A Harmonic Analysis View of Deep Network Theory

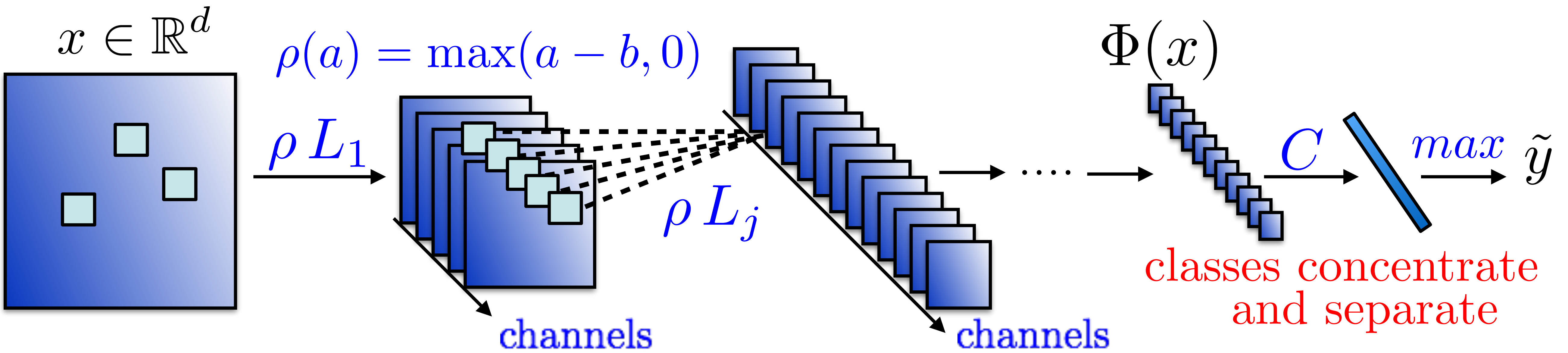
Stéphane Mallat

Flatiron Institute, CCM

Collège de France, ENS Paris

A View of Deep Network Theory

Classification with deep convolutional networks:



- Surprisingly good generalisation properties: not understood
- Issues of robustness and validation in applications: *transport, medecine, sciences...*
- Opportunity for new maths and science theories

Weekly working seminar with university collaborations *(Joan Bruna)*

Tuesdays 11am-12am EST (CCM Web page)

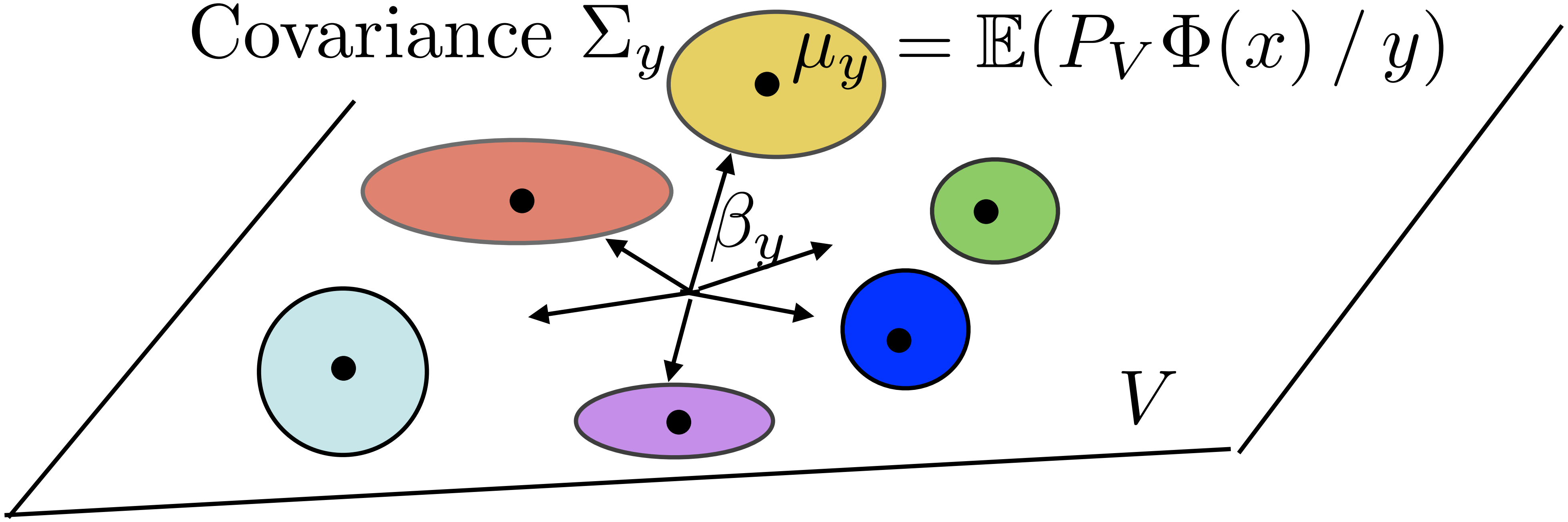
November 10th: *David Donoho* Neural Collapse

Linear Classification From $\Phi(x)$

Linear classifier: $\tilde{y} = \arg_y \max \langle \Phi(x), \beta_y \rangle + \alpha_y$

Only depends on the projection of $\Phi(x)$ on $V = Vect\{\beta_y\}_y$:

Covariance Σ_y $\mu_y = \mathbb{E}(P_V \Phi(x) \,/\, y)$



- $P_V \Phi(x)$ must have separated class means μ_y:

Fisher Ratio: $\mathrm{Trace}(\Sigma_W^{-1}\, \Sigma_B) \xrightarrow[\text{training}]{\text{Neural collapse}} \infty$

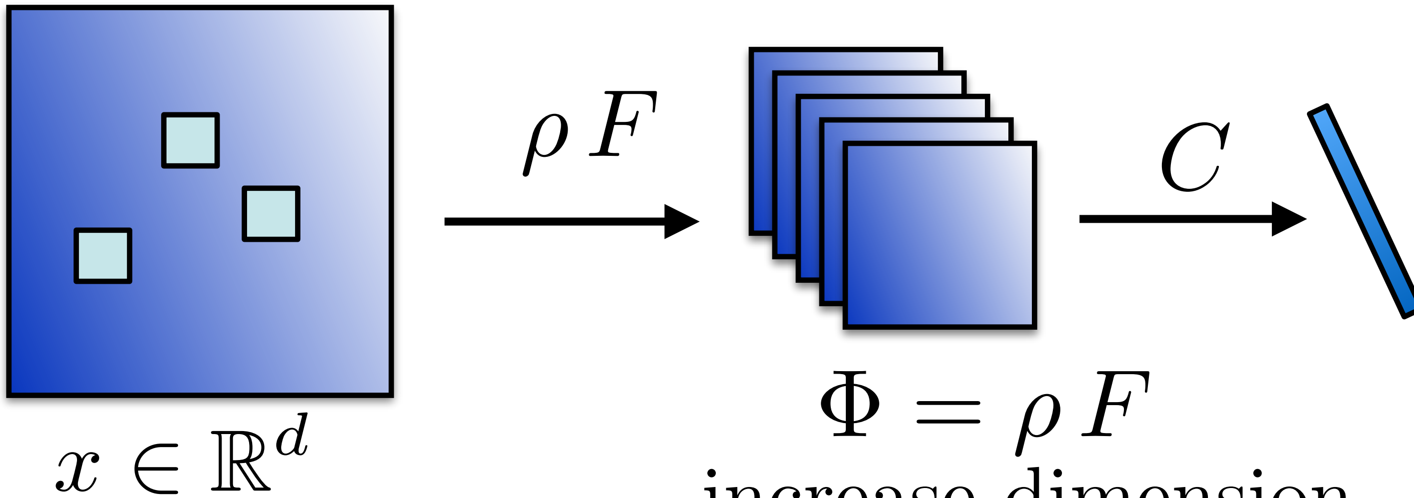
V. Papyan
X.Y. Han
D. Donoho

with $\Sigma_B = Ave_y\, (\mu_y - \overline{\mu})(\mu_y - \overline{\mu})^T$ and $\Sigma_W = Ave_y\, \Sigma_y$.

What $\Phi(x)$ achieves this concentration/separation ?

