



# Advances in Multiscale Analysis and Applications

**Wael Mattar**

School of Mathematical Sciences, Tel Aviv University

*Seminar talk at the mathematical institute, University of Oxford*

# Outline

- 1 Multiscaling overview
- 2 Pseudo-reversing in Wiener algebra
- 3 Adaptations to nonlinear spaces
- 4 Numerical applications
- 5 Image generation

# Outline

- 1 Multiscaling overview
- 2 Pseudo-reversing in Wiener algebra
- 3 Adaptations to nonlinear spaces
- 4 Numerical applications
- 5 Image generation

# Motivation (curves in nonlinear spaces)

---

Gaussian measures in  
Wasserstein spaces.

---

Visualization of a trajectory in  
the special Euclidean group.

## Subdivision schemes

In multiscale transforms we use subdivision schemes as upsampling operators.

A subdivision scheme associated with a mask  $\{g_j\}_{j \in \mathbb{Z}}$  and  $R$  is a refinement operator defined by

$$S(c)_k = \sum_{j \in \mathbb{Z}} g_{k-2j} c_j; \quad k \in \mathbb{Z};$$

for any sequence  $c = \{c_j\}_{j \in \mathbb{Z}} \in R$ .

## Subdivision schemes (cont.)

A necessary condition for the convergence of a subdivision scheme  $S$  is the constant-reproduction property

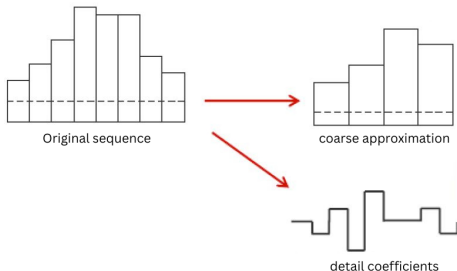
$$\sum_{j \in \mathbb{Z}} X_{2j} = \sum_{j \in \mathbb{Z}} X_{2j+1} = 1:$$

The refinement is called **interpolating** if  $\alpha_0 = 1$ . An essential tool for analyzing the convergence of  $S$  is the  $z$ -transform of  $\alpha$ , that is defined by

$$\alpha(z) = \sum_{j \in \mathbb{Z}} \alpha_j z^j; \quad z \in \mathbb{C};$$

and termed the **symbol**  $\{\alpha_j\}$  becoming a complex Fourier series on the unit circle  $T = \{z \in \mathbb{C} : |z| = 1\}$ .

# Interpolating decomposition



With an interpolating subdivision scheme  $S$ , a sequence  $\mathbf{c}^{(1)}$  associated with  $2^{-1}\mathbb{Z}$  can be decomposed into a **coarse** approximation  $\mathbf{c}^{(0)}$  and **detail** coefficients  $\mathbf{d}^{(1)}$  by

$$\mathbf{c}^{(0)} = D\mathbf{c}^{(1)}; \quad \mathbf{d}^{(1)} = \mathbf{c}^{(1)} - S\mathbf{c}^{(0)};$$

where  $(D\mathbf{c})_k = c_{2k}$ ,  $k \in \mathbb{Z}$  is the simple downsampling operator.

## Multiscaling via interpolating subdivision scheme

By construction it can be easily seen that  $d_{2^k}^{(1)} = 0$  for all  $k \in \mathbb{Z}$ , and that is a vital property for applications in multiscaling.



