Exploring the Orthogonality and Linearity of Backdoor Attacks

Kaiyuan Zhang[1*], Siyuan Cheng[1*], Guangyu Shen[1], Guanhong Tao[1], Shengwei An[1], Anuran Makur[1], Shiqing Ma[2], Xiangyu Zhang[1]
[1]Purdue University, [2]University of Massachusetts Amherst, *Equal contribution

IEEE S&P

Introduction

- Backdoor attacks embed an attacker-chosen pattern into inputs to cause model misclassification.
- A number of defense techniques proposed by the community. *Do they work for a large spectrum of attacks?*
- We study the characteristics of backdoor attacks through theoretical analysis and introduce two key properties: **orthogonality** and **linearity**.

Table 1: A Summary of Existing Attacks and Defenses

	Attack	Model Detection			Backdoor Mitigation				Input Detection			
		NC [1]	Pixel [2]	ABS [3]	Fine-Pruning [4]	NAD [5]	ANP [6]	SEAM [7]	AC [8]	SS [9]	SPECTRE [10]	SCAn [11]
Patch	BadNets [12]	●	●	●	●	●	●	●	●	●	●	●
	TrojanNN [13]	●	●	●	◍	●	●	●	◍	○	◍	●
	Dynamic [14]	◍	◍	◍	◍	◍	○	◍	○	◍	◍	○
	CL [15]	◍	◍	◍	◍	●	●	◍	○	○	◍	◍
	Input-aware [16]	○	○	○	◍	◍	●	◍	◍	◍	◍	◍
Blend	Reflection [17]	○	○	○	○	○	○	●	○	○	○	○
	Blend [18]	◍	●	◍	◍	●	◍	◍	◍	○	◍	○
	SIG [19]	○	○	○	○	○	●	◍	◍	◍	◍	◍
Filter	Instagram [3]	○	○	●	◍	◍	◍	●	◍	◍	◍	○
	DFST [20]	○	◍	○	◍	◍	○	◍	◍	◍	◍	○
Invisible	WaNet [21]	○	○	○	◍	◍	◍	◍	◍	◍	◍	○
	Invisible [22]	◍	◍	○	◍	◍	◍	◍	◍	◍	◍	○
	Lira [23]	◍	◍	○	○	○	○	◍	◍	○	◍	○
	Composite [24]	○	○	○	○	○	○	○	○	○	◍	○

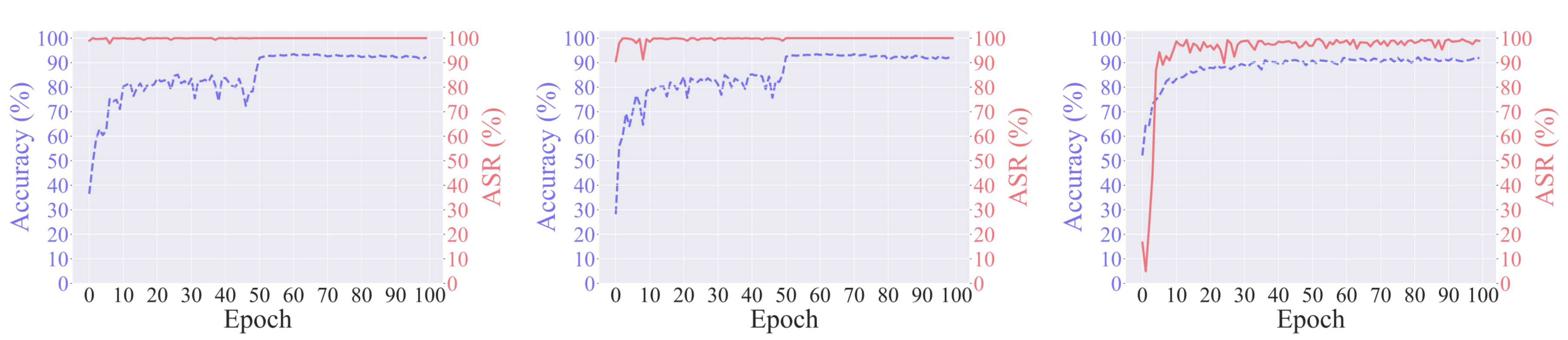
●: attacks can be defended, supported by existing works; ◍: attacks can be defended, supported by our experiments; ○: attacks cannot be defended.