Tight Frame Contraction

John Zarka, Florentin Guth



$C\Phi(x) = \Big(\langle \Phi(x), \beta_y \rangle\Big)_y$

$S = C\,\rho\, F$: 2 layer network with no bias

Tight frame: $F^T F = Id$,

contraction: $|\rho(a) - \rho(a')| \leq |a - a'|$ "Stein shrinking estimation"

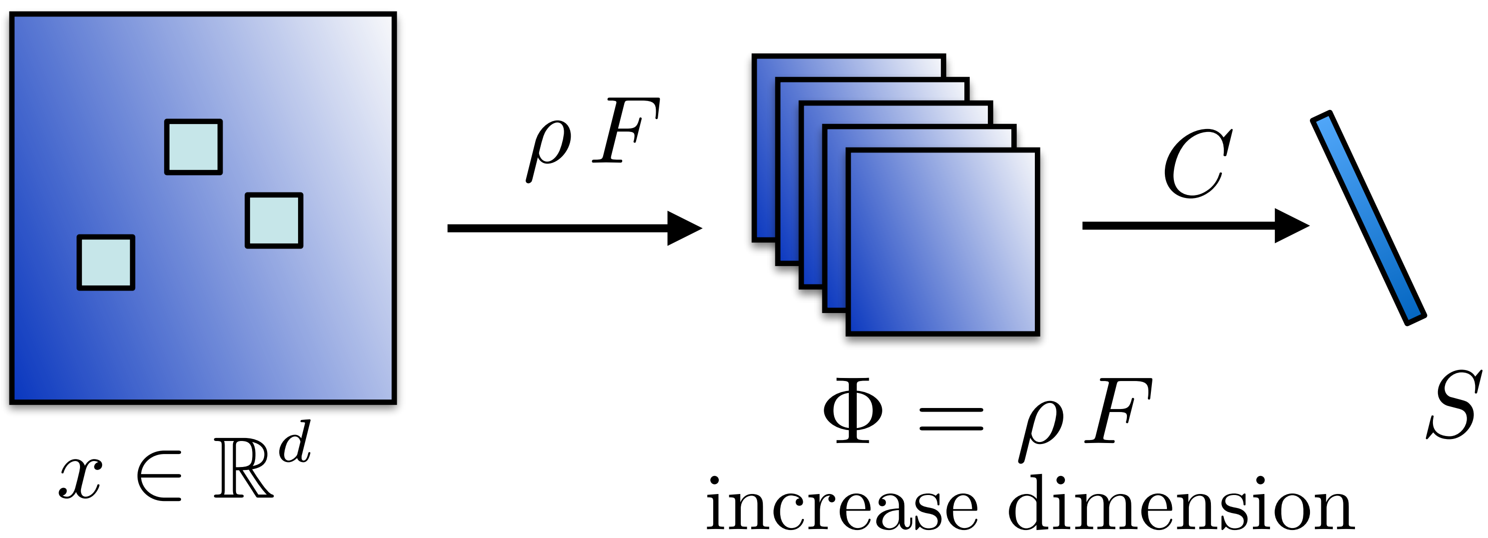
$$\Rightarrow \quad \|\Phi(x) - \Phi(x')\| \leq \|x - x'\| \ : \text{ contraction}$$

Contractions with a fixed global bias b_0:

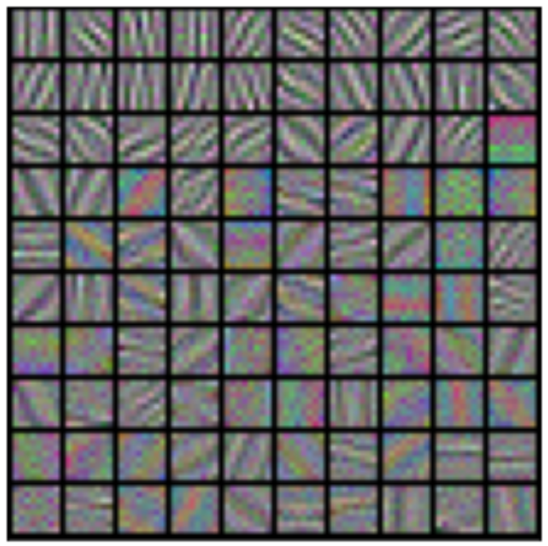
Soft-Thresh. $\rho(a) = \text{sign}(a)\, \max(|a| - b_0, 0)$ *shrinks amplitude* for noise removal

ReLu $\rho(a) = \max(a - b_0, 0)$ *shrinks amplitude and sign*

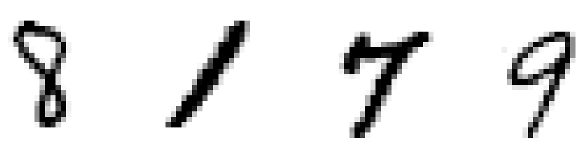
Tight Frame Contraction



Filters of F for CIFAR



- SGD optimisation

		$\Phi(x)$	x	Soft $\rho F x$	ReLu $\rho F x$
MNIST		Error Fisher	7.4% 20	1.4% 60	1.4% 60
CIFAR		Error Fisher	60% 7	39% 12	28% 15

- A soft-thresholding ρ can reduce within class variance and preserve class means μ_y if Fx is sufficiently sparse. *(Donoho Johnstone)* A ReLu ρ also modifies class means.

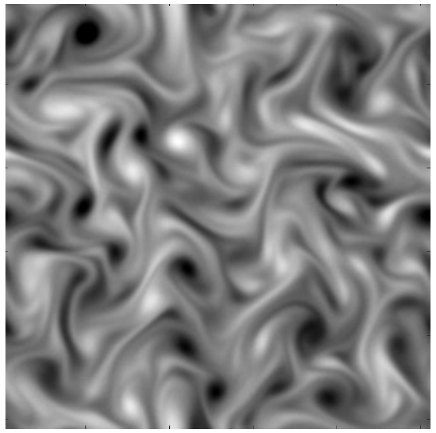
Do we need to learn the tight frame F ?

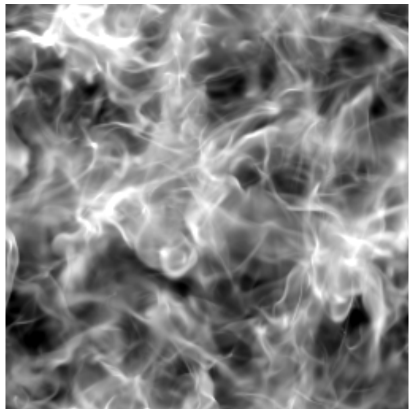
Overview

I- Concentration in Statistical Physics:

- Models of non-Gaussian processes

Turbulences:





- Wavelet separation and ReLU: scales, orientations and phases

II- Image classification by deep separation and concentration:

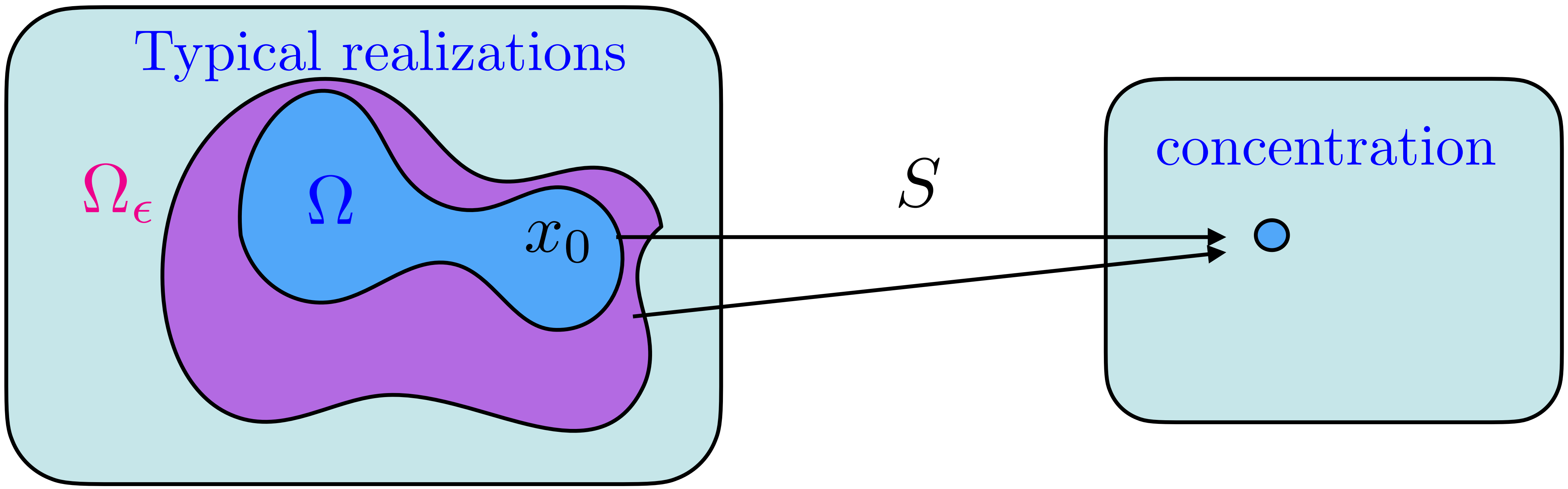
- Deep nets from priors without learning
- Learning tight frame contractions along channels only

Statistical Physics Models

Vector of statistics $S(x)$: observable

Concentration: $\text{Prob}_p\Big(\|S(x) - \mathbb{E}_p(S(x))\| > \epsilon\Big) \underset{d\to\infty}{\longrightarrow} 0$

$\Rightarrow$ a realisation x_0 satisfies $S(x_0) \approx \mathbb{E}_p(S(x))$ with high proba.



