HOW STRUCTURE OF THE DATA CAN IMPROVE OUR UNDERSTANDING AND USAGE OF DEEP LEARNING?

RAJA GIRYES

TEL AVIV UNIVERSITY

IMVC

March 6, 2018

AGENDA

- Deep learning impact.
- A sample of existing theory for deep learning.
- Data structure based theory for deep learning
  - Neural networks with random Gaussian weights.
  - Generalization error of deep neural networks.
  - Deep learning as metric learning.
  - Solving minimization problems via deep learning.

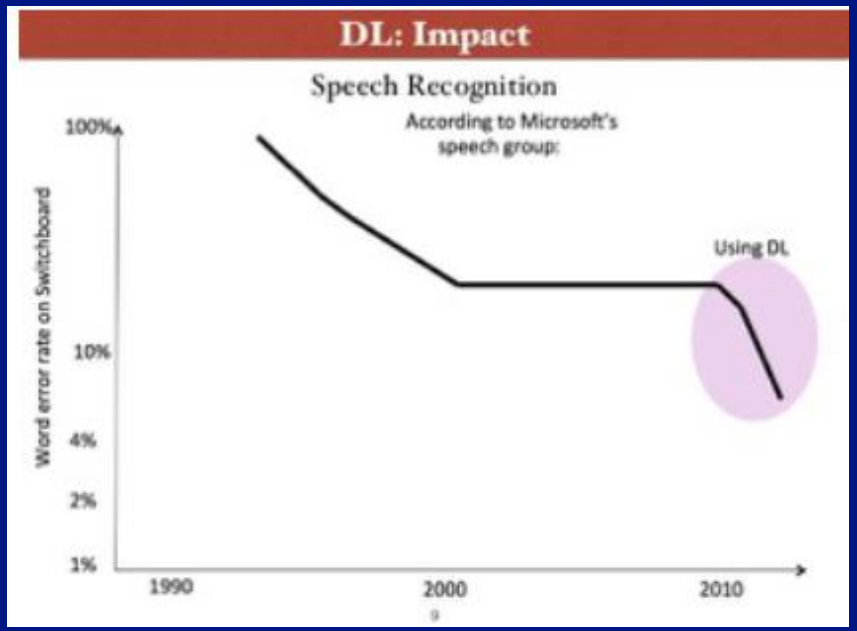
DEEP LEARNING IMPACT

DEEP LEARNING IMPACT

- Imagenet dataset
- 1,400,000 images
- 1000 categories
- 150000 for testing,
- 50000 for validation

Model	Top-1 (val)	Top-5 (val)	Top-5 (test)
SIFT + FVs [7]	—	—	*26.2%*
1 CNN	40.7%	18.2%	—
5 CNNs	38.1%	16.4%	**16.4%**
1 CNN*	39.0%	16.6%	—
7 CNNs*	36.7%	15.4%	**15.3%**

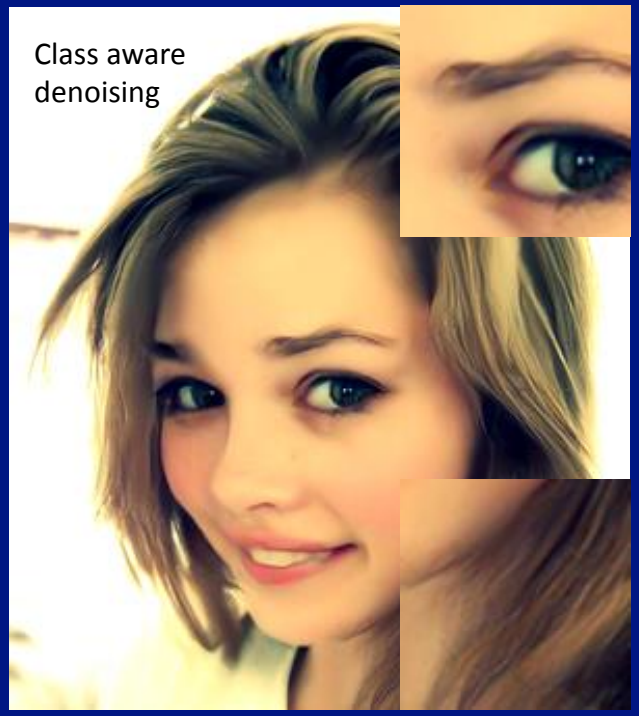
Today we get 3.5% by 152 layers

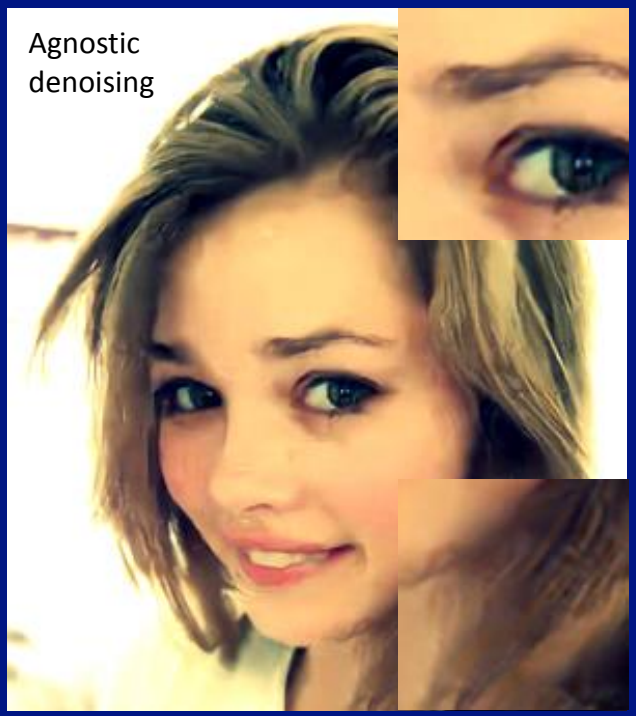


CUTTING EDGE PERFORMANCE IN MANY OTHER APPLICATIONS

- Disease diagnosis [Zhou, Greenspan & Shen, 2016].
- Language translation [Sutskever et al., 2014].
- Video classification [Karpathy et al., 2014].
- Handwriting recognition [Poznanski & Wolf, 2016].
- Sentiment classification [Socher et al., 2013].
- Image denoising [Remez et al., 2017].
- Depth Reconstruction [Haim et al., 2017].
- Super-resolution [Kim et al., 2016], [Bruna et al., 2016].
- Error correcting codes [Nahmani, 2016]
- many other applications…

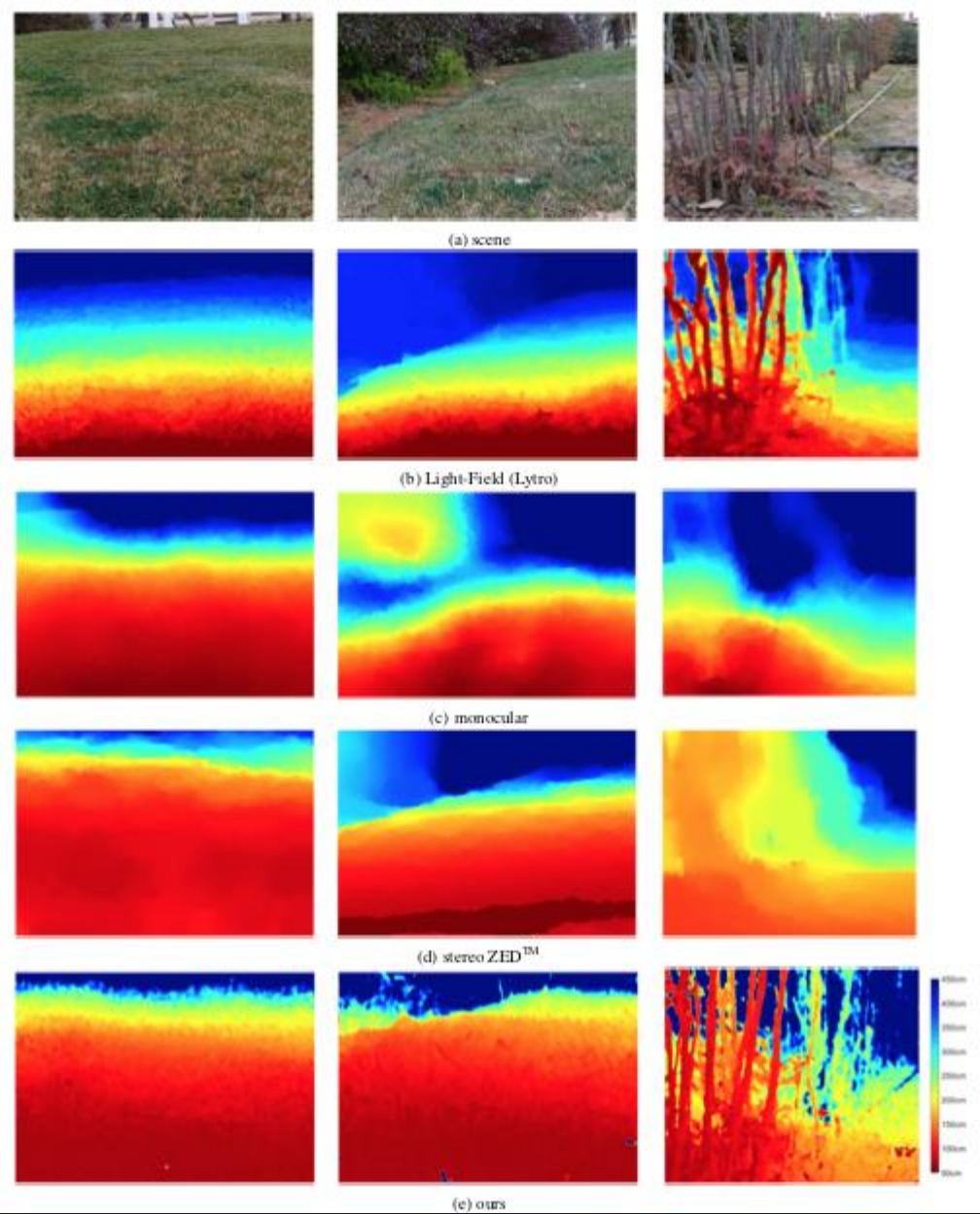
CLASS AWARE DENOISING





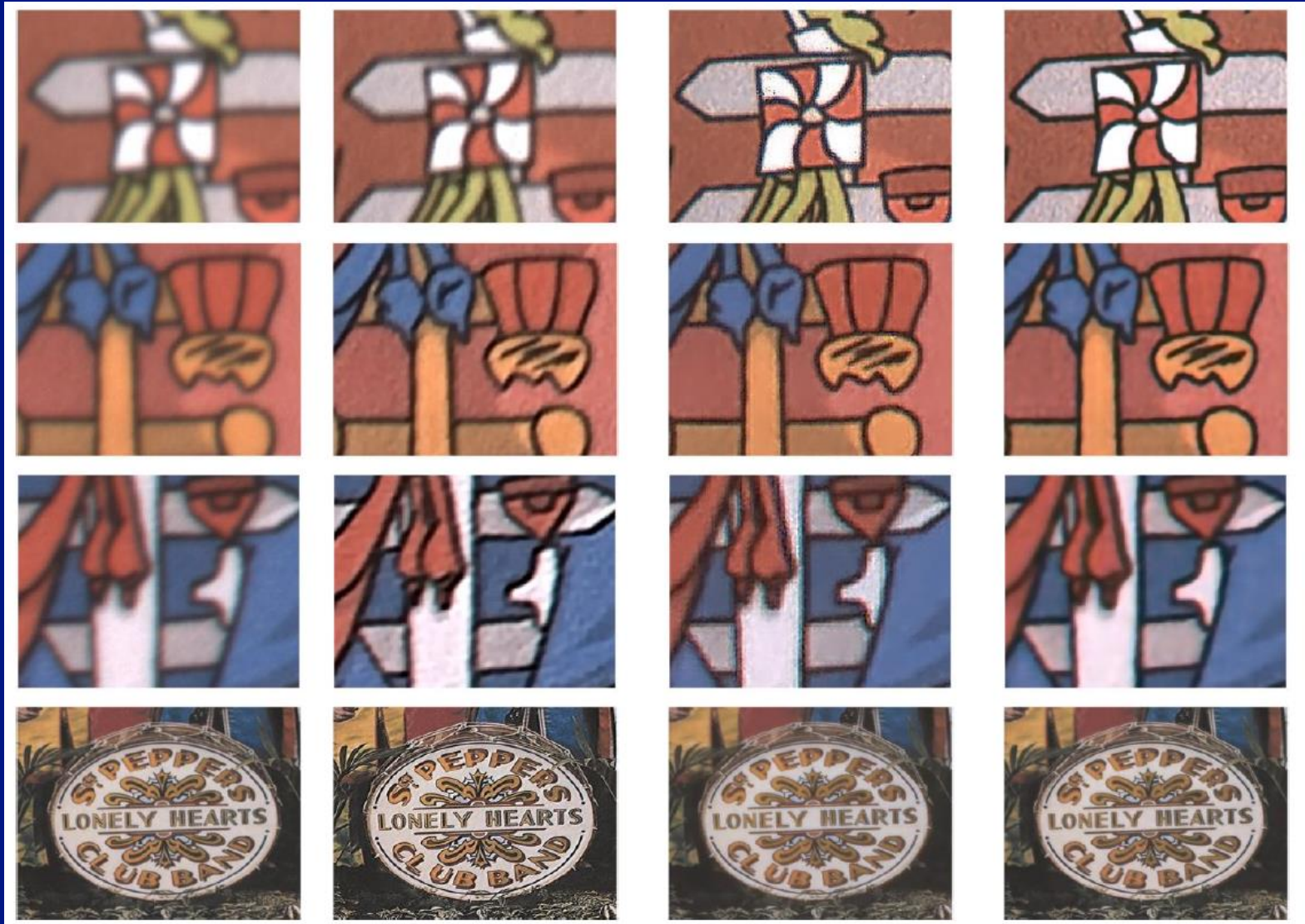
[Remez, Litani, Giryes, Bronstein, 2017]

DEPTH ESTIMATION BY PHASE CODED CUES



[Haim, Elmalem, Bronstein, Marom, Giryes, 2017]

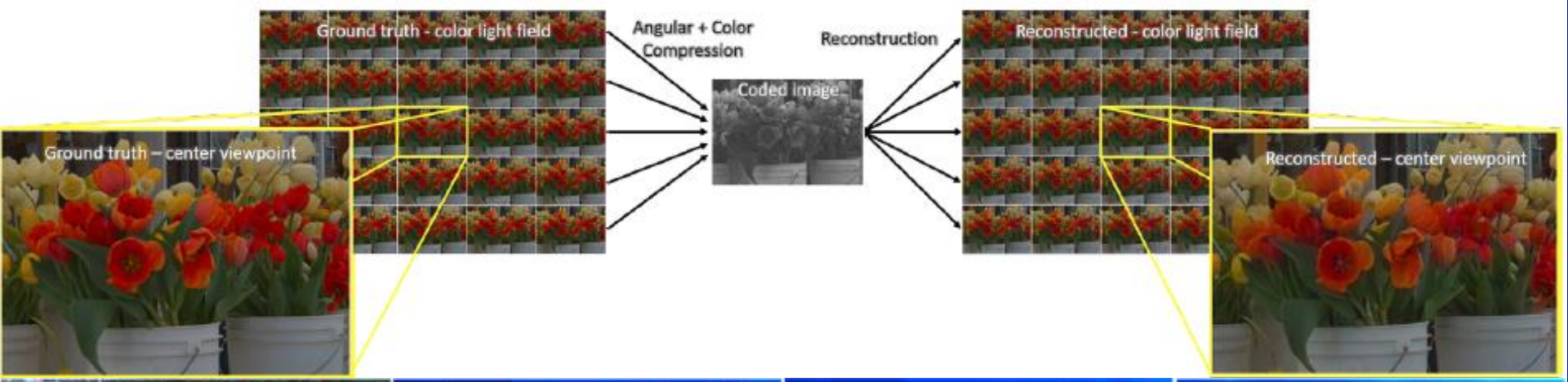
ALL-IN-FOCUS BY PHASE CODED CUES

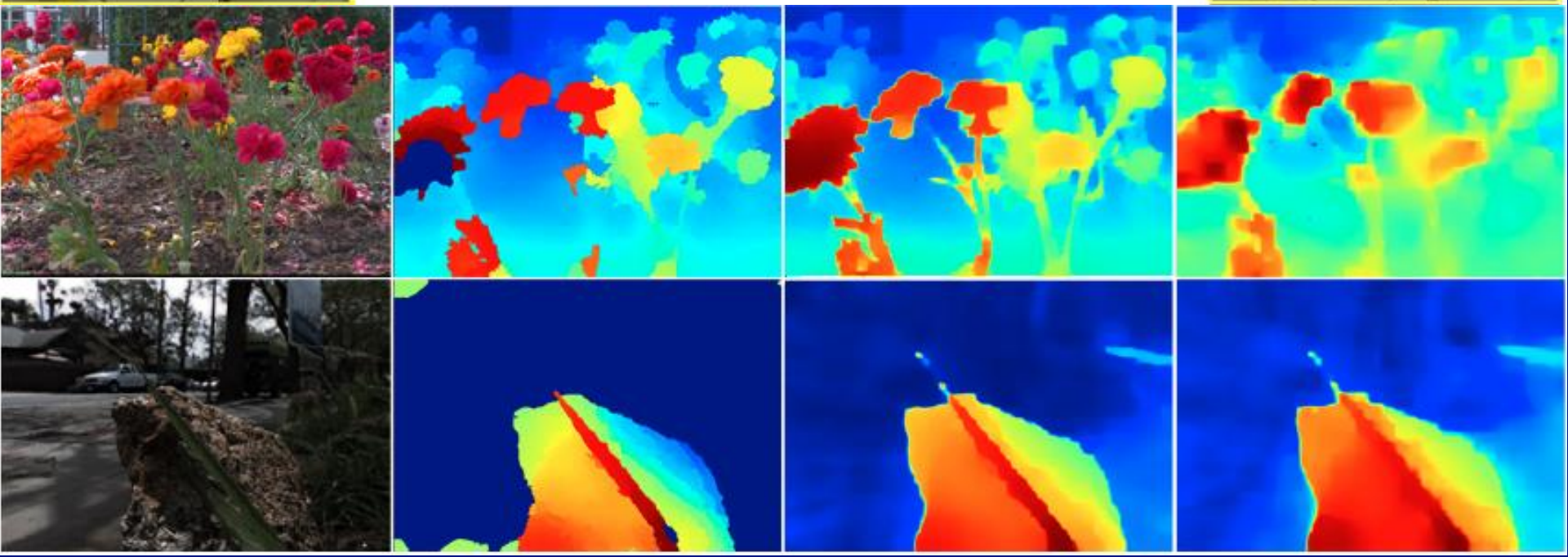


(a) Clear aperture imaging (b) Krishnan [17] on (a) (c) Our mask with [14] processing (d) Ours

[Elmalem, Marom, Giryes, 2018]

