Spectral Ergodicity in Deep Learning Architectures

Mehmet Süzen[a], Cornelius Weber[b], Joan J. Cerdà[c]
[a] FZJ
[b] University of Hamburg
[c] University of the Balearic Islands

6 December 2017, Garching

Spectral Ergodicity in Deep Learning Architectures via Surrogate Random Matrices

Mehmet Süzen, Cornelius Weber, Joan J. Cerdà

arXiv preprint arXiv:1704.08303 (2017)

dataset zenodo doi: 10.5281/zenodo.822411

code zenodo doi: 10.5281/zenodo.579642

Python package:

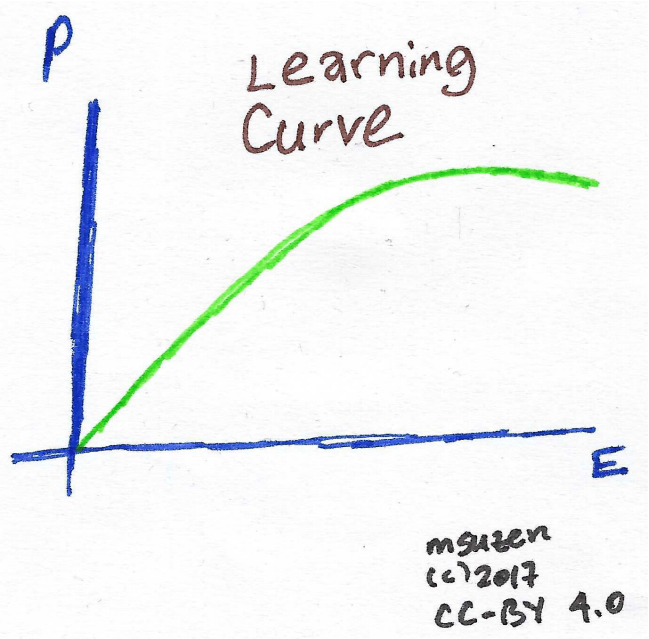
https://pypi.python.org/pypi/bristol

What is *learning* in machine learning?

*A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.*

Tom Mitchell (1997)

Learning/Forgetting curve
Hermann Ebbinghaus (1913)



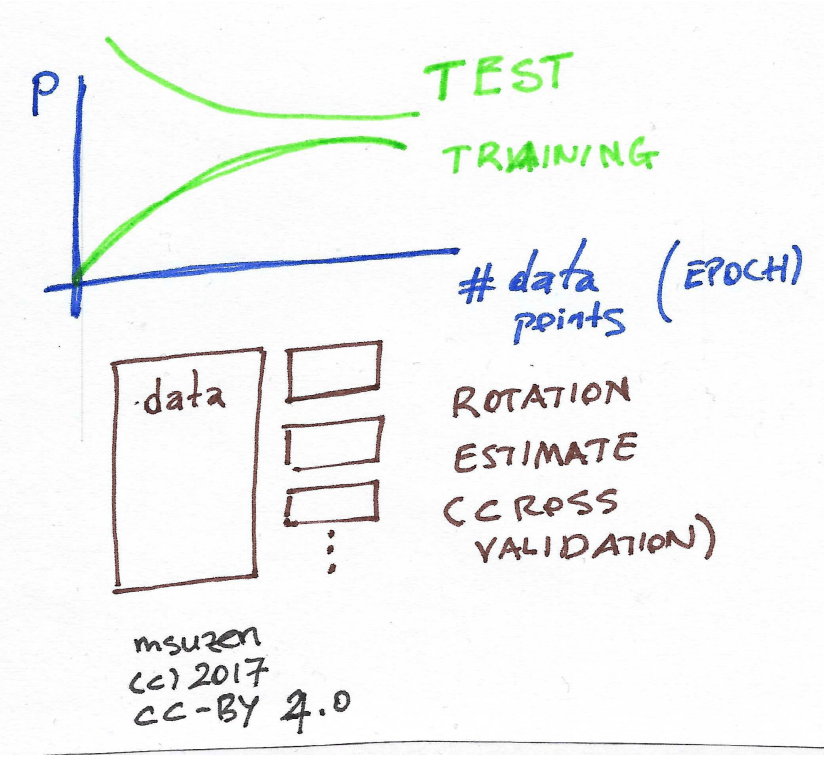
What is *supervised* learning?

