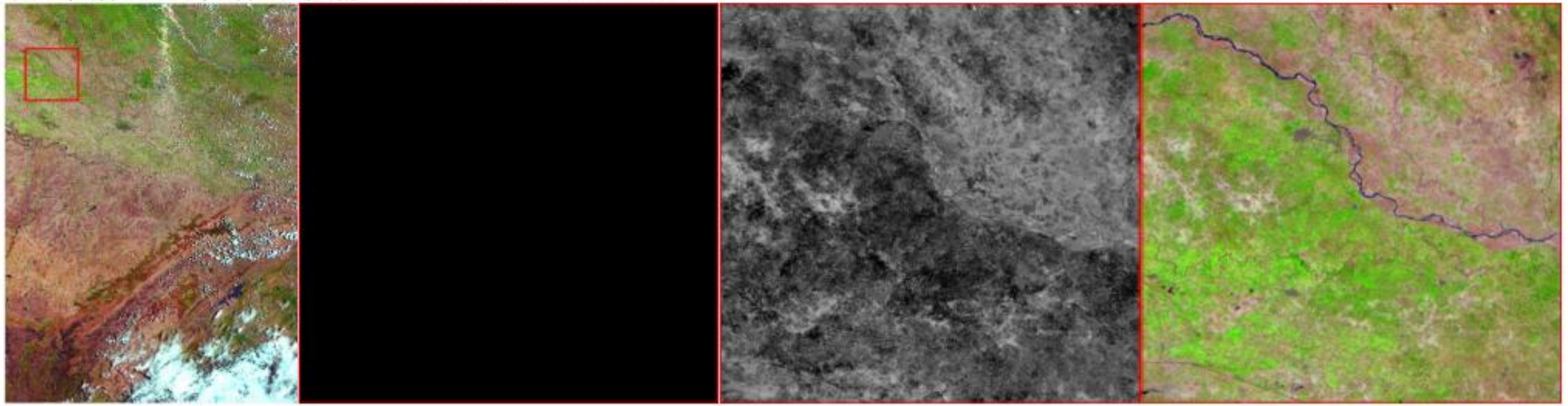




Sub-optimality Gap



Sub-optimality Gap



Metrics	Parameters	Training epochs	Training RMSE	Inference Time
AeroGAN	2 M	100	41.57	64 s
DeepAWiFS	4.7 M	1000	18.2	495 s
DSen2	4.7 M	1000	21.2	456 s
ALERT	2.8 M	50	10.3	70 s
% gain	40.42	95.00	51.41	84.64

Understanding the Role of Adversarial Regularization in Supervised Learning

- Mitigating Vanishing Gradient
- Asymptotic Iteration Complexity
- Sub-optimality Gap
- Provable Convergence
- Generalization Error (Open Problems)
- Neural Topology Analysis (Open Problems)

Provable Convergence

Augmented Objective

Theorem 6 (Strongly-convex-strongly-concave convergence). *Suppose* ***Assumption 7*** *holds. Let* $l(\theta_k) \triangleq l(\theta_k) - g(\theta_k)$ *is a* μ*-strongly convex function. Let* $\{\theta_k\}$ *be the sequence of iterates obtained using global clipping on SGD with zero momentum. Define the output to be* k*-weighted combination of iterates:* $\bar{\theta} = \frac{\sum_{k=1}^{T} k\theta_{k-1}}{\sum_{k=1}^{T} k}$. *If adaptive clipping* $\tau_k = Gk^{\frac{1}{\alpha}}\mu^{\frac{1}{\alpha}}$ *and step size* $\eta_k = \frac{5}{2\mu(k+1)}$, *then the output iterate* $\bar{\theta}$ *satisfies*

$$\mathbb{E}\left[l\left(\bar{\theta}\right)\right] - l\left(\theta^*\right) \leq \mathcal{O}\left(G^2\left(\mu\left(T+1\right)\right)^{\frac{2-2\alpha}{\alpha}} - \left(g\left(\theta^*\right) - \mathbb{E}\left[g\left(\bar{\theta}\right)\right]\right)\right).$$

Provable Convergence

Augmented Objective

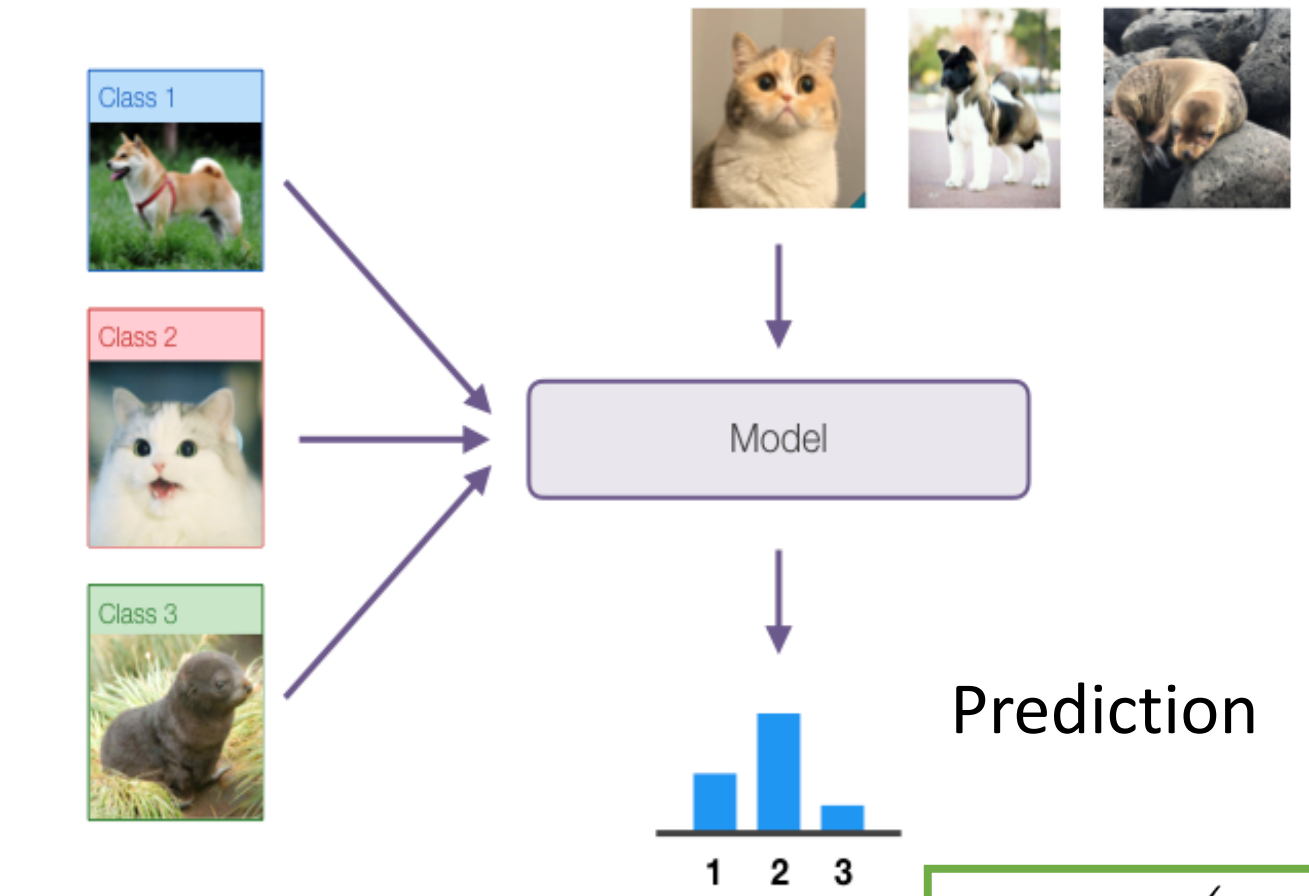
Theorem 7 (Nonconvex-nonconcave convergence). *Suppose* ***Assumption 7*** *holds. Let* $l(\theta_k) \triangleq l(\theta_k) - g(\theta_k)$ *is a possible L-smooth function and* $\{\theta_k\}$ *be the sequence of iterates obtained using global clipping on SGD with zero momentum. Given constant clipping* $\tau_k = G(\eta_k L)^{\frac{-1}{\alpha}}$ *and constant step size* $\eta_k = \left(\frac{R_0^\alpha L^{2-2\alpha}}{G^2 T^\alpha}\right)^{\frac{1}{3\alpha-2}}$, *where* $R_0 = l(\theta_0) - l(\theta^*)$, *the sequence* $\{\theta_k\}$ *satisfies*

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\|\nabla l(\theta_{k-1})\|^2\right] \leq \mathcal{O}\left(G^{\frac{2\alpha}{3\alpha-2}}\left(\frac{R_0 L}{T}\right)^{\frac{2\alpha-2}{3\alpha-2}} - \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\|\nabla g(\theta_{k-1})\|^2\right]\right).$$

Understanding the Role of Adversarial Regularization in Supervised Learning

- Mitigating Vanishing Gradient
- Asymptotic Iteration Complexity
- Sub-optimality Gap
- Provable Convergence
- Generalization Error (Open Problems)
- Neural Topology Analysis (Open Problems)