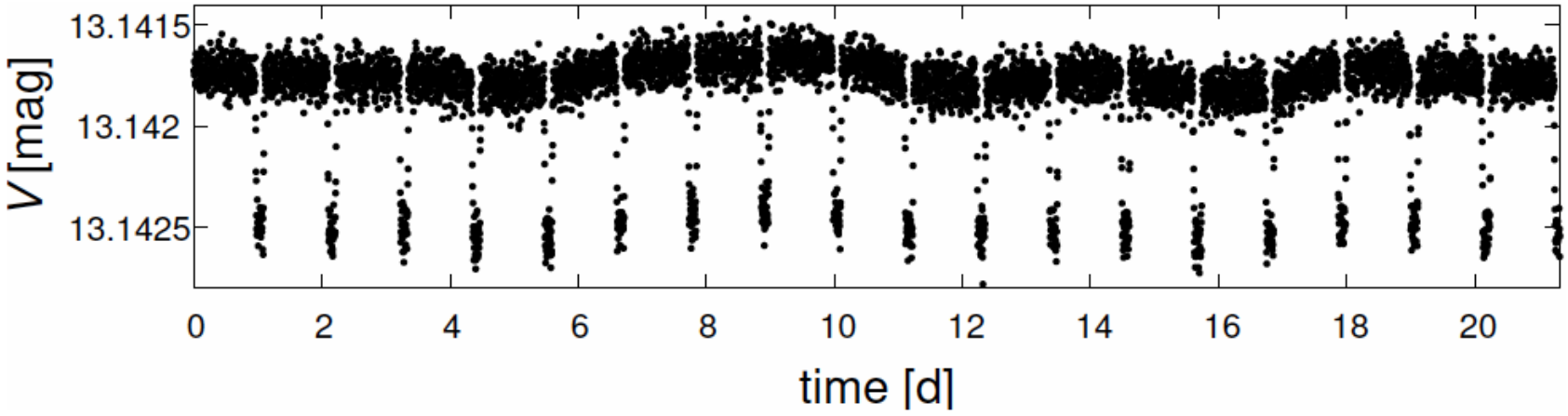
COMPRESSED COLOR LIGHT FIELD



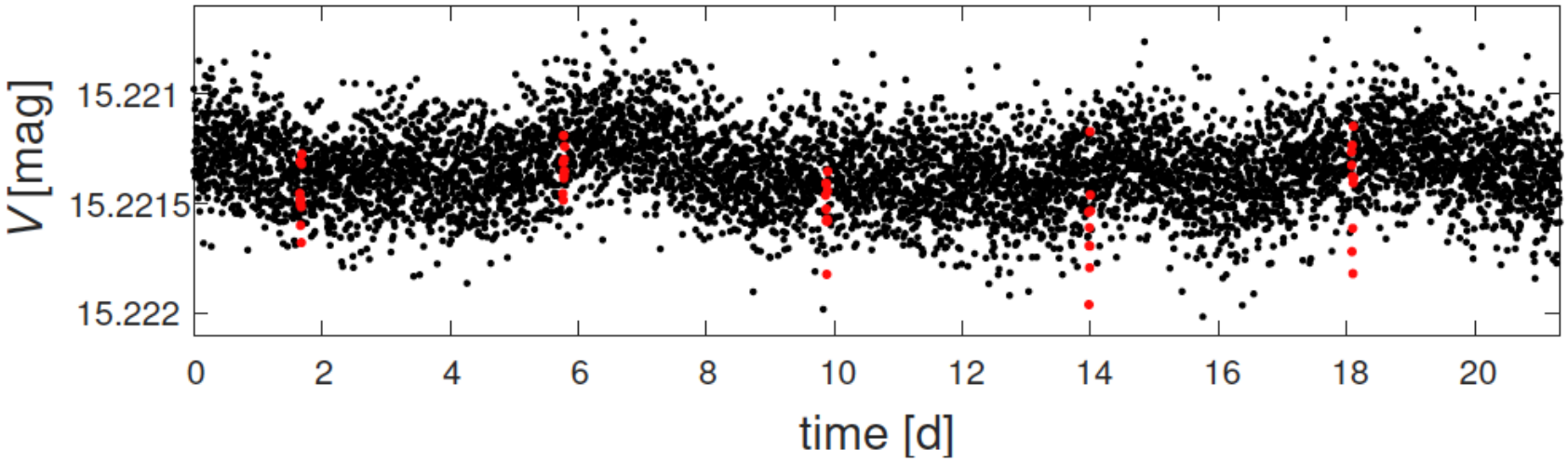


[Nabati, Mendelovic, Giryes, 2017]

EXOPLANETS DETECTION

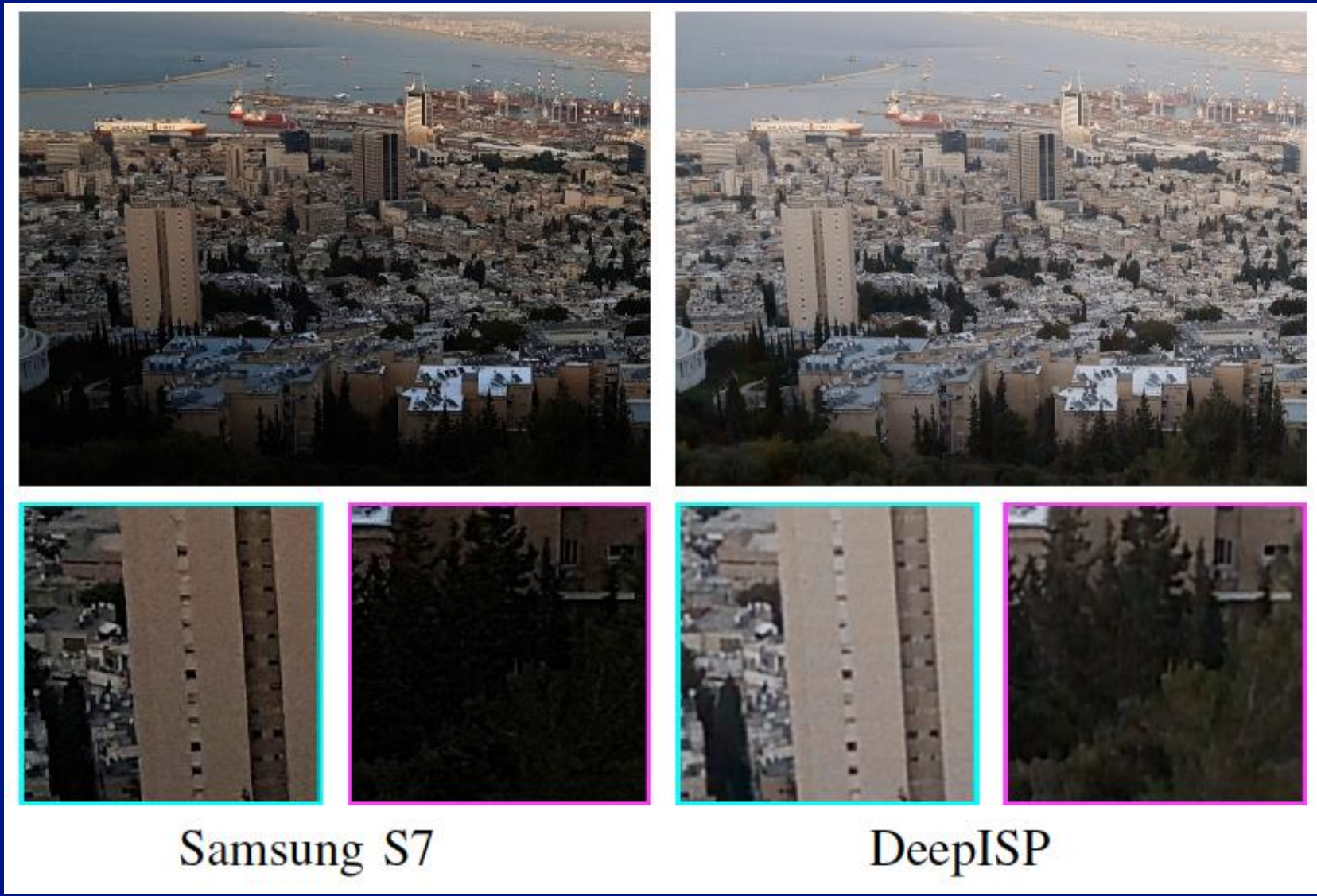


(a)



[Zucker and Giryes, 2018]

DEEP ISP



[E. Schwartz, R. Giryes and A. M. Bronstein, 2018]

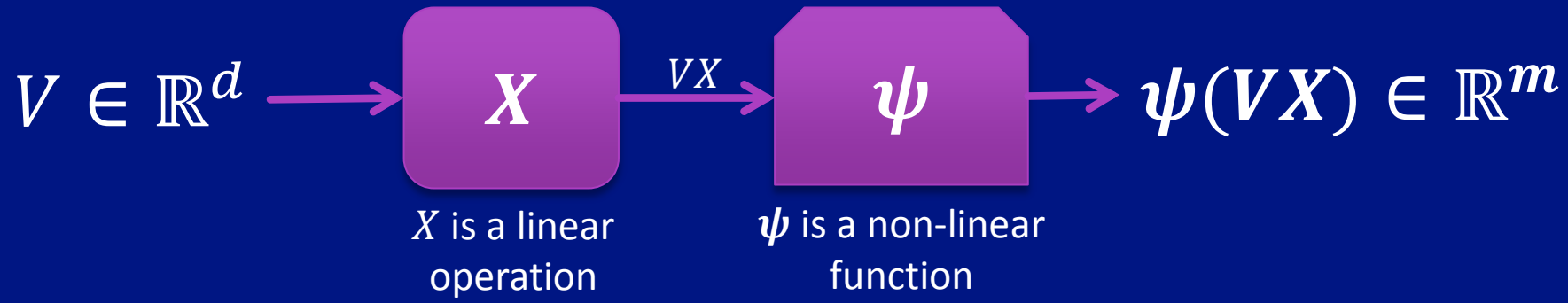
WHY THINGS WORK BETTER TODAY?

- More data – larger datasets, more access (internet)
- Better hardware (GPU)
- Better learning regularization (dropout)

- Deep learning impact and success is not unique only to image classification.
- But it is still unclear why deep neural networks are so remarkably successful and how they are doing it.

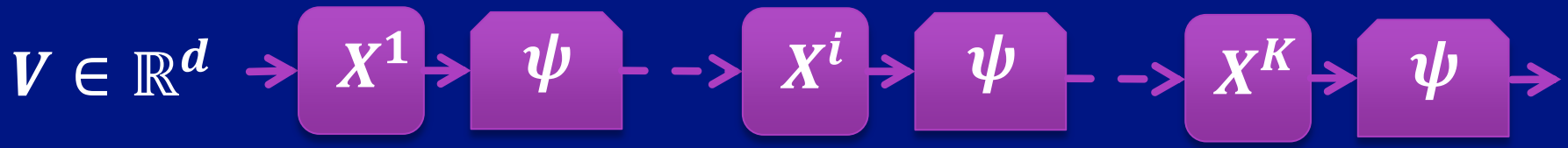
DEEP NEURAL NETWORKS (DNN)

- One layer of a neural net

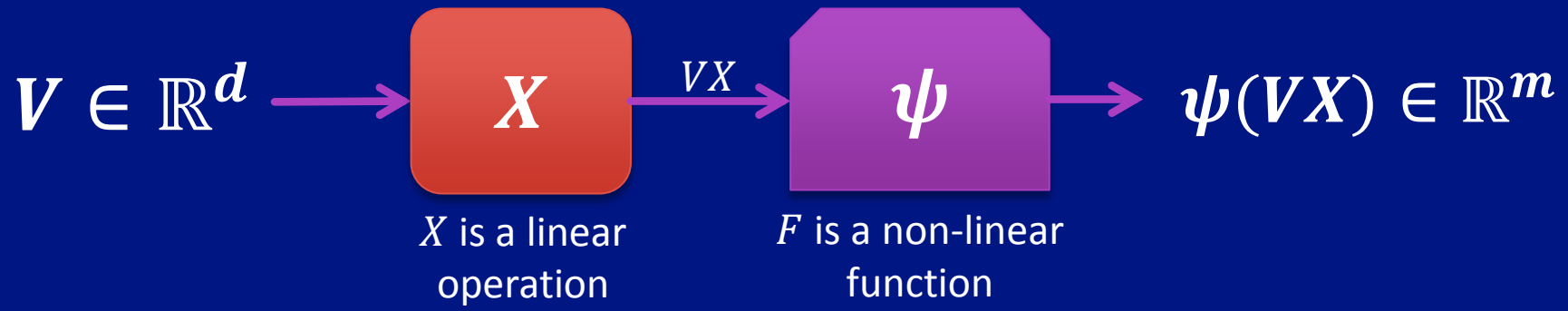


- Concatenation of the layers creates the whole net

$$\Phi(X^1, X^2, \dots, X^K) = \psi(\psi(\psi(VX^1)X^2) \dots X^K)$$



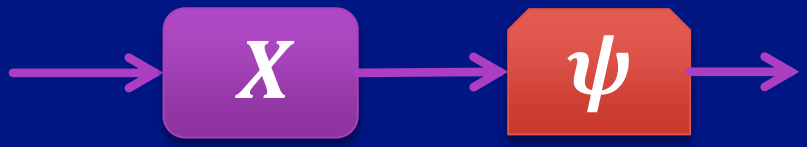
CONVOLUTIONAL NEURAL NETWORKS (CNN)



- In many cases, X is selected to be a convolution.
- This operator is shift invariant.
- CNN are commonly used with images as they are typically shift invariant.

THE NON-LINEAR PART

- Usually $\psi = g \circ f$.



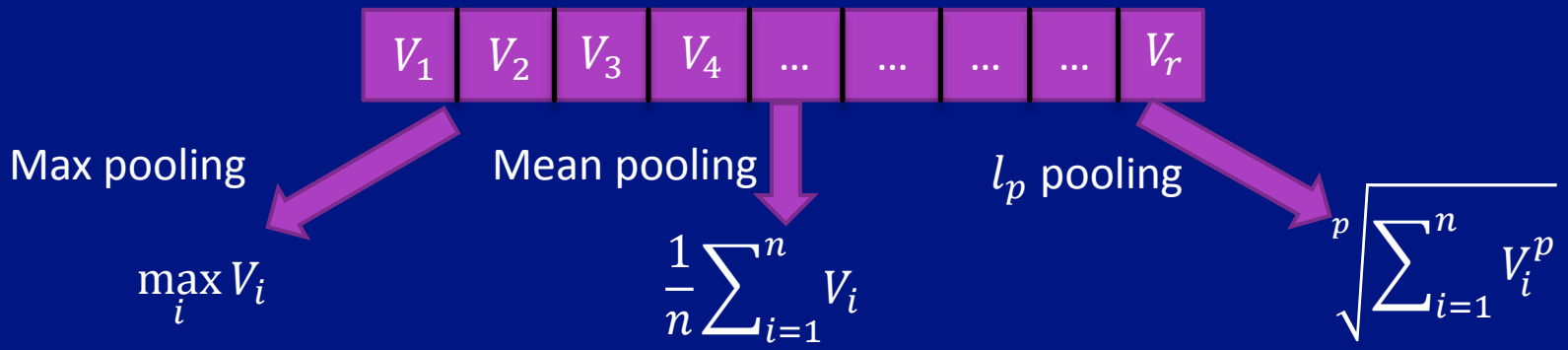
- f is the (point-wise) activation function

ReLU
$f(x) = \max(x, 0)$

Sigmoid
$f(x) = \frac{1}{1 + e^{-x}}$

Hyperbolic tangent
$f(x) = \tanh(x)$

- g is a pooling or an aggregation operator.



A SAMPLE OF EXISTING THEORY FOR DEEP LEARNING

WHY DNN WORK?

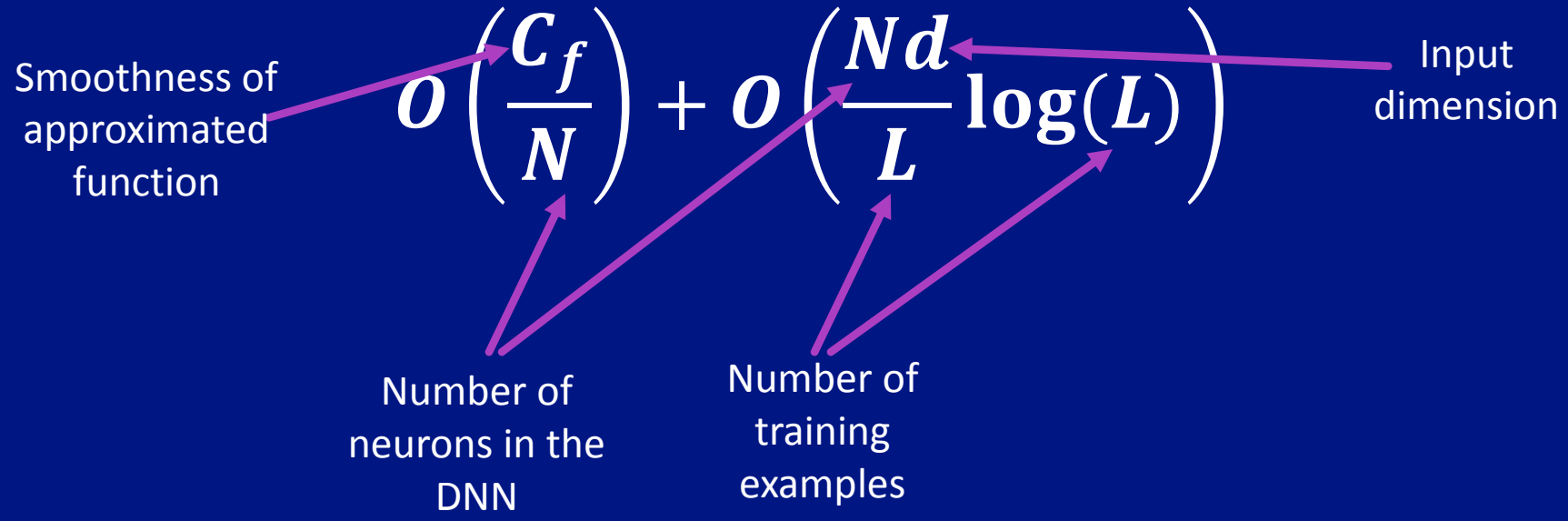

What is so special with the DNN structure?
What is the role of the depth of DNN?
What is the capability of DNN?
How many training samples do we need?
What is the role of pooling?
What is the role of the activation function?
What happens to the data throughout the layers?

REPRESENTATION POWER

- Neural nets serve as a universal approximation for any measurable Borel functions [Cybenko 1989, Hornik 1991].
- In particular, let the non-linearity ψ be a bounded, non-constant continuous function, I_d be the d-dimensional hypercube, and $C(I_d)$ be the space of continuous functions on I_d. Then for any $f \in C(I_d)$ and $\epsilon > 0$, there exists $m > 0$, and $X \in \mathbb{R}^{d \times m}$, $B \in \mathbb{R}^m$, $W \in \mathbb{R}^m$ such that the neural network
$$F(V) = \psi(VX + B)W^T$$
approximates f with a precision ϵ:
$$|F(V) - f(V)| < \epsilon, \forall V \in \mathbb{R}^d$$

ESTIMATION ERROR

- The estimation error of a function f by a neural networks scales as [Barron 1994].



DEPTH OF THE NETWORK

- Depth allow representing shallow restricted Boltzmann machines, which has an exponential number of parameters, compared to the deep one [Montúfar & Morton, 2015]
- Each DNN layer with ReLU divides the space by a hyper-plane, folding one part of it.
- Thus, the depth of the network folds the space into an exponential number of sets compared to the number of parameters [Montúfar, Pascanu, Cho & Bengio, 2014]

DEPTH EFFICIENCY OF CNN

- Function realized by CNN, with ReLU and max-pooling, of polynomial size requires super-polynomial size for being approximated by shallow network [Telgarsky 2016 ,Cohen et al., 2016].
- Standard convolutional network design has learning bias towards statistics of natural images [Cohen et al., 2016].

ROLE OF POOLING

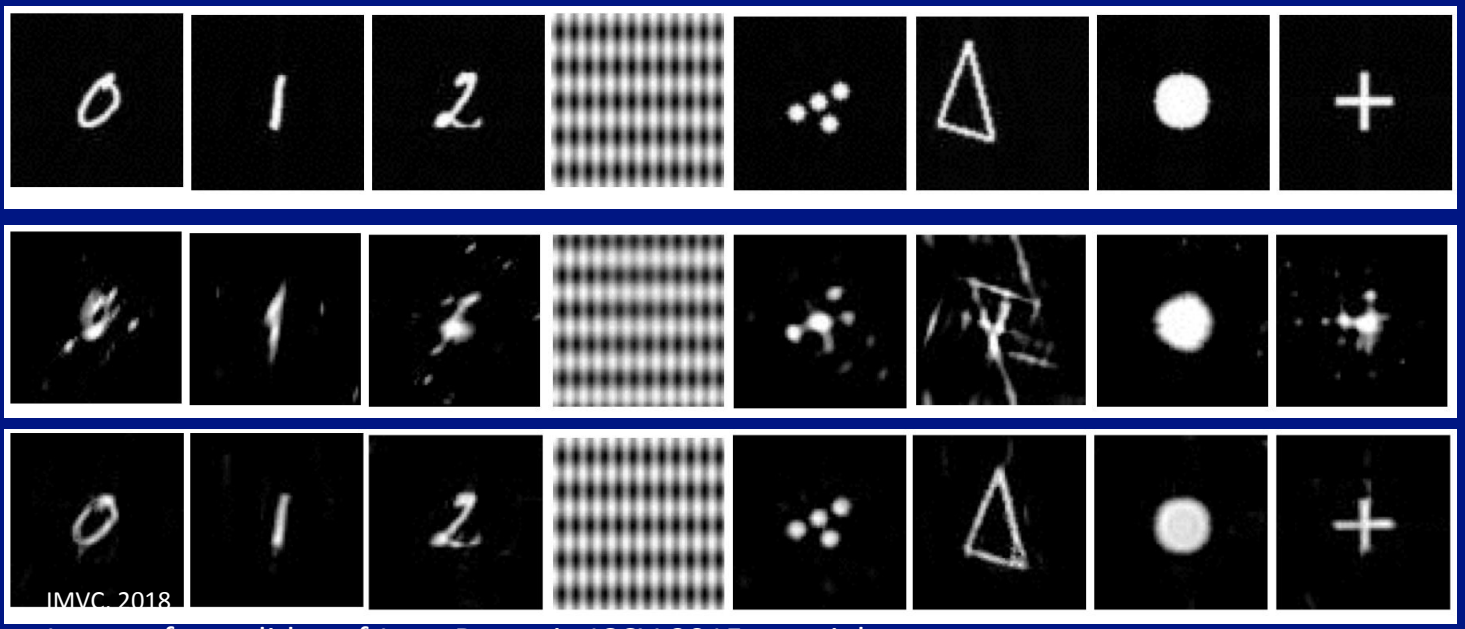
- The pooling stage provides shift invariance [Boureau et al. 2010], [Bruna, LeCun & Szlam, 2013].
- A connection is drawn between the pooling stage and the phase retrieval methods [Bruna, Szlam & LeCun, 2014].
- This allows calculating Lipchitz constants of each DNN layer $\psi(\cdot\, X)$ and empirically recovering the input of a layer from its output.
- However, the Lipchitz constants calculated are very loose and no theoretical guarantees are given for the recovery.

