Adversarial Robustness through Bias Variance Decomposition: A New Perspective for Federated Learning

Yao Zhou*[1,2], Jun Wu*[1], Haixun Wang[2], Jingrui He[1]

[1]University of Illinois at Urbana-Champaign, [2]Instacart

yaozhou3@illinois.edu, junwu3@illinois.edu, haixun@gmail.com, jingrui@illinois.edu

Problem Definition

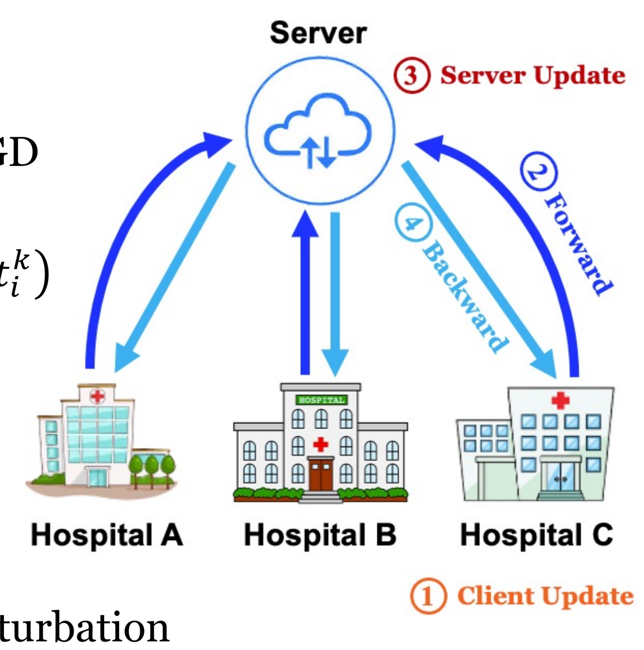
❑ Motivation

- FedAvg
  - ➢ Client update with local SGD

$$w_k \leftarrow w_k - \alpha \frac{1}{n_k} \sum_{i=1}^{n_k} L\left(f_{\mathcal{D}_k}(x_i^k; w_k), t_i^k\right)$$

  - ➢ Server update

$$w_G = \sum_{k=1}^{K} \frac{n_k}{n} w_k$$



- Vulnerable to adversarial perturbation for model inference

❑ Adversarially-robust federated learning

- Given
  - ➢ K clients with local data $\{\mathcal{D}_k\}_{k=1}^K$
  - ➢ A learning algorithm $f(\cdot)$
  - ➢ Loss function $L(\cdot,\cdot)$
  - ➢ A public auxiliary training set $\mathcal{D}_s$
- Output
  - ➢ A trained model on the central server that is **robust against adversarial perturbations** on the test set $\mathcal{D}_{test}$

Algorithm: Fed_BVA

❑ Server update

- Model aggregation: $w_G = \text{Aggregate}(w_1, w_2, \cdots, w_K)$
- Adversarial example generation: For any $x \in \mathcal{D}_s$

$$\max_{\hat{x} \in \Omega(x)} B(\hat{x}; w_1, w_2, \cdots, w_K) + V(\hat{x}; w_1, w_2, \cdots, w_K)$$

  - ➢ BV-FGSM:

$$\hat{x} \leftarrow x + \epsilon \cdot \text{sign}\left(\nabla_x\left(B(x; w_1, w_2, \cdots, w_K) + V(x; w_1, w_2, \cdots, w_K)\right)\right)$$

  - ➢ For cross-entropy loss function,

$$\nabla_x B_{CE}(x; w_1, w_2, \cdots, w_K) = \frac{1}{K} \sum_{k=1}^{K} \nabla_x L(f_{\mathcal{D}_k}(x; w_k), t)$$

$$\nabla_x V_{CE}(x; w_1, w_2, \cdots, w_K) = \frac{1}{K} \sum_{k=1}^{K} \sum_{c=1}^{C} \left(\log y_m^{(j)} + 1\right) \cdot \nabla_x f_{\mathcal{D}_k}(x; w_k)$$

❑ Backward communication

- Send global model parameters w_G and poisoned examples $\hat{x}$ to candidate client