### Definition (multiscale transform)

The **multiscale transform** of a sequence  $\mathbf{c}^{(J)}$  associated with the grid  $2^{-J}\mathbb{Z}$ ,  $J \in \mathbb{N}$  is defined by

$$\mathbf{c}^{(\ell-1)} = D\mathbf{c}^{(\ell)}; \quad \mathbf{d}^{(\ell)} = \mathbf{c}^{(\ell)} - S\mathbf{c}^{(\ell-1)}; \quad \ell = 1; \dots; J;$$

while the **inverse multiscale transform** is given by

$$\mathbf{c}^{(\ell)} = S\mathbf{c}^{(\ell-1)} + \mathbf{d}^{(\ell)}; \quad \ell = 1; \dots; J;$$

## The non-interpolating case

Iterating the multiscale transform with a non-interpolating subdivision scheme  $S$  and the elementary downsampling operator  $D$  **does not(!)** necessarily give

$$d_{2k}^{(\cdot)} = 0; \quad k \in \mathbb{Z}; \quad \cdot = 1, \dots, J:$$

## The non-interpolating case

Iterating the multiscale transform with a non-interpolating subdivision scheme  $S$  and the elementary downsampling operator  $D$  **does not(!)** necessarily give

$$d_{2k}^{(\cdot)} = 0; \quad k \in \mathbb{Z}; \quad \cdot = 1, \dots, J;$$

We **solve** this problem by replacing  $D$  with the more general decimation operator. For a sequence  $c \in \ell_1(\mathbb{Z})$ , the operator

$$(Dc)_k = \sum_{j \in 2\mathbb{Z}} c_j$$

gives zero even detail coefficients in multiscaling if

$$(D\delta) =$$

where  $\delta$  is the Kronecker delta sequence.

# Reversing

Under the  $z$ -transform, the convolution equation becomes

$$(z)(D^{-1})(z) = 1; \quad z \in \mathbb{C}:$$

In the Banach algebra of continuous functions

$$A(T) = \left\{ f(t) = \sum_{j \in \mathbb{Z}} a_j e^{int}; \quad t \in [0; 2\pi) \right\} \quad \|a\|_1 < \infty;$$

if  $f \in A(T)$  does not vanish on  $T$ , then  $1/f(t)$  is also in  $A(T)$ .

## Reversing (cont.)

In case a reverse  $\mathcal{D}^{-1}(Z)$  exists, then there exist constants  $C(\epsilon) > 0$  and  $\delta(\epsilon) \in (0, 1)$  such that

$$|j - k| \leq C^{-1} |k|; \quad k \in Z;$$

where  $\delta$  is the reversibility condition number given by

$$= \frac{\sup_{z \in Z} |(\mathcal{D}^{-1})(z)|}{\inf_{z \in Z} |(\mathcal{D}^{-1})(z)|} \in [1; \infty):$$

## Reversing (cont.)

In case a reverse  $\mathcal{D}^{-1}(Z)$  exists, then there exist constants  $C(\epsilon) > 0$  and  $\delta(\epsilon) \in (0; 1)$  such that

$$|j - k| \leq C |j^k|; \quad k \in Z;$$

where  $\epsilon$  is the reversibility condition number given by

$$= \frac{\sup_{z \in Z} |(\mathcal{D}^{-1})(z)|}{\inf_{z \in Z} |(\mathcal{D}^{-1})(z)|} \in [1; \infty):$$

However, a reverse does not always exist! For example, least-squares subdivision schemes are usually **irreversible** !

# Outline

- 1 Multiscaling overview
- 2 Pseudo-reversing in Wiener algebra**
- 3 Adaptations to nonlinear spaces
- 4 Numerical applications
- 5 Image generation

## Pseudo-reversing

Let  $p(z)$  be a polynomial of degree  $n$  satisfying  $p(1) = 1$ , and denote by  $\mathcal{R}$  the set of its roots including multiplicities. By the complete factorization theorem we rewrite  $p$  as

$$p(z) = C(p) \prod_{r \in \mathcal{R}} (z - r)$$

### Definition (pseudo-reversing)

The pseudo-reverse of  $p$  is defined by

$$p^y(z) = C(p^y) \prod_{r \in \mathcal{R}} (z - r)^y \prod_{z \in \mathcal{R}} (1 + z)^{1-y}$$

for some  $y > 0$  where  $C(p^y)$  is chosen such that  $p^y(1) = 1$ .