SUFFICIENT STATISTIC AND INVARIANCE

- Given a certain task at hand:
- Minimal sufficient statistic guarantees that we can replace raw data with a representation with smallest complexity and no performance loss.
- Invariance guarantees that the statistic is constant with respect to uninformative transformations of the data.
- CNN are shown to have these properties for many tasks [Soatto & Chiuso, 2016].
- Good structures of deep networks can generate representations that are good for learning with a small number of examples [Anselmi et al., 2016].

SCATTERING TRANSFORMS

- Scattering transform - a cascade of wavelet transform convolutions with nonlinear modulus and averaging operators.
- Scattering coefficients are stable encodings of geometry and texture [Bruna & Mallat, 2013]



Original image with d pixels

Recovery from first scattering moments: $O(\log d)$ coefficients

Recovery from 1st & 2nd scattering moments: $O(\log^2 d)$ coefficients

Images from slides of Joan Bruna in ICCV 2015 tutorial

SCATTERING TRANSFORMS AND DNN

- More layers create features that can be made invariant to increasingly more complex deformations.
- Deep layers in DNN encode complex, class-specific geometry.
- Deeper architectures are able to better capture invariant properties of objects and scenes in images
  [Bruna & Mallat, 2013], [Wiatowski & Bölcskei, 2016]

SCATTERING TRANSFORMS AS A METRIC

- Scattering transforms may be used as a metric.
- Inverse problems can be solved by minimizing distance at the scattering transform domain.
- Leads to remarkable results in super-resolution [Bruna, Sprechmann & Lecun, 2016]

SCATTERING SUPER RESOLUTION

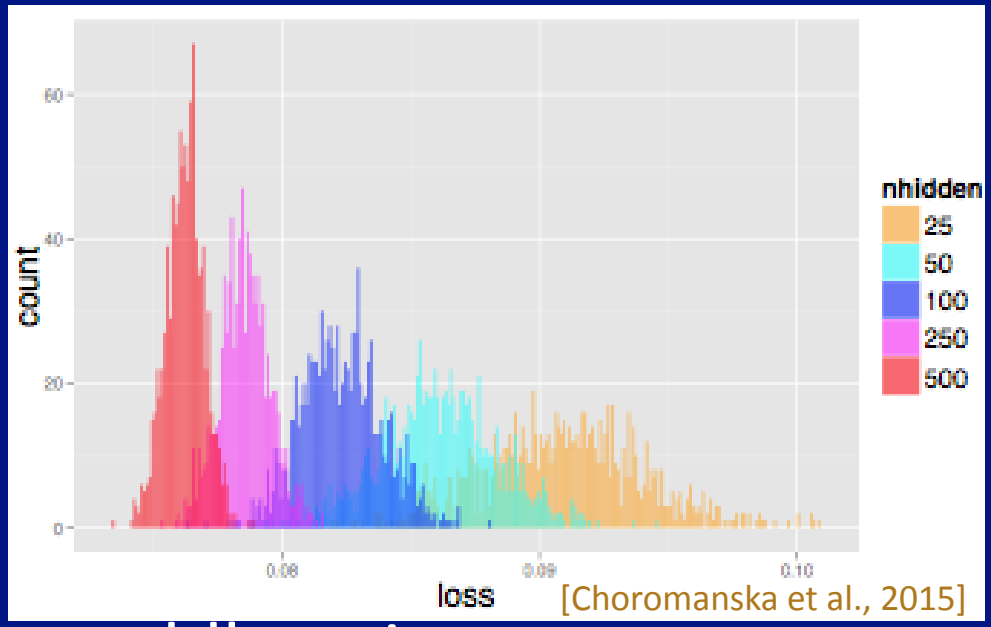
Original | Best Linear Estimate | State-of-the-art | Scattering estimate

[Bruna, Sprechmann & Lecun, 2016]

Images from slides of Joan Bruna in CVPR 2016 tutorial

MINIMIZATION

- The local minima in deep networks are not far from the global minimum.

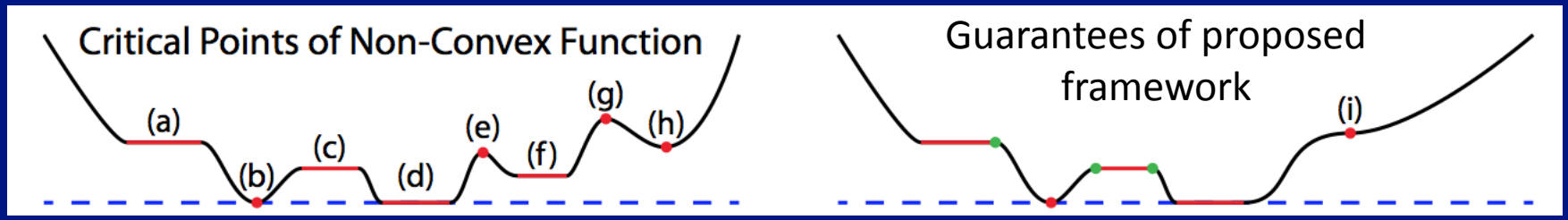


[Choromanska et al., 2015]

- saddle points are the main problem of deep Learning optimization.
- Deeper networks have more local minima but less saddle points.
  [Saxe, McClelland & Ganguli, 2014], [Dauphin, Pascanu, Gulcehre, Cho, Ganguli & Bengio, 2014] [Choromanska, Henaff, Mathieu, Ben Arous & LeCun, 2015]
- Deep learning can be viewed as a sparse recovery method [Papyan et al., 2017]

GLOBAL OPTIMALITY IN DEEP LEARNING

- Deep learning is a positively homogeneous factorization problem, i.e., $\exists p \geq 0$ such that $\forall \alpha \geq 0$ DNN obey $\Phi(\alpha X^1, \alpha X^2, \ldots, \alpha X^K) = \alpha^p \Phi(X^1, X^2, \ldots, X^K)$.
- With proper regularization, local minima are global.
- If the network is large enough, global minima can be found by local descent.



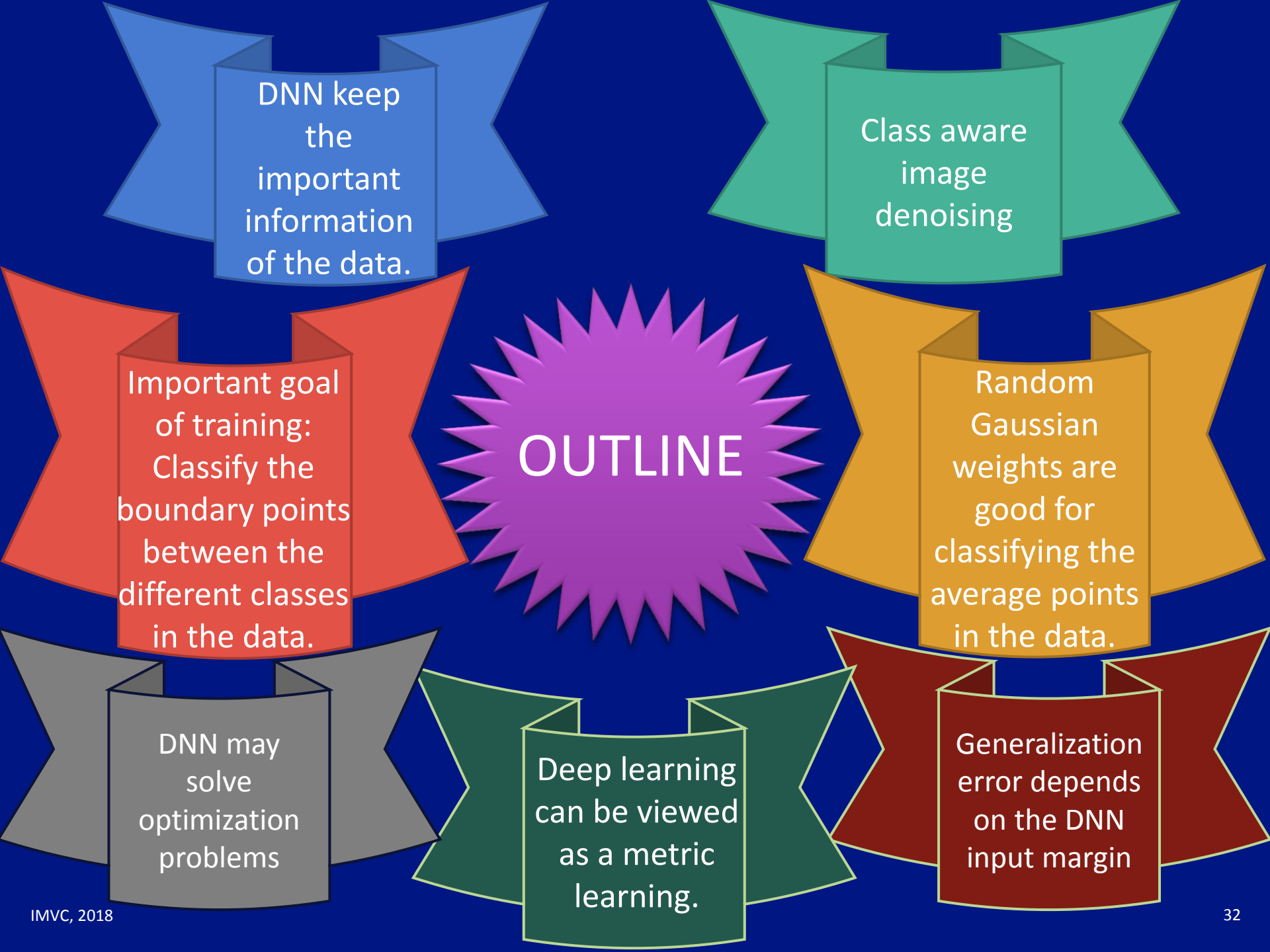
[Haeffele & Vidal, 2015]

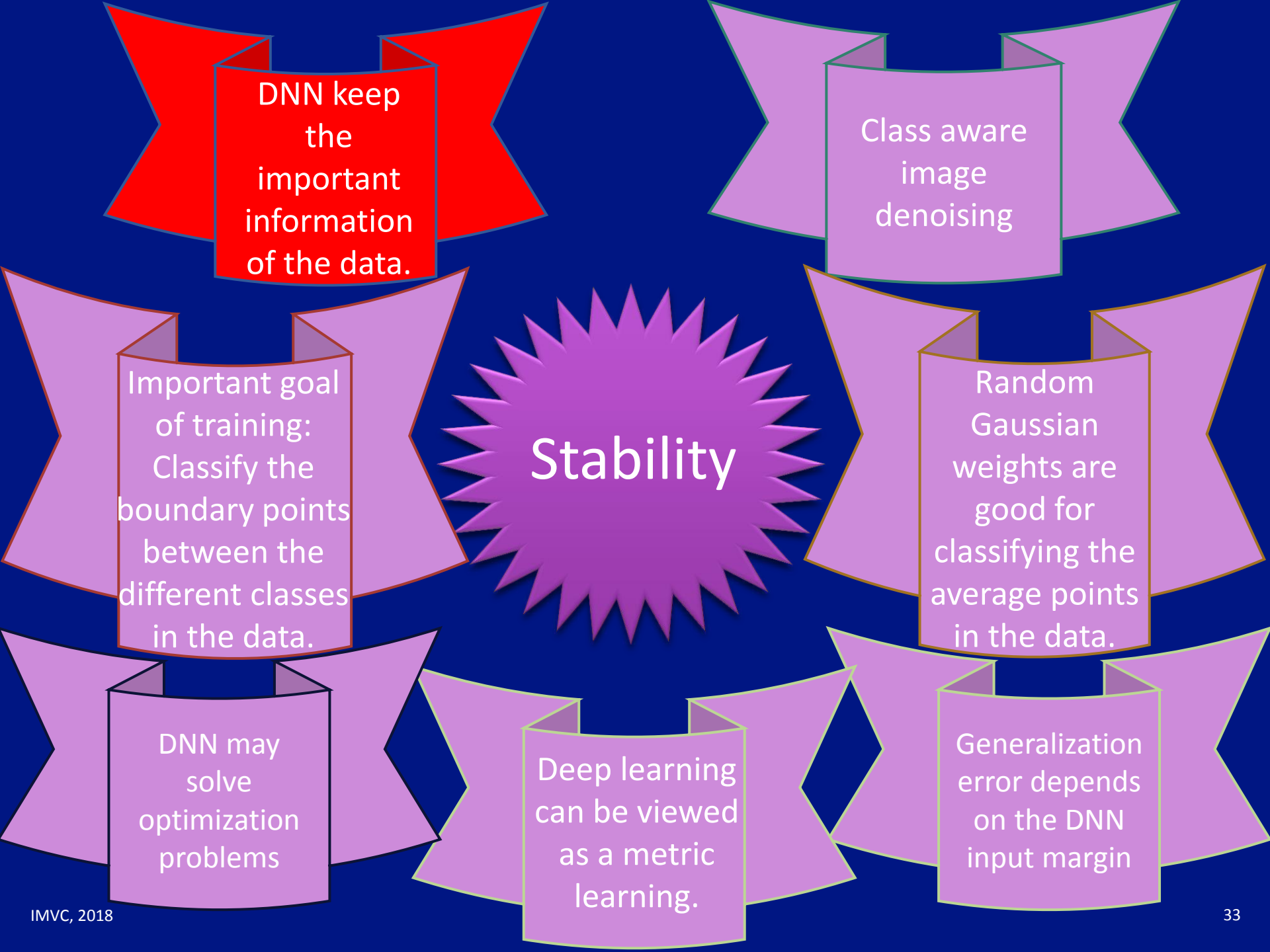
RELATIONSHIP TO SPARSE REPRESENTATION

- Forward pass of CNN can be viewed as a layerwise convolutional sparse coding
- Leads to uniqueness and stability guarantees for the representation in the layers of the network
- Novel strategies for performing the forward pass

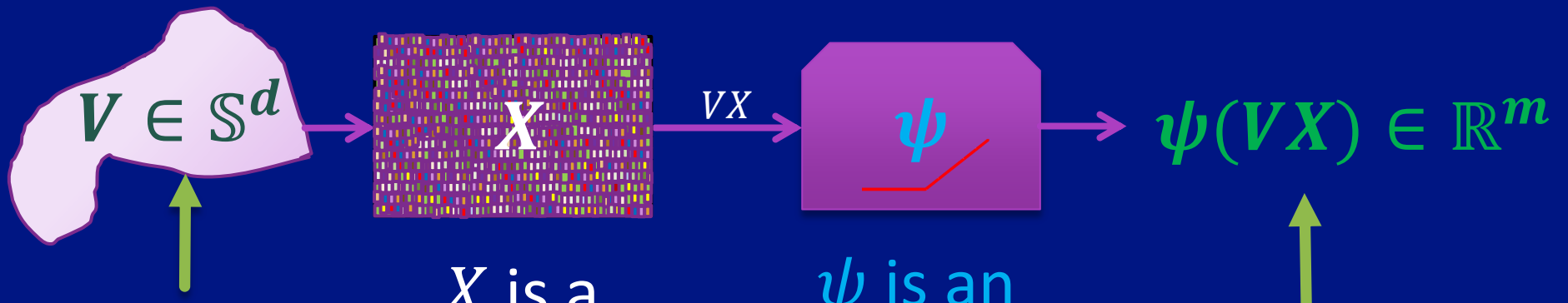
[Papyan, Romano & Elad, 2017]

DATA STRUCTURE BASED THEORY FOR DEEP LEARNING


DNN keep the important information of the data.
Class aware image denoising
Important goal of training: Classify the boundary points between the different classes in the data.
OUTLINE
Random Gaussian weights are good for classifying the average points in the data.
DNN may solve optimization problems
Deep learning can be viewed as a metric learning.
Generalization error depends on the DNN input margin

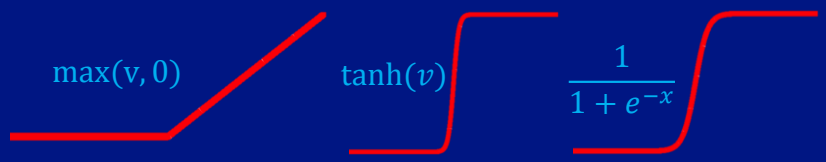

DNN keep the important information of the data.
Class aware image denoising
Important goal of training: Classify the boundary points between the different classes in the data.
Stability
Random Gaussian weights are good for classifying the average points in the data.
DNN may solve optimization problems
Deep learning can be viewed as a metric learning.
Generalization error depends on the DNN input margin

ASSUMPTIONS



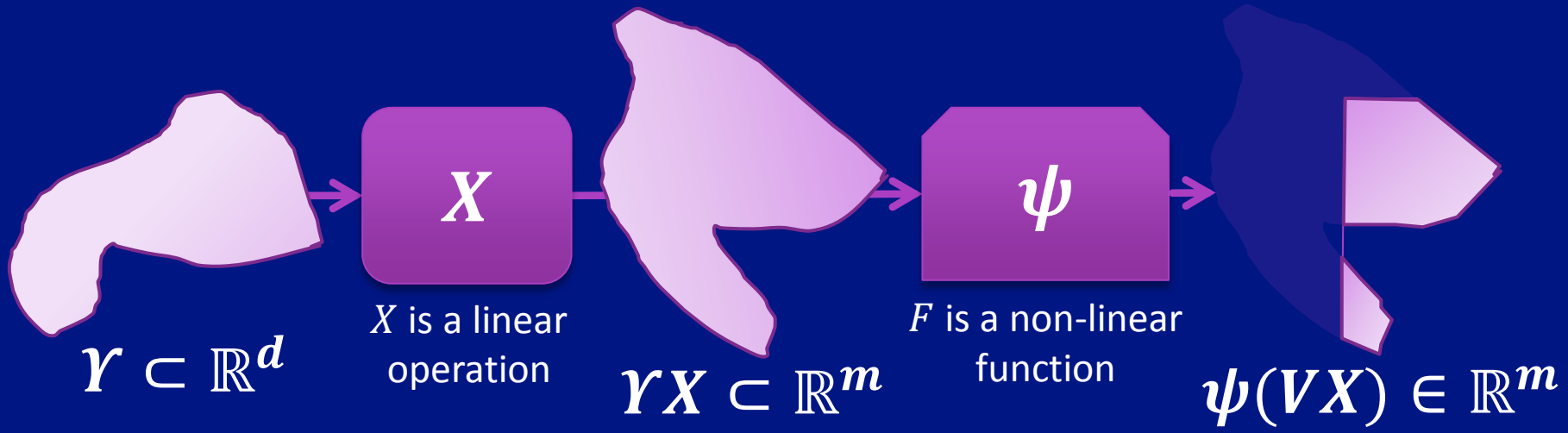
X is a random Gaussian matrix

ψ is an element wise activation function



$$m = O(\delta^{-6}\omega^2(\Upsilon))$$

GAUSSIAN MEAN WIDTH IN DNN

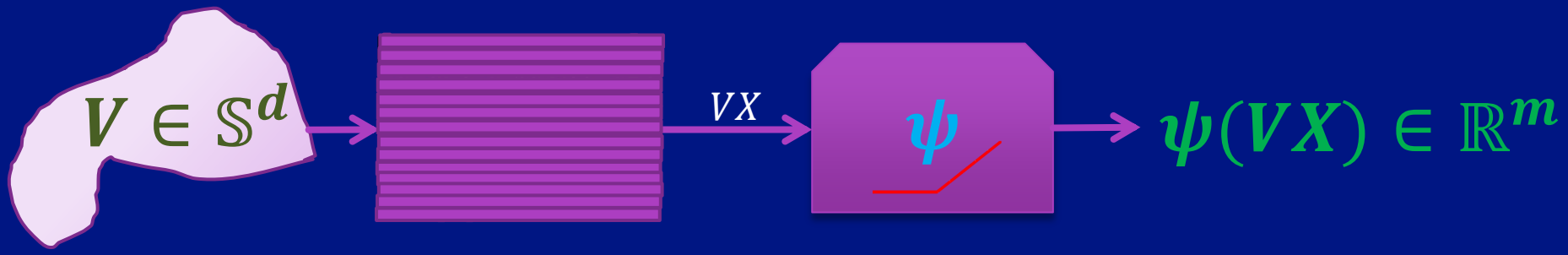


Theorem 1: small $\frac{\omega^2(\Upsilon)}{m}$ imply $\omega^2(\Upsilon) \approx \omega^2(\psi(VX))$

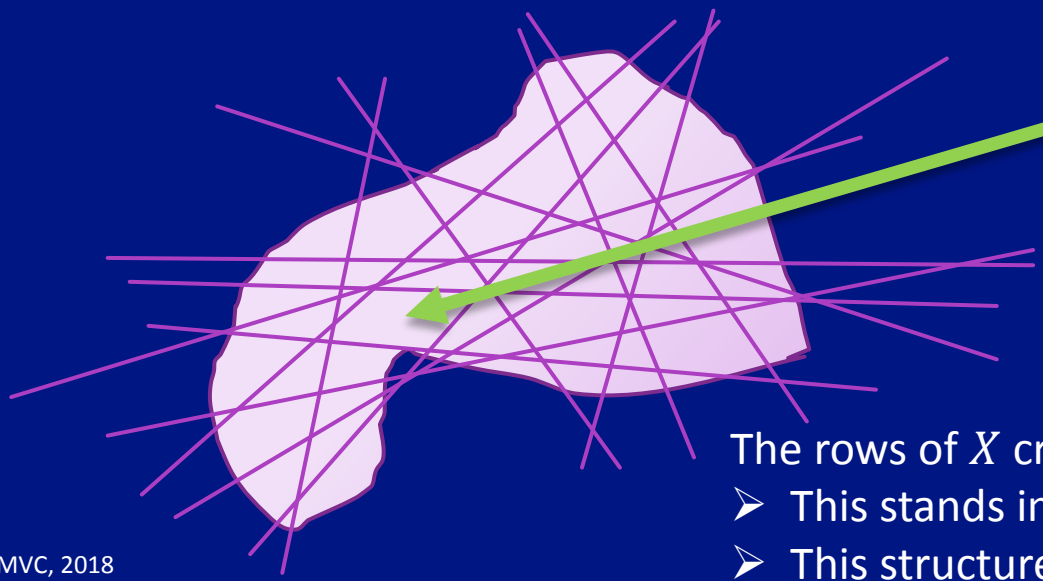
Small $\omega^2(\Upsilon)$ → Small $\omega^2(\psi(VX))$

→ It is sufficient to provide proofs only for a single layer

ISOMETRY IN A SINGLE LAYER



Theorem 2: $\psi(\cdot\, X)$ is a δ-isometry in the Gromov-Hausdorff sense between the sphere $\mathbb{S}^{d-1}$ and the Hamming cube [Plan & Vershynin, 2014, Giryes, Sapiro & Bronstein 2016].



