Advances in Multiscale Analysis and Applications

Wael Mattar

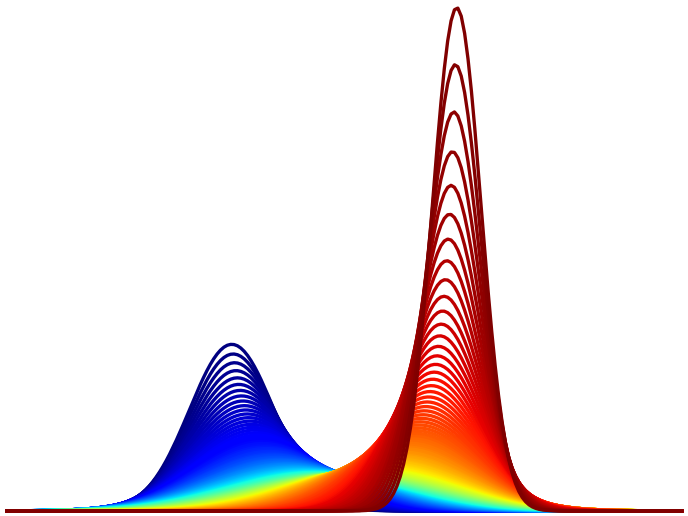
School of Mathematical Sciences, Tel Aviv University

Seminar talk at the mathematical institute, University of Oxford

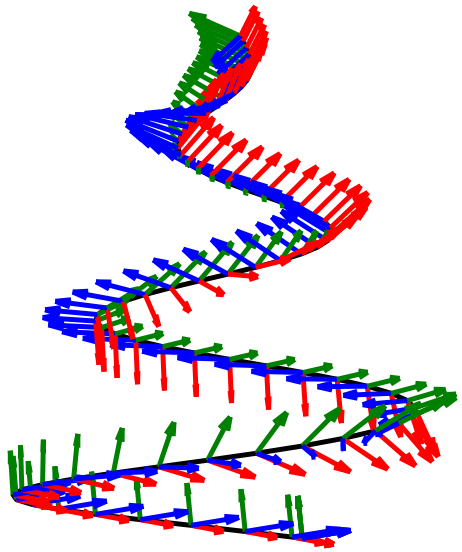
Outline

Outline

Motivation (curves in nonlinear spaces)



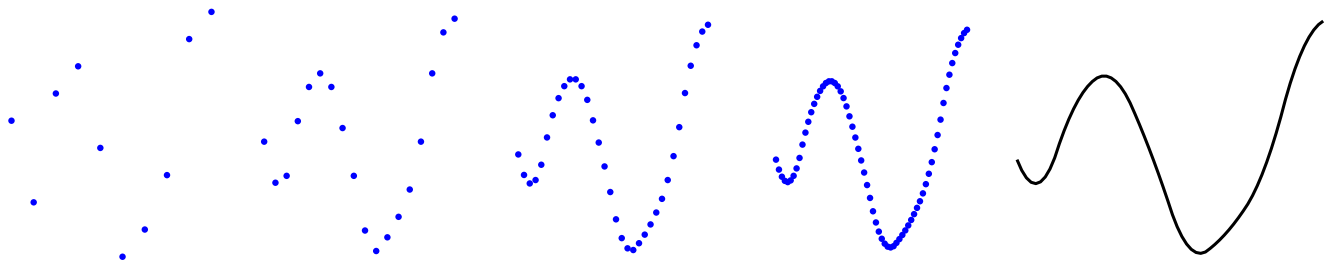
Gaussian measures in Wasserstein spaces.



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 $\boldsymbol{\alpha} = \{\alpha_j\}_{j\in\mathbb{Z}} \subset \mathbb{R}$ is a refinement operator defined by

$$\mathcal{S}_{\boldsymbol{\alpha}}(\boldsymbol{c})_k = \sum_{j\in\mathbb{Z}} \alpha_{k-2j} c_j, \quad k \in \mathbb{Z},$$

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

Subdivision schemes (cont.)

