On Mathematical Theories of Deep Learning: iv

Yuan YAO

HKUST

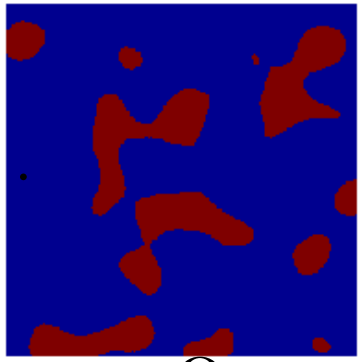
What's the Landscape of Empirical Risks and How to optimize them efficiently?

Over-parameterized models lead to simple landscapes while SGD finds flat minina.

Sublevel sets and topology

- Given loss $E(\theta)$, $\theta \in \mathbb{R}^d$, we consider its representation in terms of level sets:

$$E(\theta) = \int_0^\infty \mathbf{1}(\theta \in \Omega_u) du \ , \ \Omega_u = \{y \in \mathbb{R}^d \ ; \ E(y) \leq u\}$$



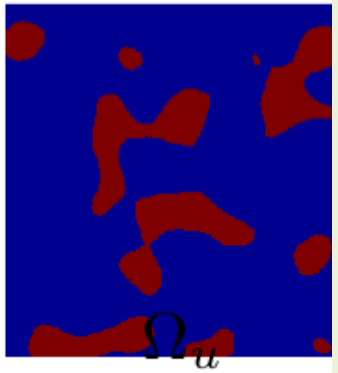
Ω_u

- A first notion we address is about the topology of the level sets .

- In particular, we ask how connected they are, i.e. how many connected components N_u at each energy level u?

Topology of Non-convex Risk Landscape

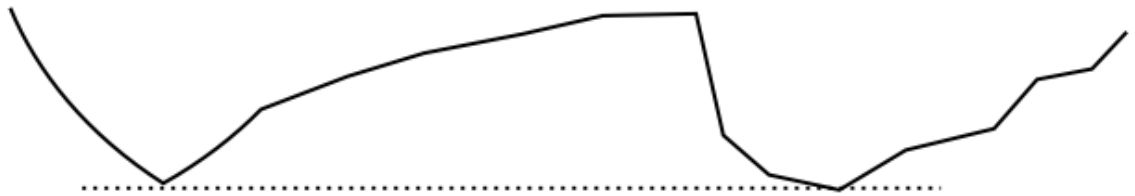
- A first notion we address is about the topology of the level sets .
  - In particular, we ask how connected they are, i.e. how many connected components N_u at each energy level u?
- This is directly related to the question of global minima:



Proposition: If $N_u = 1$ for all u then E has no poor local minima.

$$(\text{i.e. no local minima } y^* \text{ s.t. } E(y^*) > \min_y E(y))$$

- We say *E* is *simple* in that case.
- The converse is clearly not true.



Weaker: P.1, no spurious local valleys

Given a parameter space Θ and a loss function $L(\theta)$ as in (2), for all $c \in \mathbb{R}$ we define the sub-level set of L as

$$\Omega_L(c) = \{\theta \in \Theta : L(\theta) \leq c\}.$$

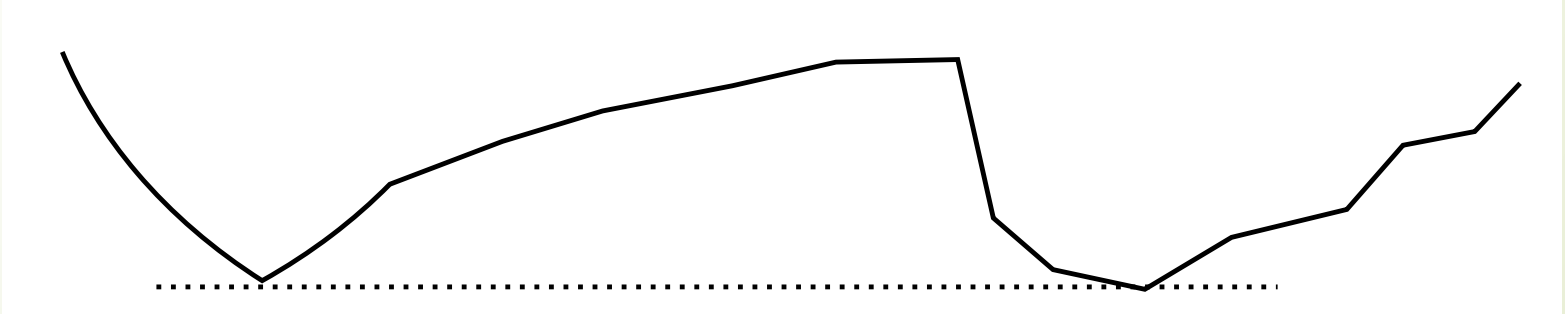
We consider two (related) properties of the optimization landscape. The first one is the following:

P.1 Given any *initial* parameter $\theta_0 \in \Theta$, there exists a continuous path $\theta : t \in [0, 1] \mapsto \theta(t) \in \Theta$ such that:

(a) $\theta(0) = \theta_0$

(b) $\theta(1) \in \arg\min_{\theta \in \Theta} L(\theta)$

(c) The function $t \in [0, 1] \mapsto L(\theta(t))$ is non-increasing.



The landscape has no spurious local valleys.

Overparameterized LN -> Single Basin

$$E(W_1, \ldots, W_K) = \mathbb{E}_{(X,Y)\sim P} \|W_K \ldots W_1 X - Y\|^2 \ .$$

Proposition: [BF'16]

1. If $n_k > \min(n, m)$, $0 < k < K$, then $N_u = 1$ for all u.

2. (2-layer case, ridge regression)
   $E(W_1, W_2) = \mathbb{E}_{(X,Y)\sim P} \|W_2 W_1 X - Y\|^2 + \lambda(\|W_1\|^2 + \|W_2\|^2)$
   satisfies $N_u = 1 \ \forall \ u$ if $n_1 > \min(n, m)$.

- We pay extra redundancy price to get simple topology.

Bruna, Freeman, 2016

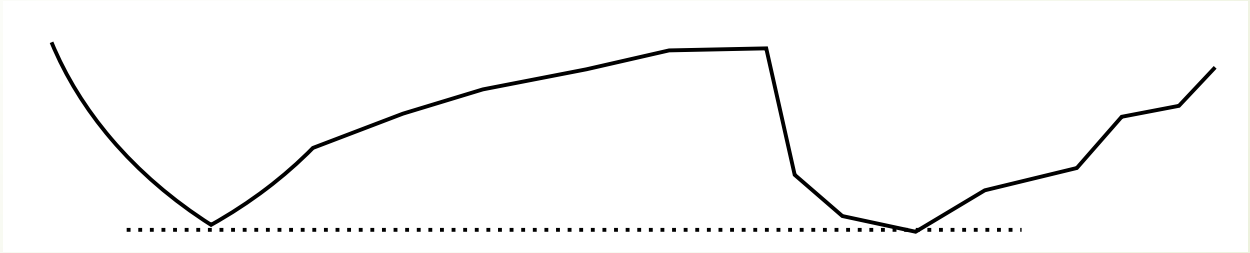
Venturi-Bandeira-Bruna'18

$$\Phi(x;\theta) = W_{K+1}\cdots W_1 x \ , \tag{13}$$

where $\theta = (W_{K+1}, W_K, \ldots, W_2, W_1) \in \mathbb{R}^{n\times p_{K+1}} \times \mathbb{R}^{p_{K+1}\times p_K} \times \cdots \mathbb{R}^{p_2\times p_1} \times \mathbb{R}^{p_1\times n}$.

