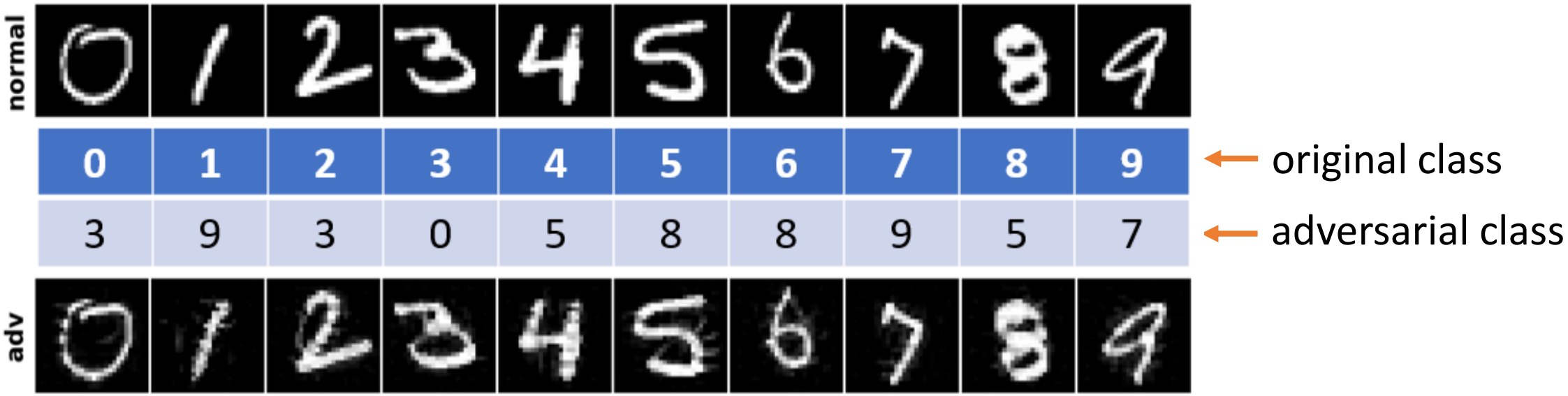
Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality

Xingjun Ma[1], Bo Li[2], Yisen Wang[3], Sarah M. Erfani[1], Sudanthi Wijewickrema[1], Grant Schoenebeck[4], Dawn Song[2], Michael E. Houle[5], James Bailey[1]

[1]The University of Melbourne; [2]University of California, Berkeley; [3]Tsinghua University; [4]University of Michigan; [5]National Institute of Informatics, Japan

Adversarial Examples:

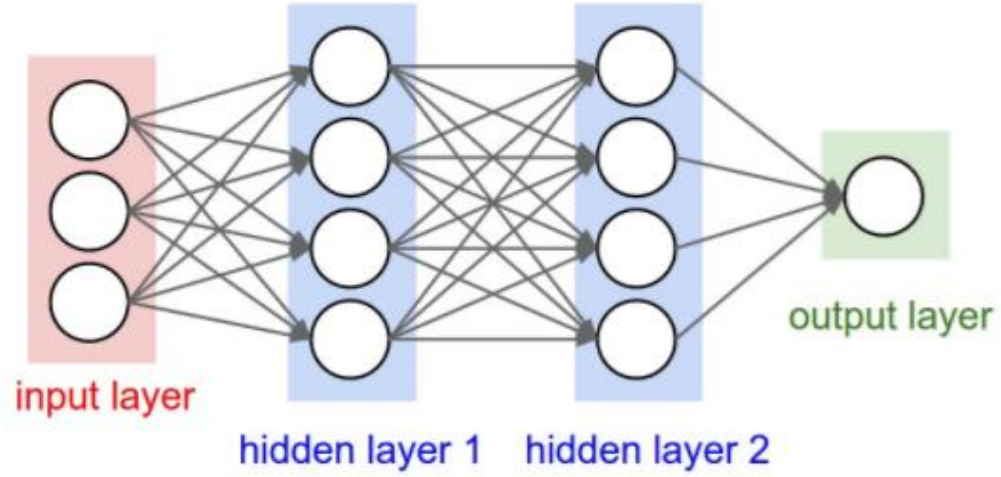
Small perturbations added to normal inputs can easily fool a deep neural network.