Generalization Error



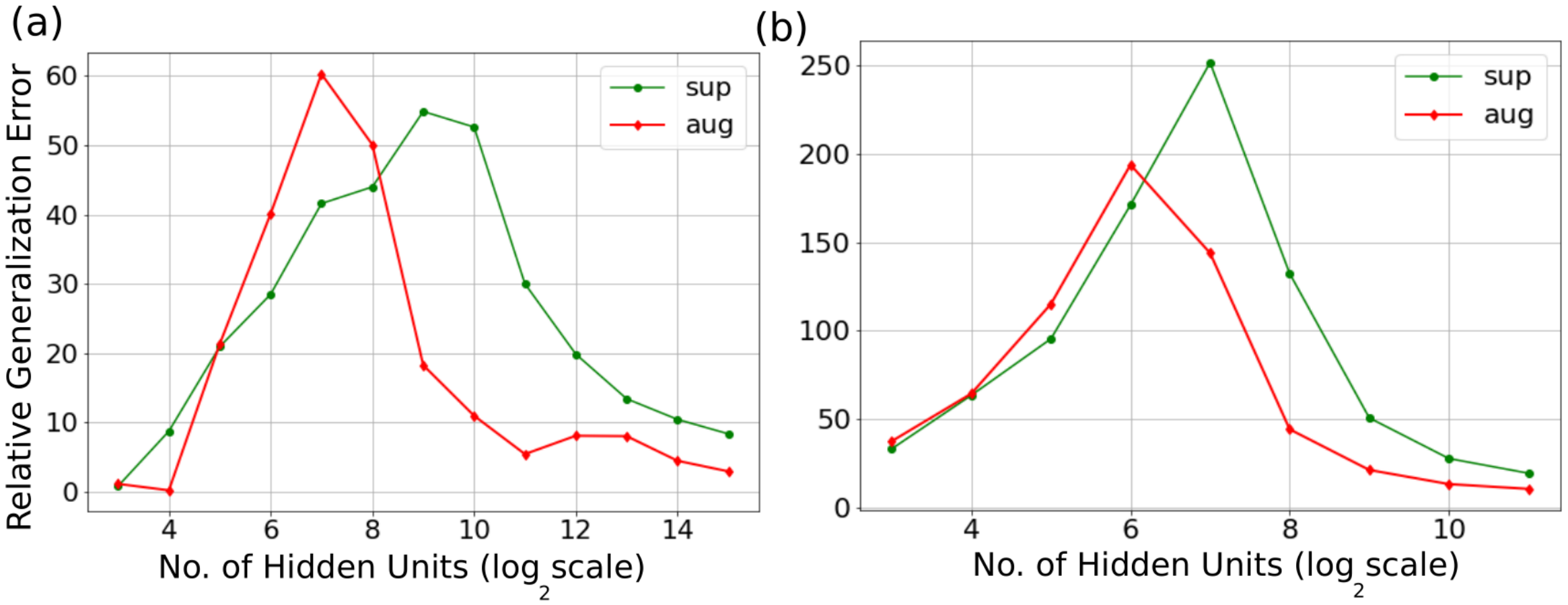
Testing Data

Prediction

Training Data

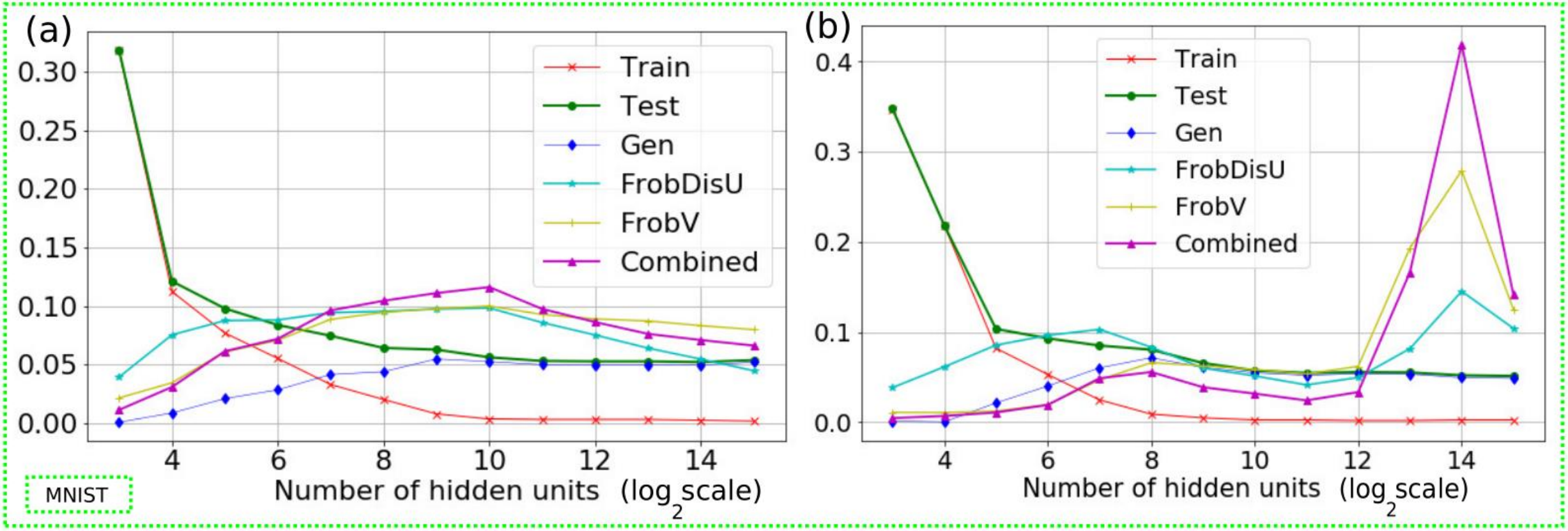
$$e_{gen} = \left(\mathbb{E}_{(x,y)\sim\mathcal{D}} [l \; o \; f] - \frac{1}{m} \sum_{i=1}^{m} l\left(f(x); y\right) \right)$$

Generalization Error



$$e_{gen,r} = \left(\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[l \; o \; f \right] - \frac{1}{m} \sum_{i=1}^{m} l\left(f(x); y \right) \right) \times \mathcal{N}$$

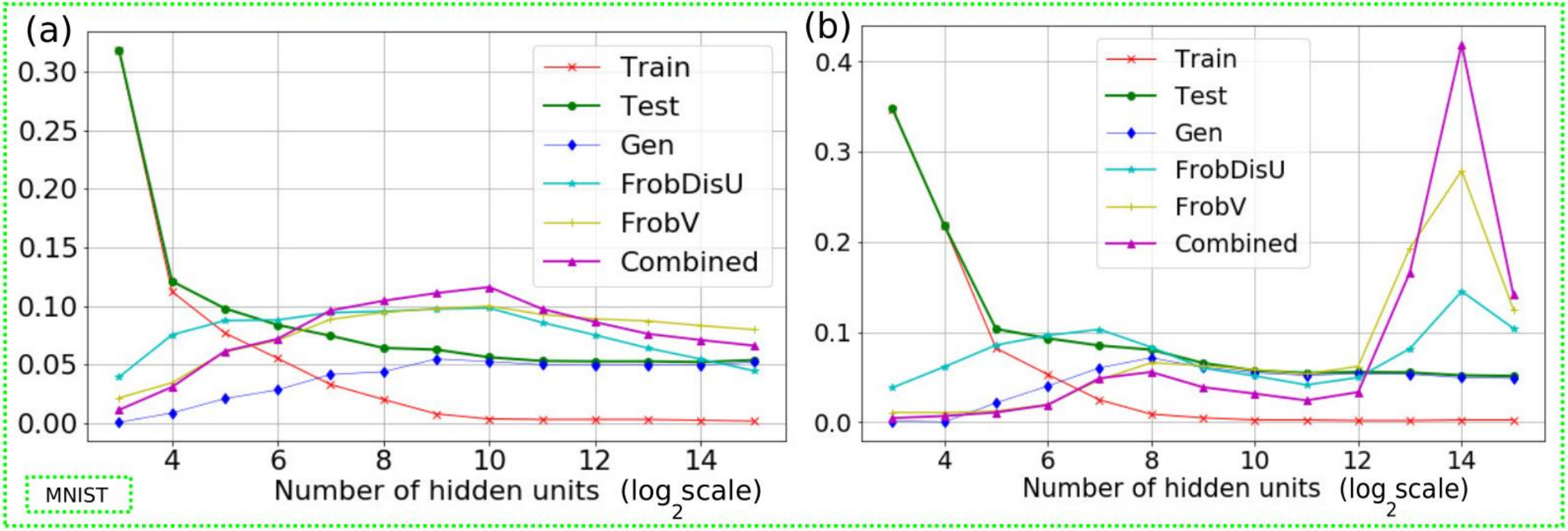
Generalization Error



Combined Metric: $\mathcal{O}\left(\left\|U^0\right\|_2 \|V\|_F + \left\|U - U^0\right\|_F \|V\|_F + \sqrt{h}\right)$

Neyshabur et al. ICLR ,2019

Generalization Error

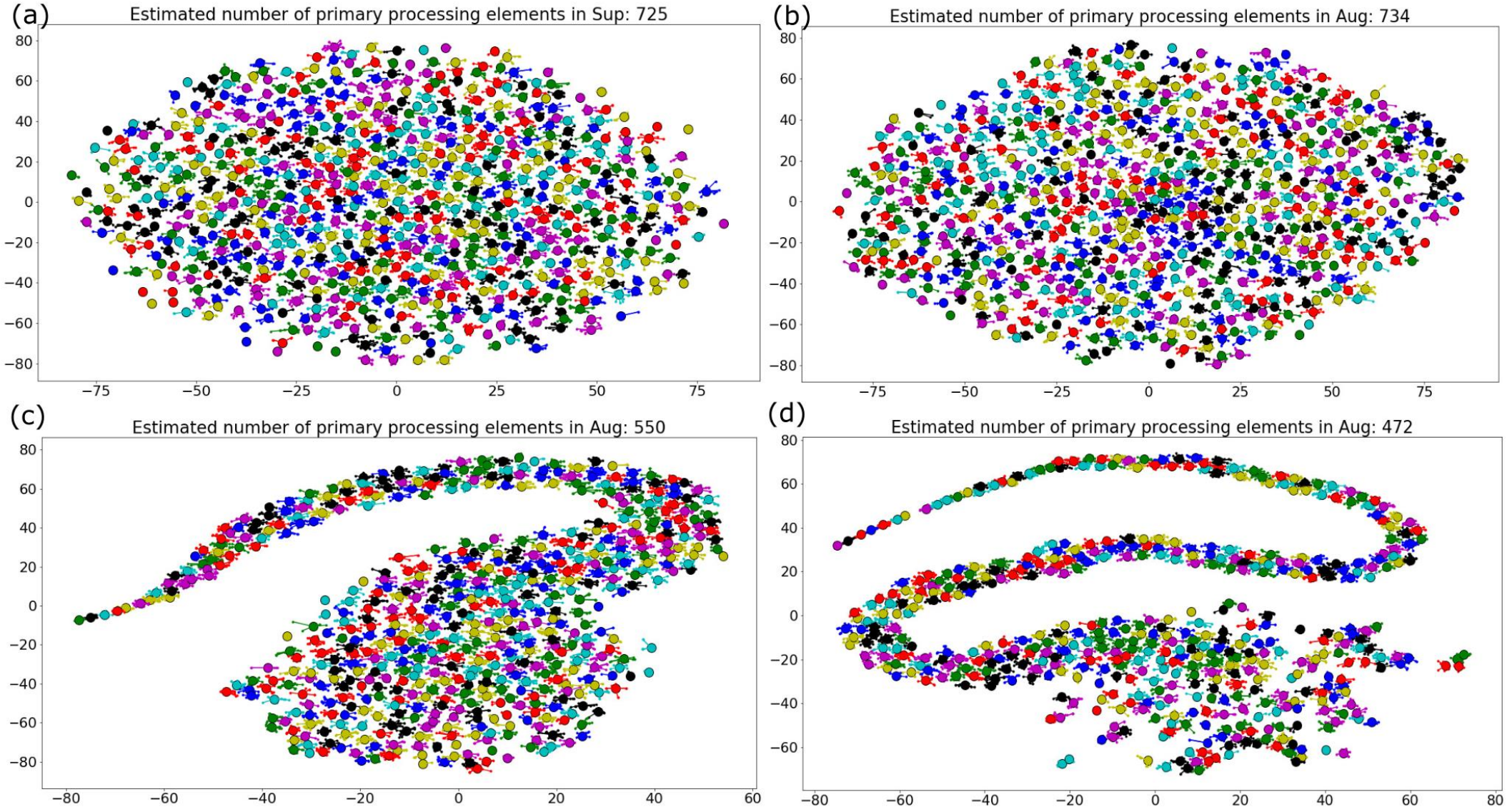


Open Problem: How to explain the generalization behaviour of the augmented objective?