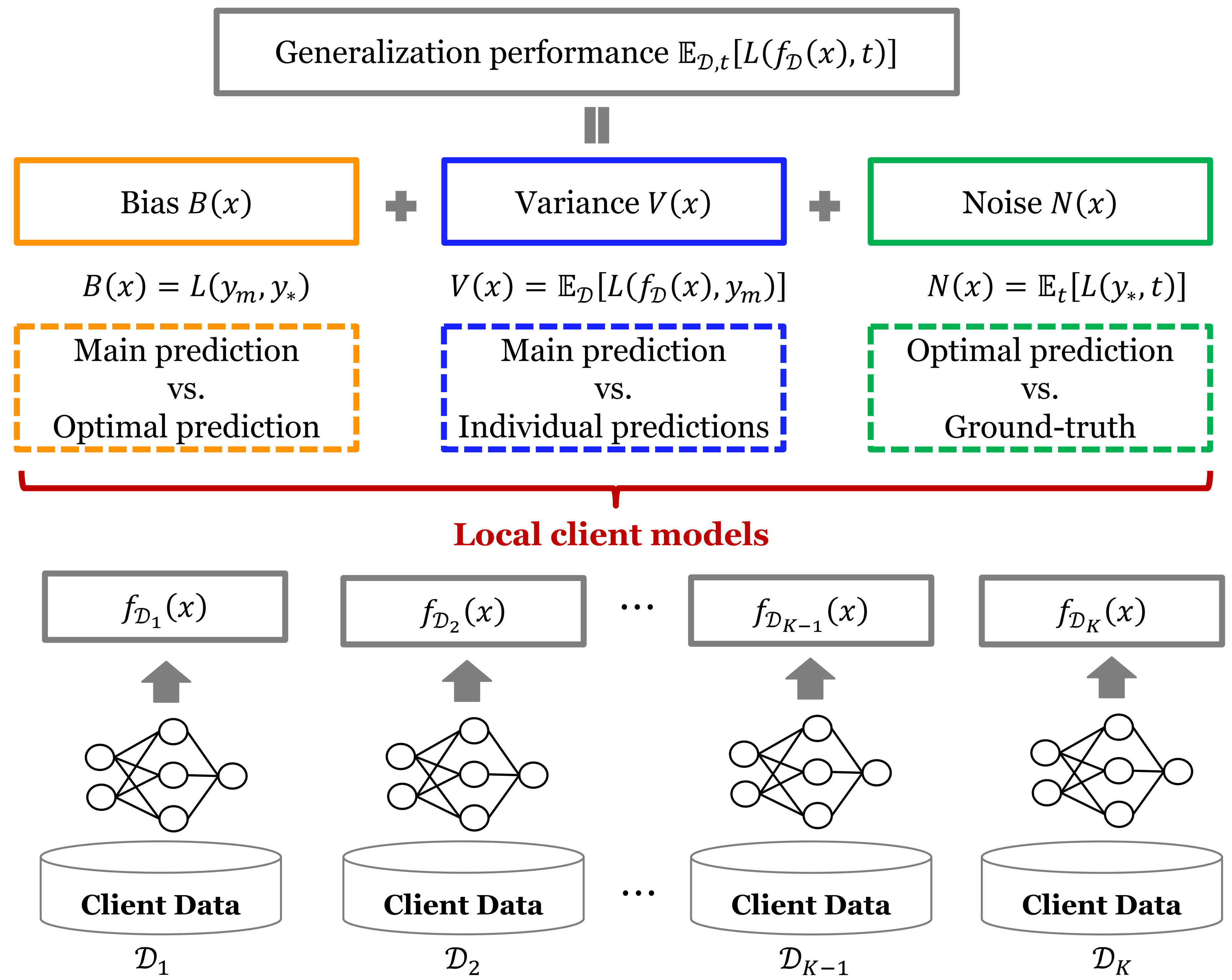
❑ Client update

Robust training:

$$\min_{w_k} \frac{1}{n_k} \sum_{i=1}^{n_k} L\left(f_{\mathcal{D}_k}(x_i^k; w_k), t_i^k\right) + \frac{1}{n_s} \sum_{j=1}^{n_s} L\left(f_{\mathcal{D}_k}(\hat{x}_j^s; w_k), t_j^s\right)$$

❑ Forward communication

- Upload local parameter updates to the server



Conclusion

- ❑ **Problem**: The adversarial robustness of federated learning is studied under the observation that federated learning model is vulnerable to evasion attacks when it is deployed.
- ❑ **Algorithm**: By investigating the generalization error of clients' local models, we propose a bias-variance oriented adversarial training algorithm Fed_BVA for robust federated learning.
- ❑ **Evaluation**: Extensive experiments confirm the effectiveness and efficiency of the Fed_BVA algorithm.

