Discrimination with Deformation for Classification

Stéphane Mallat
Centre de Mathématiques Appliquées
Ecole Polytechnique

LVA, September 2010

Classification and Non-Linear LVA

- Signals are deformed by unknown non-linear operators that carry information:
  - velocity in video, 3D shapes in stereo or shape from textures...
  - writing style, voice gender in speech, musical interpretation...

- Deformations: latent «operators»

- Classification requires estimating the amplitude of deformations

- Estimating deformations often comes with it

Classification Distance

- Major classification difficulty: find a metric to compare signals

$$\text{if } (f,g) \in \mathcal{C}^2 \text{ then } d(f,g) \text{ should be small}$$

$$\text{if } f \in \mathcal{C} \text{ and } g \in \mathcal{C}' \text{ then } d(f,g) \text{ should be big}$$

- Supervised and non-supervised classifications assign classes by minimizing distances.

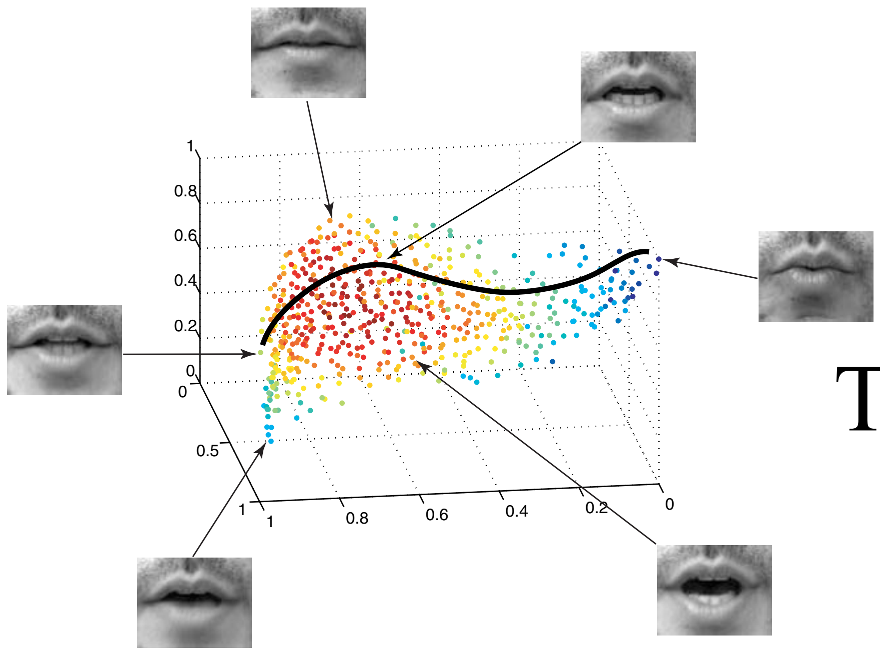
- What class of distances, how does it relate to deformations ?

- How to learn an optimized distance from training data ?

Low-Dimensional Framework

- For signals $f \in \mathbf{R}^N$ on a smooth low-dimensional manifold: can use geodesic distances.

MNIST digit data basis



Current Opinion in Neurobiology

Templates

- Deformable amplitude: length of the geodesic.

- Need to find the manifold and/or the geodesics from training data.

High Dimensional Complex Signals

- Most complex signals (audio, images...) do not belong to low dimensional manifolds:

Texture
Discrimination

Patterns
include textures

- Deformable template models do not apply.
- Not enough training data to estimate large dimensional manifolds: requires prior dimensional reduction.

Dimensionality Reduction

- In computer vision: dimensionality reduction and metrics are related to invariants to translation, rotation, scaling...
  - Histograms of wavelet coefficients: SIFT, bags of features
  - Deep learning neural networks.
- Works very well but not well understood.

- **How to build invariants from high frequencies, and measure deformations for discrimination ?**

- **Invariants in quantum physics:** specify the Lagrangian and the particle interactions in quantum field theory.

  Configurations evolve along multiple paths: not just one path as in classical mechanics.

Classification Metric Wish List

- Classification with $\mathbf{L^2}$ norm on a representation Φ:

$$d(f,g) = \|\Phi(f) - \Phi(g)\| \ .$$

- Classes are invariants to groups of operators $\{D_\tau\}_\tau$ such as rigid translations $D_\tau f(x) = f(x-\tau)$, rotations, scalings...

$$\text{if} \quad f \in \mathcal{C} \quad \text{then} \quad D_\tau f \in \mathcal{C} \quad \text{so} \quad d(f, D_\tau f) = 0 \ .$$

  hence Φ should be invariant: $\Phi(D_\tau f) = \Phi(f)$ if $\tau = cst$

- For non-rigid deformations $\tau(x)$: $D_\tau f(x) = f(x - \tau(x))$

  metrics should provide the elastic deformation amplitude

$$\|\Phi(f) - \Phi(D_\tau f)\| \sim \|f\|_a \, \|\nabla \tau\|_b \ .$$

- Metric on stationary processes: $\|E\{\Phi(F)\} - E\{\Phi(G)\}\| \ .$

Overview

- Failures of Fourier and wavelet metrics
- Interferences and invariant scattering: deep neural networks
- Mathematical properties of the invariant metrics
- Invariant metric on stationary processes
- O(N) learning and classification of patterns and textures
- Invariant scattering for general groups

Deformation Instability of Fourier

- Elastic deformation $D_\tau f(x) = f(x - \tau(x))$ with $|\nabla\tau| < 1$.

- The Fourier modulus is translation invariant:

$$\text{If } \tau(x) = cst \text{ then } |\widehat{D_\tau f}(\omega)| = |\hat{f}(\omega)| \quad : \Phi(f) = |\hat{f}| \ .$$

- High frequency instability:

$$\text{If } \tau(x) \neq cst \text{ then } \tau(x) \approx \tau(x_0) + \nabla\tau(x_0) \cdot (x - x_0) \text{ affine.}$$

$$f(x) = \theta(x)\, \mathrm{e}^{i\xi x} \Rightarrow D_\tau f(x) = \theta(x - \tau(x))\, \mathrm{e}^{i\phi(x_0)}\, \mathrm{e}^{i(Id - \nabla\tau(x_0))\xi x}$$

$$\Rightarrow \|\, |\widehat{D_\tau f}| - |\, \hat{f}|\, \| \sim \|f\|\, \|\nabla\tau \cdot \xi\|_\infty$$

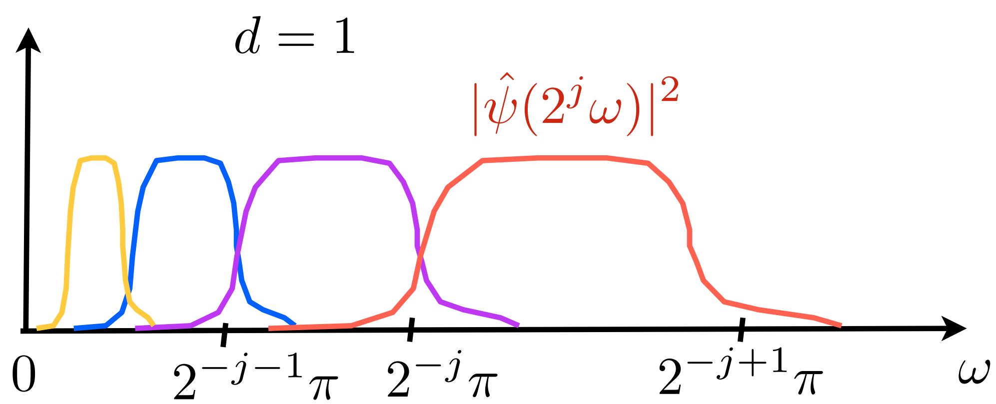
Wavelet Transform Strategy

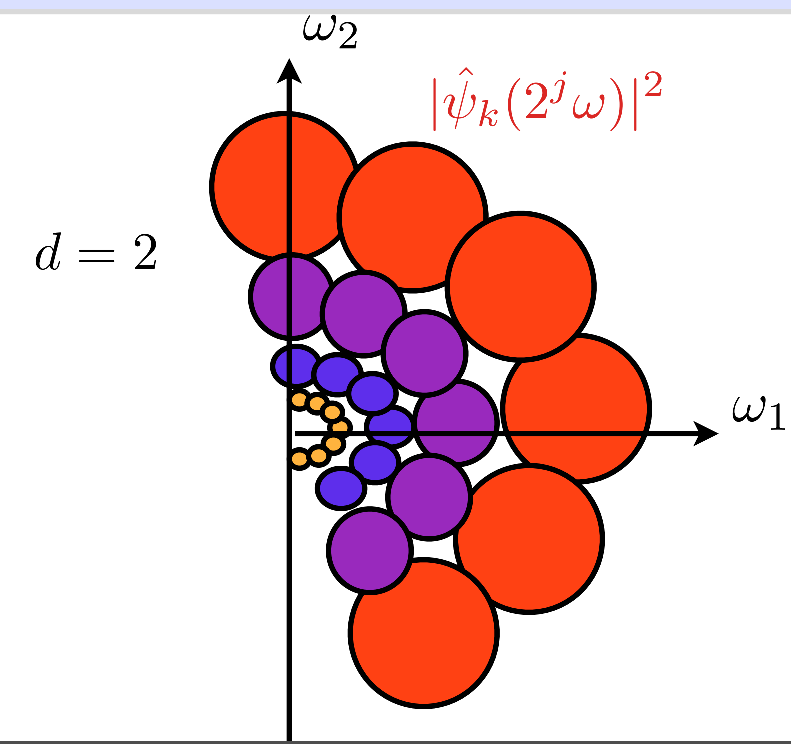
- Separates signal components in dyadic frequency bands:

$$W_{j,k} f = f \star \psi_{j,k}(x) \quad \text{with} \quad \psi_{j,k}(x) = 2^{-jd}\, \psi_k(2^{-j}x)$$

If $\forall \omega \;,\; 2(1-\delta) \leq \sum_{j,k} \left(|\hat{\psi}_k(2^j\omega)|^2 + |\hat{\psi}_k(-2^j\omega)|^2 \right) \leq 2$

then $(1-\delta)\, \|f\|^2 \leq \sum_{j,k} \|W_{j,k} f\|^2 \leq \|f\|^2$

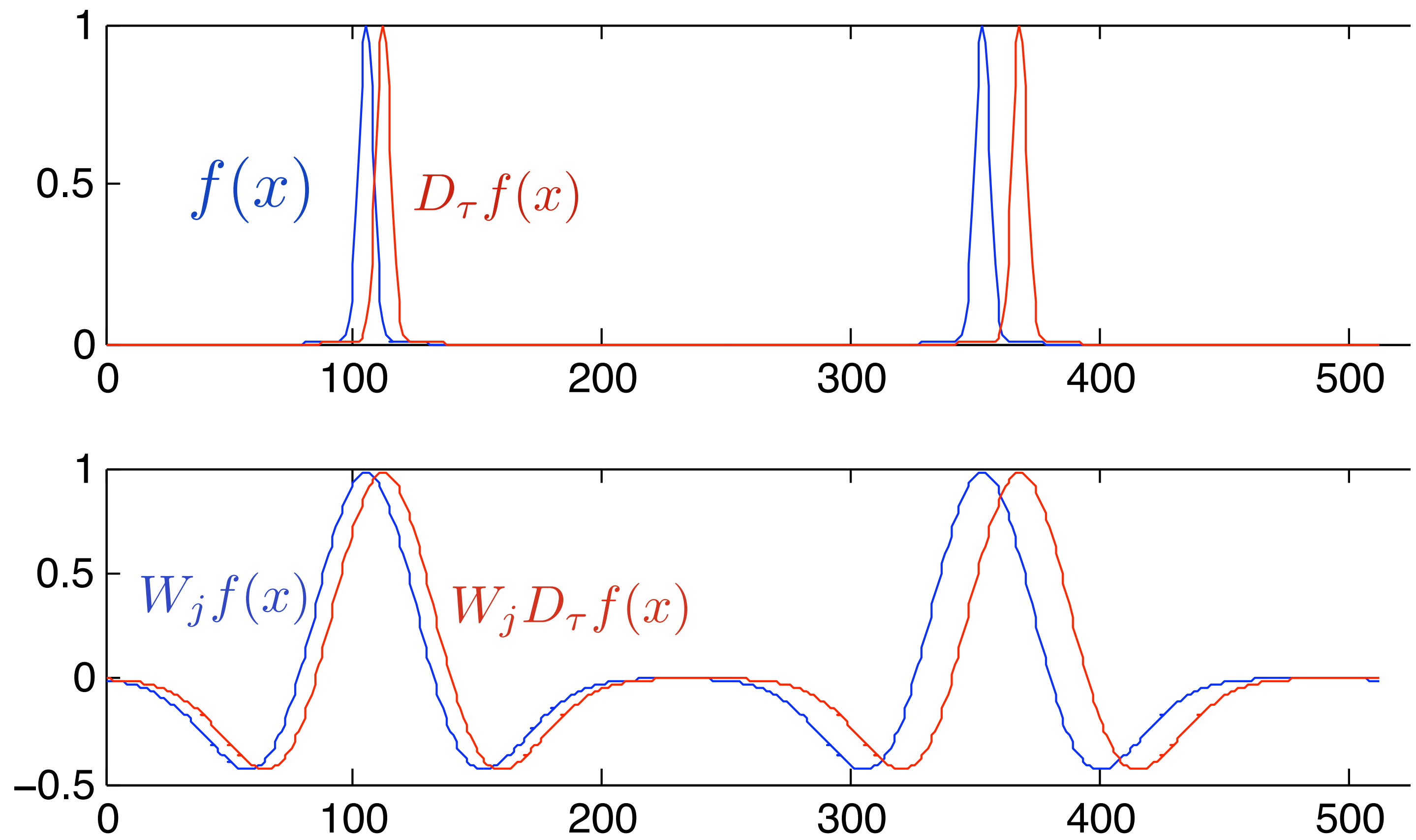




Invariants with Wavelets $\Phi = W_j$

- **Theorem** If $D_\tau f(x) = f(x - \tau(x))$ then

$$\|W_j D_\tau f - W_j f\| \leq C \, \|f\| \left(2^{-j} \, \|\tau\|_\infty + \|\nabla \tau\|_\infty\right)$$



- Near invariance if $\|\tau\|_\infty \ll 2^j$.

Wavelet Failure

- Coarse to fine strategies: begin a large scale 2^j and refine.
- Large scales keep low frequencies which are not discriminative, and hence produce big errors.
- **How to build invariants and measure deformations through high frequencies ?**
- Map high frequency wavelet coefficients to lower frequencies.

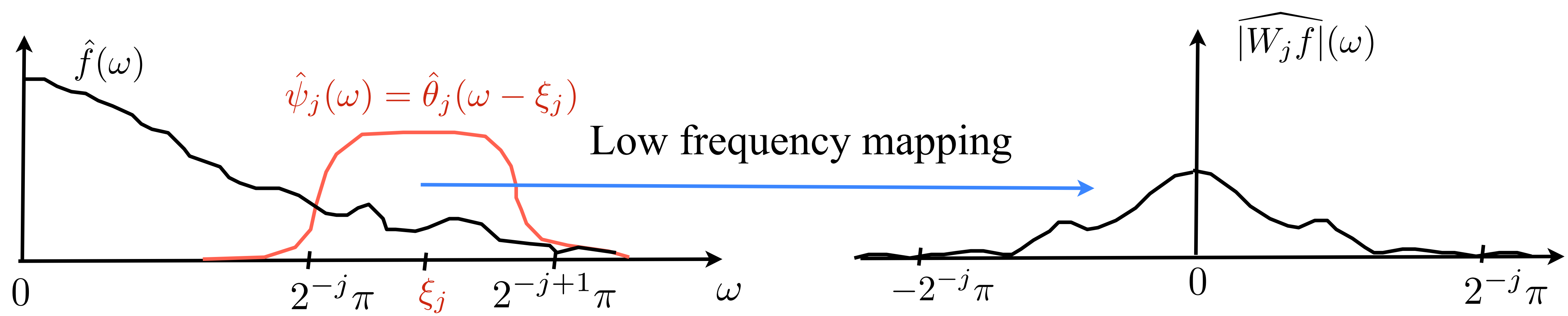
Modulus Demodulation

$$\text{If} \quad \psi(x) = \theta(x)\, e^{i\xi x} \quad \text{then} \quad \psi_j(x) = \theta_j(x)\, e^{i\xi_j x}$$

$$\text{with } \theta_j(x) = 2^{-dj}\,\theta(2^{-j}x) \quad \text{and} \quad \xi_j = 2^{-j}\xi$$

$$\text{so} \quad W_j f(x) = e^{i\xi_j x}\, f_j \star \theta_j(x) \quad \text{with} \quad f_j(x) = e^{i\xi_j x}\, f(x)$$

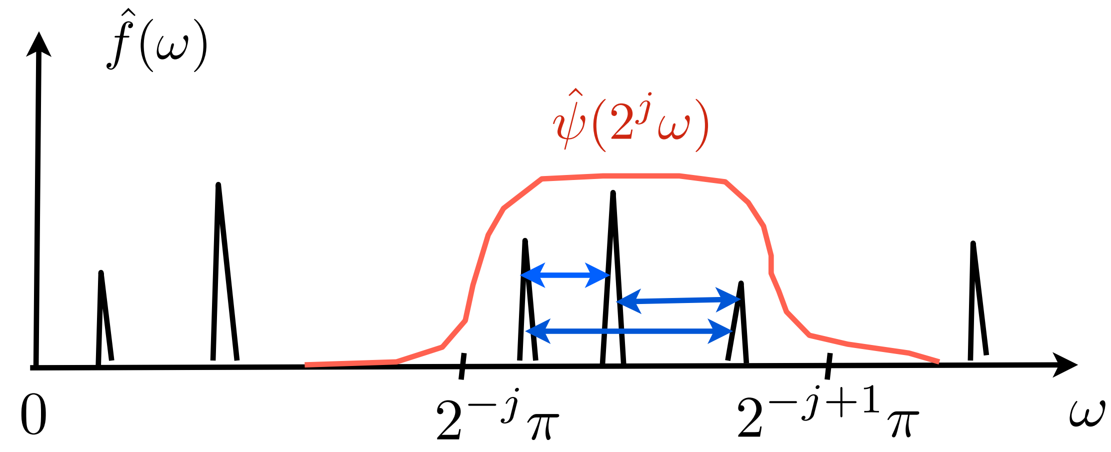
$$\text{hence} \quad |W_j f(x)| = |f_j \star \theta_j(x)| \; .$$



