MATHEMATICS OF DEEP LEARNING

RAJA GIRYES

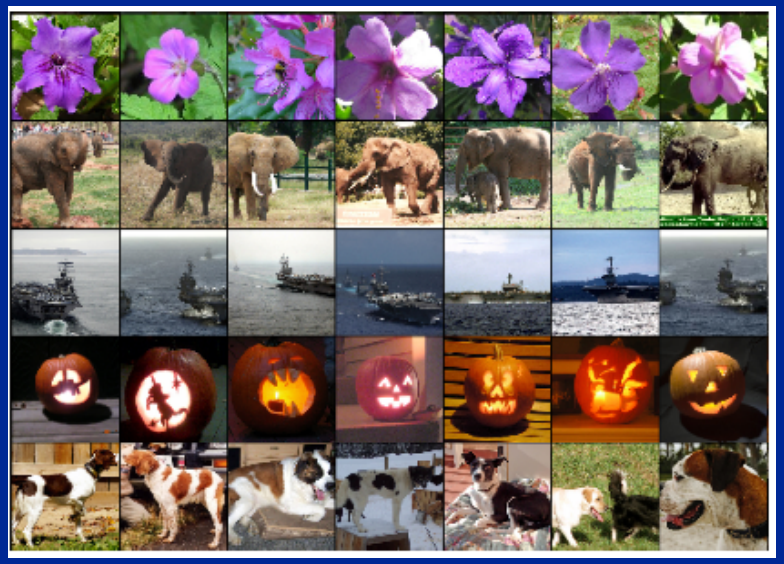
TEL AVIV UNIVERSITY

Mathematics of Deep Learning Tutorial

CDC, Melbourne, Australia

December 15, 2017

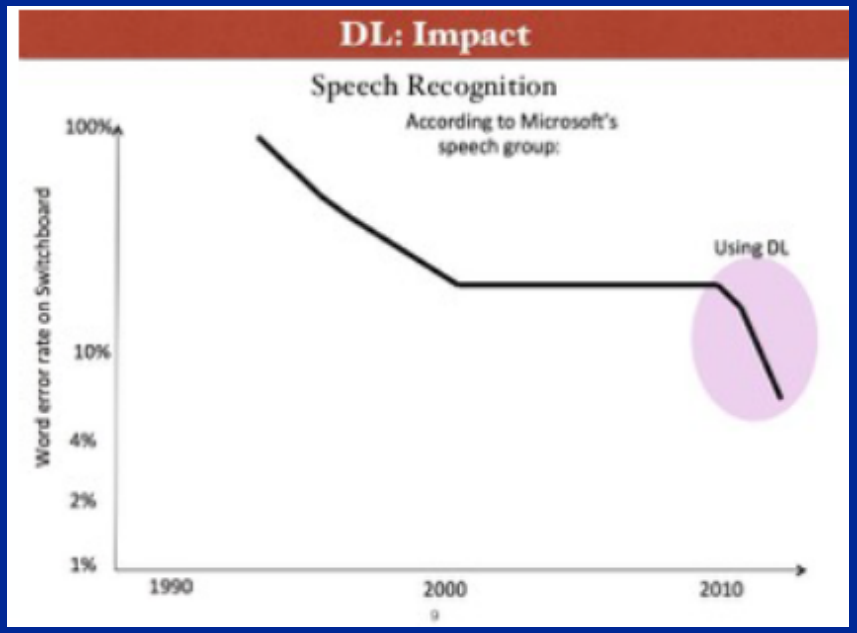
DEEP LEARNING IMPACT



- 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%**

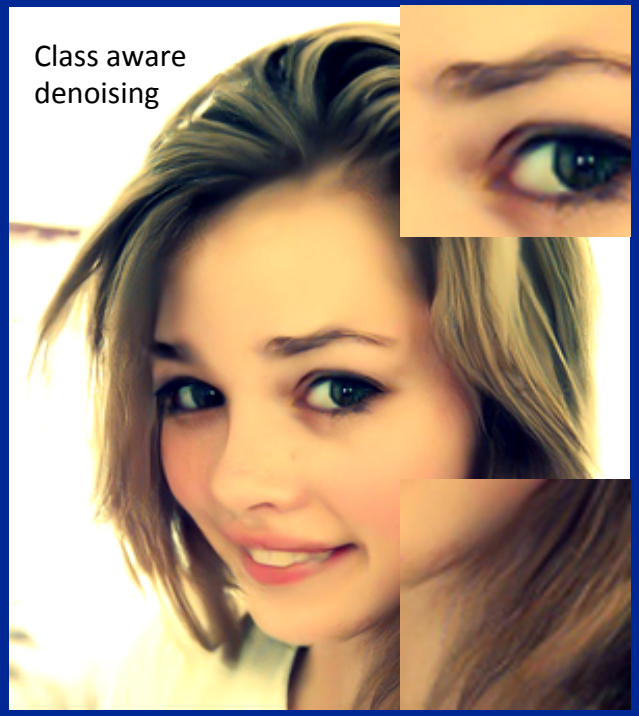
Today we get 3.5% by 152 layers

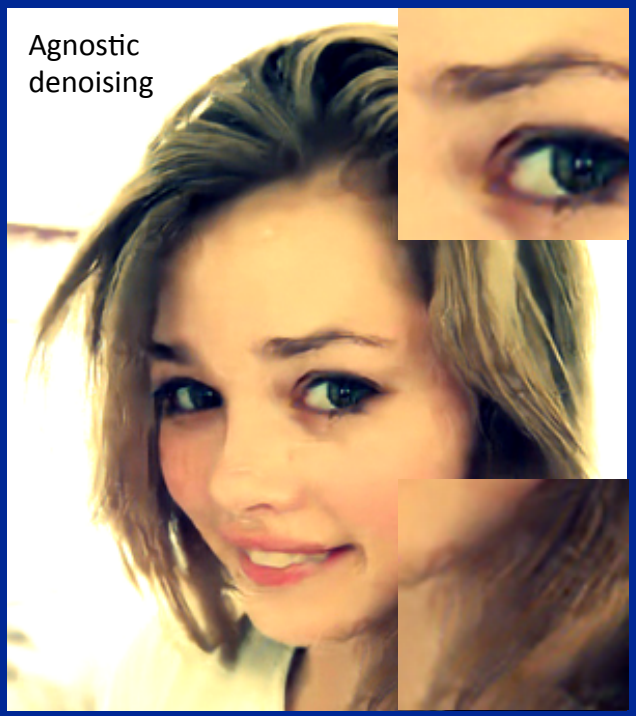


CUTTING EDGE PERFORMANCE IN MANY OTHER APPLICATIONS

- 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

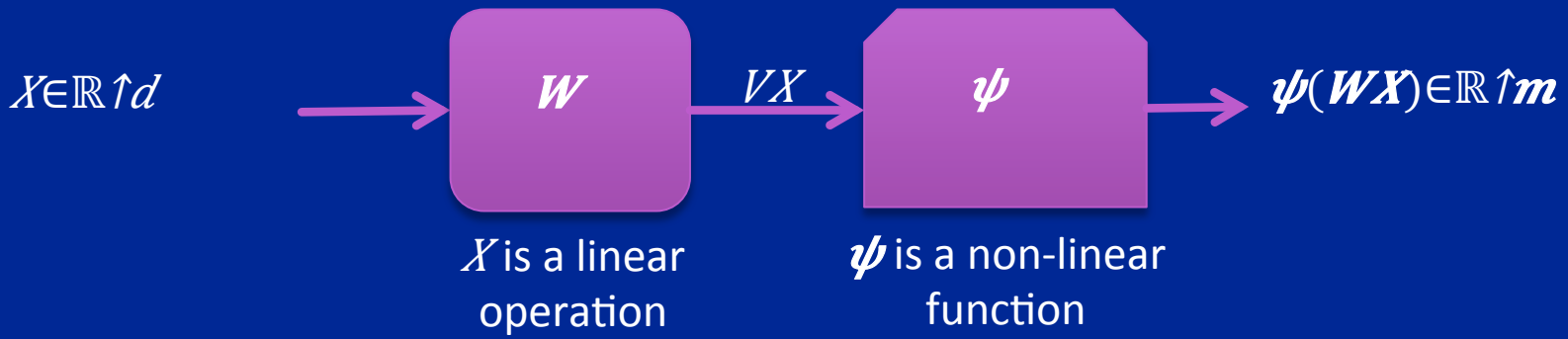




[Remez, Litani, Giryes, Bronstein, 2017]

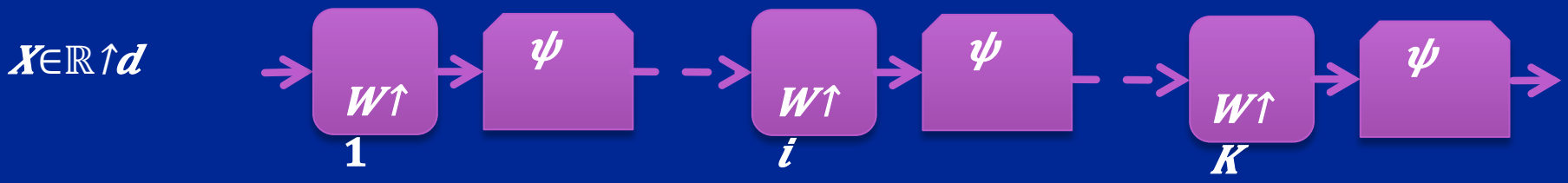
DEEP NEURAL NETWORKS (DNN)

- One layer of a neural net

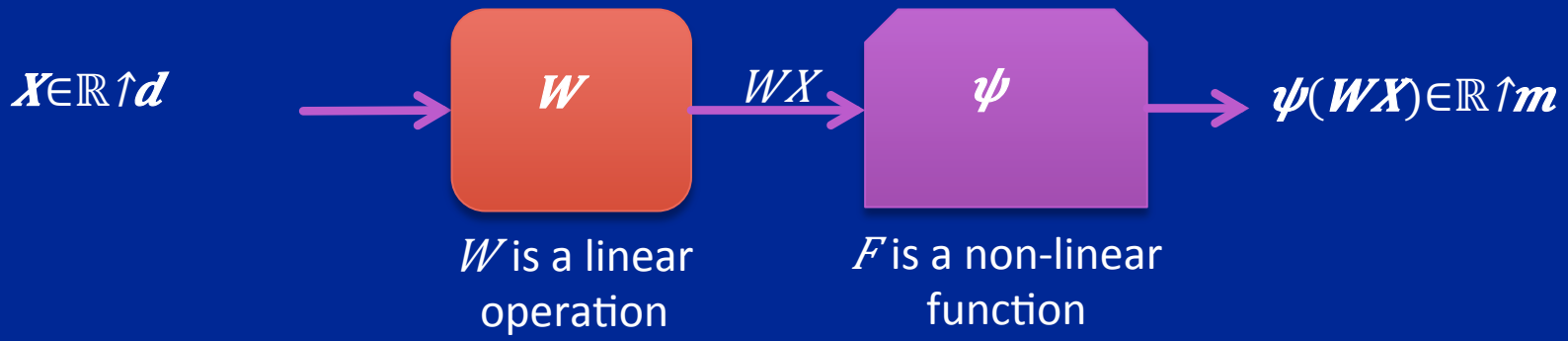


- Concatenation of the layers creates the whole net

$$\Phi(W^1, W^2, \ldots, W^K) = \psi(W^K \ldots \psi(W^2 \psi(W^1 X)))$$



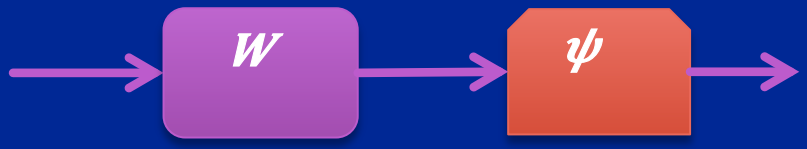
CONVOLUTIONAL NEURAL NETWORKS (CNN)



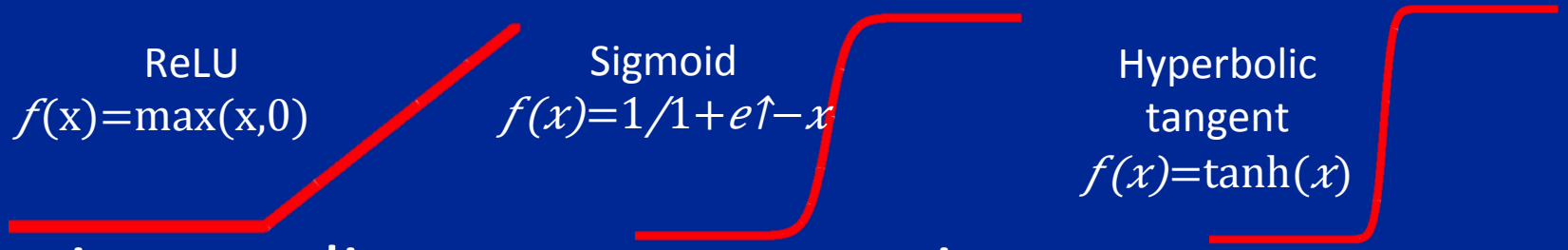
- In many cases, W is selected to be a convolution.
- This operator is shift invariant.
- CNN are commonly used with images as they are typically shift invariant.

THE NON-LINEAR PART

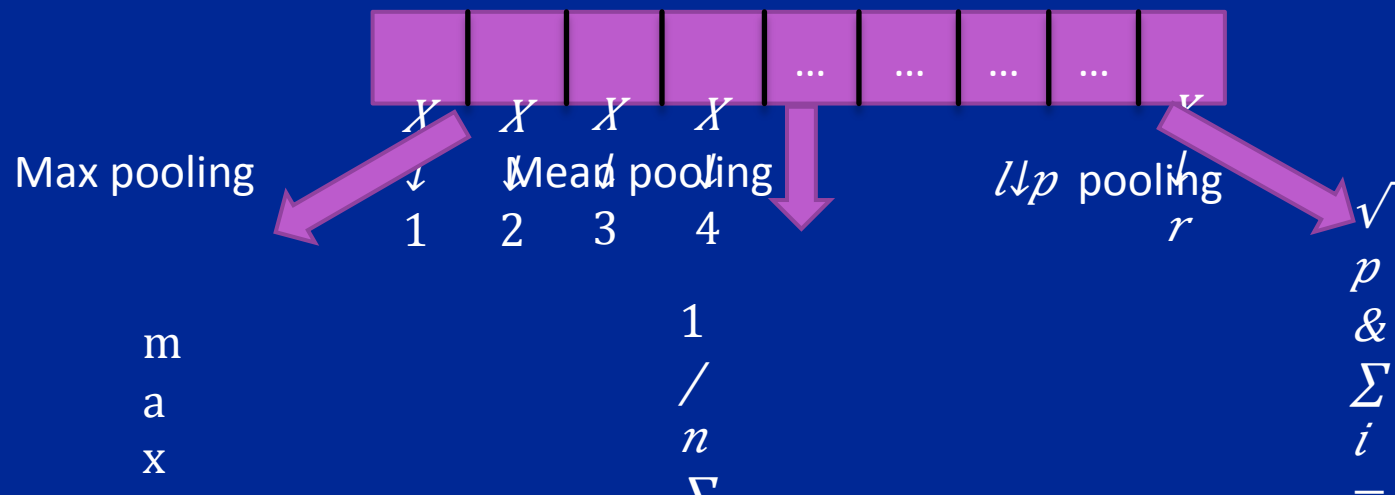
- Usually $\psi = g \circ f$.

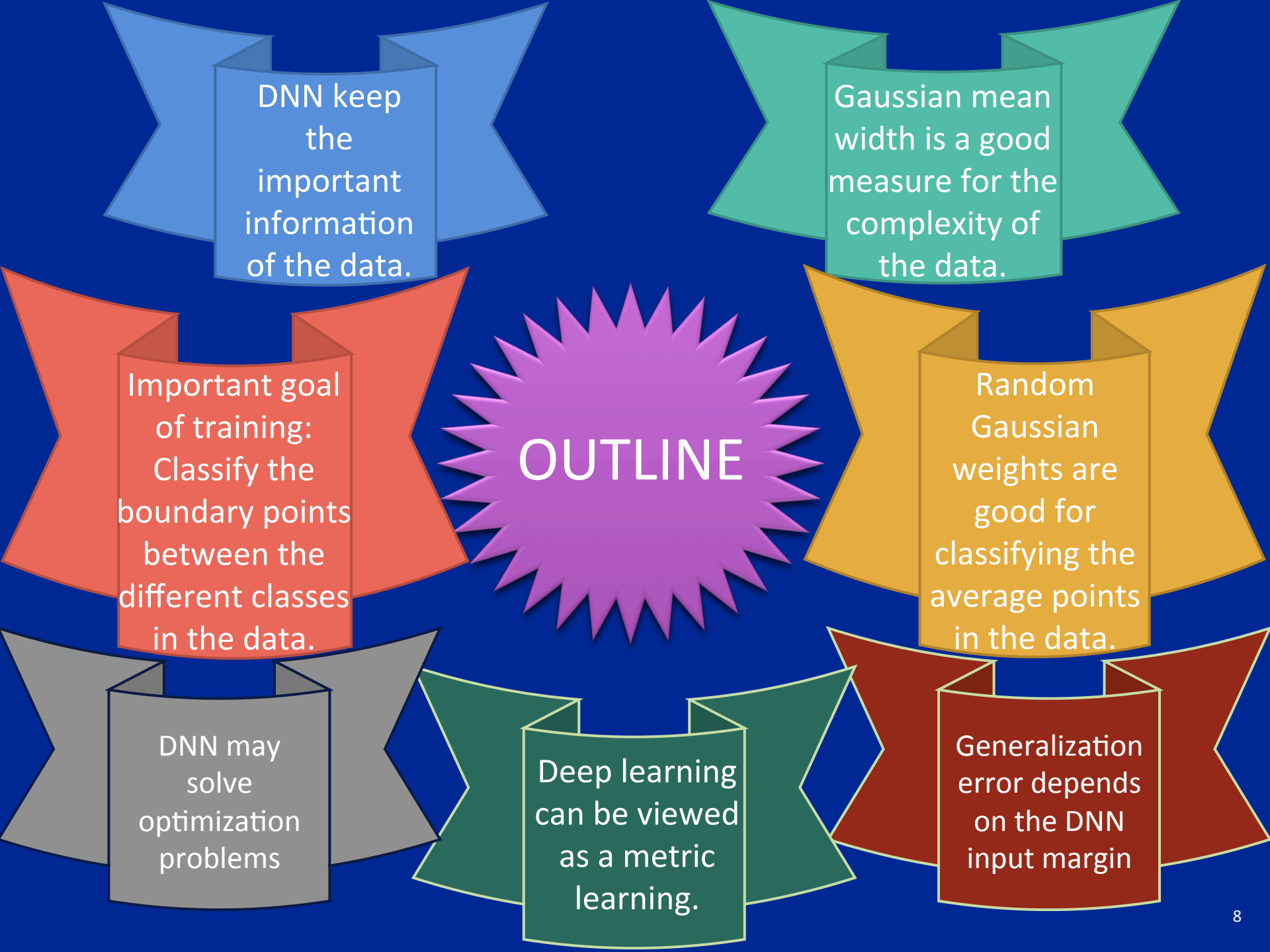


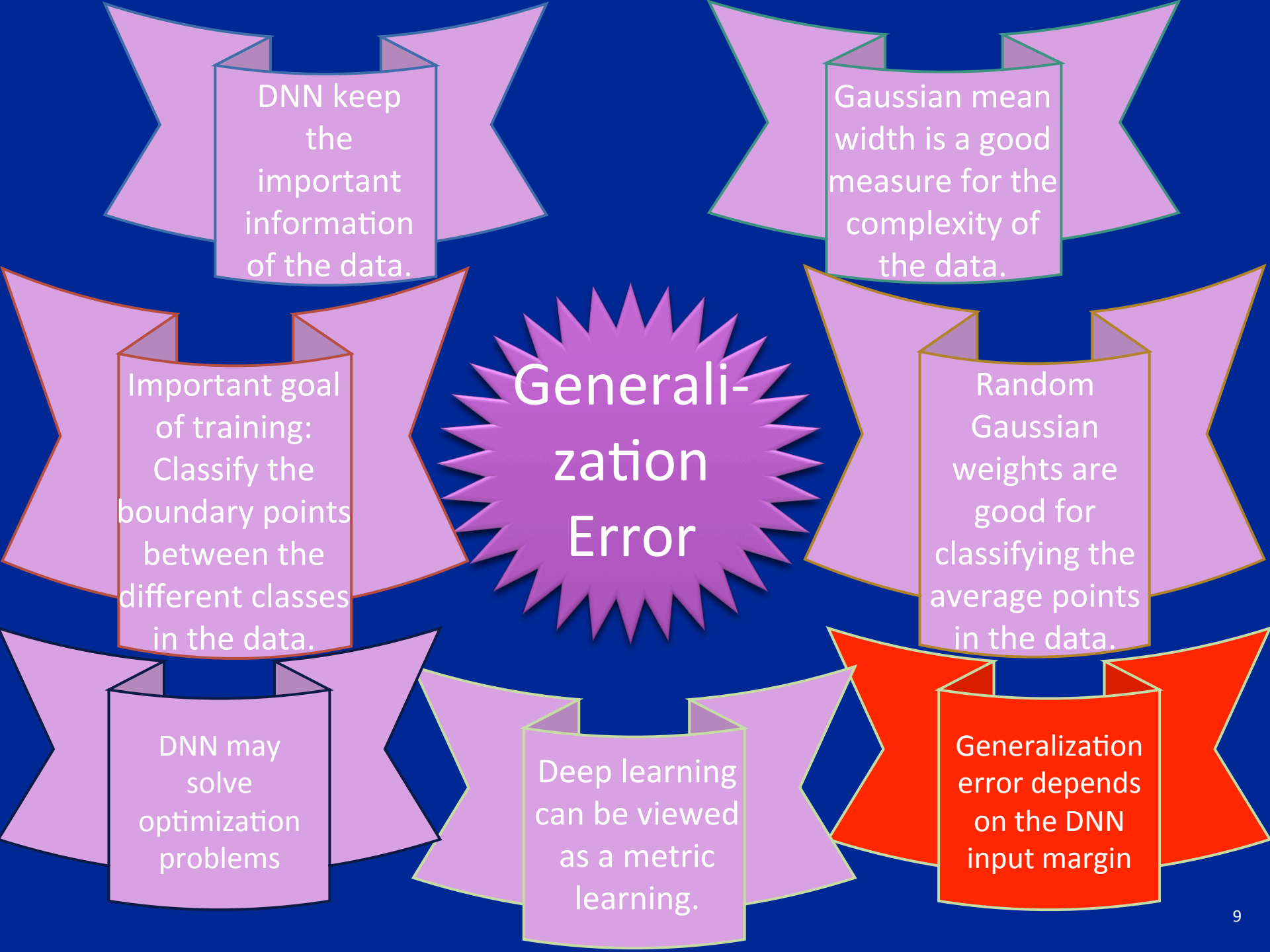
- f is the (point-wise) activation function



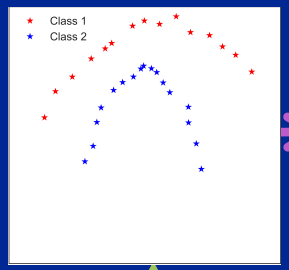
- g is a pooling or an aggregation operator.




