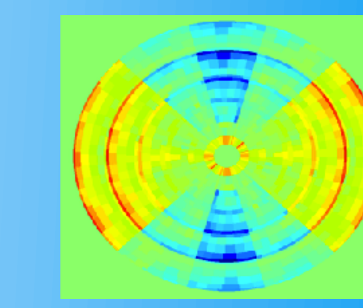
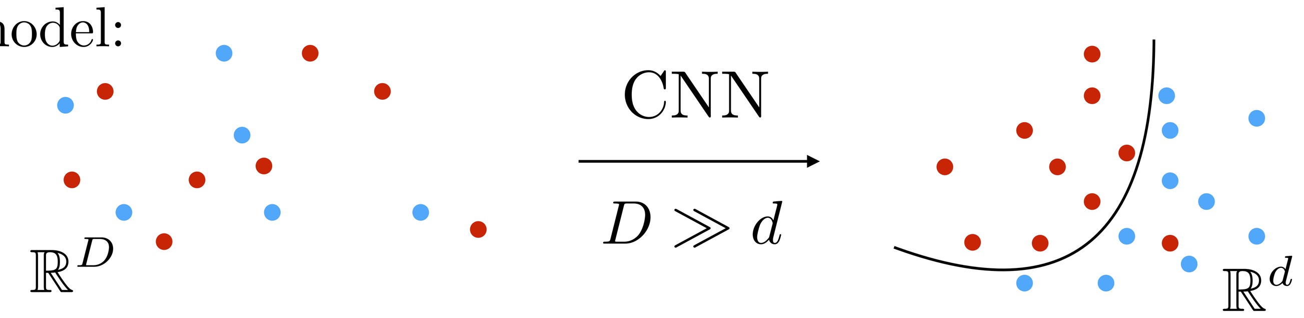
Building a Regular Decision Boundary with Deep Networks

Edouard Oyallon
DATA, Département Informatique
Ecole Normale Supérieure/PSL Research University



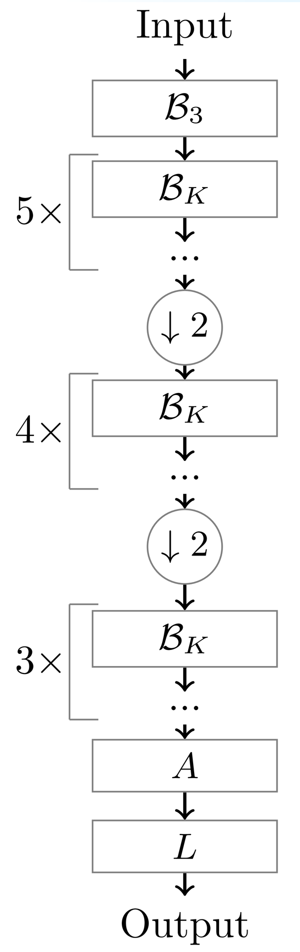
Understanding Deep Representations

- Separation, contraction are necessary properties of a successful model:

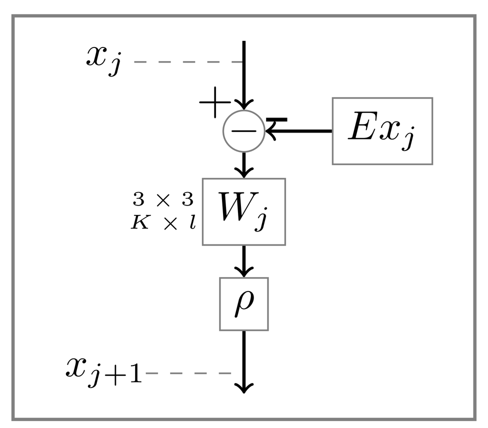


- How can we relate it to the depth? How can we design the non-linearity? How to measure the dimensionality reduction?

Simplified CNN framework



- Depends only on the width K and non-linearity ρ
- No max-pooling or ad-hoc non-linear modules



A block $\mathcal{B}_l$

- W_j: convolution with K inputs and outputs
- Only 13 layers!