Motivation

What are the underlying reasons causing defenses to fail on certain backdoor attacks?

Problem Formulation

- **Key Observation**: We observe that the backdoor task is quickly learned by the victim model (using a very few training epochs), much faster than the main task (clean).

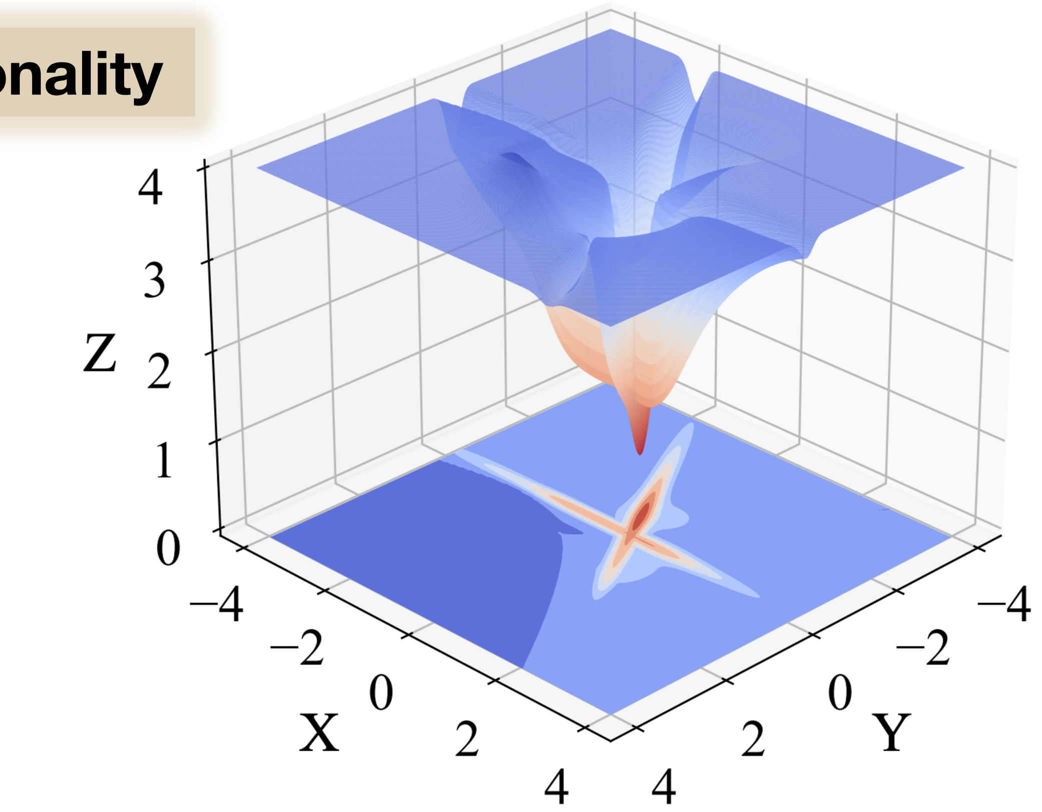


(a) BadNets (b) Blend (c) WaNet

- Based on our observation, we formulate backdoor learning as a two-task continual learning problem.

