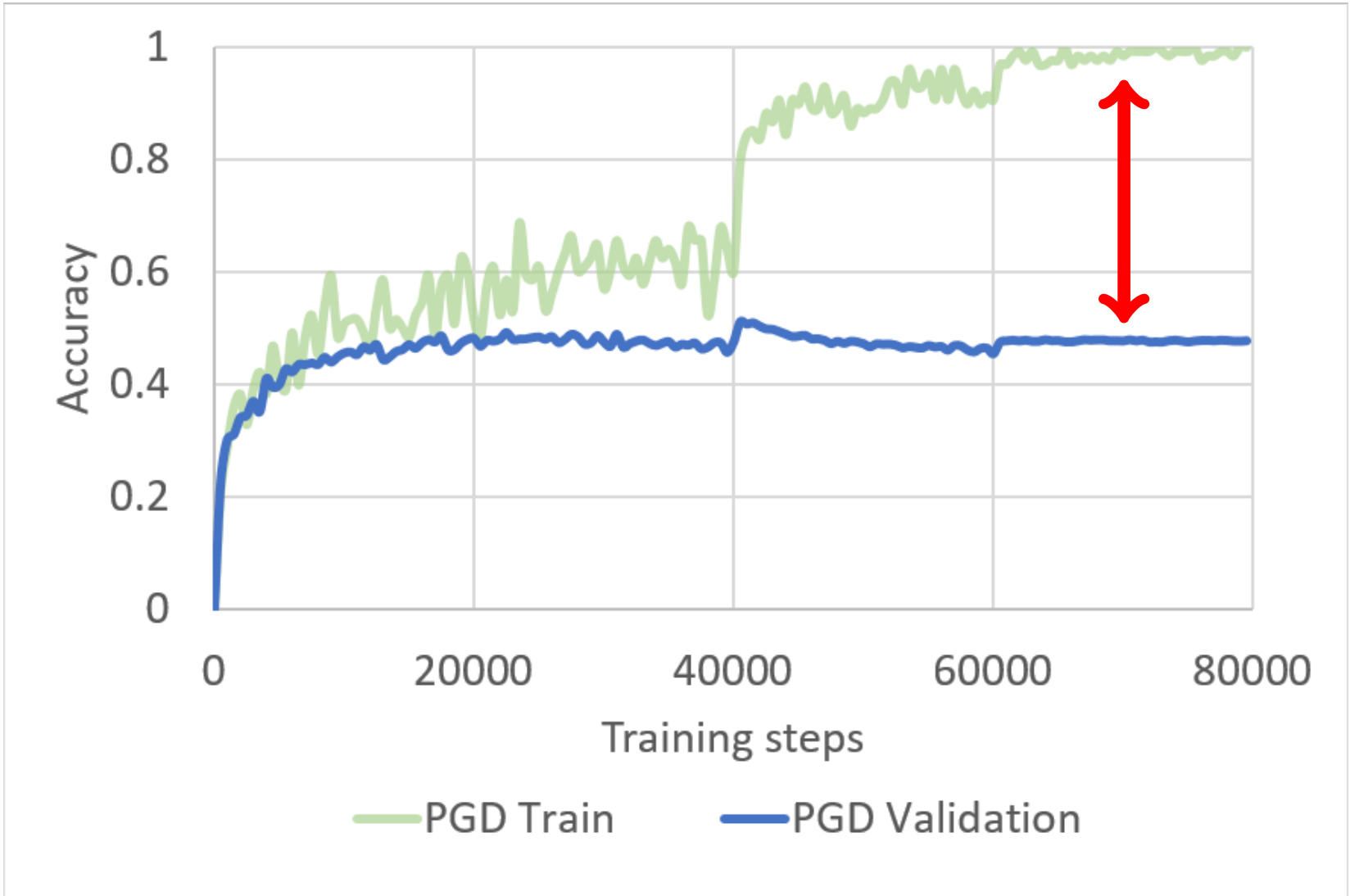
Adversarial Vertex Mixup: Toward Better Adversarially Robust Generalization