Theorem 8 *For linear networks* (13) *of any depth* $K \geq 1$ *and of any layer widths* $p_k \geq 1$, $k \in [1, K+1]$, *and input-output dimensions* n, m, *the square loss function* (2) *admits no spurious valleys.*

Symmetry $f(W_i) = f(QW_i)$ $(Q \in GL(R^{n_l}))$ helps remove the network width constraint.



2-layer Neural Networks

[Venturi, Bandeira, Bruna, 2018]

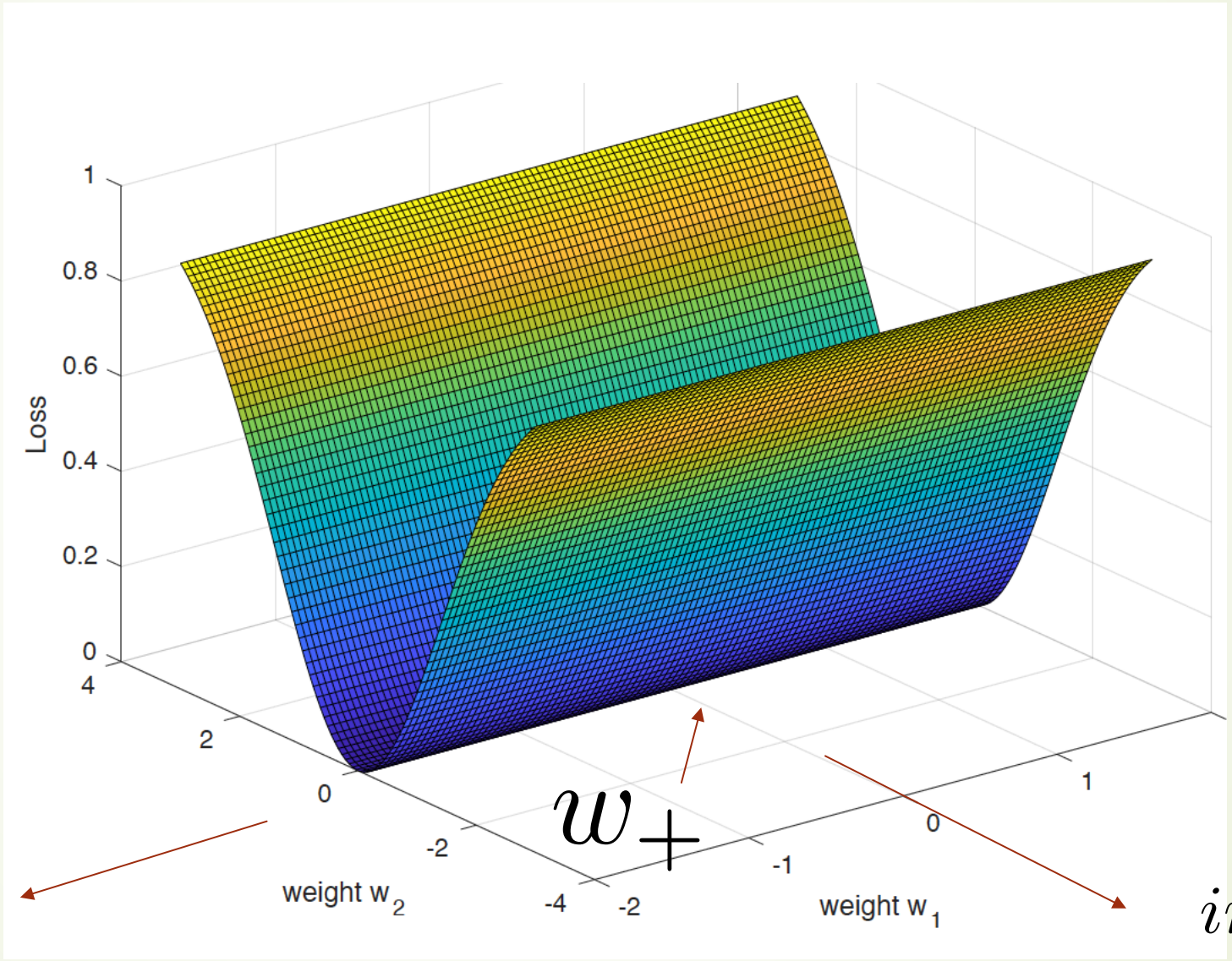
Theorem 5 *The loss function*

$$L(\theta) = \mathbb{E}\|\Phi(X;\theta) - Y\|^2$$

of any network $\Phi(x;\theta) = U\rho Wx$ *with effective intrinsic dimension* $q < \infty$ *admits no spurious valleys, in the over-parametrized regime* $p \geq q$. *Moreover, in the over-parametrized regime* $p \geq 2q$ *there is only one global valley.*

- Reproducing Kernel Hilbert Spaces (RKHS) are exploited in the proof!
- Matrix factorizations are of similar ideas.