DNN keep the important information of the data.
Gaussian mean width is a good measure for the complexity of the data.
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

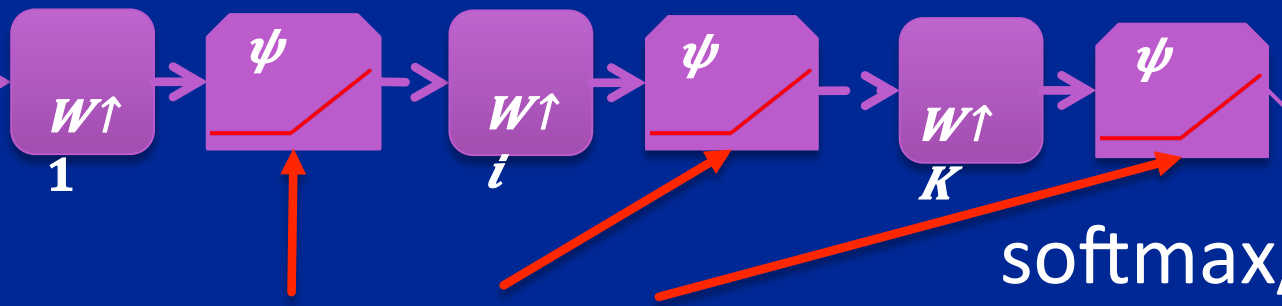

DNN keep the important information of the data.
Gaussian mean width is a good measure for the complexity of the data.
Important goal of training: Classify the boundary points between the different classes in the data.
Generali-zation Error
Random Gaussian weights are good for classifying the average 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

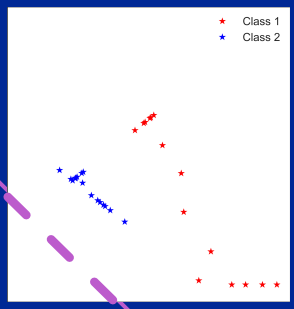


Two Classes

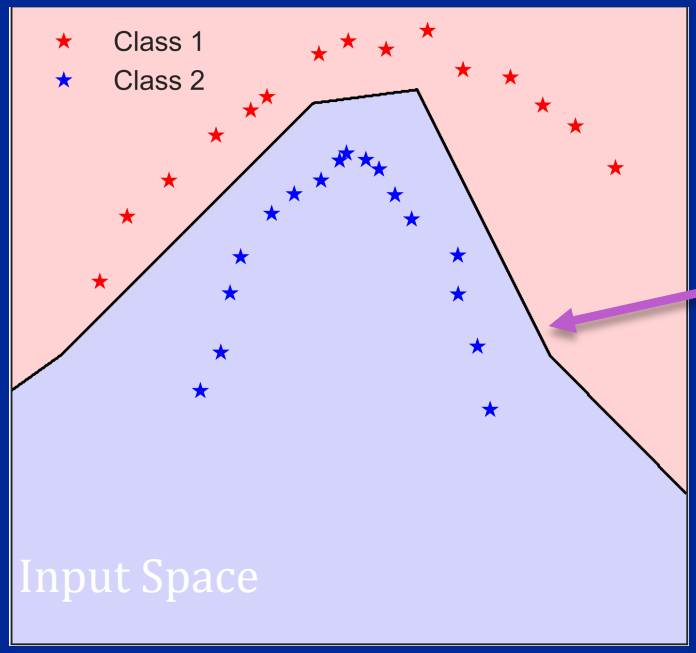
$\in r$



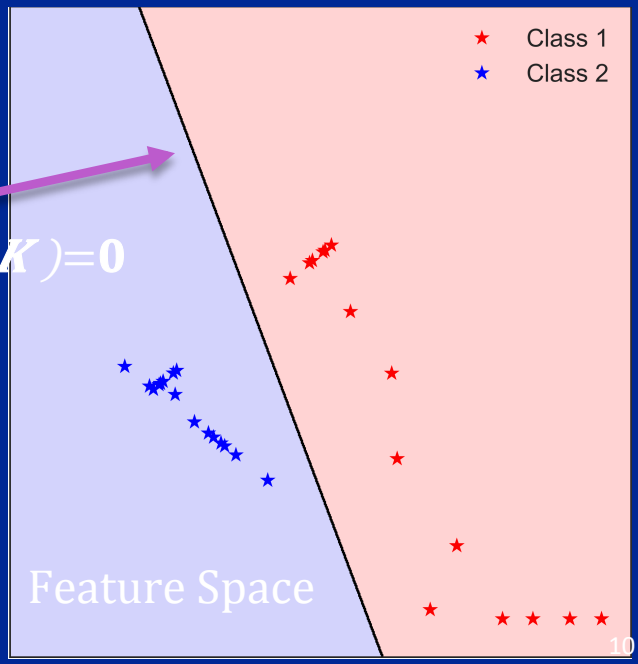
general non-linearity (ReLU, pooling,...)



softmax/ linear Classifier



$w↑T\ \Phi(W↑1, W↑2, ..., W↑K) = 0$



GENERALIZATION ERROR (GE)

- In training, we reduce the classification error $\ell\downarrow\text{training}$ of the training data as the number of training examples L increases.
- However, we are interested to reduce the error $\ell\downarrow\text{test}$ of the (unknown) testing data as L increases.
- The difference between the two is the generalization error
  $\text{GE} = \ell\downarrow\text{training} - \ell\downarrow\text{test}$
- It is important to understand the GE of DNN

ESTIMATION ERROR

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

$$O(C_f/N)+O(Nd/L\log(L))$$

Smoothness of approximated function

Input dimension

Number of neurons in the DNN

Number of training examples

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

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

  [Bartlett et al. 1998, Shalev-Shwartz and Ben-David, 2014, Bartlett 2017, Harvey et al. 2017]
- The GE was bounded also by the DNN weights

  $$\mathrm{GE} \leq 1/\sqrt{L}\ 2^{K-1} \|w\|_2 \prod_i \|W^i\|_{2,2}$$

  [Bartlett 1998, Neyshabur et al., 2015].
- Note that in both cases the GE grows with the depth

RETHINKING GENERALIZATION

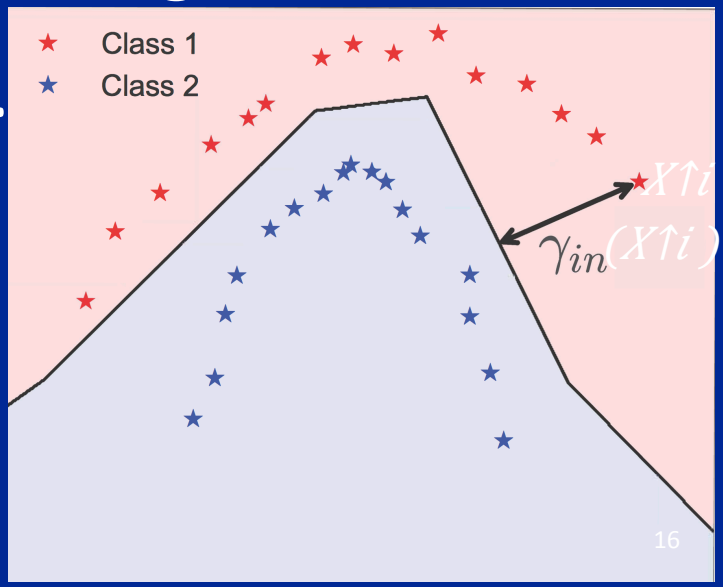
- Networks with the same architecture may generalize well with structured data but overfit if the data is given with random labels [Zhang et al., 2017].
- This phenomena is affected by explicit regularization.
- This shows that taking into account only the network structure for bouding the generalization error is misleading
- We need to seek an alternative to the Rademacher Complexity and VC-dimension based bounds

DNN INPUT MARGIN

- Theorem 6: If for every input margin $\gamma\downarrow in\ (X\uparrow i\ ) > \gamma$

  then $GE \leq \sqrt{N\downarrow \gamma/2\ (\Upsilon)\ /L}$ [Sokolic, Giryes, Sapiro, Rodrigues, 2017]

- $N\downarrow \gamma/2\ (\Upsilon)$ is the covering number of the data Υ.
- $N\downarrow \gamma/2\ (\Upsilon)$ gets smaller as γ gets larger.
- Bound is independent of depth.
- Our theory relies on the robustness framework [Xu & Mannor, 2012].



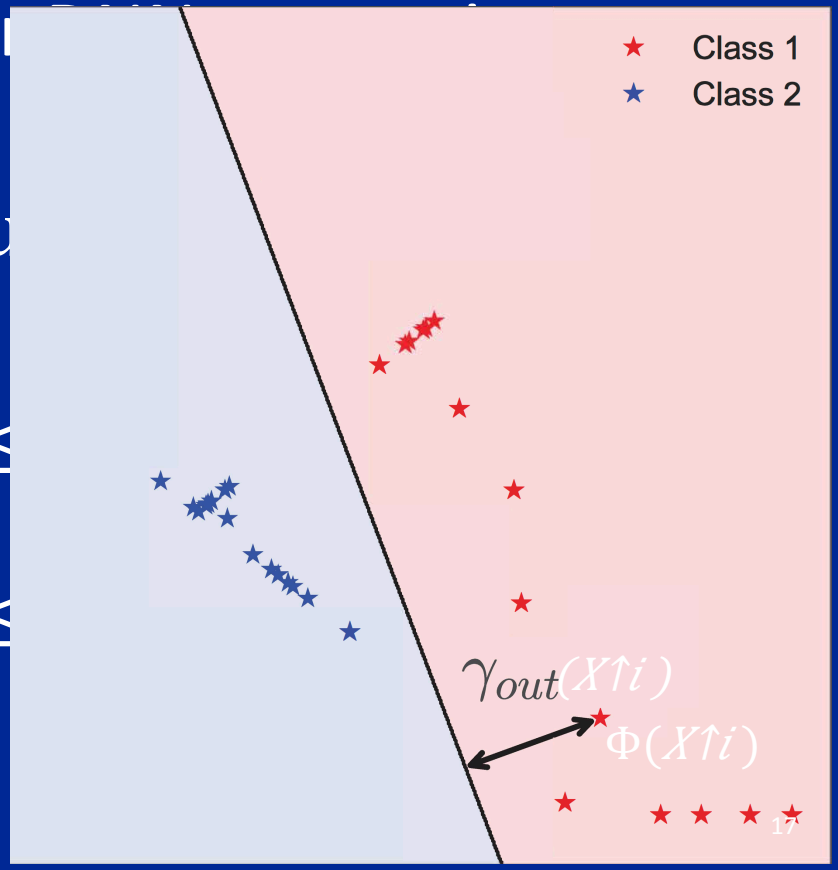
INPUT MARGIN BOUND

- Maximizing the input margin directly is hard
- Our strategy: relate the input margin to the output margin $\gamma\downarrow out\,(X\uparrow i\,)$ and other
- Theorem 7:

$$\gamma\downarrow in\,(X\uparrow i\,)\geq\gamma\downarrow out\,(X\uparrow i\,)/\mathrm{su}$$
$$J(X)\|\downarrow 2$$

$$\geq\gamma\downarrow out\,(X\uparrow i\,)/\prod 1\leq$$

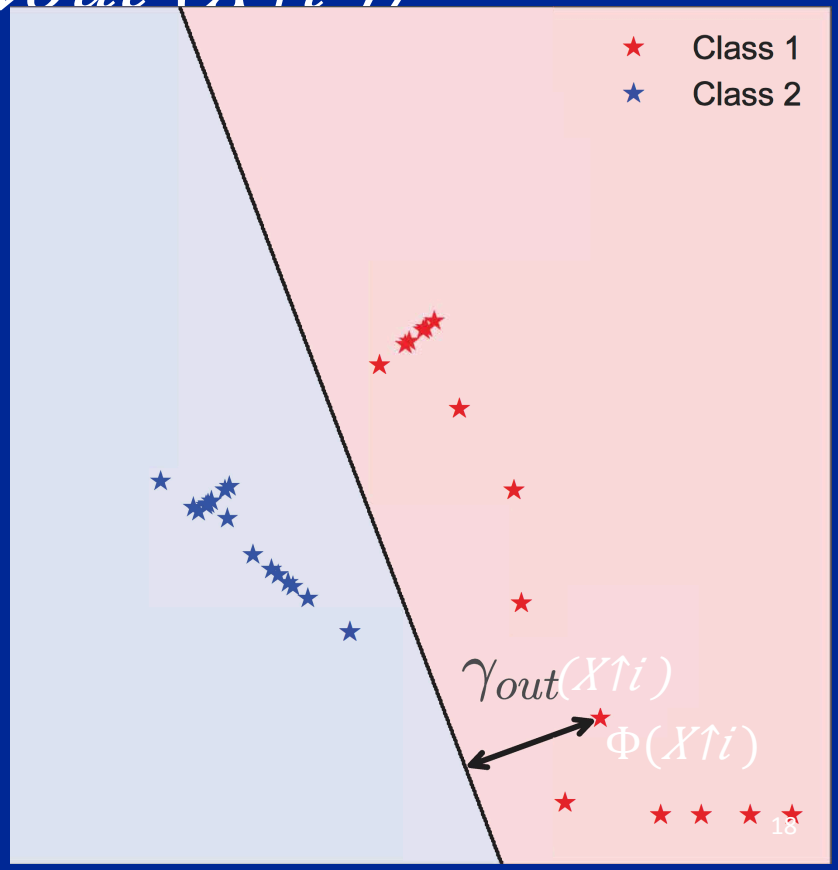
$$\geq\gamma\downarrow out\,(X\uparrow i\,)/\prod 1\leq$$



[Sokolic, Giryes, Sapiro, Rodrigues, 2017]

OUTPUT MARGIN

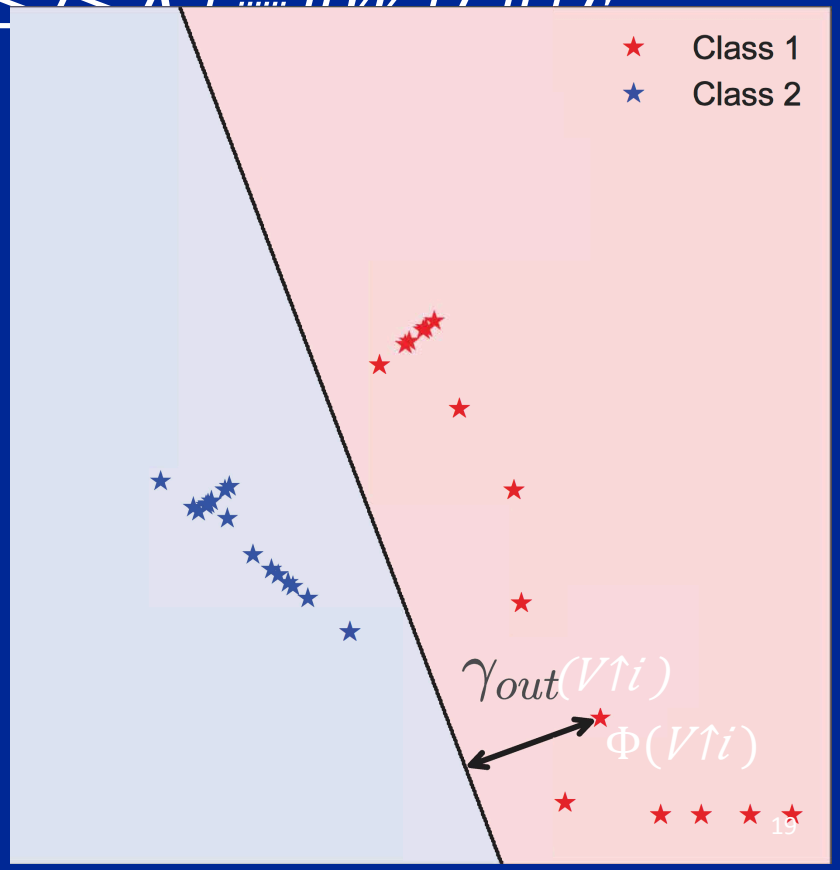
- Theorem 7: $\gamma_{in}(X^i) \geq \gamma_{out}(X^i) / \sup_{V \in Y} \|X/\|X\|_2 \; J(X)\|_2 \geq \gamma_{out}(X^i) / \prod_{1 \leq i \leq K} \|W^i\|_2 \geq \gamma_{out}(X^i) / \prod_{1 \leq i \leq K} \|W^i\|_F$
- Output margin is easier to maximize – SVM problem
- Maximized by many cost functions, e.g., hinge loss.