BatchNorm

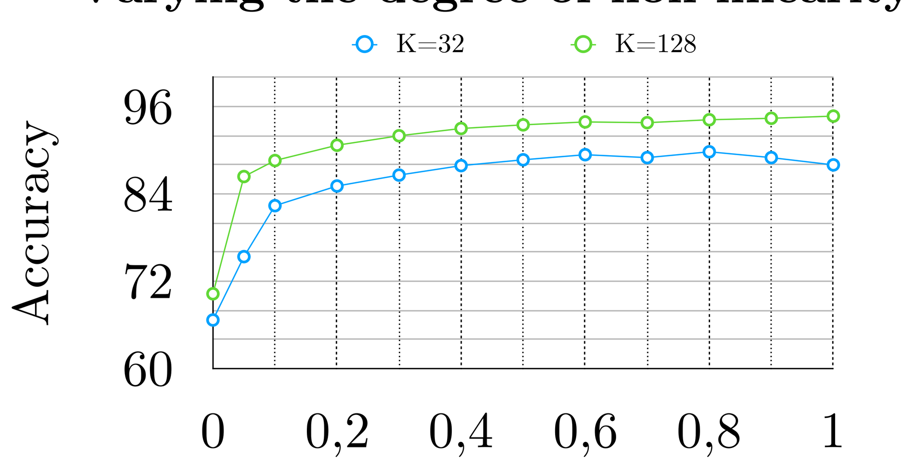
$$x_{j+1} = \rho W_j(x_j - Ex_j)$$

Benchmarking

	Depth	#params	CIFAR10	CIFAR100
Ours	13	28M	95.4	79.6
SGDR	28	150M	96.2	82.3
WResNet	28	37M	95.8	80.0
All-CNN	9	1.3M	92.8	66.6

Influence of the Hyper Parameters

Varying the degree of non-linearity



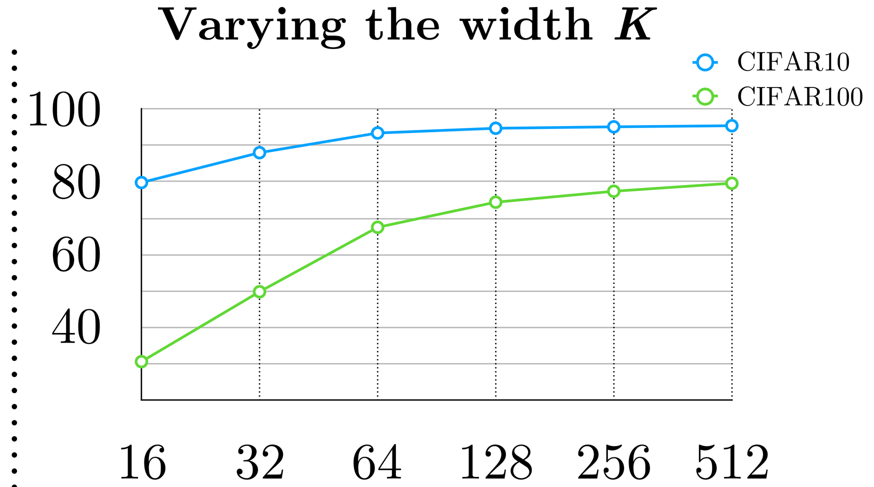
$\rho = \text{ReLU}_k^K$ Degree: $\dfrac{k}{K}$

where:

$$\text{ReLU}_k^K(x)(.,l) \triangleq \begin{cases} \text{ReLU}(x(.,l)), & \text{if } l \leq k \\ x(.,l), & \text{otherwise} \end{cases}$$

"More non-linear" is better?

Varying the width K



- It permits to study a smaller model!

Depth or width?