Over-parameterized Landscapes: as p>n, equilibria are all degenerate



$ker(X)$

$im(X^T)$

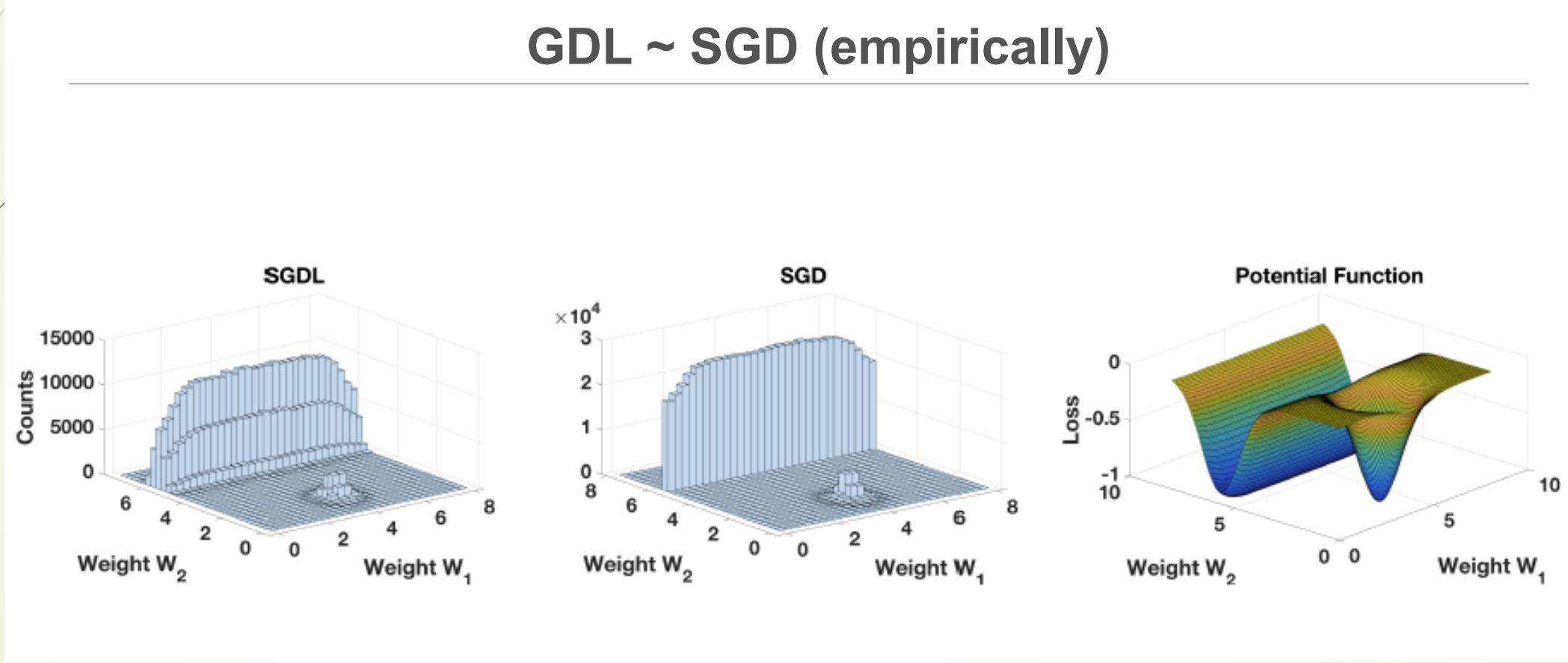
Recall: SGD behaves like Gradient Descent Langevin dynamics (GDL)

$$\frac{dw}{dt} = -\gamma_t \nabla V(w(t), z(t)) + \gamma_t' dB(t)$$

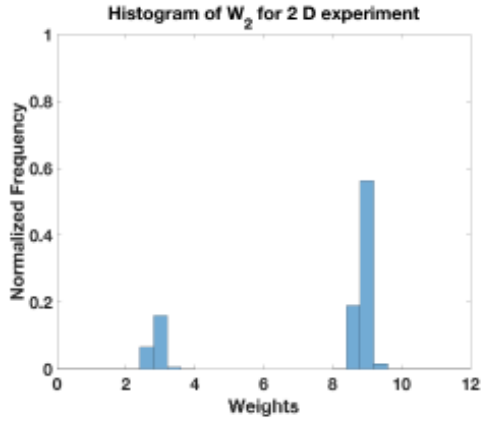
with the Boltzmann equation as asymptotic “solution”

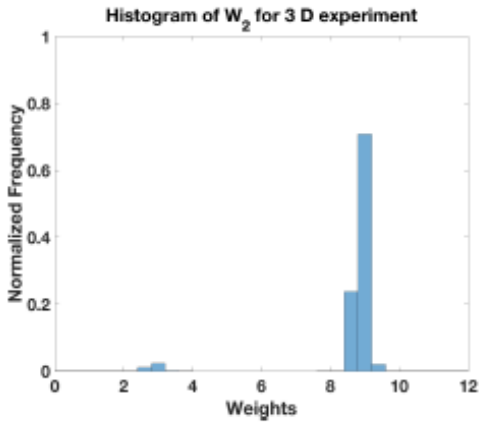
$$p(w) \sim \frac{1}{Z} = e^{-\frac{V(w)}{T}}$$

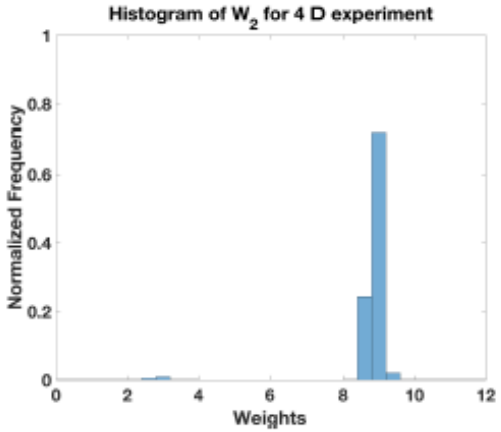
SGD/GDL selects larger volume minima e.g. degenerate

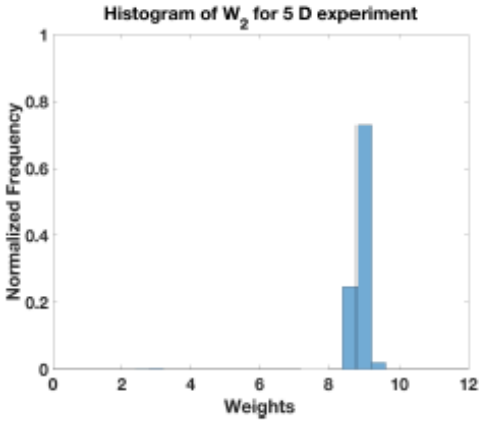


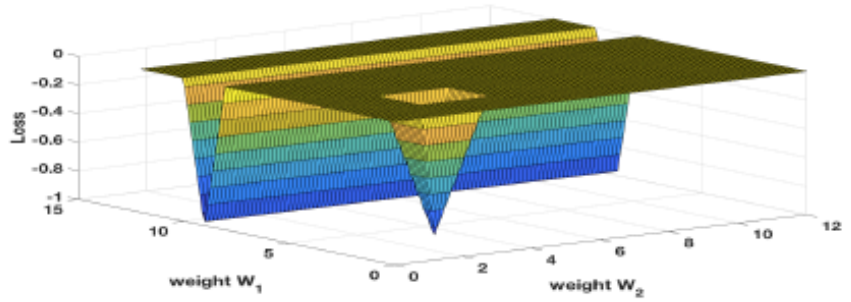
Concentration because of high dimensionality











Poggio, Rakhlin, Golovitc, Zhang, Liao, 2017

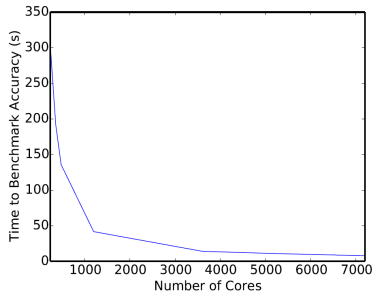
Summary

- Over-parameterization may lead to simple risk landscapes with **flat** (degenerate) global minima
- SGD tends to find **flat** global minima like Langevin Dynamics
- But SGD for multilayer neural networks suffer from vanishing gradient issue (architecture revision: ReLU, ResNet, LSTM)…