GE AND WEIGHT DECAY

- Theorem 7: $\gamma\downarrow in\,(X\uparrow i\,)\geq\gamma\downarrow out\,(X\uparrow i\,)/\sup{-}V\in\Upsilon\,\|X/\|X\|\downarrow 2\;J(X)\|\downarrow 2\;\geq\gamma\downarrow out\,(X\uparrow i\,)/\prod 1\leq i\leq K\uparrow\square\|W\uparrow i\,\|\downarrow 2\;\geq\gamma\downarrow out\,(X\uparrow i\,)/\prod 1\leq i\leq K\uparrow\square\|W\uparrow i\,\|_F$

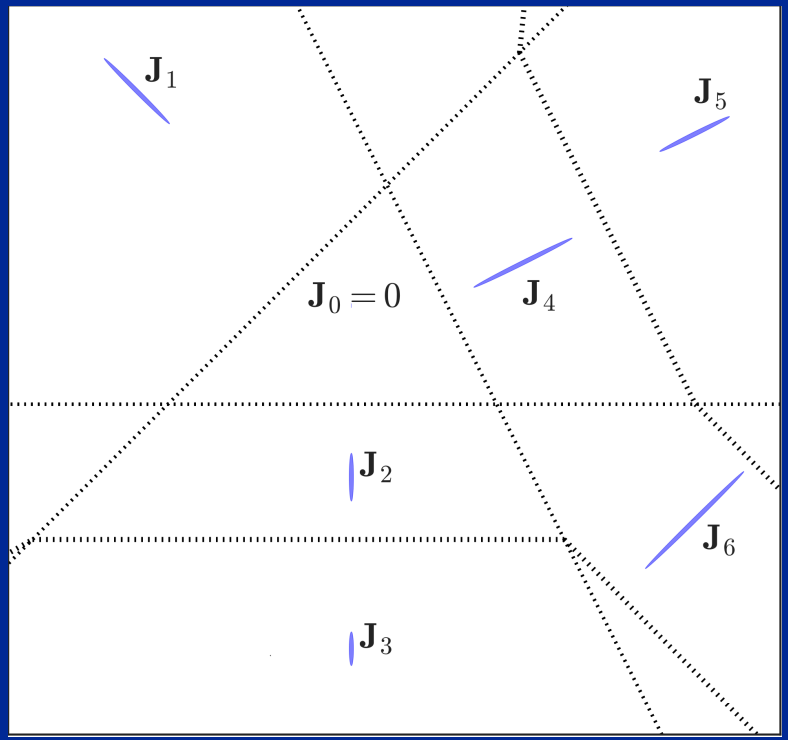
- Bounding the weights increases the input margin
- Weight decay regularization decreases the GE
- Related to regularization used by [Haeffele & Vidal, 2015]



JACOBIAN BASED REGULARIZATION

- Theorem 7: $\gamma\downarrow in\,(X\uparrow i\,)\geq\gamma\downarrow out\,(X\uparrow i\,)/\sup{\scriptstyle -}V\in\Upsilon\,\|X/\|X\|\downarrow 2\;J(X)\|\downarrow 2\;\geq\gamma\downarrow out\,(X\uparrow i\,)/\prod 1\leq i\leq K\uparrow\square\|W\uparrow i\,\|\downarrow 2\;\geq\gamma\downarrow out\,(X\uparrow i\,)/\prod 1\leq i\leq K\uparrow\square\|W\uparrow i\,\|\downarrow F$

- $J(X)$ is the Jacobian of the DNN at point X.
- $J(\cdot)$ 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.

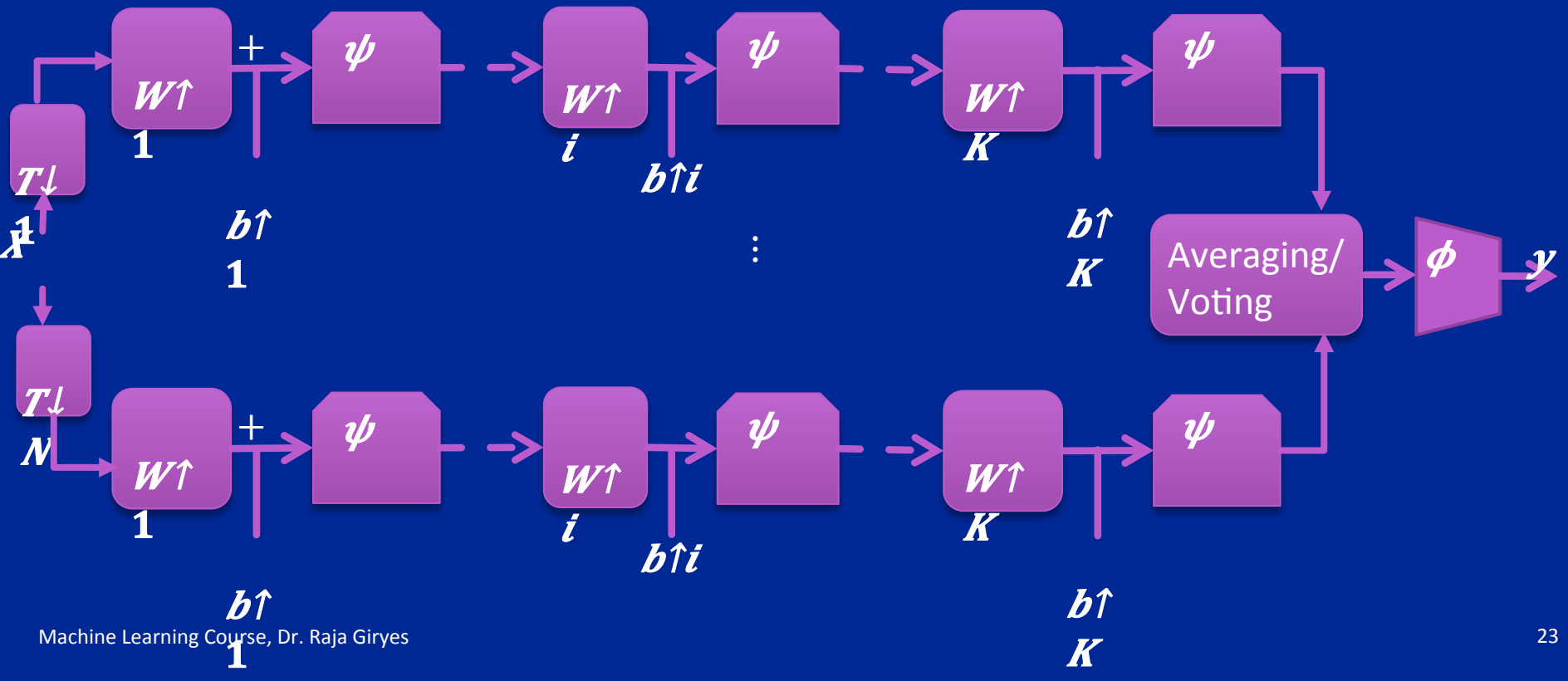
[Sokolic, Giryes, Sapiro, Rodrigues, 2017]

INVARIANCE

- Our theory extends also to study of the relation between invariance in the data and invariance in the network
- We have proposed also a new strategy to enforce invariance in the network [Sokolic, Giryes, Sapiro, Rodrigues, 2017]

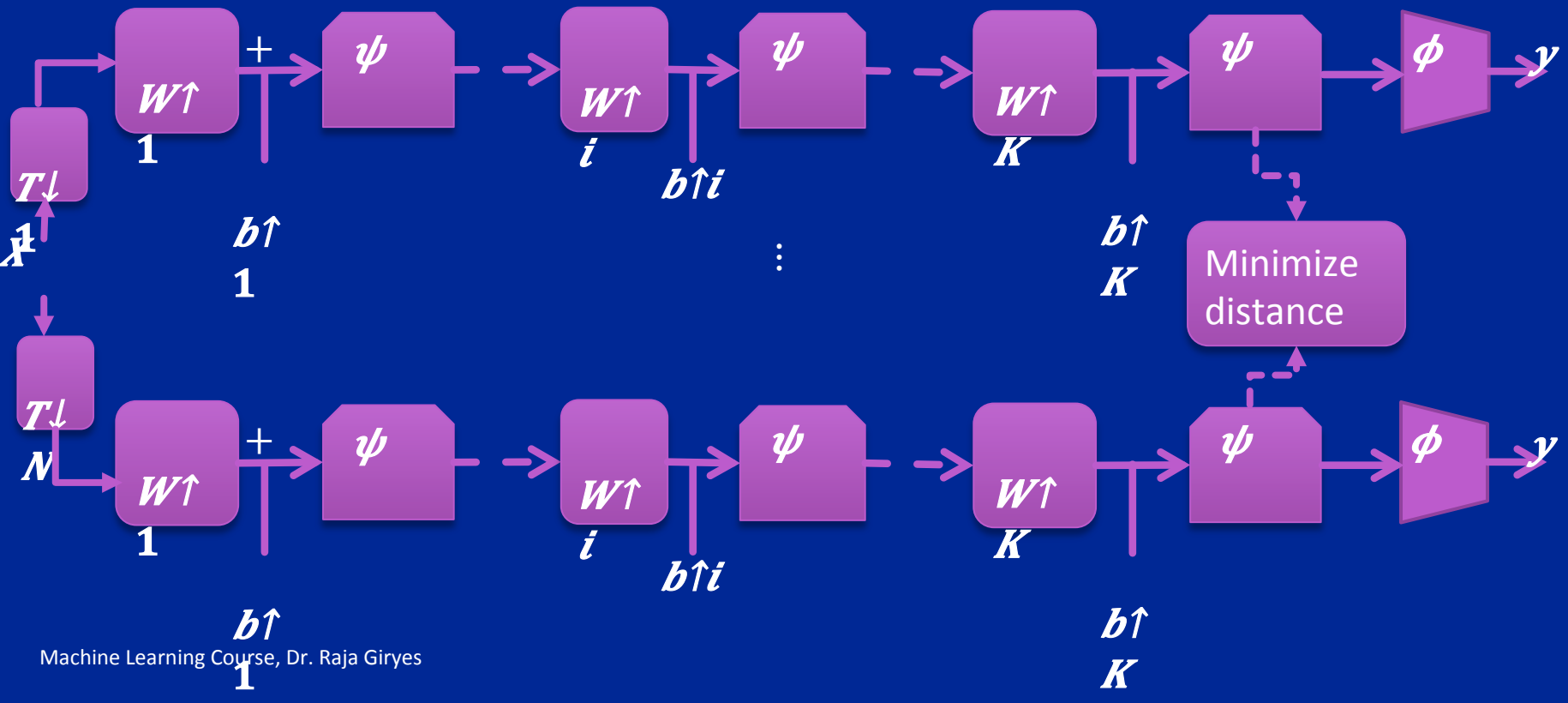
INVARIANCE SLICE

- Use transformations $T\downarrow 1$,…,$T\downarrow N$ to transform the input [Dieleman et al., 2016]
- Average the features before the soft-max layer



INVARIANCE BY REGULARIZATION

- Use transformations $T\downarrow 1$,…,$T\downarrow N$ to transform the input [Sokolic et al., 2017]
- Force features to be similar



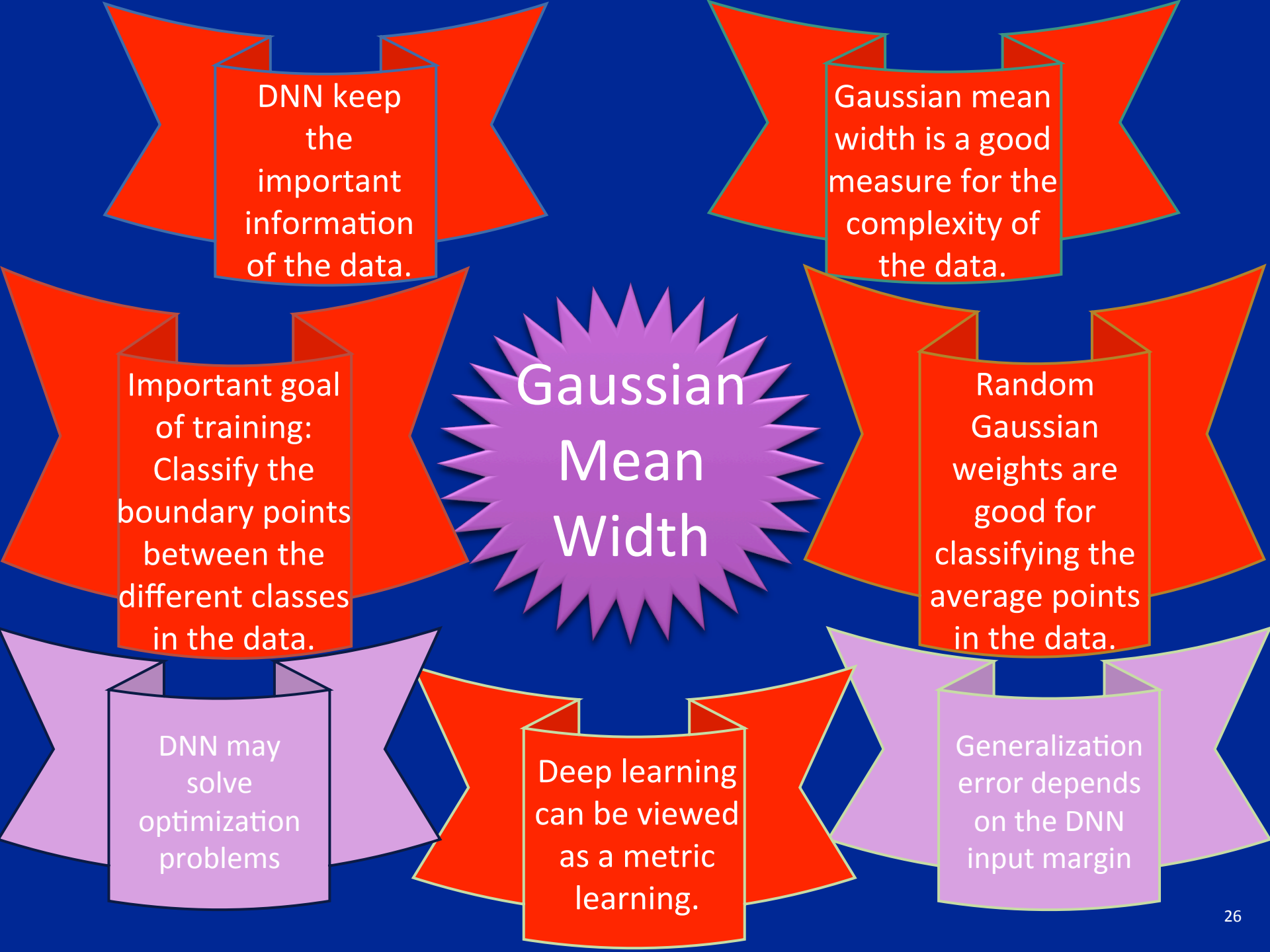
INVARIANCE

- Designing invariant DNN reduce the GE

Table 1: Classification accuracy [%] on CIFAR-10.

	number of training samples				
	2500	5000	10000	20000	50000
No reg.	68.71	76.74	85.17	87.15	93.65
Inv. Reg.	69.32	79.08	86.69	88.14	94.50
No reg. + avg.	70.59	78.40	86.05	88.13	94.26
Inv. Reg. + avg.	70.71	79.65	86.96	88.98	94.78

[Sokolić, Giryes, Sapiro & Rodrigues, 2017]


DNN keep the important information of the data.
Gaussian mean width is a good measure for the complexity of the data.
Important goal of training: Classify the boundary points between the different classes in the data.
Gaussian Mean Width
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

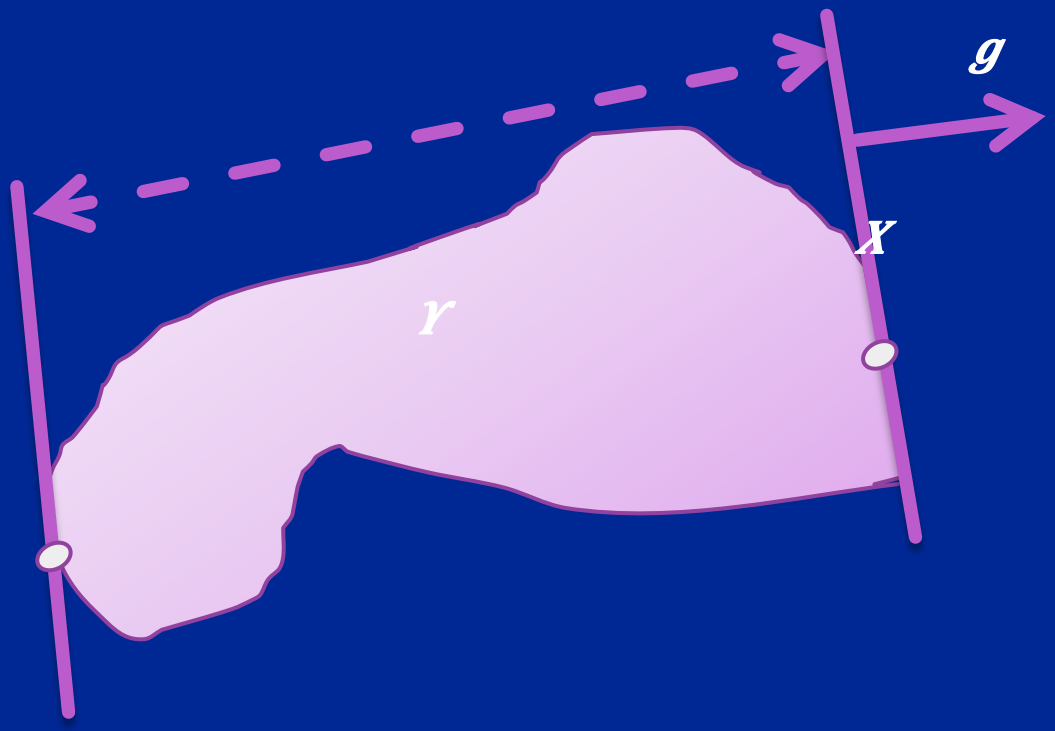
GAUSSIAN MEAN WIDTH

- Gaussian mean width:

$$\omega(Y) = E \sup_{X,Z \in Y} \langle X - Z, g \rangle,$$

$$g \sim N(0, I).$$

The width of the set r in the direction of g:



MEASURE FOR LOW DIMENSIONALITY

- Gaussian mean width:

$$\omega(Y)=E\sup_{X,Z\in Y}\langle X-Z,g\rangle,$$

$g\sim N(0,I)$.

- $\omega^2(Y)$ is a measure for the dimensionality of the data.
- Examples:

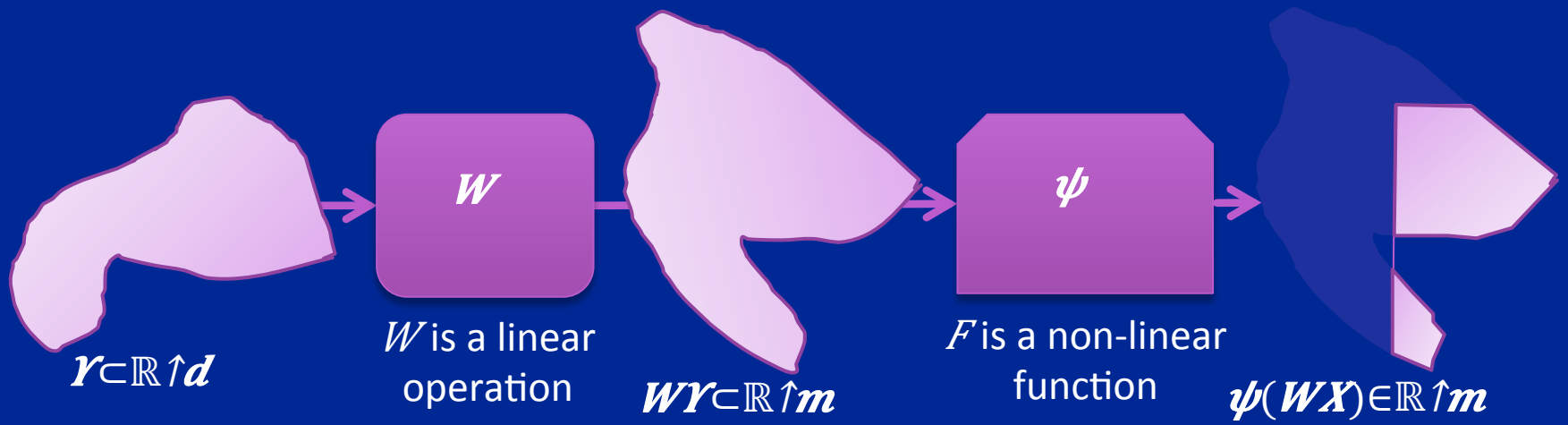
If $Y\subset\mathbb{B}^d$ is a Gaussian Mixture Model with k Gaussians then