Modulus Interferences

$$f(x) = \sum_m a_m \cos(\omega_m x)$$

$$a_{j,m} = a_m \, \hat{\psi}(2^j \omega_m)$$

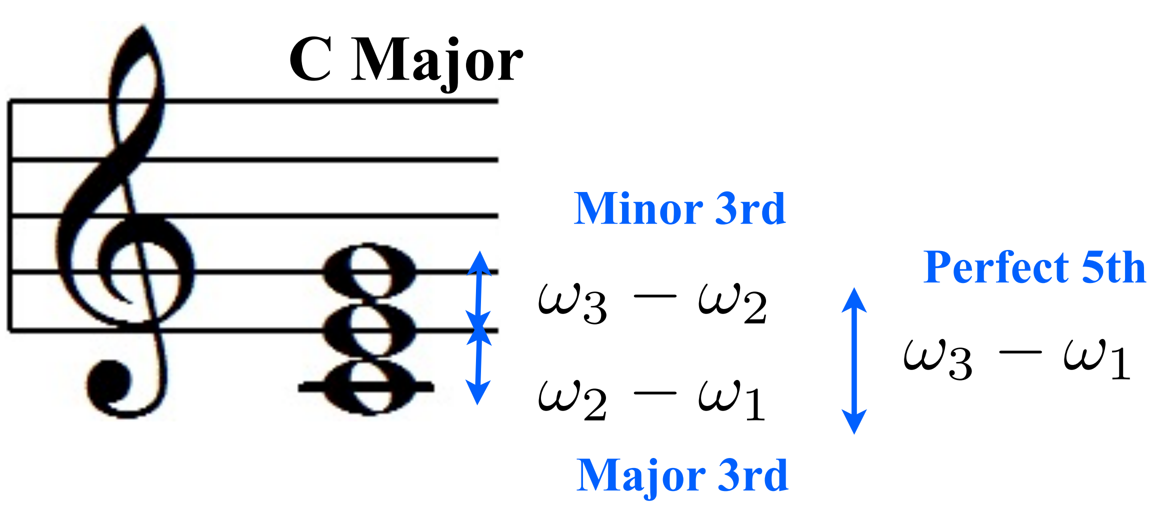


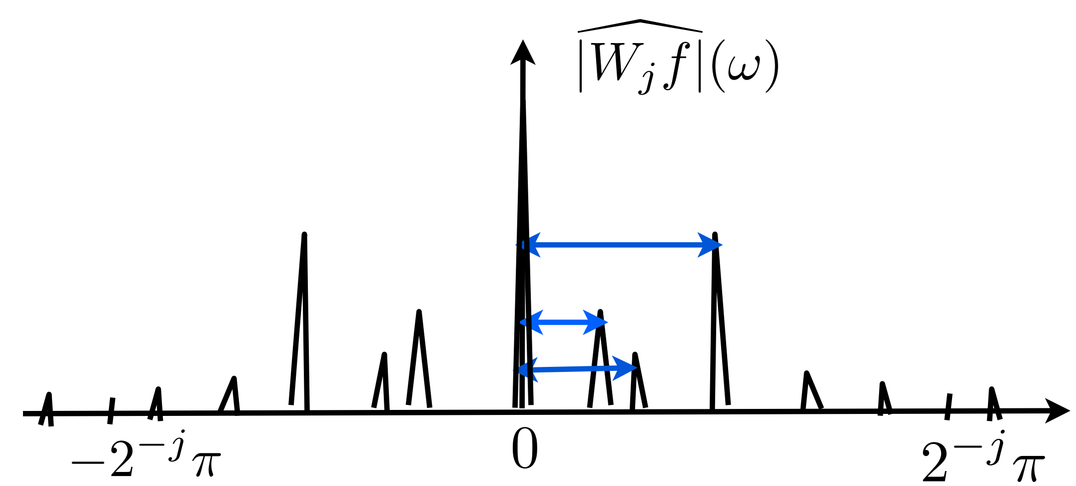
Energy : e_j^2 ***Interferences :*** $\epsilon_j(x)$

$$|W_j f(x)|^2 = \sum_m |a_{j,m}|^2 + 2 \sum_{m' \neq m} a_{j,m} a_{j,m'} \cos(\omega_m - \omega_{m'})x$$

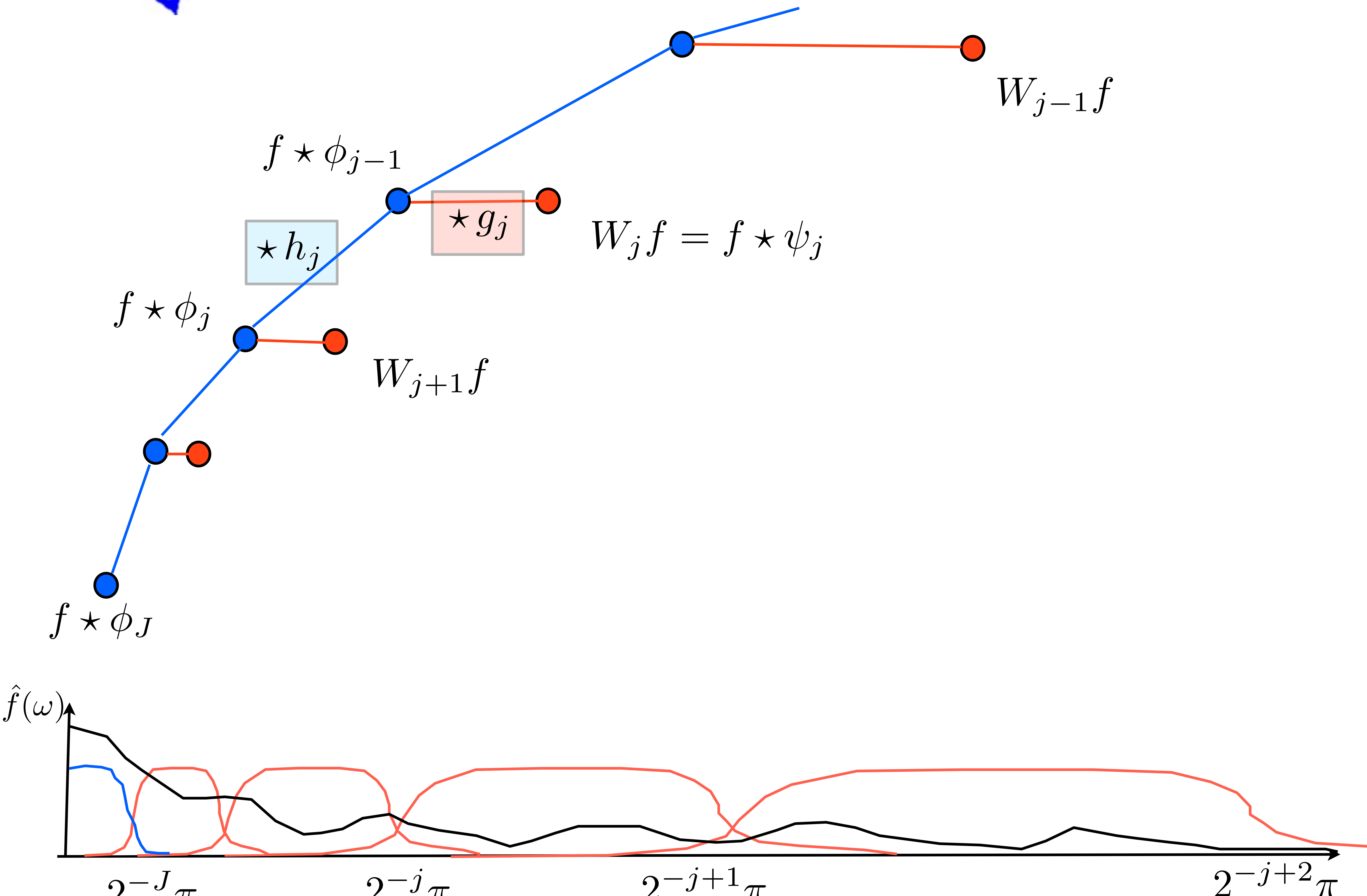
$$|W_j f(x)| = e_j + \frac{\epsilon_j(x)}{2e_j} + O\left(\frac{\epsilon_j^2(x)}{e_j^3}\right) .$$

Music chord :

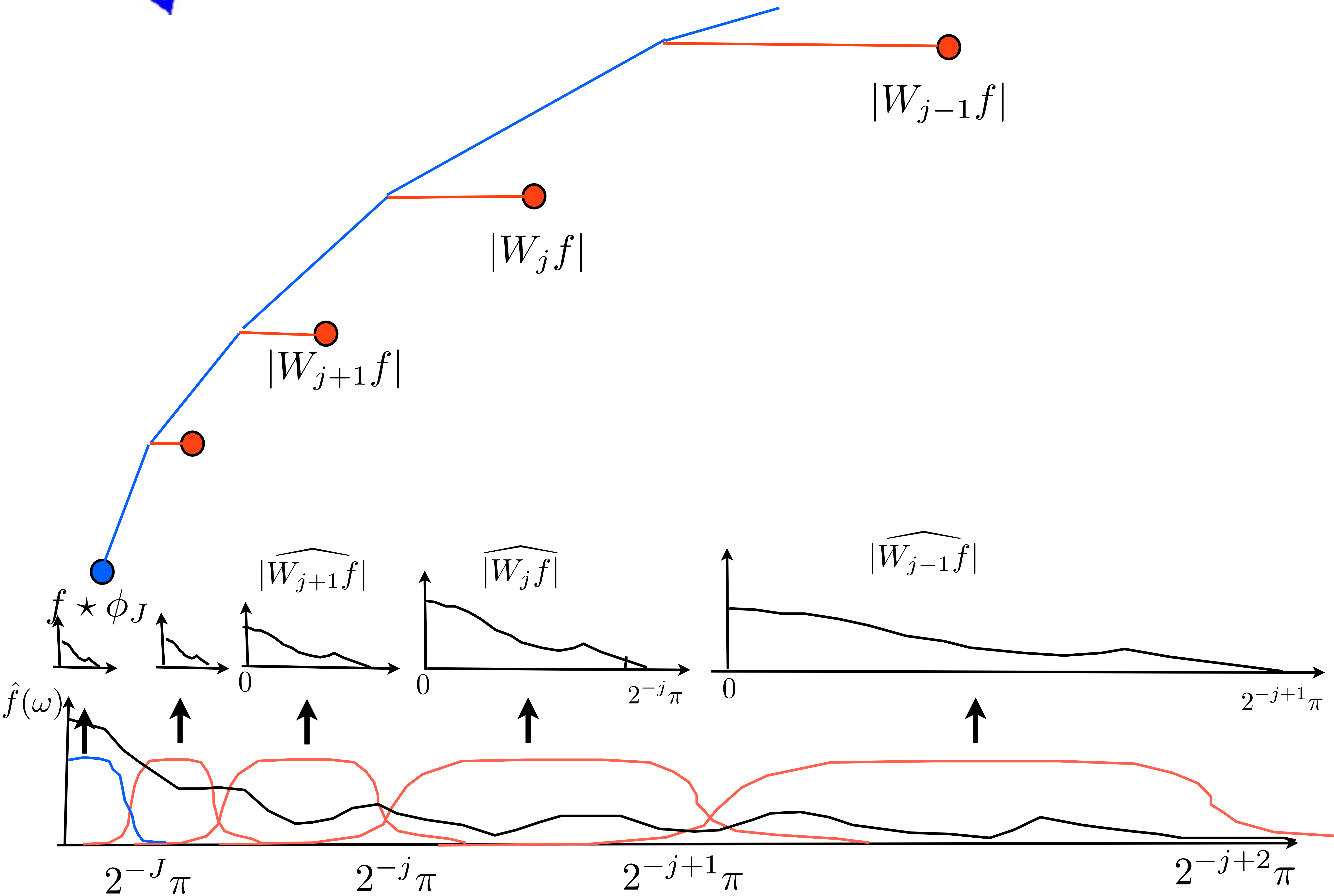




Wavelet Transform



Interference Tree



Quantum Scattering

f — *diffusion* — $f \star \phi_j$ — $f \star \phi_{j+1}$ — $U_J = f \star \phi_J$

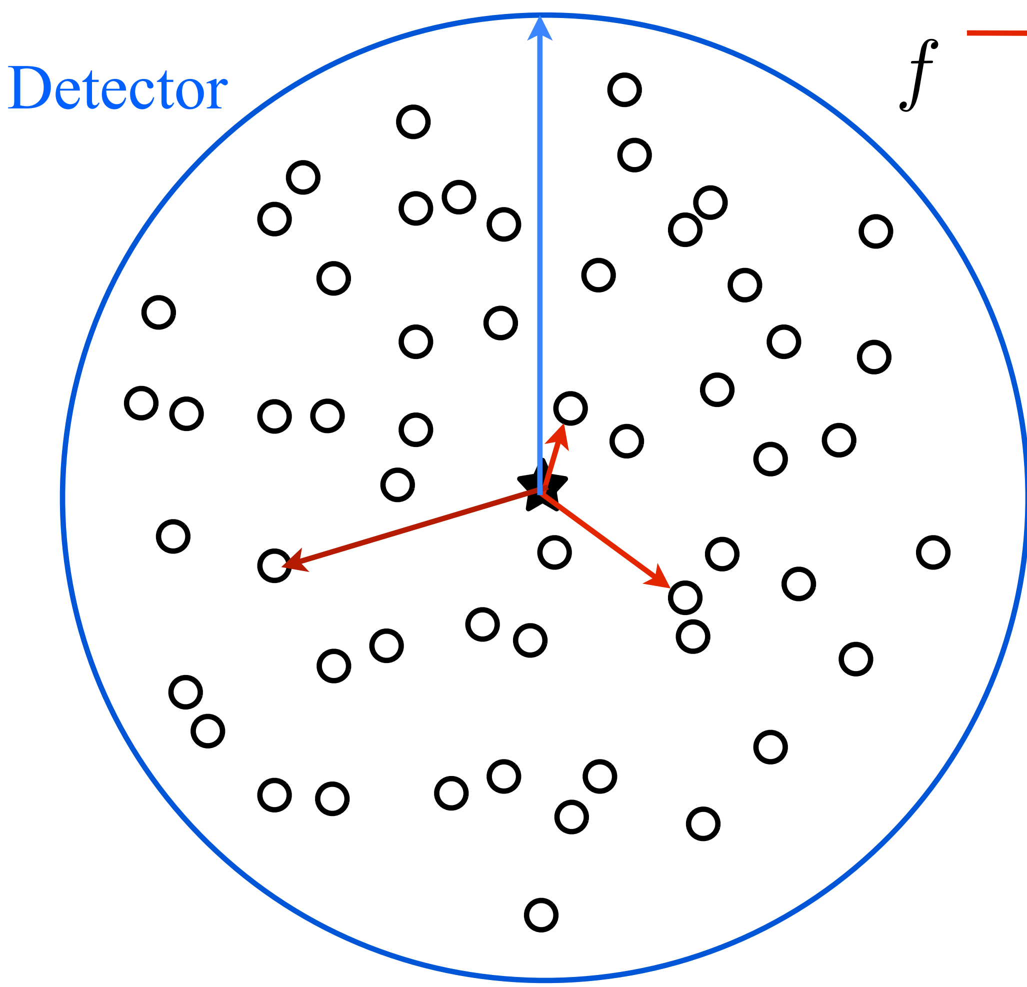
$0 \quad 2^{j-1} \quad 2^j \quad 2^{j+1} \quad 2^J \quad time$

interferences

$$U_{j,k} f = |f \star \psi_{j,k}| \qquad U_{j+1,k} f = |f \star \psi_{j+1,k}|$$

$$1 \leq k \leq K$$

$$f \xrightarrow{U} \{U_{j_1,k_1} f\}_{j_1,k_1}$$



Quantum Scattering

f $\quad$ *diffusion* $\quad f \star \phi_j \quad f \star \phi_{j+1} \quad U_J = f \star \phi_J$