- If two points belong to the same tile then their distance $< \delta$
- Each layer of the network keeps the main information of the data

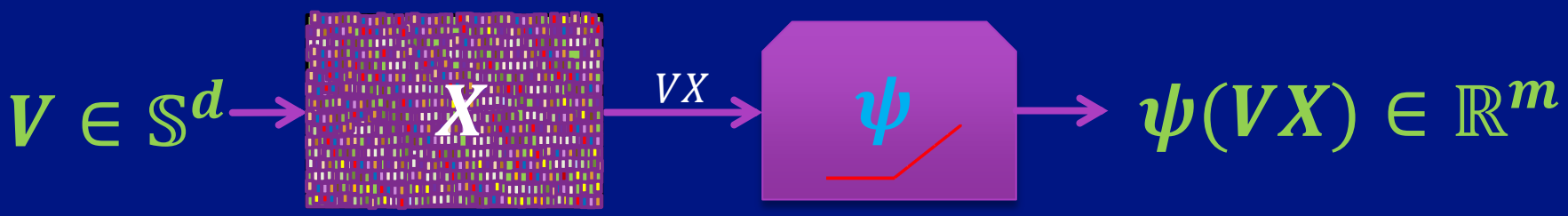
The rows of X create a tessellation of the space.

- This stands in line with [Montúfar et. al. 2014]
- This structure can be used for hashing

DNN AND HASHING

- A single layer performs a locally sensitive hashing.
- Deep network with random weights may be designed to do better [Choromanska et al., 2016].
- It is possible to train DNN for hashing, which provides cutting-edge results [Masci et al., 2012], [Lai et al., 2015].

DNN STABLE EMBEDDING



Theorem 3: There exists an algorithm $\mathcal{A}$ such that

$$\|V - \mathcal{A}(\psi(VX))\| < O\left(\frac{\omega(\Upsilon)}{\sqrt{m}}\right) = O(\delta^3)$$

[Plan & Vershynin, 2013, Giryes, Sapiro & Bronstein 2016].

- After K layers we have an error $O(K\delta^3)$
- Stands in line with [Mahendran and Vedaldi, 2015].
- DNN keep the important information of the data

Role of Training

DNN keep the important information of the data.

Class aware image denoising

Important goal of training: Classify the boundary points between the different classes in the data.

Random Gaussian weights are good for classifying the average points in the data.

DNN may solve optimization problems

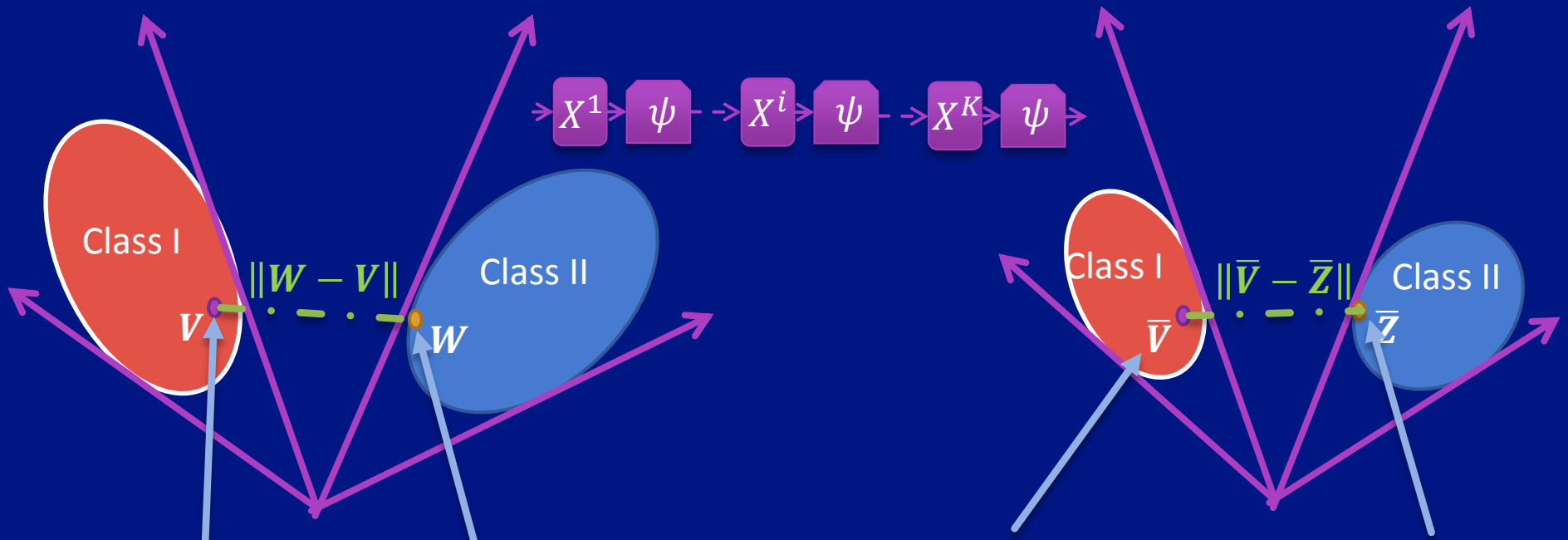
Deep learning can be viewed as a metric learning.

Generalization error depends on the DNN input margin

ROLE OF TRAINING

- Having a theory for Gaussian weights we test the behavior of DNN after training.
- We looked at the MNIST, CIFAR-10 and ImageNet datasets.
- We will present here only the ImageNet results.
- We use a state-of-the-art pre-trained network for ImageNet [Simonyan & Zisserman, 2014].
- We compute inter and intra class distances.

INTER BOUNDARY POINTS DISTANCE RATIO

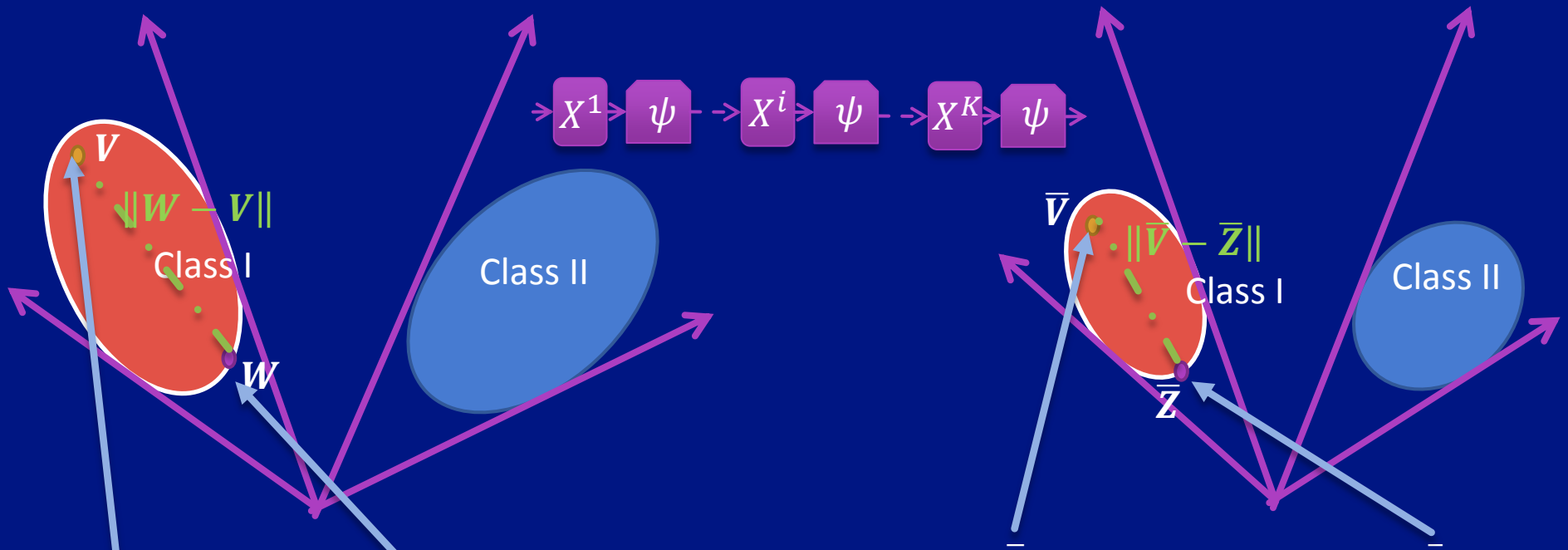


V is a random point and W its closest point from a different class.

$\bar{V}$ is the output of V and $\bar{Z}$ the closest point to $\bar{V}$ at the output from a different class.

Compute the distance ratio: $\frac{\|\bar{V}-\bar{Z}\|}{\|W-V\|}$

INTRA BOUNDARY POINTS DISTANCE RATIO



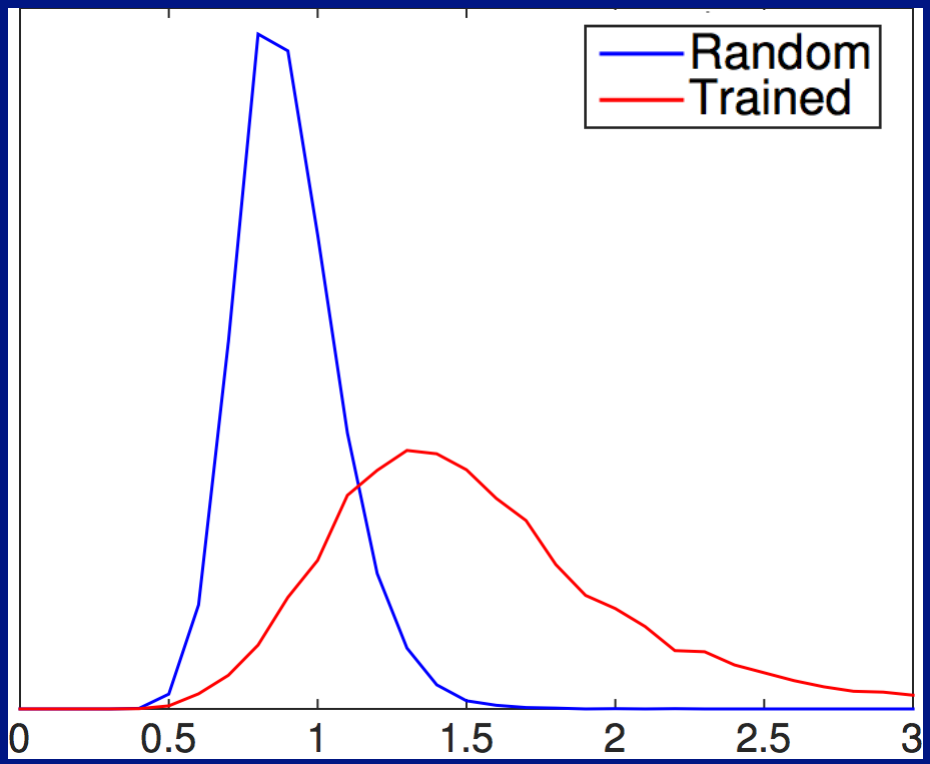
Let V be a point and W its farthest point from the same class.

Let $\bar{V}$ be the output of V and $\bar{Z}$ the farthest point from $\bar{V}$ at the output from the same class

Compute the distance ratio: $\frac{\|\bar{V}-\bar{Z}\|}{\|W-V\|}$

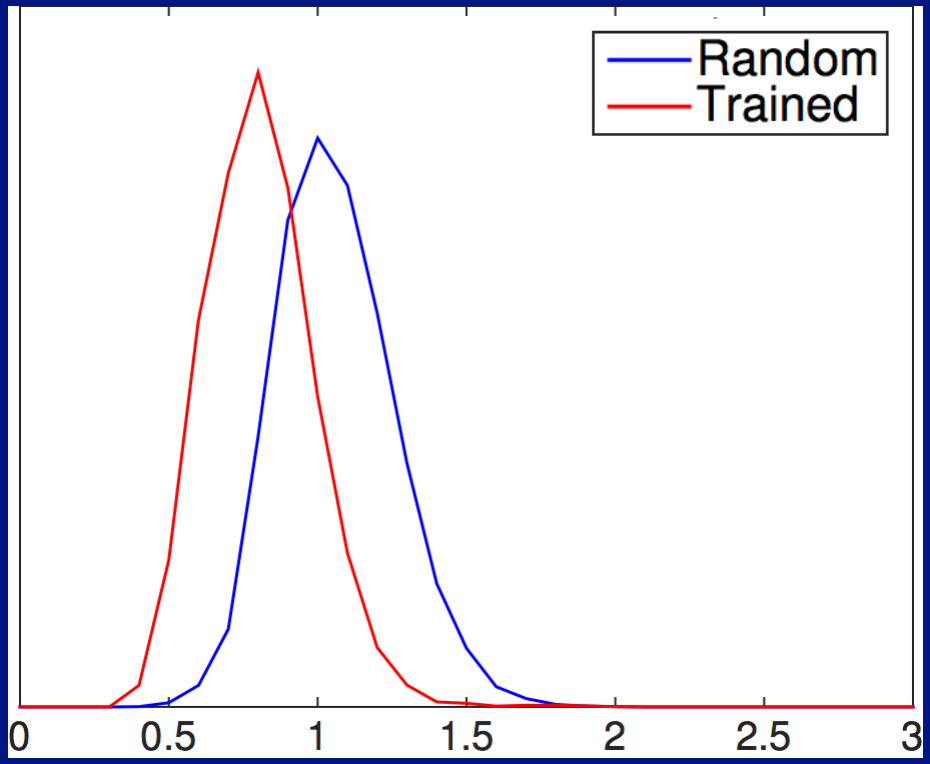
BOUNDARY DISTANCE RATIO

Inter-class



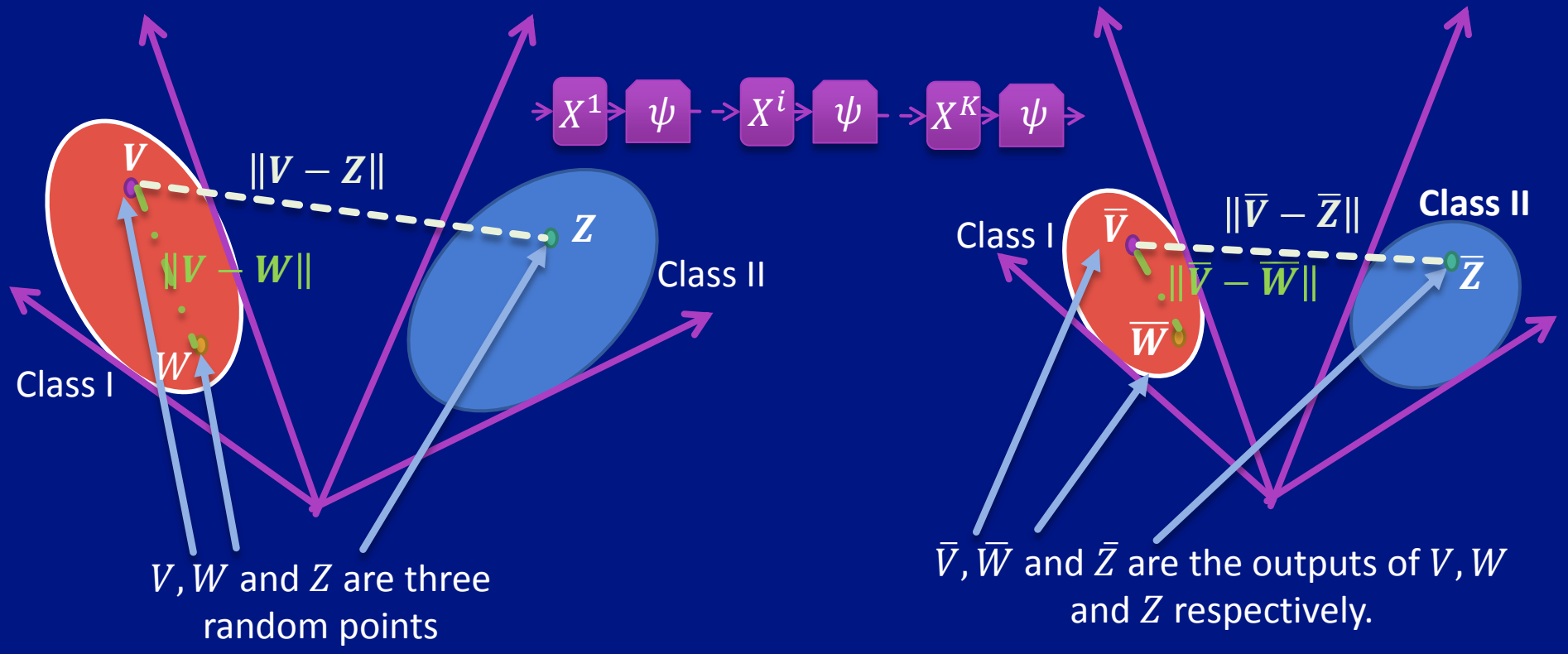
$$\frac{\|\bar{V} - \bar{Z}\|}{\|W - V\|}$$

Intra-class



$$\frac{\|\bar{V} - \bar{Z}\|}{\|W - V\|}$$

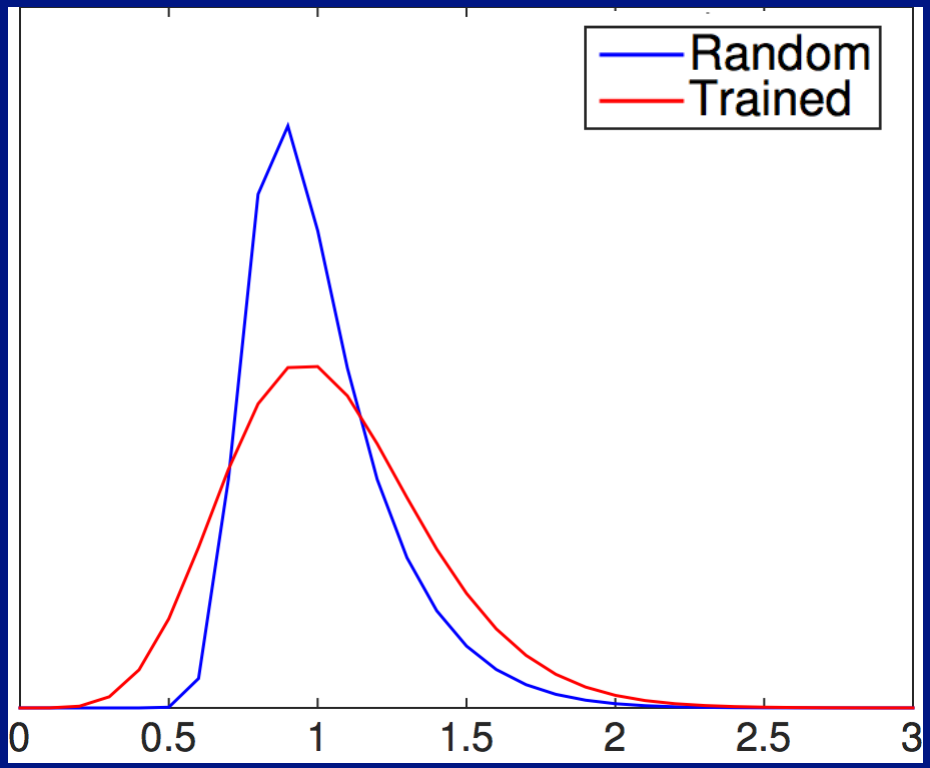
AVERAGE POINTS DISTANCE RATIO



Compute the distance ratios: $\frac{\|\bar{V}-\bar{W}\|}{\|V-W\|}, \frac{\|\bar{V}-\bar{Z}\|}{\|V-Z\|}$