A necessary condition for the convergence of a subdivision scheme $\mathcal{S}_{\boldsymbol{\alpha}}$ is the constant-reproduction property

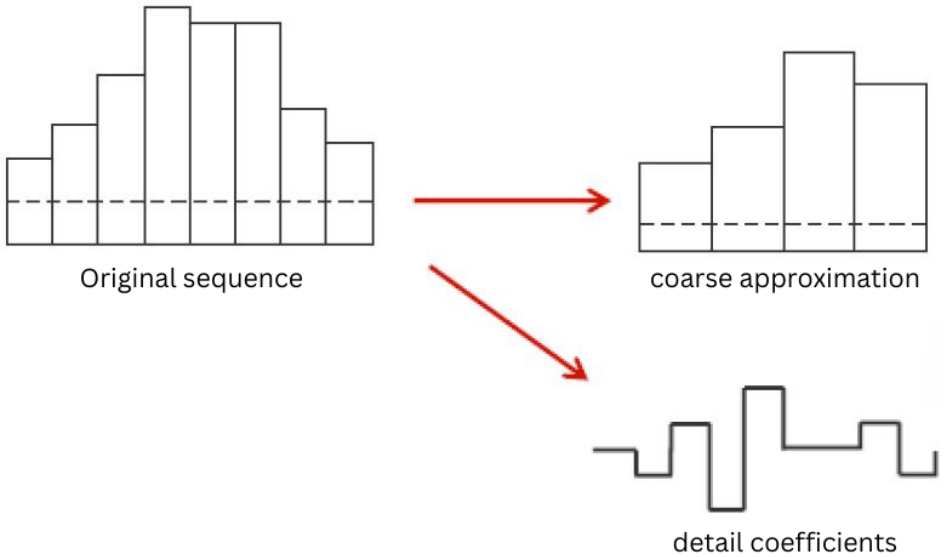
$$\sum_{j\in\mathbb{Z}} \alpha_{2j} = \sum_{j\in\mathbb{Z}} \alpha_{2j+1} = 1.$$

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

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

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

Interpolating decomposition



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

$$\boldsymbol{c}^{(0)} = \mathcal{D}\boldsymbol{c}^{(1)}, \quad \boldsymbol{d}^{(1)} = \boldsymbol{c}^{(1)} - \mathcal{S}_\alpha \boldsymbol{c}^{(0)},$$

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

Multiscaling via interpolating subdivision scheme

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

Sequence $\boldsymbol{c}^{(J)}$ $\underset{\text{reconstruction}}{\overset{\text{decomposition}}{\rightleftharpoons}}$ Pyramid $\left\{\boldsymbol{c}^{(0)}; \boldsymbol{d}^{(1)}, \boldsymbol{d}^{(2)}, \ldots, \boldsymbol{d}^{(J)}\right\}$

Definition (multiscale transform)

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

$$\boldsymbol{c}^{(\ell-1)} = \mathcal{D}\boldsymbol{c}^{(\ell)}, \quad \boldsymbol{d}^{(\ell)} = \boldsymbol{c}^{(\ell)} - \mathcal{S}_\alpha \boldsymbol{c}^{(\ell-1)}, \quad \ell = 1, \ldots, J,$$

while the inverse multiscale transform is given by

$$\boldsymbol{c}^{(\ell)} = \mathcal{S}_\alpha \boldsymbol{c}^{(\ell-1)} + \boldsymbol{d}^{(\ell)}, \quad \ell = 1, \ldots, J.$$

The non-interpolating case

Iterating the multiscale transform with a non-interpolating subdivision scheme $\mathcal{S}_\alpha$ and the elementary downsampling operator $\mathcal{D}$ does not(!) necessarily give

$$d_{2k}^{(\ell)} = 0, \quad k \in \mathbb{Z}, \quad \ell = 1, \ldots, J.$$

The non-interpolating case

Iterating the multiscale transform with a non-interpolating subdivision scheme $\mathcal{S}_{\boldsymbol{\alpha}}$ and the elementary downsampling operator $\mathcal{D}$ does not(!) necessarily give

$$d_{2k}^{(\ell)} = 0, \quad k \in \mathbb{Z}, \quad \ell = 1, \ldots, J.$$

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

$$(\mathcal{D}_{\boldsymbol{\gamma}} \boldsymbol{c})_k = \sum_{j \in \mathbb{Z}} \gamma_{k-j} c_{2j}$$

gives zero even detail coefficients in multiscaling if

$$\boldsymbol{\gamma} * (\mathcal{D}\boldsymbol{\alpha}) = \boldsymbol{\delta}$$

where $\boldsymbol{\delta}$ is the Kronecker delta sequence.