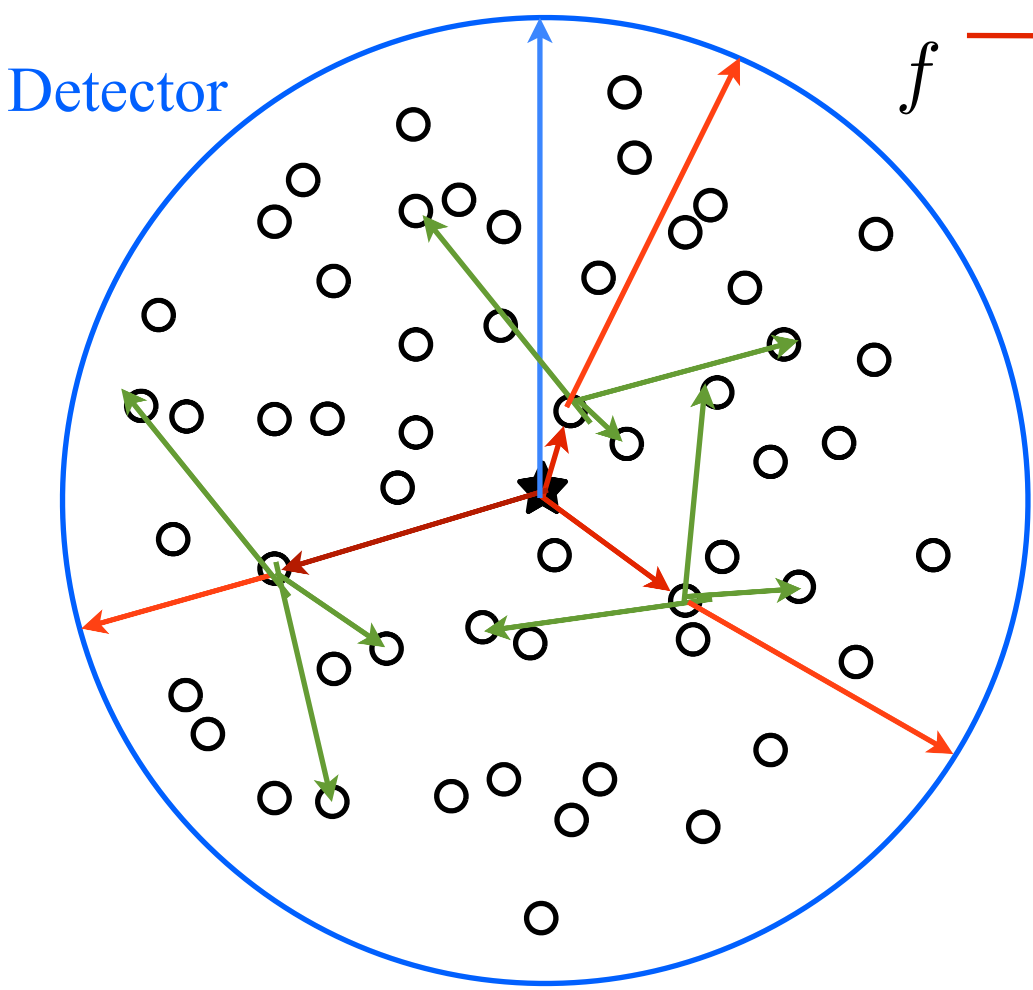
$0 \quad 2^{j-1} \quad 2^j \quad 2^{j+1} \quad 2^J$ *time*

interferences

$$U_{j,k} f = |f \star \psi_{j,k}| \qquad U_{j+1,k} f = |f \star \psi_{j+1,k}|$$

$$1 \leq k \leq K$$

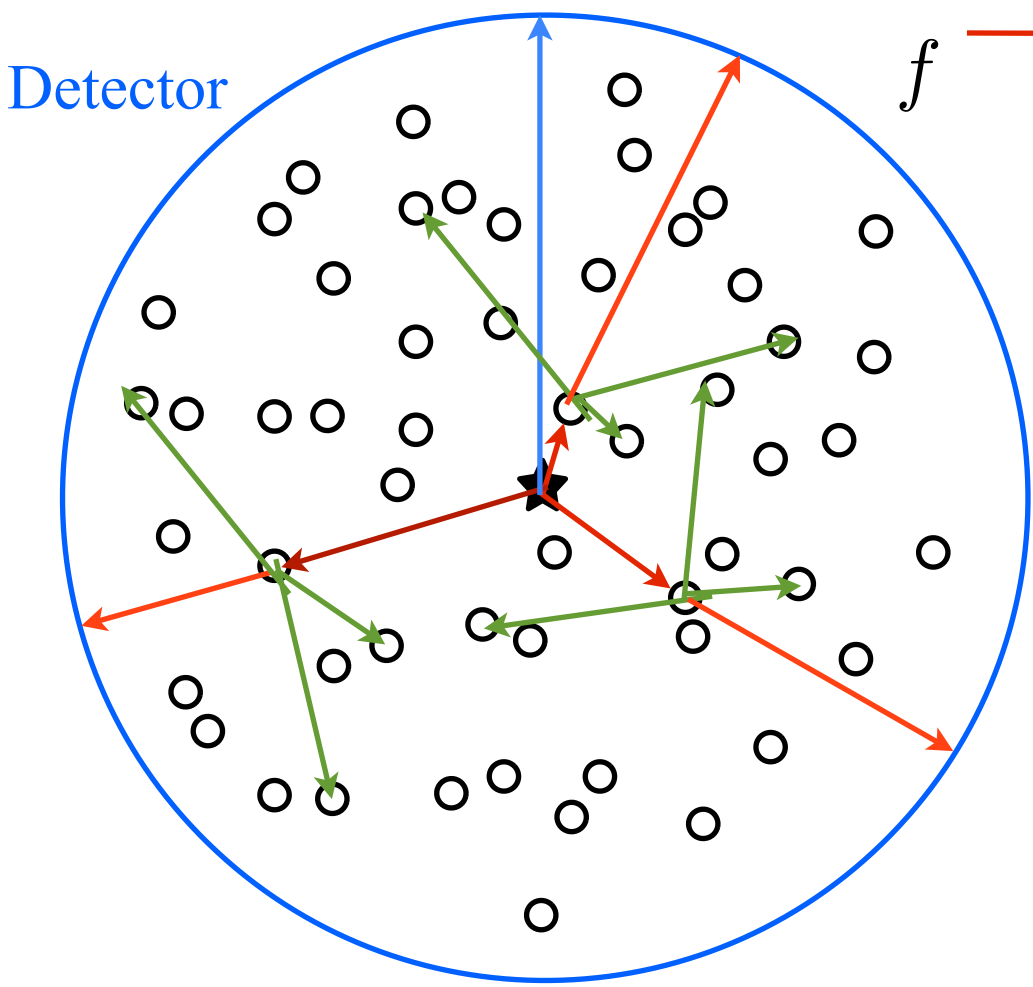
$$f \xrightarrow{U} \{U_{j_1,k_1} f\}_{j_1,k_1} \xrightarrow{U} \{U_{j_1,k_1} f \, U_{j_2,k_2} f\}_{\substack{j_1,k_1 \\ j_2,k_2}}$$



Quantum Scattering

Iteration on a unitary one-step propagator $U : U^m$

Builds paths $p = \{(j_1, k_1)\,,\, (j_2, k_2)\,, \cdots\,,\, (j_{|p|}, k_{|p|})\}$



$$f \xrightarrow{U} \{U_{j_1,k_1} f\}_{j_1,k_1} \xrightarrow{U} \{U_{j_1,k_1} f\, U_{j_2,k_2} f\}_{\substack{j_1,k_1\\ j_2,k_2}}$$

Quantum Scattering

Iteration on a unitary one-step propagator $U : U^m$

Builds paths $p = \{(j_1, k_1)\,, (j_2, k_2)\,, \cdots\,, (j_{|p|}, k_{|p|})\}$

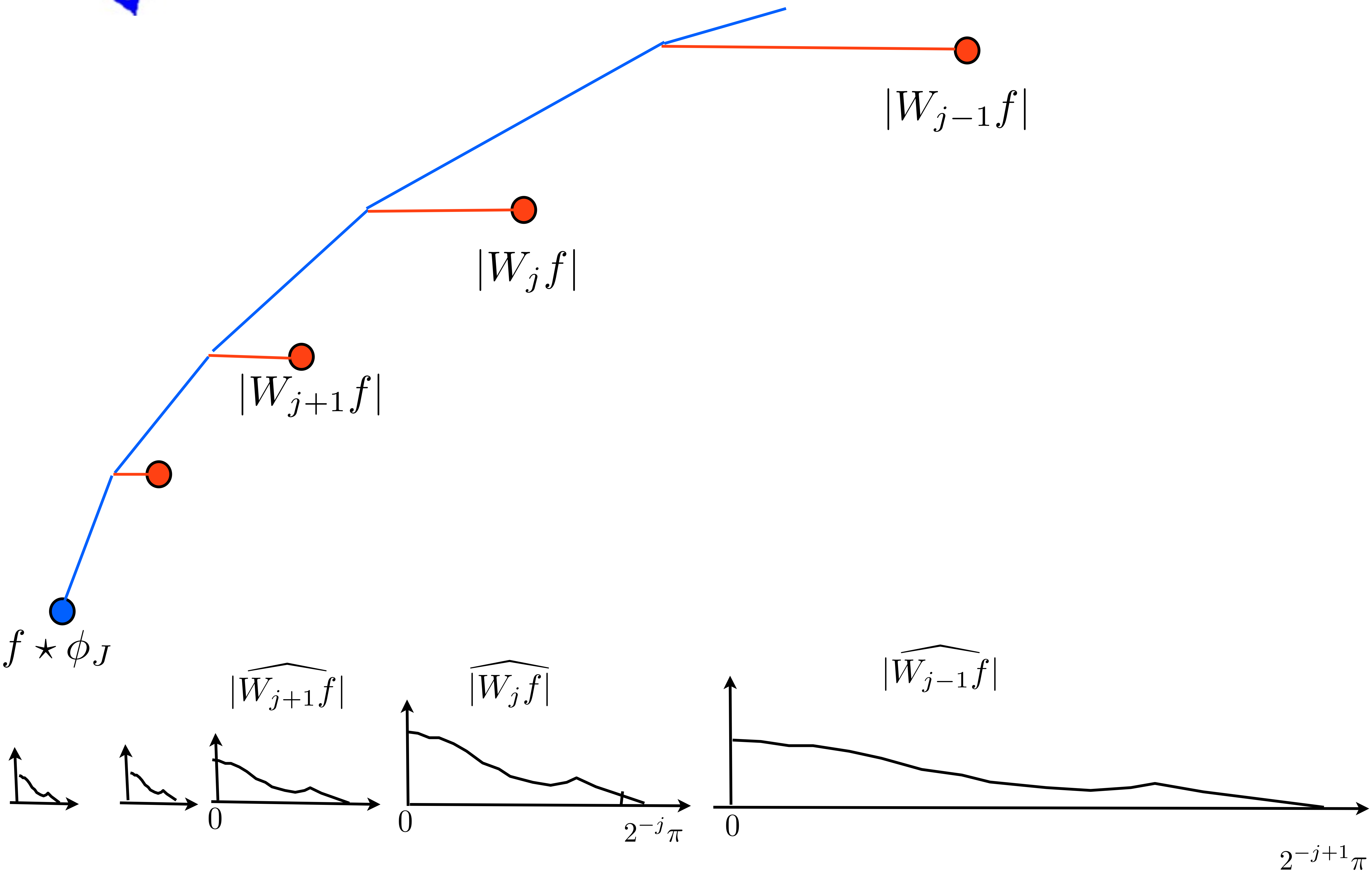
Scattering operator computes wavefunctions along paths:

$$S_J(p) f = U_J \prod_{n=1}^{|p|} U_{j_n,k_n} f = |\cdots| f \star \psi_{j_1,k_1}| \star \psi_{j_2,k_2}| \star \cdots | \star \phi_J$$

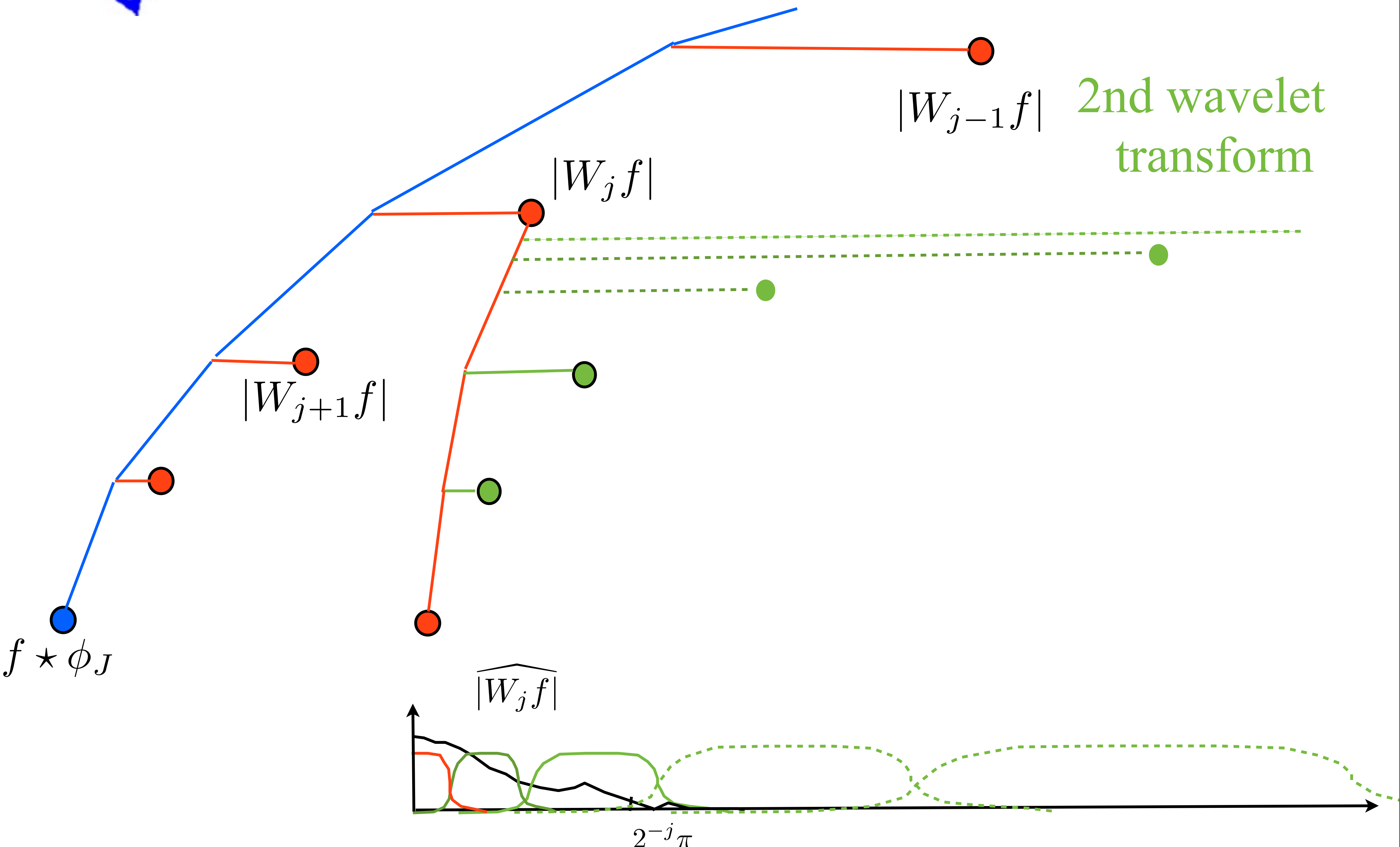
$|p|$ is the scattering order.

$\|S_J(p) f\|^2$: probability to reach the detector through the path "p".