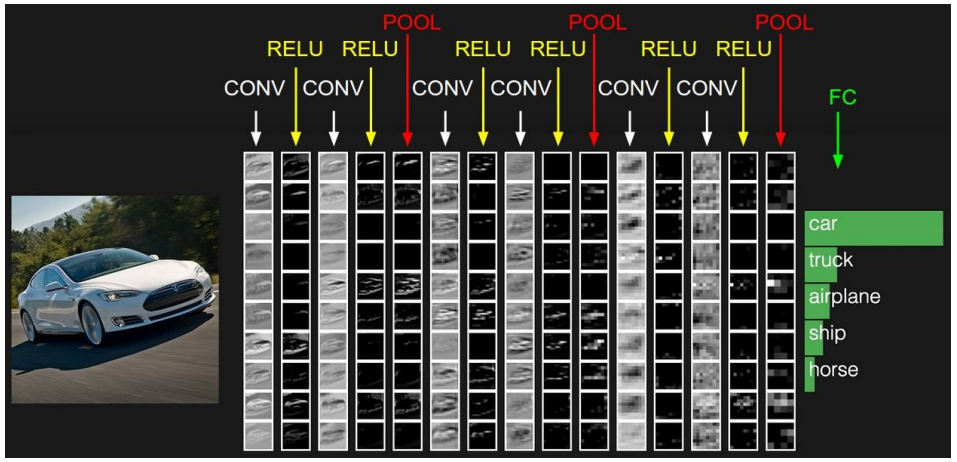


- Perturbations are small, imperceptible to human eyes.
- All networks are vulnerable to adversarial examples.
- Adversarial examples transfer across models.

Deep Neural Network Transformations:

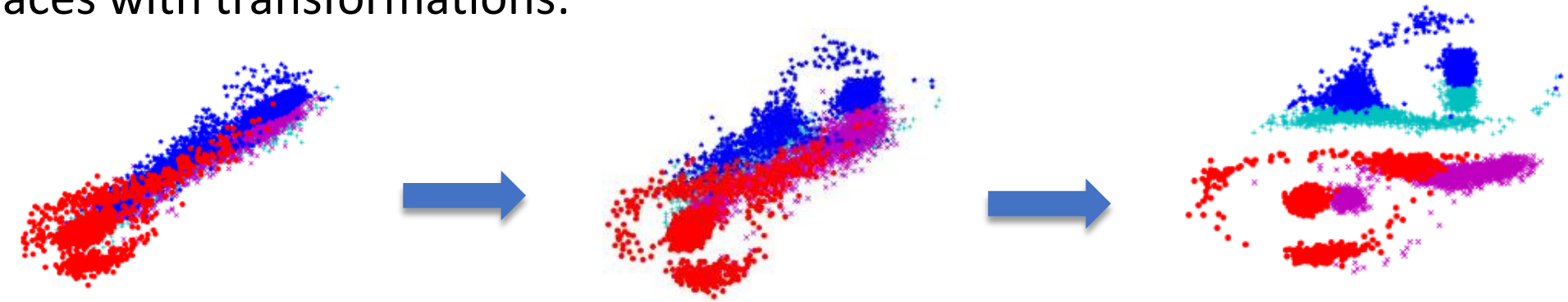
Neural network:





Krizhevsky et al. 2012, Karen et al. 2013

Evolving spaces with transformations:



1st layer

10th layer

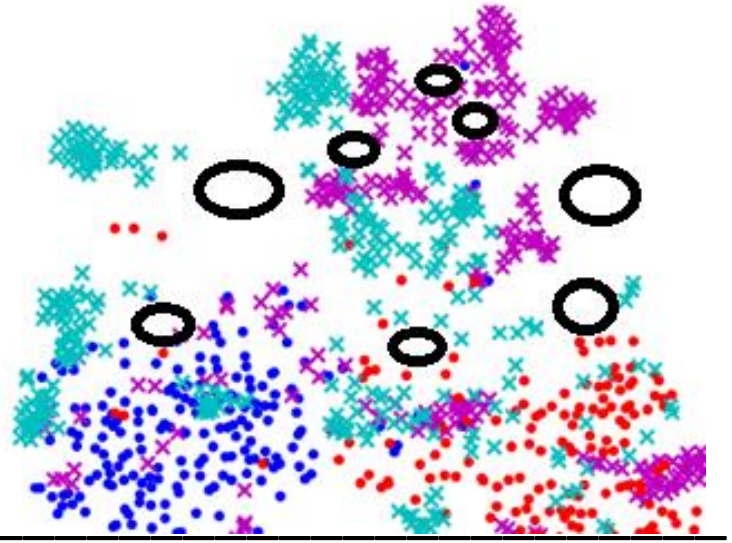
20th layer

Understanding Adversarial Subspaces:

1. Non-linear explanation (Szegedy et al. 2013):

Non-linear transformations leads to the existence of small "pockets" in the deep space:

- Regions of low probability (not naturally occurring).
- Densely scattered regions.
- Continuous regions.
- Close to normal data subspace.



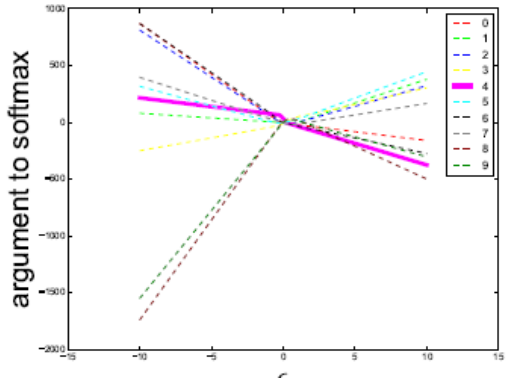
---

2. Linear explanation (Goodfellow et al. 2014):

Adversarial subspaces span a contiguous multidimensional space:

- Small changes at individual dimensions can sum up to significant change in final output: $\sum_{i=0}^{n} x_i + \epsilon$.
- Adversarial examples can always be found if ϵ is large enough.

$$w^T x + b$$



Adversarial Attack:

Given input (x, y) and a target class l, the attack generates a new example x_{adv}, so as to:

$$\text{minimize } \|x - x_{adv}\|_p$$
$$\text{subject to } f(x_{adv}) \neq f(x) \text{ or } f(x_{adv}) = l$$

- Fast Gradient Method (FGM) (*Goodfellow et al. 2014*):

$$x_{adv} = x + \varepsilon \operatorname{sign} \nabla_x J_\theta(x, y).$$

- Basic Iterative Method (BIM), an iterative version of FGM (*Kurakin et al. 2016*):

$$x_{adv}^0 = x, x_{adv}^i = x_{adv}^{i-1} + \varepsilon \operatorname{sign} \nabla_x J_\theta\left(x_{adv}^{i-1}, y\right).$$

- Jacobian-based Saliency Map Attack (JSMA) (*Papernot et al. 2016*).

- Optimization Based Attack (Opt.) (*Carlini & Wagner 2017, Liu et al. 2016*):

$$\delta = \frac{1}{2}(\tanh(\mathrm{w}) + 1) - x, \min(\|\delta\|_2^2) + c \cdot f(x + \sigma).$$

Adversarial Subspaces & Dimensional Escape:

Adversarial Subspace: local subspace surrounding an adversarial example.

Question: What are the characteristics of adversarial subspaces?

