Feature Partition Aggregation: A Fast Certified Defense Against a Union of ℓ_0 Attacks

Zayd Hammoudeh, Daniel Lowd

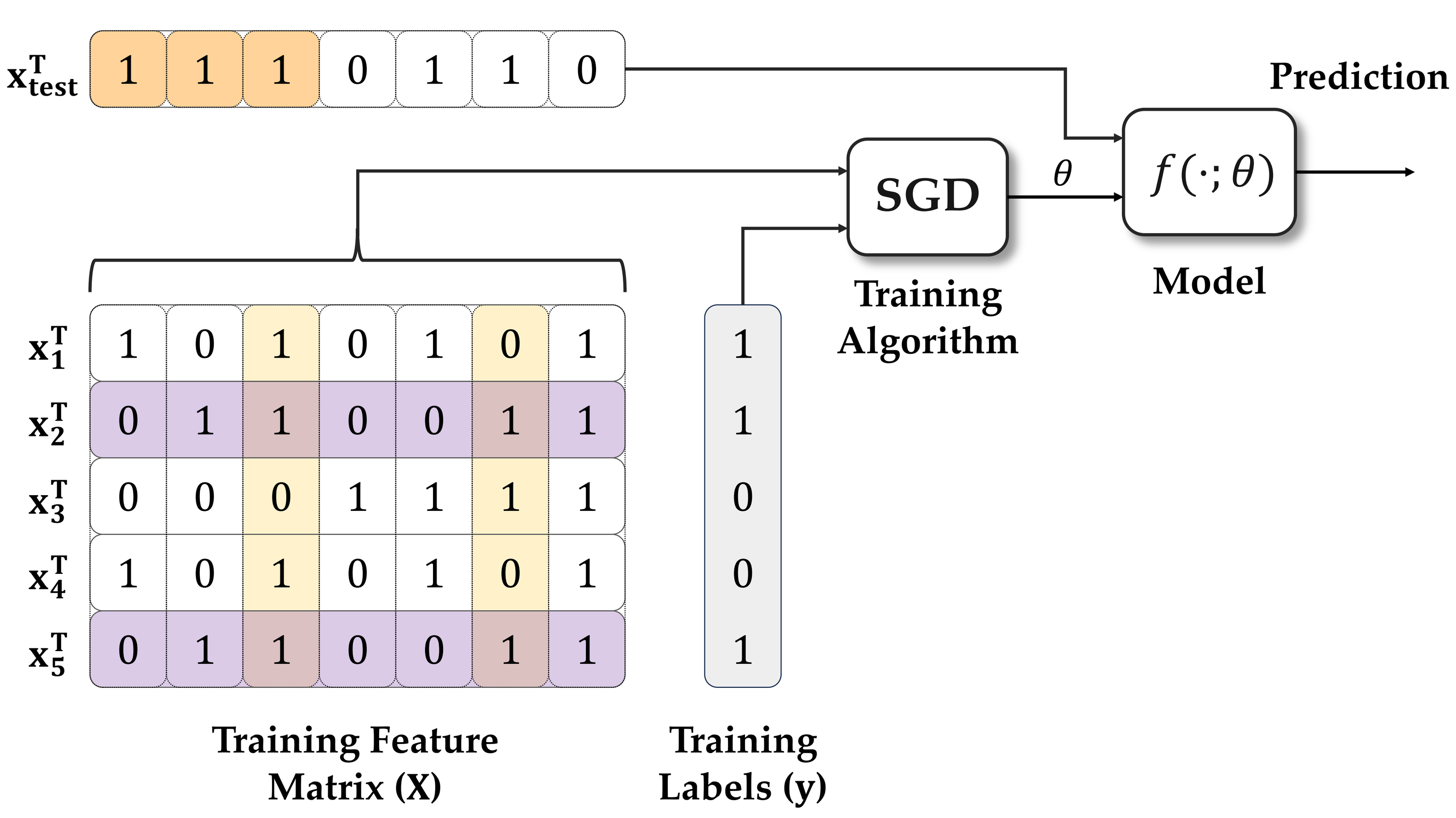
zayd@cs.uoregon.edu

ICML International Conference On Machine Learning AdvML-Frontiers Workshop

UNIVERSITY OF OREGON

Key Idea: An ensemble of submodels using disjoint feature subsets yields provable robustness to feature corruption