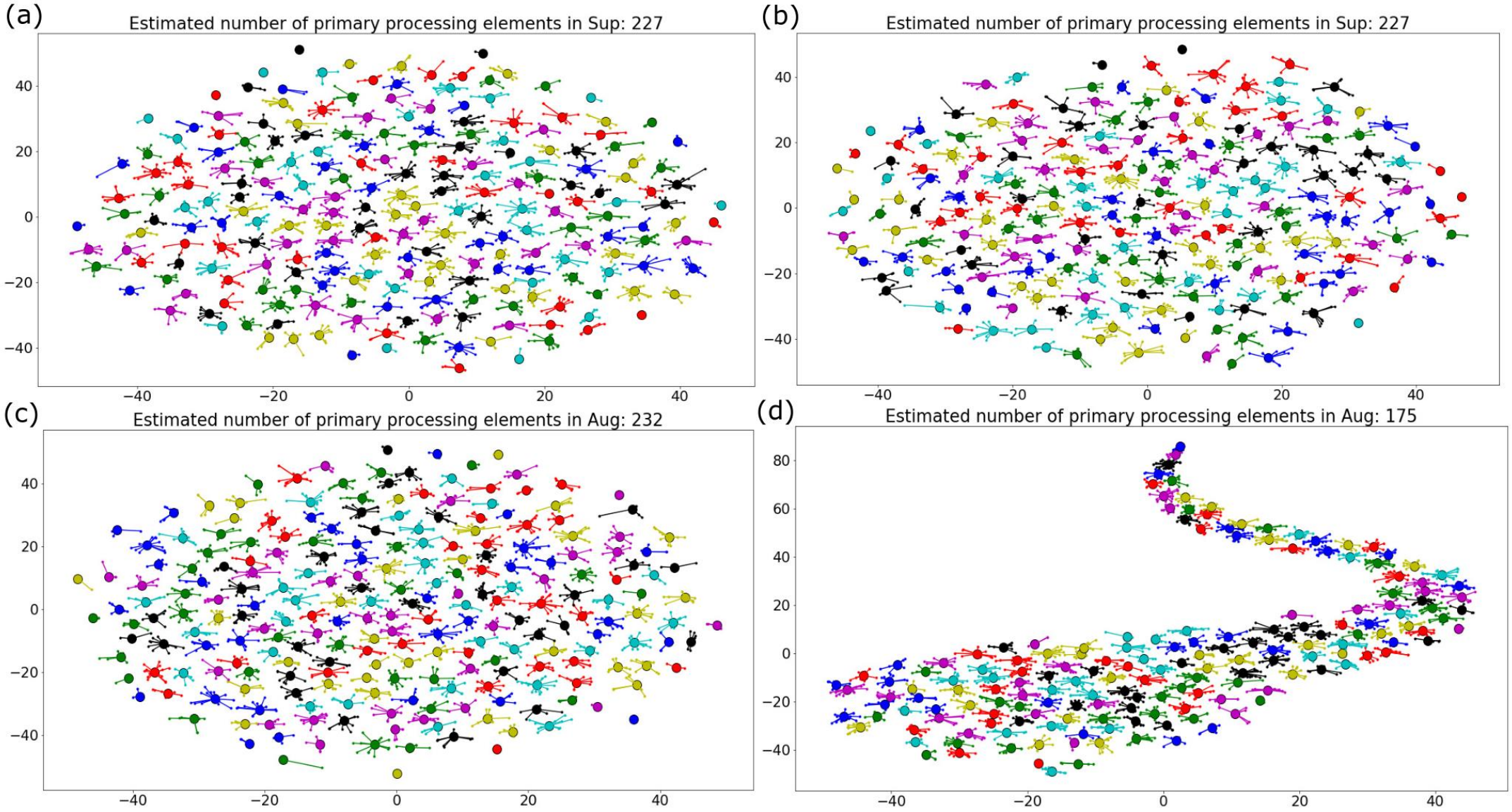
Understanding the Role of Adversarial Regularization in Supervised Learning

- Mitigating Vanishing Gradient
- Asymptotic Iteration Complexity
- Sub-optimality Gap
- Provable Convergence
- Generalization Error (Open Problems)
- Neural Topology Analysis (Open Problems)