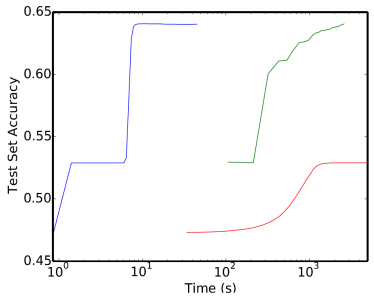
Alternative: Variable Splitting with Block Coordinate Descent

- ADMM-type: **Taylor** et al. ICML 2016
- **Proximal Propagation**, ICLR 2018
- BCD with zero Lagrangian multiplier: **Zhang** et al. NIPS 2017
- Discrete EMSA of PMP: **Qianxiao LI** et al 2017, talk on Monday in IAS workshop
  - No-vanshing gradients and parallelizable
- Global convergence can be established via Kurdyka-Lojasiewicz inequality: with **Jinshan Zeng and Tsz Kit Lau et al.**

Experiment results on Higgs dataset from Taylor et al'16

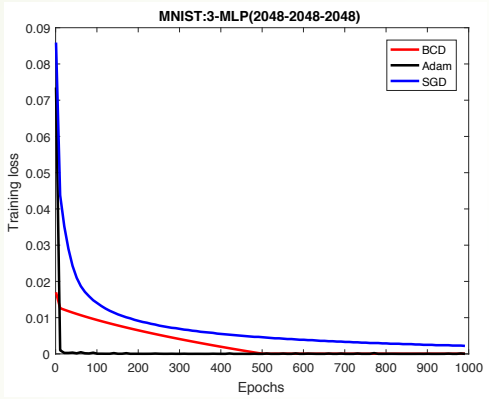


(a) **Time required for ADMM to reach 64% test accuracy when parallelized over varying levels of cores**. L-BFGS on a GPU required 181 seconds, and conjugate gradients required 44 minutes. SGD never reached 64% accuracy.

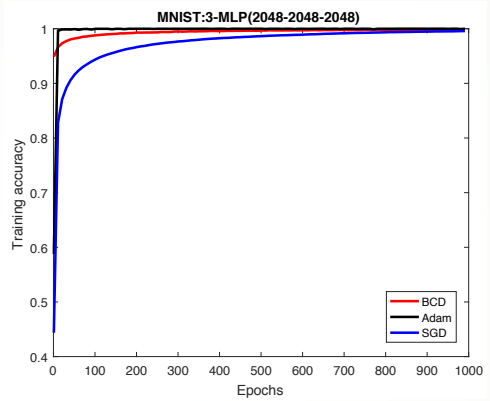


(b) **Test set predictive accuracy as a function of time** for ADMM on 7200 cores (blue), conjugate gradients (green), and SGD (red). Note the x-axis is scaled logarithmically.

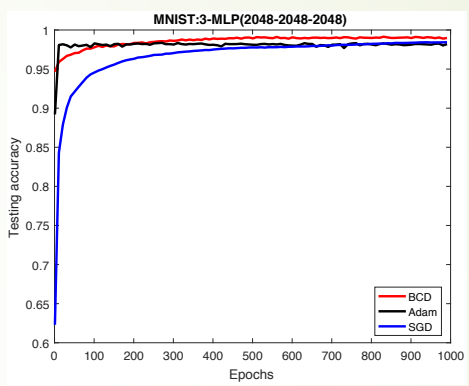
Adam, BCD, vs. SGD



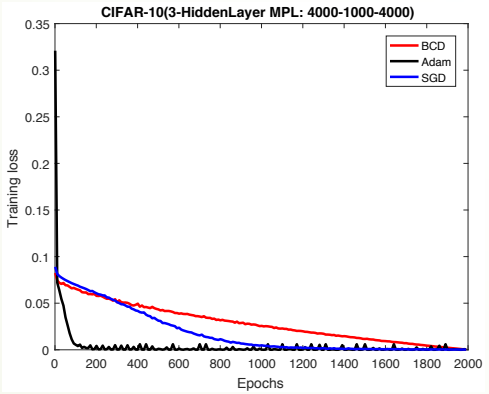
(a) MNIST (Training loss)



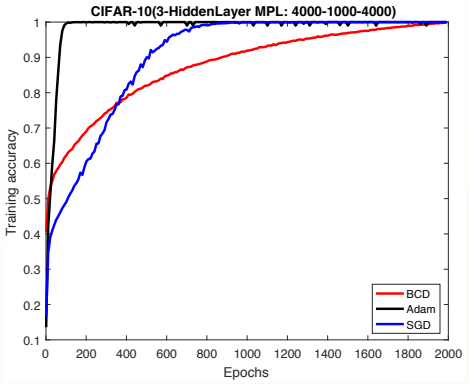
(b) MNIST (Training Acc.)



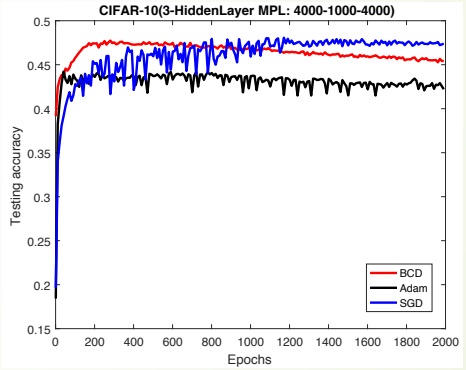
(c) MNIST (Testing Acc.)



(d) CIFAR-10 (Training loss)



(e) CIFAR-10 (Training Acc.)