Reversing

Under the z-transform, the convolution equation becomes

$$\gamma(z)(\mathcal{D}\alpha)(z) = 1, \quad z \in \mathbb{C}.$$

Wiener's lemma

In the Banach algebra of continuous functions

$$\mathcal{A}(\mathbb{T}) = \left\{ f(t) = \sum_{j\in\mathbb{Z}} a_j e^{int},\ t \in [0, 2\pi) \mid \boldsymbol{a} \in \ell_1(\mathbb{Z}) \right\},$$

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

Reversing (cont.)

In case a reverse $\gamma \in \ell_1(\mathbb{Z})$ exists, then there exist constants $C(\kappa) > 0$ and $\lambda(\kappa) \in (0, 1)$ such that

$$|\gamma_k| \leq C\lambda^{|k|}, \quad k \in \mathbb{Z},$$

where κ is the reversibility condition number given by

$$\kappa = \frac{\sup_{z\in\mathbb{T}} |(\mathcal{D}\boldsymbol{\alpha})(z)|}{\inf_{z\in\mathbb{T}} |(\mathcal{D}\boldsymbol{\alpha})(z)|} \in [1, \infty].$$

Reversing (cont.)

In case a reverse $\gamma \in \ell_1(\mathbb{Z})$ exists, then there exist constants $C(\kappa) > 0$ and $\lambda(\kappa) \in (0, 1)$ such that

$$|\gamma_k| \leq C\lambda^{|k|}, \quad k \in \mathbb{Z},$$

where κ is the reversibility condition number given by

$$\kappa = \frac{\sup_{z\in\mathbb{T}} |(\mathcal{D}\boldsymbol{\alpha})(z)|}{\inf_{z\in\mathbb{T}} |(\mathcal{D}\boldsymbol{\alpha})(z)|} \in [1, \infty].$$

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

Outline

Pseudo-reversing

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

$$p(z) = C(p) \prod_{r \in \Lambda} (z - r).$$

Definition (pseudo-reversing)

The pseudo-reverse of p is defined by

$$p_\xi^\dagger(z) = \left(C(p_\xi^\dagger) \prod_{r \in \Lambda \setminus \mathbb{T}} (z - r) \prod_{r \in \Lambda \cap \mathbb{T}} \big(z - (1+\xi)r\big) \right)^{-1}$$

for some $\xi > 0$ where $C(p_\xi^\dagger)$ is chosen such that $p_\xi^\dagger(1) = 1$.

Example & properties

Example: Consider $p(z) = (z^2 + z + 1)/3$ which vanishes for $z = 1/2 \pm i\sqrt{3}/2 \in \mathbb{T}$. The pseudo-reverse of p is

$$p_\xi^\dagger(z) = \frac{3 + 3\xi + \xi^2}{z^2 + z(1+\xi) + (1+\xi)^2}, \quad \xi > 0.$$

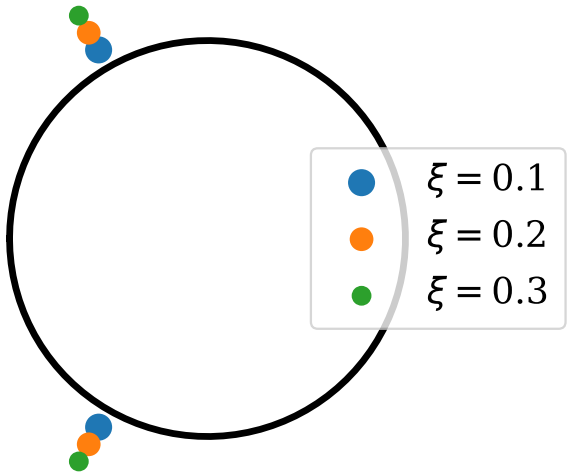
Some properties:

1. The product $p(z)p_\xi^\dagger(z)$ converges in $\mathcal{A}$-norm (the ℓ_1 norm of coefficients) to the constant 1 as $\xi \to 0^+$.
2. The function $p_\xi^\dagger(z)$ converges to 1 as $\xi \to \infty$ on every compact subset of $\mathbb{C}$, provided $\Lambda \subset \mathbb{T}$. Moreover,
3. the reversibility condition number κ of $p_\xi^{-\dagger}(z)$ satisfies