$\omega^2(Y)=O(k)$

If $Y\subset\mathbb{B}^d$ is a data with k-sparse representations then

$\omega^2(Y)=O(k\log d)$

GAUSSIAN MEAN WIDTH IN DNN

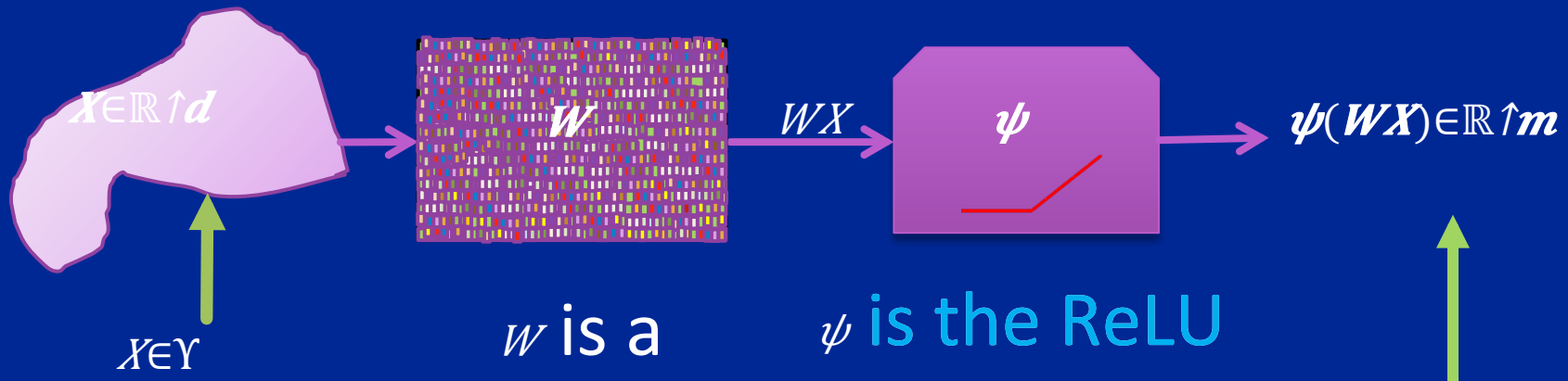


Theorem 1: small $\omega^2(X)/m$ imply $\omega^2(X) \approx \omega^2(\psi(WX))$

Small $\omega^2(X)$ → Small $\omega^2(\psi(WX))$

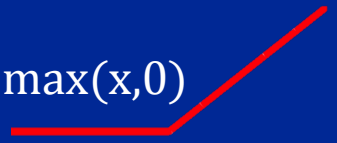
→ It is sufficient to provide proofs only for a single layer

ASSUMPTIONS



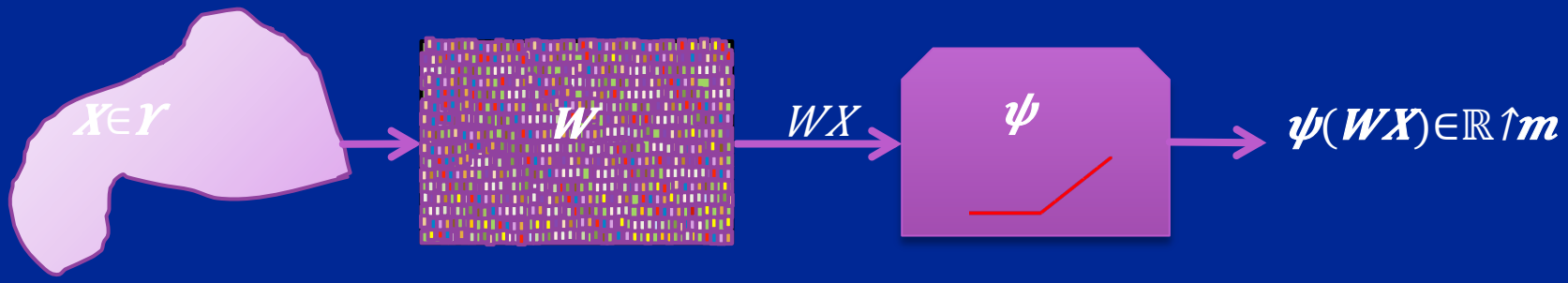
W is a random Gaussian matrix

ψ is the ReLU



$m = O(\delta^{-4} \omega^4 (Y))$

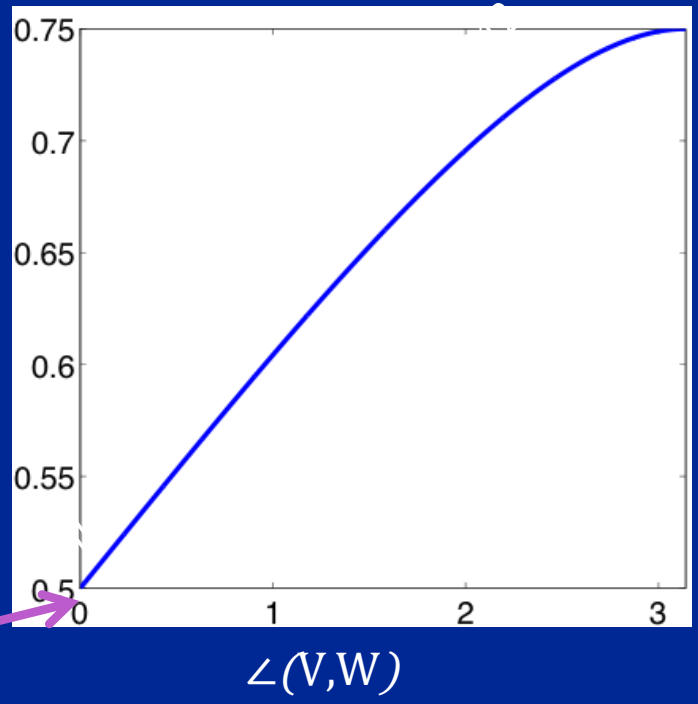
DISTANCE DISTORTION



Theorem 4: for $x, z \in \mathcal{Y}$

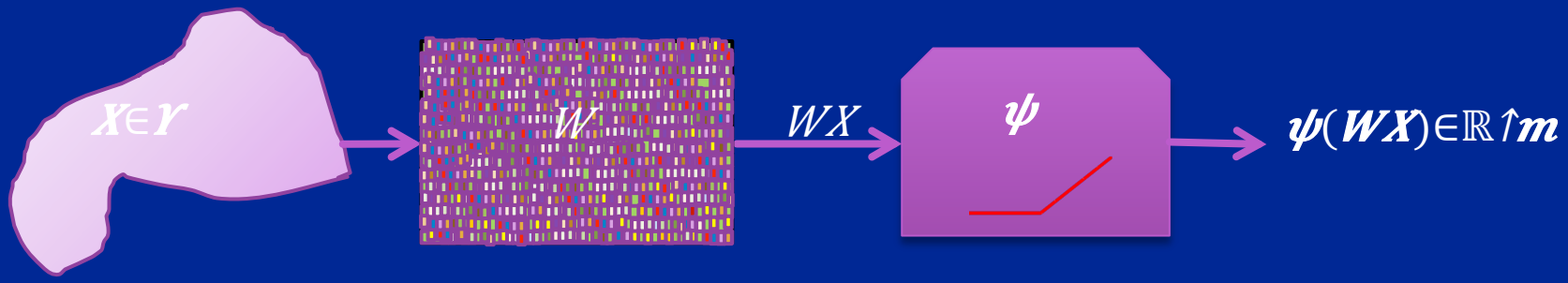
$$\|\psi(Wx) - \psi(Wz)\|_2^2 - 1/2 \|x - z\|_2^2 - \|x\|\|z\|/\pi \left(\sin\angle(x,z) - \angle(x,z)\cos\angle(x,z)\right) < \delta$$

[Giryes, Sapiro & Bronstein 2015].



The smaller $\angle(x,z)$ the smaller the distance we get between the points

ANGLE DISTORTION

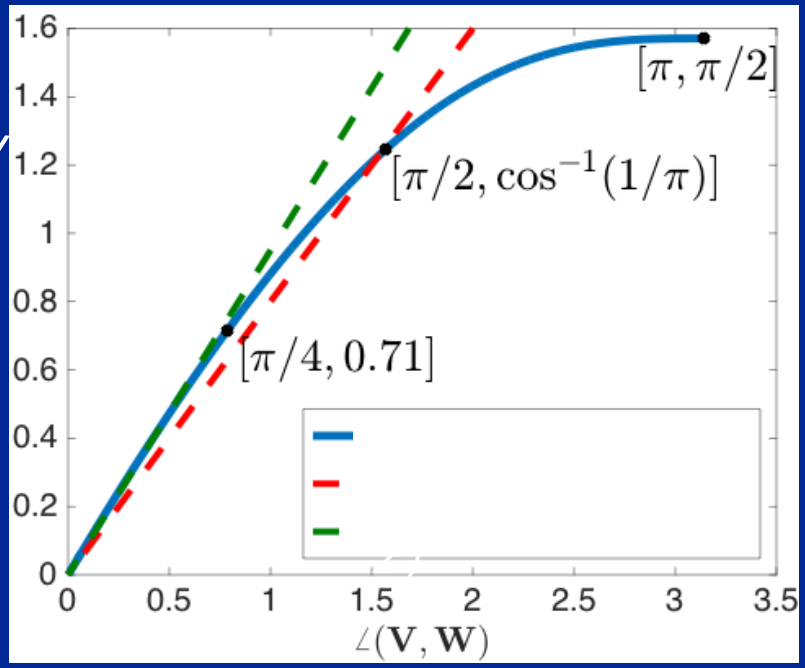


Theorem 5: for $X,Z \in \mathcal{Y}$

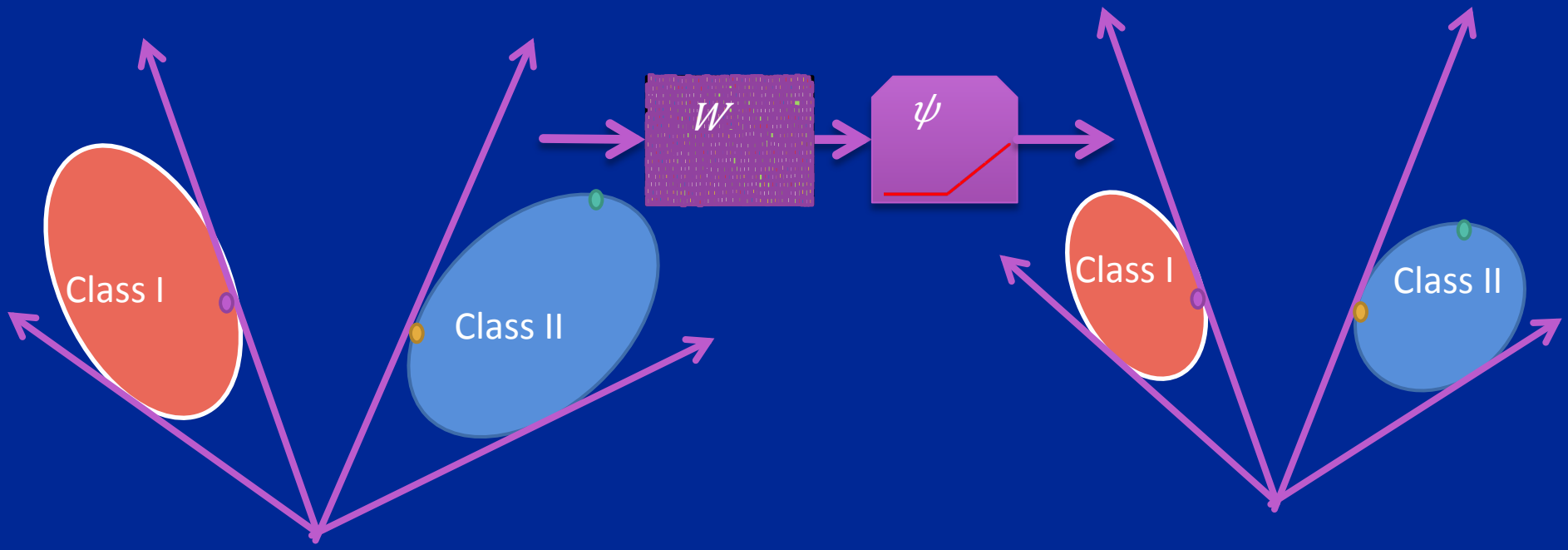
$$\cos\angle(\psi(WX),\psi(WZ)) - \cos\angle(X,Z) - 1/\pi\,(\sin\angle(X,Z) - \angle(X,Z)\cos\angle(X,Z)) < \delta$$

[Giryes, Sapiro & Bronstein 2015].

Behavior of $\angle(\psi(WX),\psi(WZ))$



DISTANCE AND ANGLES DISTORTION

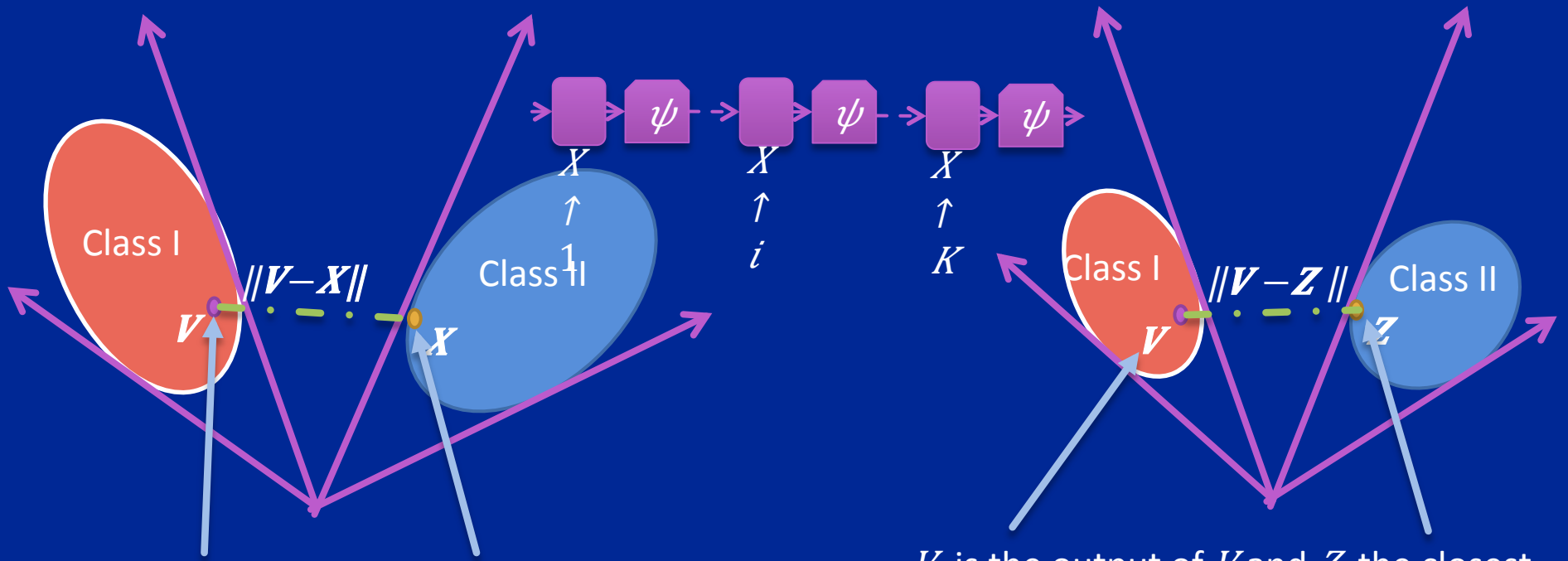


Points with small angles between them become closer than points with larger angles between them

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

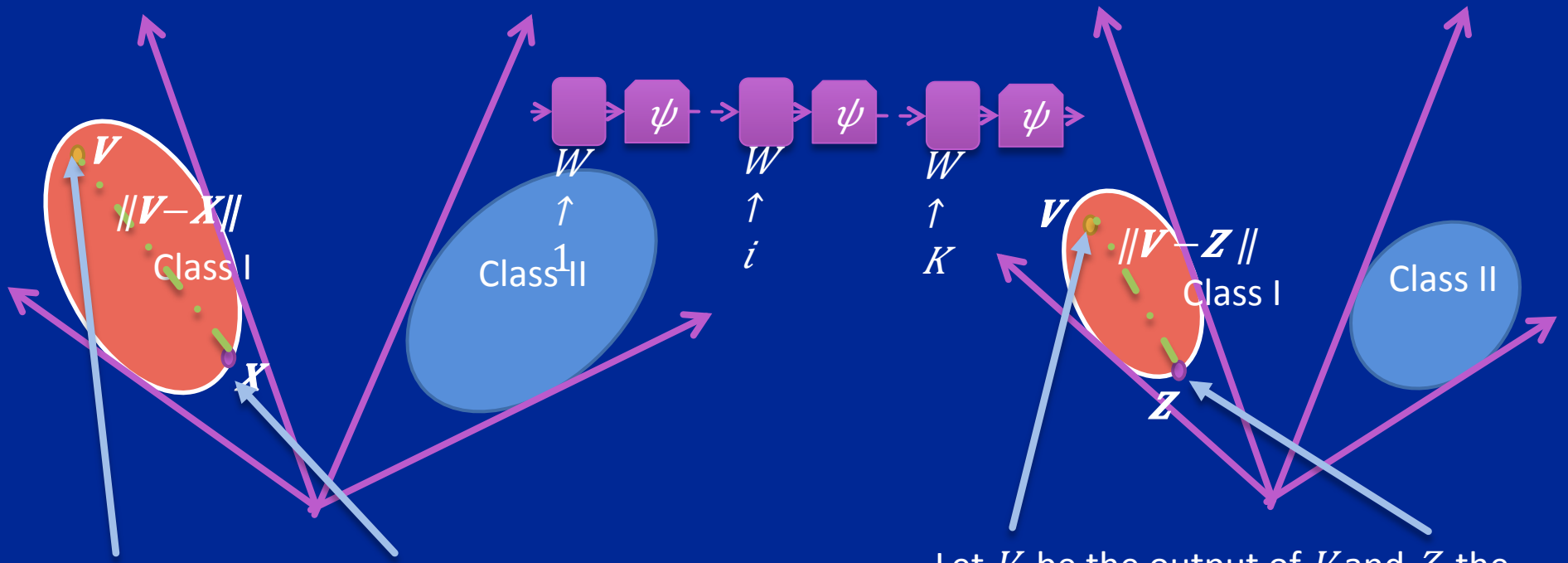


V is a random point and X its closest point from a different class.

V is the output of V and Z the closest point to V at the output from a different class.

Compute the distance ratio: $\|V-Z\|/\|V-X\|$

INTRA BOUNDARY POINTS DISTANCE RATIO



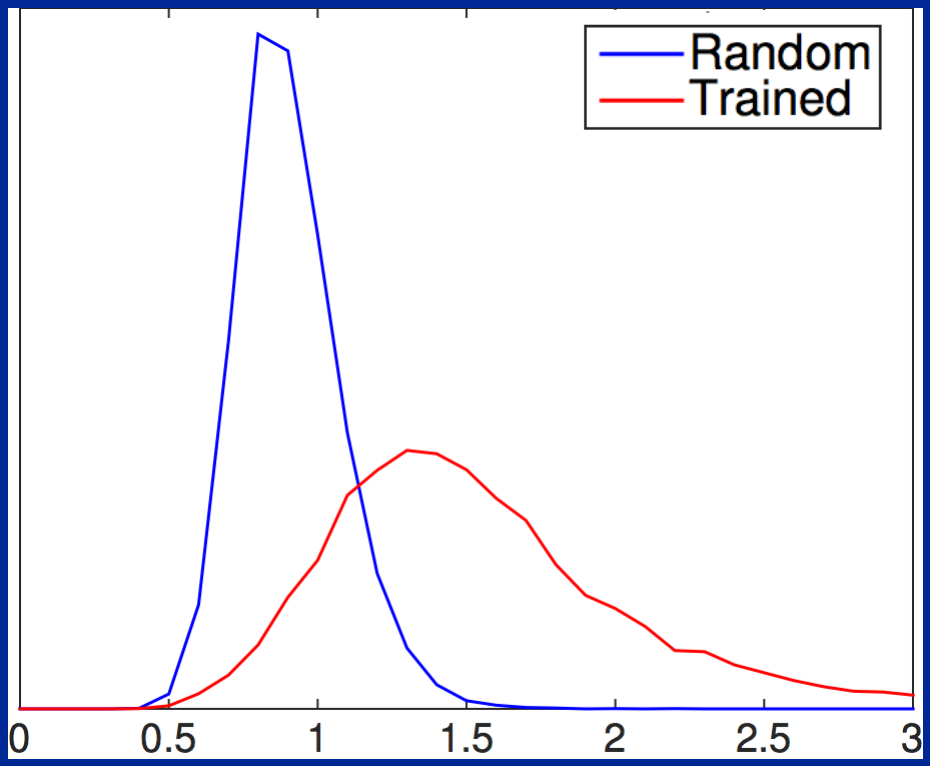
Let V be a point and X its farthest point from the same class.