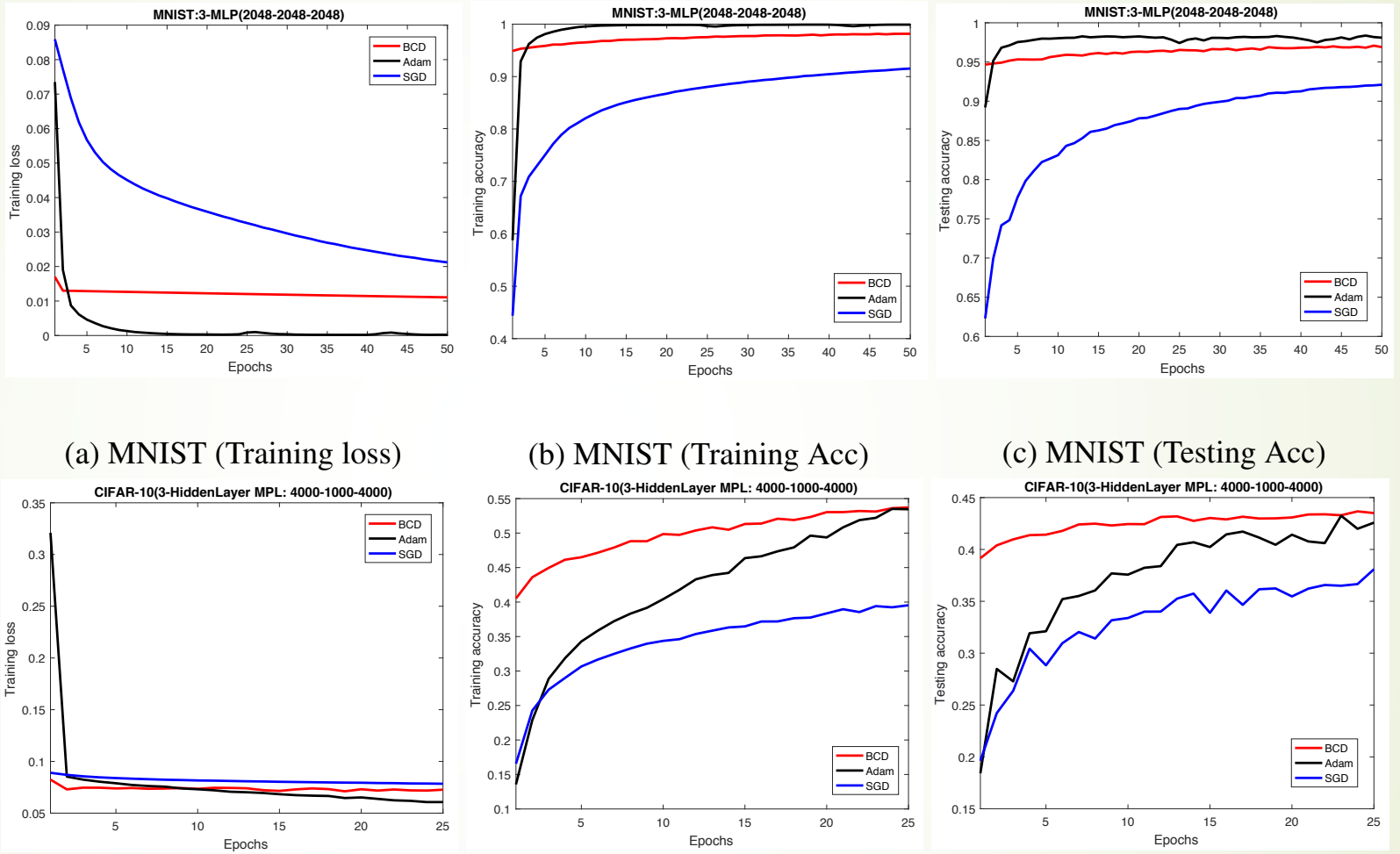


(f) CIFAR-10 (Testing Acc.)

Figure 1: Comparisons on MNIST and CIFAR-10 datasets. The first row consists of the figures including the curves of training loss, training accuracy and testing accuracy on MNIST dataset. The second row gives the associated figures on CIFAR-10 dataset.

by **Jinshan Zeng** et al.

BCD may be good initializers



(a) MNIST (Training loss) (b) MNIST (Training Acc) (c) MNIST (Testing Acc)

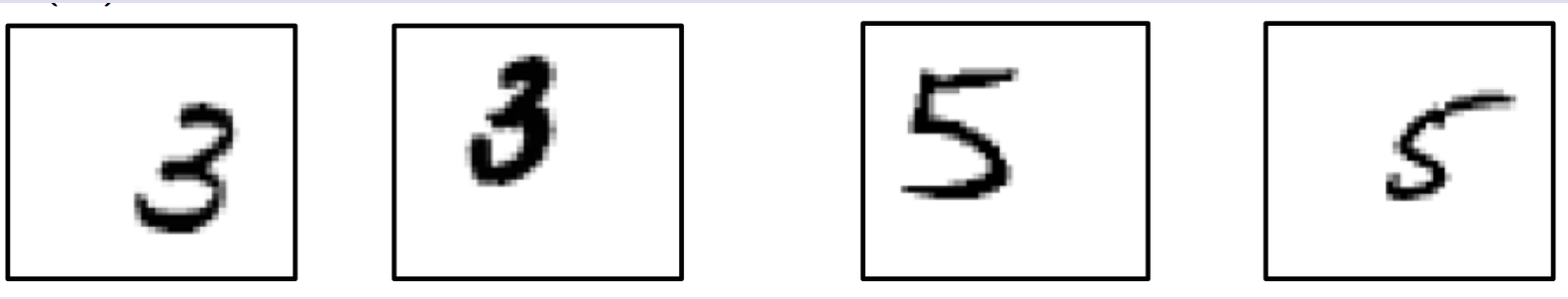
(d) CIFAR-10 (Training loss) (e) CIFAR-10 (Training Acc) (f) CIFAR-10 (Testing Acc)

Figure 2: Comparisons of the performance of three methods at the early stage.

by **Jinshan Zeng** et al.

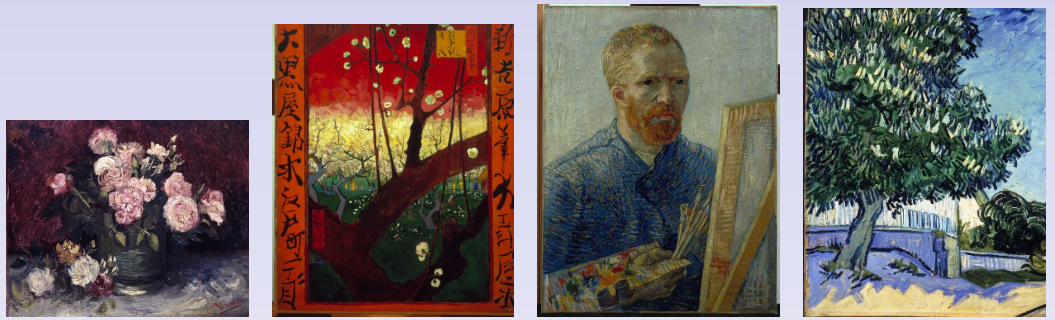
An Introduction to Wavelet Scattering Transform

MNIST Digit Classification



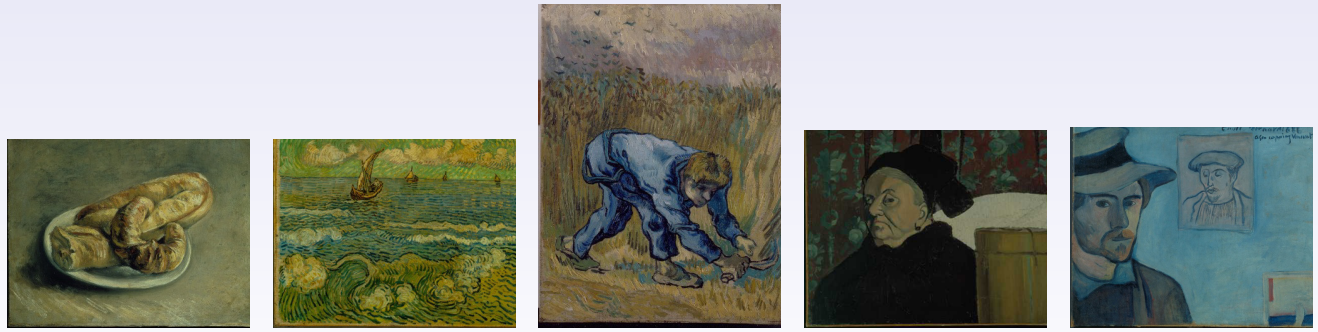
- Translation
- Deformation

van Gogh's painting vs. Forgeries



(a) f249 (b) f371 (c) f522 (d) f752

Figure: van Gogh's paintings.