Microcanonial ensemble: $\Omega_\epsilon = \{x \ : \ \|S(x) - S(x_0)\| \leq \epsilon\}$

Maximum entropy model $\tilde{p}$ supported in Ω_ϵ is uniform.

Generation by sampling $\tilde{p}$: SGD on $\|S(x) - S(x_0)\|$ from white noise

not exactly maximum entropy *(J. Bruna)*

"Sufficient statistics" if $\Omega \approx \Omega_\epsilon$: how to define S ?

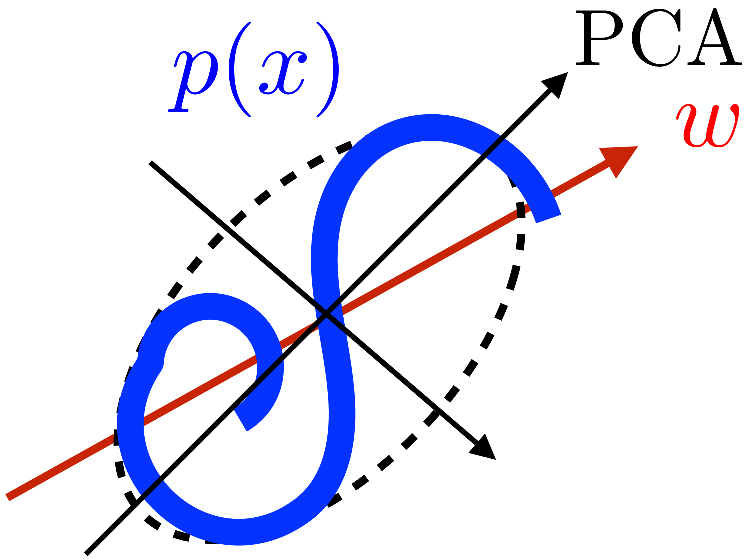
Stat. for Gaussian Stationary Models

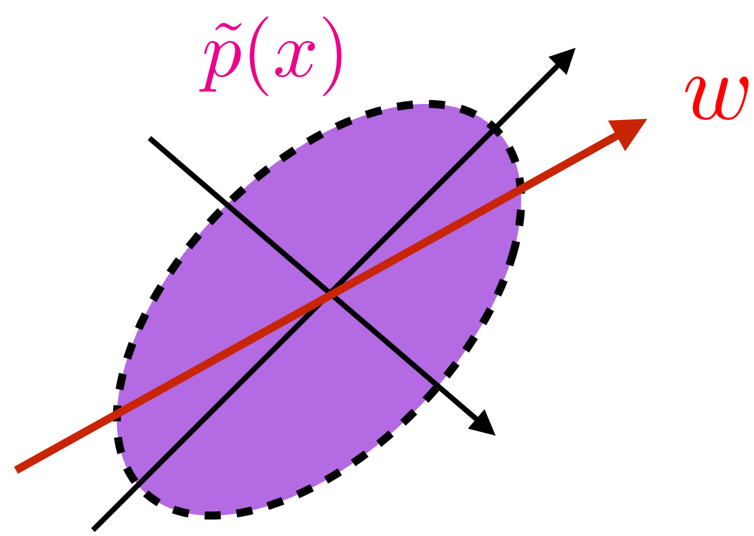
Symmetry prior: $p(x)$ is translation invariant

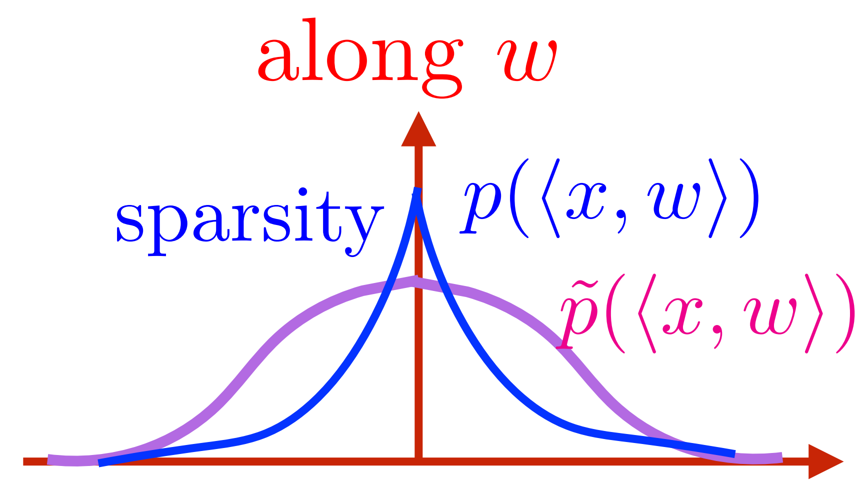
$$S(x) = \Big(d^{-1} \sum_u x(u)\, x(u-\tau) \Big)_\tau$$

empirical covariance
concentrates by spatial averaging

Maximum entropy model $\tilde p$ asymptotically Gaussian: how good ?

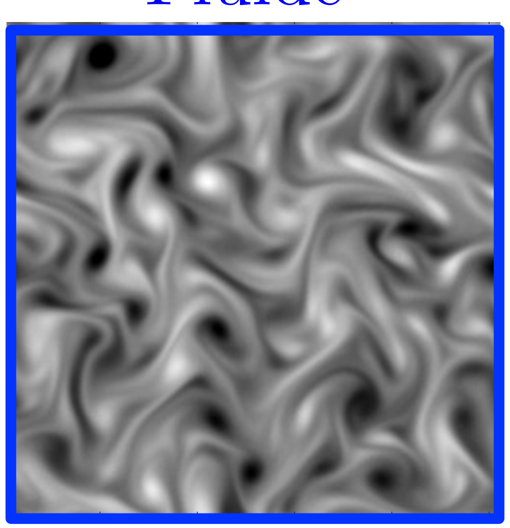




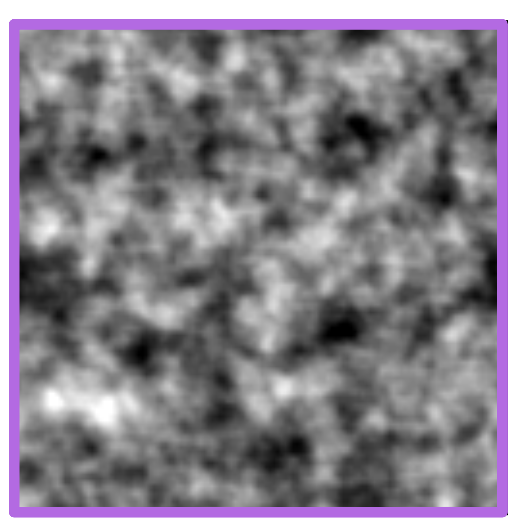


PCA basis: Fourier $\Rightarrow$ Harmonic Analysis

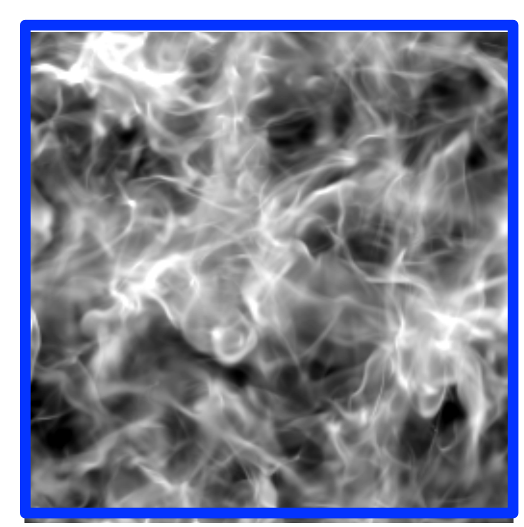
Fluide



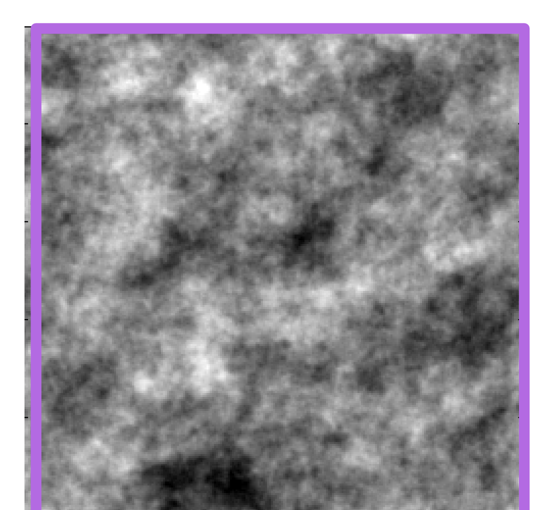
Gaussien



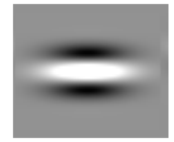
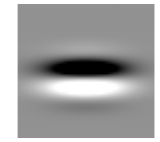
Gaz