## Example & properties

Example: Consider  $p(z) = (z^2 + z + 1) - 3$  which vanishes for  $z = 1 = 2$ . The pseudo-reverse of  $p$  is

$$p^y(z) = \frac{3 + 3z + z^2}{z^2 + z(1 + 2) + (1 + 2)^2}; \quad > 0:$$

Some properties:

- 1 The product  $p(z)p^y(z)$  converges in  $A$ -norm (the  $\ell_1$  norm of coefficients) to the constant 1 as  $\epsilon \rightarrow 0^+$ .
- 2 The function  $p^y(z)$  converges to 1 as  $\epsilon \rightarrow 1$  on every compact subset of  $\mathbb{C}$ , provided  $\epsilon \in T$ . Moreover,
- 3 the reversibility condition number of  $p^y(z)$  satisfies

$$(1 + 2\epsilon)^n:$$

## Illustration

Pseudo-reversing a least-squares-based subdivision scheme with  
 $\mathbf{M} = \frac{1}{12} \begin{bmatrix} 3 & 4 & 3 & 4 & 3 & 4 & 3 \end{bmatrix}$  supported on  $\mathbb{Z} \setminus \{3\}$ .

(a) Roots displacement

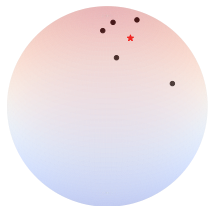
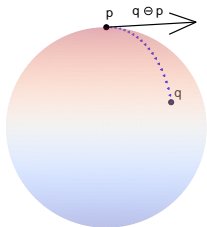
(b) Pseudo-reverse coefficients

# Outline

- 1 Multiscaling overview
- 2 Pseudo-reversing in Wiener algebra
- 3 Adaptations to nonlinear spaces**
- 4 Numerical applications
- 5 Image generation

# Adaptations to Riemannian manifolds

Let  $(M; \cdot)$  be a Riemannian manifold.



Adaptations of " " and "+":

$$q \quad p = \text{Log}_p(q) \in T_p M$$

$$p \quad v = \text{Exp}_p(v) \in M$$

$$\boxed{p \quad (q \ominus p) = q}$$

Euclidean CoM  $x = \frac{1}{n} \sum_{j=1}^n x_j$   
is generalized via:

$$x \in \text{argmin}_{x \in M} \sum_{j=1}^n d^2(x; x_j)$$

## Multiscaling in Wasserstein spaces

We adapt the multiscaling to the space of probability measures

$$P_p(\mathbb{R}^d) = \left\{ P \in \mathcal{P}(\mathbb{R}^d) \mid \int_{\mathbb{R}^d} \|x\|^p dP(x) < \infty \right\}$$

endowed with the Wasserstein metric  $W_p$  defined by

$$W_p(\mu; \nu) = \left( \min_{\gamma \in \Pi(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p d\gamma(x; y) \right)^{1/p}$$

where  $\Pi(\mu; \nu)$  is the set of all probability measures on  $\mathbb{R}^d \times \mathbb{R}^d$  with marginals  $\mu$  and  $\nu$ .

## Weighted averaging

McCann's interpolants can be utilized to define a weighted average.

The weighted average between  $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$  is defined by

$$M(\mu, \nu; t) = (\mu \# \nu)_t; \quad t \in [0; 1];$$

where  $\mu \# \nu$  is an optimal transport plan pushing  $\mu$  onto  $\nu$ , and the map  $\mu \# \nu : \mathbb{R}^d \rightarrow \mathbb{R}^d$  is given by  $(\mu \# \nu)(x; y) = (1 - t)x + ty$ .

McCann's interpolants are the only **constant-speed** geodesics in the metric space  $(\mathcal{P}_p(\mathbb{R}^d); W_p)$ .