Let V be the output of V and Z the farthest point from V at the output from the same class

Compute the distance ratio: $\|V-Z\|/\|V-X\|$

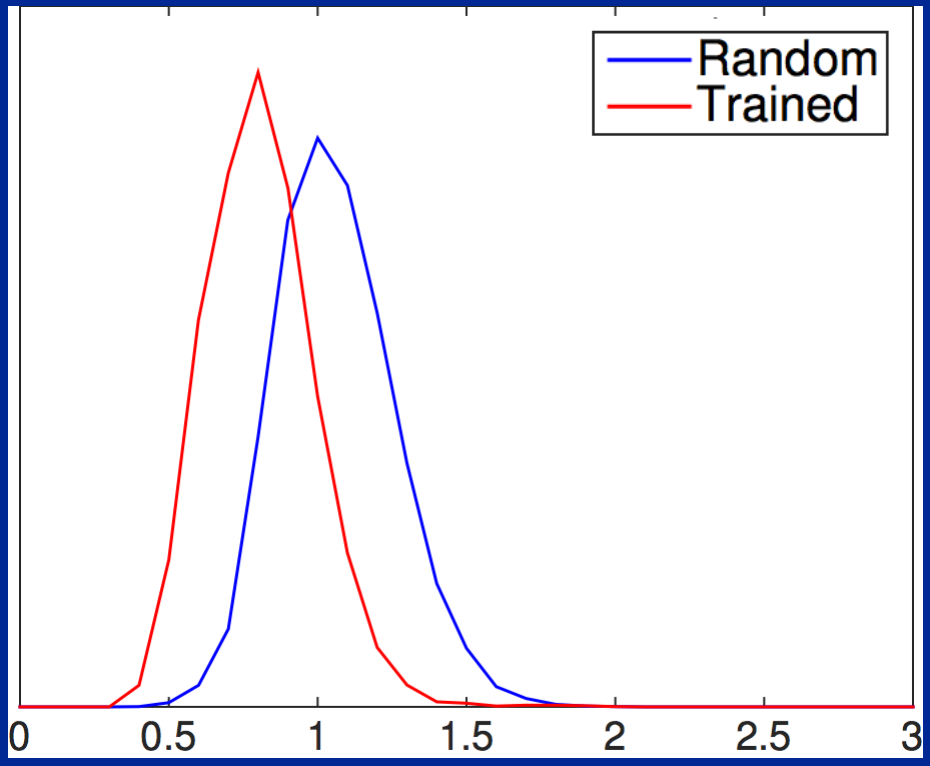
BOUNDARY DISTANCE RATIO

Inter-class



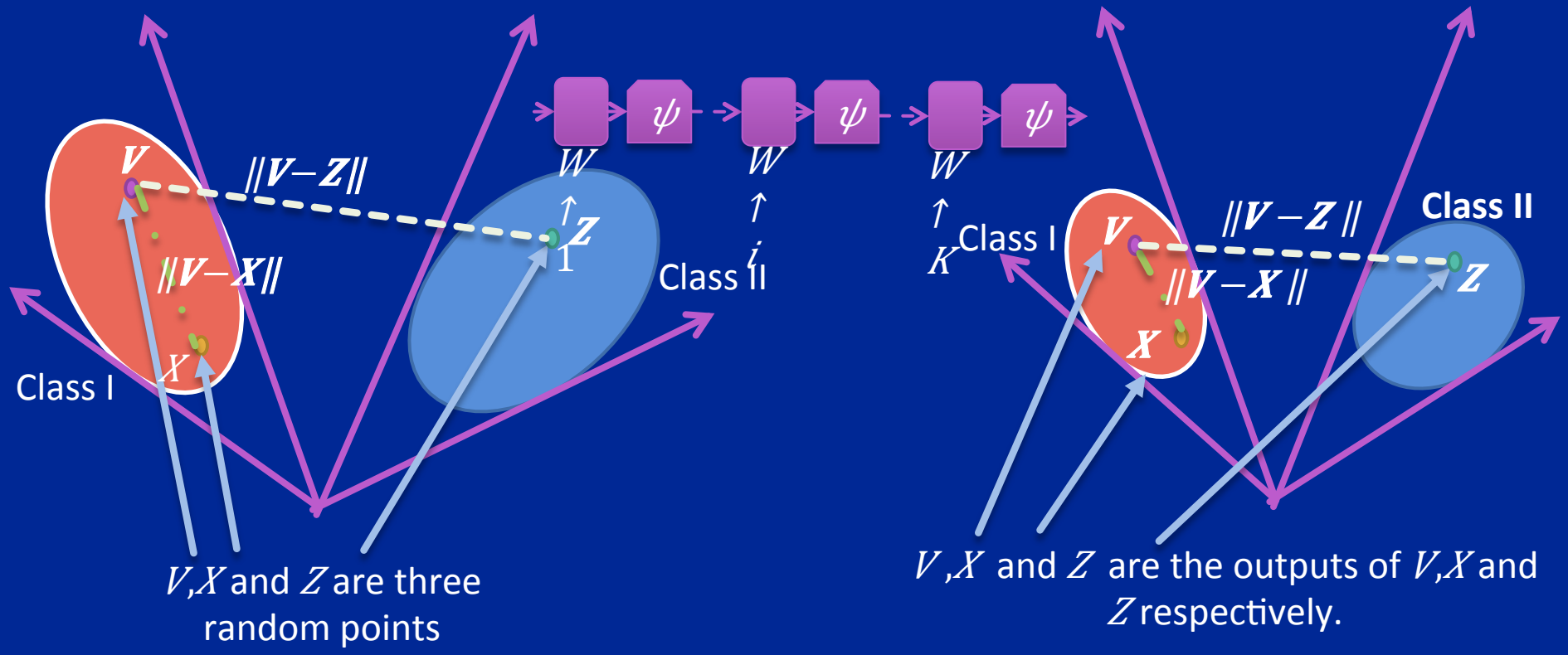
$\|V-Z\|/\|V-X\|$

Intra-class



$\|V-Z\|/\|V-X\|$

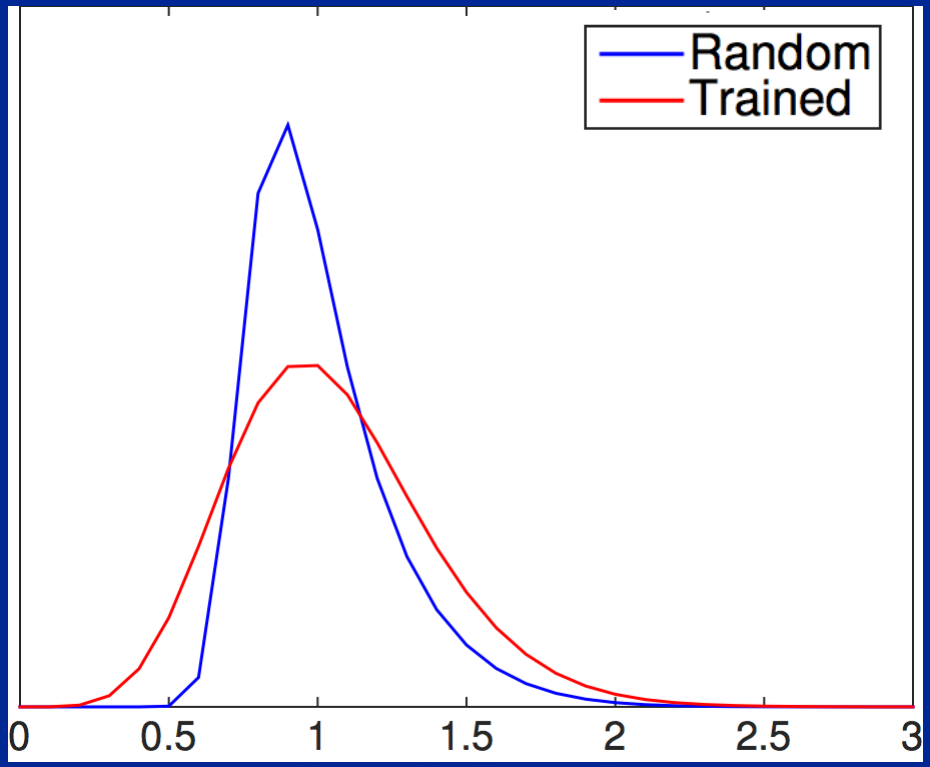
AVERAGE POINTS DISTANCE RATIO



Compute the distance ratios: $\|V-X\|/\|V-X\|$, $\|V-Z\|/\|V-Z\|$

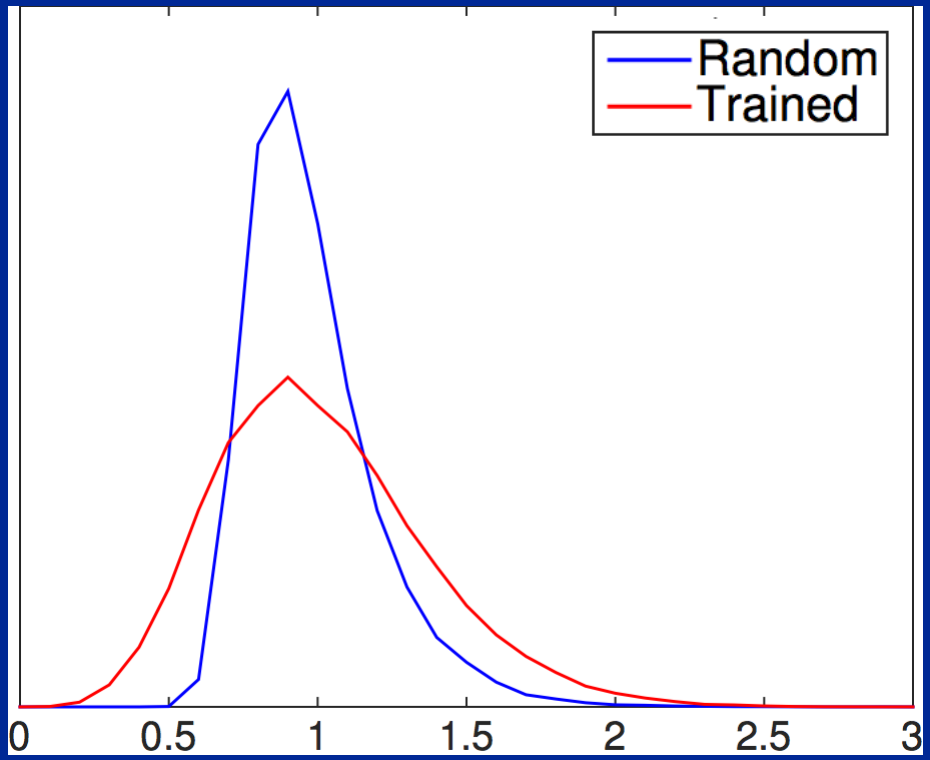
AVERAGE DISTANCE RATIO

Inter-class



$\|V-Z\|/\|V-Z\|$

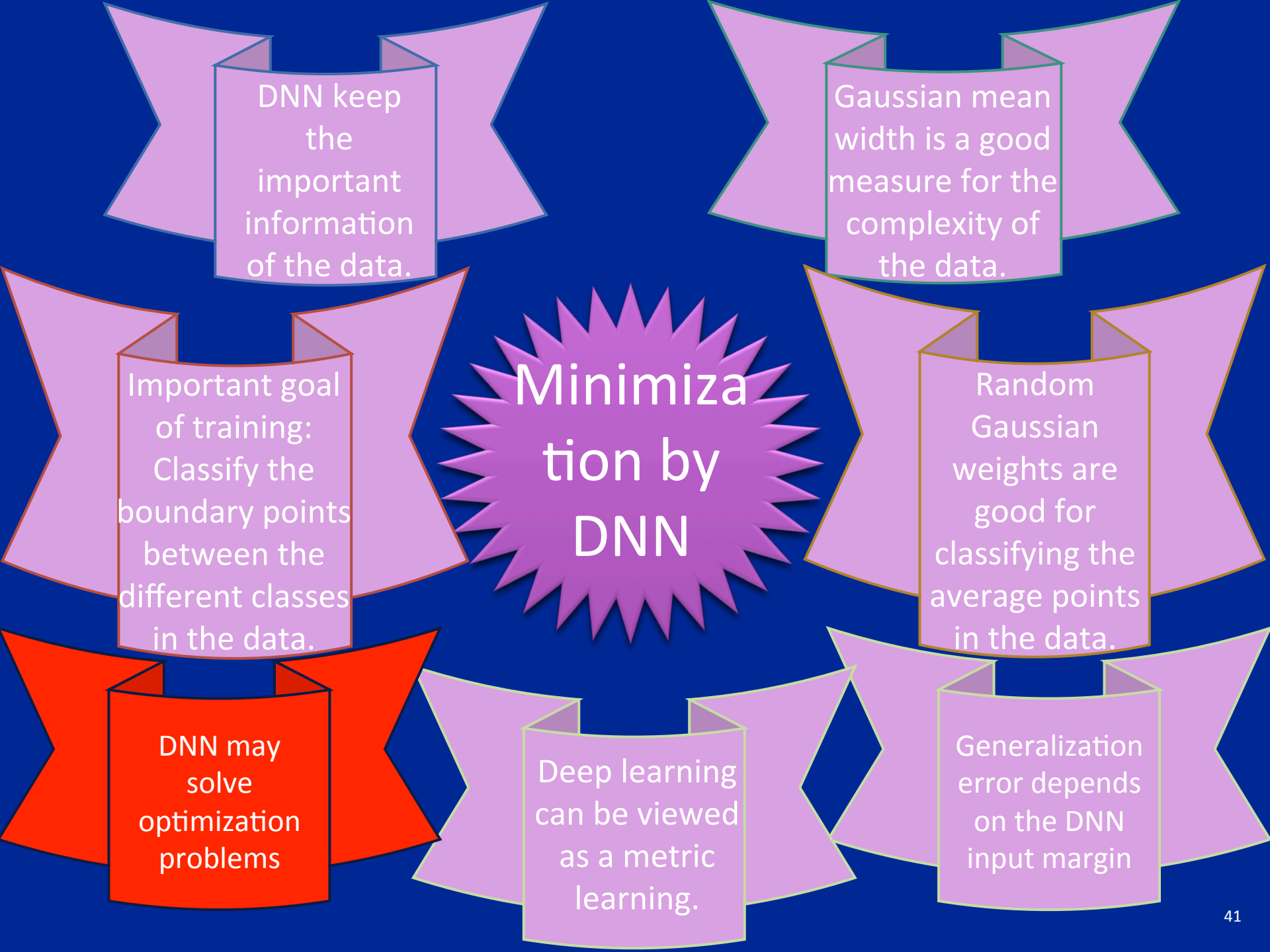
Intra-class



$\|V-X\|/\|V-X\|$

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.
Gaussian mean width is a good measure for the complexity of the data.
Important goal of training: Classify the boundary points between the different classes in the data.
Minimiza tion by DNN
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

INVERSE PROBLEMS

- We are given $X = AZ + E$

Given set of measurements

linear operator

Unknown signal

noise

- Standard technique for recovery

$$\min_Z \|X - AZ\|_2 \quad \text{s.t.} \quad Z \in Y$$

- Unconstrained form

$$\min_Z \|X - AZ\|_2^2 + \lambda f(Z)$$

Z resides in a low dimensional set Y

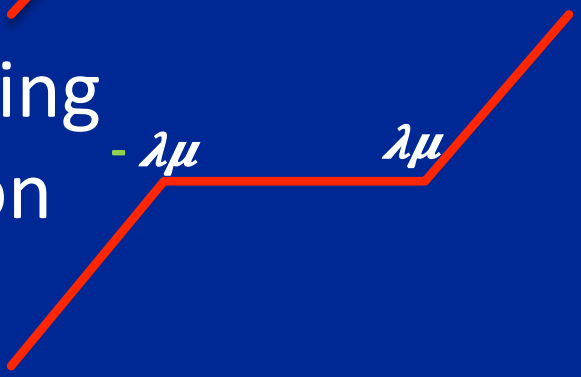
Regularization parameter

f is a penalty function

ℓ_1 MINIMIZATION CASE

- Unconstrained form
  $\min_Z \|X-AZ\|_2^2 + \lambda\|Z\|_1$
- Can be solved by proximal gradient, e.g., iterative shrinkage and thresholding technique (ISTA)
  $Z^{t+1} = \psi_{\lambda\mu}(Z^t + \mu A^T(X-AZ^t))$

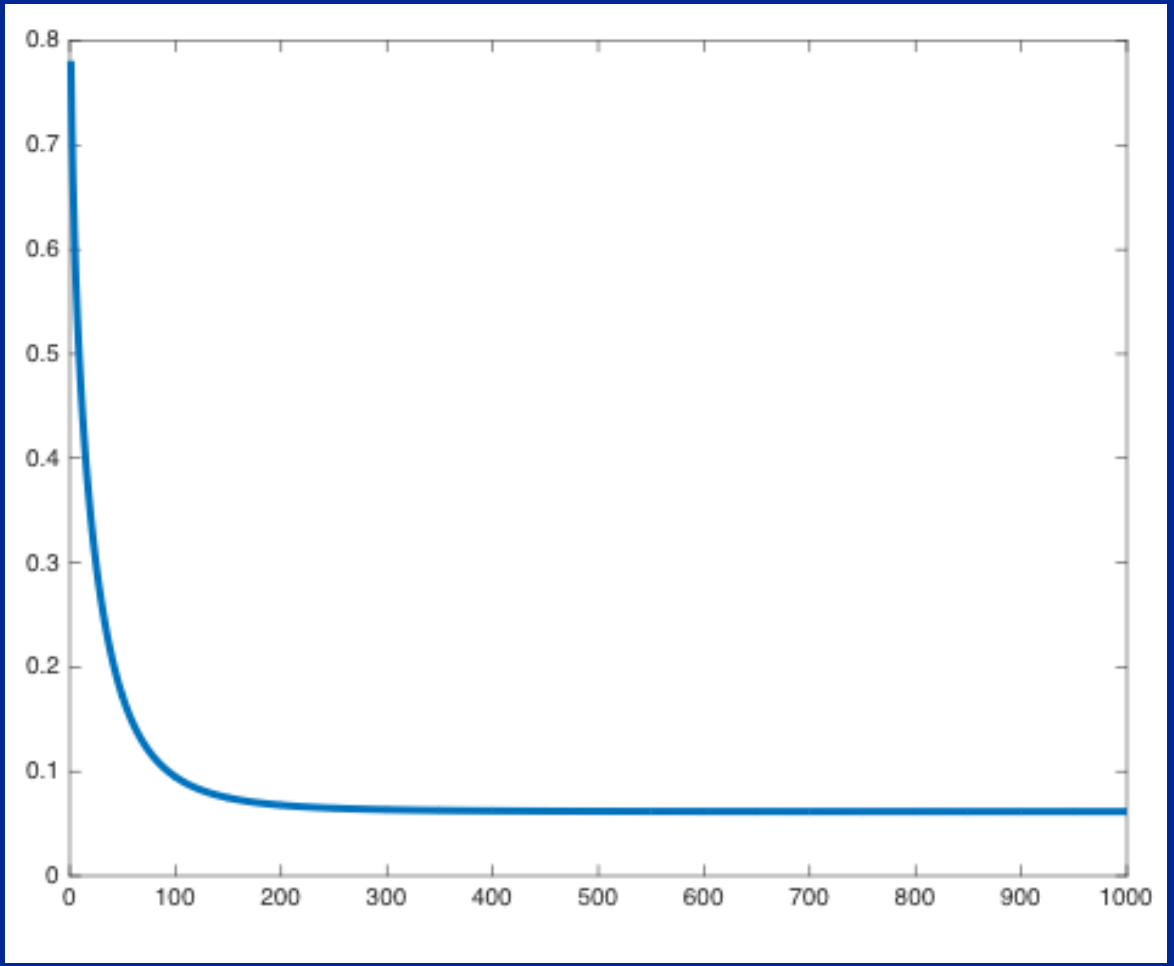
Soft thresholding operation



μ is the step size

ISTA CONVERGENCE

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



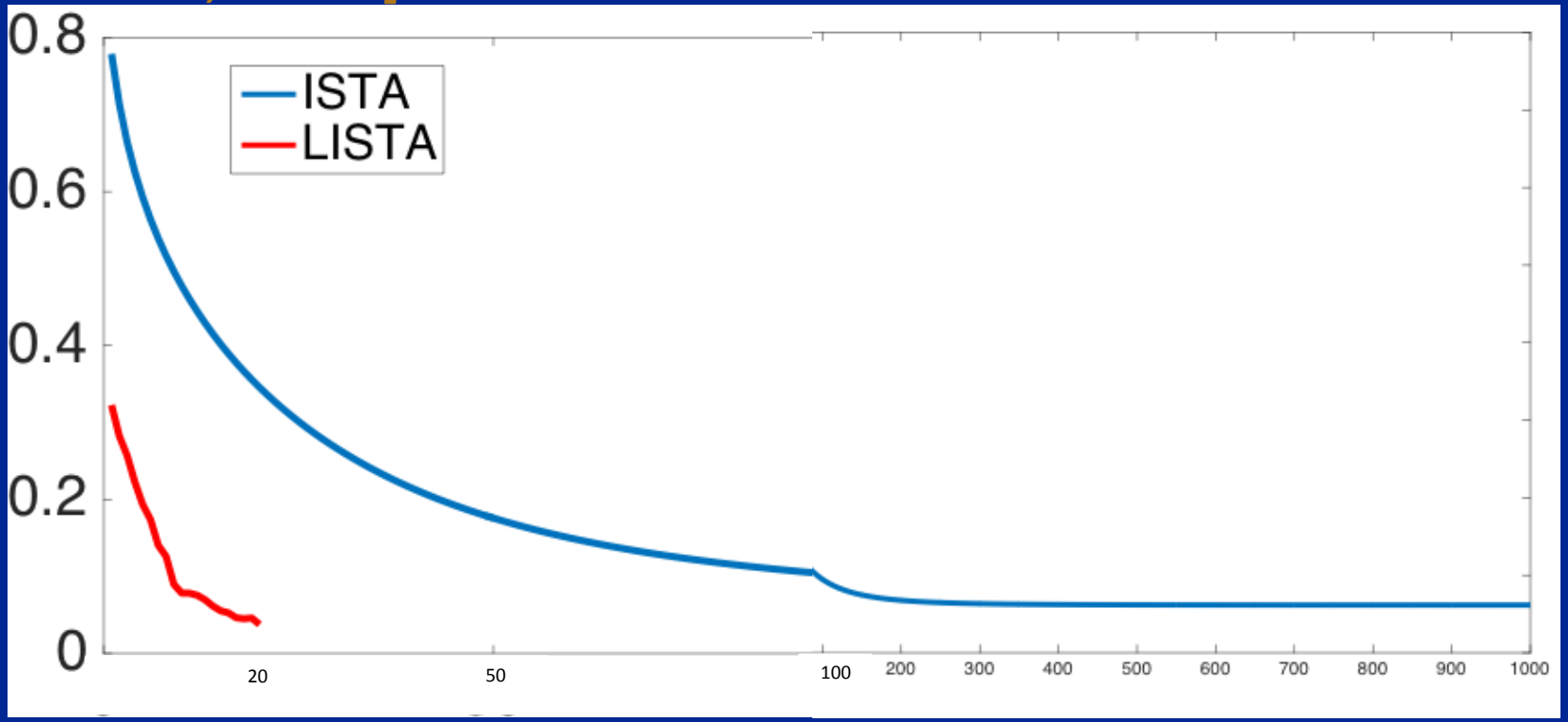
LISTA

- ISTA
  $$\boldsymbol{z^{t+1} = \psi_{\lambda\mu}\left(z^{t} + \mu A^{T}\left(x - A z^{t}\right)\right)}$$
- Rewriting ISTA:
  $$\boldsymbol{z^{t+1} = \psi_{\lambda\mu}\left(\left(I - \mu A^{T} A\right) z^{t} + \mu A^{T} x\right)}$$
- Learned ISTA (LISTA):
  $$\boldsymbol{z^{t+1} = \psi_{\lambda}\left(W z^{t} + S x\right)}$$