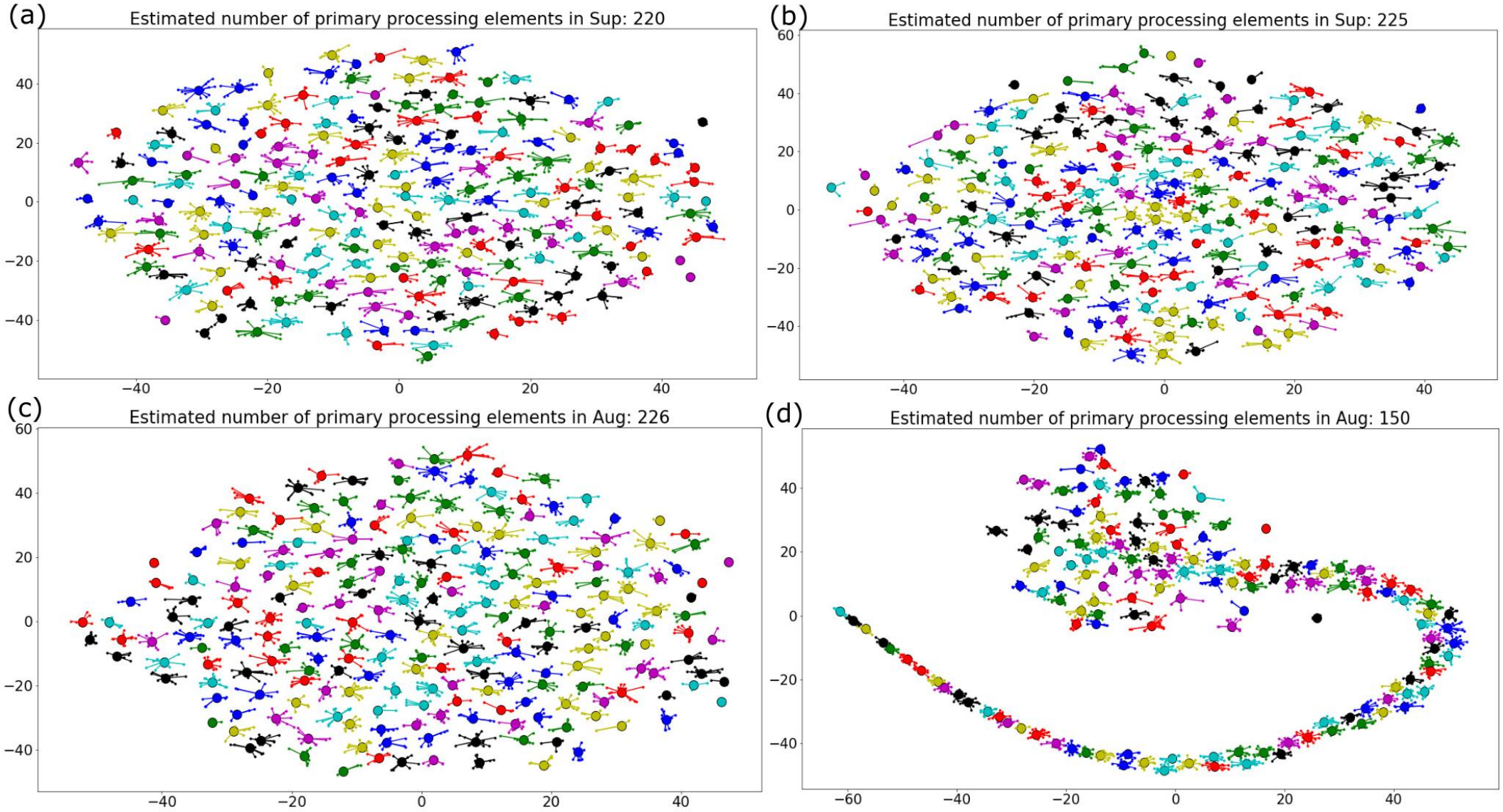
Neural Topology Analysis



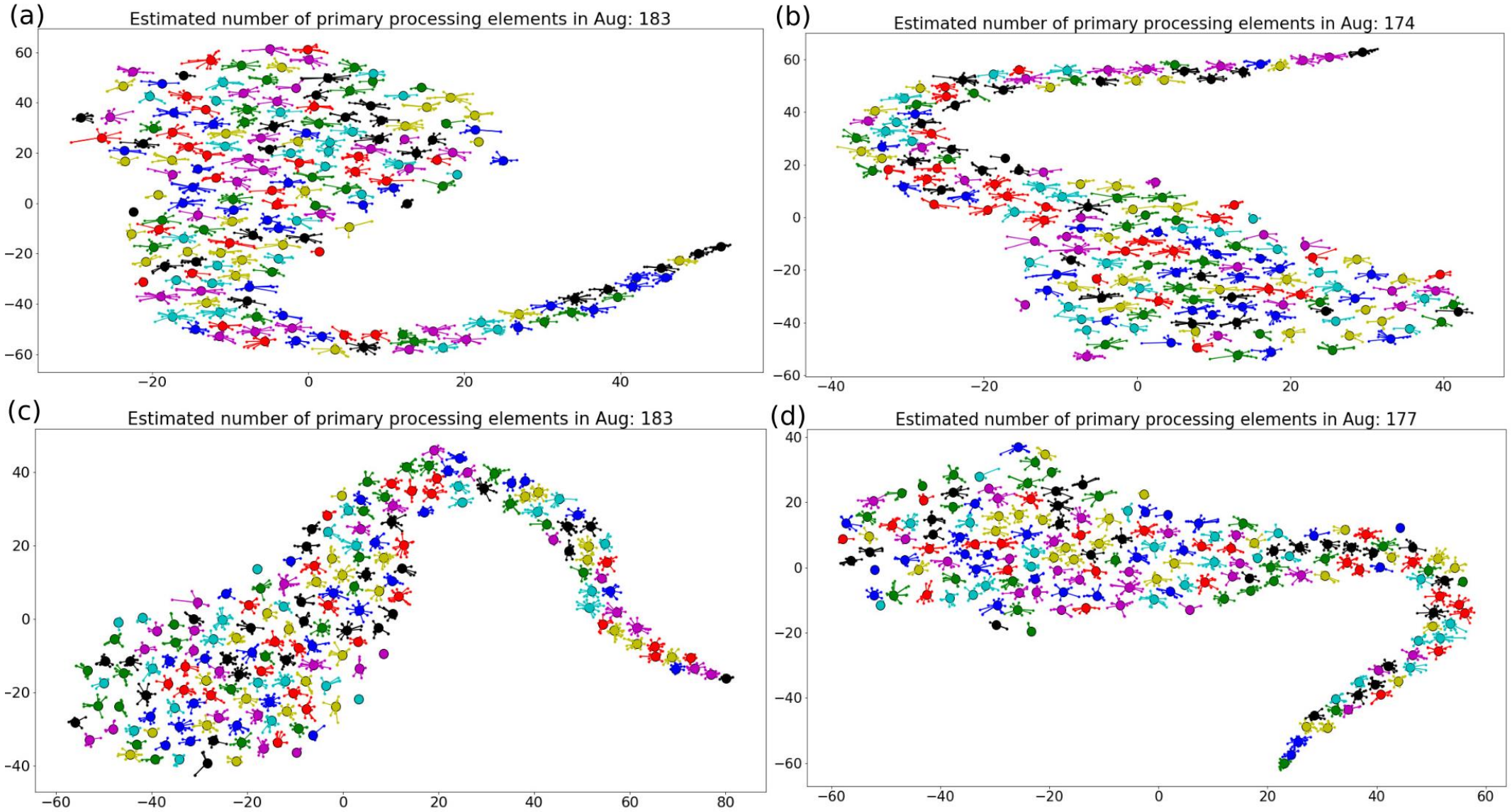
Neural Topology Analysis: Hidden Layer



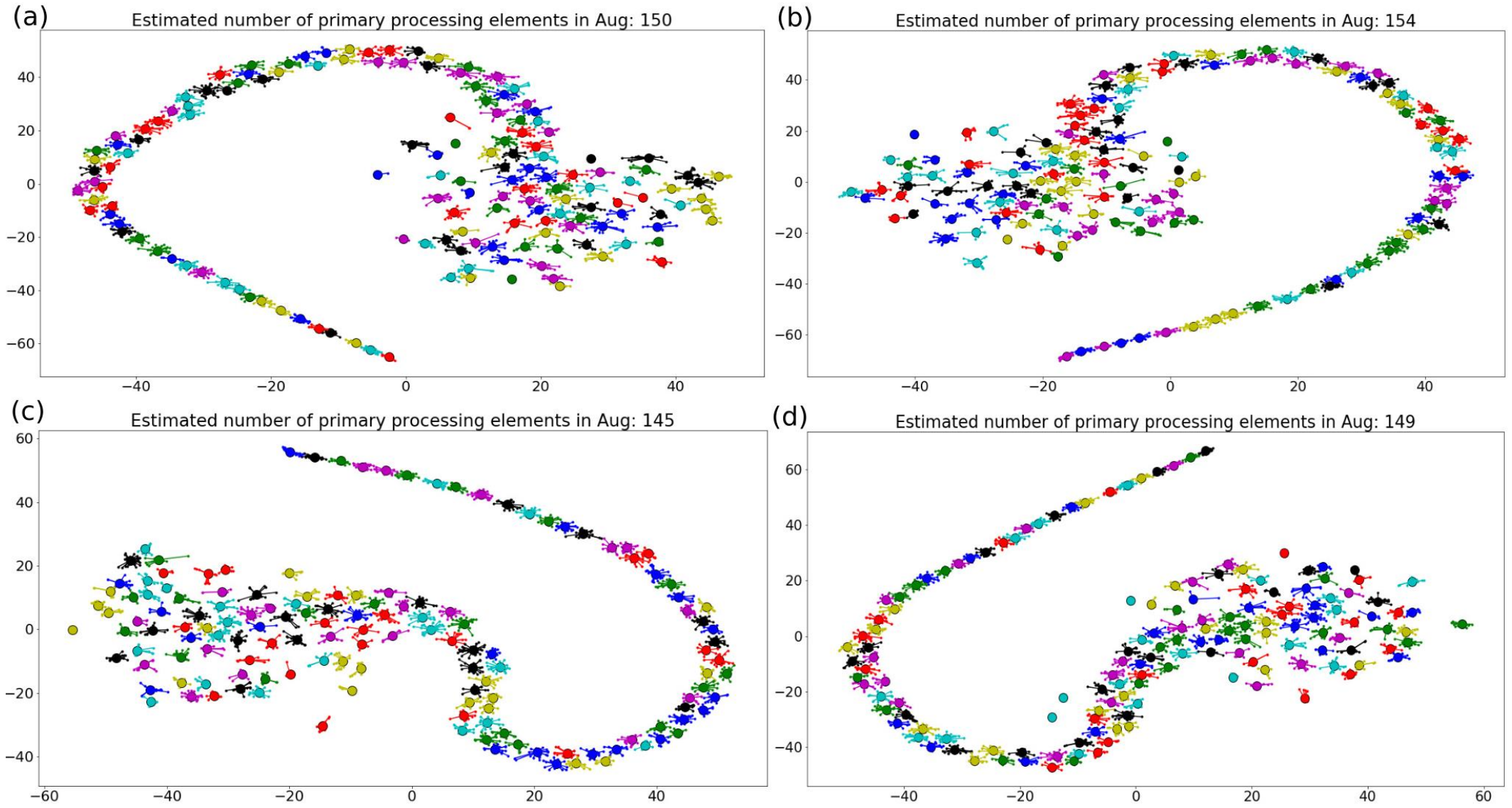
Neural Topology Analysis: Top Layer



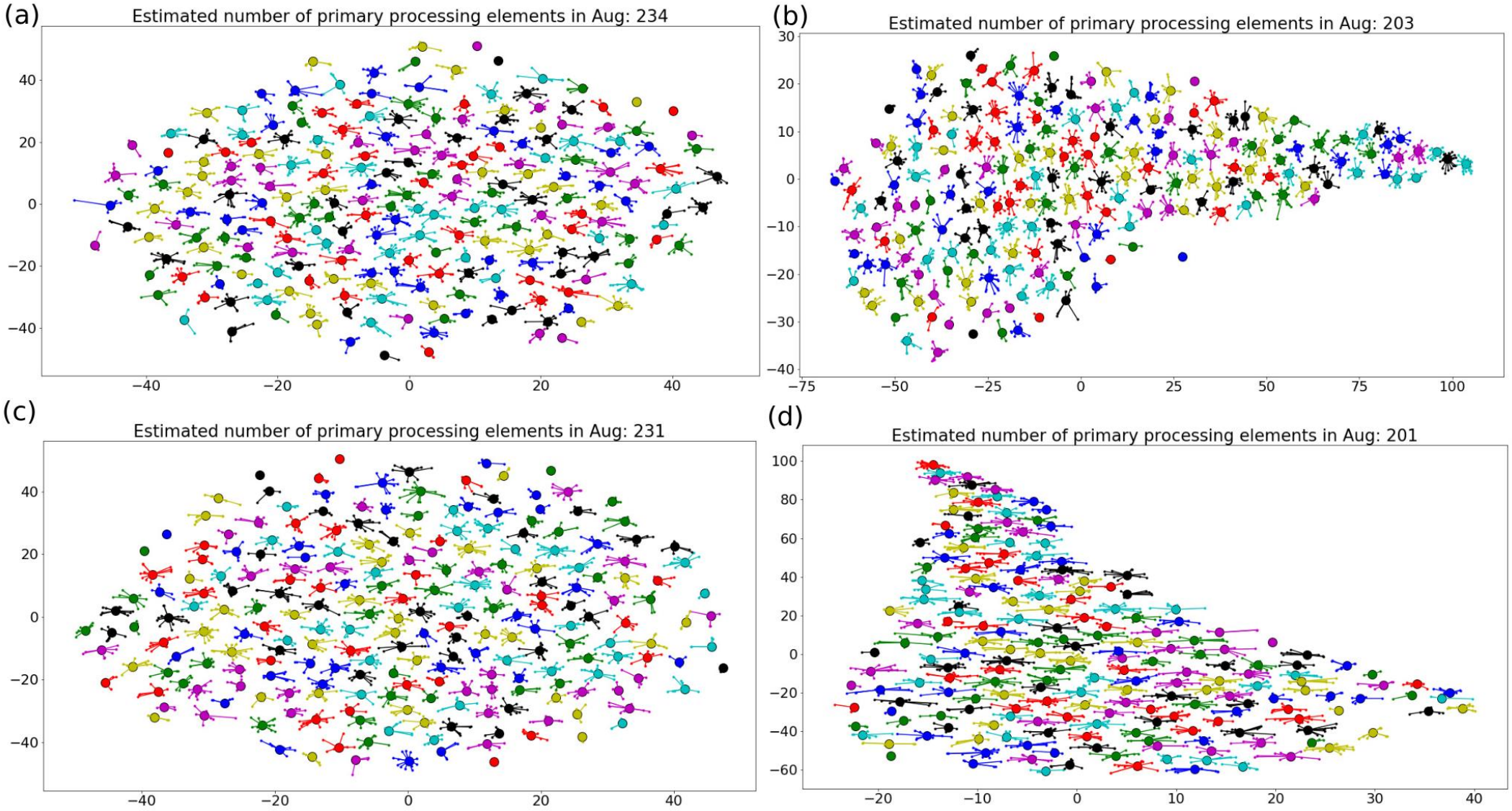
Neural Topology Analysis: Hidden Layer Subsets



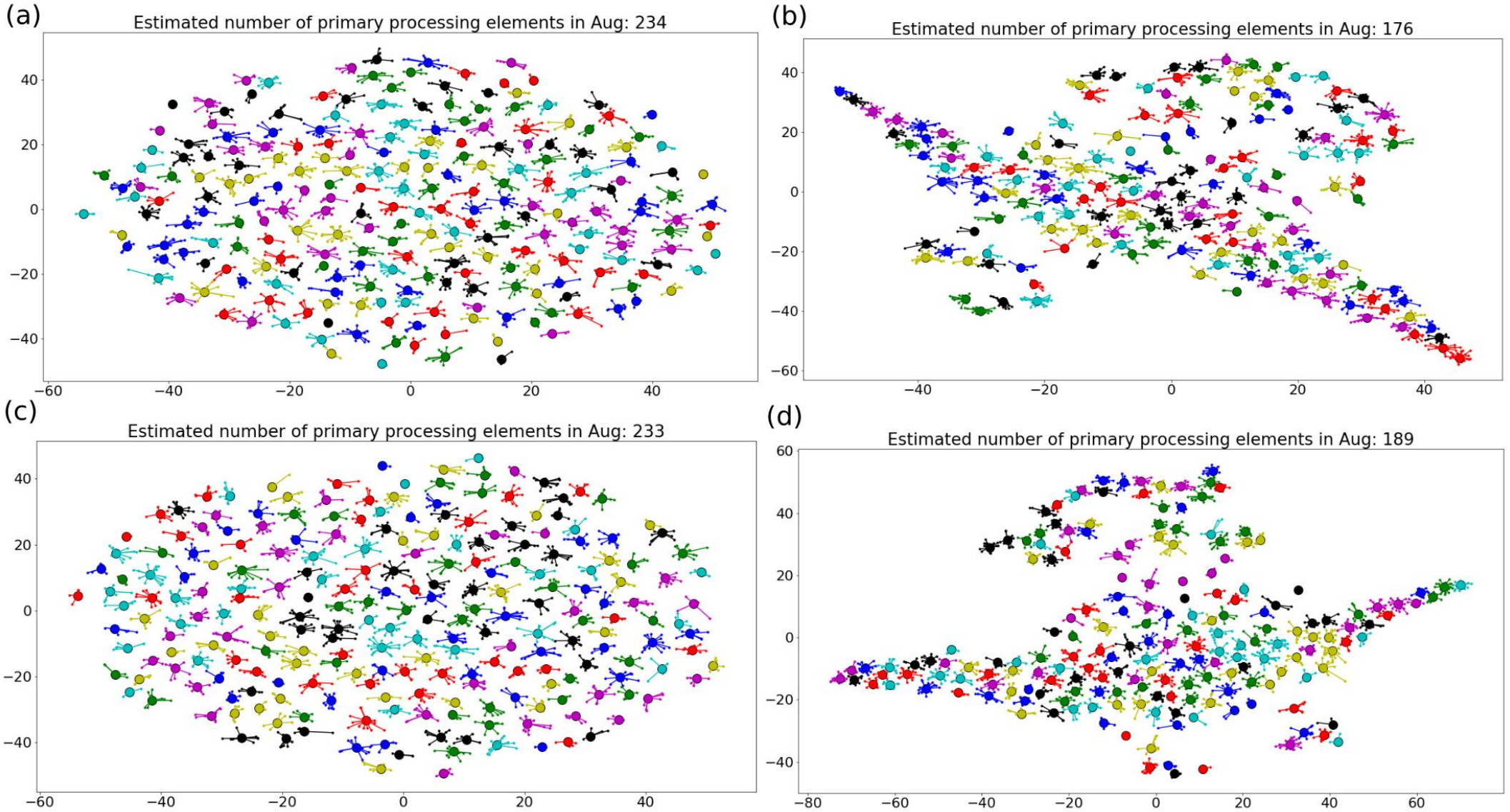
Neural Topology Analysis: Top Layer Subsets



Neural Topology Analysis: Hidden Layer FMNIST



Neural Topology Analysis: Top Layer FMNIST



Interesting Problems

- Why does adversarial interaction create non-homogeneous patterns?
- How do we predict the shape of the non-homogeneous patterns?
- What happens to the parameters if the supervised cost is augmented in the discriminator objective?
- Is gradient descent or backpropagation the optimal learning algorithm?

Interesting Problems

- Why does adversarial interaction create non-homogeneous patterns?
- How do we predict the shape of the non-homogeneous patterns?
- What happens to the parameters if the supervised cost is augmented in the discriminator objective?
- Is gradient descent or backpropagation the optimal learning algorithm?