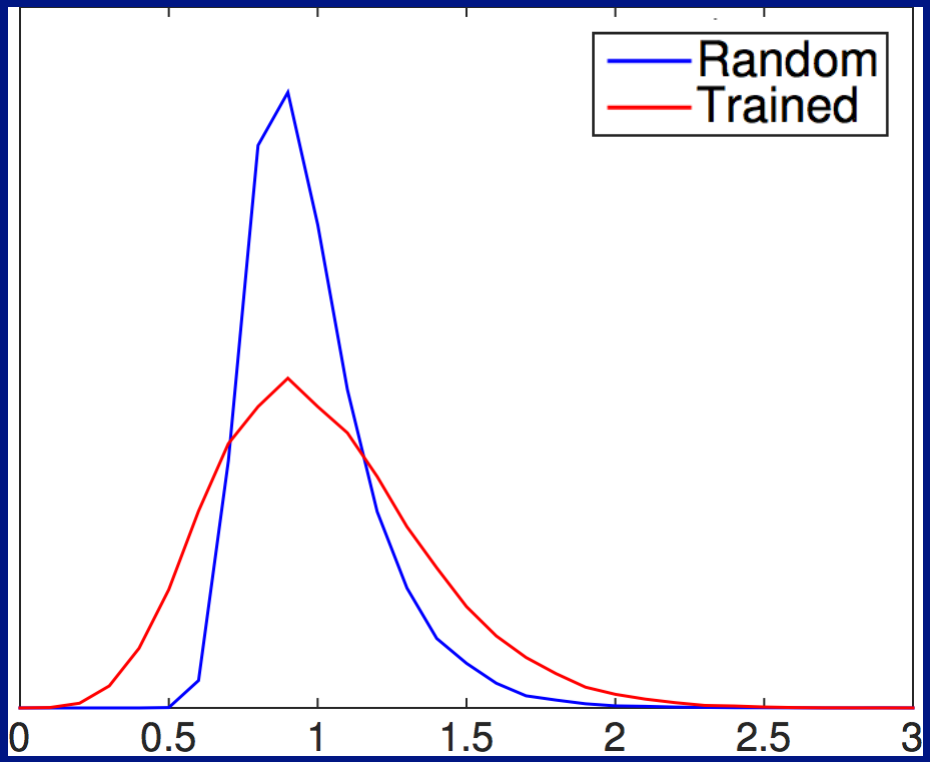
AVERAGE DISTANCE RATIO

Inter-class



$$\frac{\|\bar{V}-\bar{Z}\|}{\|V-Z\|}$$

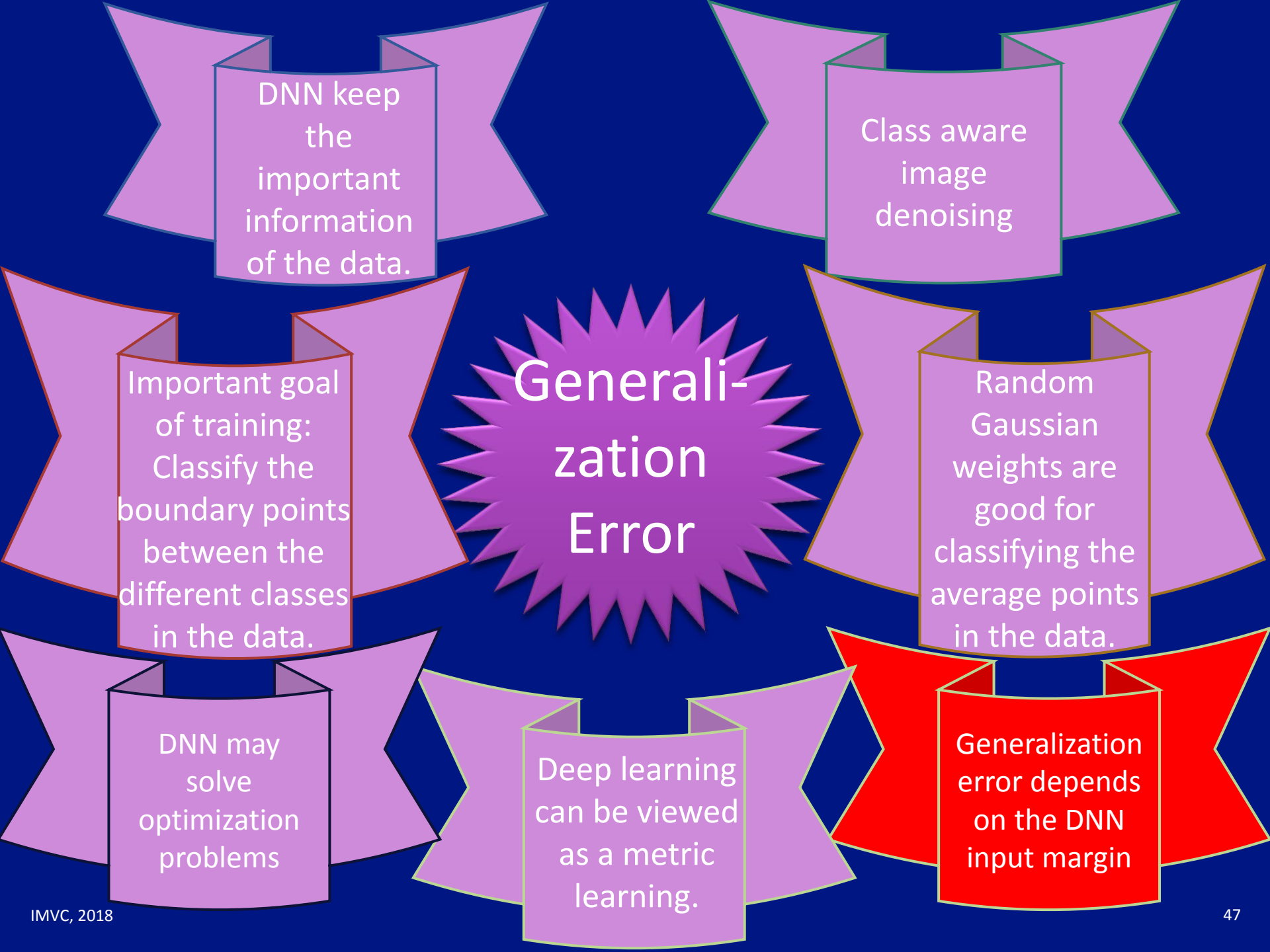
Intra-class



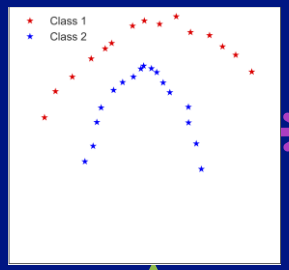
$$\frac{\|\bar{V}-\bar{W}\|}{\|V-W\|}$$

ROLE OF TRAINING

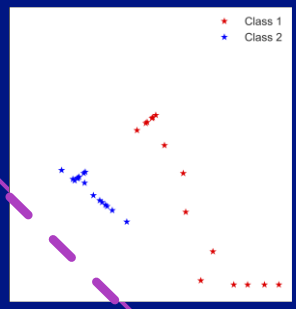
- On average distances are preserved in the trained and random networks.
- The difference is with respect to the boundary points.
- The inter distances become larger.
- The intra distances shrink.


DNN keep the important information of the data.
Class aware image denoising
Generali-zation Error
Important goal of training: Classify the boundary points between the different classes in the data.
Random Gaussian weights are good for classifying the average points in the data.
DNN may solve optimization problems
Deep learning can be viewed as a metric learning.
Generalization error depends on the DNN input margin

ASSUMPTIONS



X^1 → ψ → X^i → ψ → X^K → ψ

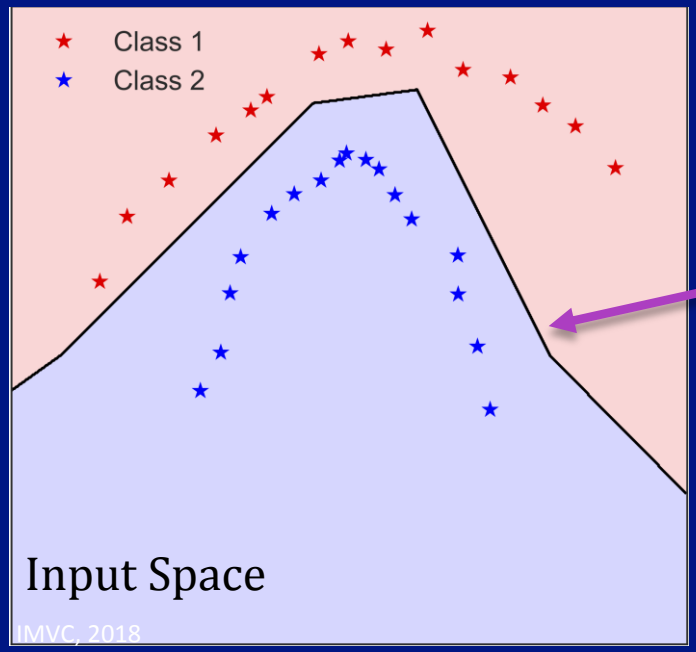


Two Classes $\in \Upsilon$

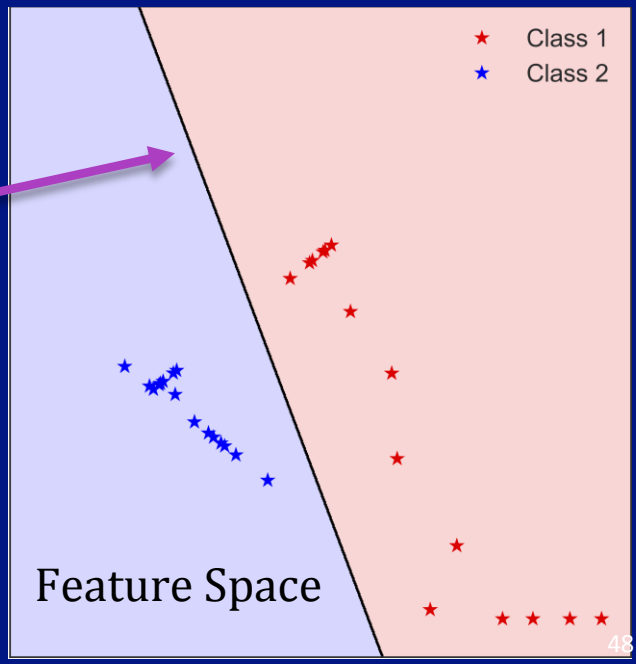
general non-linearity (ReLU, pooling,…)

softmax/ linear classifier

W



$w^T \Phi(X^1, X^2, \ldots, X^K) = 0$



GENERALIZATION ERROR (GE)

- In training, we reduce the classification error ℓ_{training} of the training data as the number of training examples L increases.
- However, we are interested to reduce the error ℓ_{test} of the (unknown) testing data as L increases.
- The difference between the two is the generalization error

$$\text{GE} = \ell_{\text{training}} - \ell_{\text{test}}$$

⇨ It is important to understand the GE of DNN

REGULARIZATION TECHNIQUES

- Weight decay – penalizing DNN weights [Krogh & Hertz, 1992].
- Dropout - randomly drop units (along with their connections) from the neural network during training [Hinton et al., 2012], [Baldi & Sadowski, 2013], Srivastava et al., 2014].
- DropConnect – dropout extension [Wan et al., 2013]
- Batch normalization [Ioffe & Szegedy, 2015].
- Stochastic gradient descent (SGD) [Hardt, Recht & Singer, 2016].
- Path-SGD [Neyshabur et al., 2015].
- And more [Rifai et al., 2011], [Salimans & Kingma, 2016], [Sun et al, 2016].

A SAMPLE OF GE BOUNDS

- Using the VC dimension it can be shown that

$$\mathrm{GE} \leq O\left(\sqrt{\text{DNN params} \cdot \frac{\log(L)}{L}}\right)$$

[Shalev-Shwartz and Ben-David, 2014].

- The GE was bounded also by the DNN weights

$$\mathrm{GE} \leq \frac{1}{\sqrt{L}} 2^K \|w\|_2 \prod_i \left\|X^i\right\|_{2,2}$$

[Neyshabur et al., 2015].

A SAMPLE OF GE BOUNDS

- Using the VC dimension it can be shown that

$$\mathrm{GE} \leq O\left(\sqrt{\text{DNN params} \cdot \frac{\log(L)}{L}}\right)$$

[Shalev-Shwartz and Ben-David, 2014].

- The GE was bounded also by the DNN weights

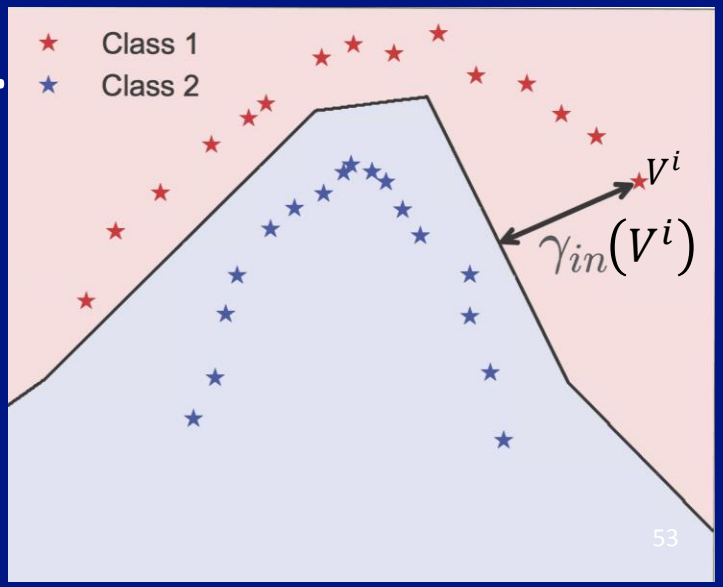
$$\mathrm{GE} \leq \frac{1}{\sqrt{L}} 2^{K} \|w\|_2 \prod_i \left\|X^i\right\|_{2,2}$$

[Neyshabur et al., 2015].

- Note that in both cases the GE grows with the depth

DNN INPUT MARGIN

- Theorem 6: If for every input margin $\gamma_{in}(V^i) > \gamma$ then $GE \leq \sqrt{N_{\gamma/2}(\Upsilon)}/\sqrt{L}$ [Sokolic, Giryes, Sapiro, Rodrigues, 2016]
- $N_{\gamma/2}(\Upsilon)$ is the covering number of the data Υ.
- $N_{\gamma/2}(\Upsilon)$ gets smaller as γ gets larger.
- Bound is independent of depth.
- Our theory relies on the robustness framework [Xu & Mannor, 2012].



INPUT MARGIN BOUND

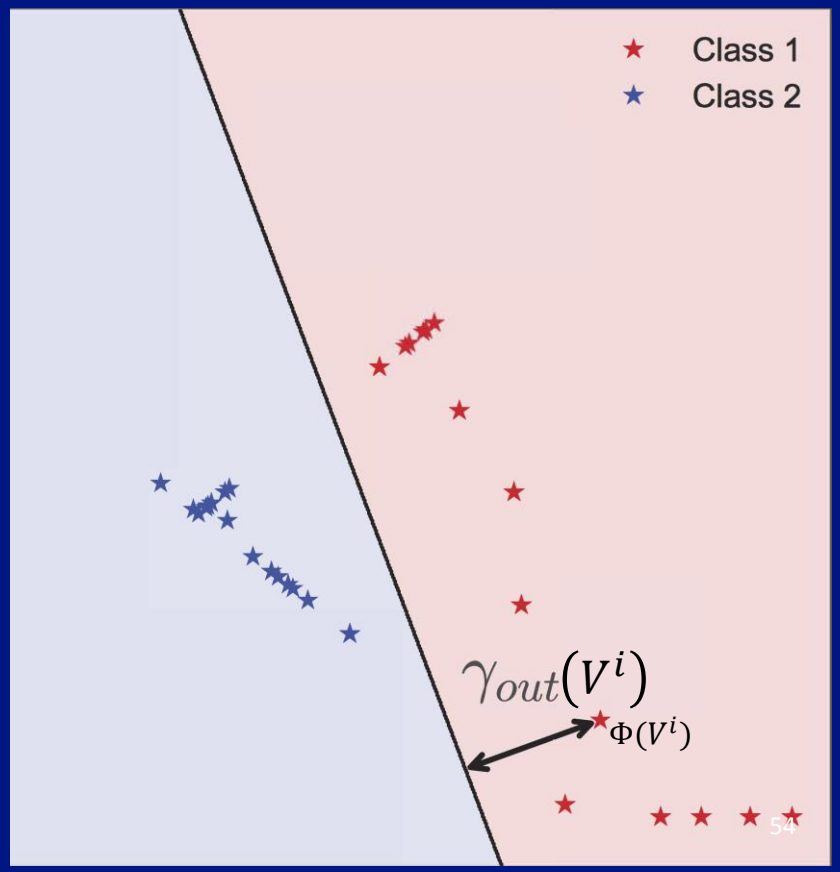
- Maximizing the input margin directly is hard
- Our strategy: relate the input margin to the output margin $\gamma_{out}(V^i)$ and other DNN properties
- Theorem 7:

$$\gamma_{in}(V^i) \geqslant \frac{\gamma_{out}(V^i)}{\sup_{V \in \Upsilon} \left\| \frac{V}{\|V\|_2} J(V) \right\|_2}$$

$$\geqslant \frac{\gamma_{out}(V^i)}{\prod_{1 \leq i \leq K} \|X^i\|_2}$$

$$\geqslant \frac{\gamma_{out}(V^i)}{\prod_{1 \leq i \leq K} \|X^i\|_F}$$

[Sokolic, Giryes, Sapiro, Rodrigues, 2016]



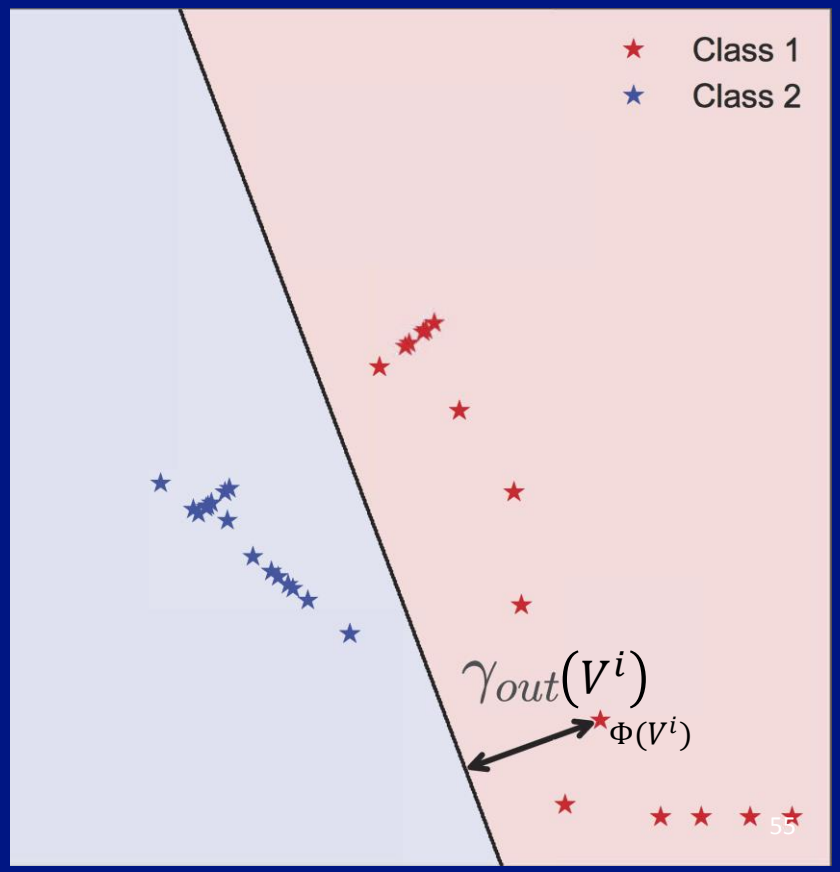
OUTPUT MARGIN

- Theorem 7:

$$\gamma_{in}(V^i) \geq \frac{\gamma_{out}(V^i)}{\sup_{V \in \mathcal{Y}} \left\| \frac{V}{\|V\|_2} J(V) \right\|_2} \geq \frac{\gamma_{out}(V^i)}{\prod_{1 \leq i \leq K} \|X^i\|_2}$$

$$\geq \frac{\gamma_{out}(V^i)}{\prod_{1 \leq i \leq K} \|X^i\|_F}$$

- Output margin is easier to maximize – SVM problem
- Maximized by many cost functions, e.g., hinge loss.