Types of Adversarial Attacks



Evasion Attack: Modifies test ($\mathbf{x}_{\text{test}}$) features only

Instance-wise Data Poisoning: Modifies specific training instances (rows of $\mathbf{X}$), including the labels ($\mathbf{y}$)

Feature-wise Data Poisoning: Modifies both training features (columns of $\mathbf{X}$) and test ($\mathbf{x}_{\text{test}}$) features

Backdoor Attack: Modifies both training and test data

Patch Attack: Evasion attack where the perturbation is restricted to a specific shape

What is an ℓ_0 Adversarial Attack?

ℓ_0 (Sparse) Attack: Adversary arbitrarily controls an unknown subset of the feature set

When is ℓ_0 Robustness Analysis Appropriate?

- Heterogenous feature types (e.g., both numerical and categorical features)
- Different feature scales
- Tabular data
- Certified patch robustness regardless of patch shape or number of patches

Certified Feature Robustness

Pointwise Certified Robustness: Provable guarantee of an individual prediction's robustness against an adversarial attack

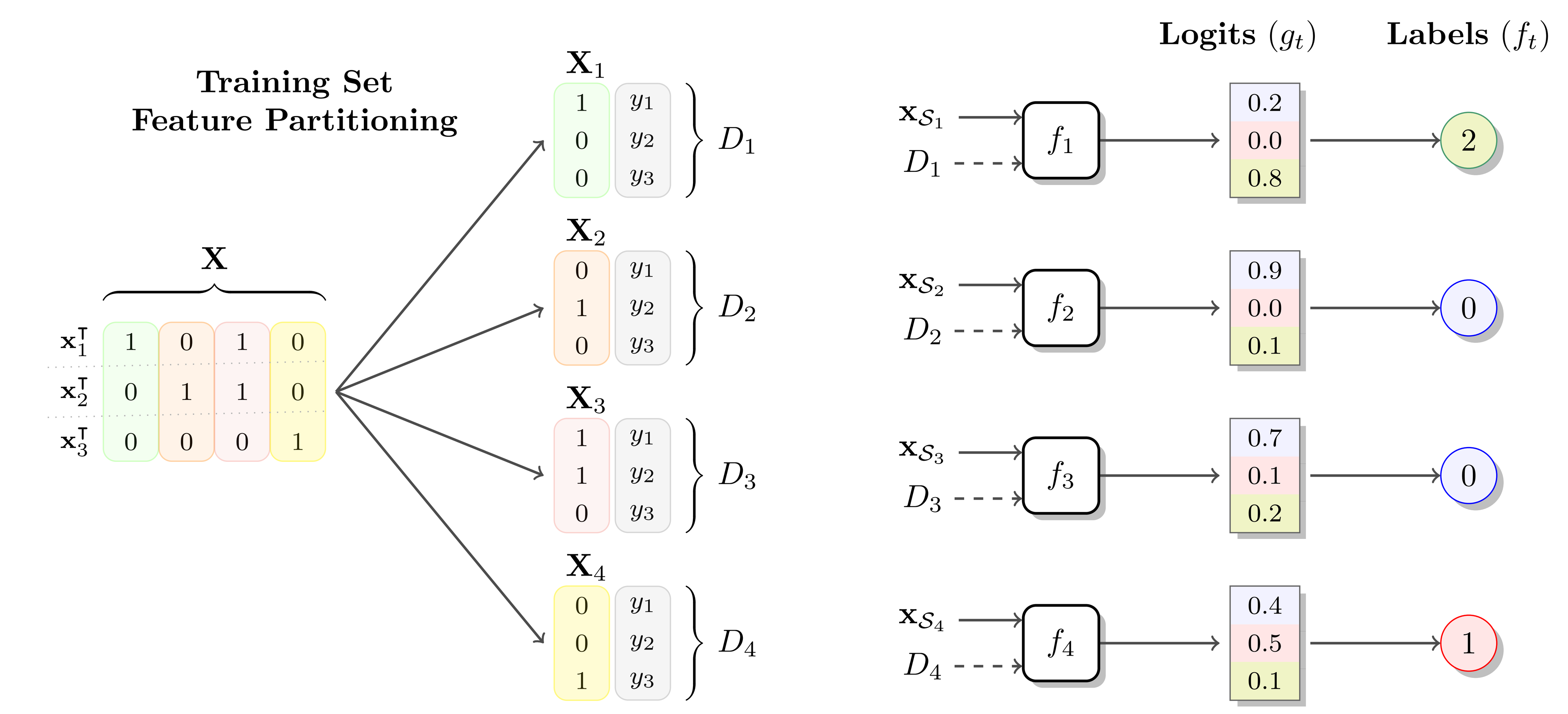
Certified Feature Robustness: Given model f trained on $(\mathbf{X}, \mathbf{y})$, model f' trained on $(\mathbf{X}', \mathbf{y})$, and feature vector $\mathbf{x}'$, a deterministic guarantee $r \in \mathbb{N}$ w.r.t. $(\mathbf{x}, y)$ where