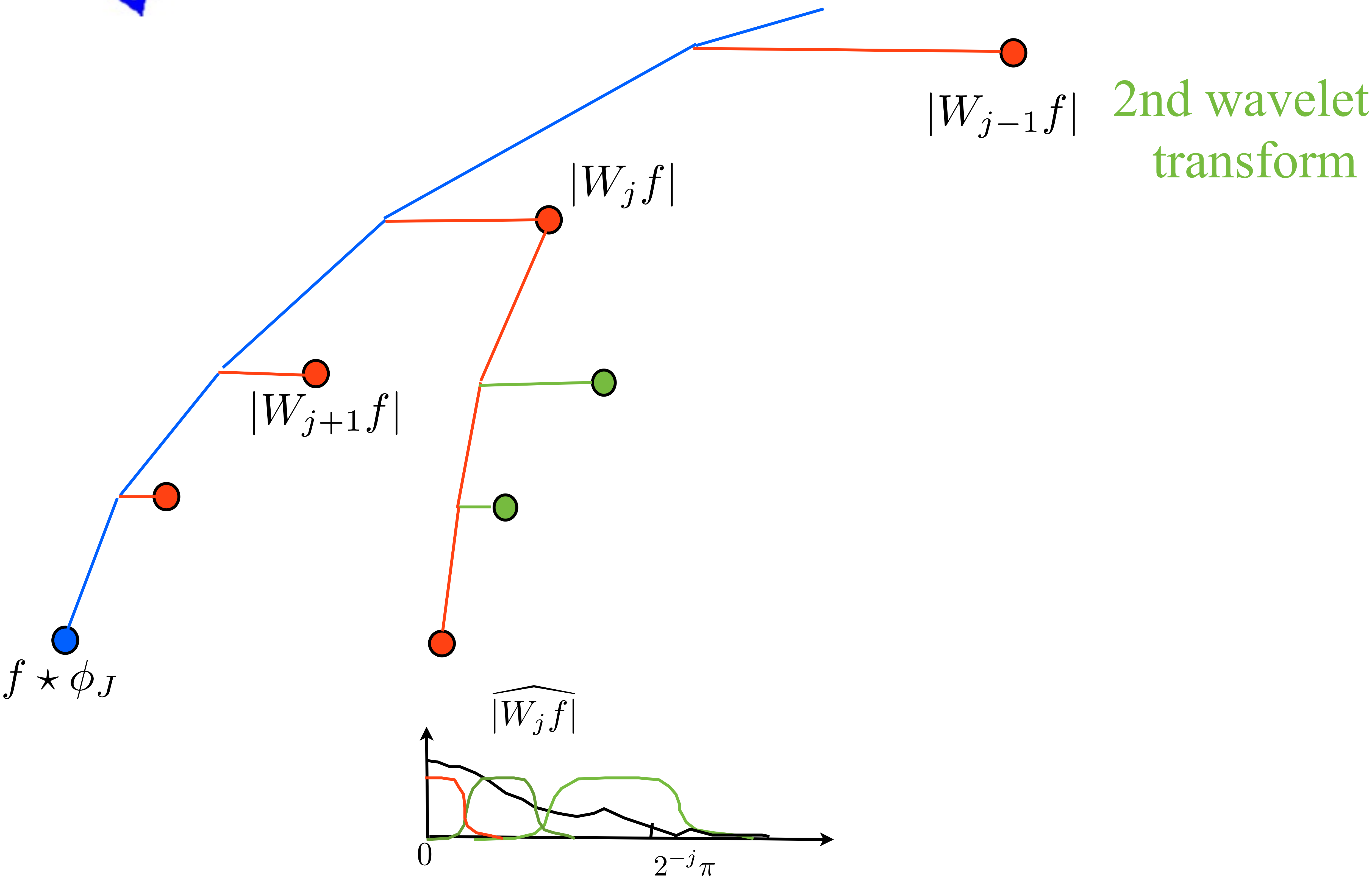
One-step propagator U



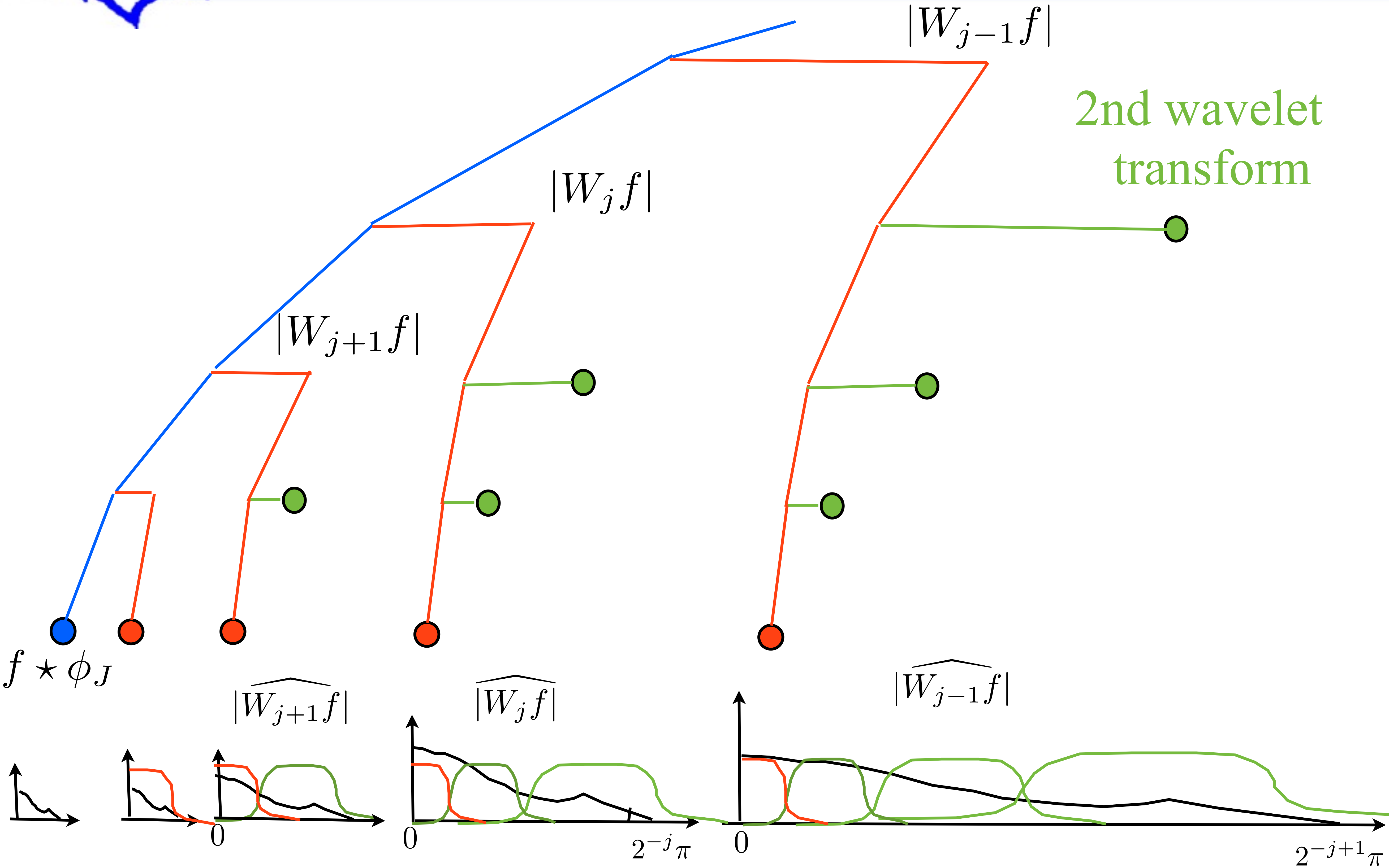
One-step propagator U



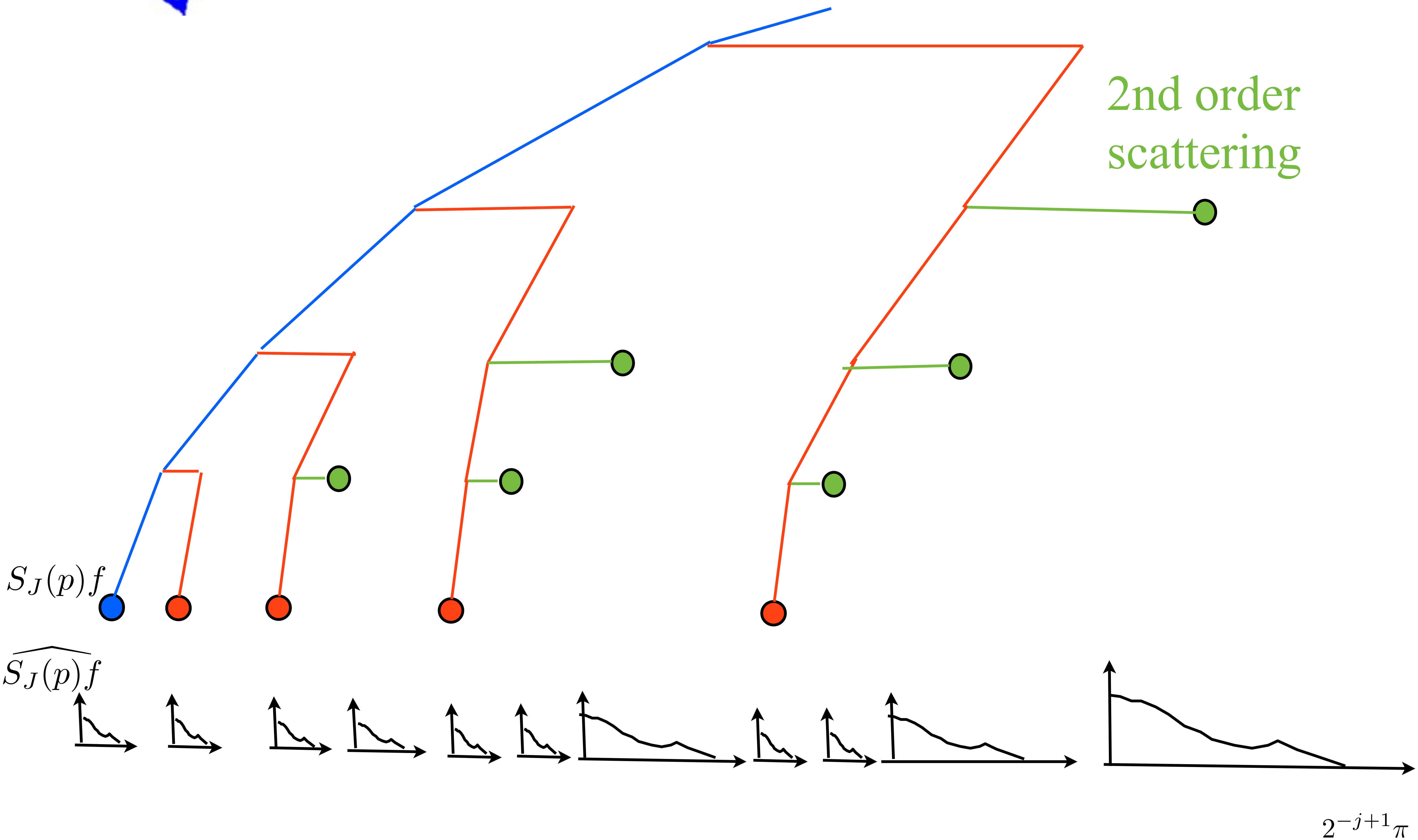
Progressive paths: $j_{n+1} > j_n$



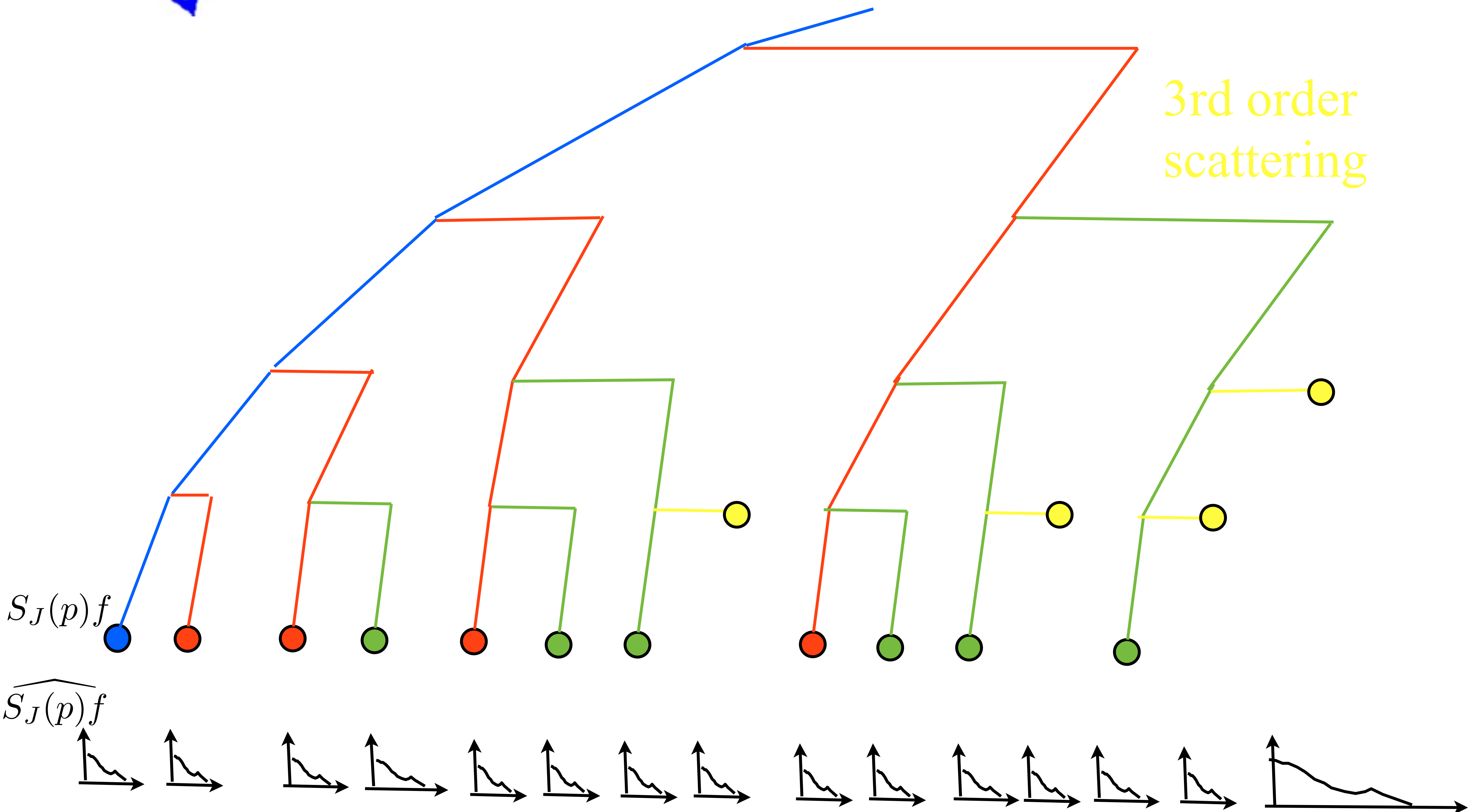
Progressive paths: $j_{n+1} > j_n$



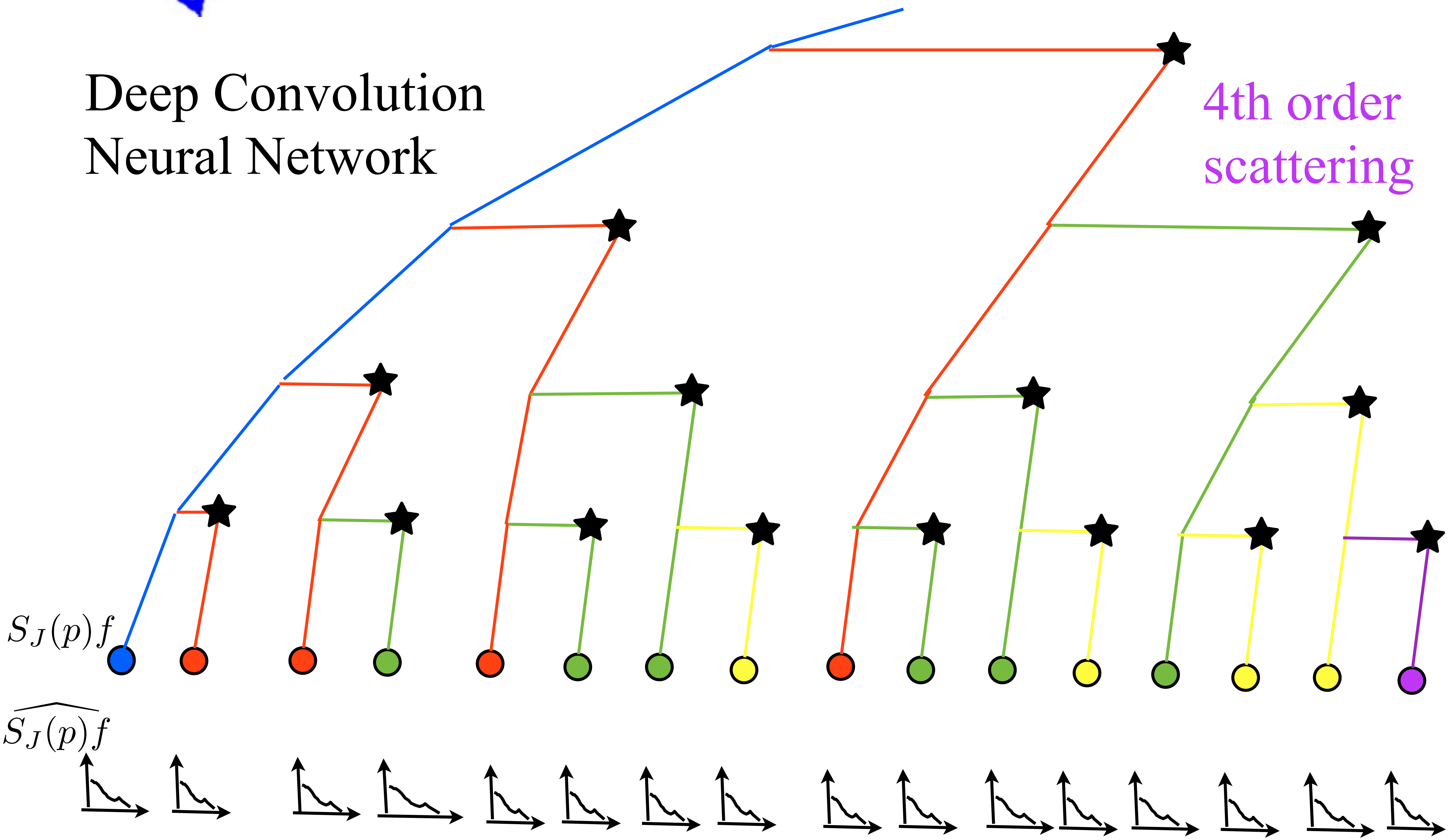
Progressive paths: $j_{n+1} > j_n$



Progressive paths: $j_{n+1} > j_n$



Progressive paths: $j_{n+1} > j_n$

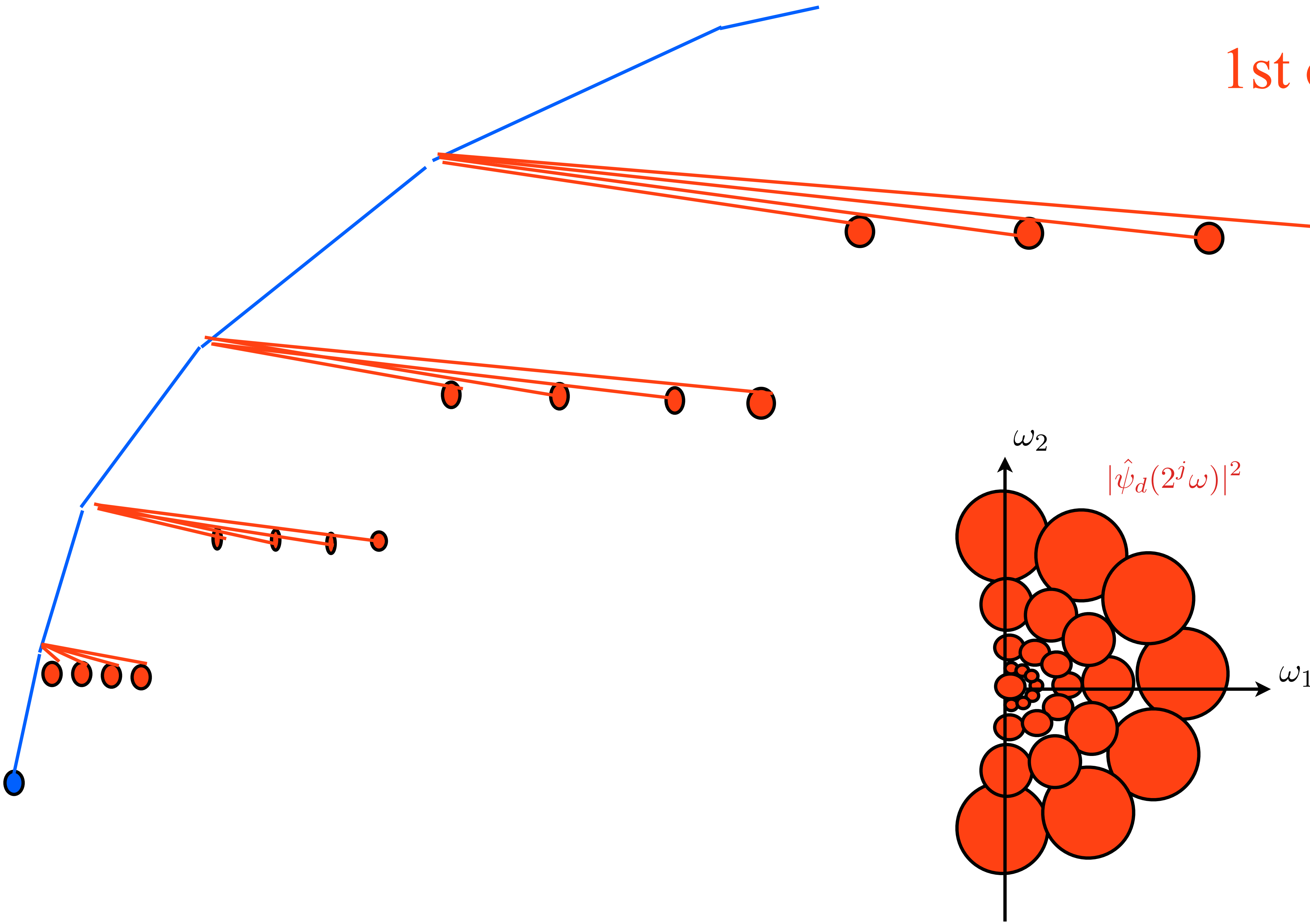


When J increases, $S_J(p)f$ converges to the $\mathbf{L}^1$ norm along the path:

$$\lim_{J\to\infty} 2^{Jd}\, S_J(p)f(x) = \int |\cdots||f \star \psi_{j_1,k_1}| \star \psi_{j_2,k_2}| \star \cdots|\, dx$$