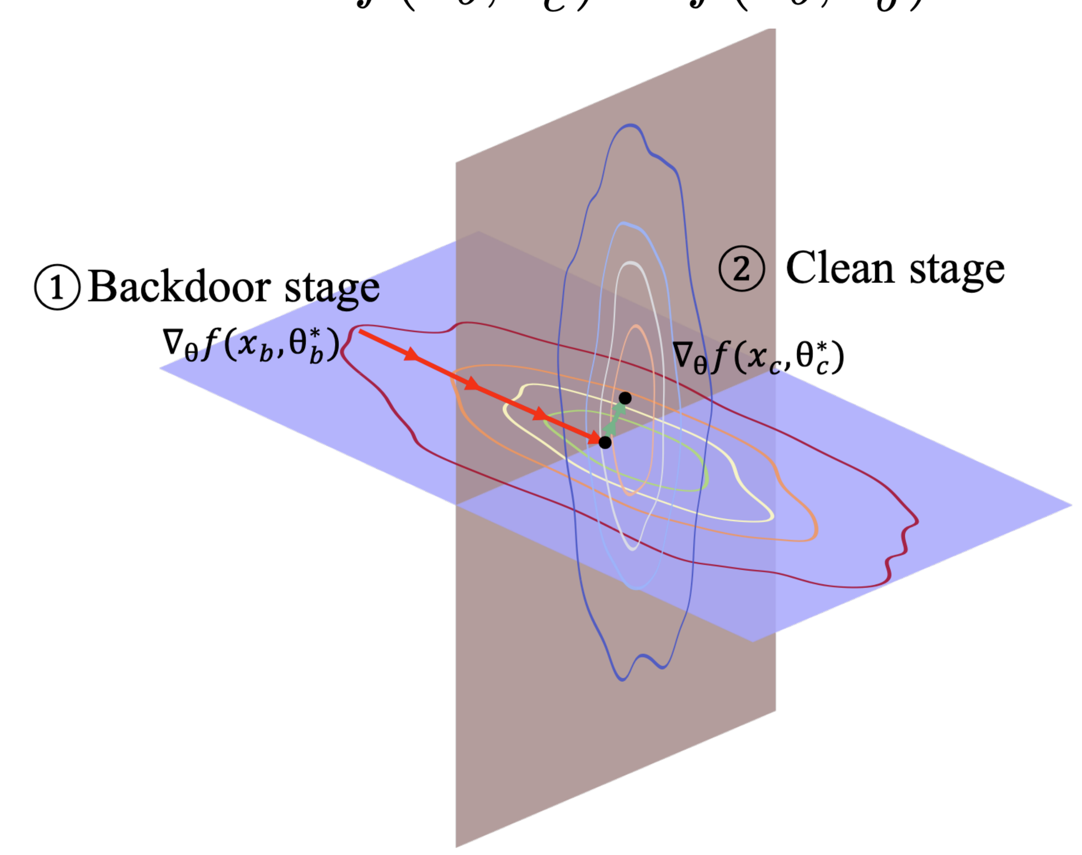
Theoretical Results

Orthogonality

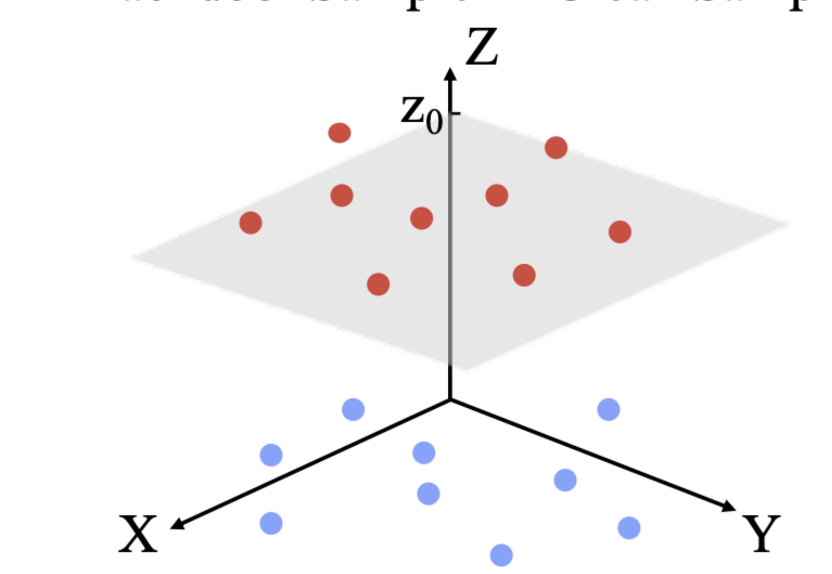


Theorem 3.4. ***(Backdoor Stays under Orthogonal Gradient Descent)*** *Let $f(x, \theta_b^\star)$ and $f(x, \theta_c^\star)$ represent the converged neural network associated with the backdoor and clean tasks, respectively, parameterized by converged backdoor model parameters $\theta_b^\star$ and converged clean model parameters $\theta_c^\star$. Given a sample of backdoor training data (x_b, y_b) derived from a prior backdoor task b and following the distribution $\mathcal{D}_b$, we can establish that*