# Operators on discrete measures (illustrations)

# Operators on discrete measures (illustrations)

# Operators on discrete measures (illustrations)

## Refinements in Wasserstein spaces

We focus on multiscaling measures with the most elementary subdivision scheme  $S$  that is given by the rules

$$\left( \begin{array}{l} (Sc)_{2k} = c_k; \\ (Sc)_{2k+1} = \frac{1}{2}c_k + \frac{1}{2}c_{k+1}; \end{array} \right. \quad k \in \mathbb{Z};$$

for a given  $\mathbb{R}$ -valued sequence  $c$ . Therefore, for a  $P_p(\mathbb{R}^d)$ -valued sequence  $\mu$ , the adaptation of  $S$  becomes

$$\left( \begin{array}{l} (S\mu)_{2k} = \mu_k; \\ (S\mu)_{2k+1} = M(\mu_k; \mu_{k+1}; 1=2); \end{array} \right. \quad k \in \mathbb{Z};$$

## The $\#$ and $\#_p$ operators

For A.C. measures  $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$  the difference is defined via

$$\mu \# \nu = T \# I;$$

where  $T$  is the optimal transport map, and  $I$  is the identity map. With this definition, we have that  $\mu \# 0 = 0$  the zero map, and

$$k \mu \# \nu = \int_{\mathbb{R}^d} k(T(x)) \nu(x) dx = W_p^p(\mu; \nu);$$

Moreover, the **compatible** operator is defined via

$$\mu \#_p \nu = (\mu + \nu) \#;$$

for any Borel measurable map  $T: \mathbb{R}^d \rightarrow \mathbb{R}^d$ .

## The $\otimes$ and $\otimes^*$ operators (for discrete measures)

For discrete probability measures  $\mu \in \mathcal{P}_p(\mathbb{R}^d)$  given by

$$\mu = \sum_{i=1}^m p_i \delta_{x_i} \quad \text{and} \quad \nu = \sum_{j=1}^n p_j \delta_{x_j};$$

the difference is defined via

$$\mu \otimes \nu = \sum_{i=1, \dots, m} \sum_{j=1, \dots, n} p_i p_j \delta_{x_i + x_j};$$

where  $C$  is the coupling matrix between  $\mu$  and  $\nu$ . In this case, the **compatible** operator is defined via

$$\mu \otimes^* \nu = \sum_{i=1}^m \sum_{j=1}^n p_i p_j C_{ij} \delta_{x_i + x_j};$$

for any tensor pair  $(\mu, \nu)$  where  $\mu \in \mathcal{P}_p(\mathbb{R}^m)$  and  $\nu \in \mathcal{P}_p(\mathbb{R}^k)$ .

## Theoretical results

Multisampling a sequence  $(f_j)$  in  $P_p(\mathbb{R}^d)$  with the above notations via the interpolating elementary  $S$  becomes straightforward.

### Theorem (decay of detail coefficients)

Assume  $(f_j)$  is sampled over the dyadic grid parametrization  $2^{-j}Z$  from an absolutely continuous curve  $f$  in  $P_p(\mathbb{R}^d)$  with a finite metric derivative  $\|f\|_{L^p(\mathbb{R}^d)} = \sup_{t \in \mathbb{R}} |f'(t)| < 1$ . Then, the resulting detail coefficients  $(d_k^{(j)})$  satisfy

$$\|d_k^{(j)}\|_{k_1} \leq 2^{-j|k|}; \quad j = 1, \dots, J;$$

where

$$\|d_k^{(j)}\|_{k_1} = \sup_{k \in \mathbb{Z}} \|d_k^{(j)}\|_{L^p(\mathbb{R}^d)};$$

## Theoretical results

The refinement  $S$  is called **stable** if there exists  $K > 0$  such that

$$W_p(S; S) \leq K W_p(\cdot; \cdot)$$

for any sequences  $\{k_n\}$  and  $\{k'_n\}$  where  $W_p(\cdot; \cdot) = \sup_{k \in \mathbb{Z}} W_p(\cdot; k)$ .