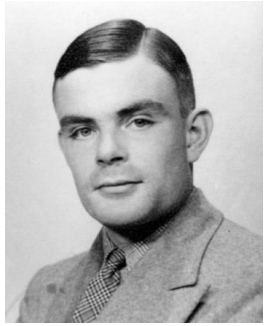
Why Adversarial Interaction Creates Non-Homogeneous Patterns: A Pseudo-Reaction Diffusion Model for Turing Instability

- Adversarial Interaction
  - Generative Adversarial Networks (GANs)
  - Application of conditional GANs
- Non-Homogeneous Patterns
  - Homogeneous patterns
  - Supervised learning
- Reaction-Diffusion
  - Turing's RD model (1952)
  - Gray-Scott RD model (1984)
- Turing Instability
  - Reaction dynamics
  - Diffusion dynamics

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

- Adversarial Interaction
  - Generative Adversarial Networks (GANs)
  - Application of conditional GANs
- Non-Homogeneous Patterns
  - Homogeneous patterns
  - Supervised learning
- Reaction-Diffusion
  - Turing's RD model (1952)
  - Gray-Scott RD model (1984)
- Turing Instability
  - Reaction dynamics
  - Diffusion dynamics

Turing's Reaction-Diffusion Model



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

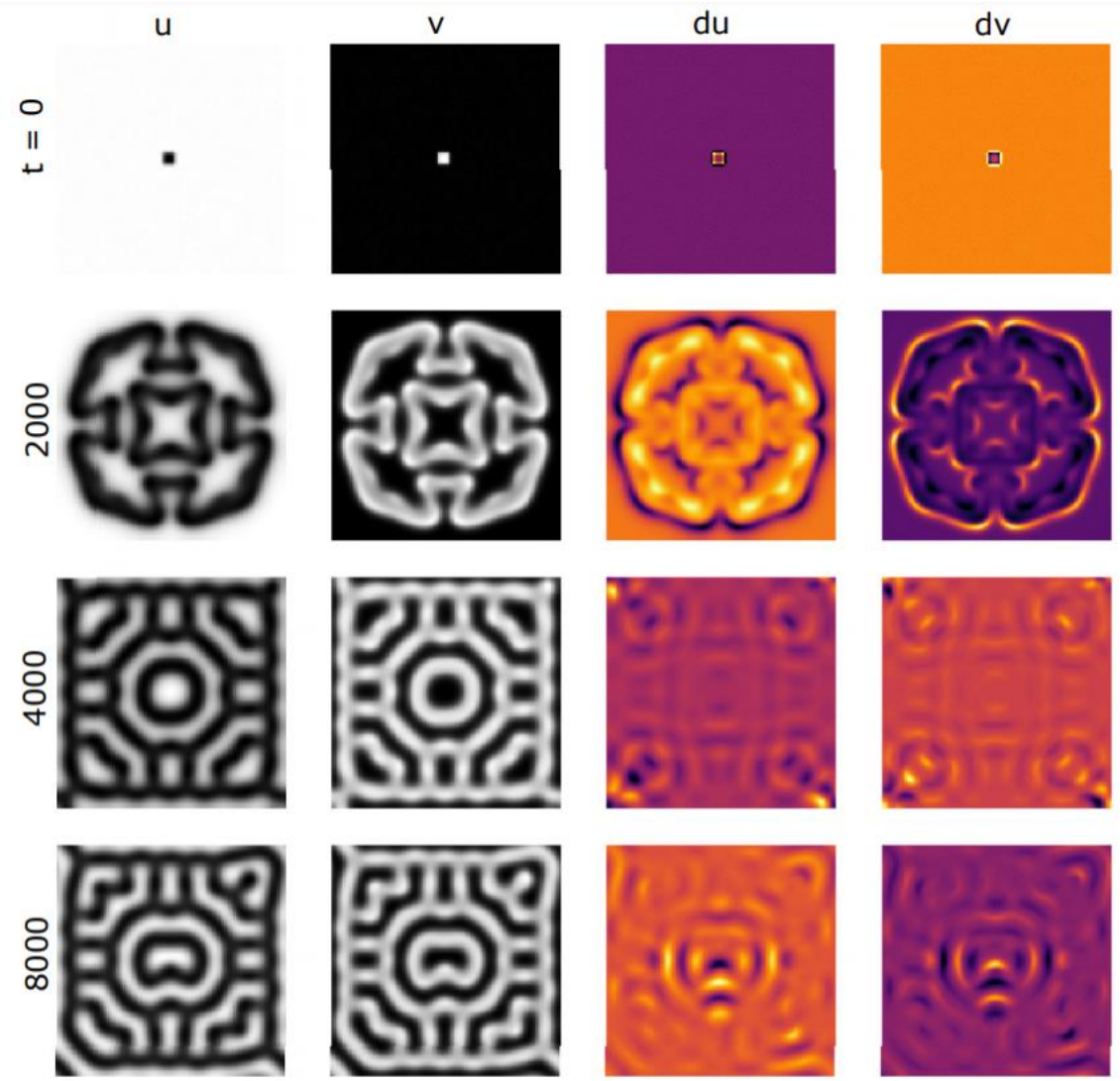
$$\frac{d\boldsymbol{v}_j}{dt} = \mathfrak{R}_j^{\boldsymbol{v}}\left(\boldsymbol{u}_j, \boldsymbol{v}_j\right) + \mathfrak{D}_j^{\boldsymbol{v}}\left(\nabla^2 \boldsymbol{v}_j\right)$$

A. M. Turing, 1952

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

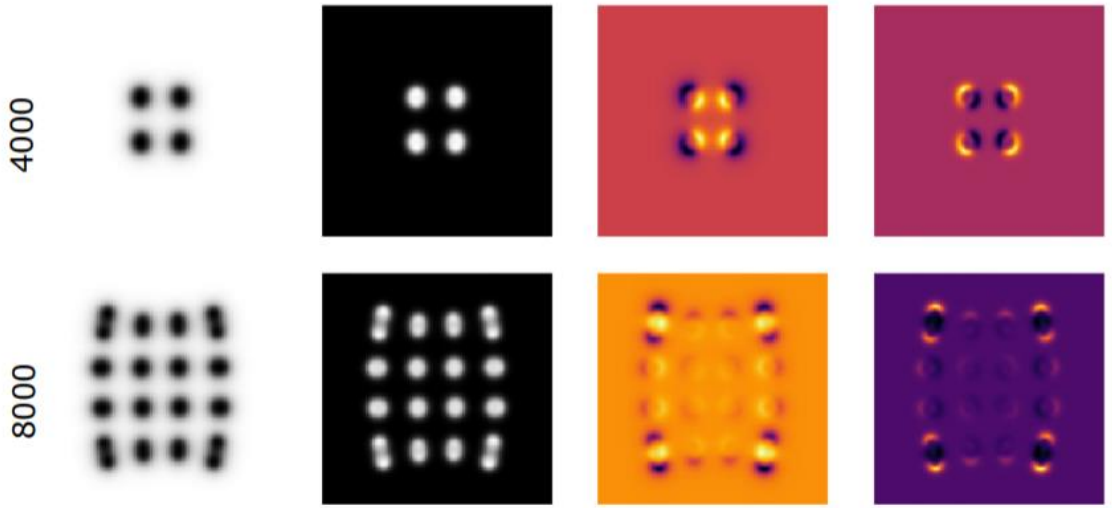
- Adversarial Interaction
  - Generative Adversarial Networks (GANs)
  - Application of conditional GANs
- Non-Homogeneous Patterns
  - Homogeneous patterns
  - Supervised learning
- Reaction-Diffusion
  - Turing's RD model (1952)
  - Gray-Scott RD model (1984)
- Turing Instability
  - Reaction dynamics
  - Diffusion dynamics

Gray-Scott Reaction-Diffusion Model



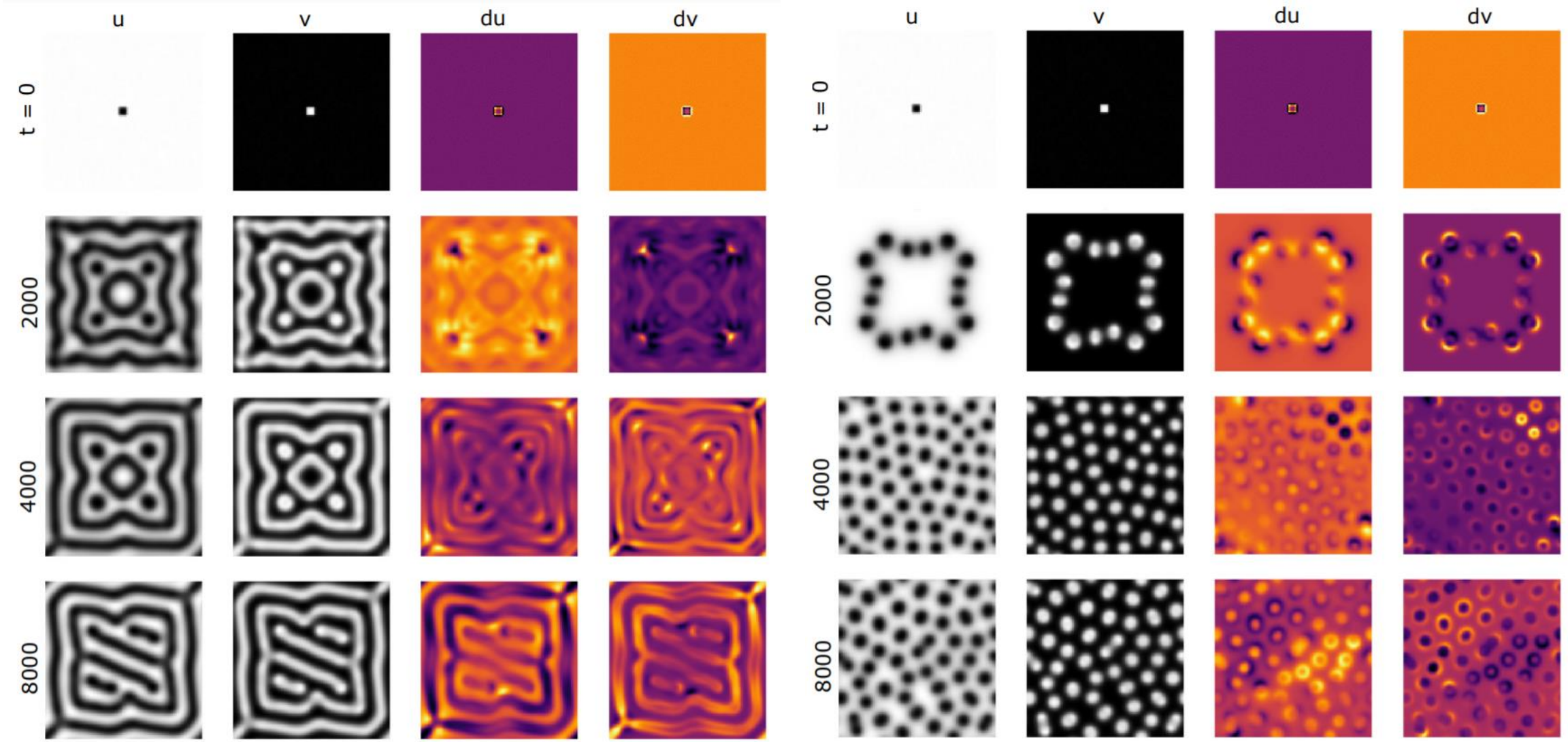
$$\frac{\partial u}{\partial t} = F(1-u) - uv^2 + \mu'\nabla^2 u$$