- Not only a function optimization or function inference.

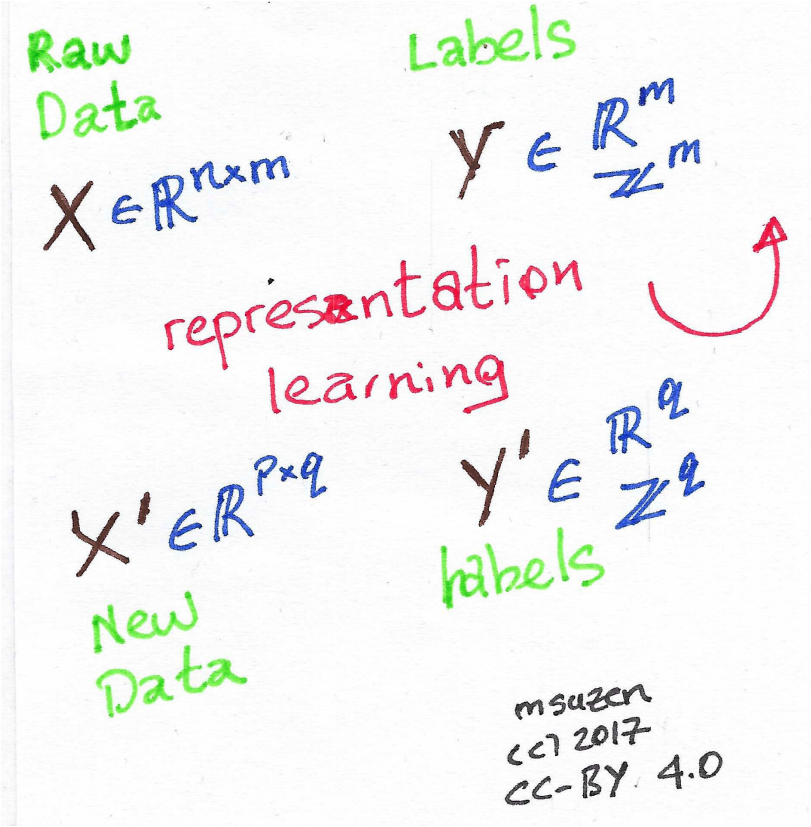
- Combination of two

  - *Algorithm is generalized.*
    - ***Avoid Overtraining***

  - *Algorithm obeys some form of Occam's Razor*
    - ***Avoid Overfitting***



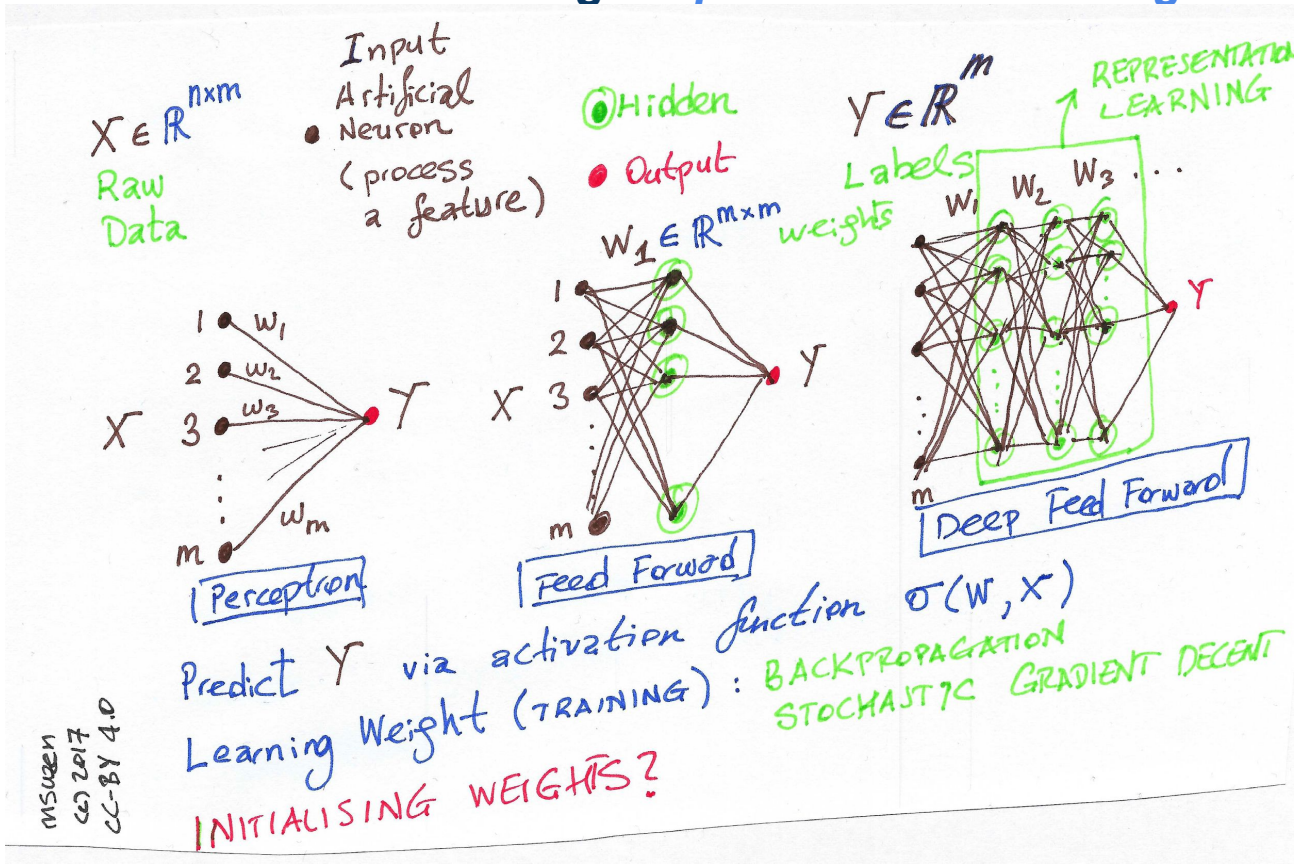
Main activity in machine learning: *Feature engineering*



<u>Manual</u> Representation learning

- Transformation
  - Encode categorical information: one-hot-encode, binning.
  - Nonlinear functions, convolutions.
- Combine features
- Feature selection
  - Forward, backward
- Dimensional reduction
  - PCA, ICA, Gaussian Random Projection, CS

Modern machine learning: *Representation learning*



Deep learning architectures: Success

- Dramatic increase in learning accuracy:
  LeCun, Bengio, Hinton
  Nature, 521, 436 (2015)

- Example in computer vision: **RoboCup Soccer**
  Robocup soccer using convolutional neural networks (CNNs), Speck, Weber, Wermter (2016)
  - *Can a robot reliably locate the ball?*
    https://www2.informatik.uni-hamburg.de/wtm/publications/2016/SBWW16/Barros_speck_Ball_Localization_Robocup_Soccer_Convolutional_Neural_Networks_2016.pdf