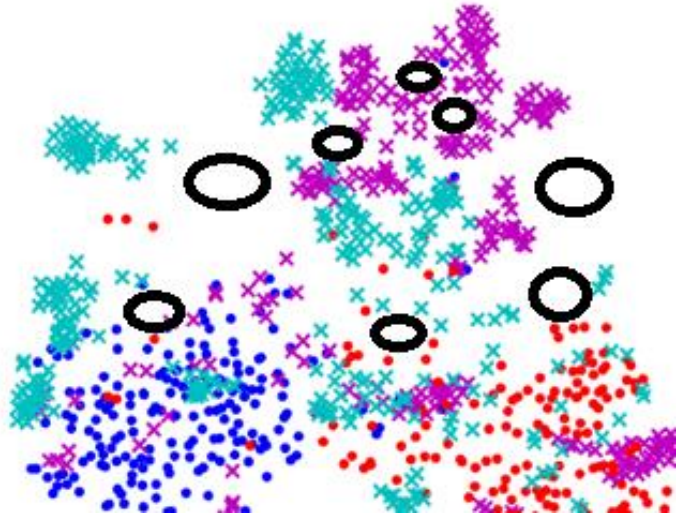
What we do know about adversarial subspaces?

- Low probability and continuous regions, close to normal data submanifold, …

Intuitively:

- Close in distance, yet semantically far from original data subspace.
- "Escape" to adversarial subspace with only a small perturbation.

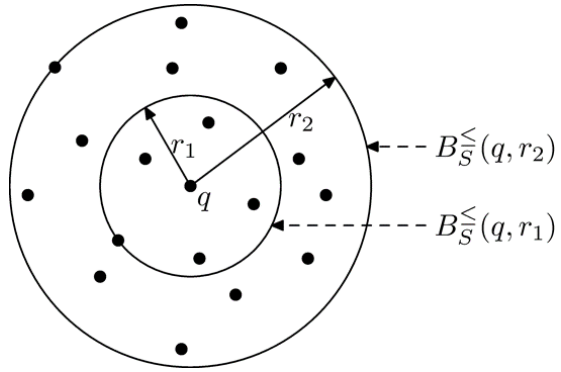
Dimensional Escape!



Dimensionality and Adversarial Subspaces:

Intuition:

Adversarial subspaces are of high intrinsic dimensionality.



Dimensional Escape:

- Density will change
- Uncertainty will change
- Fundamentally, dimensionality increases!
- Revealed using estimators of intrinsic dimensionality.

Expansion Dimension:

- Two balls of differing radii r_1 and r_2, dimension m can be deduced from ratios of volumes:

$$\frac{\mathrm{V}_2}{V_1} = \left(\frac{r_2}{r_1}\right)^m \Rightarrow m = \frac{\ln(V_2/V_1)}{\ln(r_2/r_1)}$$

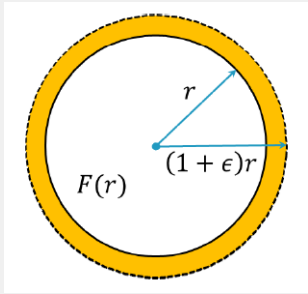
- Related to the Expansion Dimension (*Karger and Ruhl 2002, Houle et al. 2012*)
- V_1 and V_2 estimated by the numbers of points contained in the two balls.

Local Intrinsic Dimensionality (LID):

Definition (Local Intrinsic Dimensionality)

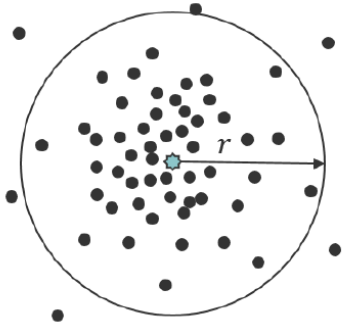
Given a data sample $x \in X$, let $r > 0$ be a random variable denoting the distance from x to other data samples. The *local intrinsic dimension* of x at distance r is

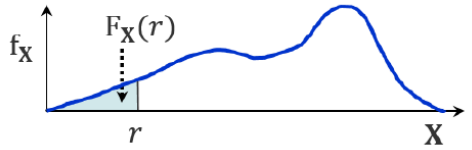
$$\mathrm{LID}_F(r) \triangleq \lim_{\epsilon \to 0^+} \frac{\ln(F((1+\epsilon)\cdot r)/F(r))}{\ln(1+\epsilon)} = \frac{r \cdot F'(r)}{F(r)},$$