Gaussien



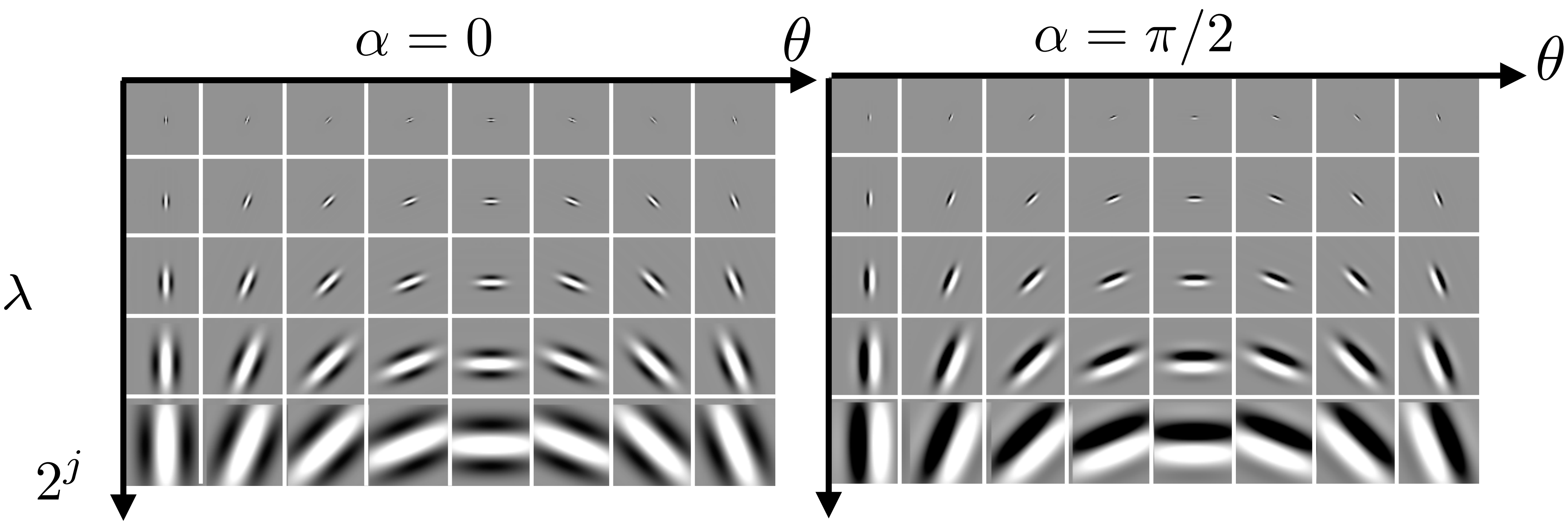
Separation with Wavelets

- Wavelet filter $\psi^\alpha(u)$:  

  phase α: 0 $\pi/2$

Scales 2^j, angles θ, phases α: $\psi_\lambda(u) = 2^{-2j}\,\psi^\alpha(2^{-j} r_\theta u)$

Wavelets ψ_λ

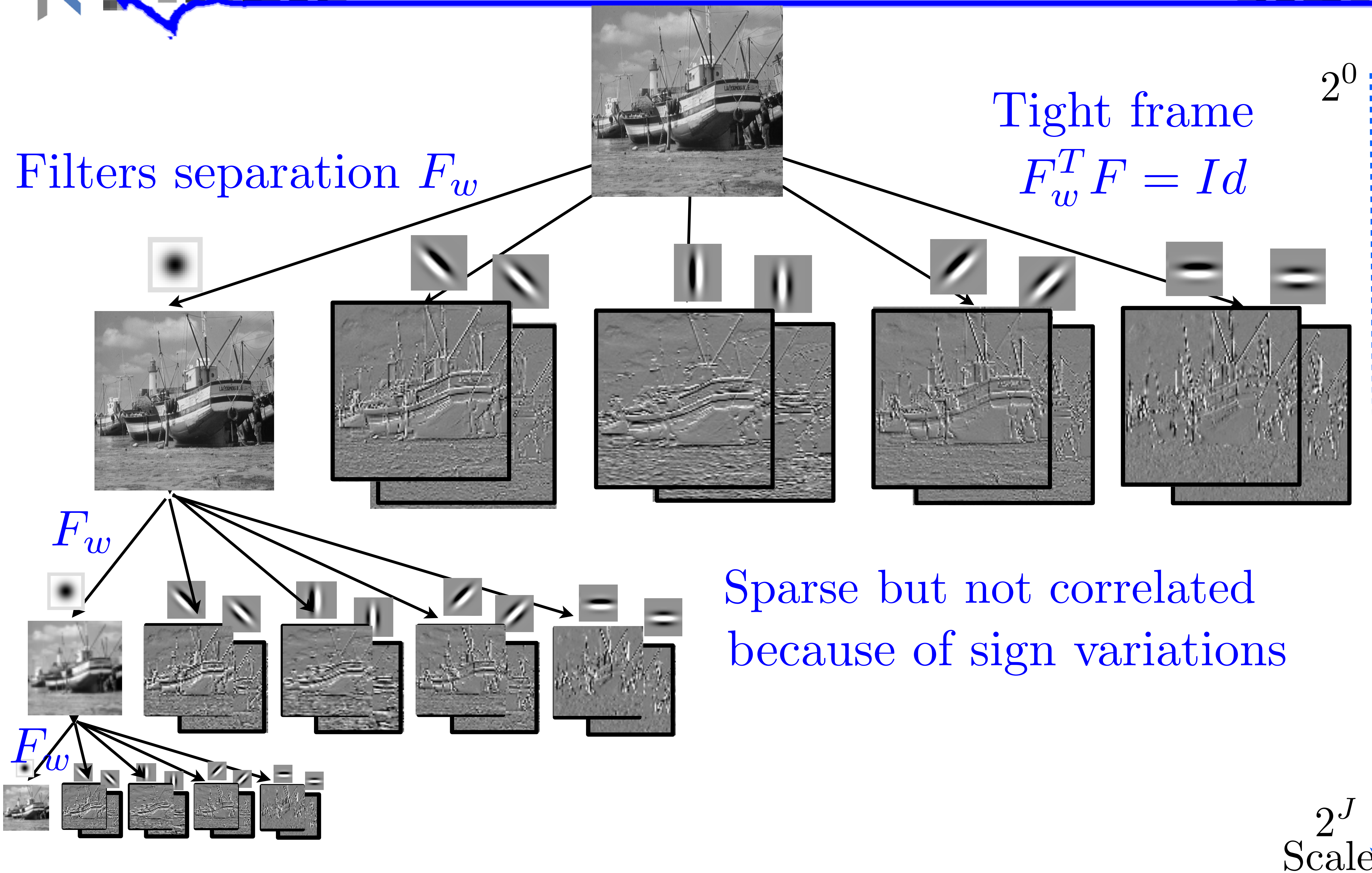


- Wavelet tight frame separation: $Wx(u,\lambda) = x \star \psi_\lambda(u)$

- Not correlated across "channels" if x is stationary:

$$\mathbb{E}\Big(Wx(u,\lambda)\,Wx(u,\lambda')\Big) \approx 0 \ \text{ if } \ \lambda \neq \lambda'$$

Wavelet Filter Bank Algorithm

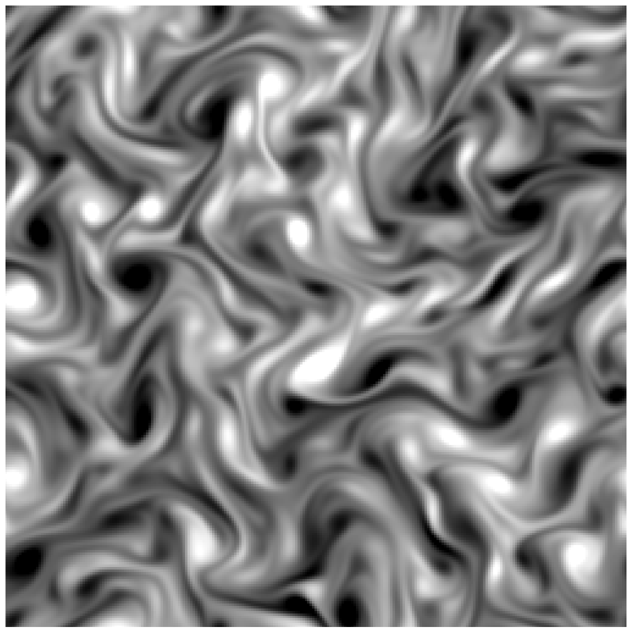


How to capture dependance across scales, angles, phases channels ?