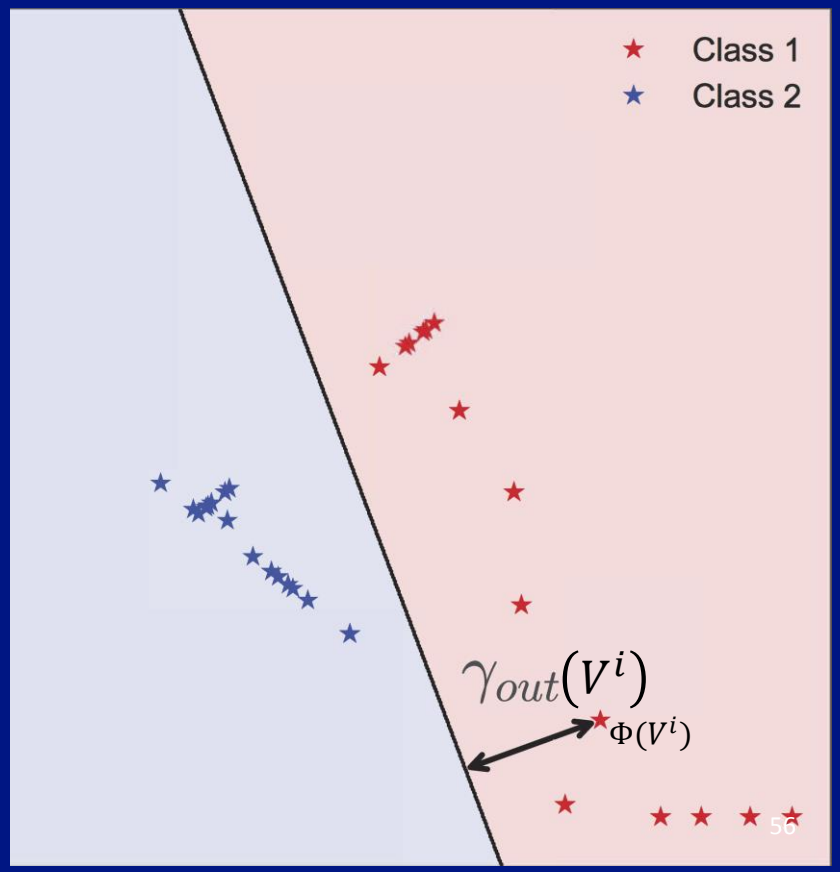


GE AND WEIGHT DECAY

- Theorem 7: $\gamma_{in}(V^i) \geq \frac{\gamma_{out}(V^i)}{\sup_{V \in \mathcal{V}} \left\| \frac{V}{\|V\|_2} J(V) \right\|_2} \geq \frac{\gamma_{out}(V^i)}{\prod_{1 \leq i \leq K} \|X^i\|_2}$

  $\geq \frac{\gamma_{out}(V^i)}{\prod_{1 \leq i \leq K} \|X^i\|_F}$

- Bounding the weights increases the input margin
- Weight decay regularization decreases the GE
- Related to regularization used by [Haeffele & Vidal, 2015]

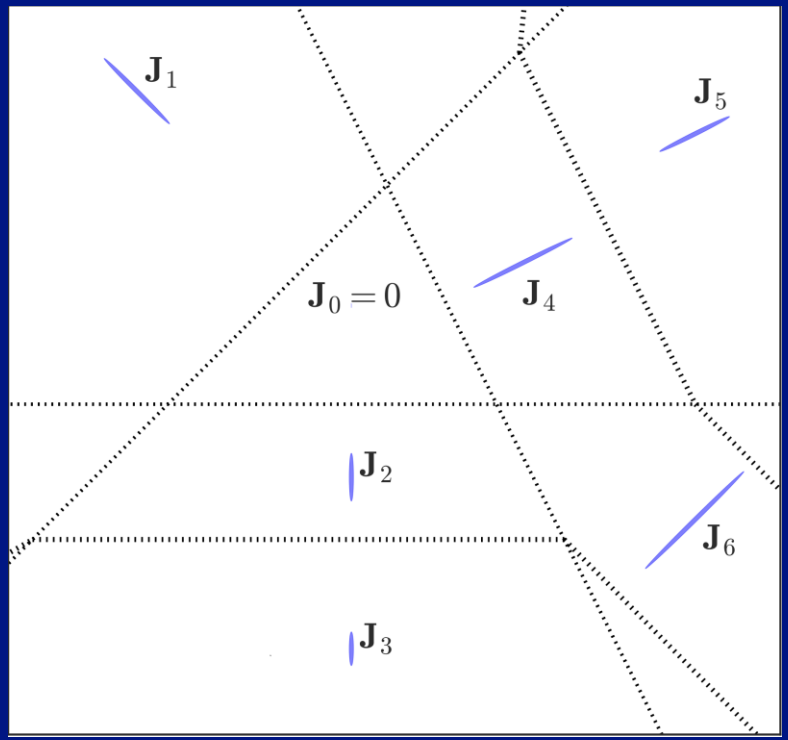


JACOBIAN BASED REGULARIZATION

- Theorem 7: $\gamma_{in}(V^i) \geq \frac{\gamma_{out}(V^i)}{\sup_{V \in \Upsilon} \left\| \frac{V}{\|V\|_2} J(V) \right\|_2} \geq \frac{\gamma_{out}(V^i)}{\prod_{1 \leq i \leq K} \|X^i\|_2}$

  $\geq \frac{\gamma_{out}(V^i)}{\prod_{1 \leq i \leq K} \|X^i\|_F}$

- $J(V)$ is the Jacobian of the DNN at point V.
- $J(\cdot)$ is piecewise constant.
- Using the Jacobian of the DNN leads to a better bound.
- ⇨ New regularization technique.



RESULTS

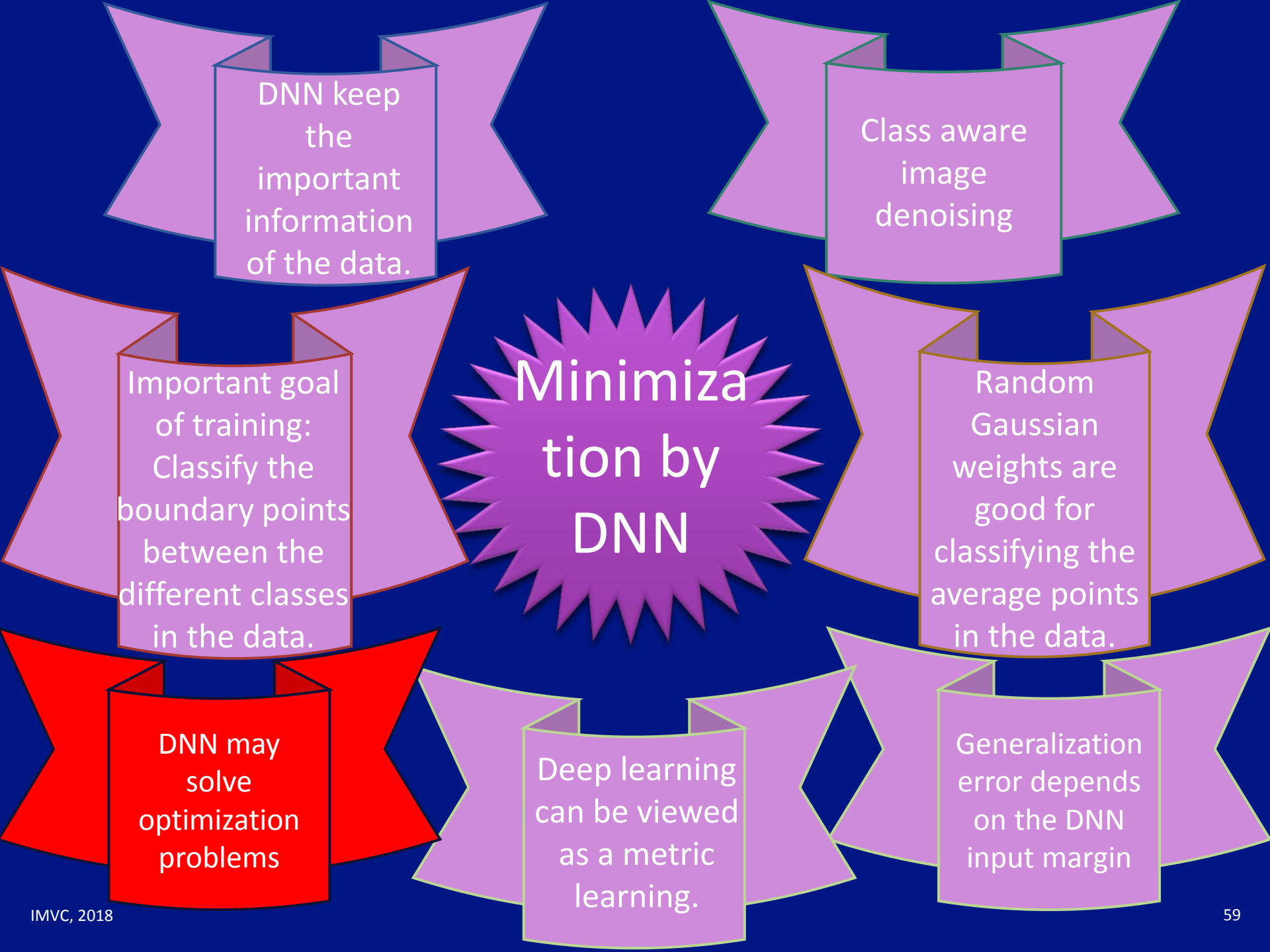
- Better performance with less training samples

MNIST Dataset

		256 samples			512 samples			1024 samples		
loss	# layers	no reg.	WD	LM	no reg.	WD	LM	no reg.	WD	LM
hinge	2	88.37	89.88	**93.83**	93.99	94.62	95.49	95.79	96.57	97.45
hinge	3	87.22	89.31	93.22	93.41	93.97	**95.76**	95.46	96.45	**97.60**
CCE	2	88.45	88.45	92.77	92.29	93.14	95.25	95.38	95.79	96.89
CCE	3	89.05	89.05	93.10	91.81	93.02	95.32	95.11	95.86	97.14

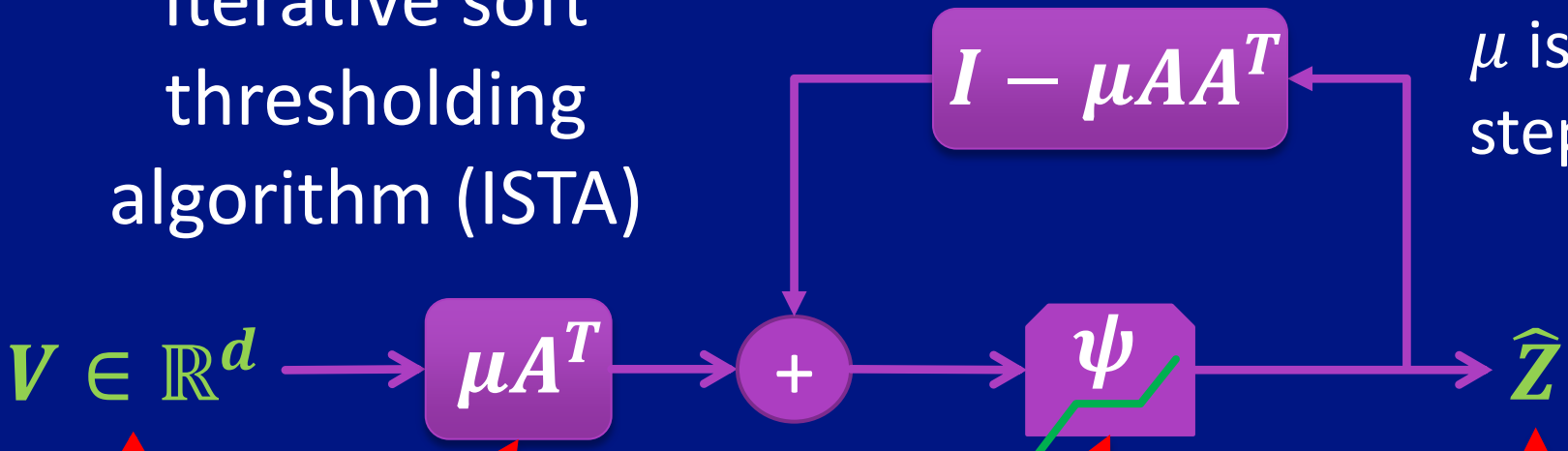
- CCE: the categorical cross entropy.
- WD: weight decay regularization.
- LM: Jacobian based regularization for large margin.
- Note that hinge loss generalizes better than CCE and that LM is better than WD as predicted by our theory.

[Sokolic, Giryes, Sapiro, Rodrigues, 2016]


DNN keep the important information of the data.
Class aware image denoising
Minimiza tion by DNN
Important goal of training: Classify the boundary points between the different classes in the data.
Random Gaussian weights are good for classifying the average points in the data.
DNN may solve optimization problems
Deep learning can be viewed as a metric learning.
Generalization error depends on the DNN input margin

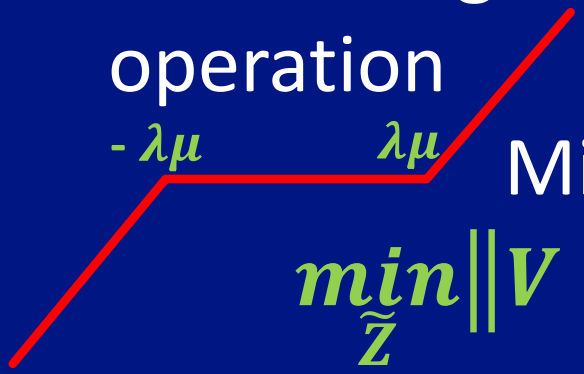
UNCONSTRAINED ℓ_1-MINIMIZATION

Iterative soft thresholding algorithm (ISTA)



Step size μ obeys $\frac{1}{\mu} \geq \|A\|$

Soft thresholding operation



Minimizer of

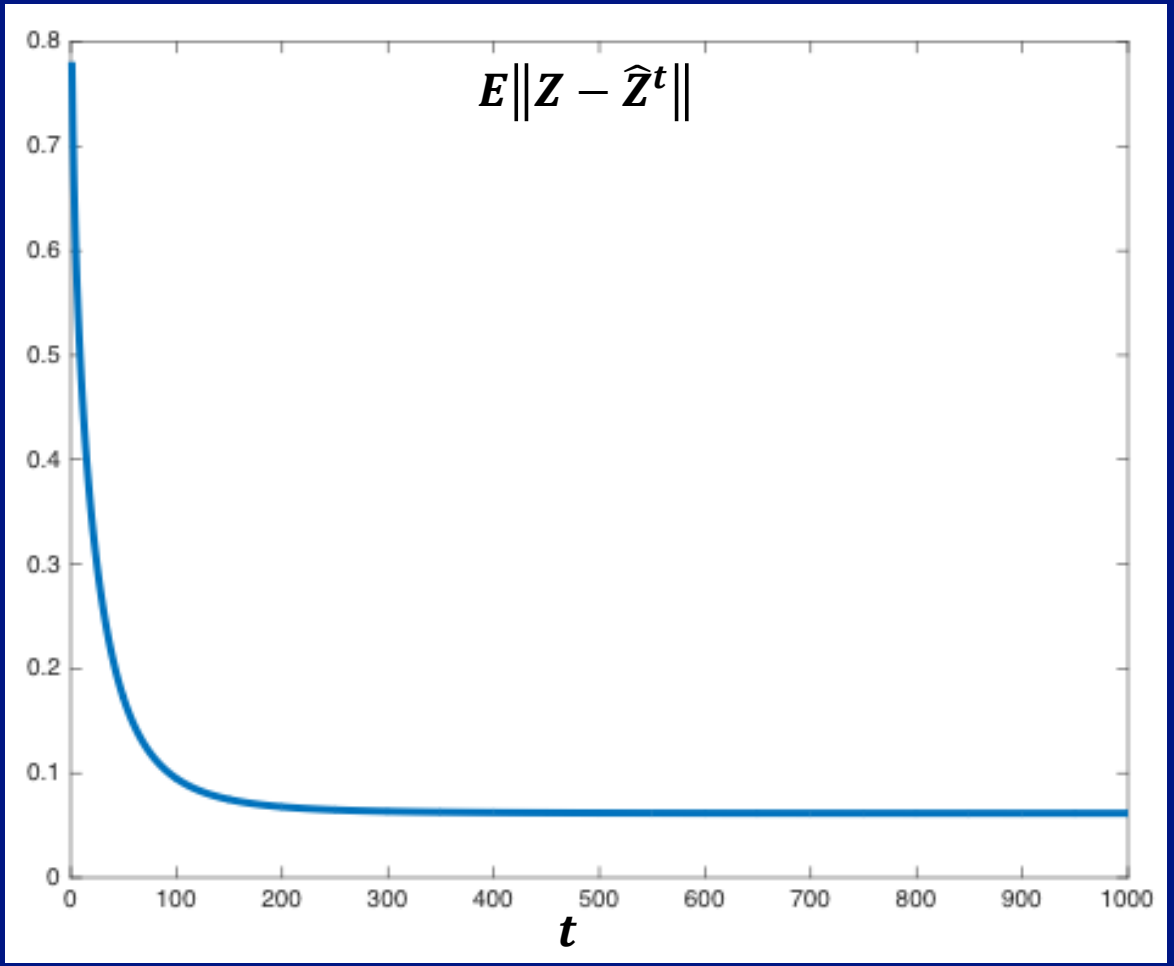
$$\min_{\widetilde{Z}} \|V - \widetilde{Z}A\| + \lambda\|\widetilde{Z}\|_1$$

$$V = ZA + E$$

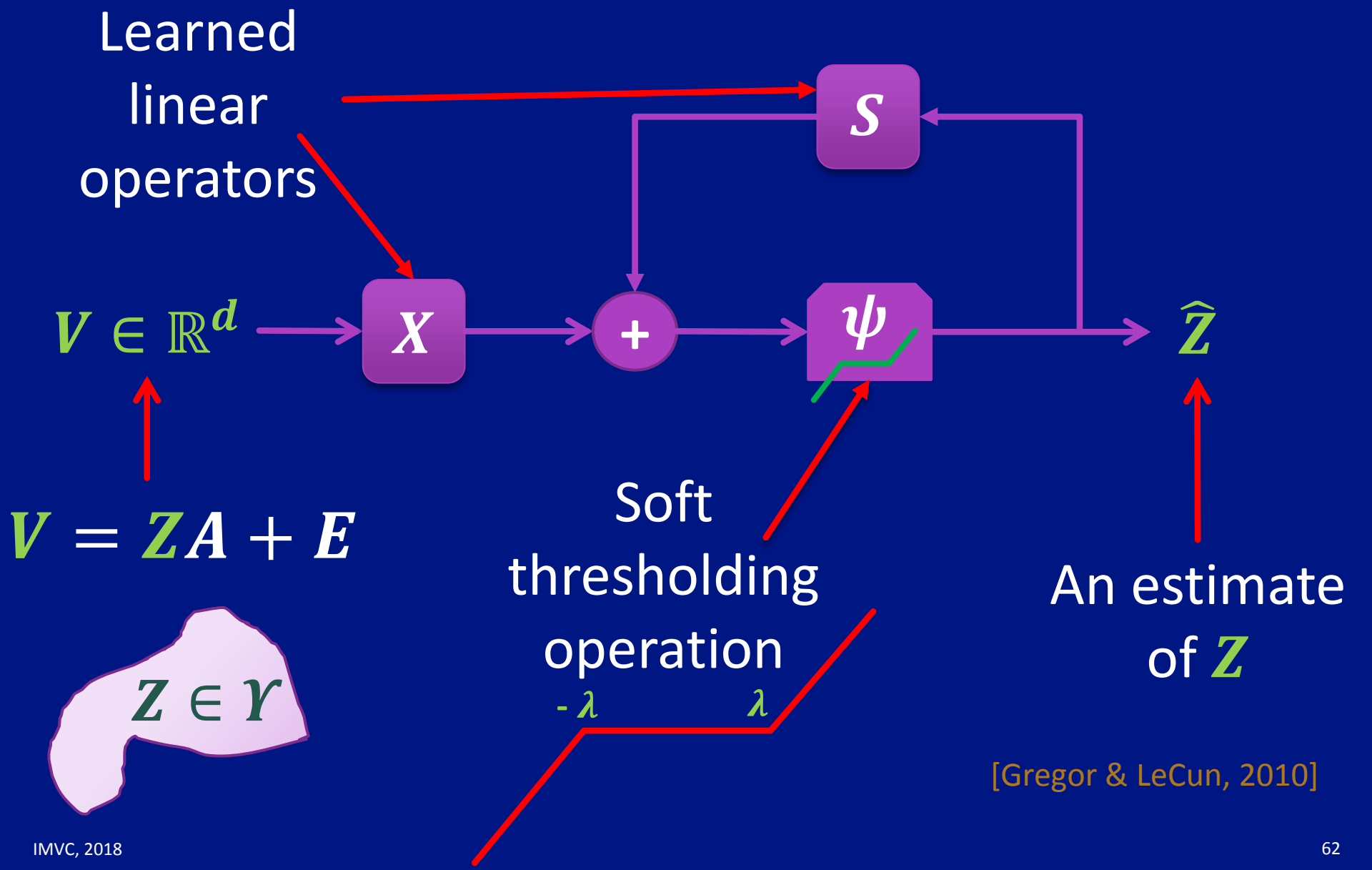
[Daubechies, Defrise & Mol, 2004], [Beck & Teboulle, 2009]