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



Gray and Scott, 1984

Gray-Scott Reaction-Diffusion Model



Gray and Scott, 1984

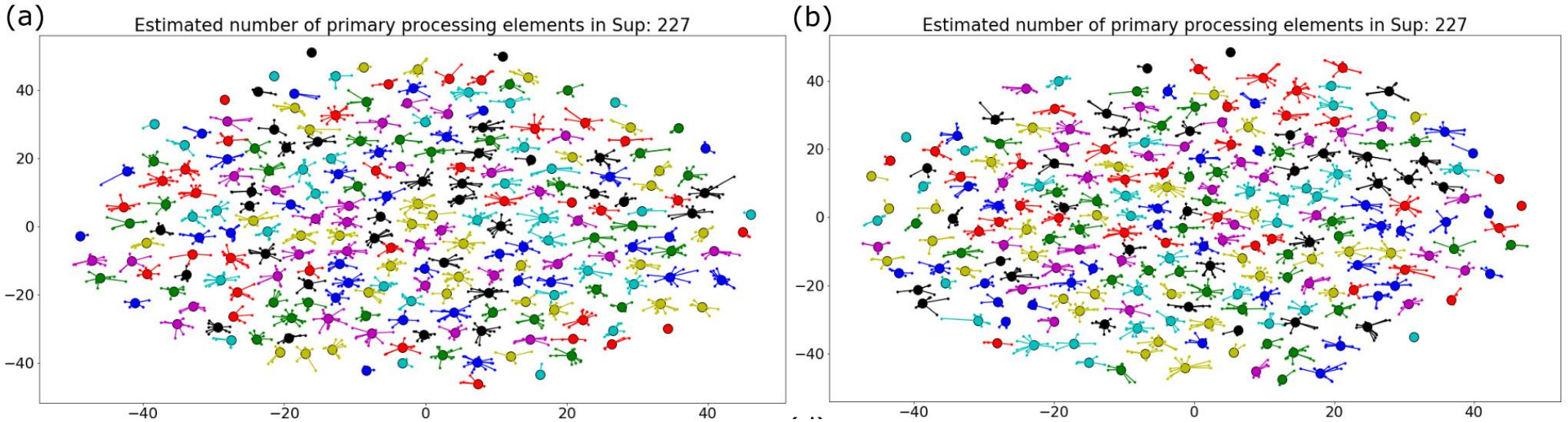
Why Adversarial Interaction Creates Non-Homogeneous Patterns: A Pseudo-Reaction Diffusion Model for Turing Instability

- Adversarial Interaction
  - Generative Adversarial Networks (GANs)
  - Application of conditional GANs
- Non-Homogeneous Patterns
  - Homogeneous patterns
  - Supervised learning
- Reaction-Diffusion
  - Turing's RD model (1952)
  - Gray-Scott RD model (1984)
- Turing Instability
  - Reaction dynamics
  - Diffusion dynamics

Symmetry and Homogeneity

Supervised Learning

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



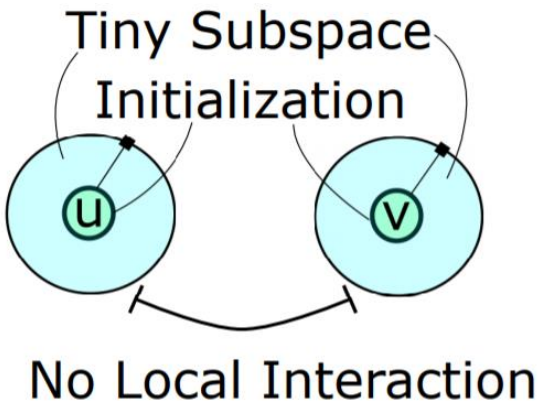
Symmetry and Homogeneity

Theorem (Symmetry and Homogeneity)

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

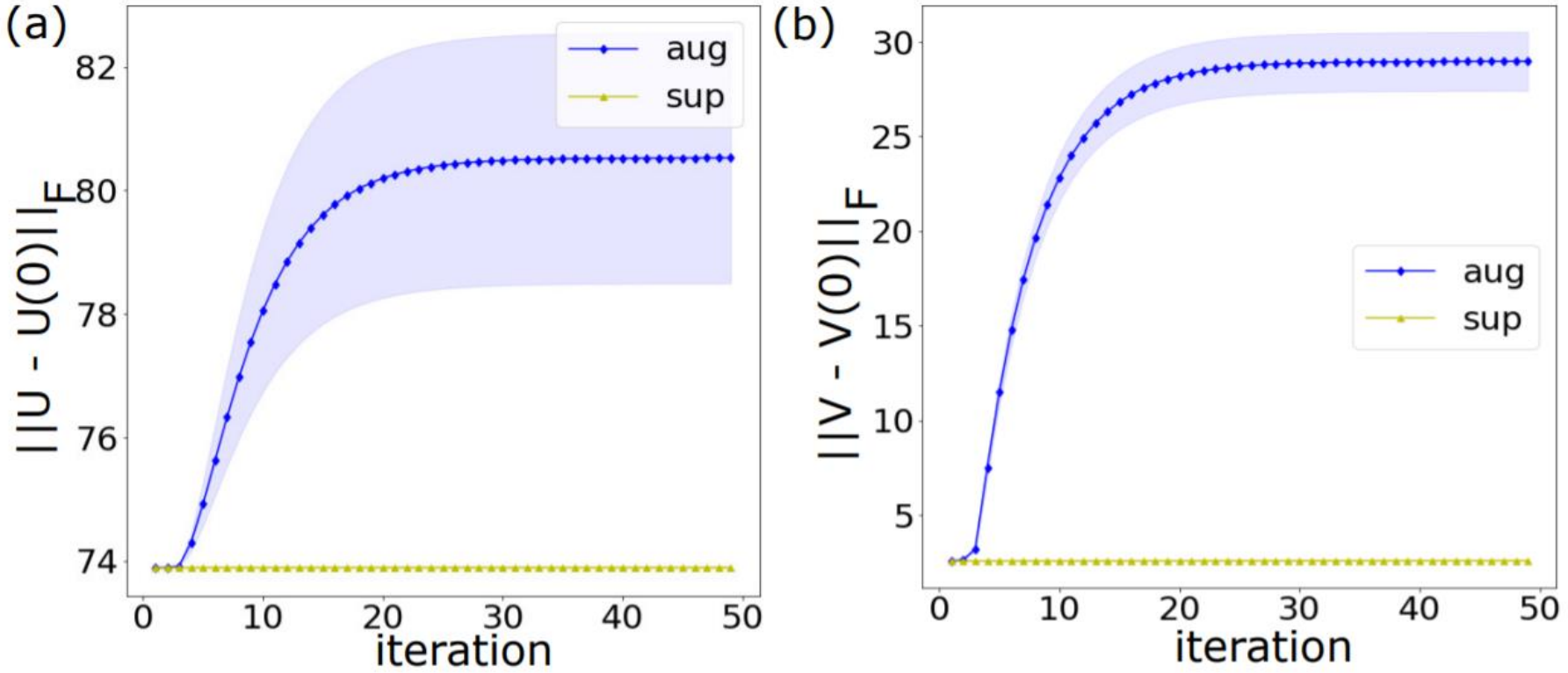
$$\|\boldsymbol{u}_j(t) - \boldsymbol{u}_j(0)\|_2 \leq \mathcal{O}\left(\frac{n^{3/2}}{\sqrt{m}\lambda_0\delta}\right),$$

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



Symmetry and Homogeneity

Supervised Learning



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

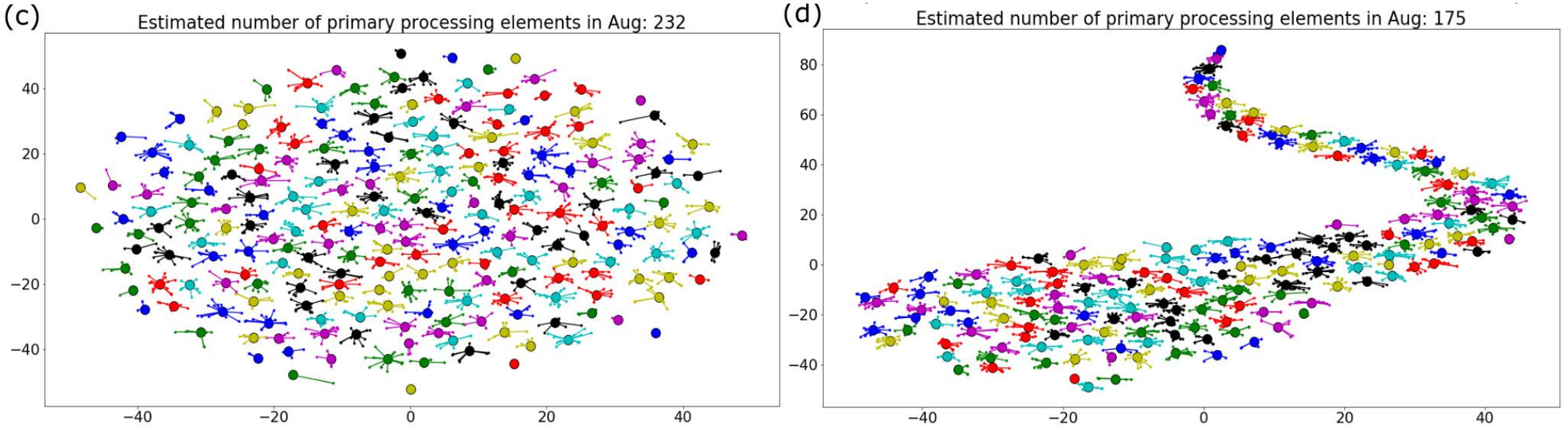
- Adversarial Interaction
  - Generative Adversarial Networks (GANs)
  - Application of conditional GANs
- Non-Homogeneous Patterns
  - Homogeneous patterns
  - Supervised learning
- Reaction-Diffusion
  - Turing's RD model (1952)
  - Gray-Scott RD model (1984)
- Turing Instability
  - Reaction dynamics
  - Diffusion dynamics

Breakdown of Symmetry and Homogeneity

Adversarial Regularization

$$\mathcal{L}_{aug}(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\boldsymbol{a}) = \underbrace{\frac{1}{2}\left\|\frac{1}{\sqrt{d_{out}m}}\boldsymbol{V}\sigma(\boldsymbol{U}\boldsymbol{X})-\boldsymbol{Y}\right\|_F^2}_{\mathcal{L}_{sup}}$$

$$-\frac{1}{m\sqrt{d_{out}}}\sum_{p=1}^{n}\boldsymbol{a}^T\sigma(\boldsymbol{W}\boldsymbol{V}\sigma(\boldsymbol{U}\boldsymbol{x}_p))$$



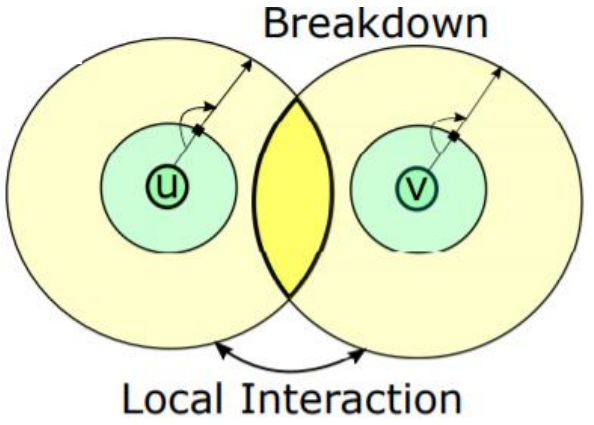
Breakdown of Symmetry and Homogeneity

Theorem (Breakdown of Symmetry and Homogeneity)