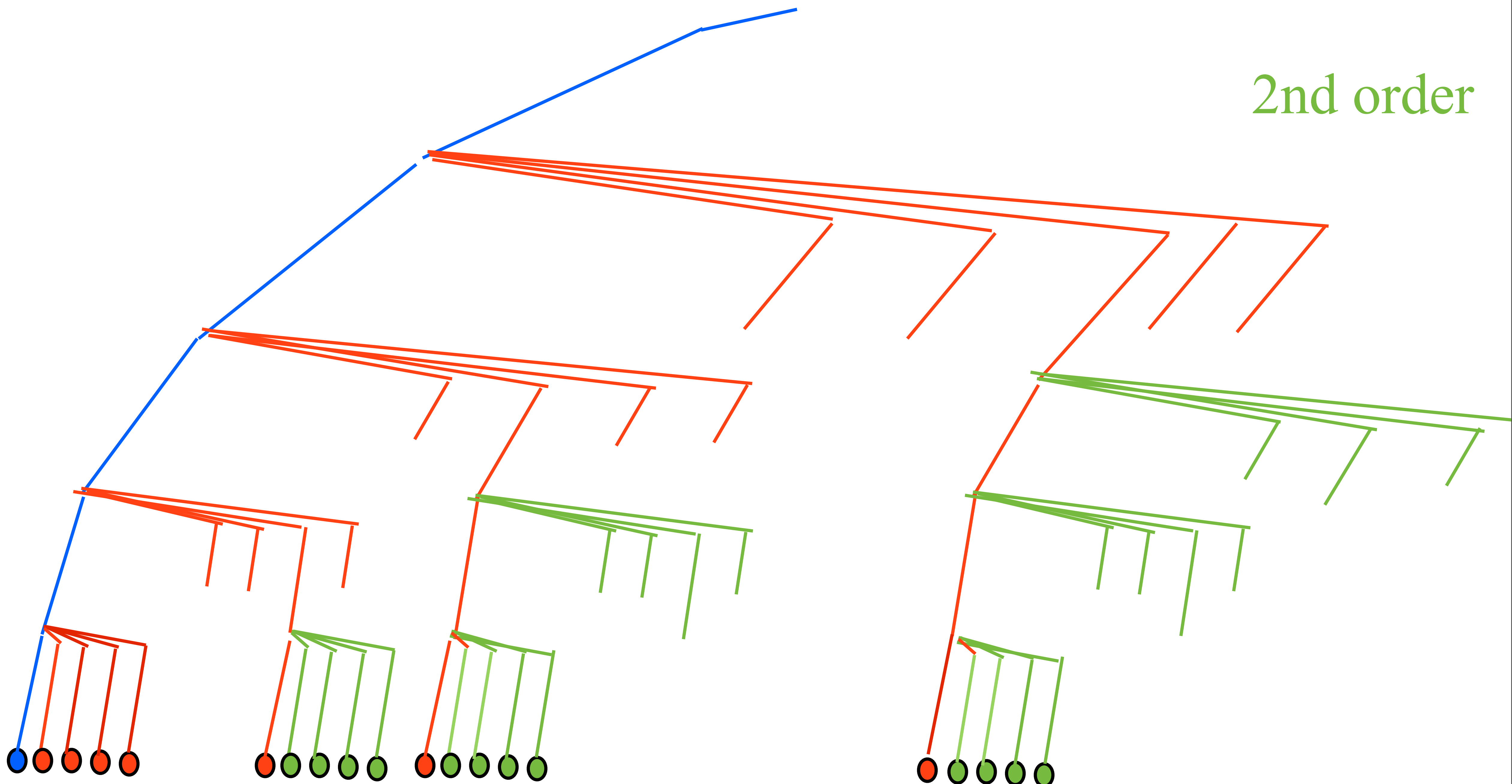
Multiple Mother Wavelets

Multiple mother wavelet transform: $\psi_k(x)$ for $1 \leq k \leq K$.



Multiple Mother Wavelets

Multiple mother wavelet transform: $\psi_k(x)$ for $1 \leq k \leq K$.

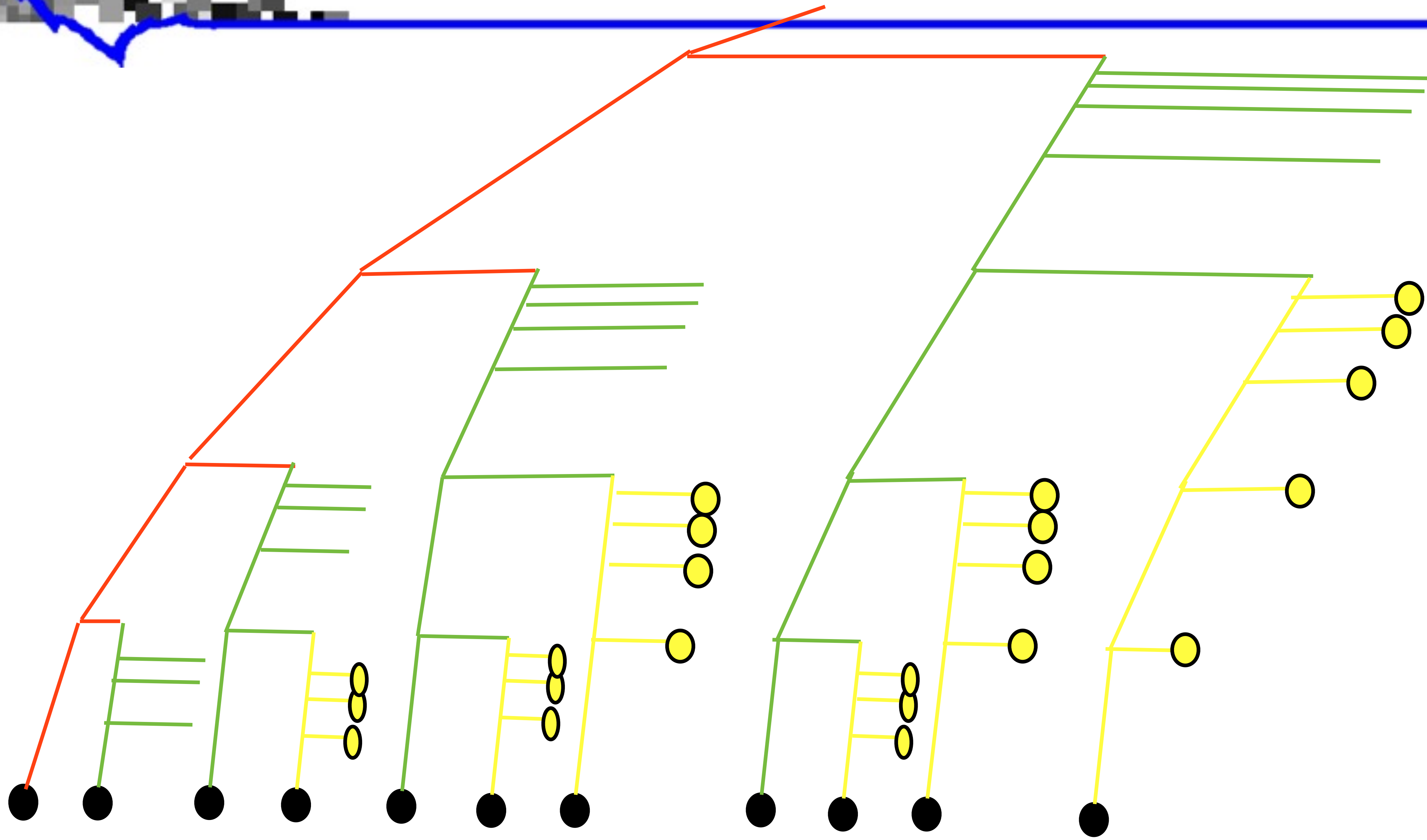


Metric over Scattering Paths

What are the properties of the scattering metric ?

$$
\begin{aligned}
\|S_J f - S_J g\|^2 &= \sum_p \int |S_J(p) f(x) - S_J(p) g(x)|^2 \, dx \\
&= \sum_p \|S_J(p) f - S_J(p) g\|^2
\end{aligned}
$$

Contraction

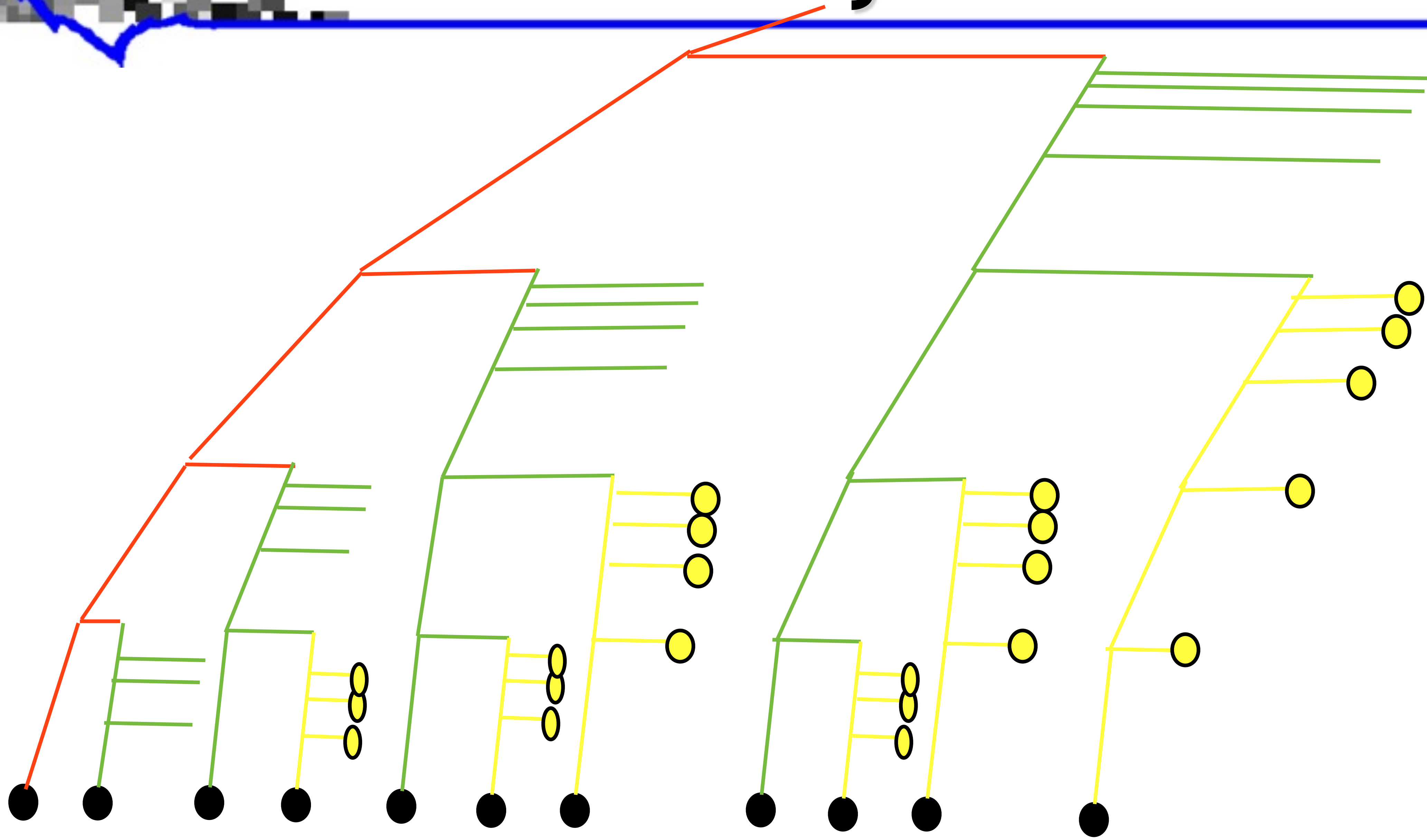


If $|p| \leq m$ then $S_J(p)f = U^m(p)f$ where U is contractivee and unitary so

S_J is contractive:

$$\|S_J f - S_J g\|^2 = \sum_p \|S_J(p)f - S_J(p)g\|^2 \leq \|f - g\|^2 \ .$$

Unitary



Theorem: For appropriate **complex** wavelets, S_J is unitary:

$$\|S_J f\|^2 = \sum_p \|S_J(p) f\|^2 = \|f\|^2 \ .$$

High energy paths are low order progressive paths.

Limit Metric

Theorem For $(f,g) \in \mathbf{L^2}(\mathbf{R^2})$:

$$\|S_{J+1} f - S_{J+1} g\| \leq \|S_J f - S_J g\|$$

$$\text{so} \quad \lim_{J\to+\infty} \|S_J f - S_J g\| = d(f,g) \leq \|f-g\|$$

$$\text{and} \quad d(f,0) = \|f\| \ .$$

If f is supported in $[0,2^L]^d$ then $J \leq L$:

$$S_L(p) f = 2^{-dL} \int_{[0,2^L]^d} |\cdots|f \star \psi_{j_1,k_1}| \star \psi_{j_2,k_2}| \star \cdots|\, dx$$

$$d^2(f,g) = \|S_L f - S_L g\|^2 = \sum_p |S_L(p) f - S_L(p) g|^2 \ .$$

If $f \in \mathbf{L^2}(\mathbf{R}^d)$ then $d(f,g)$ is an integral over a path variable.

Translation Invariance

If $D_\tau f(x) = f(x - \tau)$ is a translation then

$$S_J(p) D_\tau f(x) = S_J f(x - \tau) = D_\tau \, S_J f(x) \ .$$

Theorem:

$$\lim_{J \to \infty} \|S_J D_\tau f - S_J f\| = d(D_\tau f \ , \ f) = 0 \ .$$

Elastic Deformations

Theorem If $D_\tau f(x) = f(x - \tau(x))$ with $\|\nabla \tau\|_\infty < 1$

then for $J > \log \frac{\|\tau\|_\infty}{\|\nabla\tau\|_\infty}$

$$\|S_J D_\tau f - S_J f\| \leq C \, \|f\|_w \, \log^{3/2}\left(\frac{\|\tau\|_\infty}{\|\nabla\tau\|_\infty}\right) \|\nabla\tau\|_\infty$$

$$\text{with } \|f\|_w = \sum_{j=0}^{+\infty} \|W_j f\|^2 + \sum_{j=-\infty}^{0} |j| \, \|W_j f\|^2 \; .$$

Proof:

$$\|S_J D_\tau f - S_J f\| \leq \|D_\tau S_J f - S_J f\| + \|D_\tau S_J f - S_J D_\tau f\|$$

Key element: $\|[W, D_\tau]\|^2 = \|\sum_j [W_j, D_\tau] \, [W_j, D_\tau]^* \|$

Linearisation of Deformations

Theorem If $D_\tau f(x) = f(x - \tau(x))$ with $\|\nabla \tau\|_\infty < 1$

then for $J > \left(\log \frac{\|\tau\|_\infty}{\|\nabla\tau\|_\infty}\right)^{1/2}$

$$\|S_J D_\tau f - S_J f + \tau \cdot \nabla S_J(p) f\| \leq C\,\|f\|_w \log\left(\frac{\|\tau\|_\infty}{\|\nabla\tau\|_\infty}\right) \|\nabla\tau\|_\infty$$

- Deformations are linearized: possibility to learn classification metrics through affine projections.

- Deformations $\tau(x)$ (optical flow, stereo disparity) can be estimated with a system of linear equations:

$$\forall p\ ,\ S_J(p) D_\tau f(x) - S_J(p) f(x) + \tau(x) \cdot \nabla S_J(p) f(x) \approx 0\ .$$

Scattering Stationary Processes

- **Theorem:** If $F(x)$ is a stationary then $S_J(p)F(x)$ is stationary.

$$E\{S_J(p)F(x)\} = E\{|\cdots|F \star \psi_{j_1}|\cdots| \star \psi_{j_{|p|}}(x)|\}$$

and $\quad \mathrm{var}(S_J(p)F(x)) \leq \mathrm{var}(F(x))\,\beta^{|p|}$ with $\beta < 1$.