Saehyung Lee Hyungyu Lee Sungroh Yoon

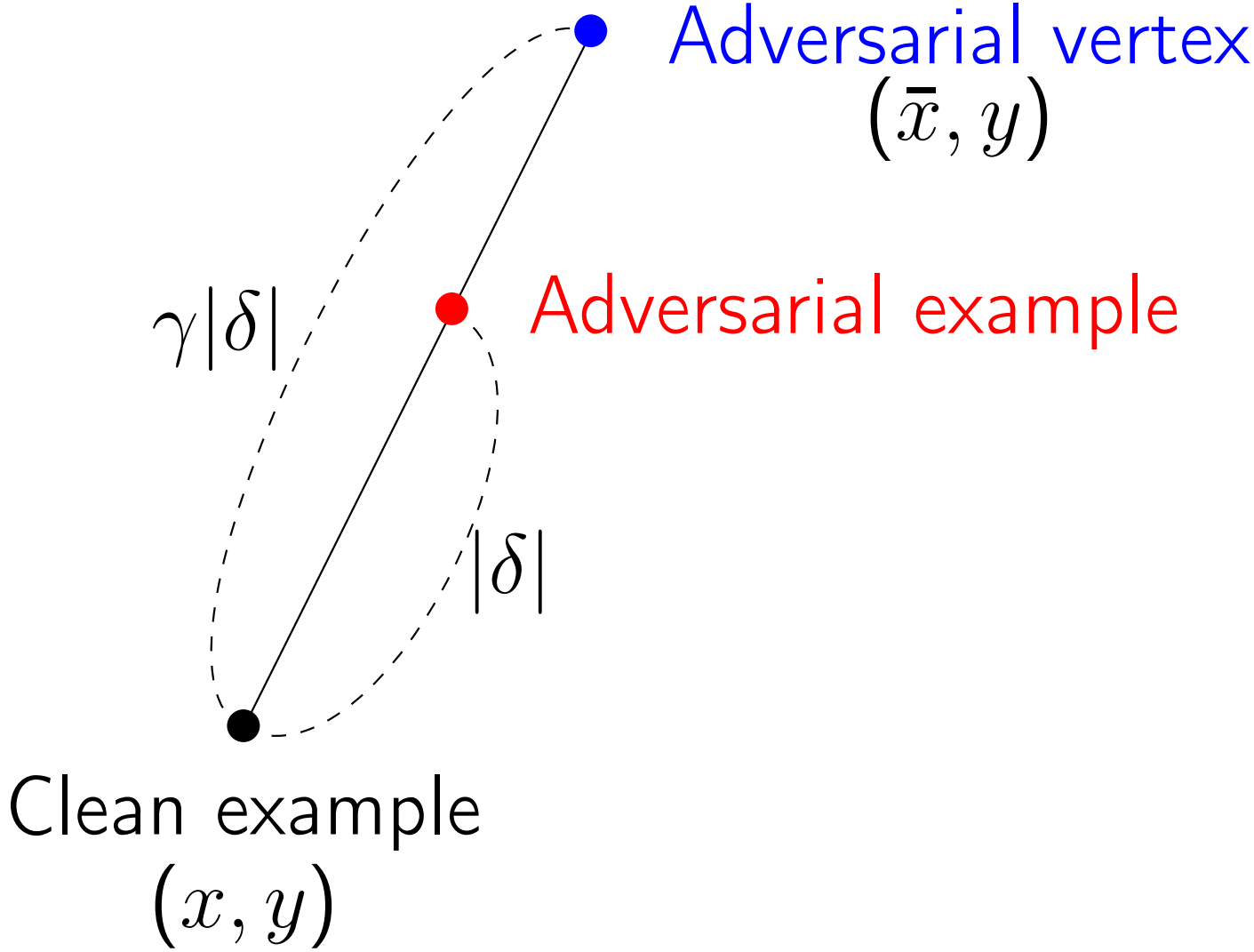
SEOUL
NATIONAL
UNIVERSITY

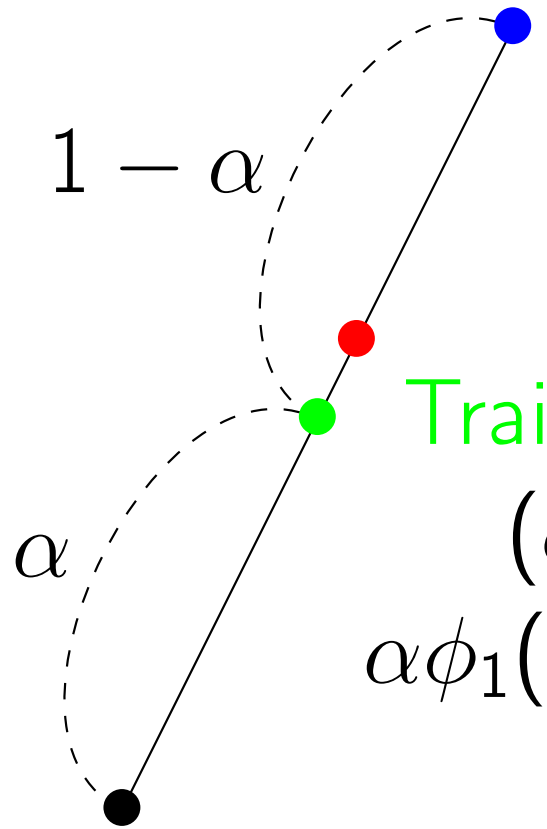
Problem statement



Adversarial Vertex mixup

Adversarial Vertex mixup (AVmixup)
is a soft-labeled data augmentation method
for improving