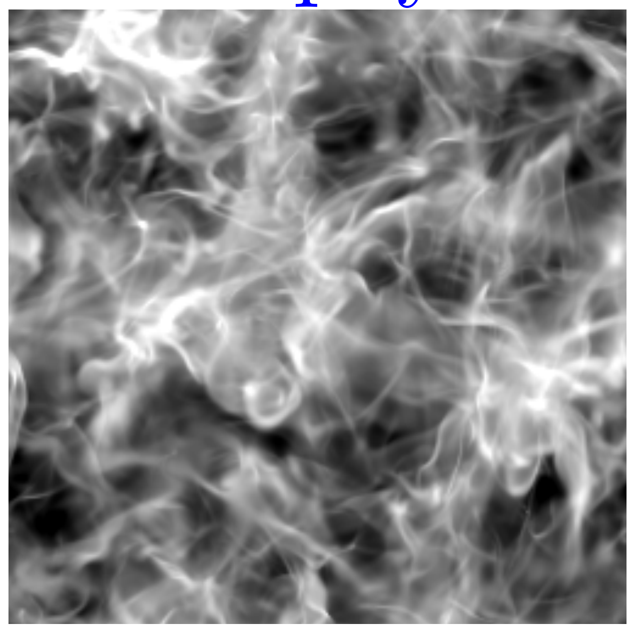
Models of Stationary Processes

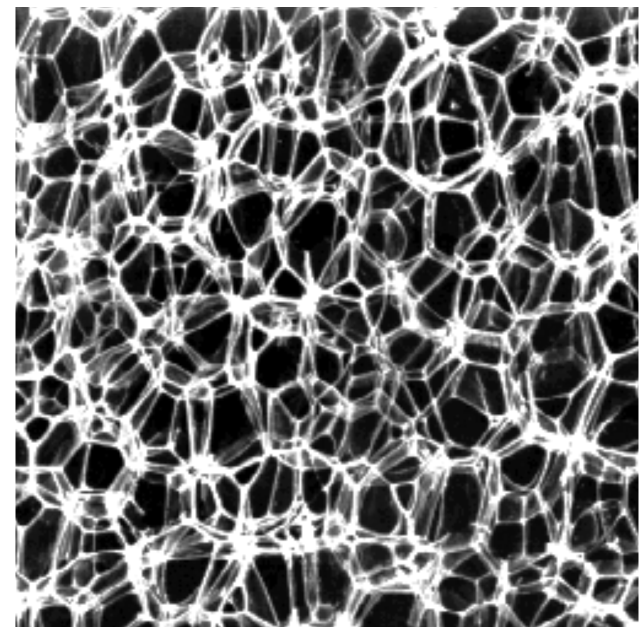
Sixin Zhang

x_0

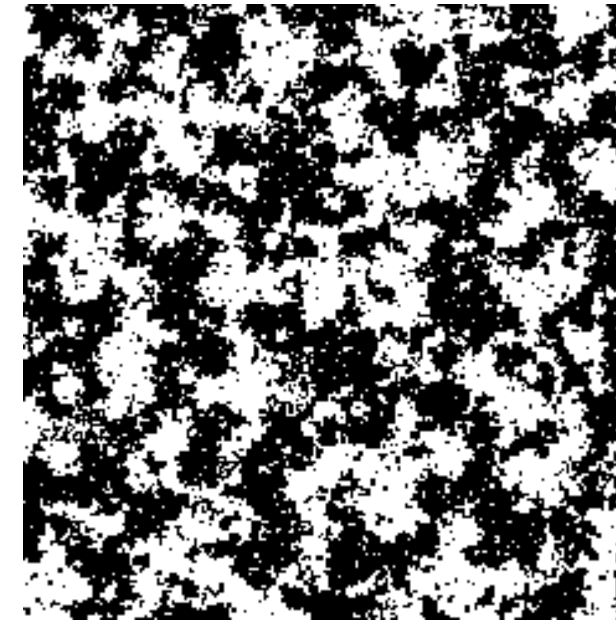


Astrophysics





Ising-critical



Correlations across scales/orientations/phases $\lambda = (2^j, \theta, \alpha)$

are created by a ReLu $\rho(a) = \max(a, 0)$ (no bias) which shrinks sign:

$$S(x) = d^{-1} \sum_u \rho W x(u)\, \rho W x(u)^T \quad : \text{ empirical correlation}$$