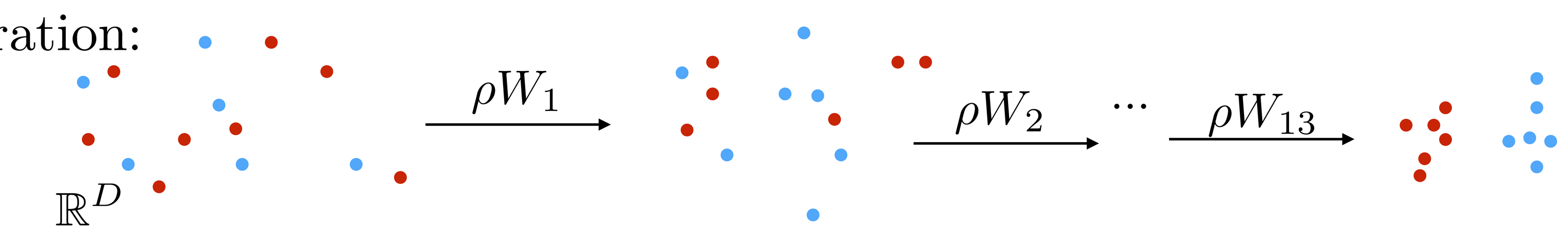
Necessary Conditions for ρ?

- Non-linearity can be designed thanks to mathematical considerations:
  - Non-expansivity
  - Phase removing
  - Continuity
- We show that a non-continuous, non-Lipschitz and sign-preserving non-linearity obtains 89.0% on CIFAR10:

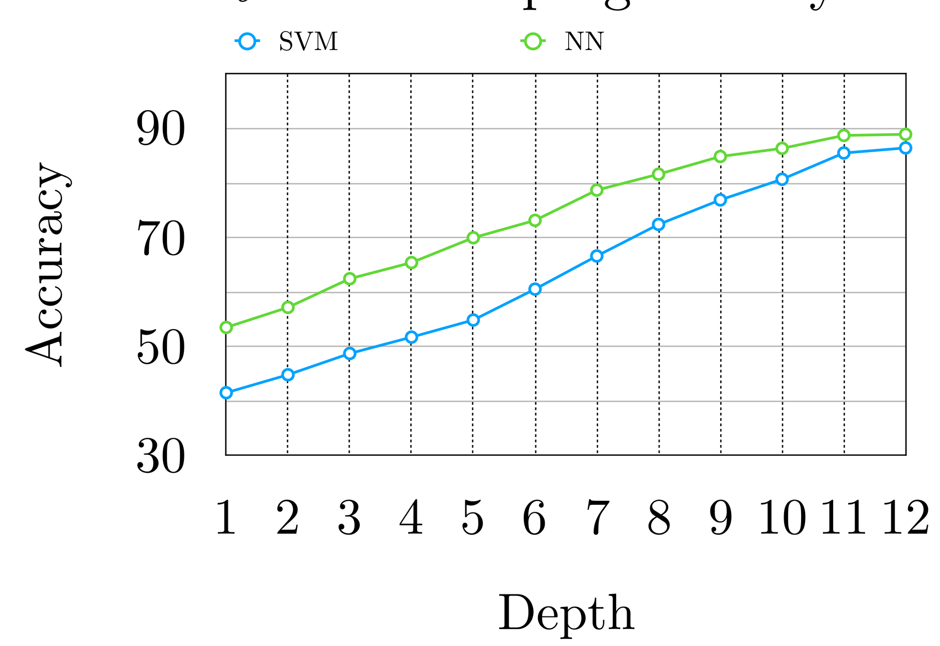
Contradiction! $\rho(x) = \text{sign}(x)(\sqrt{|x|} + 0.1)$

Progressive Improvements of Local Classifiers

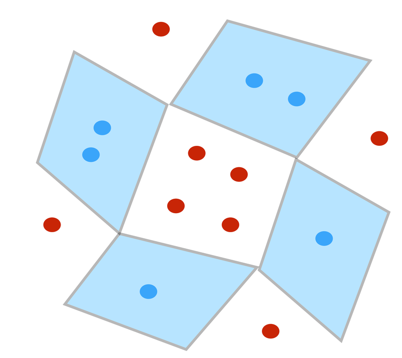
- We aim to show the cascade permits a progressive contraction & separation:



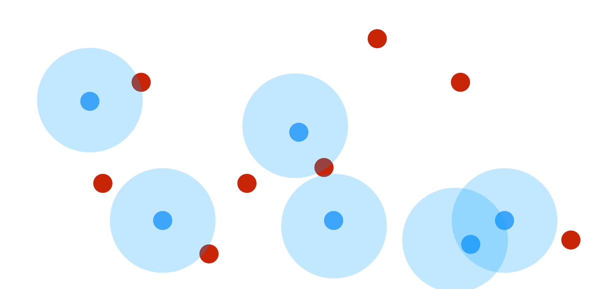
- Accuracy of local classifiers progressively improves with depth: the ℓ^2 metric is progressively more meaningful.



Nearest Neighbor (NN)



Gaussian SVM

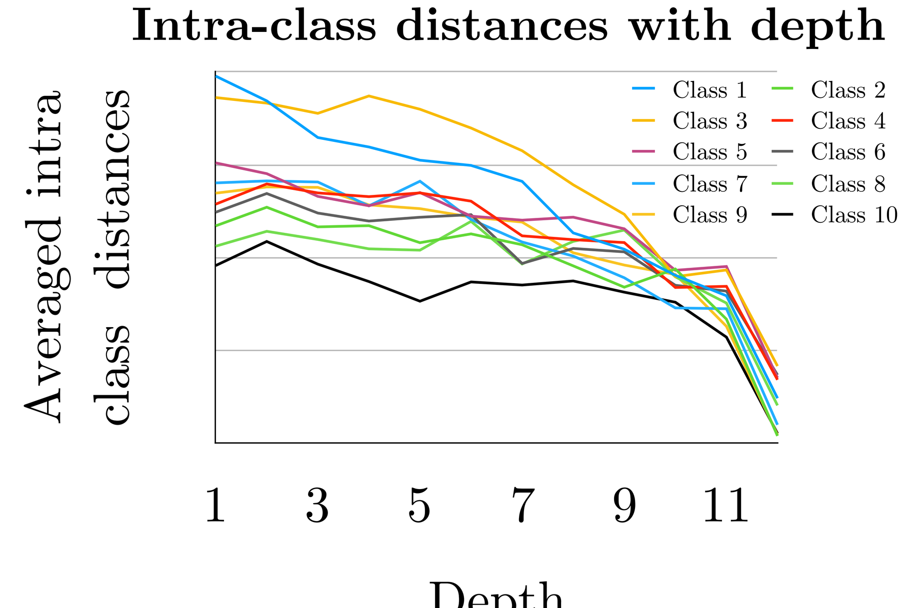


$x^{(k)}$: k-th nearest neighbour
$y(x)$: class of x

Global Contraction and Local Separation

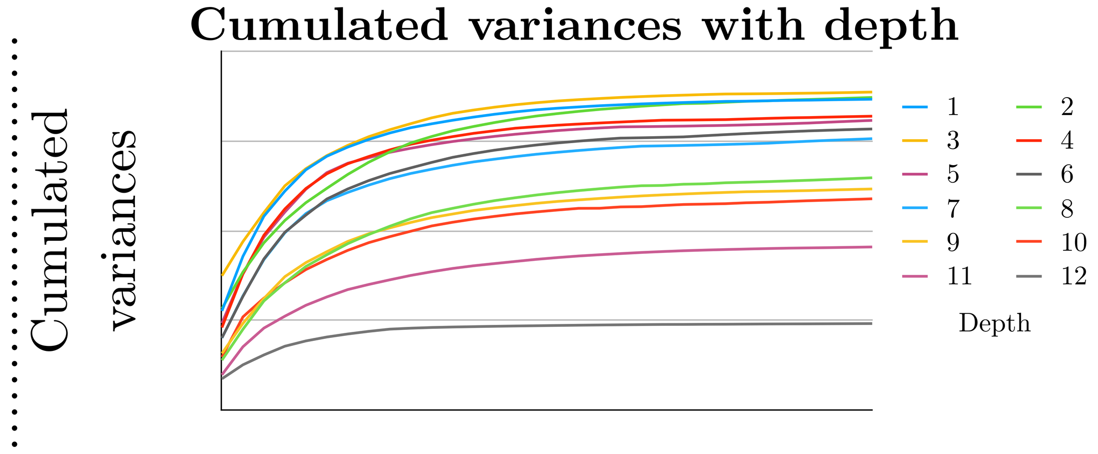
- Progressive contraction:

Intra-class distances with depth



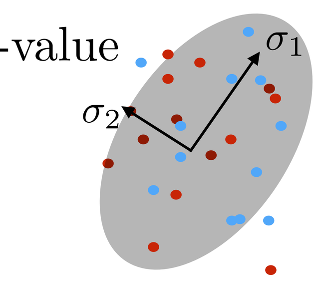
$$\frac{1}{5000^2} \sum_{y(x_j)=y(\tilde{x}_j)=c} \|x_j - \tilde{x}_j\|$$

Cumulated variances with depth



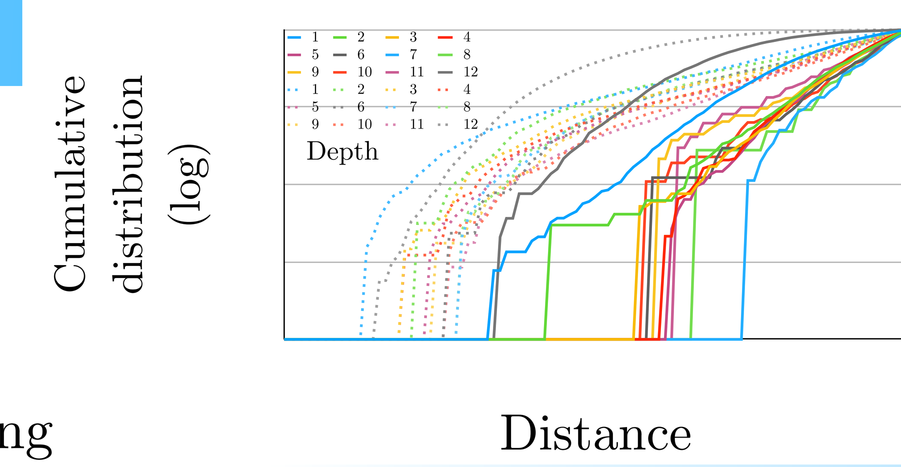
σ_p the p-th PCA ordered eigen-value

We study: $\Sigma_P = \sum_{p \leq P} \sigma_p$



Refining those measures?

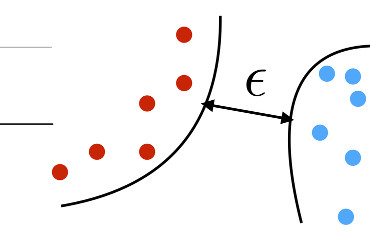
- Margin:



Cumulative distributions of:

$\{\|x_j^{(1)} - x_j\|\}_{y(x_j^{(1)})=y(x_j)}$ ....

$\{\|x_j^{(1)} - x_j\|\}_{y(x_j^{(1)})\neq y(x_j)}$ ———



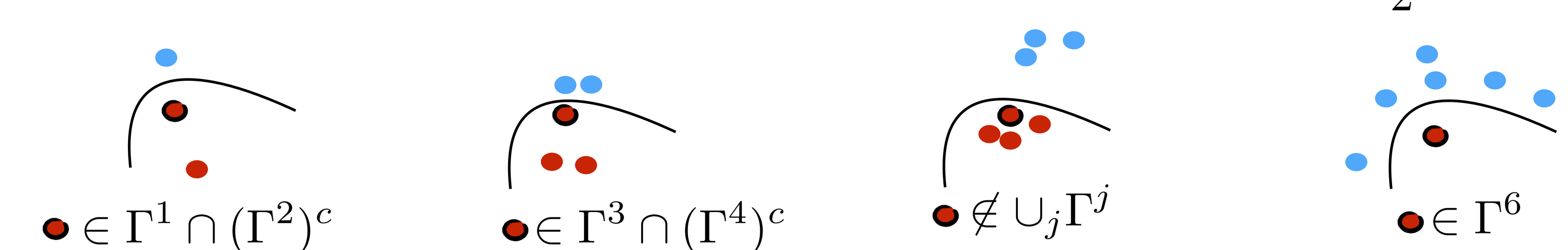
Local Support Vectors: Exploring Regularity