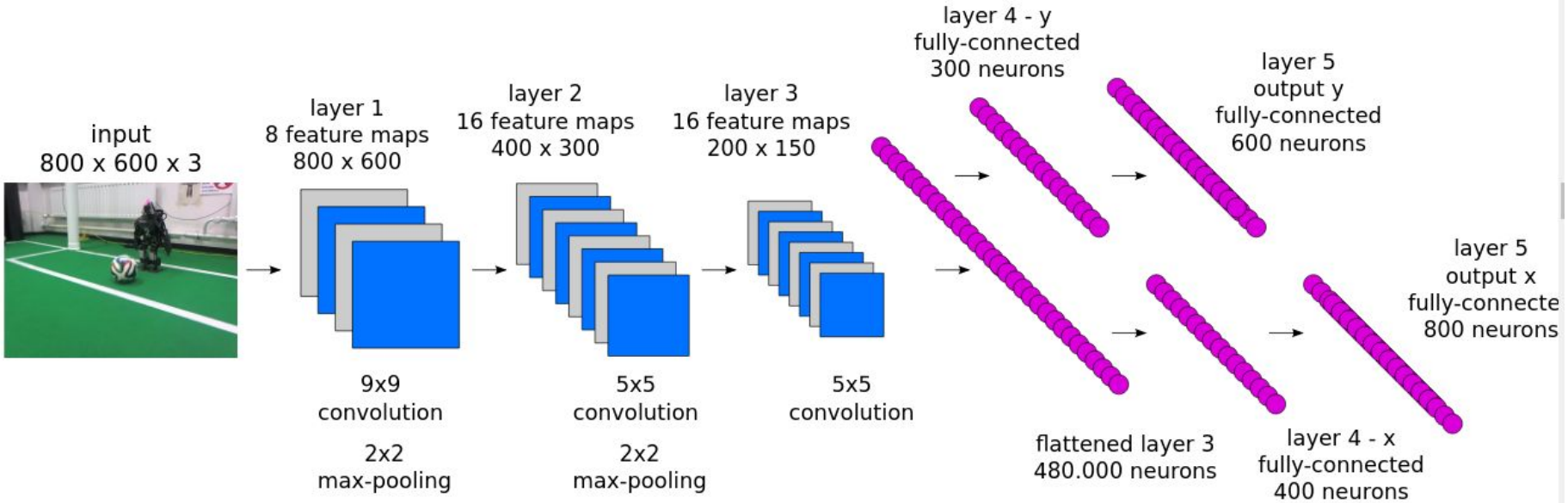
Deep learning architectures: Success

Red dominates due to lighting

Moving ball

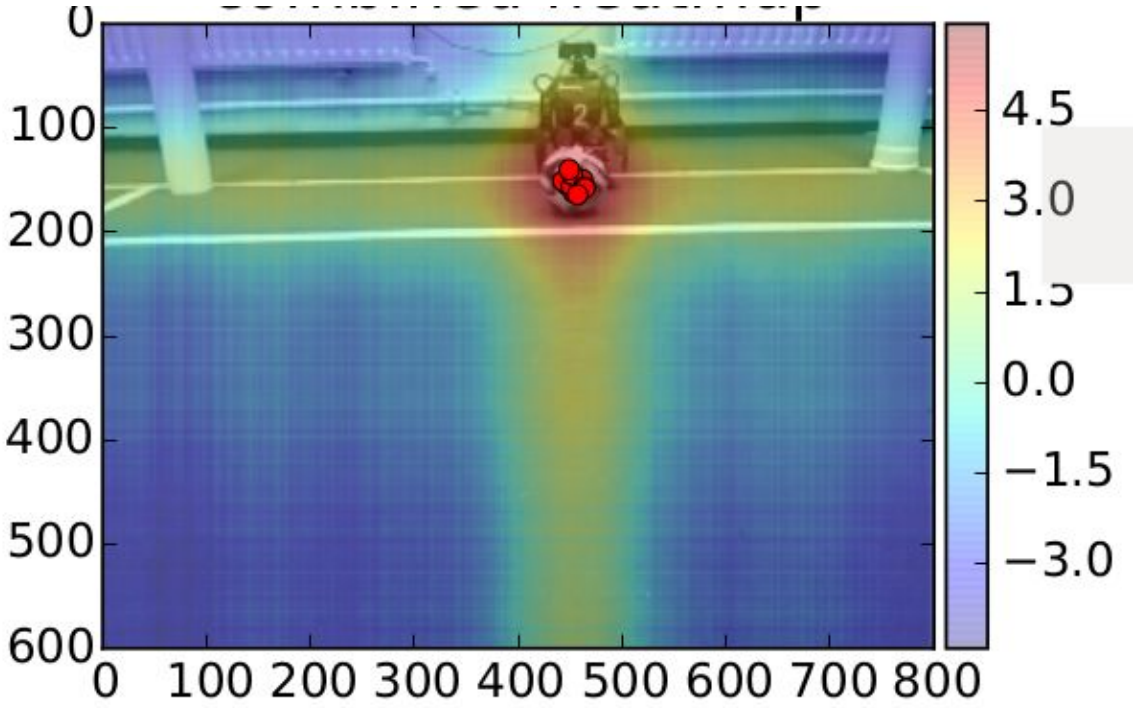
Speck, Weber, Wermter (2016)