Acknowledgments

This work is supported by National Science Foundation under Award No. IIS-1947203, IIS-2117902, IIS-2137468, and Agriculture and Food Research Initiative (AFRI) grant no. 2020-67021-32799/project accession no.1024178 from the USDA National Institute of Food and Agriculture. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.

Experimental Results

❑ Performance comparison

Method	IID			non-IID		
	Clean	FGSM	PGD-20	Clean	FGSM	PGD-20
Centralized	**0.991**$_{\pm 0.000}$	0.689$_{\pm 0.000}$	0.182$_{\pm 0.000}$	n/a	n/a	n/a
FedAvg	0.989$_{\pm 0.001}$	0.669$_{\pm 0.009}$	0.267$_{\pm 0.014}$	**0.980**$_{\pm 0.002}$	0.491$_{\pm 0.067}$	0.158$_{\pm 0.074}$
FedAvg_AT	0.988$_{\pm 0.000}$	0.802$_{\pm 0.001}$	0.512$_{\pm 0.042}$	0.974$_{\pm 0.005}$	0.649$_{\pm 0.066}$	0.363$_{\pm 0.066}$
Fed_Bias	0.986$_{\pm 0.000}$	0.812$_{\pm 0.009}$	0.583$_{\pm 0.036}$	0.971$_{\pm 0.004}$	0.679$_{\pm 0.040}$	0.394$_{\pm 0.103}$
Fed_Variance	0.985$_{\pm 0.001}$	0.803$_{\pm 0.007}$	0.572$_{\pm 0.019}$	0.973$_{\pm 0.005}$	0.684$_{\pm 0.004}$	0.395$_{\pm 0.049}$
Fed_BVA	0.986$_{\pm 0.001}$	0.818$_{\pm 0.003}$	0.613$_{\pm 0.020}$	0.969$_{\pm 0.002}$	0.705$_{\pm 0.009}$	0.469$_{\pm 0.031}$
EAT	0.981$_{\pm 0.000}$	**0.902**$_{\pm 0.001}$	0.811$_{\pm 0.004}$	0.972$_{\pm 0.002}$	0.789$_{\pm 0.016}$	0.415$_{\pm 0.035}$
EAT+Fed_BVA	0.980$_{\pm 0.001}$	0.901$_{\pm 0.006}$	**0.821**$_{\pm 0.013}$	0.965$_{\pm 0.005}$	**0.811**$_{\pm 0.020}$	**0.670**$_{\pm 0.014}$

Robustness on MNIST under IID and non-IID settings

❑ Ablation study

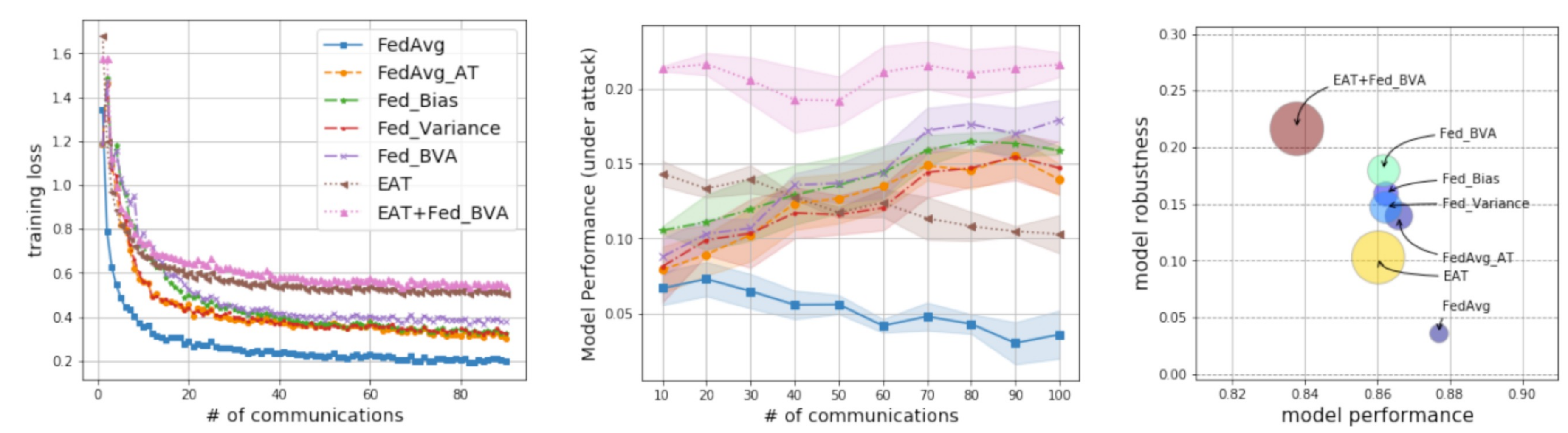
Loss	Clean	Fed_BVA		
		BiasOnly	VarianceOnly	BVA
CE	0.588(38.13s)	**0.763**(47.58s)	**0.759**(63.46s)	**0.776**(63.67s)
MSE	**0.601**(39.67s)	0.711(65.03s)	0.711(162.40s)	0.712(179.60s)