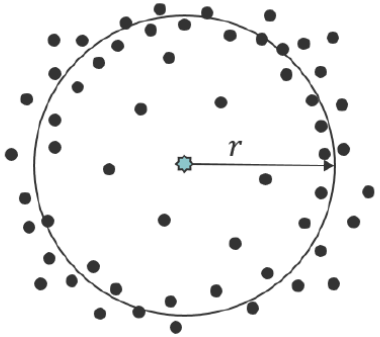


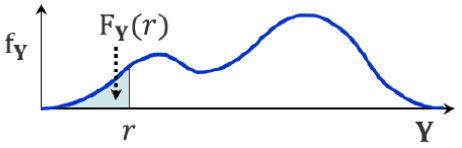
wherever the limit exists.

- $F(r)$: cumulative distribution function.
- $F(r)$ is analogous to volume V in Euclidean space, where r_1 and r_2 can be allowed to tend to a single value r.









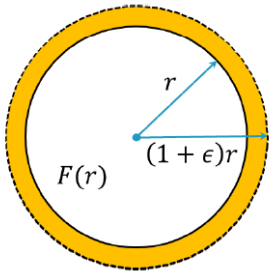
Estimation of LID:

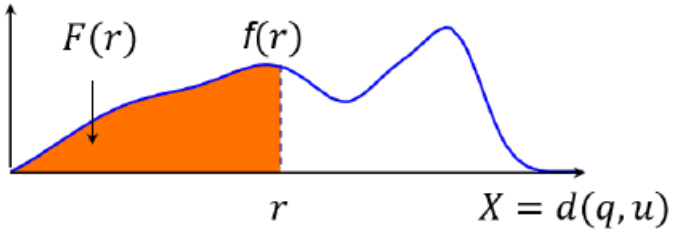
Estimators of LID already available:

- Hill (MLE) estimator (*Hill 1975, Amsaleg et al. 2015*):

$$\widehat{\mathrm{LID}}(x) = -\left(\frac{1}{k}\sum_{i=1}^{k}\log\frac{r_i(x)}{r_k(x)}\right)^{-1},$$

r_i is the distance of x to its i^{th} nearest neighbor.