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

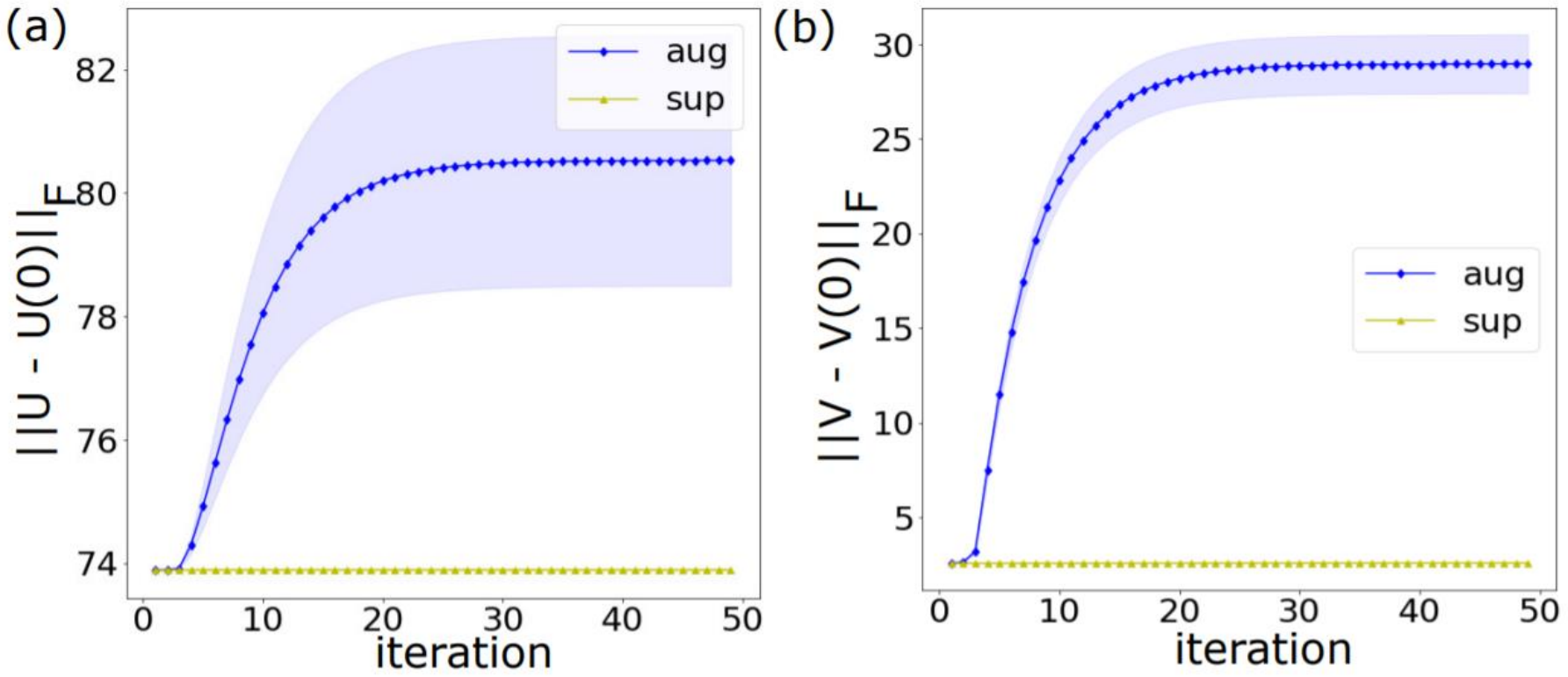
$$\|\boldsymbol{u}_j(t) - \boldsymbol{u}_j(0)\|_2 \leq \mathcal{O}\left(\frac{n^{3/2}}{\sqrt{m}\lambda_0\delta} + \left(\frac{\mu\left(1+\kappa\sqrt{n}\right)}{\sqrt{m}}\right)t\right),$$