adversarially robust generalization.


Adversarial vertex
$(\bar{x}, y)$
$\gamma|\delta|$
Adversarial example
$|\delta|$
Clean example
(x, y)



Adversarial vertex

$(\bar{x}, \phi_2(y))$

Training example

$(\alpha x + (1-\alpha)\bar{x},$

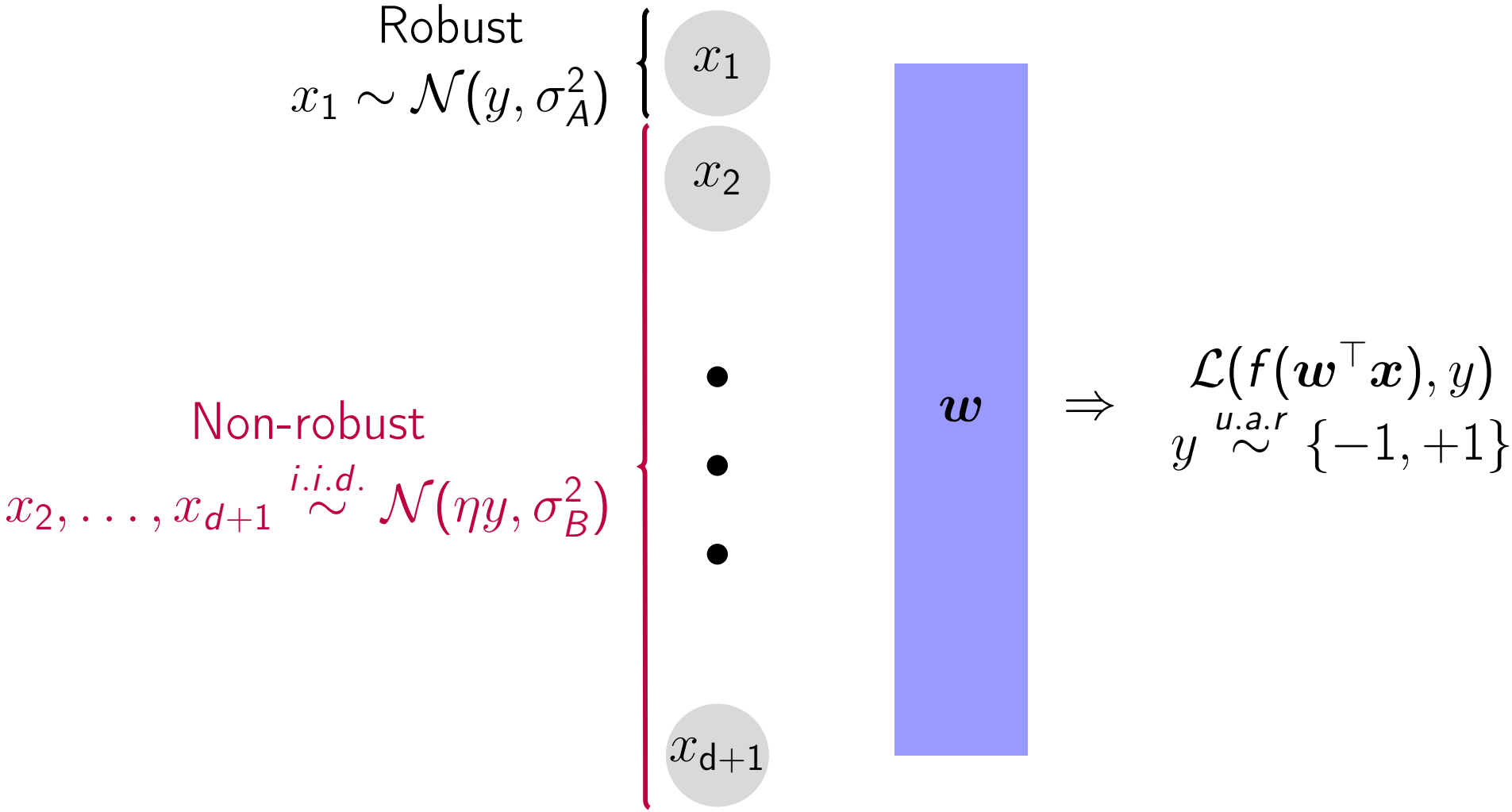
$\alpha\phi_1(y) + (1-\alpha)\phi_2(y))$