Let  $f^{(0)}; f^{(1)}; \dots; f^{(J)}$  and  $e^{(0)}; e^{(1)}; \dots; e^{(J)}$  be two pyramid representations of  $f^{(J)}$  and  $e^{(J)}$ , respectively. Assume that  $\|f^{(j)}\|_{k, \text{Lip}} \leq C$  for all  $j = 1; \dots; J$  and  $k \in \mathbb{Z}$ . If  $S$  is stable with constant  $K$ , then for  $L = \max\{1; (KC)^J\}$  we have

$$W_p(f^{(J)}; e^{(J)}) \leq L W_p(f^{(0)}; e^{(0)}) + \sum_{j=1}^J \|f^{(j)}\|_{k_1} \|e^{(j)}\|_{k_1}$$

# Outline

- 1 Multiscaling overview
- 2 Pseudo-reversing in Wiener algebra
- 3 Adaptations to nonlinear spaces
- 4 Numerical applications**
- 5 Image generation

# Application 1: Denoising

Multiscaling A.C. measures in the Wasserstein space  $P_2(\mathbb{R})$ .

Figure: Sequence of Gaussian measures and its multiscale transform.

# Application 1: Denoising

Multiscaling A.C. measures in the Wasserstein space  $P_2(\mathbb{R})$ .

Figure: Contamination with noise.

# Application 1: Denoising

Multiscaling A.C. measures in the Wasserstein space  $P_2(\mathbb{R})$ .

Figure: Denoising result obtained by thresholding with  $\epsilon = 0.01$ .

## Application 2: Anomaly detection

Multiscaling A.C. measures in the Wasserstein space  $P_2(\mathbb{R})$ .

Figure: Detecting jump discontinuities via multiscaling.

## Application 3: Analyzing NN learning dynamics

Multiscaling discrete measures in the Wasserstein space  $\mathcal{P}_2(\mathbb{R})$ .

**Figure:** Analyzing the learning dynamics of a simple neural network on MNIST dataset. On the left, the prediction of the digit \ 3" across epoch iterations. On the right, the multiscale transform of the resulted measure sequence. The convergence is clear on coarse scales.

## Application 4: Contrast enhancement

The manifold of interest is  $M = \text{SO}(3)$ .

**Figure:** Contrast enhancement of a sequence of rotation matrices. On the left, illustration of the original sequence applied to the standard basis of  $\mathbb{R}^3$ . On the right, the enhanced sequence obtained by scaling the top 20% of detail coefficients with a factor of 40%.

## Application 5: Data compression

The manifold of interest is  $M = \text{SO}(3) \times \mathbb{R}^3$ .

**Figure:** Data compression of a rigid body motion. On the left, illustration of 641 matrices representing the special Euclidean sequence. On the right, the decompressed trajectory that is obtained by 41 matrices in addition to 12 detail coefficients.

# Outline

- 1 Multiscaling overview
- 2 Pseudo-reversing in Wiener algebra
- 3 Adaptations to nonlinear spaces
- 4 Numerical applications
- 5 Image generation**

## Bivariate multiscale transform

Multivariate multiscale transforms can be constructed by applying tensor products to  $\mathcal{H}_1$  and  $\mathcal{H}_2$ .

Figure: Sketch for decomposing a  $4 \times 4$  grayscale image.

# Multiscaling an image

Figure: Multiscaling an image of the author's own brain (MRI).

# Motivation

We take a new approach to autoregressive image generation that is based on two main ingredients:

- 1 Wavelet image coding, and
- 2 an LLM transformer with a re-designed architecture.

Wavelets are all you need for autoregressive image generation.

Mattar W, Levy I, Sharon N, Dekel S. Pure and Functional Analysis.

# Wavelet scanning order

For images we scan wavelet coefficients in the following pattern

## Embedded wavelet tokenization

We describe how to tokenize wavelet images with **only 7 tokens**.

Let  $f \in [0; 1]^{M \times M}$  be an image, and assume  $M = 2^m$  for some  $m \in \mathbb{N}$ .