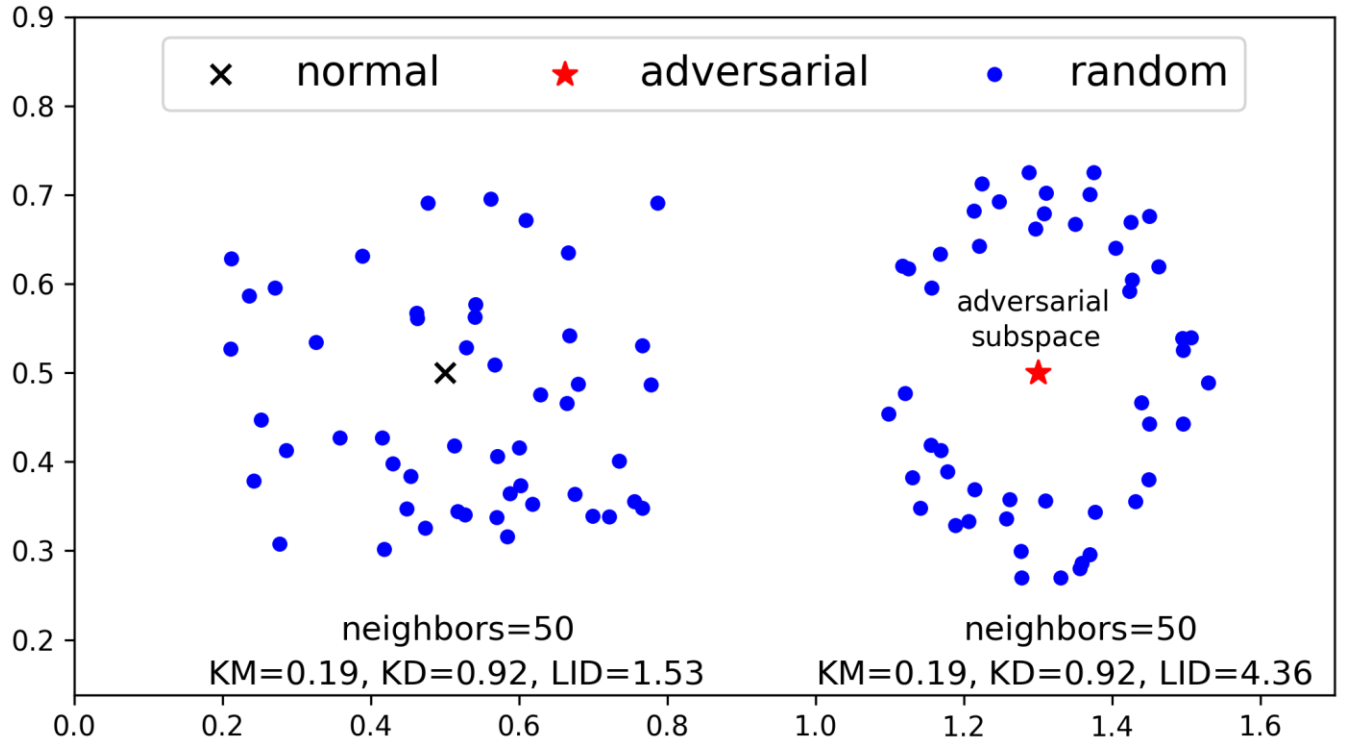




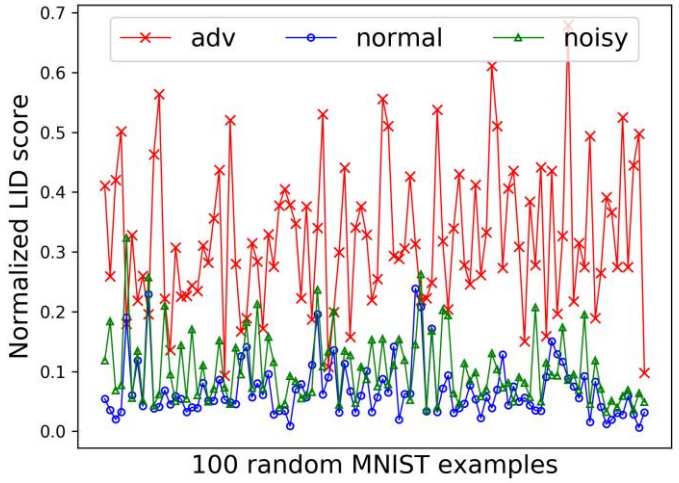
- Based on Extreme Value Theory:
  - Nearest neighbor distances are extreme events.
  - Lower tail distribution follows Generalized Pareto Distribution (GPD).
- Other estimators: e.g. *Amsaleg et al. 2015, Gomes et al. 2008*.

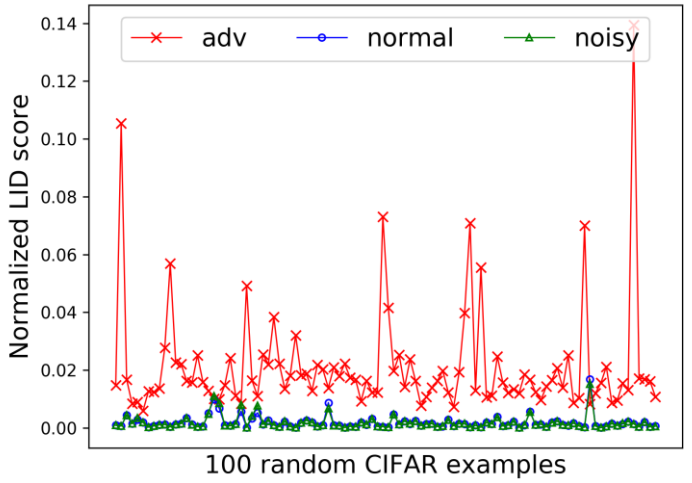
Interpretation of LID for Adversarial Subspaces:

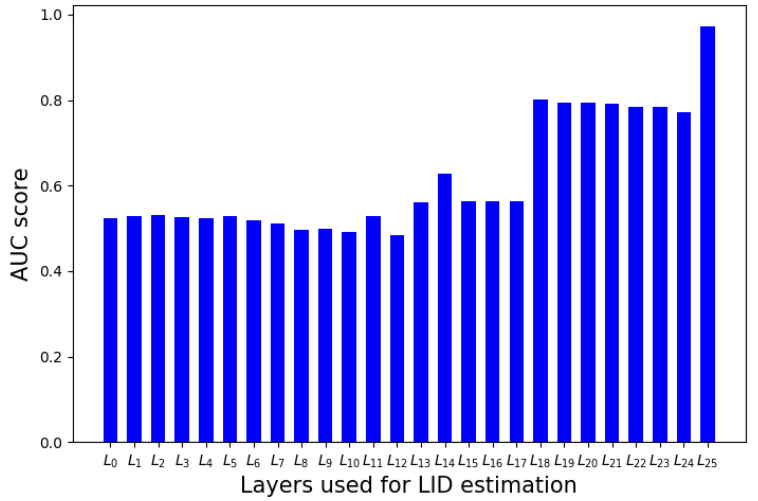
- LID directly measures expansion rate of local distance distributions.
- The expansion of adversarial subspace is higher than normal data subspace.
- Adversarial examples are dimensionally outliers.



LID of Adversarial Subspaces:







We found:

- Higher dimensionality: In general, adversarial examples are of higher dimensionality (LID).
- Minibatch efficiency: Estimation of LID within a **minibatch** of 100 samples can help discriminate adversarial examples with high success rate, although larger batch size leads to further improvement.
- Deeper layers: LID difference is more pronounced at deeper layers.

Previous Work on Characterization of Adversarial Examples:

- Kernel Density (KD) (*Feinman et al. 2017)*

$$KD(x) = \frac{1}{|X_c|}\sum_{s \in X_c} \exp\left(\frac{|F^{n-1}(x) - F^{n-1}(s)|^2}{\sigma^2}\right),$$

X_c: the set of samples in class c;
$F^{n-1}(x)$: the final hidden layer output.

- Bayesian Uncertainty (BU) (*Feinman et al. 2017)*

$$BU(x) = \left(\frac{1}{L}\sum_{r=1}^{L} \|F_r(x)\|\right) - \left\|\frac{1}{L}\sum_{r=1}^{L} F_r(x)\right\|,$$

F_r: dropout randomized network;
L: the number of randomization.

Experiments & Results:

Table 1: A comparison of the discrimination power (AUC score (%) of a logistic regression classifier) among LID, Kernel Density (KD), Bayesian Uncertainty (BU). AUC is computed for 5 attack strategies on 3 datasets. Best results are in **bold**.

Dataset	Feature	FGM	BIM-a	BIM-b	JSMA	Opt
MNIST	KD	78.12	98.14	98.61	68.77	95.15
	BU	32.37	91.55	25.46	88.74	71.30
	KD+BU	82.43	99.20	98.81	90.12	95.35
	LID	**96.89**	**99.60**	**99.83**	**92.24**	**99.24**
CIFAR-10	KD	64.92	68.38	98.70	85.77	91.35
	BU	70.53	81.60	97.32	87.36	91.39
	KD+BU	70.40	81.33	98.90	88.91	93.77
	LID	**82.38**	**82.51**	**99.78**	**95.87**	**98.94**
SVHN	KD	70.39	77.18	99.57	86.46	87.41
	BU	86.78	84.07	86.93	91.33	87.13
	KD+BU	86.86	83.63	99.52	93.19	90.66
	LID	**97.61**	**87.55**	**99.72**	**95.07**	**97.60**

LID characteristics can help discriminate adversarial examples residing in the adversarial subspace.