ISTA CONVERGENCE

- Reconstruction mean squared error (MSE) as a function of the number of iterations



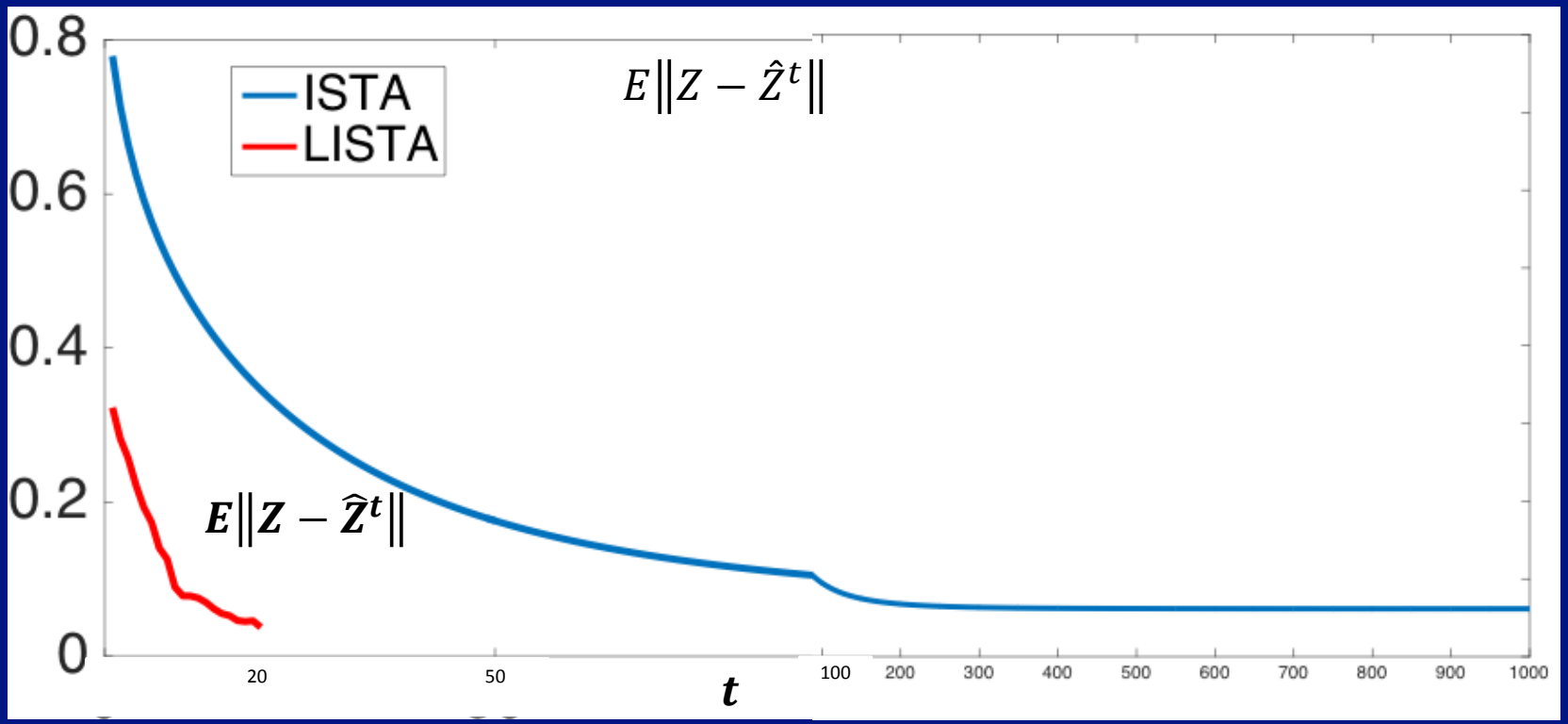
LEARNED ISTA (LISTA)



[Gregor & LeCun, 2010]

LISTA CONVERGENCE

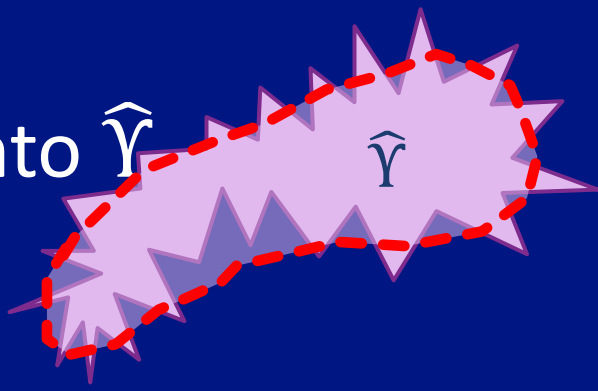
- Replacing $I - \mu AA^T$ and μA^T in ISTA with the learned X and S improves convergence [Gregor & LeCun, 2010]

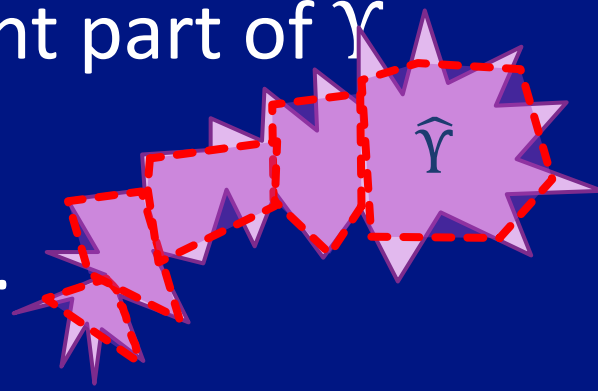


- Extensions to other models [Sprechmann, Bronstein & Sapiro, 2015], [Remez, Litani & Bronstein, 2015], [Tompson, Schlachter, Sprechmann & Perlin, 2016].

LISTA MIXTURE MODEL

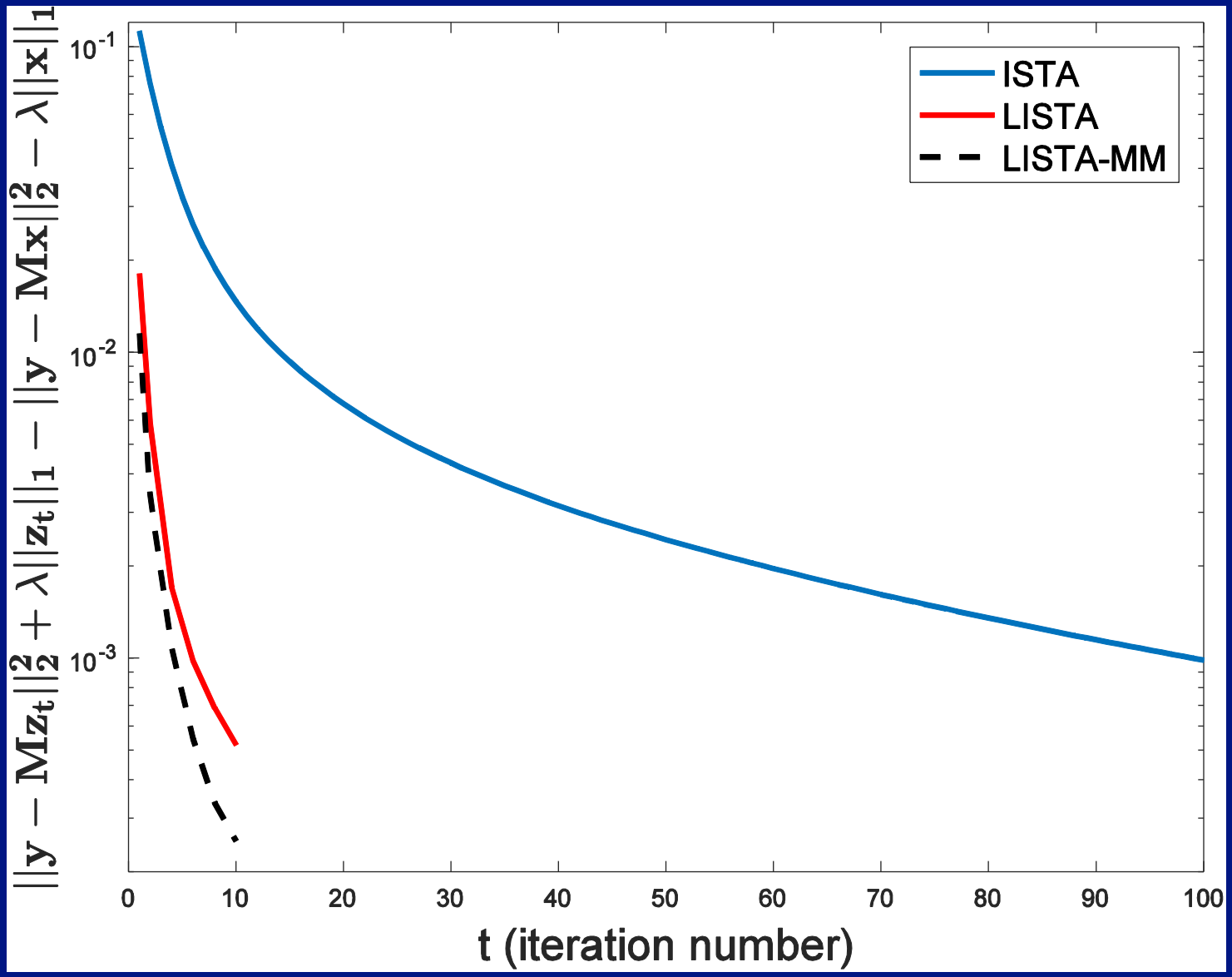
- Approximation of the projection onto $\widehat{\Upsilon}$ with one linear projection may not be accurate enough.
- This requires more LISTA layers/iterations.
- Instead, one may use several LISTA networks, where each approximates a different part of $\widehat{\Upsilon}$
- Training multiple LISTA networks accelerate the convergence further.



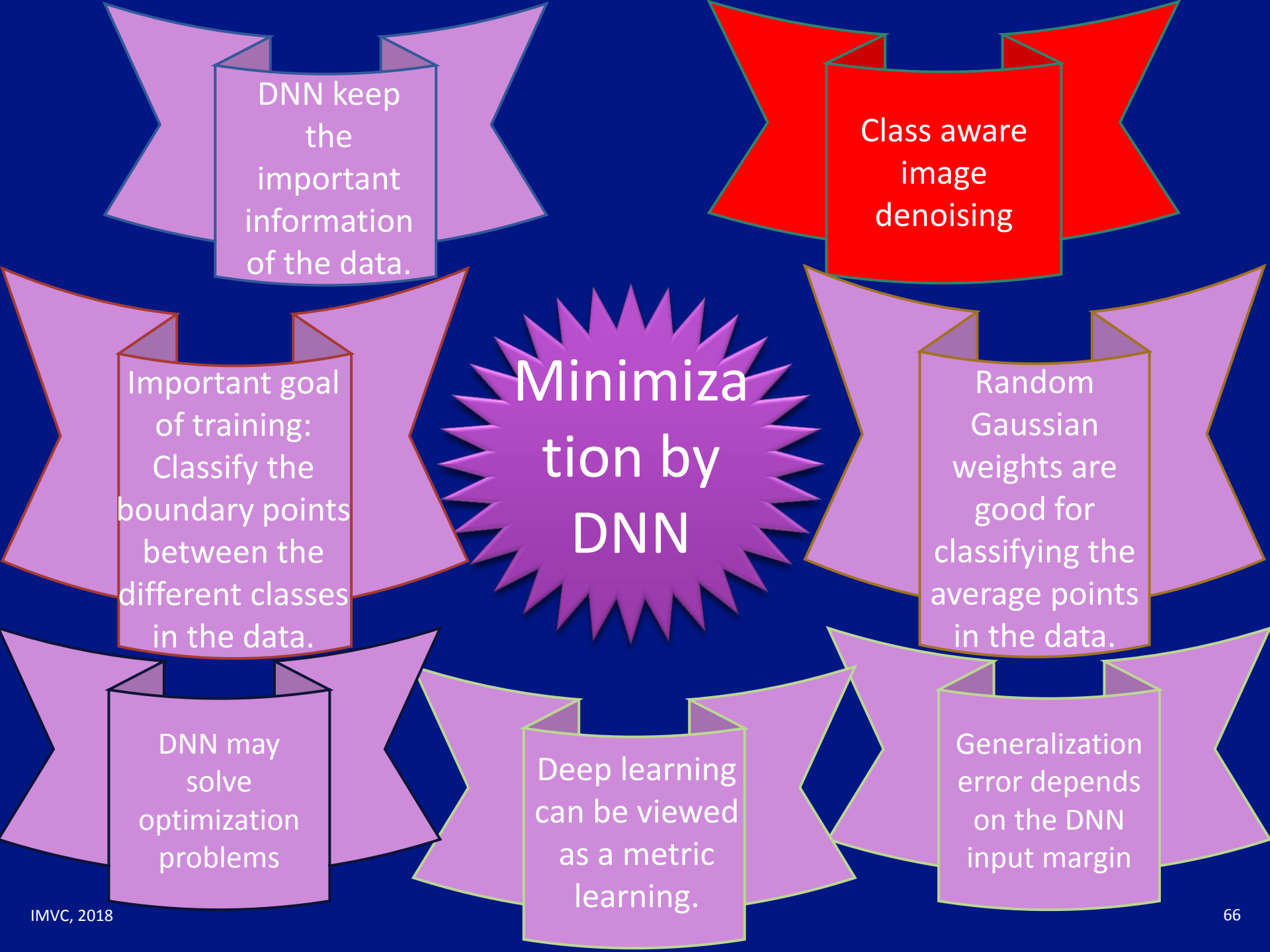


[Giryes et al., 2018]

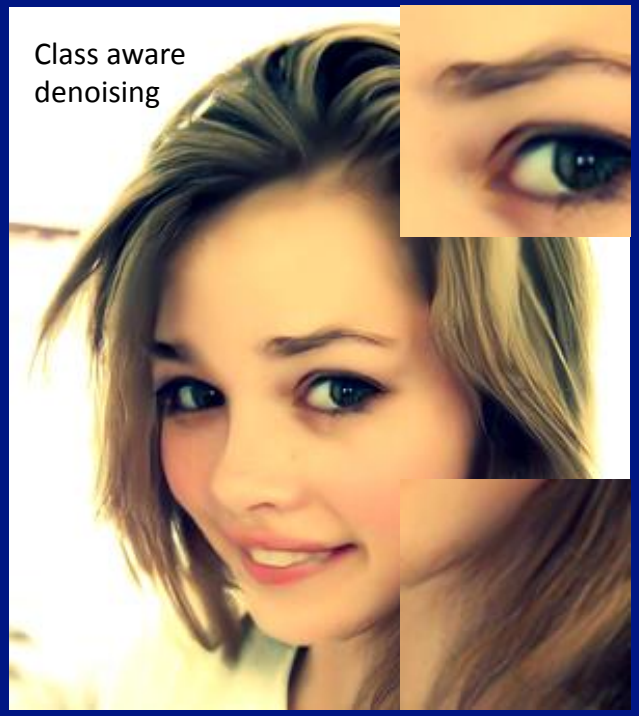
LISTA MIXTURE MODEL

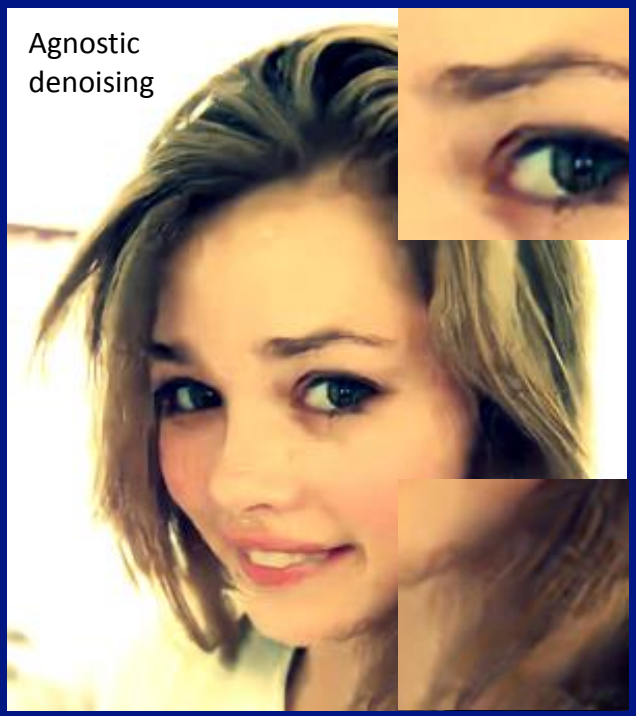


[Giryes et al., 2018]


DNN keep the important information of the data.
Class aware image denoising
Minimiza tion by DNN
Important goal of training: Classify the boundary points between the different classes in the data.
Random Gaussian weights are good for classifying the average points in the data.
DNN may solve optimization problems
Deep learning can be viewed as a metric learning.
Generalization error depends on the DNN input margin