- Indeed, if $F(x)$ is stationary then $F \star \psi_j(x)$ and $|F \star \psi_j(x)|$ are stationary and the modulus reduces the variance:

$$\frac{\mathrm{var}(|F \star \psi_j|)}{\mathrm{var}(F \star \psi_j)} = 1 - \frac{\pi}{4} \quad \text{if } F \text{ is Gaussian} .$$

Invariant Metric on Processes

$$\|E\{S_J F\} - E\{S_J G\}\|^2 = \sum_p |E\{S_J(p)F\} - E\{S_J(p)G\}|^2$$

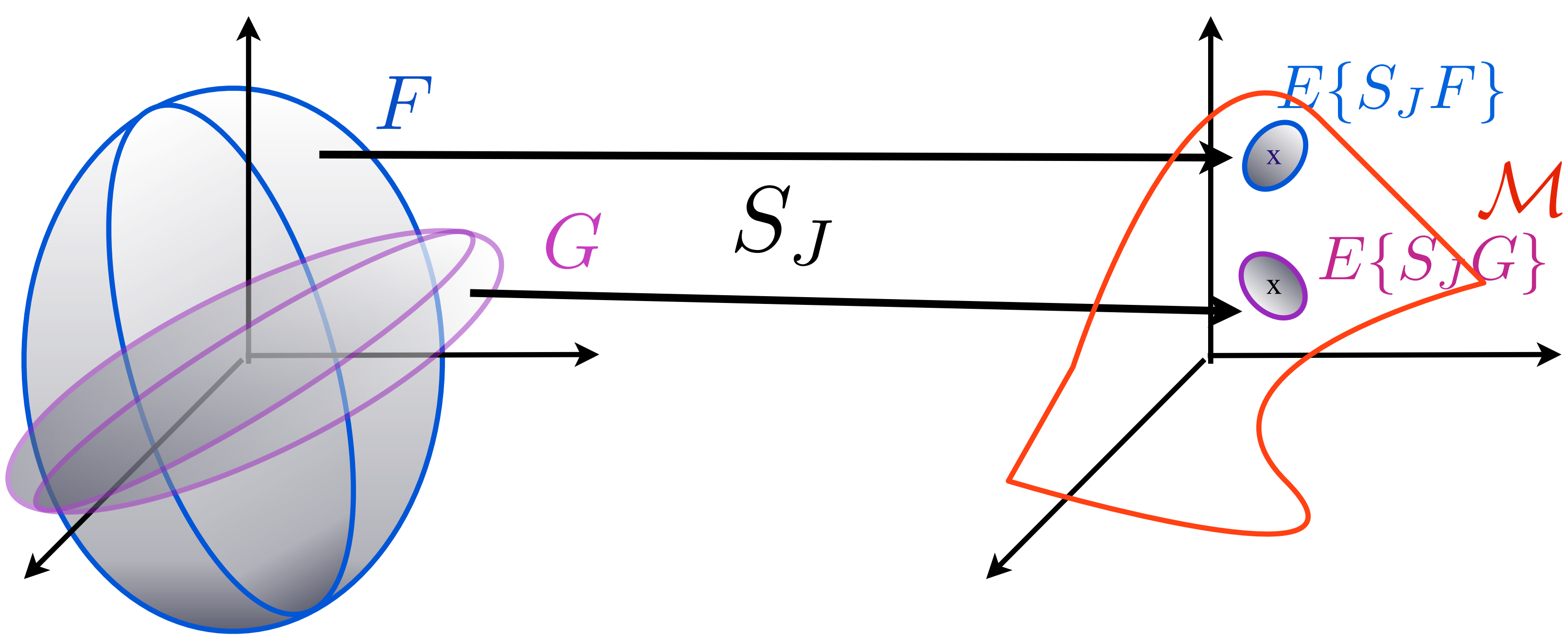
Theorem: If F and G are stationary then

$$\|E\{S_J F\} - E\{S_J G\}\|^2 \leq E\{|F(x) - G(x)|^2\} \, .$$

For Gaussian white noise and wide classes of processes:

$$\lim_{J\to+\infty} \sum_p \mathrm{var}(S_J(p)F) = 0$$

$$\lim_{J\to+\infty} \|S_J F - S_J G\|^2 = \|E\{S_J F\} - E\{S_J G\}\|^2 \, .$$



Computational Complexity

If $f(n)$ is of size N then $S_J(p)f(n)$ is of size $N\,2^{-dJ}$

For K mother wavelets:

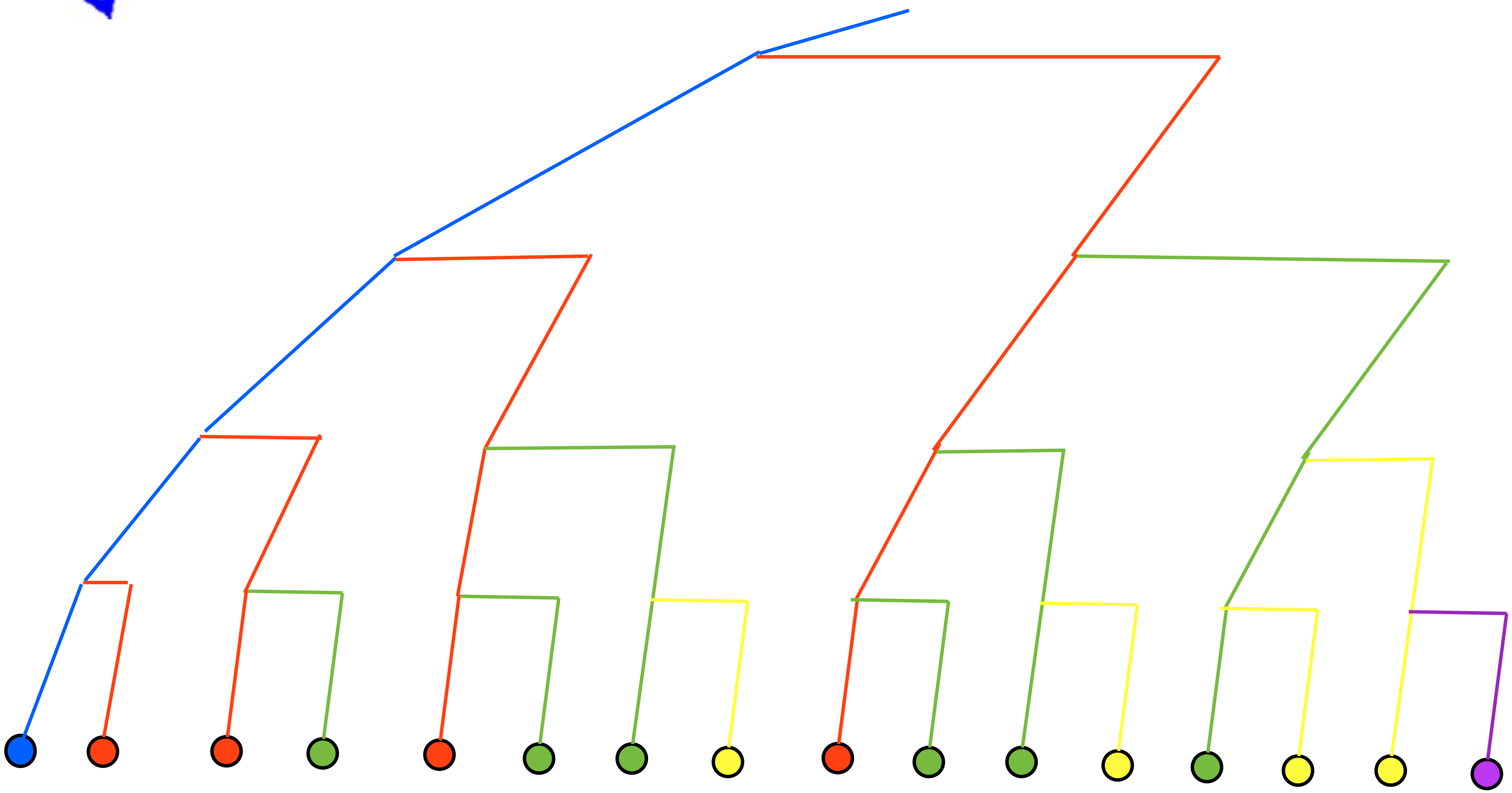
$O(K^m\,J^m)$ progressive paths of order $|p| \leq m$.

If $J = d^{-1}\log_2 N$ there are $O(d^{-m}\,K^m\,(\log_2 N)^m)$ coefficients.

For images experiments: $d = 2$, $K = 4$, $m = 3$.

Computations: $O(N)$.

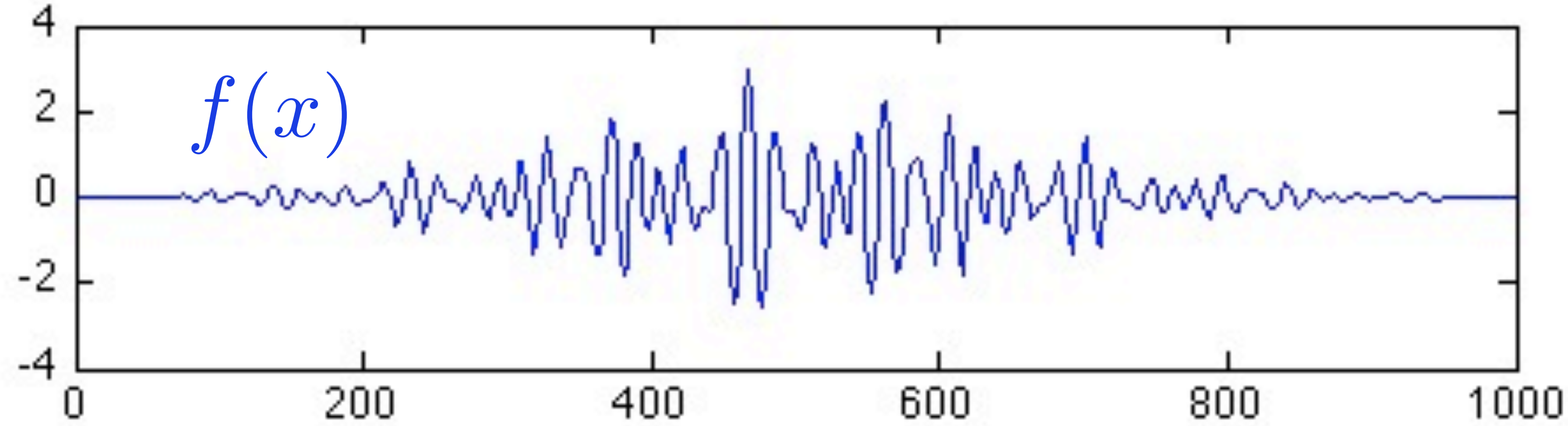
Path Ordering



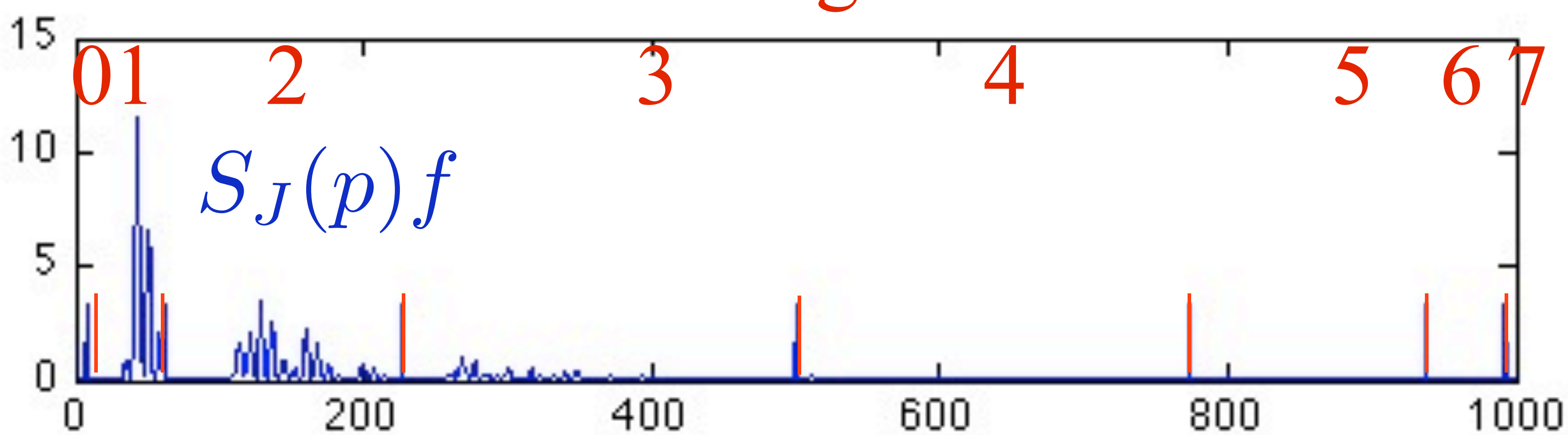
Scattering order reordering:

$|p| = 1$ $|p| = 2$ $|p| = 3$ $|p| = 4$

Musical Chord

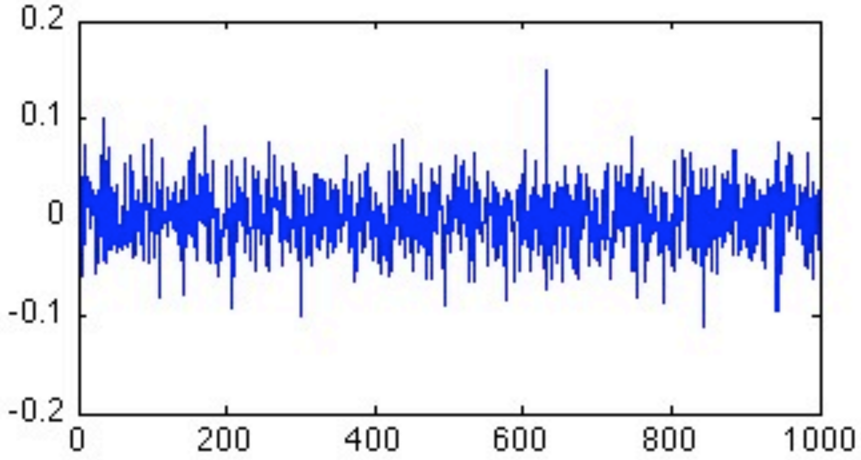


Scattering order

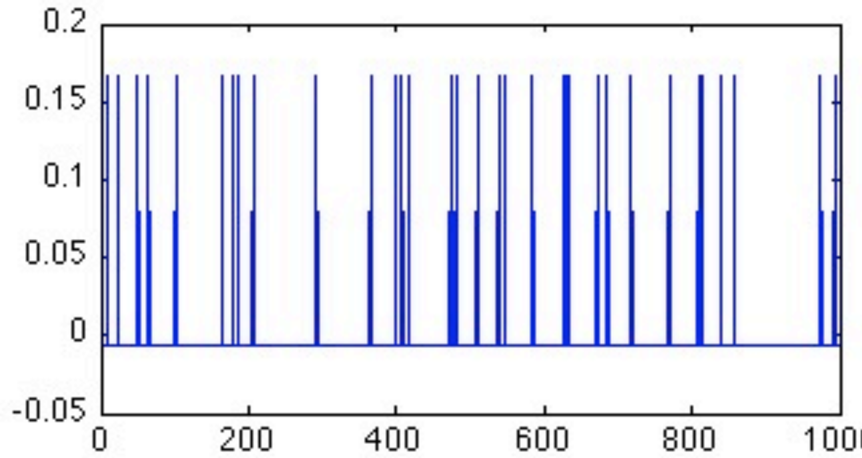


Gaussian White and Bernouilli

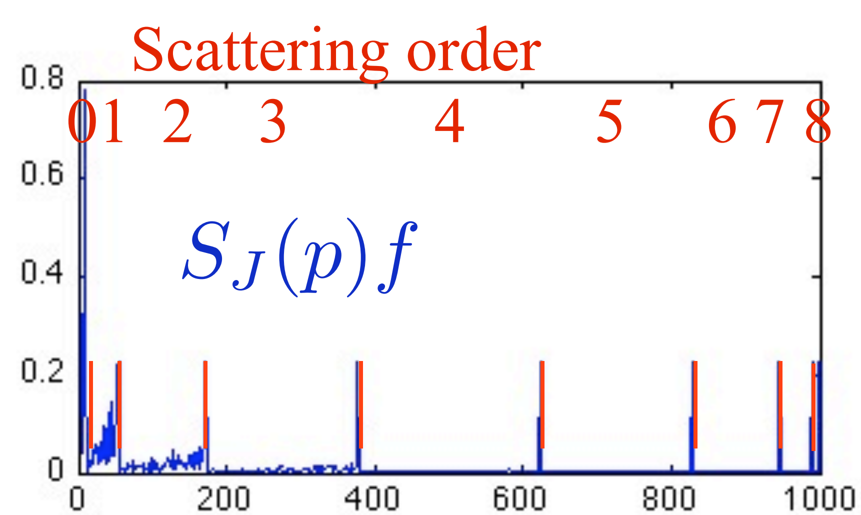
Gaussian White Noise

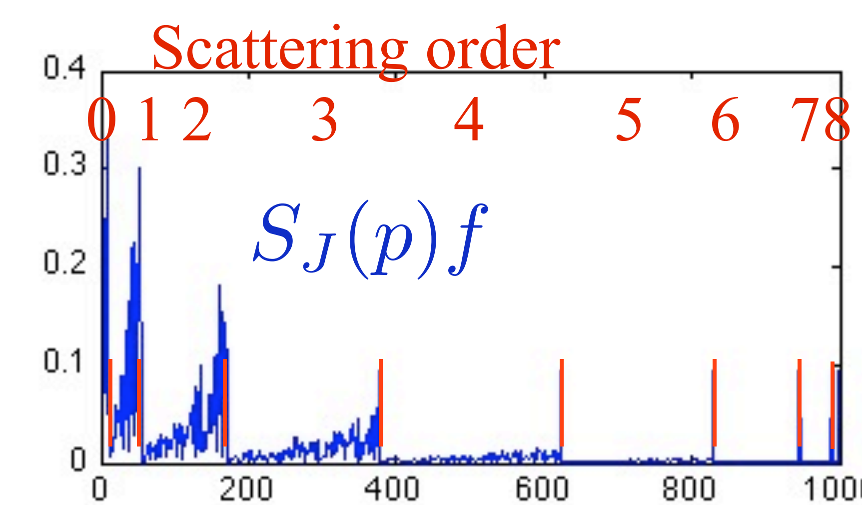


Bernouilli Process



$F(x)$





Classification of Deformed Processes

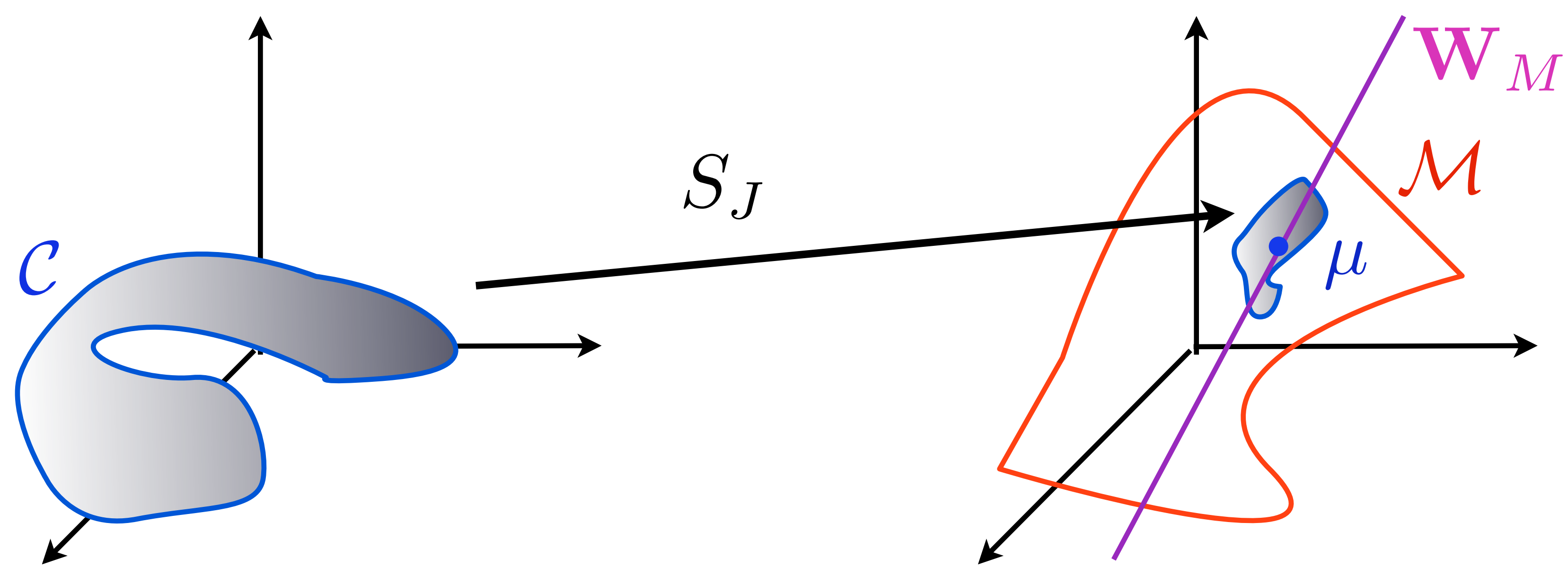
Joan Bruna

- Low dimensional affine space models in the scattering domain:
  - deformations are linearized: principal deformation directions
  - principal directions of residual variability for processes

- For realizations F of a class $\mathcal{C}$ let $\mu(p) = E\{S_J(p)F\}$,

the class distance $d(f, \mathcal{C}) = \|S_J f - \mu\|^2$ is replaced by

$$d(f, \mathcal{C}) = \|S_J f - \mu - P_{\mathbf{W_M}}(S_J f - \mu)\|^2$$



Affine Space Selection

- Affine space **learning** with PCA : O(training coefficients)
  - For each class $\mathcal{C}_k$: compute the mean μ_k and covariance Σ_k of $S_J f_n$ for all training signals $f_n \in \mathcal{C}_k$
  - Best approximation space $\mathbf{W}_{k,M}$ of dimension *M* : space generated by the *M* eigenvectors of largest eigenvalues.