$$|\mathbf{X} \ominus \mathbf{X}' \cup \mathbf{x} \ominus \mathbf{x}'| \leq r \Rightarrow y = f'(\mathbf{x}').$$

Feature robustness guarantees are over the **union of ℓ_0 evasion, backdoor, and poisoning attacks**.

Feature Partition Aggregation's Model Architecture

Ensemble of T submodels each trained on and evaluating a disjoint subset of the features set



Key Insight: Any adversarially perturbed feature (training or test) affects at most one submodel prediction

How to Partition the Feature Set?

Answer: Any way you want

Random Partitioning: Assign features to submodels uniformly at random

Deterministic Partitioning: Use domain-specific knowledge to craft a better feature partition

Benefits of FPA over Previous Work

Stronger Guarantees: Deterministic guarantee + robustness over the union of ℓ_0 evasion, backdoor, and poisoning attacks

Faster: Certify predictions orders of magnitude faster than randomized ablation

Model Architecture Agnostic: FPA supports any submodel architecture (e.g., random forests, neural networks, etc.)

Calculating FPA's Robustness Guarantee

Depends on the Decision Function

Plurality Voting: [LF21]

- **Plurality Label:** $f(\mathbf{x}) = y_{\text{pl}} \coloneqq \operatorname{argmax}_y \sum_i \mathbf{1}_{\{y = f_t(\mathbf{x})\}}$
- **Runner-Up Label:** $y_{\text{ru}} \coloneqq \operatorname{argmax}_{y \neq y_{\text{pl}}} \sum_i \mathbf{1}_{\{y = f_t(\mathbf{x})\}}$

$$r_{\text{pl}} = \left\lfloor \frac{\sum_i \mathbf{1}_{\{y_{\text{pl}} = f_t(\mathbf{x})\}} - \sum_i \mathbf{1}_{\{y_{\text{ru}} = f_t(\mathbf{x})\}} - \mathbf{1}_{\{y_{\text{ru}} < y_{\text{pl}}\}}}{2} \right\rfloor$$

Run-Off-Election: Two-round voting election for multiclass classification [Rez+23]

- **Round #1**: Identify plurality and runner-up labels
- **Round #2**: Submodels revote but only for either the plurality and runner-up labels

$$f(x) = \begin{cases} y_{\text{pl}} & \sum_i \mathbf{1}_{\{g_t(\mathbf{x}, y_{\text{pl}}) > g_t(\mathbf{x}, y_{\text{ru}})\}} - \mathbf{1}_{\{y_{\text{ru}} < y_{\text{pl}}\}} > \frac{T}{2} \\ y_{\text{ru}} & \text{Otherwise} \end{cases}$$