Learned operators

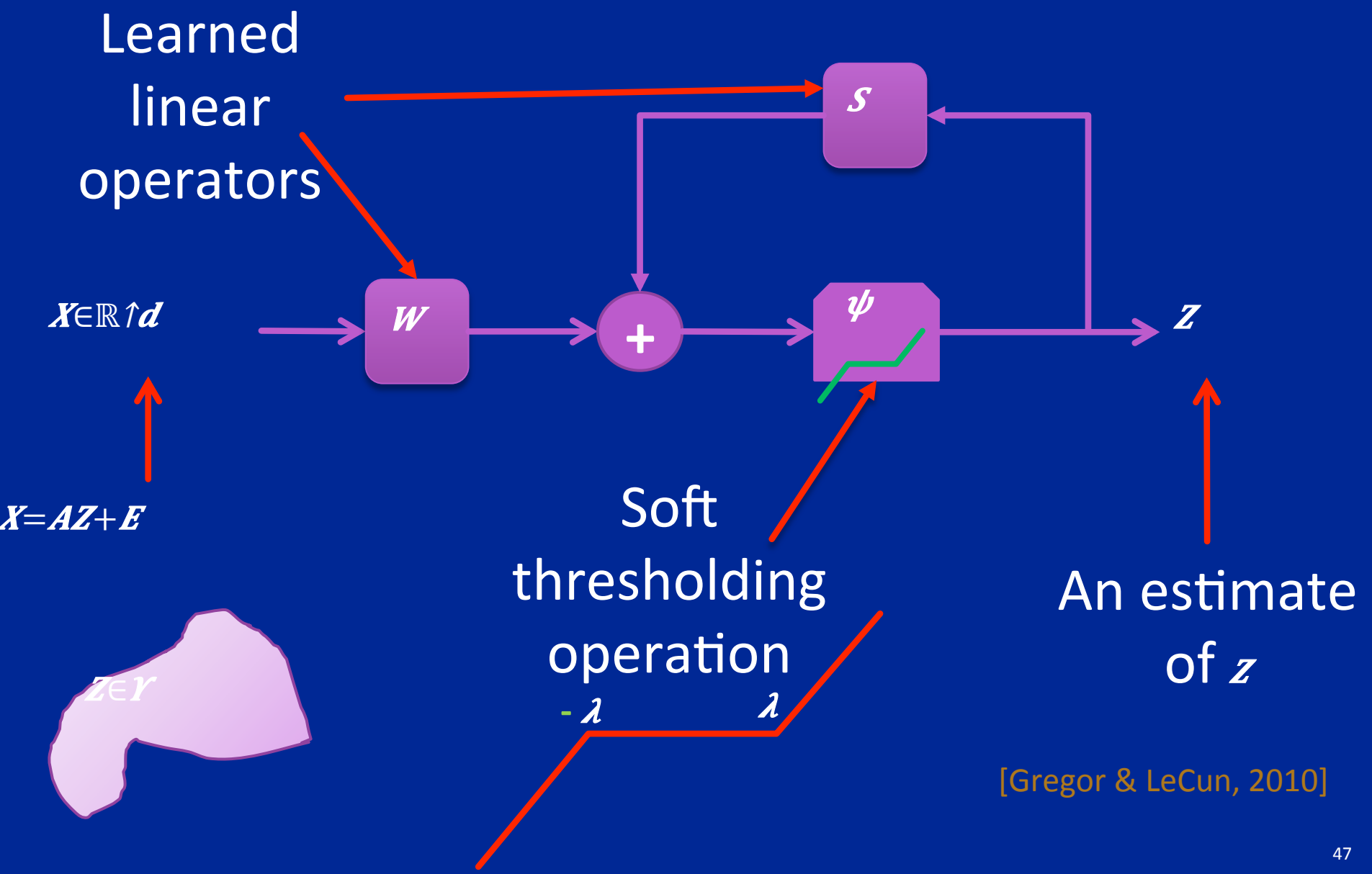
LISTA CONVERGENCE

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



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

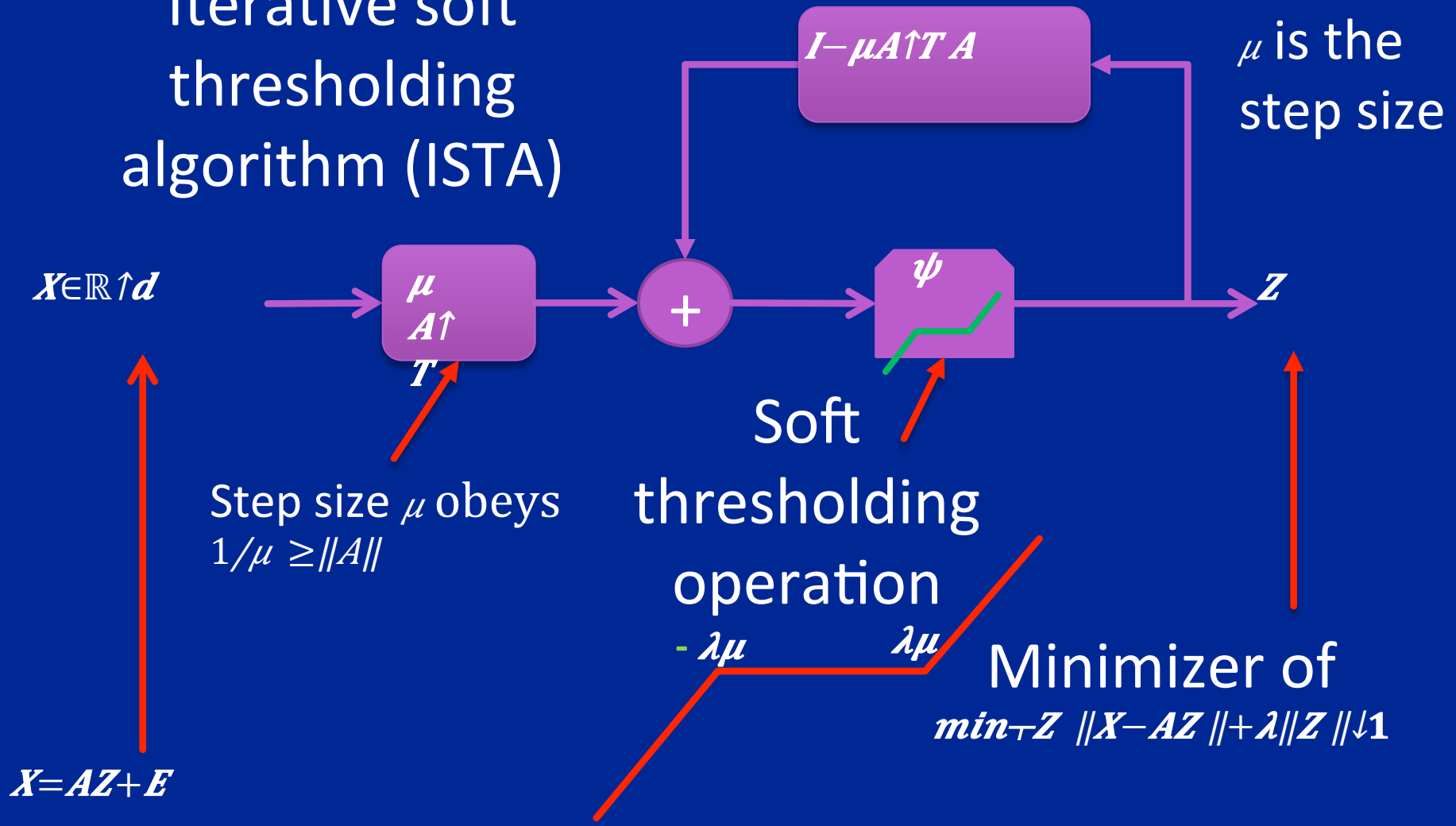
LISTA AS A NEURAL NETWORK



[Gregor & LeCun, 2010]

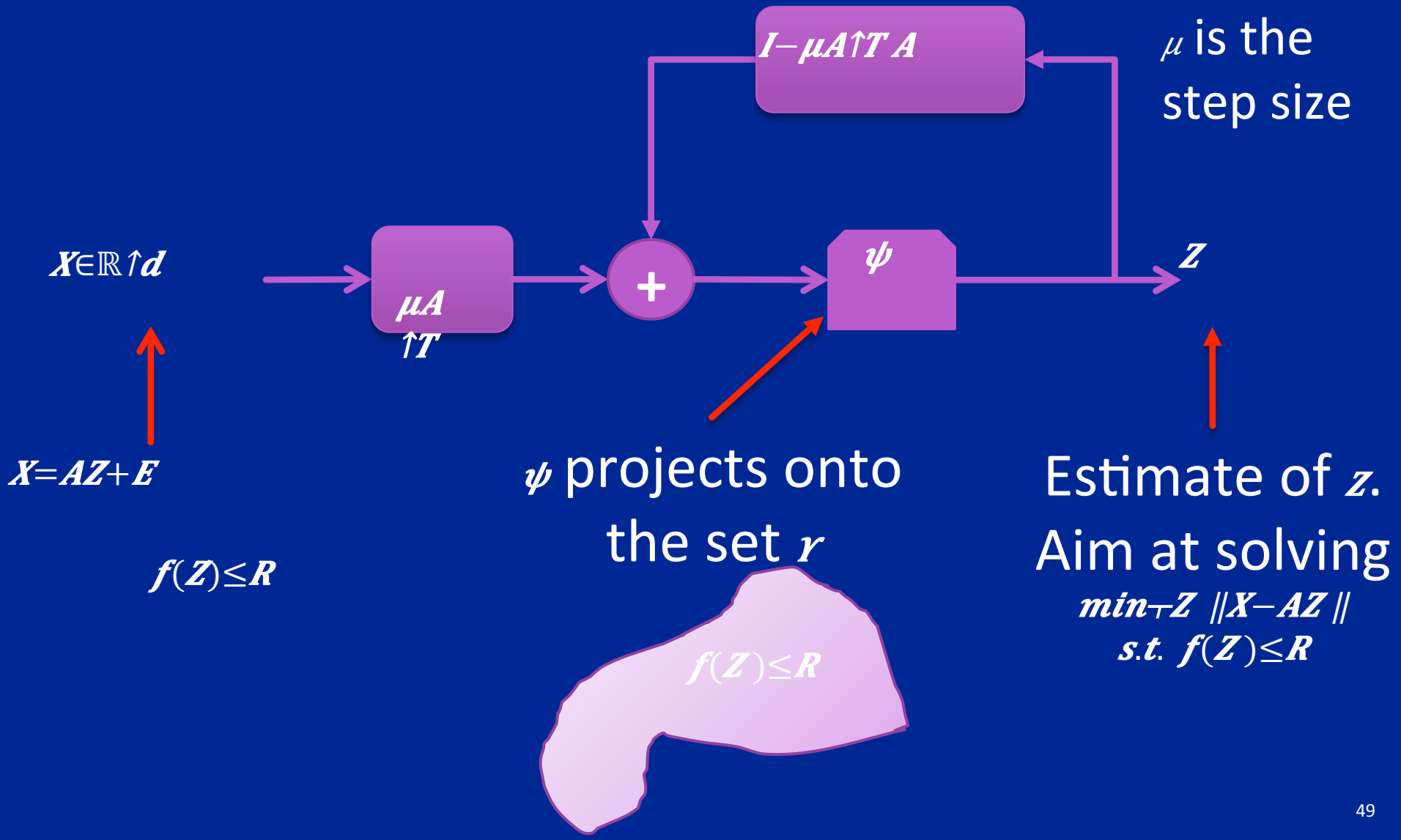
ISTA

Iterative soft thresholding algorithm (ISTA)



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

PROJECTED GRADIENT DESCENT (PGD)



THEORY FOR PGD

- Theorem 8: Let $Z \in \mathbb{R}^d$, $f: \mathbb{R}^d \to \mathbb{R}$ a proper function, $f(Z) \leq R$, $C_f(Z)$ the tangent cone of f at point Z, $A \in \mathbb{R}^{m \times d}$ a random Gaussian matrix and $X = AZ + E$. Then the estimate of PGD at iteration t, Z^t, obeys
$\|Z^t - Z\| \leq (\kappa_f \rho)^t \|Z\|$,

  where $\rho = \sup_{U,V \in C_f(Z) \cap \mathcal{B}^d} U^T (I - \mu A^T A) V$
  and $\kappa_f = 1$ if f is convex and $\kappa_f = 2$ otherwise.
  [Oymak, Recht & Soltanolkotabi, 2016].

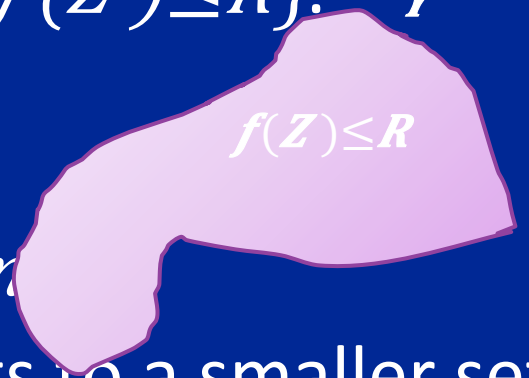
PGD CONVERGENCE RATE

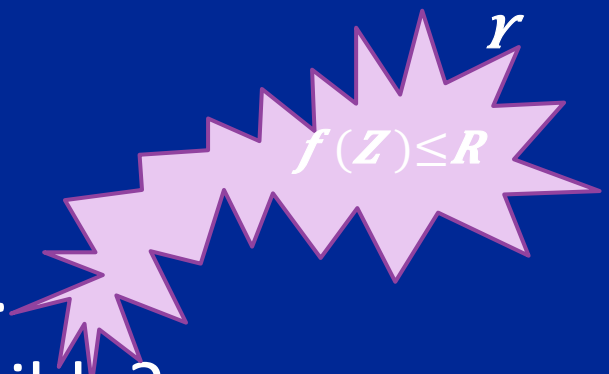
- $\rho=\sup_{U,V\in C\downarrow f(Z)\cap B\uparrow d} U\uparrow T(I-\mu A\uparrow T A)V$ is the convergence rate of PGD.
- Let ω be the Gaussian mean width of $C\downarrow f(Z)\cap B\uparrow d$.
- If $\mu=1/(\sqrt{m}+\sqrt{d})\uparrow 2 \simeq 1/d$ then $\rho=1-O(\sqrt{m}-\omega/\sqrt{m+d})$.
- If $\mu=1/m$ then $\rho=O(\omega/\sqrt{m})$.
- For the k-sparse model $\omega\uparrow 2 =O(k\log(\mathrm{d}))$
- For GMM with k Gaussians $\omega\uparrow 2 =O(k)$.
- How may we cause ω to become smaller for having a better convergence rate?

INACCURATE PROJECTION

- PGD iterations projects onto $\Upsilon = \{Z : f(Z) \leq R\}$.
- Smaller $\Upsilon \Rightarrow$ Smaller ω.

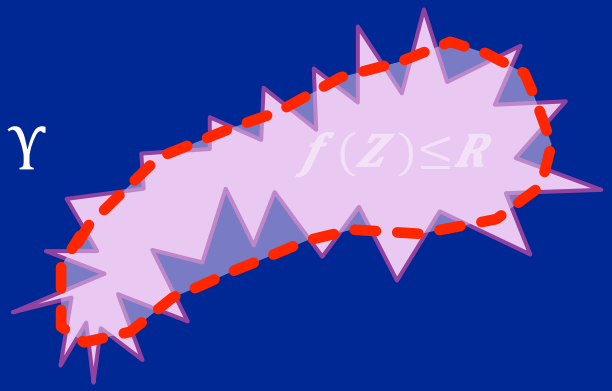
  Faster convergence as $\rho = 1 - O(\sqrt{m - \omega/m + d})$ or $O(\omega/\sqrt{m}$
- Let us assume that our signal belongs to a smaller set $\Upsilon = \{Z : f(Z) \leq R\}$ with $\omega \ll \omega$.
- Ideally, we would like to project onto Υ instead of Υ.
- This will lead to faster convergence.
- What if such a projection is not feasible?