Concentration by spatial averaging: dimension $O(\log^2 d)$

Maximum entropy models conditioned by $S(x_0)$

Sampling from Max Entropy Model

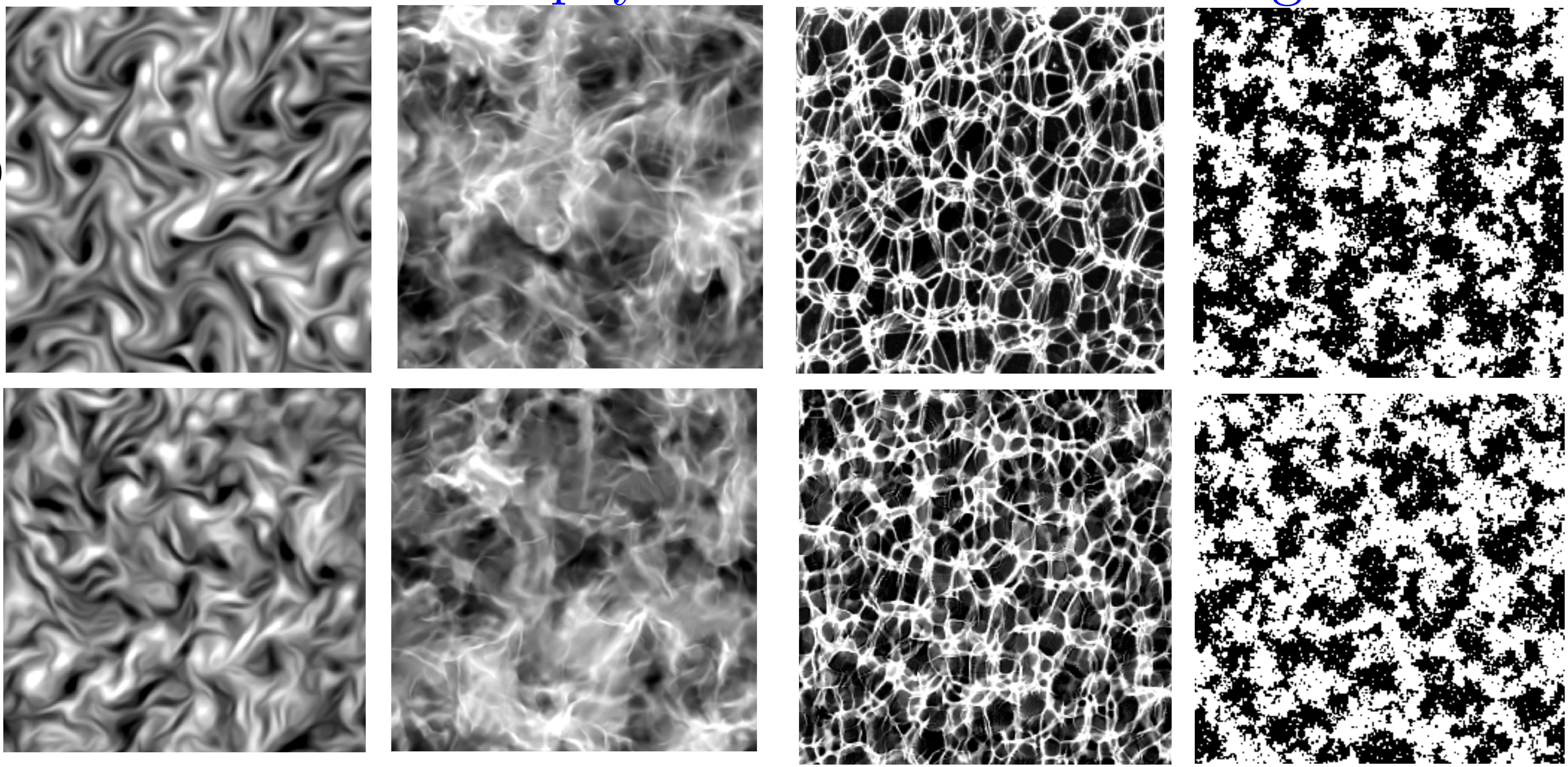
Sixin Zhang

$d = 6\,10^4$

Astrophysics   Ising-critical

x_0

x



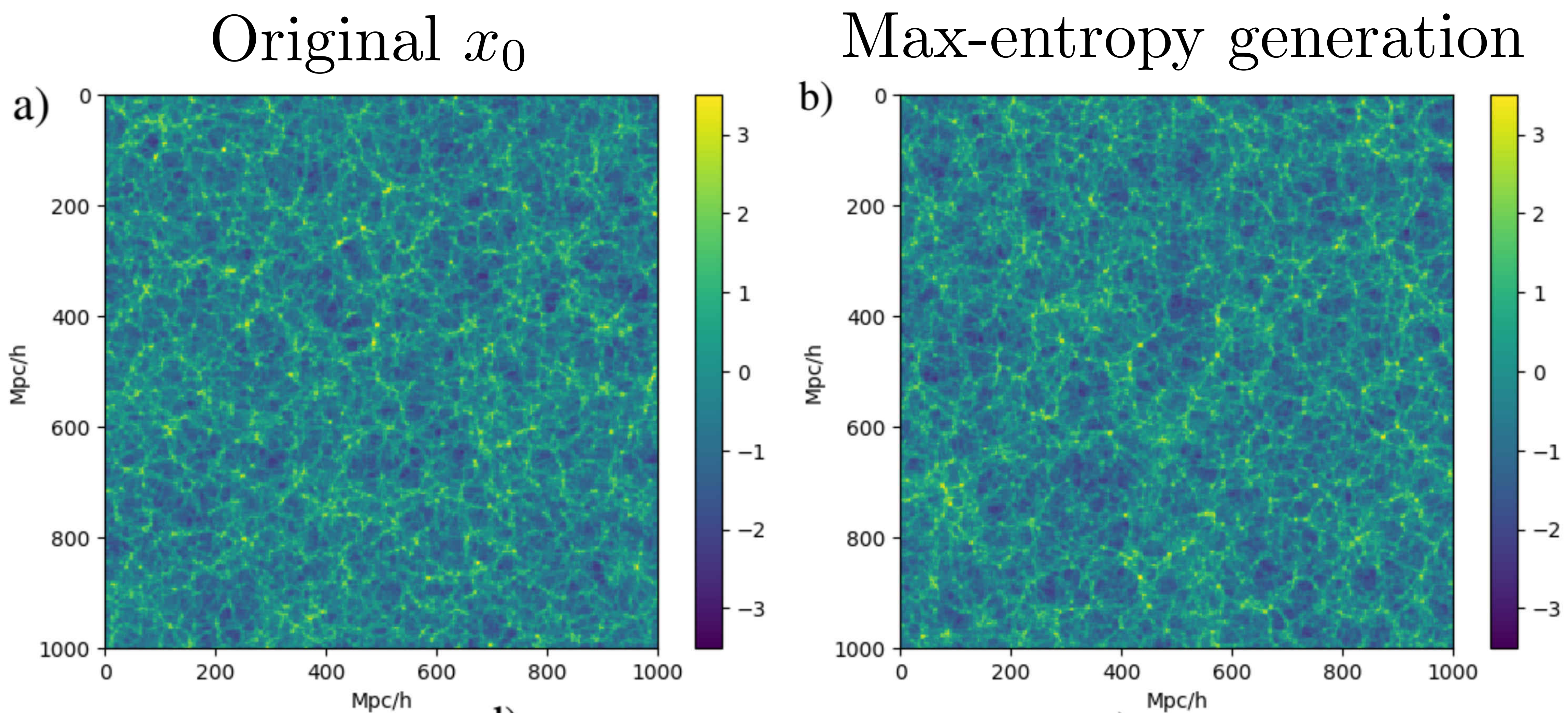
$S(x_0)$ has $2\,10^3$ empirical covariances

Sampled from $S(x_0)$ with SGD algorithm

Generation of Cosmological Models

E. Allys, T. Marchand, J.F. Cardoso, F. Villaescusa, S. Ho, S. Mallat

Generation of matter density fields from rectified wavelet covariances:



- Reproduces high order moments
- Accurate regression of 6 cosmological parameters from $S(x_0)$

II - Image Classification

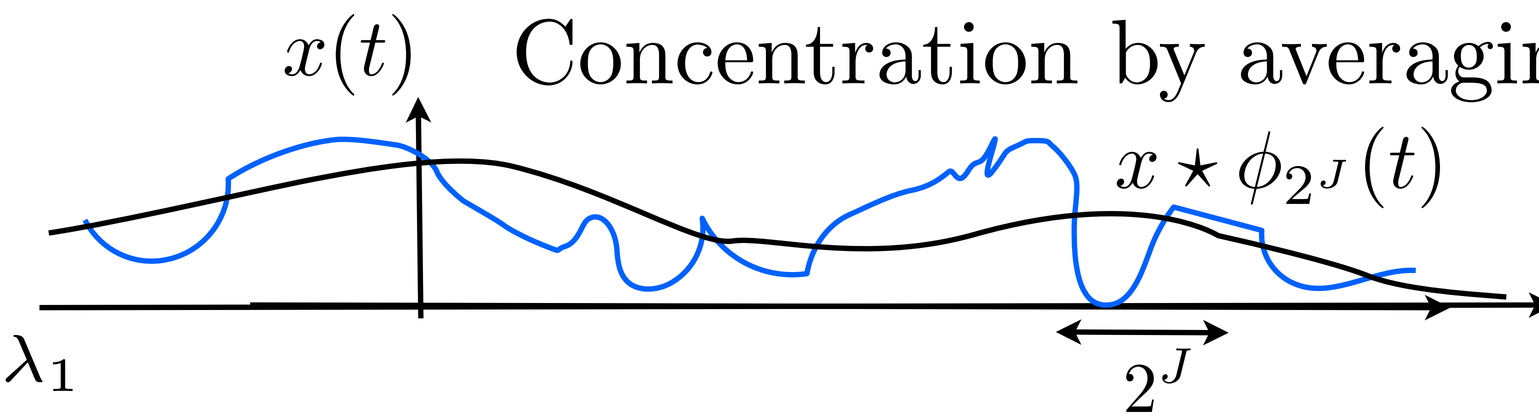
- A deep network progressively separates and concentrates
    - Can we do it from prior without learning ?
    - If not, what needs to be learned ?

Concentration and Scale Separation

Wavelet separation:

$$\rho\, \boldsymbol{W}_1 x = \begin{pmatrix} x \star \phi_{2^J} \\ \rho(x \star \psi_{\lambda_1}) \end{pmatrix}_{\lambda_1}$$

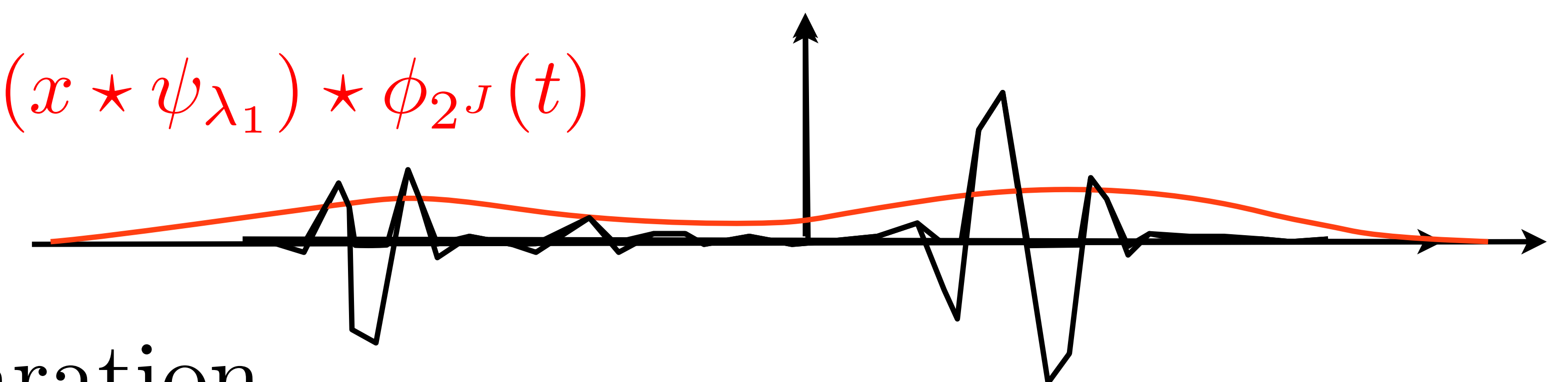
Concentration by averaging



Lost high frequencies: $x \star \psi_{\lambda_1}(t)$ but $(x \star \psi_{\lambda_1}) \star \phi_J = 0$