Experiments & Results:

Table 2: AUC (%) is computed for a logistic regression classifier trained on features (KD, BU, LID) of FGM attack, then tested on other forms of attacks (BIM-a, BIM-b, JSMA and Opt). The best results are in **bold**.

Train \ Test attack		FGM	BIM-a	BIM-b	JSMA	Opt
FGM	KD	64.92	69.15	89.71	85.72	91.22
	BU	70.53	81.67	2.65	86.79	91.27
	LID	**82.38**	**82.30**	**91.61**	**89.93**	**93.32**

LID characteristics of simple attacks can help to discriminate other attacks.

Integrating LID into the Adversarial Objective:

$$\text{minimize } \underbrace{\|x_{adv} - x\|_2^2}_{\text{Small perturbation}} + \alpha \cdot (\underbrace{\ell(x_{adv})}_{\text{Change of class}} + \underbrace{\ell(\text{LID}(x_{adv}))}_{\text{Low LID}})$$

Table 3: The failure rate (%) of an adaptive attack targeting low intrinsic dimensionality (LID score).

	MNIST	CIFAR-10	SVHN
Scenario 1 (low LID at all layers): Attack Failure Rate	100	100	100
Scenario 2 (low LID at one layer): Attack Failure Rate	100	95.7	97.2

It is difficult to explicitly generate adversarial examples residing in subspaces with low intrinsic dimensionality.

Conclusion:

- We characterize the dimensional characteristics of adversarial subspaces.
- Adversarial subspaces tend to possess higher intrinsic dimensionality than normal data subspaces.
- The dimensional characteristics (LID) can be leveraged to recognize adversarial examples residing in adversarial subspaces.

Future Work:

- Dimensionality-driven adversarial defense.
- Better estimation of LID: Larger batch size / neighborhood size.
- Better understanding of DNNs in terms of intrinsic dimensionality.

References:

Michael E. Houle. Local intrinsic dimensionality I: an extreme-value-theoretic foundation for similarity applications. In SISAP, pp. 64–79, 2017a.

Michael E. Houle. Local intrinsic dimensionality II: multivariate analysis and distributional support. In SISAP, pp. 80–95, 2017b.

Laurent Amsaleg, Oussama Chelly, Teddy Furon, Stéphane Girard, Michael E. Houle, Ken-ichi Kawarabayashi, and Michael Nett. Estimating local intrinsic dimensionality. In SIGKDD, pp.29–38. ACM, 2015.

Laurent Amsaleg, James Bailey, Dominique Barbe, Sarah Erfani, Michael E. Houle, Vinh Nguyen, and Miloš Radovanović. The vulnerability of learning to adversarial perturbation increases with intrinsic dimensionality. In WIFS, 2017.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In S&P, pp. 582–597, 2016d.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

Thank you!

Poster: Tue May 1st **11:00 AM -- 01:00 PM** @ East Meeting level; 1,2,3 **#41**

Limitations of LID for detection:

- We are focusing on characterizing adversarial subspaces instead of proposing a perfect detection method.
- We run additional experiments with DNNs trained using batch normalization, and tested the discrimination power of LID on attacks (in **bold**) analysed in *Lu et all. 2018* and *Athalye et all. 2018*.

Dataset	%	FGM	BIM	**PGD**	Deepfool	EAD-0	**EAD-40**	Opt-0	**Opt-40**
CIFAR-10	AUC	88.55	95.28	**94.45**	98.78	98.85	**98.82**	98.75	**98.45**
	Accuracy	80.89	87.74	**86.80**	95.98	93.23	**94.58**	95.61	**94.02**
	Precision	82.21	77.55	**77.10**	95.98	94.25	**95.45**	95.75	**94.42**
	Recall	80.10	88.98	**85.92**	96.20	92.45	**93.91**	95.70	**96.48**

We found that the LID will be more robust in regularized space (batch normalization), and it can be used to distinguish high confidence (40) attacks.

There are still space to improve LID based analysis to better understand adversarial spaces.