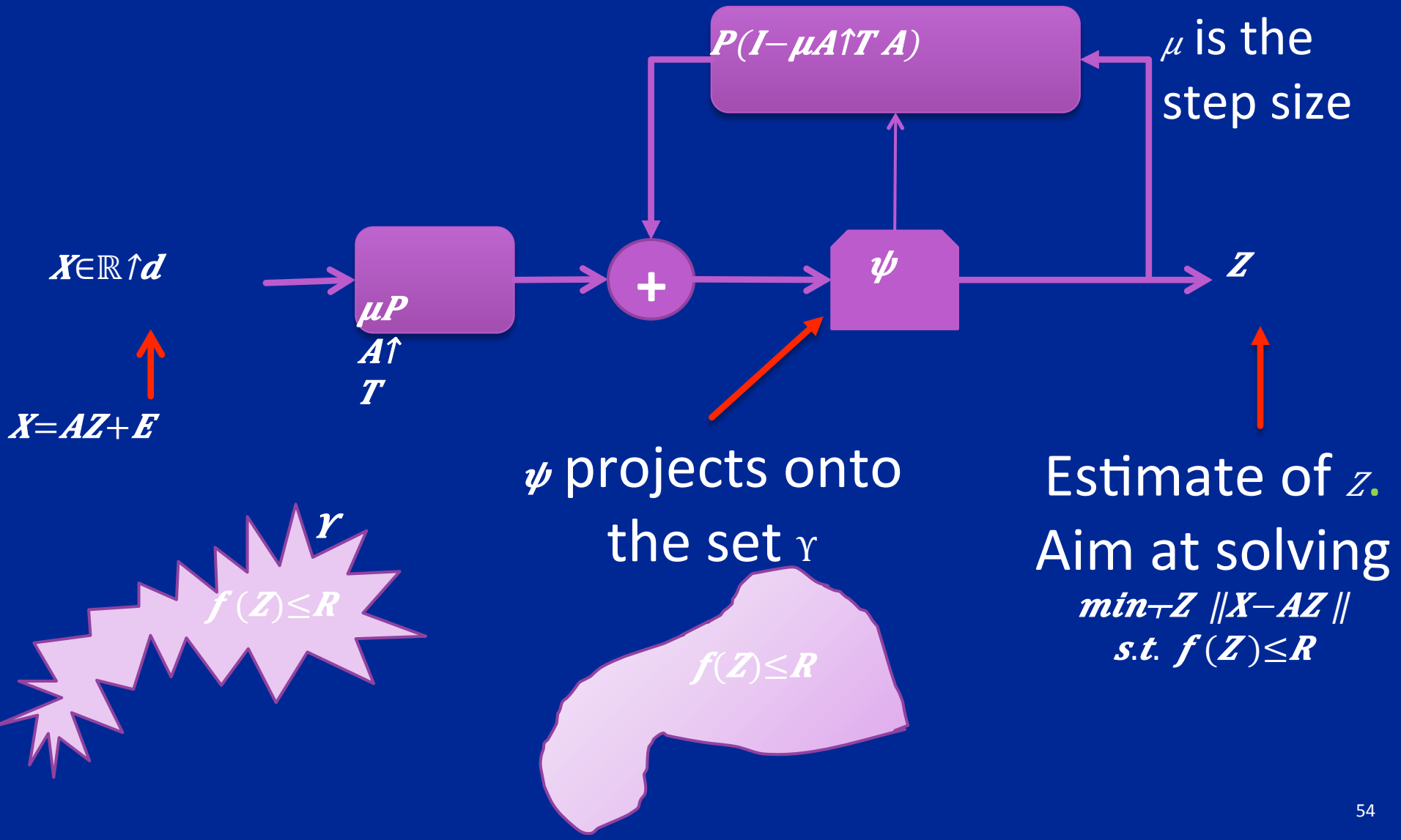
INACCURATE PROJECTION

- We will estimate the projection onto Υ by
  - A linear projection P
  - Followed by a projection onto Υ
- Assumptions:
  - $\|\wp\downarrow\Upsilon\ (PZ)-Z\|\leq\epsilon$



Projection of the target vector Z
onto P and then onto Υ

INACCURATE PGD (IPGD)



THEORY FOR IPGD

- Theorem 9: Let $Z \in \mathbb{R}^d$, $f: \mathbb{R}^d \to \mathbb{R}$ a proper convex* function, $f(Z) \leq R$, $C_f(Z)$ the tangent cone of f at point Z, $A \in \mathbb{R}^{d \times m}$ a random Gaussian matrix and $X = AZ + E$. Then the estimate of IPGD at iteration t, Z^t, obeys

$$\|Z^t - Z\| \leq ((\rho_P)^t + 1 - (\rho_P)^t / 1 - \rho_P \epsilon) \|Z\|,$$

where $\rho_p = \sup_{U,V \in C_f(Z) \cap \mathcal{B}^d} U^T P(I - \mu A^T A) PV$

and $\epsilon = (2 + \rho_p)\epsilon$.

[Giryes, Eldar, Bronstein & Sapiro, 2016]

*We have a version of this theorem also when f is non-proper or non-convex function

CONVERGENCE RATE COMPARISON

- PGD convergence:

$(\rho)\uparrow t$

- IPGD convergence:

$(\rho\downarrow P)\uparrow t+1-(\rho\downarrow P)\uparrow t/1-\rho\downarrow P\ (2+\rho\downarrow p)\epsilon$

$\tilde{}\simeq \perp(a)\ (\rho\downarrow P)\uparrow t+\epsilon\tilde{}\simeq \perp(b)\ (\rho\downarrow P)\uparrow t\ \tilde{}\ll \perp(c)\ (\rho)\uparrow t$

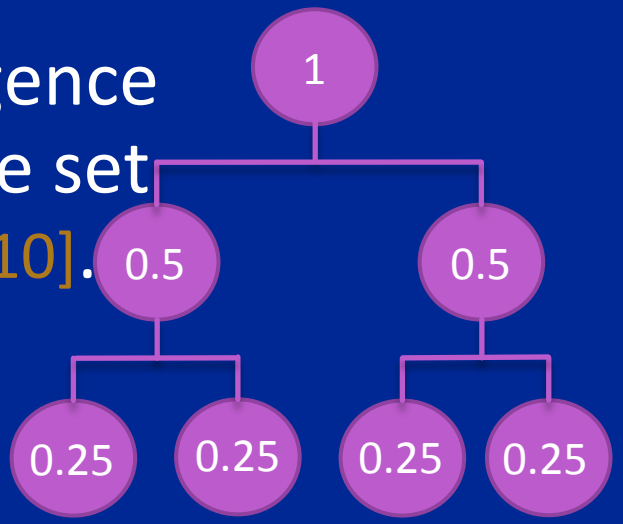
$(a)\epsilon$ is negligible compared to $\rho\downarrow P$

(b) For small values of t (early iterations).

(c) Faster convergence as $\rho\downarrow P\ll\rho$ (because $\omega\downarrow p\ll\omega$).

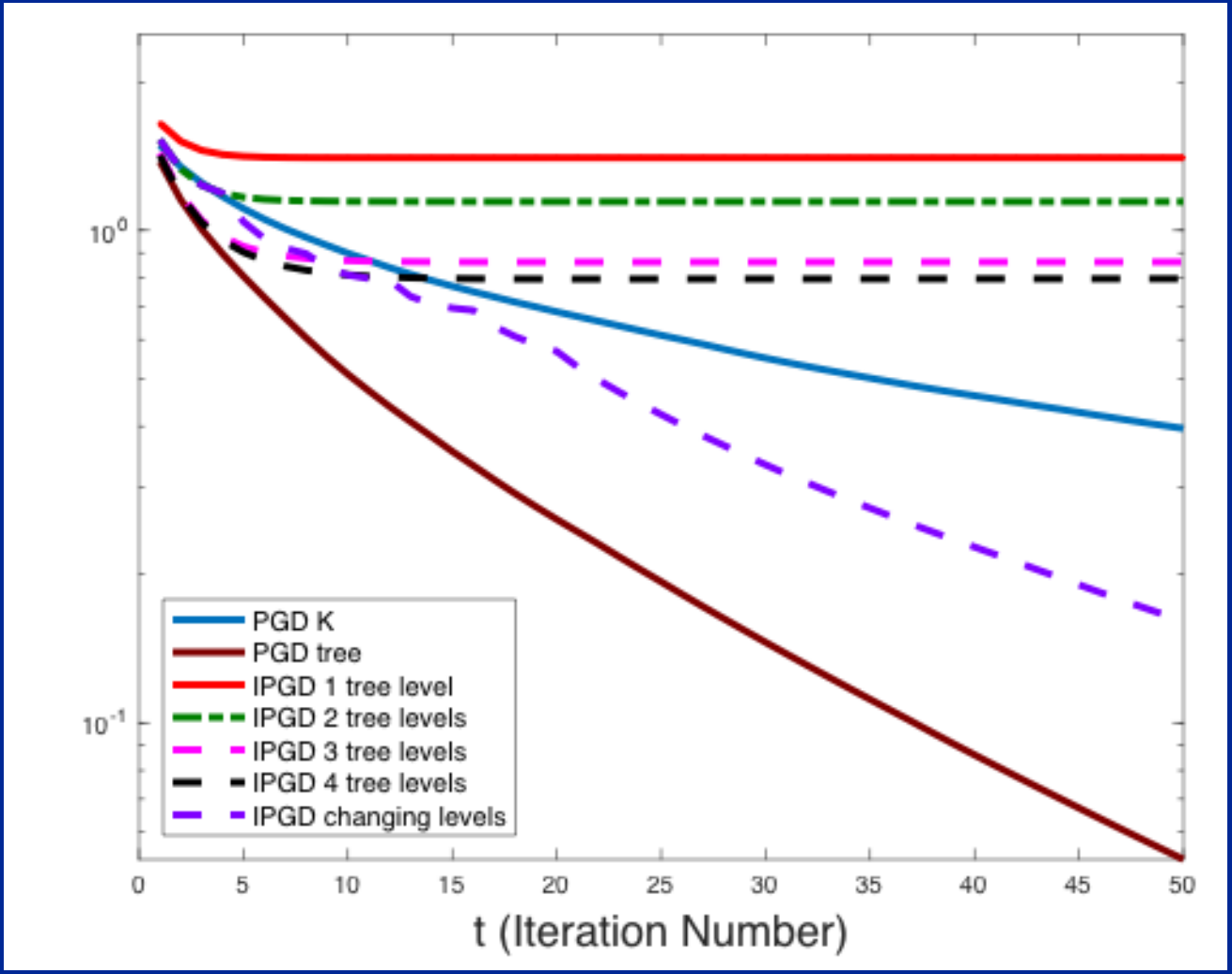
MODEL BASED COMPRESSED SENSING

- Υ is the set of sparse vectors with sparsity patterns that obey a tree structure.
- Projecting onto Υ improves convergence rate compared to projecting onto the set of sparse vectors Υ [Baraniuk et al., 2010].
- The projection onto Υ is more demanding than onto Υ.



- Note that the probability of selecting atoms from lower tree levels is smaller than upper ones.
- P will be a projection onto certain tree levels – zeroing the values at lower levels.

MODEL BASED COMPRESSED SENSING

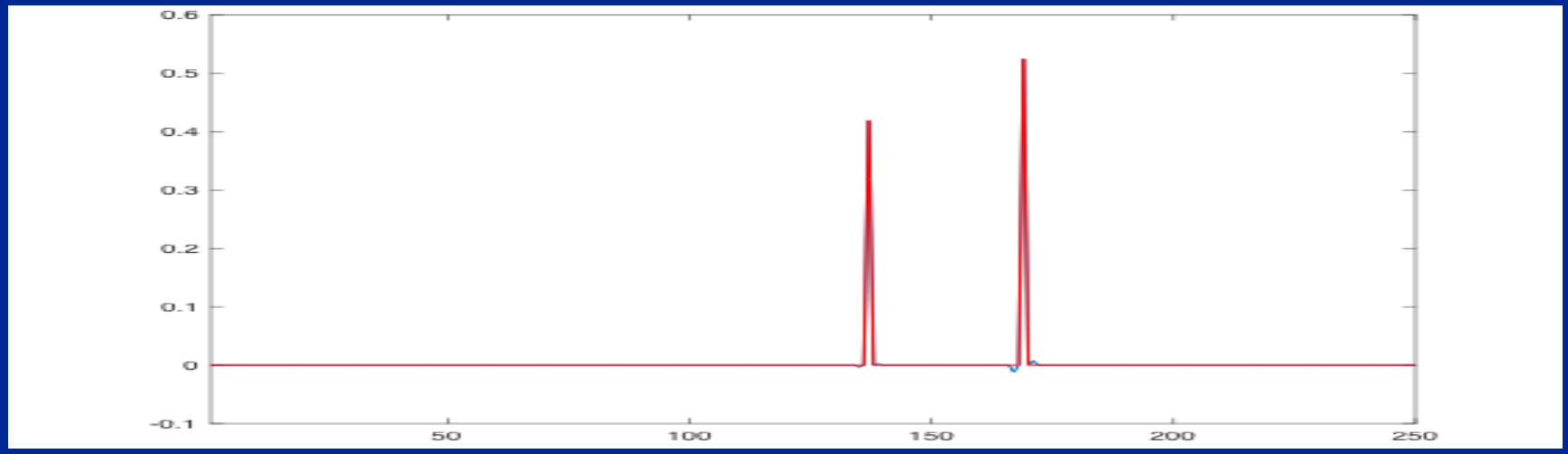


Non-zeros picked entries has zero mean random Gaussian distribution with variance:

- 1 at first two levels
- 0.5^2 at the third level
- 0.2^2 at the rest of the levels

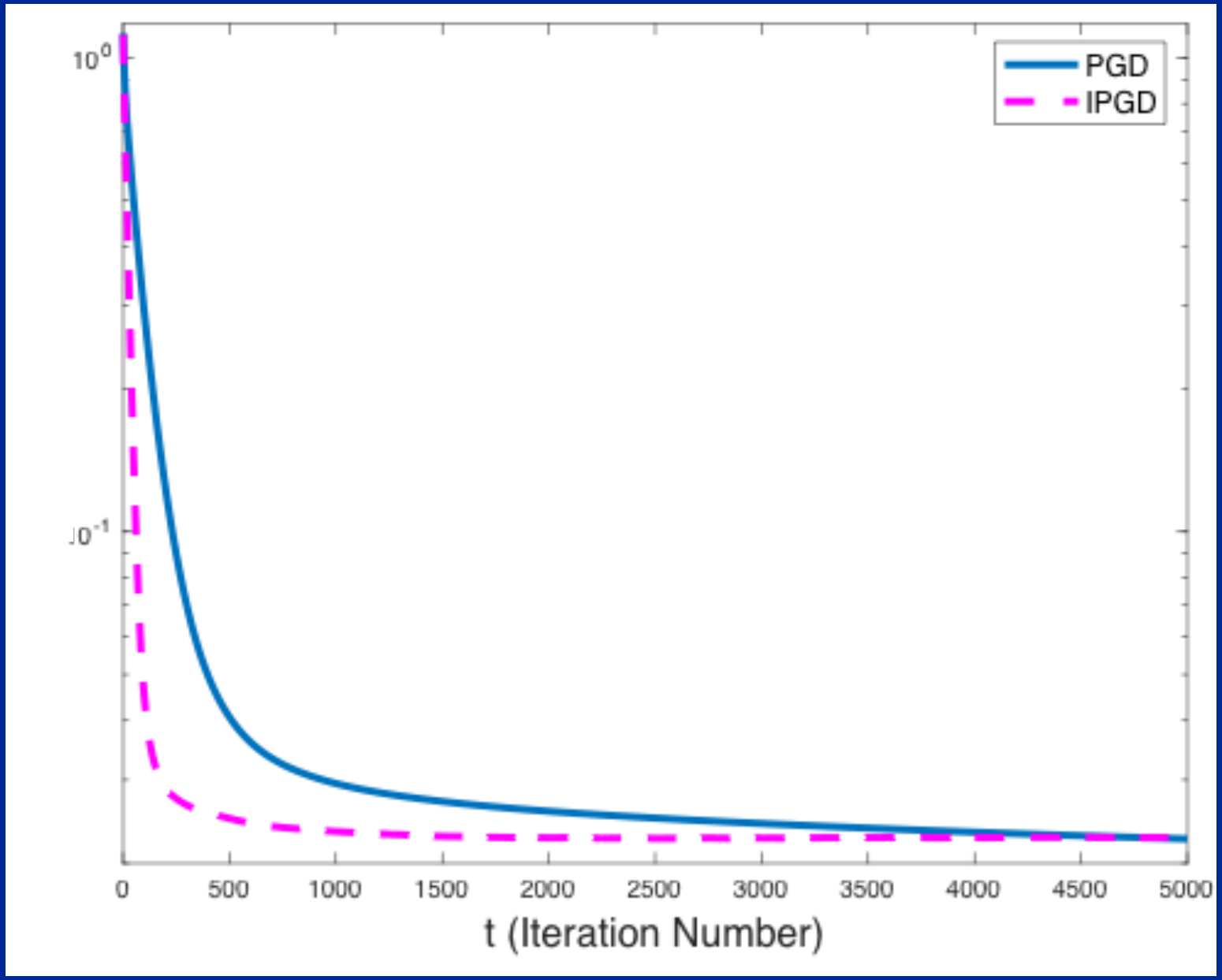
SPECTRAL COMPRESSED SENSING

- Υ is the set of vectors with sparse representation in a 2-times redundant DCT dictionary such that:



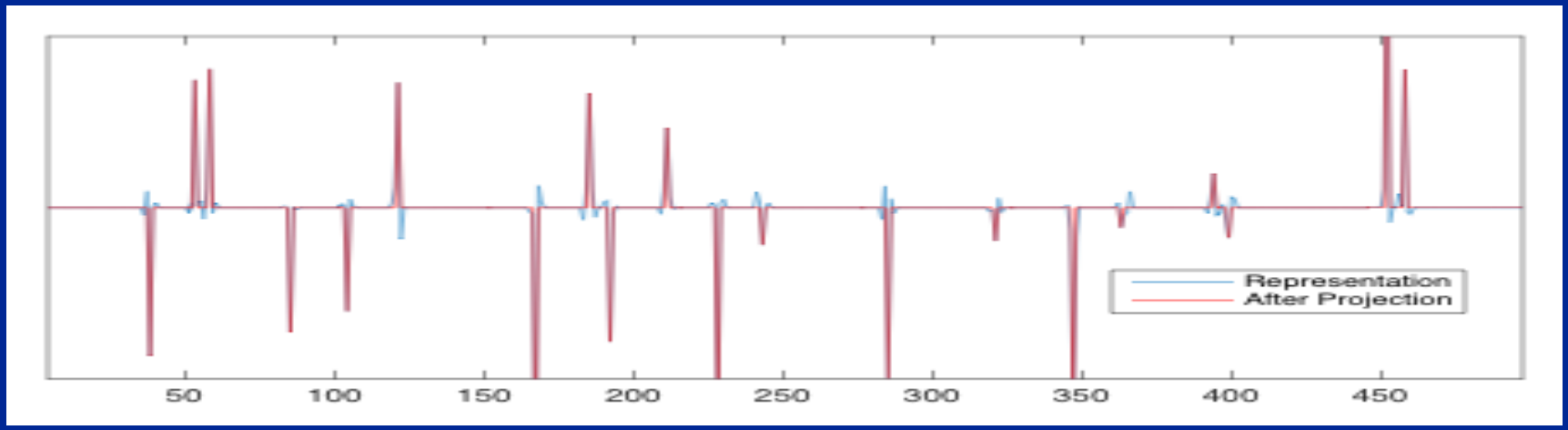
- We set P to be a pooling-like operation that keeps in each window of size 3 only the largest value.

SPECTRAL COMPRESSED SENSING



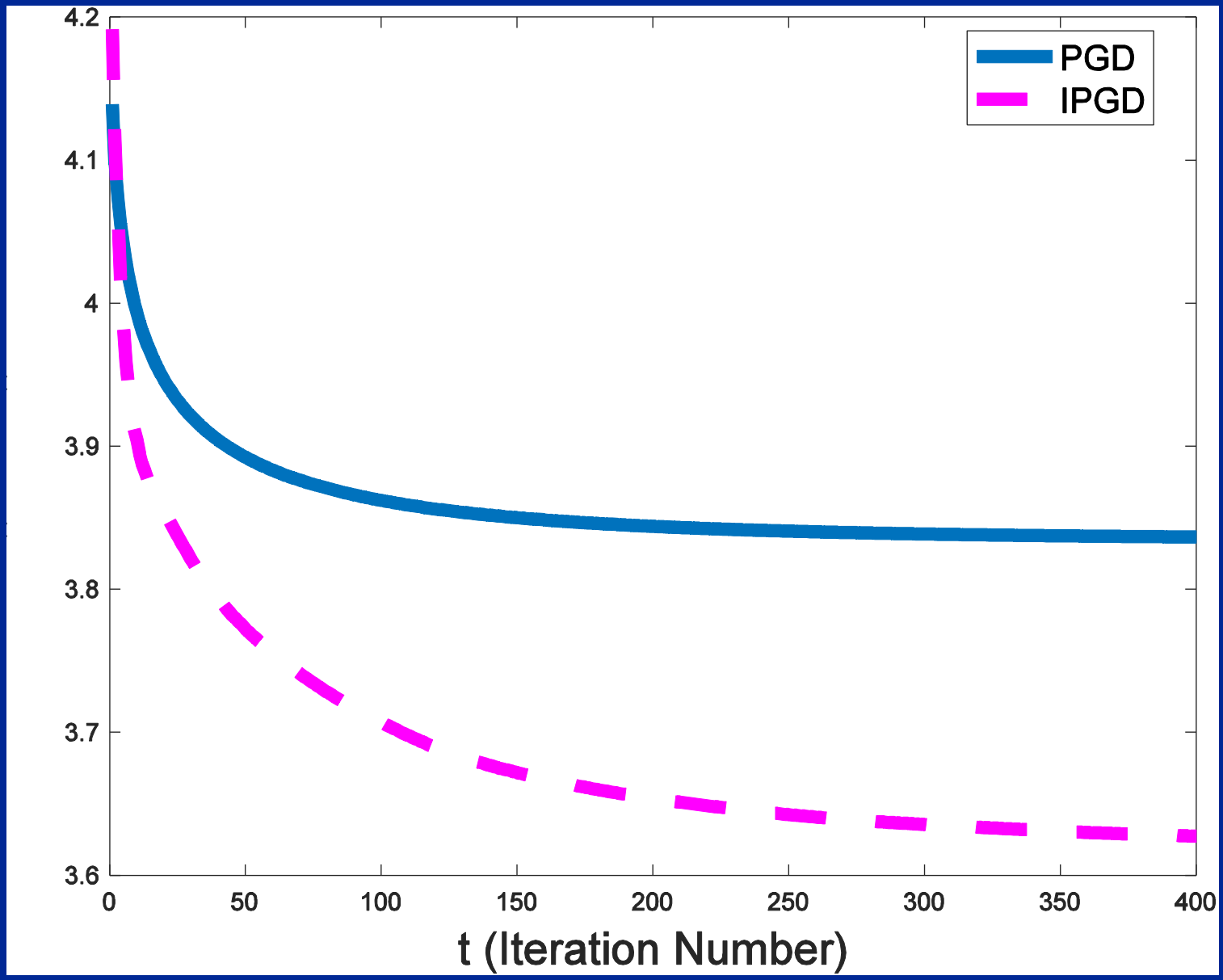
SPECTRAL COMPRESSED SENSING

- Υ is the set of vectors with sparse representation in a 4-times redundant DCT dictionary such that:



- We set P to be a pooling-like operation that keeps in each window of size 5 only the largest value.