$\alpha \in [0, 1]$

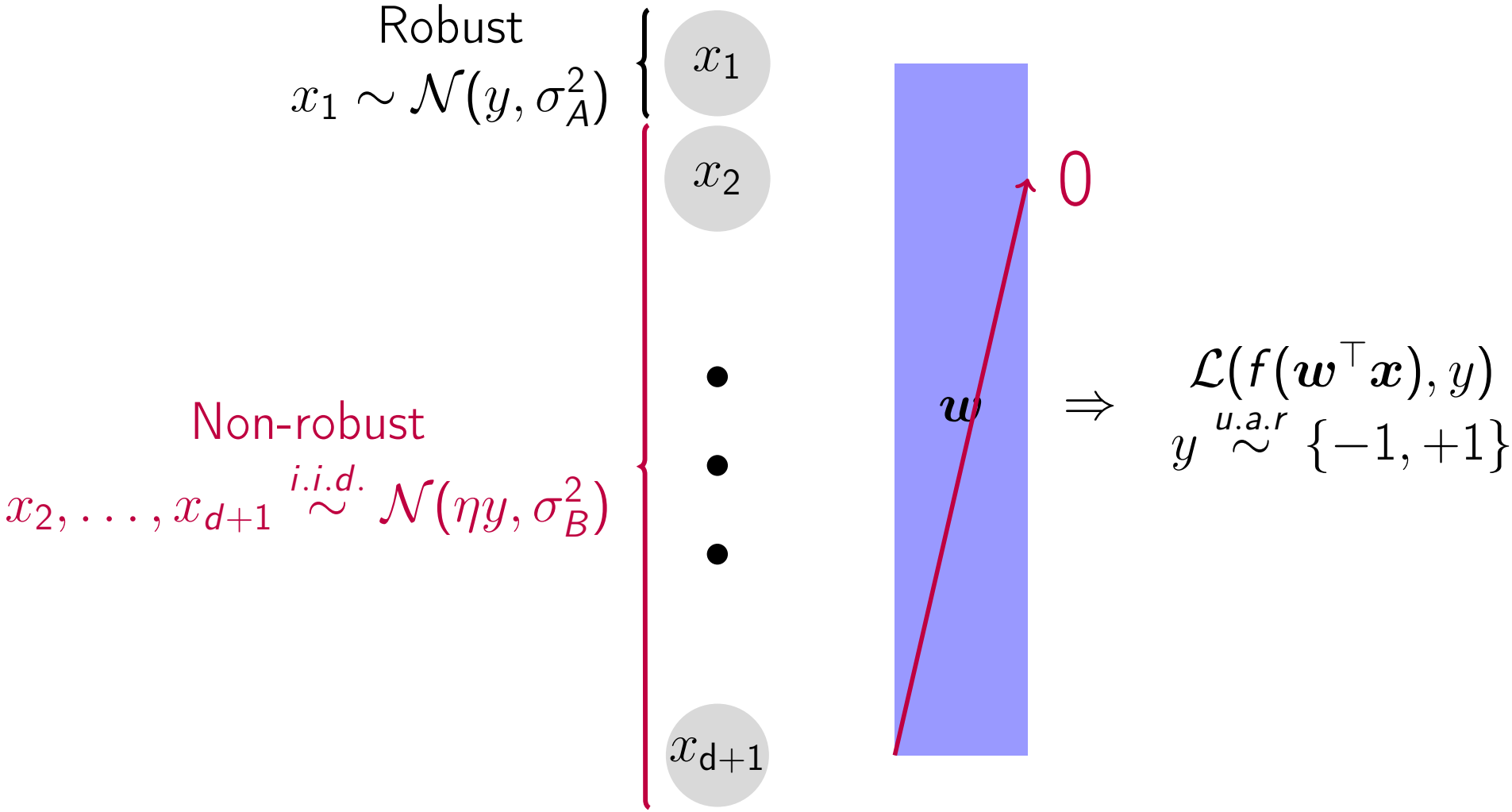
Clean example

$(x, \phi_1(y))$

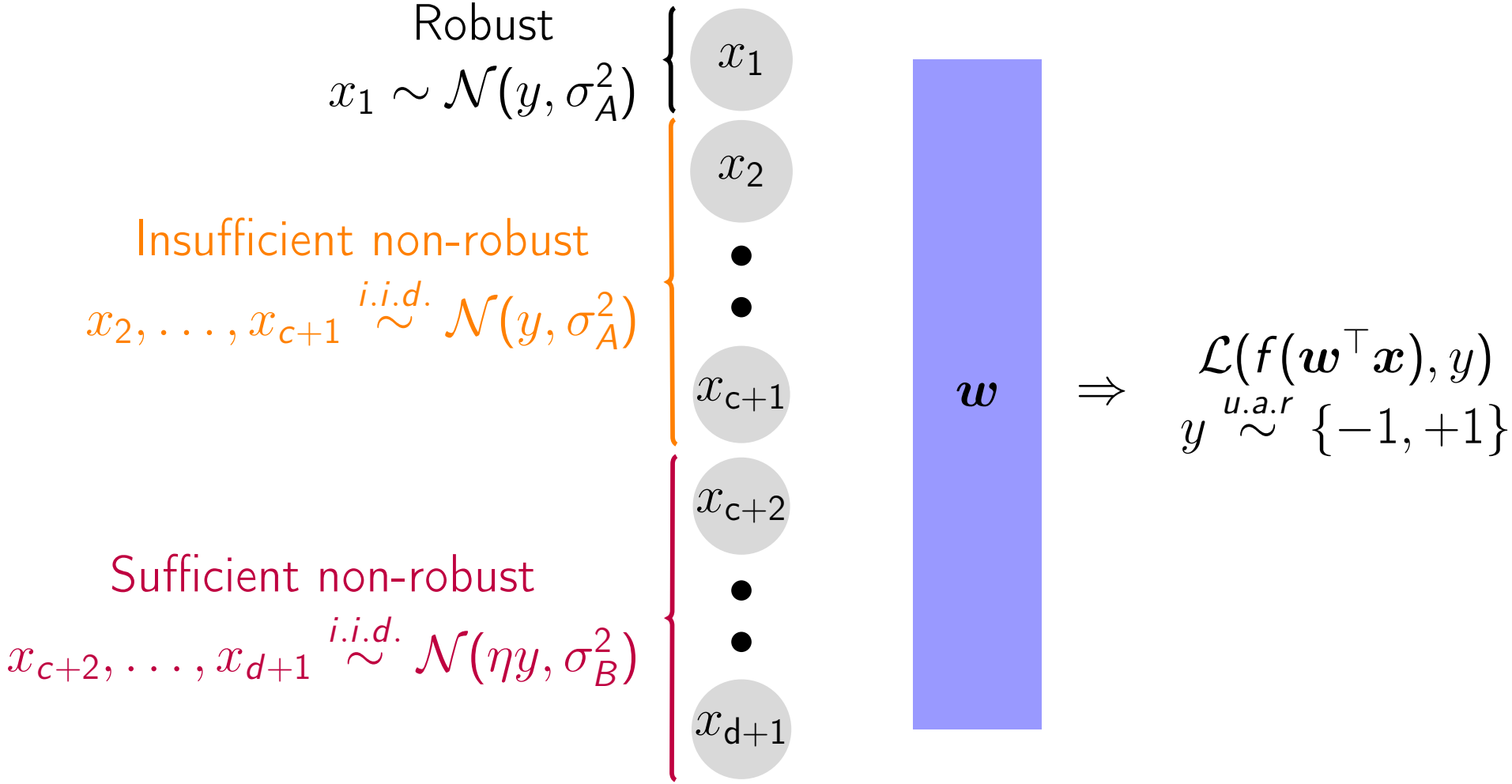
Adversarial Feature Overfitting



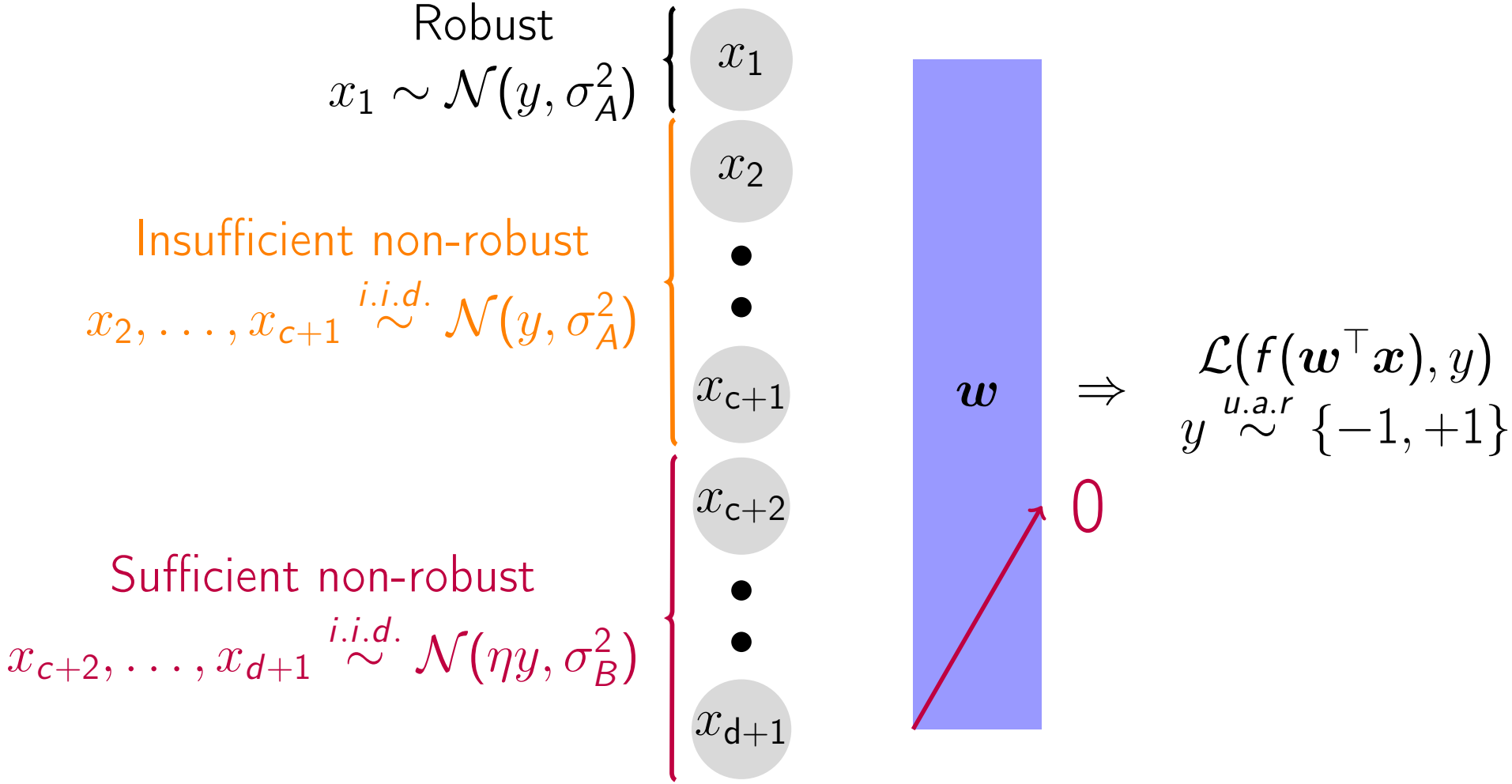
Adversarial Feature Overfitting



Adversarial Feature Overfitting



Adversarial Feature Overfitting



Linearization

$$|f(\boldsymbol{x}+\boldsymbol{\delta})-f(\boldsymbol{x})| = \left|\boldsymbol{\delta}^{\top}\nabla_{\boldsymbol{x}} f(\boldsymbol{x})\right|$$

Experimental results

Experimental results

White-box attack evaluation (acc%)

Model	Clean	FGSM	PGD10	PGD20	CW20
Standard	95.48	7.25	0.0	0.0	0.0