(a) Cross-Entropy (CE) vs. Mean Squared Error (MSE)

Method	IID			non-IID		
	FGSM	PGD-10	PGD-20	FGSM	PGD-10	PGD-20
FedAvg	0.588	0.620	0.205	0.147	0.525	0.089
Fed_BVA$_{(BV\text{-}FGSM)}$	**0.776**	0.793	0.570	**0.670**	0.695	0.472
Fed_BVA$_{(BV\text{-}PGD)}$	0.757	**0.840**	**0.632**	0.659	**0.784**	**0.575**

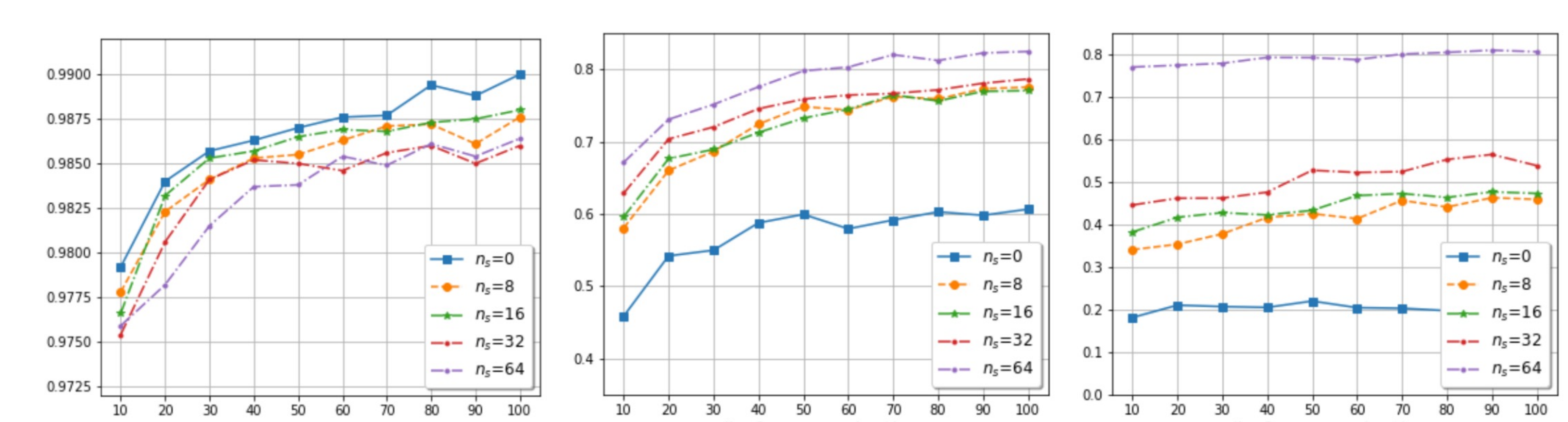
(b) BV-FGSM vs. BV-PGD

❑ Model analysis



(a) Convergence (b) Performance (c) Efficiency

❑ Hyperparameter sensitivity – size of public data set n_s



(a) Clean training (b) Under FGSM attack (b) Under PGD attack

ILLINOIS