- **Classification** by penalized estimation: $O(K\,(\log N)^m)$
  - The class of f is estimated by minimizing the class distance, penalized by its dimension:

$$k(f) = \arg\min_{1 \leq k \leq K} \min_{M} \Big(\| S_J f - \mu_k - P_{\mathbf{W}_{k,M}} (S_J f - \mu_k) \|^2 + \lambda\, M \Big)$$

  - Cross validation estimation of λ and J during learning.

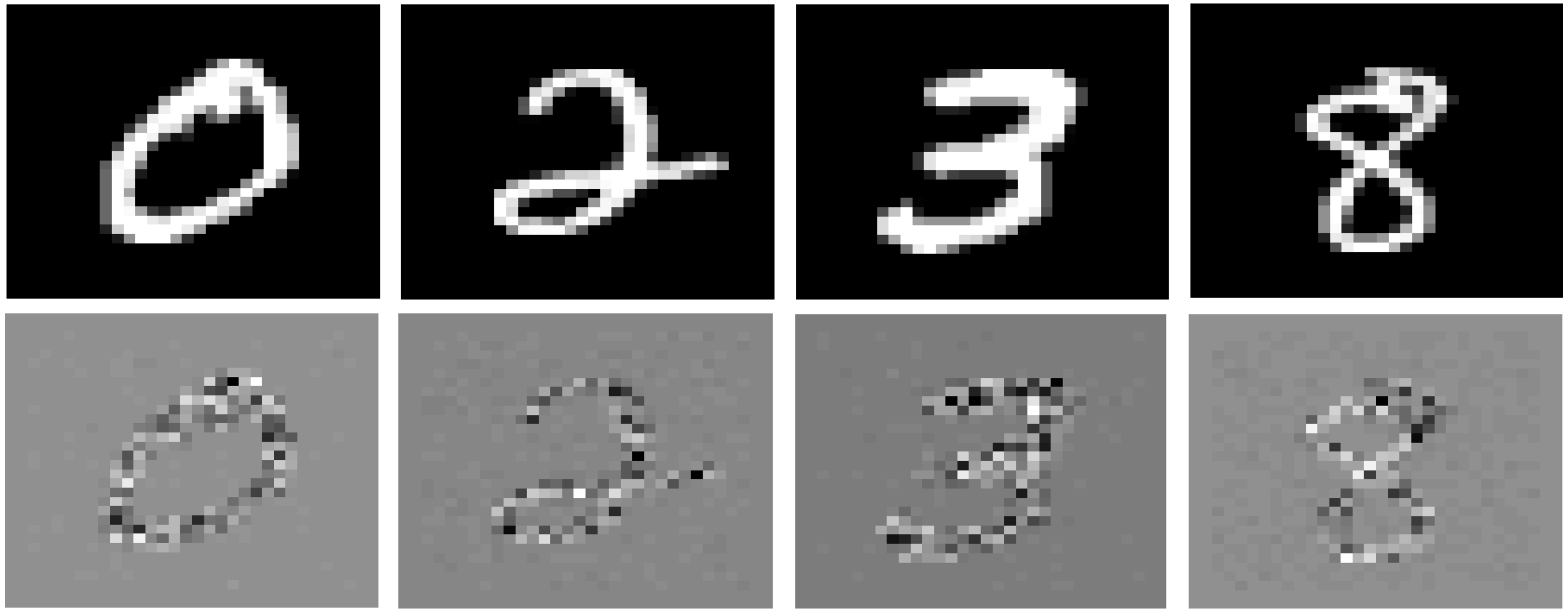
Digit Classification: MNIST

$S_J f$ with $J = 3$ and $m = 3$.

Training Size	SVM	Deep Net	Scattering
100	28%		15%
500	**12**%	**6.0** %	**3**.4%
1000	8.5%	3.21%	2.2%
5000	**4.2**%	**1.52** %	**1.3**%
10000	3.1%	0.85%	1.2%
30000	1.8%	0.7 %	0.85%
60000	**1.4**%	**0.64** %	**0.78**%

Training Size	SVM: N_s	Scattering M_{av}
100	45	9
500	**120**	**40**
1000	200	80
5000	500	100
10000	800	100
30000	1500	140
60000	**2000**	**160**

Textured Digit Classification

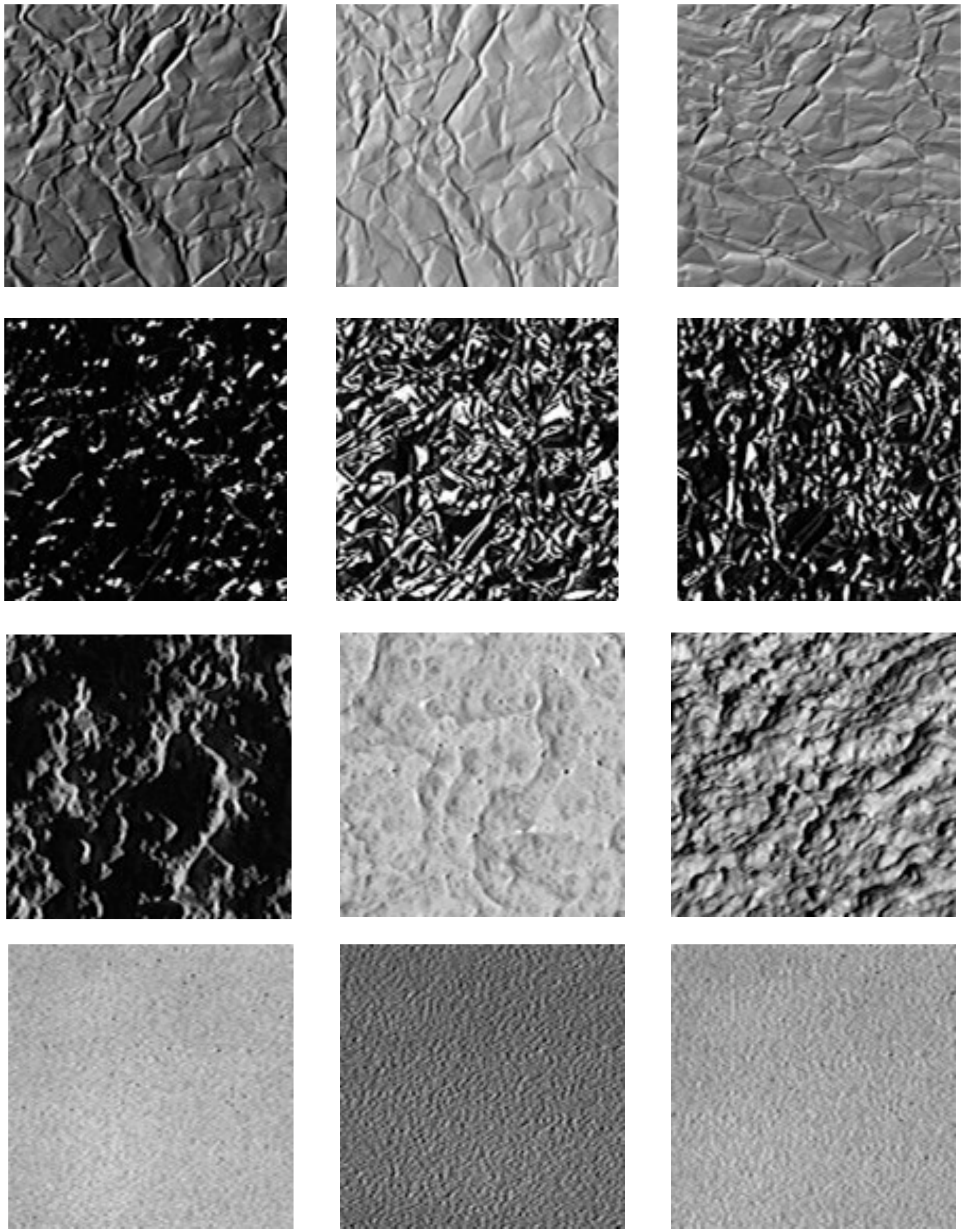


$S_J f$ with $J = 3$ and $m = 3$.

Training Size	SVM	Scattering
100	80%	41%
500	80%	23%
1000	80%	18%
5000	65%	10%
20000	-	8%

Training Size	SVM: N_s	Scattering M_{av}
100	500	10
500	500	40
1000	1000	70
5000	2000	160
20000	-	100

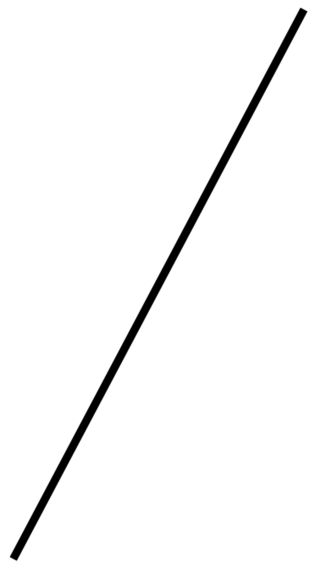
Classification of Textures

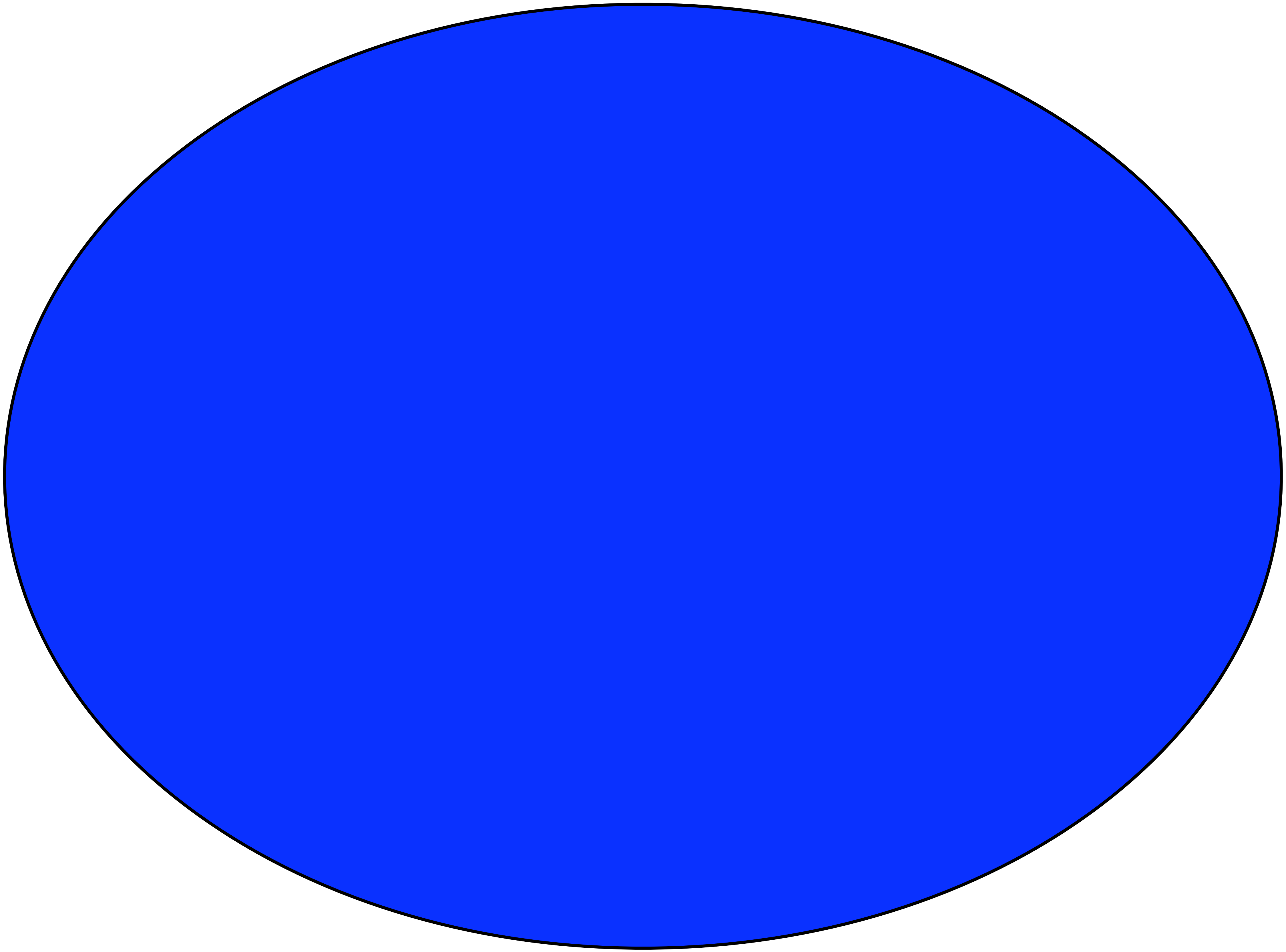


61 classes
Training size per class: 46
Testing size per class: 46

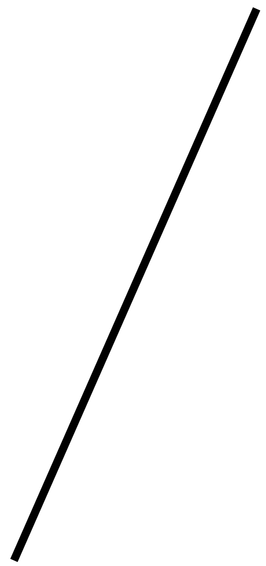
Malik (3D Textons): **5.35%**
Zisserman (MRF): **2.57%**
Scattering error: **0.4%**

General Group Invariance

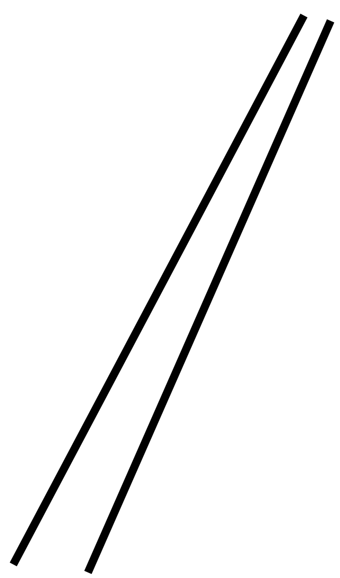




General Group Invariance



General Group Invariance



General Group Invariance

- Need invariance to other groups of deformations $\mathcal{G} = \{G_k\}_k$ rotations, scaling...