Relu non-linearity: $\rho(x \star \psi_{\lambda_1}(t))$ to remove sign

Concentration: $\rho(x \star \psi_{\lambda_1}) \star \phi_{2^J}(t)$



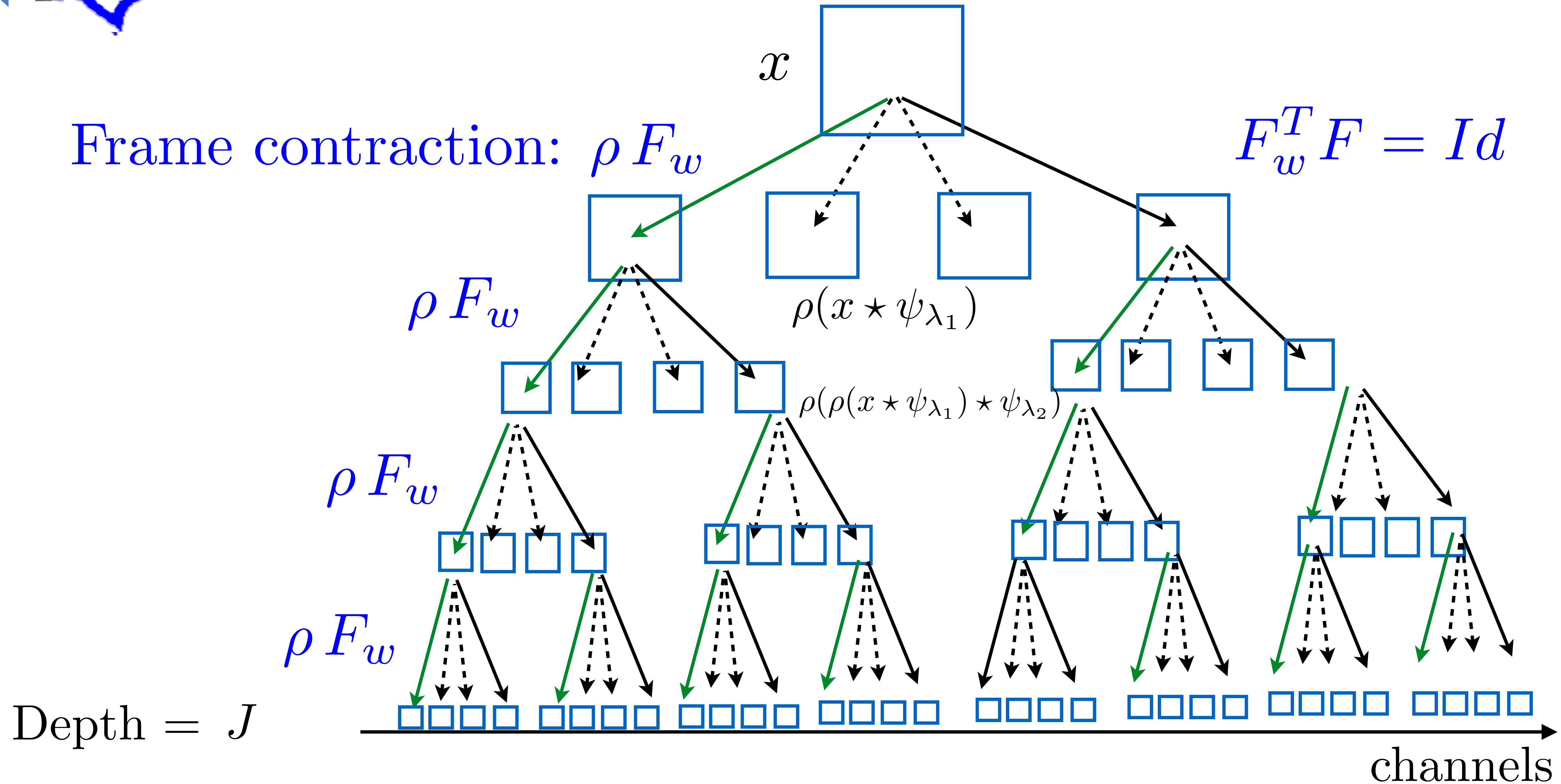
To preserve separation

Need to recover lost high frequencies: $\rho(x \star \psi_{\lambda_1}) \star \psi_{\lambda_2}(t)$

Concentration with Relu and averaging:

$$\begin{pmatrix} \rho(x \star \psi_{\lambda_1}) \star \phi_{2^J}(t) \\ \rho(\rho(x \star \psi_{\lambda_1}) \star \psi_{\lambda_2}) \star \phi_{2^J}(t) \end{pmatrix}_{\lambda_2}$$

Wavelet Scattering Network



$\Phi = (\rho\, F_w)^J$: iterated frame contractions

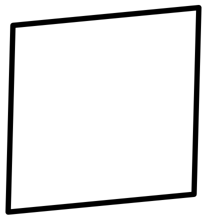
Scatters along progressively more channels

A convolution tree: no channel interactions, no learning.

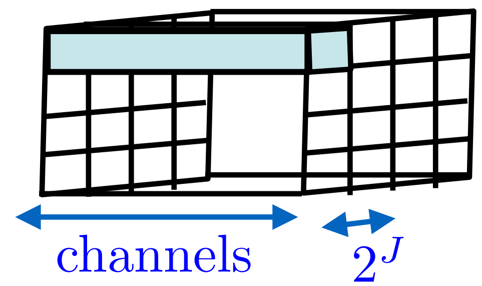
Image Classification

Scat-Net$_J$: J layers

$$x \rightarrow \boxed{\rho(\rho(x \star \psi_\lambda) \star \psi_{\lambda'}) \star \phi_J} \rightarrow \Phi(x) \rightarrow \boxed{C} \rightarrow \tilde{y}$$



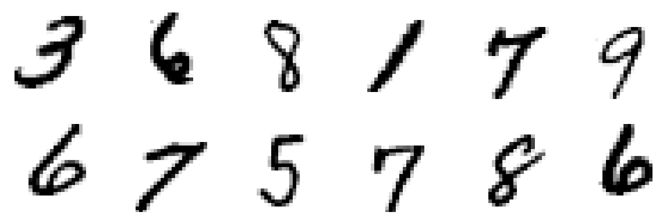
no learning



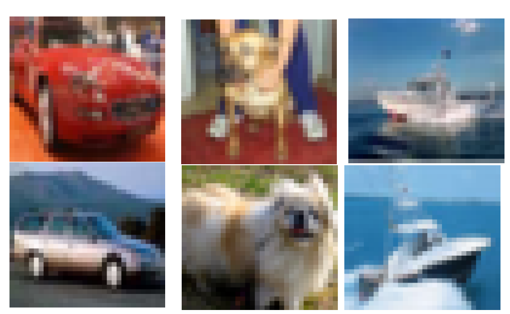
supervised learning

Errors:

MNIST: 28^2
10 classes



CIFAR: 32^2
10 classes



ImageNet: 228^2
10^3 classes
1 million training

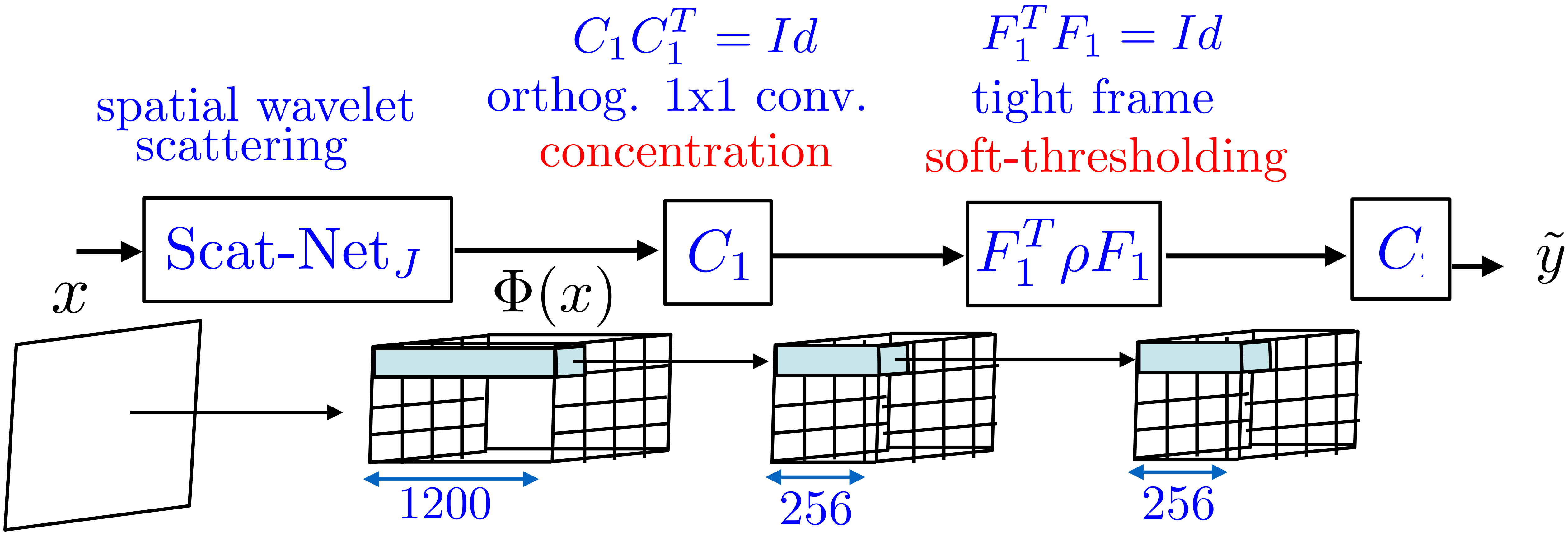
	Scattering	Deep Nets.
$J = 3$	0.5 %	0.5 %
$J = 4$	23%	ResNet-18: 8% ResNet-50: 7.6%
$J = 6$	52 %	AlexNet-7: 20% ResNet-18: 11% Res-Net 50: 7%

What is learned ?