Deep learning architectures: Success



Speck, Weber, Wermter (2016)

Deep learning architectures: Success



Speck, Weber, Wermter (2016)

Core issues in learning Performance:

- Initialization of weight matrices.
- How to design a deep learning architecture?

Deep learning architectures: Success

- No theoretical explanation but success of deep learning attributed to
  - *network topology,* multiple processing layers,
  - availability of *large labelled datasets*,
  - leap in *computing capacity*.

Deep learning architectures: A generic approach

A generic approach to identify success in deep learning:
Use ensemble of random matrices as a surrogate for weight matrices

- Remember: Initializing Weights problems: Dynamical Isometry
- independent of learning algorithm and dataset
- generic in network topology and node characteristics
- conceptually easy to simulate

Deep learning architectures: A generic approach

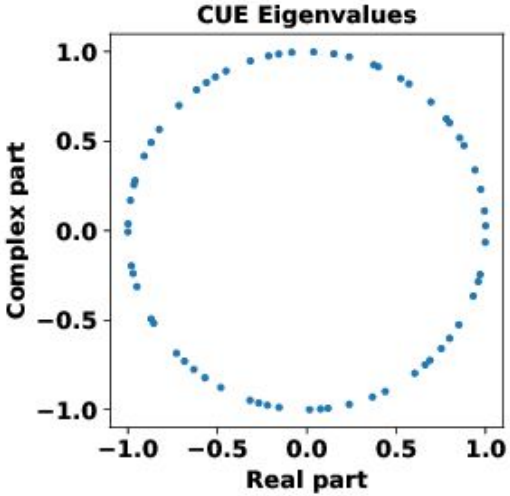
A generic approach to identify success in deep learning:
Use ensemble of random matrices as a
surrogate for weight matrices

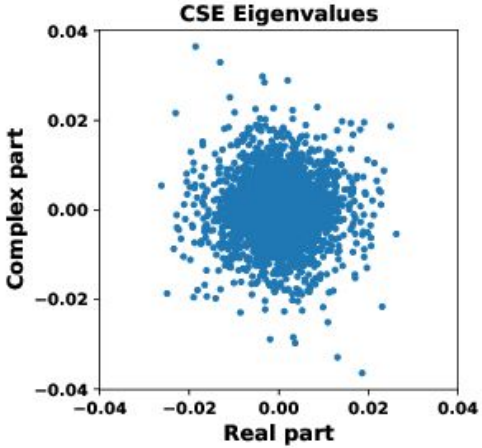
- interpretation:
  - ensemble of weight matrices for different layers
  - ensemble of weight matrices for a given topology

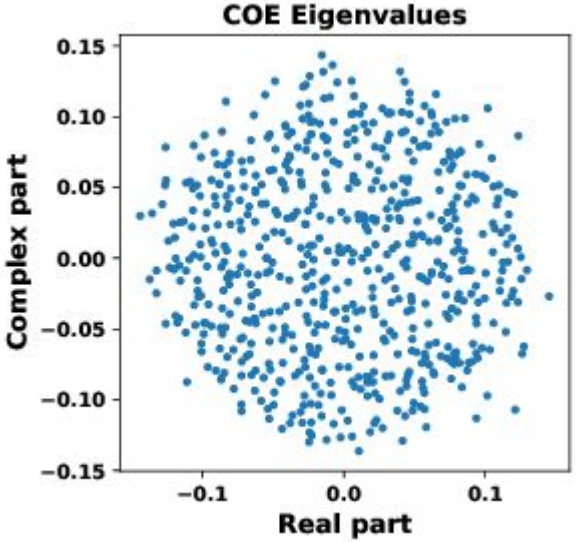
- unit spectral radius: Circular ensembles
  - Circular Orthogonal Ensemble (COE)
  - Circular Unitary Ensemble (CUE)
  - Circular Symplectic Ensemble (CSE)