- Estimating the intrinsic dimension of the classification boundary is hard *(curse of dimensionality!)*: we introduce local support vectors:

$$\Gamma_j^1 = \{x_j | y(x_j^{(1)}) \neq y(x_j)\}$$

- It permits to measure the regularity of the classification boundary at depth j:

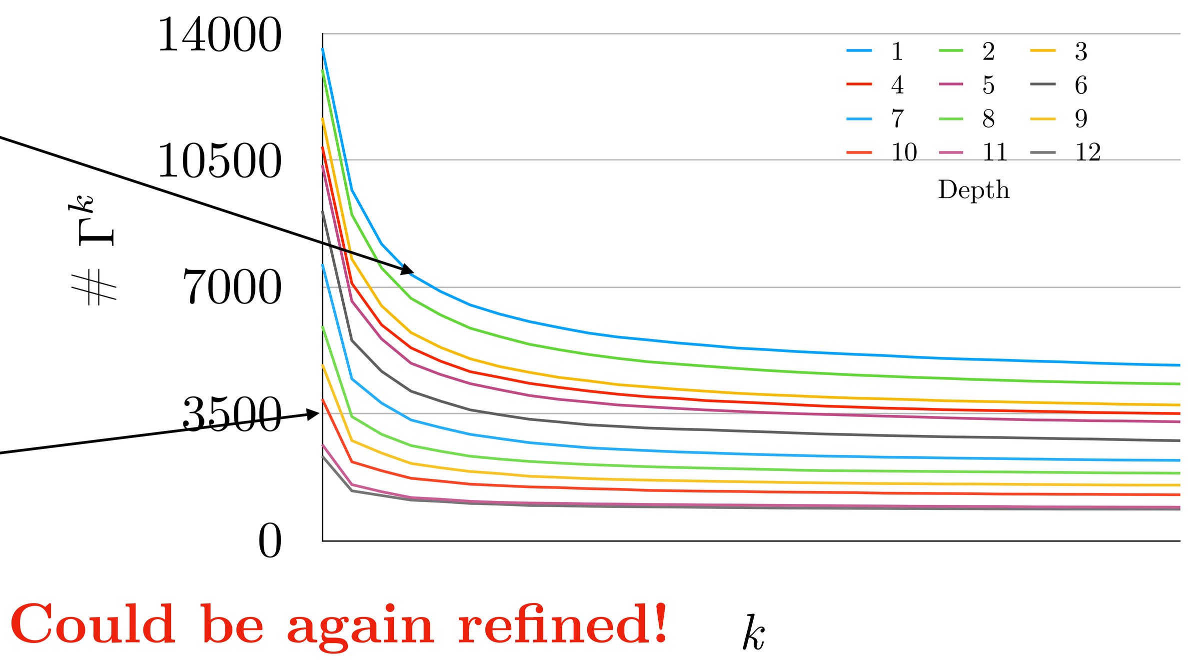
$$\Gamma_j^{k+1} = \{x_j \in \Gamma_j^k | \text{card}\{y(x_j) \neq y(x_j^{(l)}), l \leq k+1\} > \frac{k+1}{2}\}$$



Fast decay to stationary value indicates contraction
(only the nearest neighbour helps to classify)

Small amount of local support vectors shows separation
(the boundary of classification does not require many support vectors: simple to separate)

Regularity of Local Support Vectors with depth



Could be again refined!

Conclusion

- Which mechanisms permit the dimensionality reduction to occur? *Linearization* of complex variabilities?
- *Theoretical guarantees* are necessary to engineer and understand better deep networks.

Contacts

- Email: edouard.oyallon@ens.fr
- Website (codes): http://www.di.ens.fr/~oyallon/
- Website of my team: http://www.di.ens.fr/data/