(a) f253a (b) f418 (c) f687 (d) s205 (e) s206v

Figure: Forgeries.

Liu, Chan, Yao, ACHA, 2016

Van Gogh's Painting dataset

- **Liu, Chan, Yao**, Geometric Tight Frame based Stylometry for Art Authentication of van Gogh Paintings, *ACHA*, 2016
- **79** paintings authenticated by experts
- **64** genuine paintings and **15** forgeries
- Forgeries are 'quite' genuine with 6 historically wrongly attributed to van Gogh
- High-resolution professional images provided by van Gogh Museum and Kroller-Müller Museum
- Design an algorithm to determine if a painting is from van Gogh or NOT

Raphael's Painting or Not?

Image 1

Image 7

Image 10

Image 20

Image 23

Image 25

Image 26

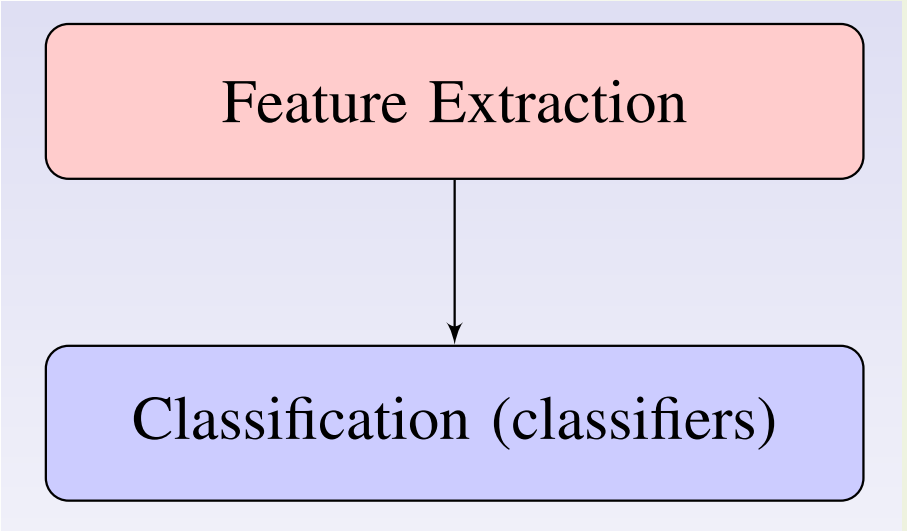
Data provided by Prof. Yang WANG

Raphael's Drawing dataset

- Dataset provided a museum in Boston via Prof. Yang WANG
- **28** digital images (.jpeg or .tiff) of drawings:
  - **12** by Raphael
  - **9** non-Raphael
  - **7** disbuted?

Image classification problem:

- Feature Extraction
  - Fourier Transform
  - **Wavelet**
  - EMD
  - Tight frame
  - Neural Networks
  - Decision trees, etc.
- Classification or Visualization
  - Logistic Regression
  - SVM
  - PCA/Manifold Learning, etc.



Wavelet Scattering Transform [Stephane Mallat et al.]

AIM: Classify correctly although translation and deformation, i.e.,

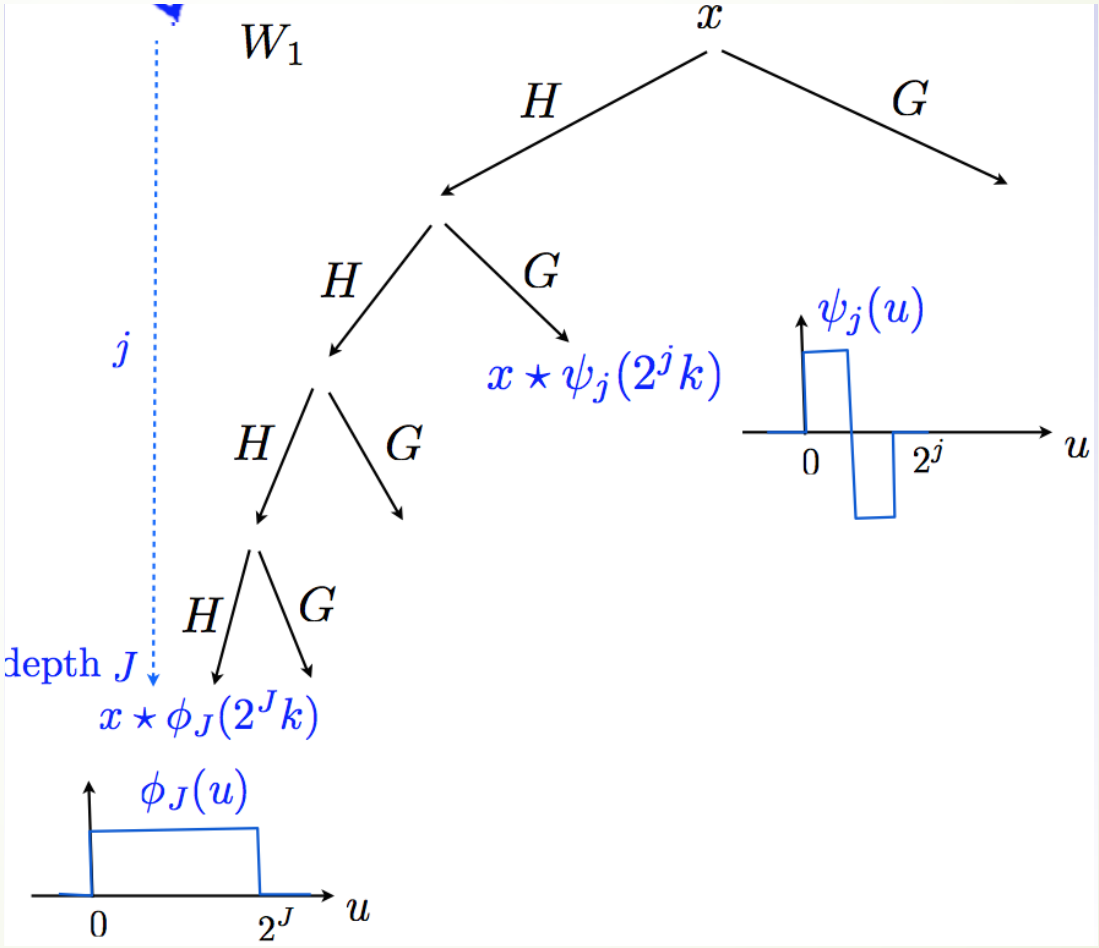
- Globally invariant to the translation group
- Locally invariant to small deformation