Denote by  $\{j\}$  its MRA decomposition. Namely,

$$\{j\}_{(i_1; i_2)} = \left( \begin{array}{l} hf; e_{m; (i_1; i_2)} i; \quad 1 \quad i_1; i_2 \quad 2; \\ hf; e_{j; k} i; \quad 3 \quad i_1; i_2 \quad M; \quad 1 \quad j \quad m; \quad e = 1; 2; 3; \end{array} \right)$$

After  $m - 1$  iterations of the bivariate DWT we have

$$\max_{(i_1; i_2)} |j\}_{(i_1; i_2)}| \leq 2^{m-1};$$

Now, compute  $\alpha = \log_2 \max_{(i_1; i_2)} |j\}_{(i_1; i_2)}|$  and initialize the first threshold  $T = 2^{\alpha - 1}$  for the beginning of the tokenization.

## First bit-plane

Beginning with the initial bit-plane  $[T=2; T]$ :

If  $j!_{(i_1; i_2)} j \geq T=2$  ) the coefficient will be reported with '**NowSignificantPos**' or '**NowSignificantNeg**' and gets the approximation value  $3T=4$  depending on its sign.

If  $j!_{(i_1; i_2)} j < T=2$  ) the coefficient will be reported as '**Insignificant**' and gets the approximation value 0.

## Next Accuracy bits

Next, the bit-plane is updated to  $[T=4; T=2]$ :

Previously reported significant coefficients get either `'NextAccuracy0'` or `'NextAccuracy1'` depending on the comparison between their true value and the encoded approximation. Their App. will be updated with  $T=8$ .

If  $T=4$   $j \neq (i_1; i_2)$   $T=2$  ) the coefficient will be reported with `'NowSignificantPos'` or `'NowSignificantNeg'` and gets the approximation value  $3T=8$  depending on its sign.

## Zero blocks

We introduce two additional tokens to shorten the tokenization:

The token `'Group2x2'` replaces each square block of 4 `'Insignicant'` tokens falling in the same subband if the top left position indices are divisible by 2.

The token `'Group4x4'` replaces each square block of 16 `'Insignicant'` tokens falling in the same subband if the top left position indices are divisible by 4.

The token sequences are then concatenated in the natural order of the bit-planes. Let us illustrate the process visually.

# Tokenization process (illustration)

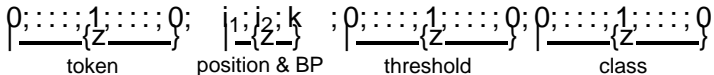
## Dataset structure

The resulting dataframe of the MNIST dataset tokenization with the Haar wavelet. The tokenization terminated at  $T = 2^3$ .

## Architecture modification

The transformer architecture is modified as follows:

The classical positional encoding is cancelled. Instead, at the transformer's embedding layer, we concatenate the one-shot vector representations of the tokens with the corresponding positions, and then with the threshold and the image class id.



## Conditional inference

We impose conditions on the acceptance of the next predicted token. In particular, at any given position,

`NowSigni cant' tokens can only be preceded by `Insigni cant' or `Group' tokens and followed by `NextAccuracy' tokens.

`NextAccuracy' tokens can only be preceded by `NextAccuracy' or `NowSigni cant' tokens.

`Insigni cant' can only be preceded by either `Insigni cant' or `Group' tokens.

`Group2x2' can only be preceded by `Group' tokens.

`Group4x4' can only be preceded by a `Group4x4' token.

## Empirical results

Some results of the Wavelet Generative Transformer trained on the MNIST dataset, with Haar wavelet tokenization and  $2^{-3}$  minimal threshold.

# Empirical results

Some results of the Wavelet Generative Transformer trained on the FashionMNIST dataset, with bior4.4 wavelet tokenization and  $2^4$  minimal threshold.

## Empirical results

Two separate DistilGPT2 models were trained on the two datasets. We used an NVIDIA A100 GPU with 80GB; MNIST occupied 22GB while FashionMNIST occupied 61GB. Both models were trained for few days. Here are some extra results.