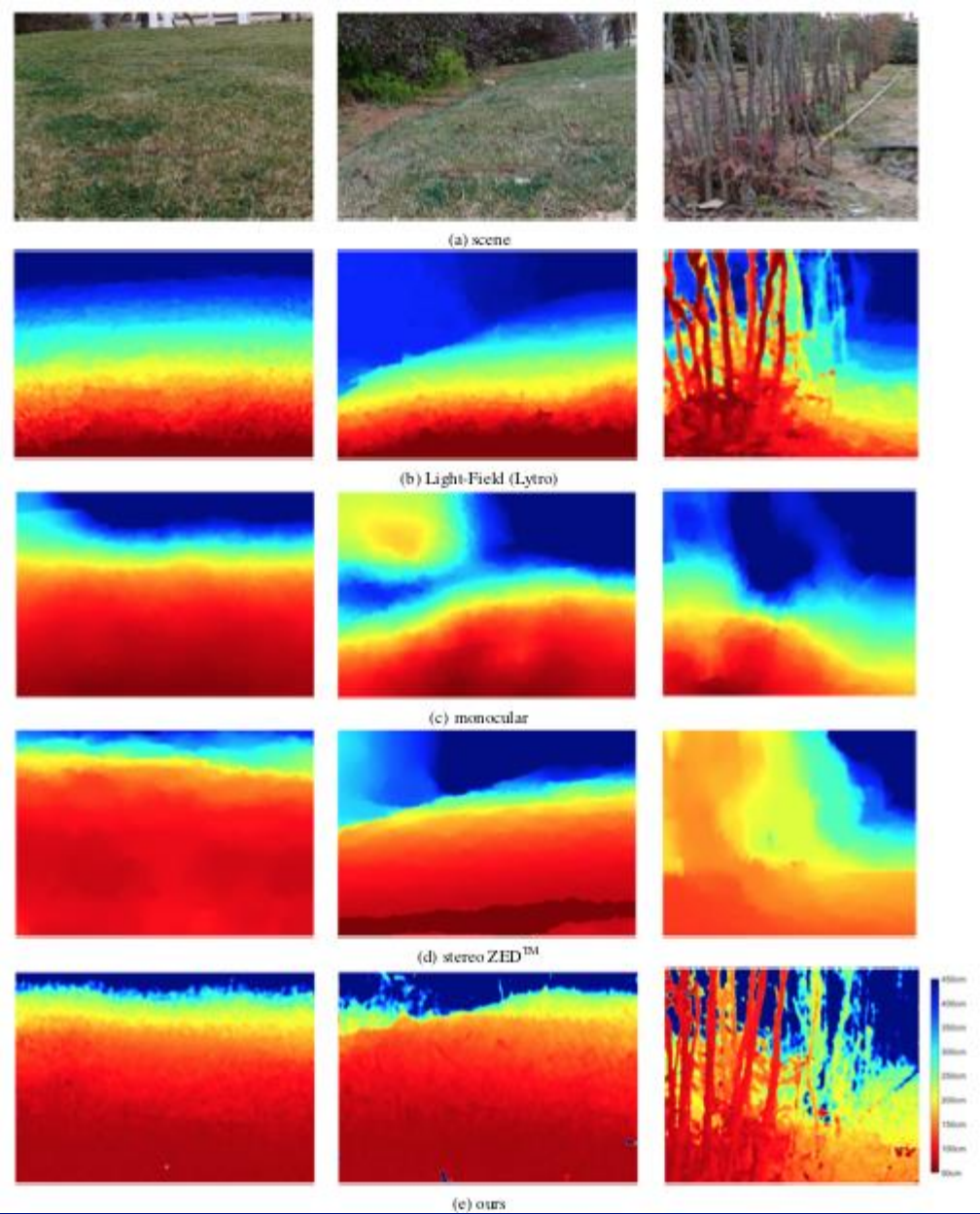
CLASS AWARE DENOISING



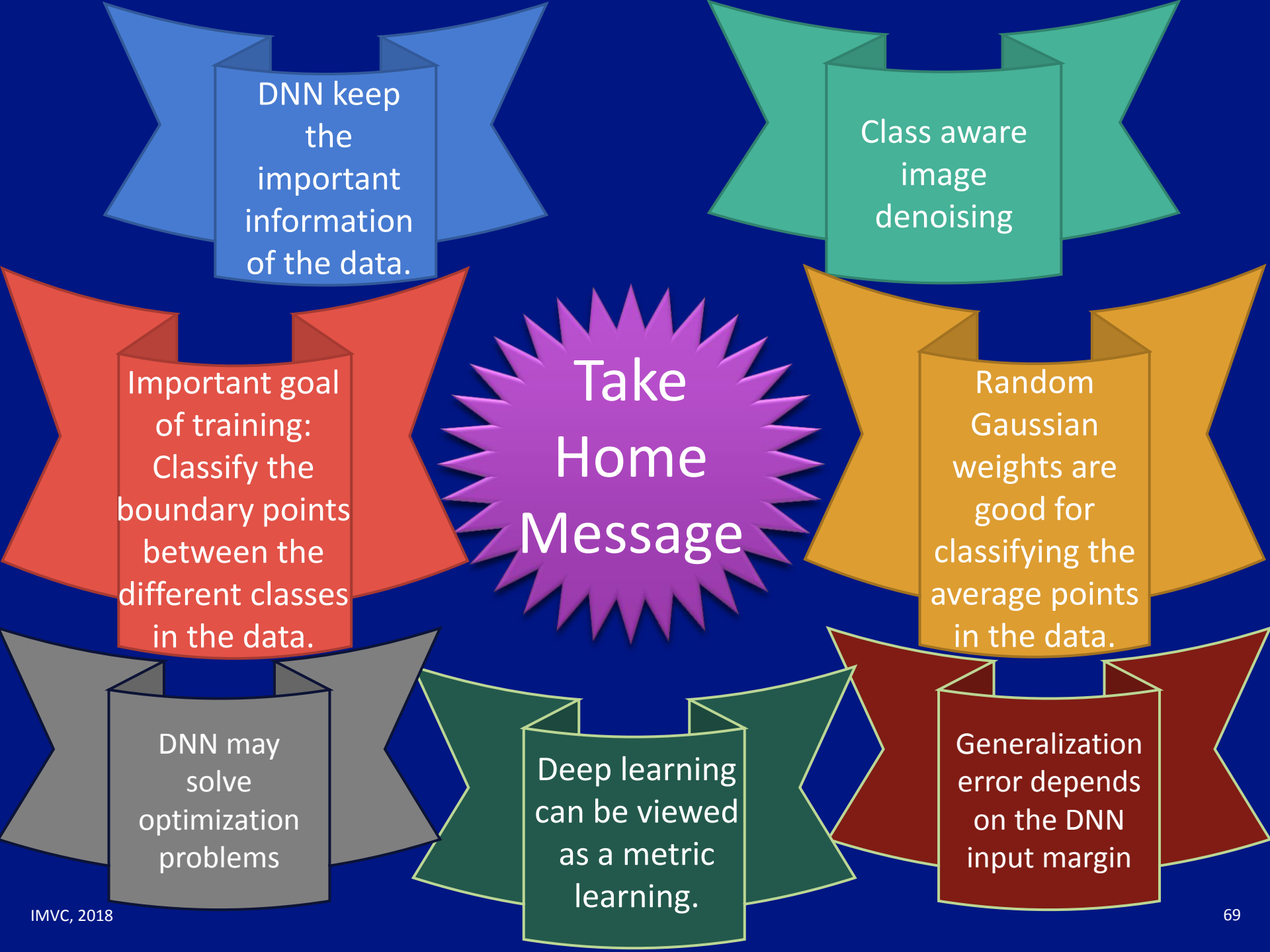


[Remez, Litani, Giryes, Bronstein, 2017]

DEPTH ESTIMATION BY PHASE CODED CUES



[Haim, Elmalem, Bronstein, Marom, Giryes, 2017]


DNN keep the important information of the data.
Class aware image denoising
Take Home Message
Important goal of training: Classify the boundary points between the different classes in the data.
Random Gaussian weights are good for classifying the average points in the data.
DNN may solve optimization problems
Deep learning can be viewed as a metric learning.
Generalization error depends on the DNN input margin

QUESTIONS?

WEB.ENG.TAU.AC.IL/~RAJA

FULL REFERENCES 1

- A. L. Samuel, “Some studies in machine learning using the game of checkers,” *IBM Journal*, vol. 3, no. 3, pp. 535–554, 1959.
- D. H. Hubel & T. N. Wiesel, “Receptive fields of single neurones in the cat's striate cortex”, J Physiol., vol. 148, no. 3, pp. 574-591, 1959.
- D. H. Hubel & T. N. Wiesel, “Receptive fields, binocular interaction and functional architecture in the cat's visual cortex”, J Physiol., vol. 160, no. 1, pp. 106-154, 1962.
- K. Fukushima, “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position”, Biological Cybernetics, vol. 36, no. 4, pp. 93-202, 1980.
- Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard & L. D. Jackel, “Backpropagation Applied to Handwritten Zip Code Recognition”, Neural Computation, vol. 1, no. 4, pp. 541-551, 1989.
- Y.LeCun, L. Bottou, Y. Bengio & P. Haffner, “Gradient Based Learning Applied to Document Recognition”, Proceedings of IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.
- C. Farabet, C. Couprie, L. Najman & Y. LeCun, “Learning Hierarchical Features for Scene Labeling,” IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 35, no. 8, pp. 1915-1929, Aug. 2013.
- O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge”, International Journal of Computer Vision, vol. 115, no. 3, pp. 211-252, 2015
- A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks”, NIPS, 2012.
- K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition”, CVPR, 2016.

FULL REFERENCES 2

- M.D. Zeiler & R. Fergus, “Visualizing and Understanding Convolutional Networks”, ECCV, 2014.
- D. Yu & L. Deng, “Automatic Speech Recognition: A Deep Learning Approach”, Springer, 2014.
- J. Bellegarda & C. Monz, “State of the art in statistical methods for language and speech processing,” Computer Speech and Language, vol. 35, pp. 163–184, Jan. 2016.
- C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke & A. Rabinovich, “Going Deeper with Convolutions”, CVPR, 2015.
- V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra & M. Riedmiller, “Playing Atari with Deep Reinforcement Learning”, NIPS deep learning workshop, 2013.
- V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg & D. Hassabis, “Human-level control through deep reinforcement learning”, Nature vol. 518, pp. 529–533, Feb. 2015.
- D. Silver, A. Huang, C. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel & D. Hassabis, “Mastering the Game of Go with Deep Neural Networks and Tree Search”, Nature, vol. 529, pp. 484–489, 2016.
- S. K. Zhou, H. Greenspan, D. Shen, “Deep Learning for Medical Image Analysis”, Academic Press, 2017.
- I. Sutskever, O. Vinyals & Q. Le, “Sequence to Sequence Learning with Neural Networks”, NIPS 2014.
- A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, “Large-scale Video Classification with Convolutional Neural Networks”, CVPR, 2014.

FULL REFERENCES 3

- F. Schroff, D. Kalenichenko & J. Philbin, “FaceNet: A Unified Embedding for Face Recognition and Clustering”, CVPR, 2015.
- A. Poznanski & L. Wolf, “CNN-N-Gram for Handwriting Word Recognition”, CVPR, 2016.
- R. Socher, A. Perelygin, J. Wu, J. Chuang, C. Manning, A. Ng & C. Potts, “Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, EMNLP, 2013.
- H. C. Burger, C. J. Schuler & S. Harmeling, Image denoising: Can plain Neural Networks compete with BM3D?, CVPR, 2012.
- J. Kim, J. K. Lee, K. M. Lee, “Accurate Image Super-Resolution Using Very Deep Convolutional Networks”, CVPR, 2016.
- J. Bruna, P. Sprechmann, and Y. LeCun, “Super-Resolution with Deep Convolutional Sufficient Statistics”, ICLR, 2016.
- V. Nair & G. E. Hinton, “Rectified linear units improve restricted Boltzmann machines”, ICML, 2010.
- L. Deng & D. Yu, “Deep Learning: Methods and Applications”, Foundations and Trends in Signal Processing, vol. 7 no. 3-4, pp. 197–387, 2014.
- Y. Bengio, “Learning Deep Architectures for AI”, Foundations and Trends in Machine Learning, vol. 2, no. 1, pp. 1–127, 2009.
- Y. LeCun, Y. Bengio, & G. Hinton. Deep learning. Nature, vol. 521, no. 7553, pp. 436–444, 2015.
- J. Schmidhuber, “Deep learning in neural networks: An overview”, Neural Networks, vol. 61, pp. 85–117, Jan. 2015.
- I. Goodfellow, Y. Bengio & A. Courville, “Deep learning”, Book in preparation for MIT Press, 2016.

FULL REFERENCES 4

- G. Cybenko, “Approximation by superpositions of a sigmoidal function,” Math. Control Signals Systems, vol. 2, pp. 303–314, 1989.
- K. Hornik, “Approximation capabilities of multilayer feedforward networks,” Neural Netw., vol. 4, no. 2, pp. 251–257, 1991.
- A. R. Barron, Approximation and estimation bounds for artificial neural networks, Machine Learning, vol. 14, no. 1, pp. 115–133, Jan. 1994.
- G. F. Montúfar & J. Morton, “When does a mixture of products contain a product of mixtures”, SIAM Journal on Discrete Mathematics (SIDMA), vol. 29, no. 1, pp. 321-347, 2015.
- G. F. Montúfar, R. Pascanu, K. Cho, & Y. Bengio, “On the number of linear regions of deep neural networks,” NIPS, 2014.
- N. Cohen, O. Sharir & A. Shashua, “Deep SimNets,” CVPR, 2016.
- N. Cohen, O. Sharir & A. Shashua, “On the Expressive Power of Deep Learning: A Tensor Analysis,” COLT, 2016.
- N. Cohen & A. Shashua, “Convolutional Rectifier Networks as Generalized Tensor Decompositions,” ICML, 2016
- M. Telgarsky, “Benefits of depth in neural networks,” COLT, 2016.
- R. Eldan and O. Shamir, “The power of depth for feedforward neural networks.,” COLT, 2016.
- N. Cohen and A. Shashua, “Inductive Bias of Deep Convolutional Networks through Pooling Geometry,” arXiv abs/ 1605.06743, 2016.
- J. Bruna, Y. LeCun, & A. Szlam, “Learning stable group invariant representations with convolutional networks,” ICLR, 2013.
- Y-L. Boureau, J. Ponce, Y. LeCun, Theoretical Analysis of Feature Pooling in Visual Recognition, ICML, 2010.

FULL REFERENCES 5

- J. Bruna, A. Szlam, & Y. LeCun, “Signal recovery from lp pooling representations”, ICML, 2014.
- S. Soatto & A. Chiuso, “Visual Representations: Defining properties and deep approximation”, ICLR 2016.
- F. Anselmi, J. Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, and T. Poggio, “Unsupervised learning of invariant representations in hierarchical architectures,” Theoretical Computer Science, vol. 663, no. C, pp. 112-121, Jun. 2016.
- J. Bruna and S. Mallat, “Invariant scattering convolution networks,” IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI), vol. 35, no. 8, pp. 1872–1886, Aug 2013.
- T. Wiatowski and H. Bölcskei, “A Mathematical Theory of Deep Convolutional Neural Networks for Feature Extraction,” arXiv abs/1512.06293, 2016
- A. Saxe, J. McClelland, and S. Ganguli, “Exact solutions to the nonlinear dynamics of learning in deep linear neural network”, ICLR, 2014.
- Y. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, “Identifying and attacking the saddle point problem in high dimensional non-convex optimization,” NIPS, 2014.
- A. Choromanska, M. B. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, “The loss surfaces of multilayer networks,” in International Conference on Artificial Intelligence and Statistics (AISTATS), 2015.
- B. D. Haeffele and R. Vidal. Global Optimality in Tensor Factorization, Deep Learning, and Beyond. arXiv, abs/1506.07540, 2015.
- S. Arora, A. Bhaskara, R. Ge, and T. Ma, “Provable bounds for learning some deep representations,” in Int. Conf. on Machine Learning (ICML), 2014, pp. 584–592.

FULL REFERENCES 6

- A. M. Bruckstein, D. L. Donoho, & M. Elad, “From sparse solutions of systems of equations to sparse modeling of signals and images”, SIAM Review, vol. 51, no. 1, pp. 34–81, 2009.
- G. Yu, G. Sapiro & S. Mallat, “Solving inverse problems with piecewise linear estimators: From Gaussian mixture models to structured sparsity”, IEEE Trans. on Image Processing, vol. 21, no. 5, pp. 2481 –2499, May 2012.
- N. Srebro & A. Shraibman, “Rank, trace-norm and max-norm,” COLT, 2005.
- E. Cand`es & B. Recht, “Exact matrix completion via convex optimization,” Foundations of Computational mathematics, vol. 9, no. 6, pp. 717– 772, 2009.
- R. G. Baraniuk, V. Cevher & M. B. Wakin, "Low-Dimensional Models for Dimensionality Reduction and Signal Recovery: A Geometric Perspective," Proceedings of the IEEE, vol. 98, no. 6, pp. 959-971, 2010.
- Y. Plan and R. Vershynin, “Dimension reduction by random hyperplane tessellations,” Discrete and Computational Geometry, vol. 51, no. 2, pp. 438–461, 2014.
- Y. Plan and R. Vershynin, “Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach,” IEEE Trans. Inf. Theory, vol. 59, no. 1, pp. 482–494, Jan. 2013.
- R. Giryes, G. Sapiro and A.M. Bronstein, “Deep Neural Networks with Random Gaussian Weights: A Universal Classification Strategy? “, IEEE Transactions on Signal Processing, vol. 64, no. 13, pp. 3444-3457, Jul. 2016.
- A. Choromanska, K. Choromanski, M. Bojarski, T. Jebara, S. Kumar, Y. LeCun, “Binary embeddings with structured hashed projections”, ICML, 2016.
- J. Masci, M. M. Bronstein, A. M. Bronstein and J. Schmidhuber, “Multimodal Similarity-Preserving Hashing”, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 36, no. 4, pp. 824-830, April 2014.

FULL REFERENCES 7

- H. Lai, Y. Pan, Y. Liu & S. Yan, “Simultaneous Feature Learning and Hash Coding With Deep Neural Networks”, CVPR, 2015.
- A. Mahendran & A. Vedaldi, “Understanding deep image representations by inverting them,” CVPR, 2015.
- K. Simonyan & A. Zisserman, “Very deep convolutional networks for large-scale image recognition”, ICLR, 2015
- A. Krogh & J. A. Hertz, “A Simple Weight Decay Can Improve Generalization”, NIPS, 1992.
- P. Baldi & P. Sadowski, “Understanding dropout”, NIPS, 2013.
- N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, & R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.
- L. Wan, M. Zeiler, S. Zhang, Y. LeCun & R. Fergus, “Regularization of Neural Networks using DropConnect”, ICML, 2013.
- S. Ioffe & C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” ICML, 2015.
- M. Hardt, B. Recht & Y. Singer, “Train faster, generalize better: Stability of stochastic gradient descent”, ICML, 2016.
- B. Neyshabur, R. Salakhutdinov & N. Srebro, “Path-SGD: Path-normalized optimization in deep neural networks,” NIPS, 2015.
- S. Rifai, P. Vincent, X. Muller, X. Glorot, & Y. Bengio. “Contractive auto-encoders: explicit invariance during feature extraction,” ICML, 2011.
- S. Zucker & R. Giryes, “Shallow Transits – Deep Learning I: Feasibility Study of Deep Learning to Detect Periodic Transits of Exoplanets”, to appear in The Astronomical Journal, 2018.

FULL REFERENCES 8

- T. Salimans & D. Kingma, “Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks”, arXiv abs/1602.07868, 2016.
- S. Sun, W. Chen, L. Wang, & T.-Y. Liu, “Large margin deep neural networks: theory and algorithms”, AAAI, 2016.
- S. Shalev-Shwartz & S. Ben-David. “Understanding machine learning: from theory to algorithms”, Cambridge University Press, 2014.
- P. L. Bartlett & S. Mendelson, “Rademacher and Gaussian complexities: risk bounds and structural results”. The Journal of Machine Learning Research (JMLR), vol 3, pp. 463–482, 2002.
- B. Neyshabur, R. Tomioka, and N. Srebro, “Norm-based capacity control in neural networks,” COLT, 2015.
- J. Sokolic, R. Giryes, G. Sapiro, M. R. D. Rodrigues, “Robust Large Margin Deep Neural Networks”, IEEE Transactions on Signal Processing, 2017.
- J. Sokolic, R. Giryes, G. Sapiro, M. R. D. Rodrigues, “Generalization Error of Invariant Classifiers”, AISTATS, 2017.
- H. Xu and S. Mannor. “Robustness and generalization,” JMLR, vol. 86, no. 3, pp. 391–423, 2012.
- J. Huang, Q. Qiu, G. Sapiro, R. Calderbank, “Discriminative Geometry-Aware Deep Transform”, ICCV 2015
- J. Huang, Q. Qiu, G. Sapiro, R. Calderbank, “Discriminative Robust Transformation Learning”, NIPS 2016.
- T. Blumensath & M.E. Davies, “Iterative hard thresholding for compressed sensing”, Appl. Comput. Harmon. Anal, vol. 27, no. 3, pp. 265 – 274, 2009.
- I. Daubechies, M. Defrise, and C. De Mol, “An iterative thresholding algorithm for linear inverse problems with a sparsity constraint”, Communicationson Pure and Applied Mathematics, vol. 57, no. 11, pp. 1413–1457, 2004.

FULL REFERENCES 9

- K. Gregor & Y. LeCun, “Learning fast approximations of sparse coding”, ICML, 2010.
- P. Sprechmann, A. M. Bronstein & G. Sapiro, “Learning efficient sparse and low rank models”, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1821–1833, Sept. 2015.
- T. Remez, O. Litany, & A. M. Bronstein, “A picture is worth a billion bits: Real-time image reconstruction from dense binary pixels”, ICCP, 2015.
- J. Tompson, K. Schlachter, P. Sprechmann & K. Perlin, “Accelerating Eulerian Fluid Simulation With Convolutional Networks”, arXiv, abs/1607.03597, 2016.
- S. Oymak, B. Recht, & M. Soltanolkotabi, “Sharp time–data tradeoffs for linear inverse problems”, IEEE Transactions on Information Theory, 2018.
- R. Giryes, Y. C. Eldar, A. M. Bronstein, G. Sapiro, “Tradeoffs between Convergence Speed and Reconstruction Accuracy in Inverse Problems”, IEEE Transactions on Signal Processing, 2018.
- R.G. Baraniuk, V. Cevher, M.F. Duarte & C. Hegde, “Model-based compressive sensing”, IEEE Trans. Inf. Theory, vol. 56, no. 4, pp. 1982–2001, Apr. 2010.
- T. Remez, O. Litany, R. Giryes, A. M. Bronstein, “Class aware image denoising”, arXiv, 2017
- H. Haim, S. Elmalem, A. M. Bronstein, E. Marom & R. Giryes, “Depth Estimation from a Single Phase Coded Aperture Image using Deep Fully Convolutional Networks”, 2017
- J. Bruna & T. Moreau, Adaptive Acceleration of Sparse Coding via Matrix Factorization, ICLR, 2017.
- O. Nabati, D. Mendelovic & R. Giryes, Fast and Accurate Reconstruction of Compressed Color Light Field, ICCP, 2018
- E. Schwartz, R. Giryes and A. M. Bronstein, “DeepISP: Learning End-to-End Image Processing Pipeline”, arXiv, 2018