- Wavelets $\{\psi_k = G_k\,\psi\}_k$ dilated $\psi_{j,k}(x) = 2^{-dj}\psi_k(2^{-j}x)$

$$W_j f(k,x) = f \star \psi_{j,k}(x)$$

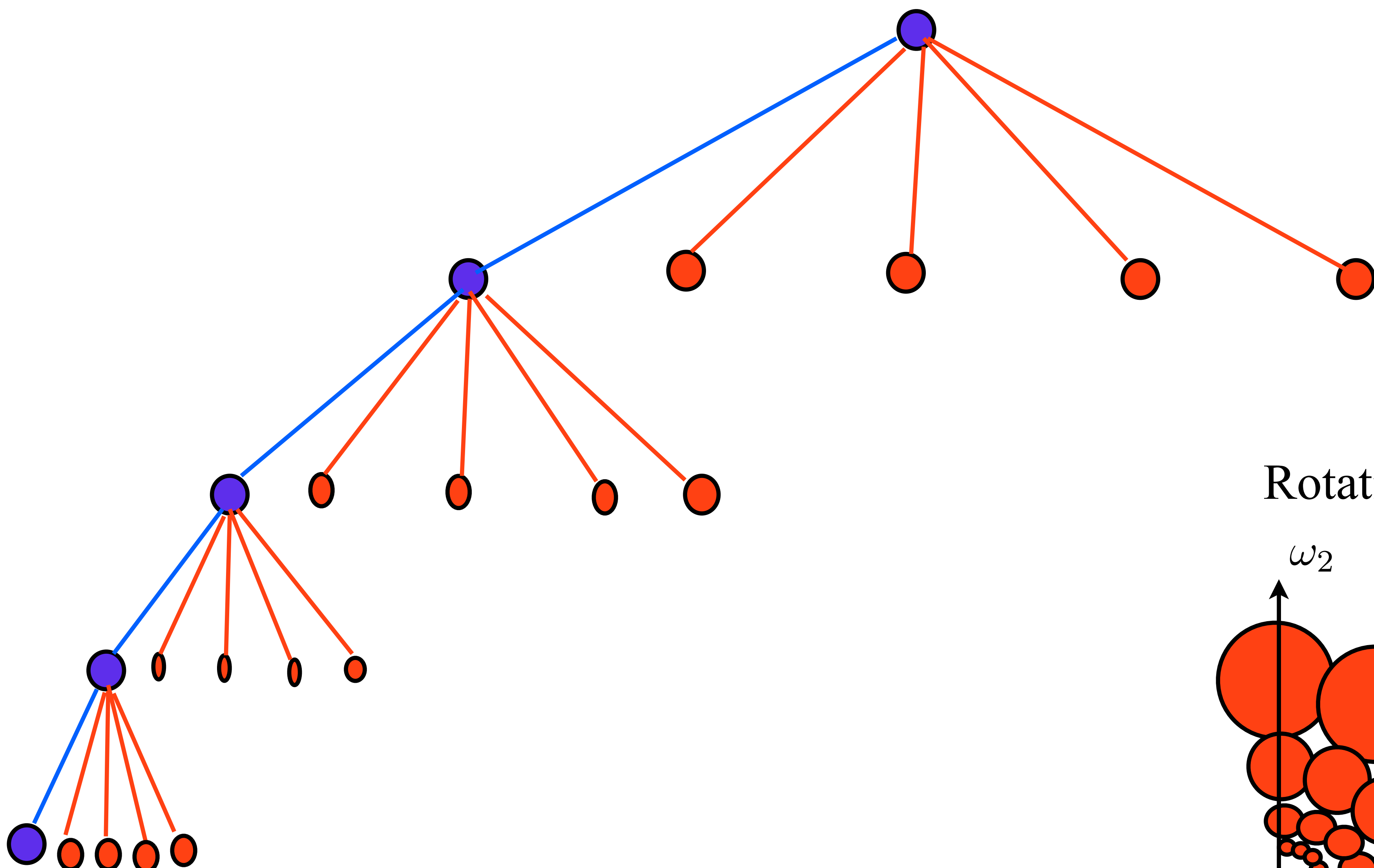
$$W_j(G_a f)(k,x) = G_a W_j f(k-a,x)\ .$$

- Invariance by invariant scattering along *k* .

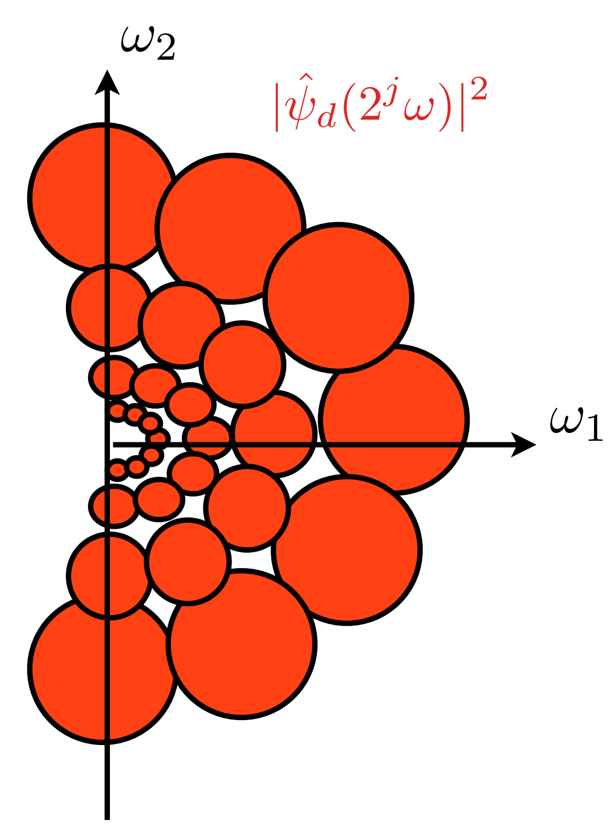
Group Interference Tree

Group transformed wavelets: $\{\psi_k = G_k\,\psi\}_k$

1st order

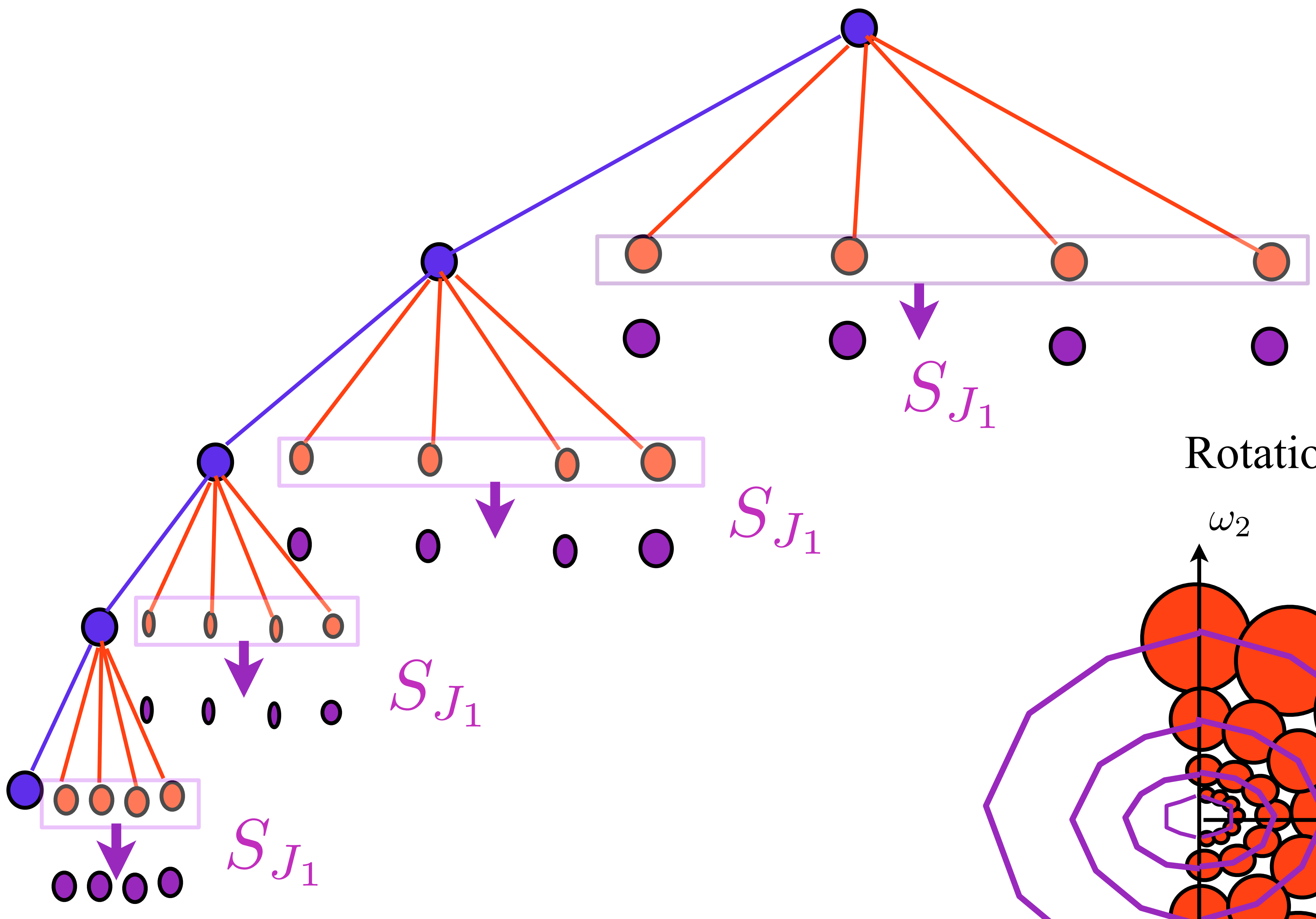


Rotation example

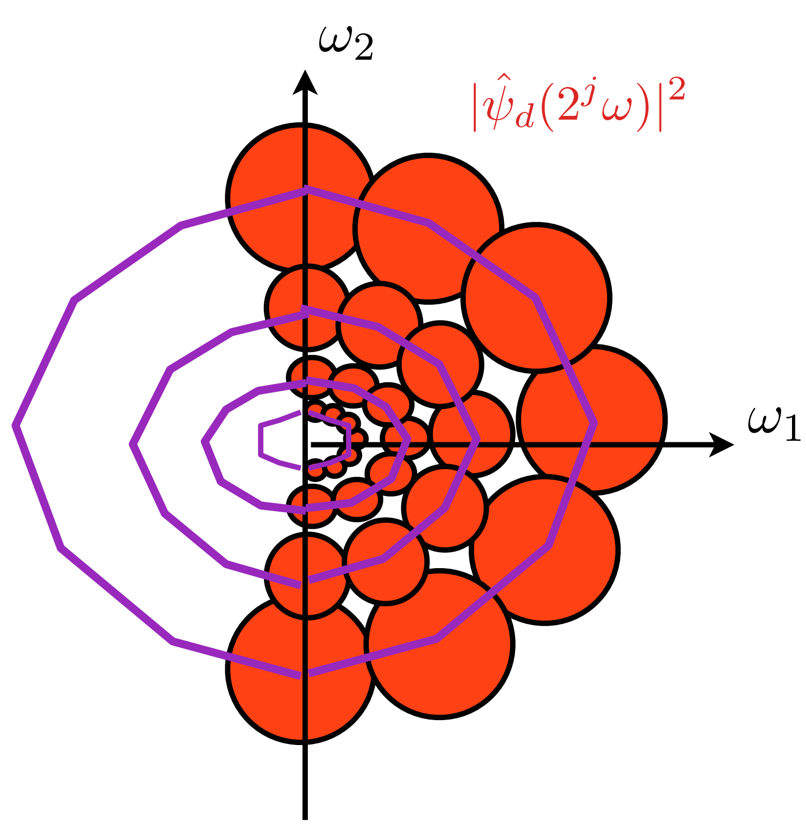


Group Interference Tree

Group transformed wavelets: $\{\psi_k = G_k \, \psi\}_k$

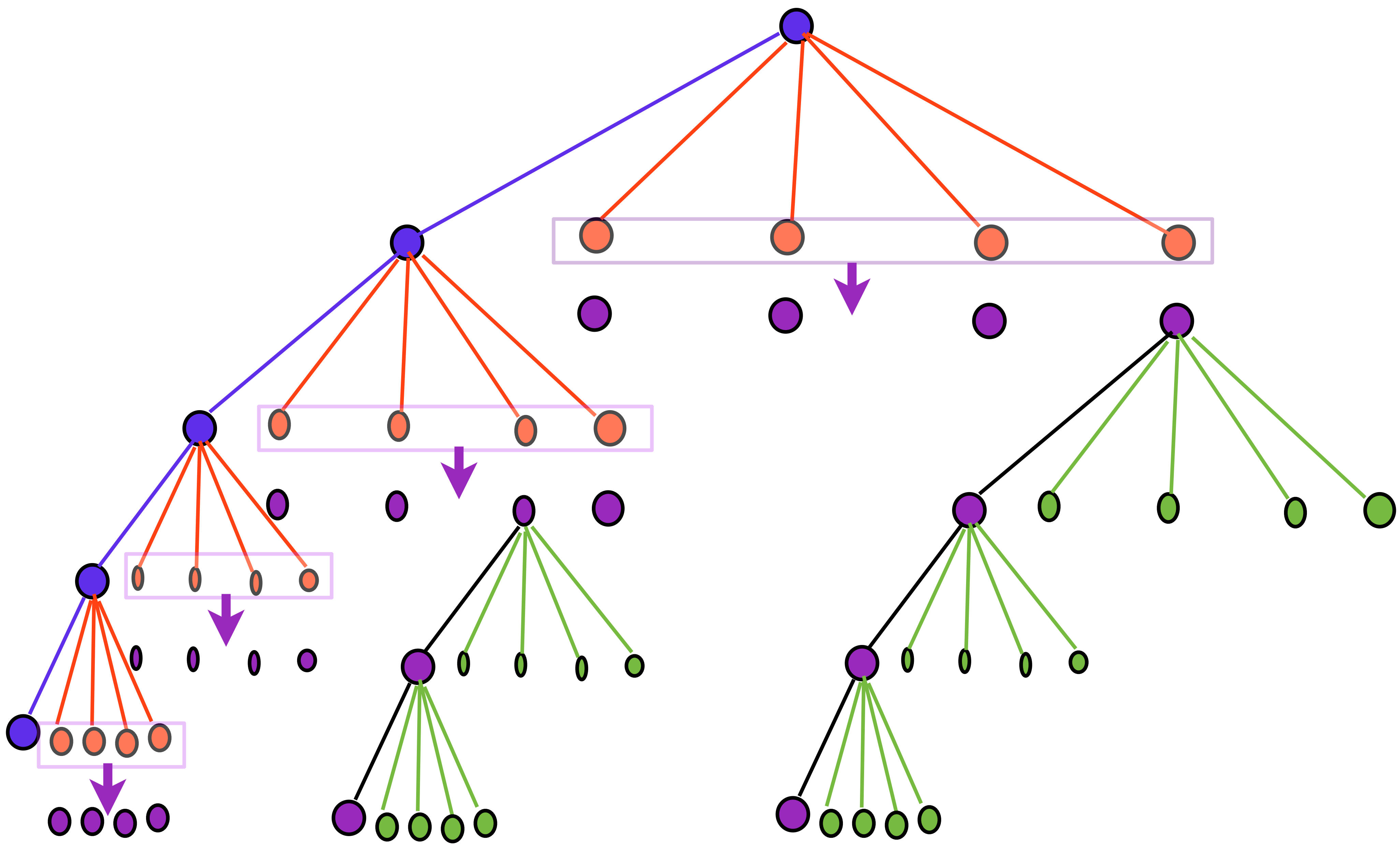


Rotation example



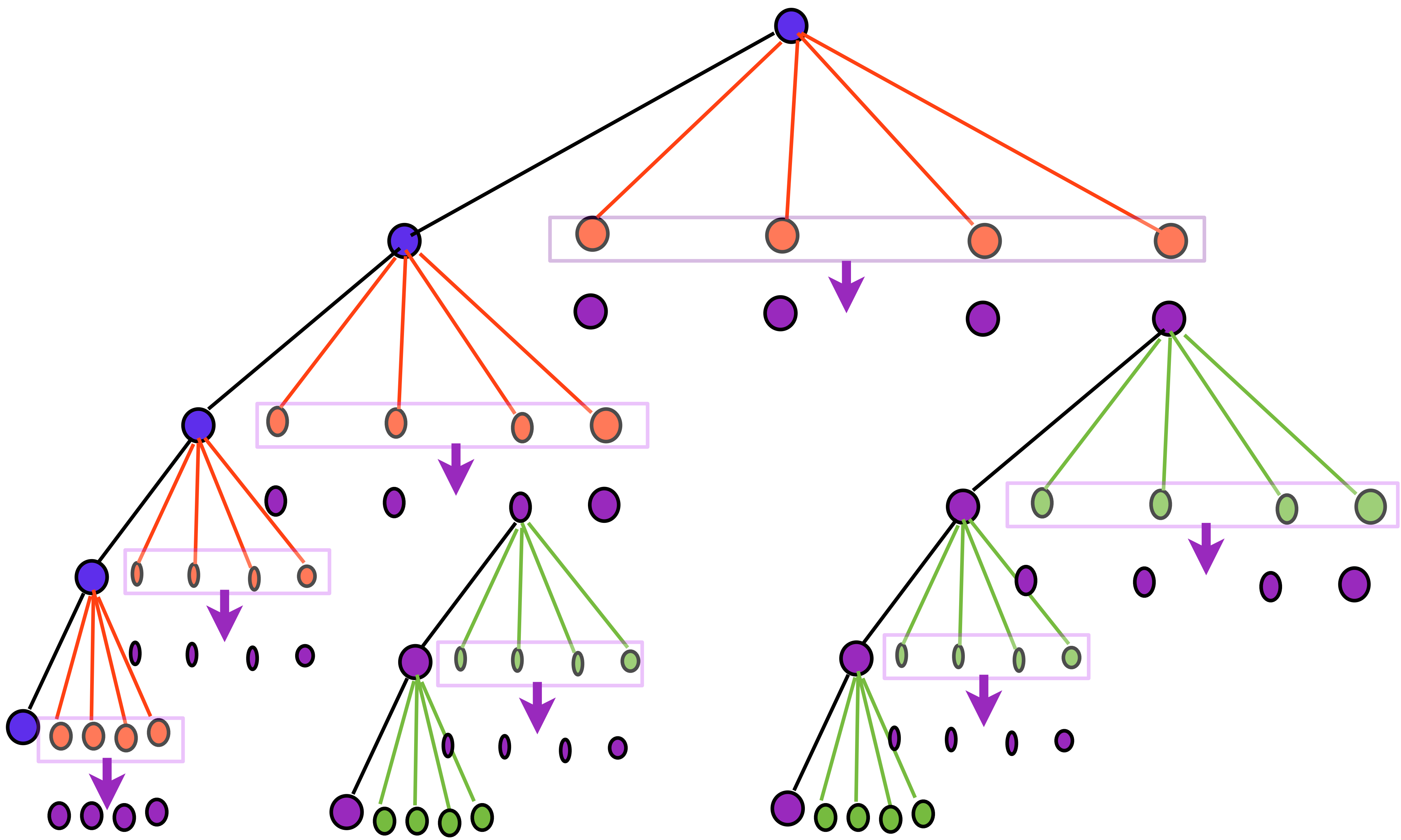
Group Interference Tree

Group transformed wavelets: $\{\psi_k = G_k\,\psi\}_k$



Group Interference Tree

Group transformed wavelets: $\{\psi_k = G_k\, \psi\}_k$



Conclusion

- The properties of invariant scattering come as a suprise, with many open questions:
- **Mathematics.**
  - Characterization of the metric on stationary processes
  - Dimensionality and «size» of the attractor manifold.
- **Applications.**
  - Image, audio (attacks) and generic classification
  - Estimation of deformations, mouvements...
  - Building and understand neural networks
  - Biological plausible models of complex cells for perception ?
  - Relations with quantum physics.