Deep learning architectures: Circular ensembles

Spectra of COE, CUE and CSE are generated via *Bristol Python* package. 40 matrices for each of the ensemble with matrix sizes N=64, 128, 256, 512, 768, 1024.







Deep learning architectures: Spectral ergodicity

Spectra histogram bin

single matrix spectral density

$$\Omega_k^N \equiv \Omega^N(b_k) = \frac{1}{M \cdot N} \sum_{j=1}^{M} \left[\rho_j(b_k) - \bar{\rho}(b_k)\right]^2 ,$$

Spectral ergodicity given N and binning

M: ensemble size
N: matrix size

ensemble averaged spectral density

Deep learning architectures: Approach to spectral ergodicity

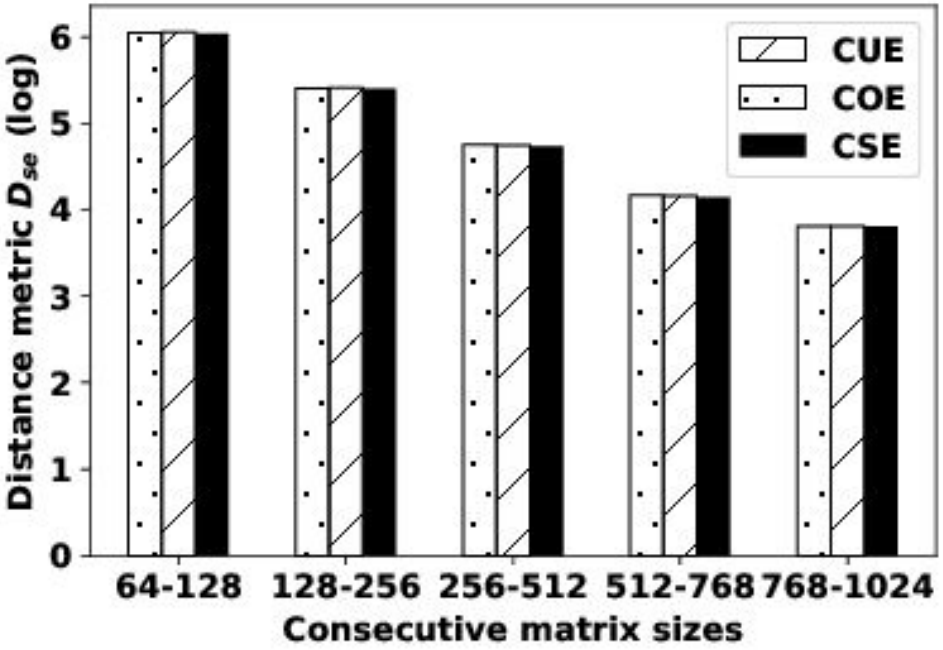
Distance metric: Reminiscent of Kullback-Leibler divergence

Measures the effect of depth in different topologies

$$D_{KL}(\Omega^{N_a}||\Omega^{N_b}) = \sum_{k=1}^{K} \Omega_k^{N_a} \log_2(\Omega_k^{N_a}/\Omega_k^{N_b})$$

Spectral ergodicity given different lengths.

Deep learning architectures: Approach to spectral ergodicity



Conclusions and outlook

- We attribute the success of deep learning to spectral ergodicity of a given topology and learning algorithm using surrogate random matrices.
- Interpretation of ensemble can be different layers or entire network.
- The work can be extended to synaptic matrices in the context of spiking neurons.

Thank you

Q & A