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

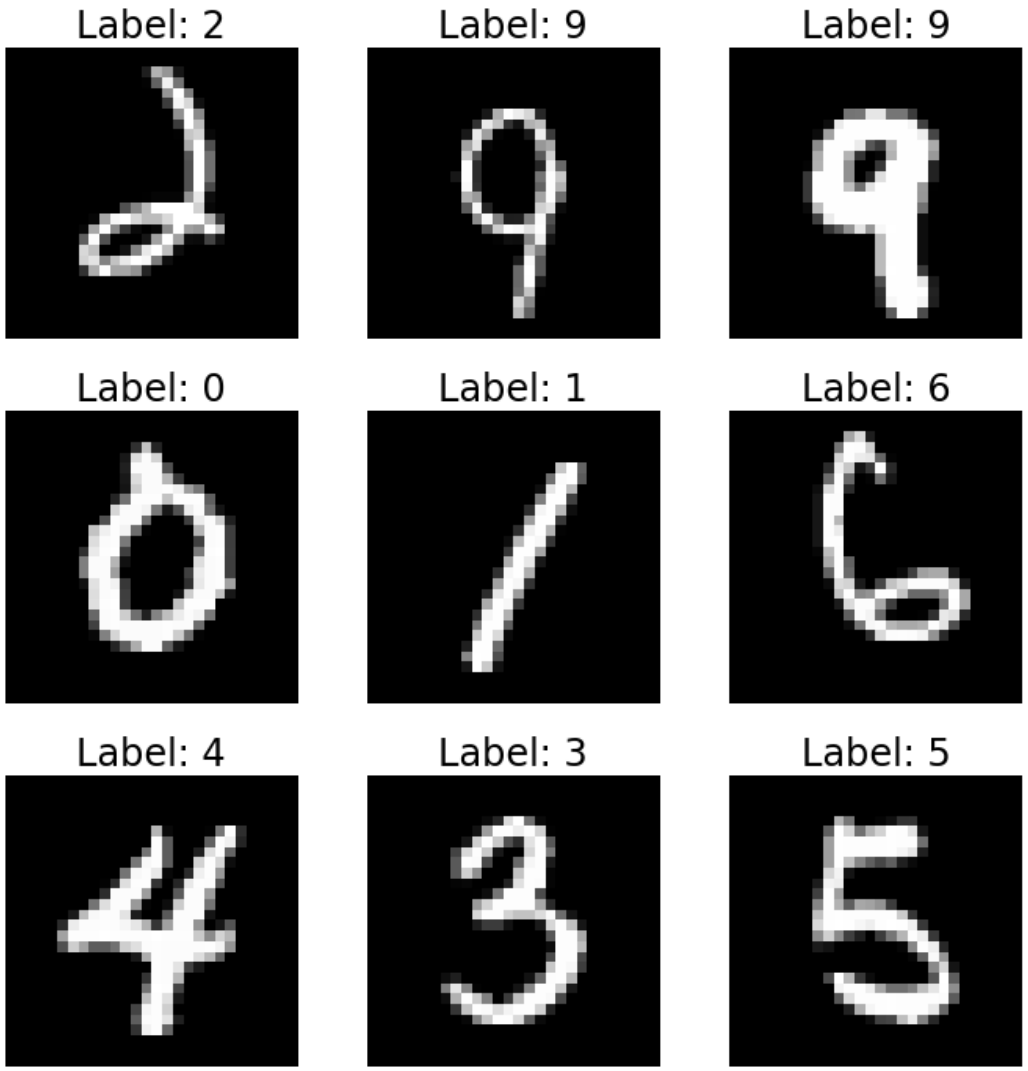


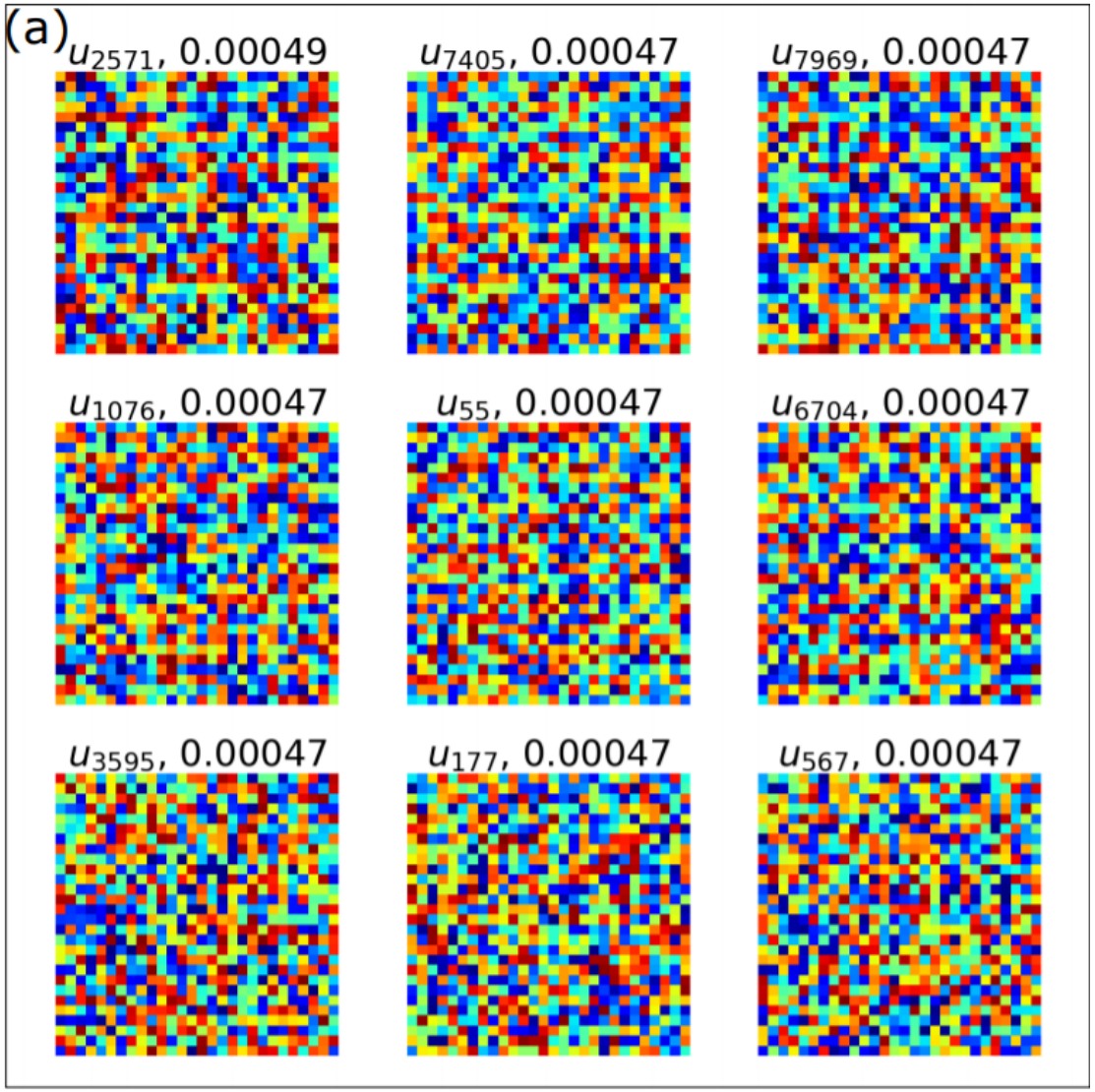
Breakdown of Symmetry and Homogeneity

Adversarial Regularization

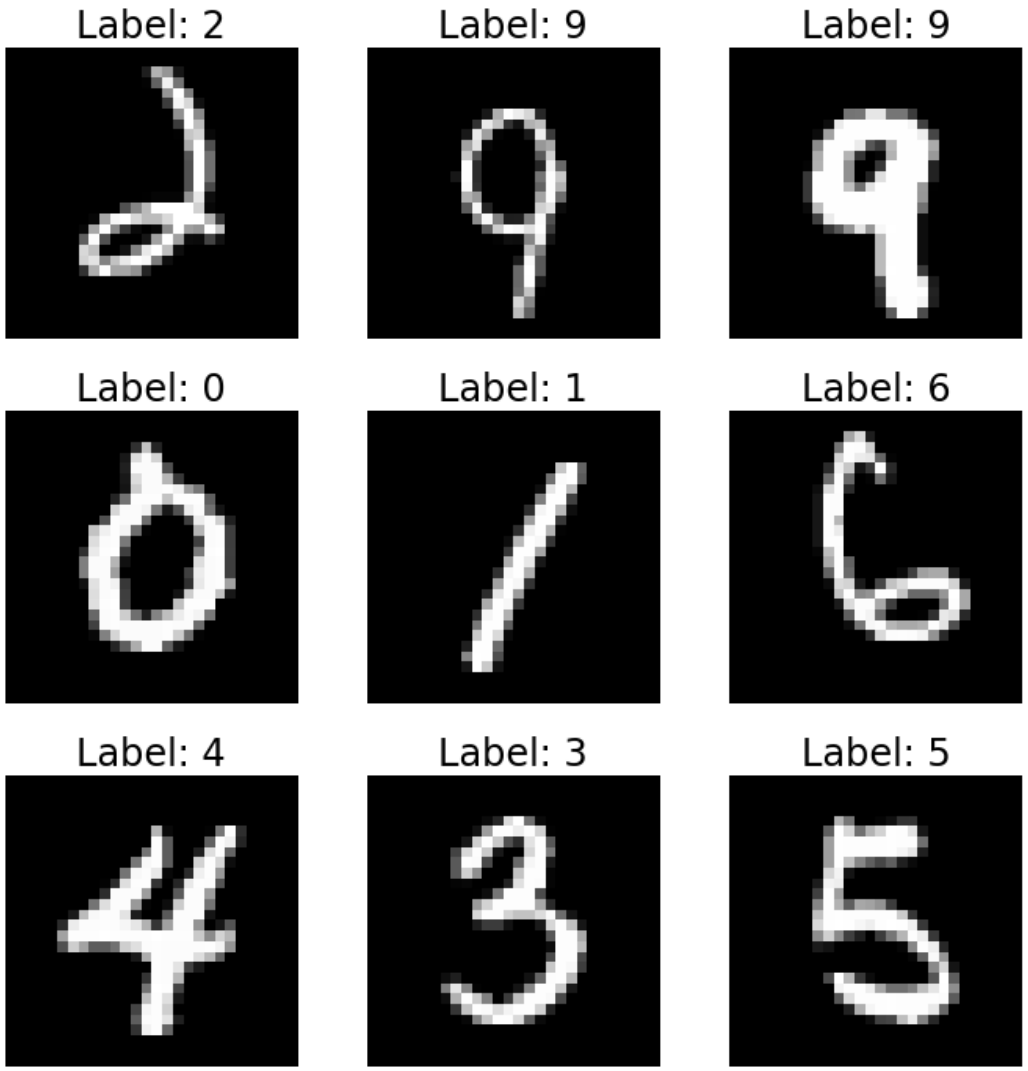


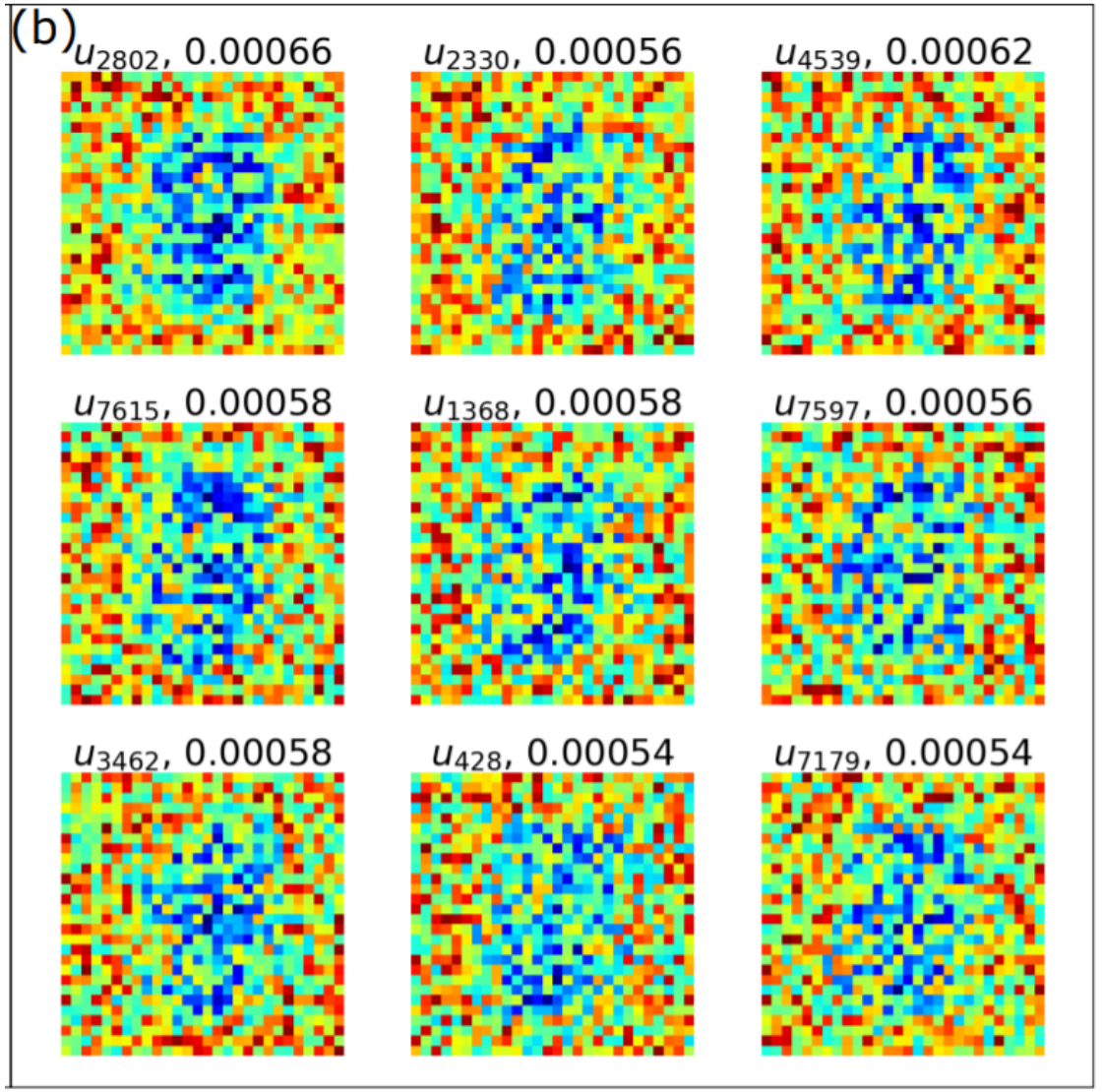
Interpretability of Supervised Learning



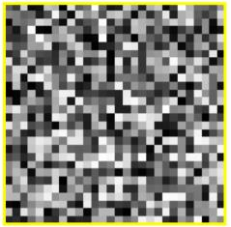


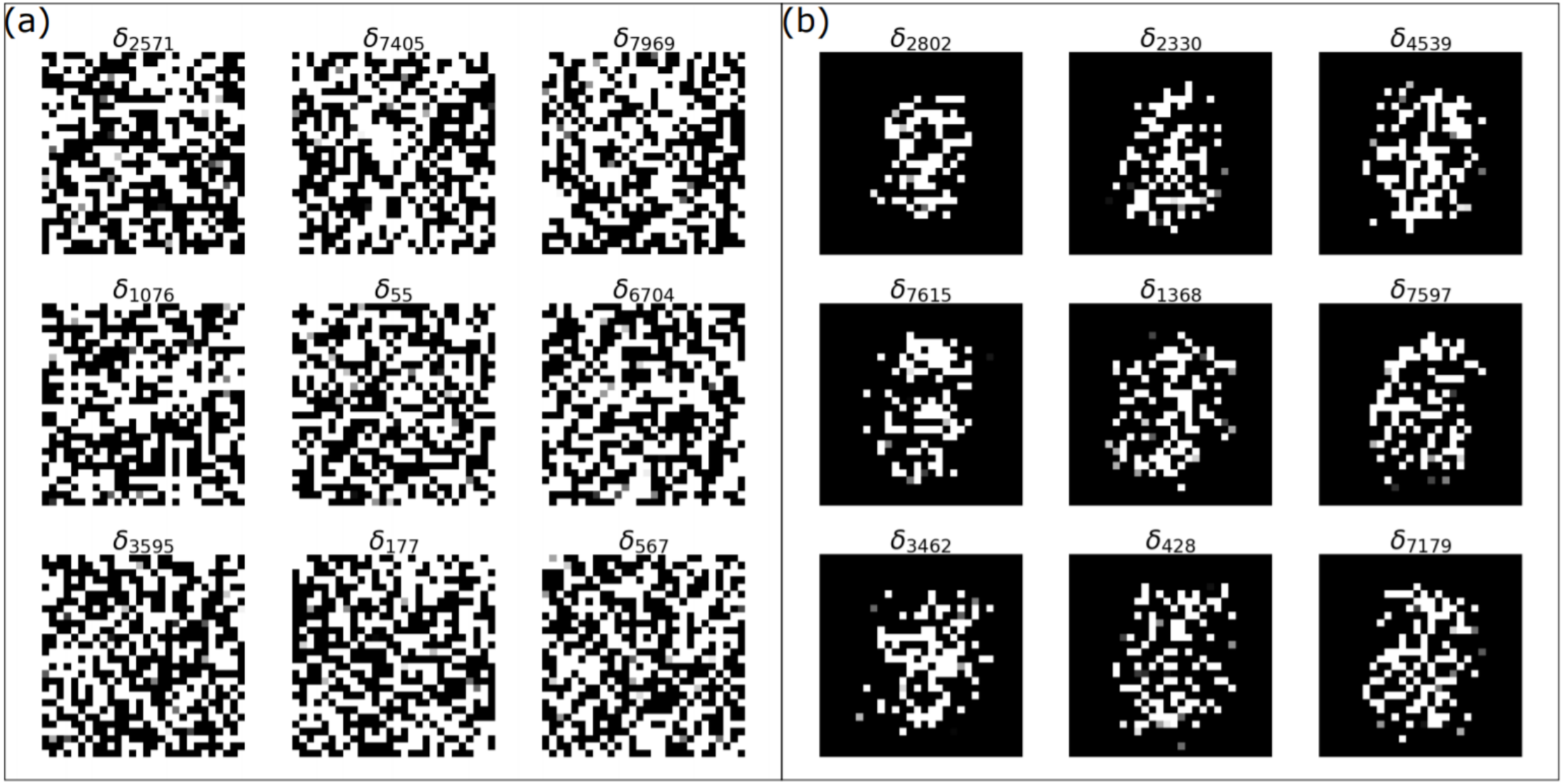
Interpretability of Adversarial Regularization





Interpretability of Adversarial Regularization





Jointly Training Both Layers

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

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

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

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

Bernoulli Differential Equation

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

Natural Phenomena

- Population Growth
- Tumour Growth
- Fermi-Dirac Statistics
- Diffusion of Innovation

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

J. Bernoulli, 1695

Bernoulli Differential Equation

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

Natural Phenomena

- Population Growth
- Tumour Growth
- Fermi-Dirac Statistics
- Diffusion of Innovation

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

Adversarial Regularization

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

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

- Adversarial Interaction
  - Generative Adversarial Networks (GANs)
  - Application of conditional GANs
- Non-Homogeneous Patterns
  - Homogeneous patterns
  - Supervised learning
- Reaction-Diffusion
  - Turing's RD model (1952)
  - Gray-Scott RD model (1984)
- Turing Instability
  - Reaction dynamics
  - Diffusion dynamics

Interesting Problems

- Why does adversarial interaction create non-homogeneous patterns?
- How do we predict the shape of the non-homogeneous patterns?
- What happens to the parameters if the supervised cost is augmented in the discriminator objective?
- Is gradient descent or backpropagation the optimal learning algorithm?