Wavelet Scattering Transform

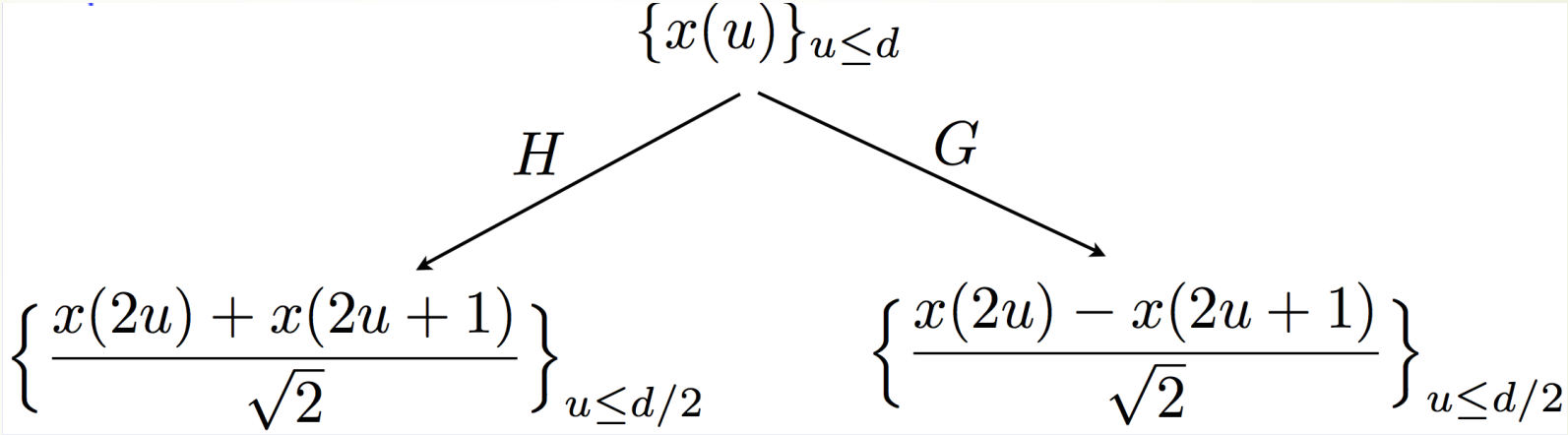
Some advantages of Wavelet Scattering Transform:

- Share hierarchical structure of DNNs
- replace data-driven filters by wavelets
- have strong theoretical support
- better performance for small-sample data

Haar Wavelet Transform



Haar Filtering



$$Hx(u) = x * h(2u) \text{ and } Gx(u) = x * g(2u)$$

where h is a low frequency and g is a high frequency.

Multiscale Wavelet Transform

wavelet filters $\{\psi_\lambda\}_\lambda$

- Dilated Wavelets: $\psi_\lambda(t) = 2^j\psi(2^j t)$ with $\lambda = 2^j$.
- Multiscale and oritented wavelet filters

$$\psi_\lambda = 2^j\psi(2^j\theta x)$$

where $\theta \in \mathcal{R}(\mathbb{R}^2)$ be a rotation matrix and $\lambda = (2^j, \theta)$.

$$x * \psi_\lambda(\omega) = \int x(u)\psi_\lambda(\omega - u) \Rightarrow \widehat{x * \psi_\lambda}(\omega) = \widehat{x} \cdot \widehat{\psi_\lambda}$$

- Wavelet transform:

$$Wx = \begin{bmatrix} x * \phi_{2^J}(t) \\ x * \psi_\lambda(t) \end{bmatrix}_{\lambda \leq 2^J}$$

Why Wavelets?

- Wavelets are uniformly **stable to deformations** (but not Fourier Transform)

$$\text{if } \psi_{\lambda,\tau}(t) = \psi_\lambda(t - \tau(t)) \text{ then}$$

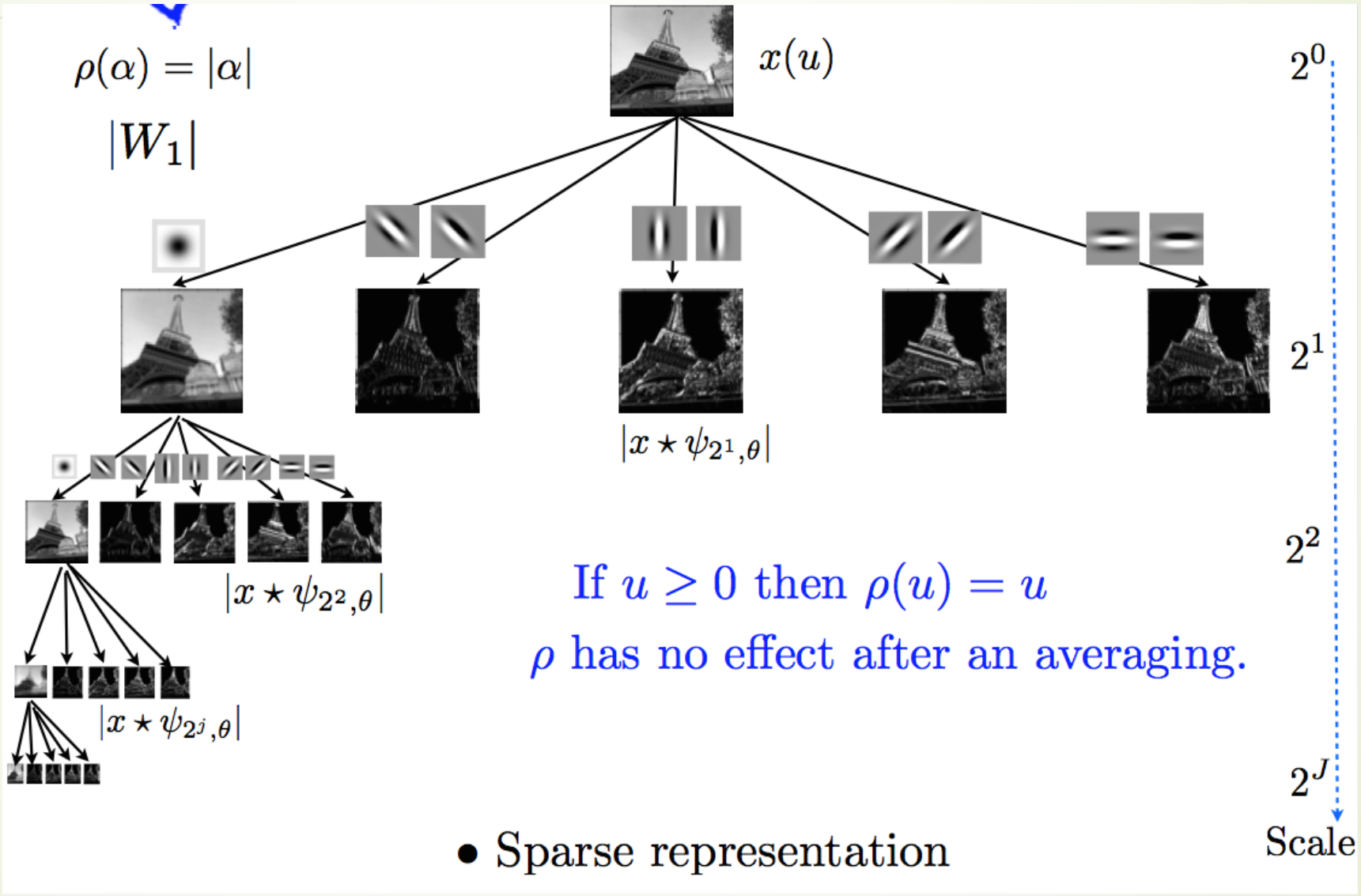
$$\|\psi_\lambda - \psi_{\lambda,\tau}\| \leq C \sup_t |\nabla\tau(t)| \ .$$