PGD	86.88	62.68	47.69	46.34	47.35
LS0.8	87.28	66.09	53.49	50.87	50.60
LS0.9	87.64	65.96	52.82	50.29	50.30
AVmixup	**93.24**	**78.25**	**62.67**	**58.23**	**53.63**

Transfer-based black-box attack evaluation (acc%)

Defense model	Attack model			
	Standard	PGD	LS0.8	LS0.9
PGD	85.6	-	65.70	64.91
LS0.8	86.03	63.60	-	64.83
LS0.9	86.40	63.74	65.78	-
AVmixup	**89.53**	**68.51**	**71.48**	**70.50**

Experimental results

Dataset	Model	Clean	FGSM	PGD20	PGD100	CW20	CW100
CIFAR10	PGD	85.7	54.9	44.9	44.8	45.7	45.4
	Feature Scatter	90.22	78.19	69.74	67.35	60.77	58.29
	Feature Scatter + AVmixup	**92.37**	**83.49**	**82.31**	**81.88**	**71.88**	**69.50**
CIFAR100	PGD	59.9	28.5	22.6	22.3	23.2	23.0
	Feature Scatter	74.9	72.99	45.29	42.77	27.35	24.89
	Feature Scatter + AVmixup	**78.62**	**78.92**	**47.28**	**46.29**	**33.20**	**31.22**
SVHN	PGD	93.9	68.4	47.9	46.0	48.7	47.3
	Feature Scatter	**96.42**	**95.92**	58.67	46.98	51.23	38.89
	Feature Scatter + AVmixup	96.07	95.26	**73.65**	**70.24**	**67.06**	**62.01**

Conclusion

- Identify Adversarial Feature Overfitting (AFO), which may cause poor adversarially robust generalization.
- Propose Adversarial Vertex mixup (AVmixup), a soft-labeled data augmentation approach for improving adversarially robust generalization.
- The results of experiments indicate that AVmixup substantially reduces the generalization gap in state-of-the-art adversarial training methods.

Thank you!

Poster #000

Saehyung Lee

Hyungyu Lee

Sungroh Yoon

Acknowledgements: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) [2018R1A2B3001628], the Brain Korea 21 Plus Project in 2020, Samsung Research Funding & Incubation Center of Samsung Electronics under Project Number SRFC-IT1901-12, and AIR Lab (AI Research Lab) in Hyundai Motor Company through HMC-SNU AI Consortium Fund.