$$f(x_b, \theta_c^\star) = f(x_b, \theta_b^\star) \quad (4)$$



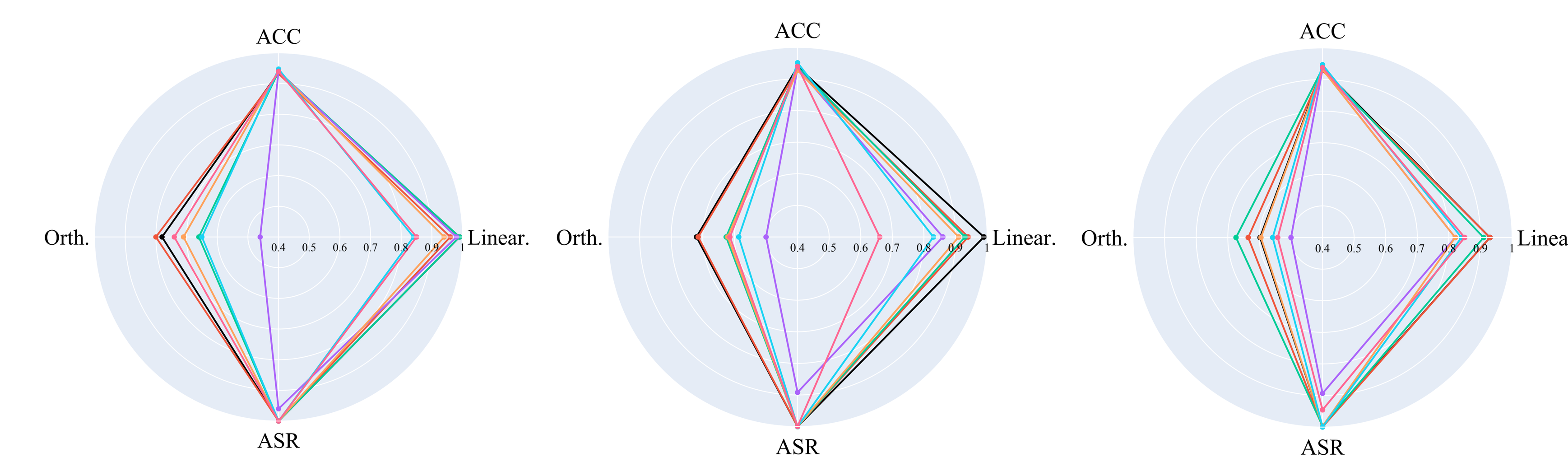
Linearity



Proposition 3.9. ***(Linearity Perspective of Backdoor Learning)*** *For a well-poisoned model $f : X \to Y$ with a near 100% attack success rate, there exists a specific hyperplane $\{\mathbf{Wx} - \mathbf{b} = 0\}$, which capable of capturing the Trojan behavior in the backdoor learning phase, and this trojan hyperplane persists in the clean learning phase.*

Numerical Results

- Evaluate our theoretical analysis and hypotheses on 14 attacks and 12 defenses.
- Investigate the impact of 6 key factors that affect the orthogonality and linearity.
- Offer insights on When and Why do defenses fail or succeed against various attacks.



(a). BadNets (b). Blend (c). WaNet

Take Away

- We systematically explore why existing defenses fail on certain backdoor attacks.
- We provide a theoretical analysis on two critical properties orthogonality and linearity.
- Unlock new insights — TRYOUT our NEW measurements beyond ASR and ACC!

$$Orth. = \arccos\left(\frac{\mathcal{L}(\theta_b^\star) \cdot \mathcal{L}(\theta_c^\star)}{\|\mathcal{L}(\theta_b^\star)\| \, \|\mathcal{L}(\theta_c^\star)\|}\right) \qquad Linear. = LR(\Delta\gamma, \Delta\rho)$$

Backdoor gradient ↑ Clean gradient ↑ Inputs fluctuations ↑ Outputs fluctuations ↑

Project Page

Personal Page