- Wavelets are **sparse** representations of functions
- Wavelets separate **multiscale** information
- Wavelets can be locally **translation invariant**

Toward scattering

- Modulus improves invariance
- Fourier transform on translated function, modulus lead to translation invariance

$$|W|x = \begin{bmatrix} x * \phi_{2^J}(t) \\ |x * \psi_\lambda(t)| \end{bmatrix}_{\lambda \leq 2^J}$$


$\rho(\alpha) = |\alpha|$
$|W_1|$
$x(u)$
2^0
2^1
2^2
2^J
$|x \star \psi_{2^1,\theta}|$
$|x \star \psi_{2^2,\theta}|$
$|x \star \psi_{2^j,\theta}|$
If $u \geq 0$ then $\rho(u) = u$
ρ has no effect after an averaging.
• Sparse representation
Scale

Multilayer Scattering coefficients

- first-layer scattering coefficients

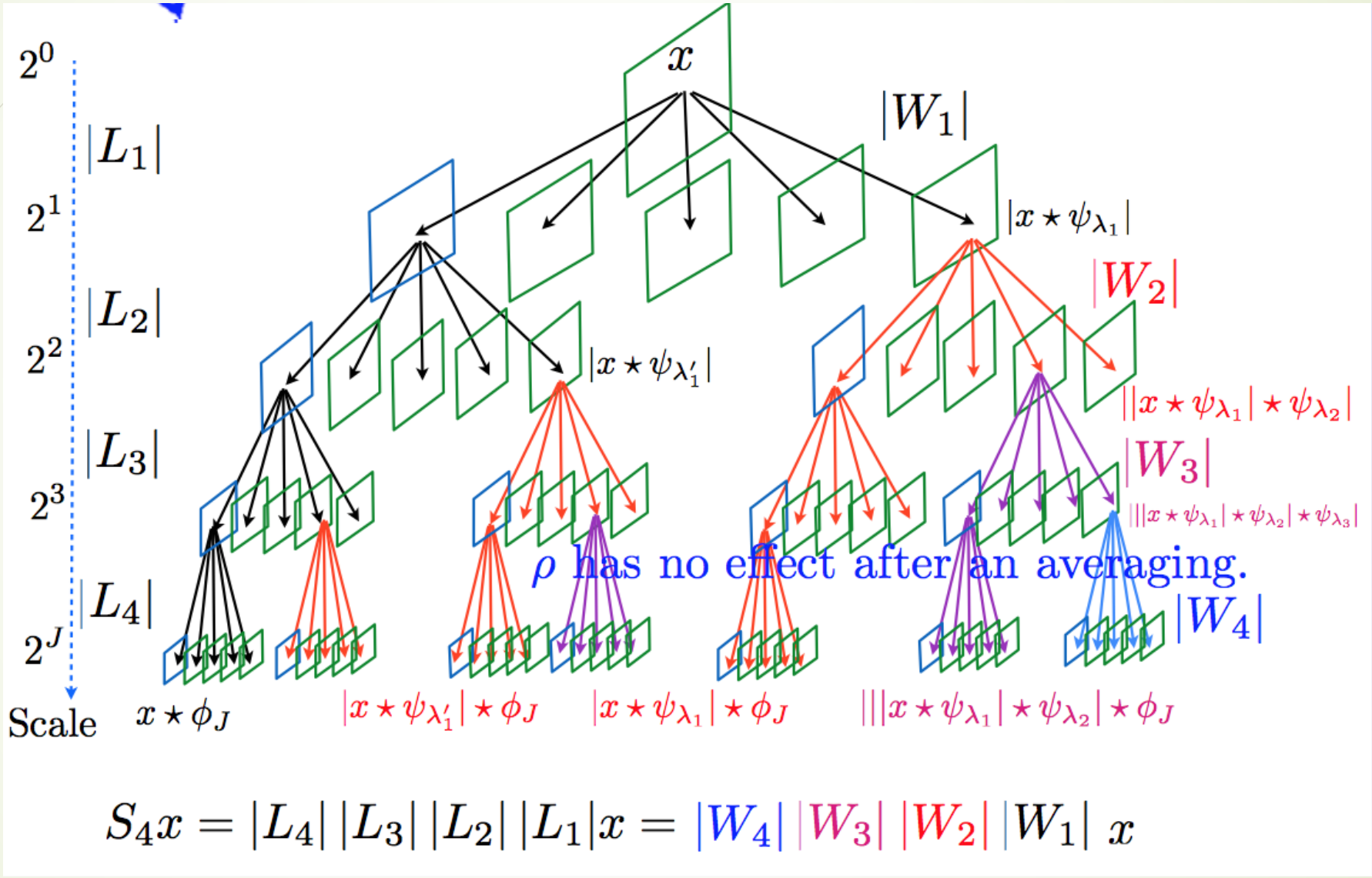
$$S_{1,J}((\lambda_1),x) = |X * \psi_{\lambda_1}| * \phi_J(x)$$

- second-layer scattering coefficients

$$S_{2,J}((\lambda_1,\lambda_2),x) = ||X * \psi_{\lambda_1}| * \psi_{\lambda_2}| * \phi_J(x)$$

- m-th layer scattering coefficients

$$S_{2,J}((\lambda_1,\lambda_2,\cdots,\lambda_m),x) = ||X * \psi_{\lambda_1}| \cdots * \psi_{\lambda_m}| * \phi_J(x)$$



$$S_4 x = |L_4|\,|L_3|\,|L_2|\,|L_1| x = |W_4|\,|W_3|\,|W_2|\,|W_1|\,x$$

Renormalization

$$\tilde{S}_{1,J}((\lambda_1)) = S_{1,J}((\lambda_1))$$

and

$$\tilde{S}_{2,J}((\lambda_1,\lambda_2)) = \frac{S_{2,J}((\lambda_1,\lambda_2))}{S_{1,J}((\lambda_1))}$$

Paper *Deep Scattering Spectrum* points out second coefficients can be decorrelated to increase their invariance through a renormalization.

Spatial invariant features based on scattering coefficients:

One choice is to take spatial averages of scattering coefficients

$$\bar{S}_{m,J} = \sum_x \tilde{S}_{m,J}((\lambda_1, \cdots, \lambda_m), x).$$

- dimension reduction
- destroy the spatial information contained in scattering coefficients

Software

- Matlab:
  - https://www.di.ens.fr/data/software/
  - ScatNet is recommended
- Python:
  - PyScatWave: https://github.com/edouardoyallon/pyscatwave