One Concentrated Scattering

John Zarka, Florentin Guth

Frame soft-thresholding along scattering channels:



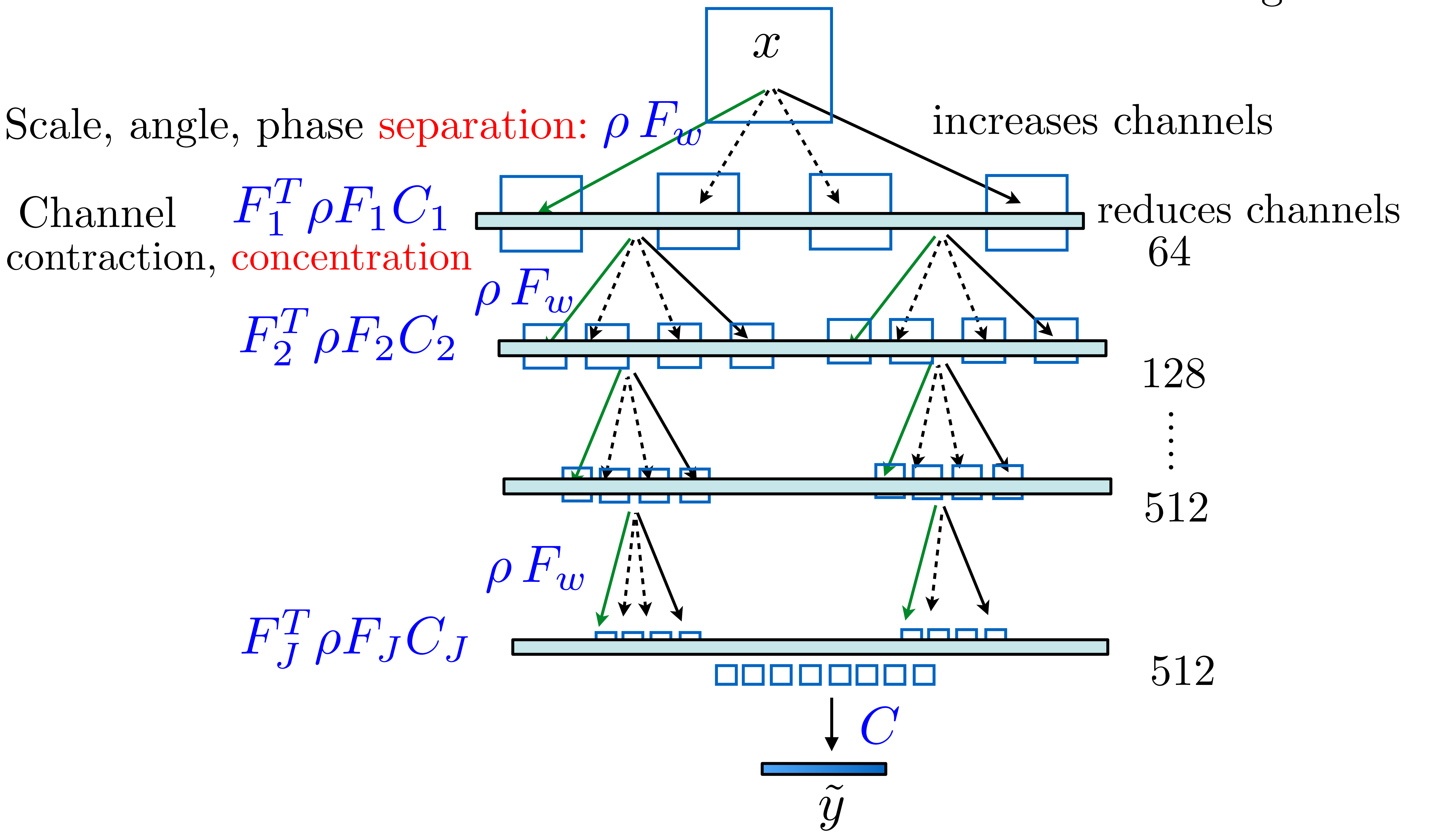
- SGD optimisation

	$\Phi(x)$	Scat.	1CoScat	ResNet-18
CIFAR	Error	27%	18%	8%
	Fisher	22	30	
ImageNet Top 5	Error	60%	30%	11%
	Fisher	2.9	3.4	

Multiscale Concentrated Scattering

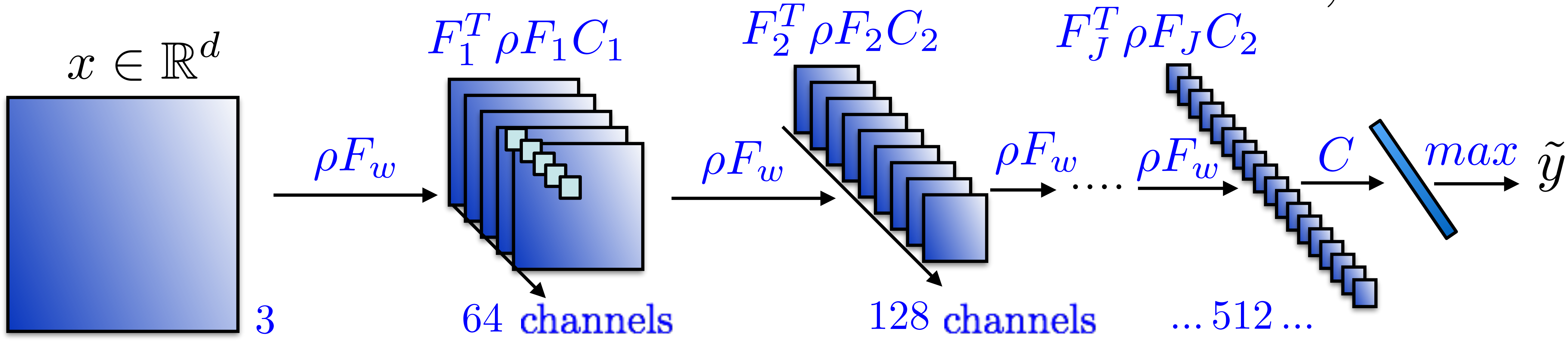
Wavelet frame contraction: $\rho \, F_w$ spatial conv, shrinks sign

Concentrated frame contraction: $F_j^T \rho F_j \, C_j$ shrinks amplitude
1x1 conv. along channels



Concentrated Scattering

J. Zarka, F. Guth



- Network without learning bias and semi-orthogonal operators
- Learning 1x1 convolutions across scattering channels
- SGD optimisation

	$\Phi(x)$	1CoScat	CoScat	ResNet-18
CIFAR	Error	18%	7.8%	8%
	Fisher	30	70	
	Depth	5	8	18
ImageNet Top 5	Error	30%	13%	11%
	Fisher	3.4	7.2	
	Depth	7	12	18

Mathematical control of Fisher ratios ?

Conclusion

- **Deep network separate and concentrate: what mechanism ?**

- Variance can be reduced with tight frame contractions
- Spatial filtering can be handled with wavelet frame which separate scale, angle and phase channels.
- Learning contractions along channels can reach ResNet accuracy

- Control of ***Fisher ratios*** is an open math. problem.

New Interpretable Statistics for Large Scale Structure Analysis and Generation

Allys, Marchand, Cardoso, Villaescusa, Ho, Mallat, arXiv:2006.06298, Phys. Rev.

Tight Frame Contractions in Deep Networks

J. Zarka, F. Guth,, S. Mallat, OpenReview, ICLR 2021