SPECTRAL COMPRESSED SENSING

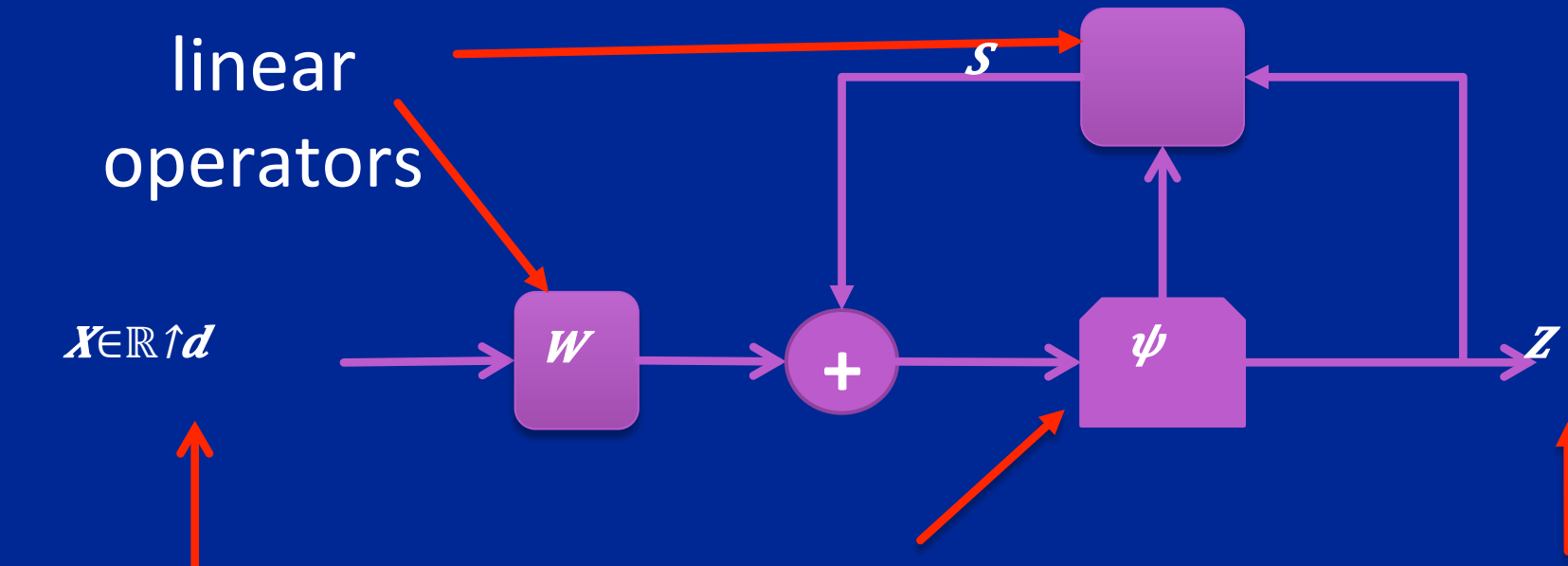


LEARNING THE PROJECTION

- If we have no explicit information about Υ it might be desirable to learn the projection.
- Instead of learning P, it is possible to replace $P(I-\mu A\uparrow T A)$ and $\mu PA\uparrow T$ with two learned matrices S and W respectively.
- This leads to a very similar scheme to the one of LISTA and provides a theoretical foundation for the success of LISTA.

LEARNED IPGD

Learned linear operators

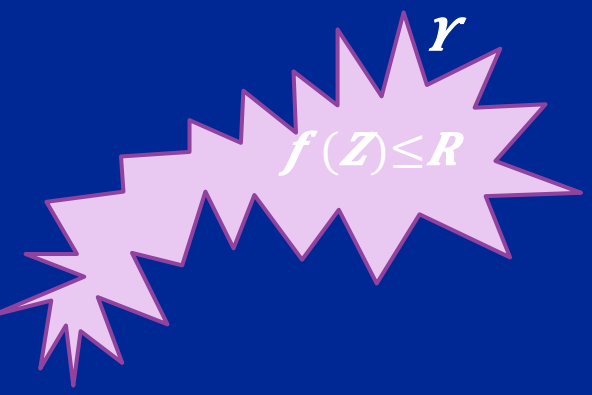


$X=AZ+E$

ψ projects onto the set r

Estimate of z. Aim at solving

$$\min_Z \|X-AZ\| \quad \text{s.t. } f(Z)\le R$$

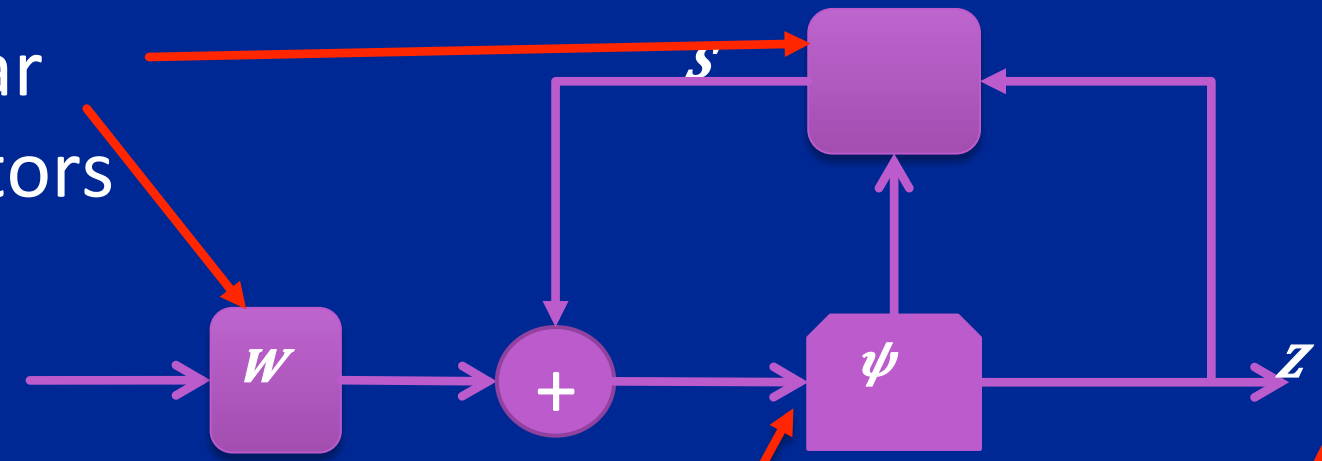


SUPER RESOLUTION

- A popular super-resolution technique uses a pair of low-res and high-res dictionaries [Zeyde et al. 2012]
- The original work uses OMP with sparsity 3 to decode the representation of patches in low-res image
- Then the representation is used to reconstruct the patches of the high-res image
- We replace OMP with LIPGD with 3 levels but higher target sparsity
- This leads to better reconstruction results (with up to 0.5dB improvement)

LISTA

Learned linear operators



$X \in \mathbb{R}\uparrow d$

$X = AZ + E$

ψ is a proximal mapping.

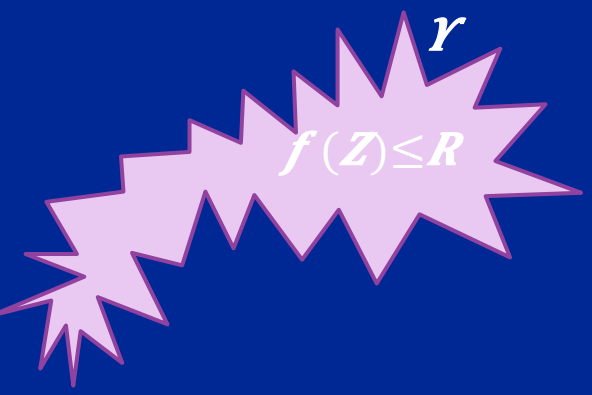
$\psi(U) =$

$\mathbf{argmin}_{Z \in \mathbb{R}\uparrow d} \|U - Z\| + \lambda f(Z)$

Estimate of z.

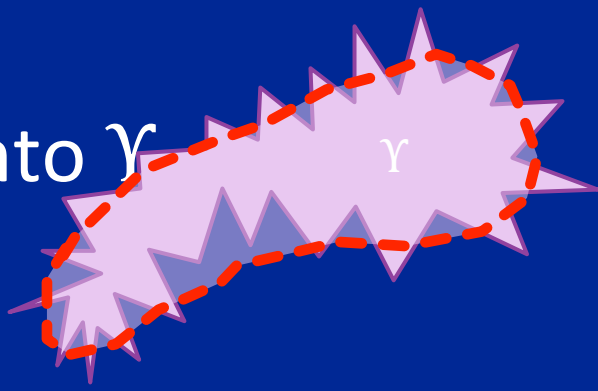
Aim at solving

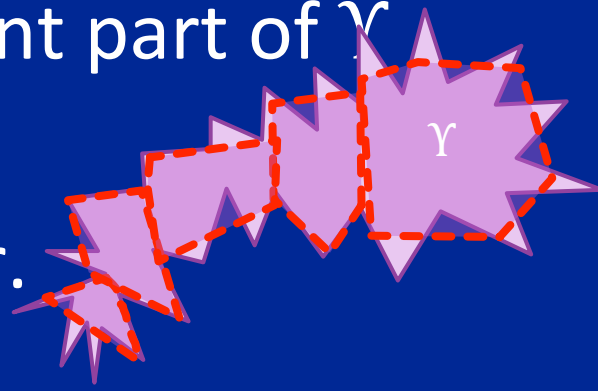
$\min_Z \|X - AZ\| + \lambda f(Z)$



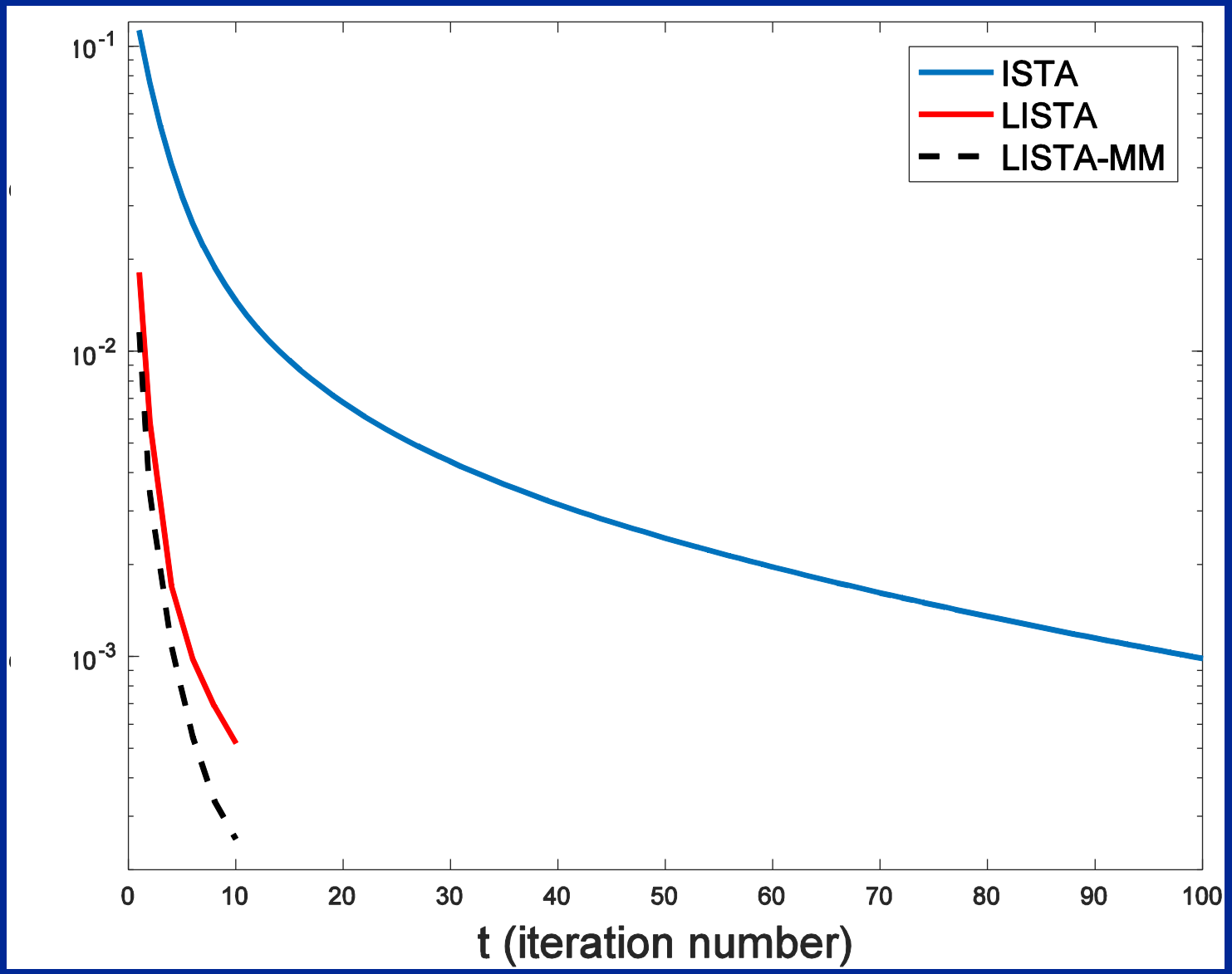
LISTA MIXTURE MODEL

- Approximation of the projection onto Υ 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 Υ
- Training multiple LISTA networks accelerate the convergence further.



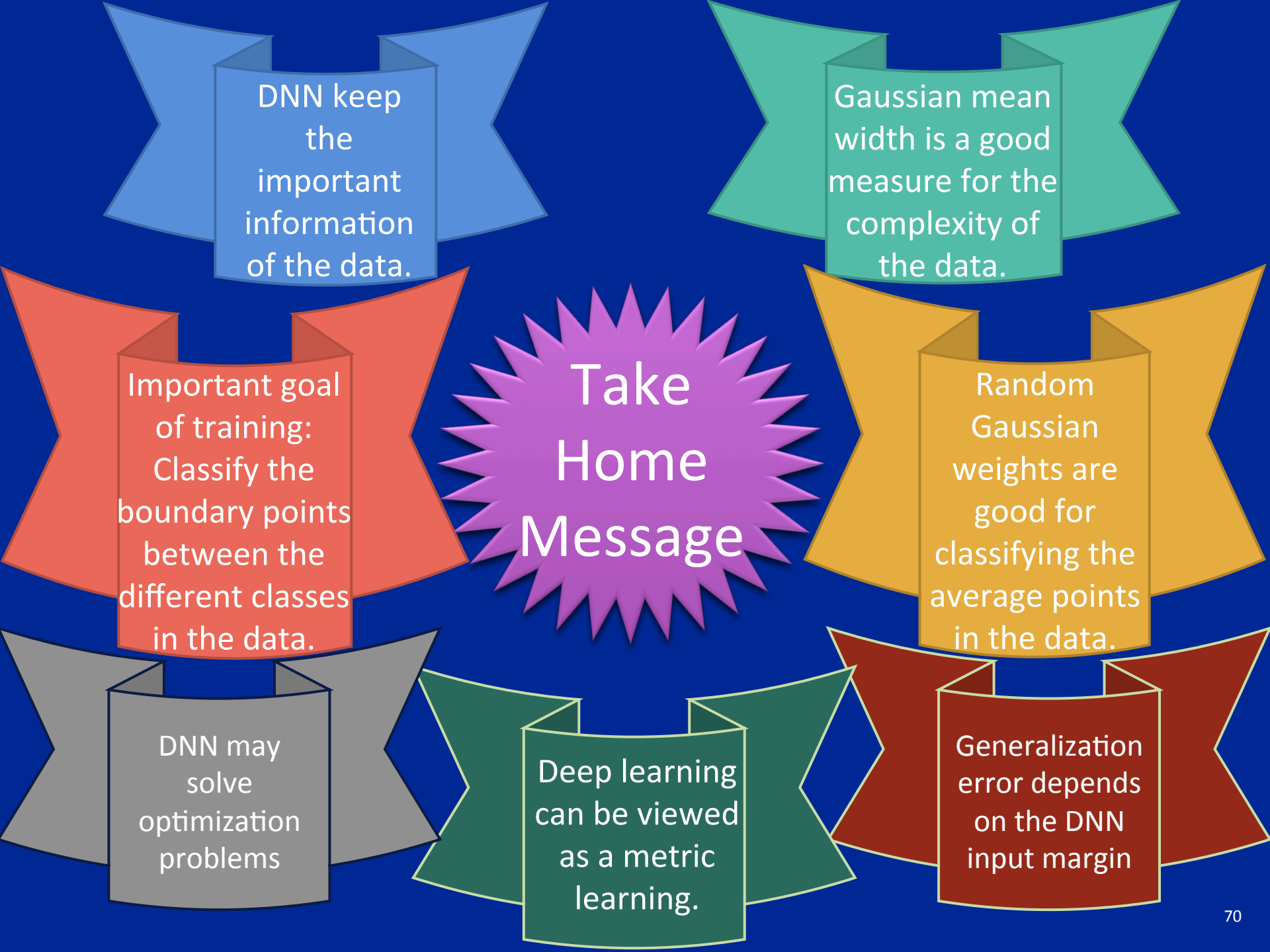


LISTA MIXTURE MODEL



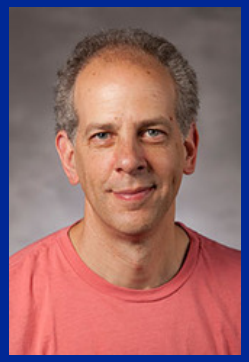
RELATED WORKS

- In [Bruna et al. 2017] it is shown that a learning may give a gain due to better preconditioning of *A.*
- In [Xin et al. 2016] a relation to the restricted isometry property (RIP) is drawn
- In [Borgerding & Schniter, 2016] a connection is drawn to approximate message passing (AMP).
- All these works consider only the sparsity case


DNN keep the important information of the data.
Gaussian mean width is a good measure for the complexity of the data.
Important goal of training: Classify the boundary points between the different classes in the data.
Take Home Message
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

ACKNOWLEDGEMENTS

Yonina C. Eldar
Technion



Guillermo Sapiro
Duke University

Jure Sokolic
UCL

Alex M. Bronstein
Technion

Miguel Rodrigues
UCL

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 onSignal 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.

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, “Margin Preservation of Deep Neural Networks”, arXiv, abs/1605.08254, 2016.
- 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

- A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Img. Sci., 2(1):183–202, Mar. 2009.
- 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”, arXiv, abs/1507.04793, 2016.
- R. Giryes, Y. C. Eldar, A. M. Bronstein, G. Sapiro, “Tradeoffs between Convergence Speed and Reconstruction Accuracy in Inverse Problems”, arXiv, abs/1605.09232, 2016.
- 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.
- M. F. Duarte & R. G. Baraniuk, “Spectral compressive sensing”, Appl. Comput. Harmon. Anal., vol. 35, no. 1, pp. 111 – 129, 2013.
- J. Bruna & T. Moreau, Adaptive Acceleration of Sparse Coding via Matrix Factorization, arXiv abs/1609.00285, 2016.