Run-off Feature Robustness: Minimum certified robustness of either rounds #1 and #2

Empirical Evaluation

Baseline: Randomized Ablation [LF20b, Jia+22b]

ℓ_0 evasion defense based on randomized smoothing

Median Certified Robustness: Median robustness value across a dataset's entire test set

Dataset	Dim. (d)	FPA (ours)		Random. Ablate.	
		Plural	Run-Off	[LF20b]	[Jia+22b]
CIFAR10	1024	11	**13**	7	10
MNIST	784	9	**12**	8	10
Weather	128	**4**	–	0	1
Ames	352	**3**	–	1	1

Takeaway: FPA provides larger and stronger median robustness guarantees than the baseline

Classification Accuracy

Fraction of correctly classified test predictions

Dataset	FPA (ours)				Rand. Abl.	
	r_{med}	Acc.	r_{med}	Acc.	ρ_{med}	Acc.
CIFAR10	13	62.4	10	**75.0**	10	64.7
MNIST	12	87.2	10	**96.1**	10	93.1
Weather	4	76.1	1	**85.3**	1	75.2
Ames	3	65.5	1	**84.6**	1	67.2

Takeaway: FPA's median robustness gains come at little to no cost in model accuracy.

Prediction Certification Time

Mean time in seconds to certify a single prediction

Dataset	RA [Jia+22b]		FPA (ours)		Speedup
	e	Time	T	Time	
CIFAR10	15	5.4E+0	115	7.3E−3	**743**×
MNIST	25	6.8E−1	60	2.9E−3	**235**×
Weather	45	3.1E−1	21	1.0E−4	**3,134**×
Ames	60	3.8E−1	21	3.5E−4	**1,082**×

Takeaway: FPA certifies predictions 2 to 3 orders of magnitude faster than the baseline.

FPA as a Certified Patch Defense

CIFAR10 Certified Patch Accuracy: Fraction of correctly classified test instances satisfying the robustness criterion

Method	24 Pixel Rect.		Square
	Min.	Max.	5 × 5
FPA Plurality ($T = 180$, ours)	⟵ 38.53 ⟶		37.77
FPA Run-Off ($T = 180$, ours)	⟵ 41.60 ⟶		40.95
Randomized Ablation [LF20b]	⟵ 28.95 ⟶		28.21
Randomized Ablation [Jia+22b]	⟵ 37.31 ⟶		36.43
(De)Random. Smoothing [LF20a]	0.0	72.68	57.69
BagCert [MY21]	**43.11**	60.17	59.95
Patch IBP [Chi+20b]	—	—	30.30

Takeaway: FPA provides strong certified patch robustness with fewer assumptions

References

[Chi+20b] P. Chiang, R. Ni, A. Abdelkader, C. Zhu, C. Studor, and T. Goldstein. "Certified Defenses for Adversarial Patches," ICLR, 2020.

[LF20a] A. Levine and S. Feizi. "(De)Randomized Smoothing for Certifiable Defense against Patch Attacks," NeurIPS, 2020.

[LF20b] A. Levine and S. Feizi. "Robustness Certificates for Sparse Adversarial Attacks by Randomized Ablation," AAAI, 2020.

[LF21] A. Levine and S. Feizi, "Deep Partition Aggregation: Provable Defenses Against General Poisoning Attacks," ICLR, 2021.

[MY21] J. Metzen and M. Yatsura. "Efficient Certified Defenses Against Patch Attacks on Image Classifiers," ICLR, 2021.

[Jia+22] J. Jia, B. Wang, X. Cao, H. Liu, and N. Gong. "Almost Tight ℓ_0-norm Certified Robustness of Top-k Predictions against Adversarial Perturbations," ICLR, 2022.

[HL23] Z. Hammoudeh and D. Lowd. Reducing Certified Regression to Certified Classification for General Poisoning Attacks," SaTML, 2023.

[Rez+23] K. Rezaei, K. Banihashem, A. Chegini, and S. Feizi. "Run-Off Election: Improved Provable Defense against Data Poisoning Attacks," ICML, 2023.