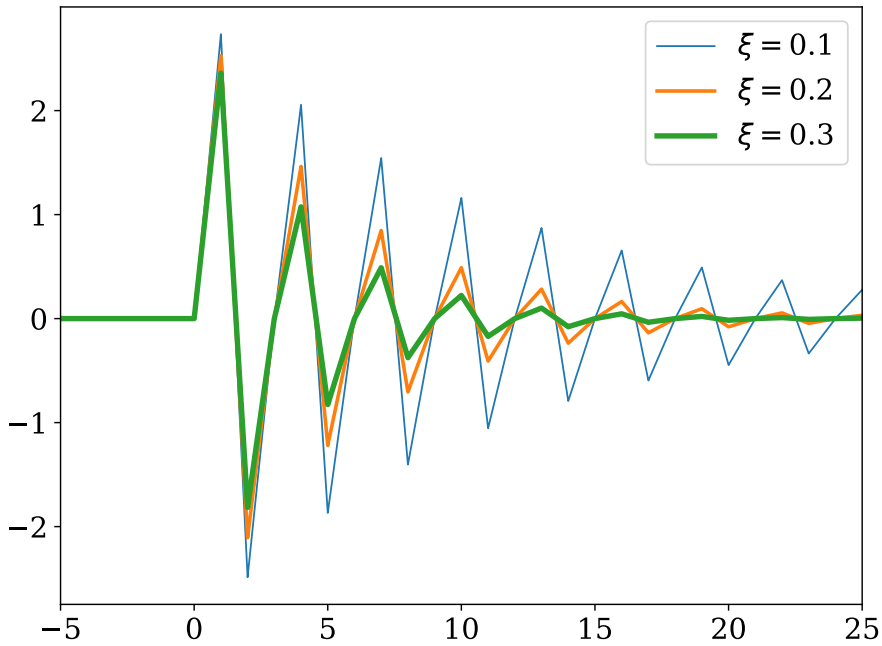
$$\kappa \leq (1 + 2/\xi)^n.$$

Illustration

Pseudo-reversing a least-squares-based subdivision scheme with $\boldsymbol{\alpha} = 1/12\ [3,\ 4,\ 3,\ 4,\ 3,\ 4,\ 3]$ supported on $[-3, 3] \cap \mathbb{Z}$.



(a) Roots displacement

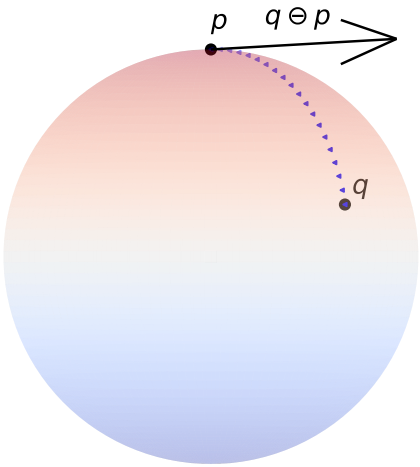


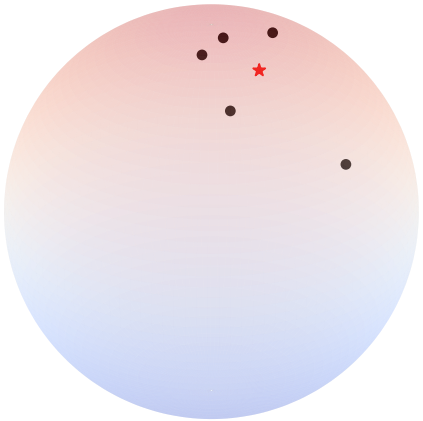
(b) Pseudo-reverse coefficients

Outline

Adaptations to Riemannian manifolds

Let $(\mathcal{M}, \rho)$ be a Riemannian manifold.





Adaptations of "$-$" and "$+$":

$$q \ominus p = \mathrm{Log}_p(q) \in T_p\mathcal{M}$$

$$p \oplus v = \mathrm{Exp}_p(v) \in \mathcal{M}$$

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

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

$$x^* \in \mathrm{argmin}_{x \in \mathcal{M}} \sum_{j=1}^n \beta_j \rho^2(x, x_j)$$

Multiscaling in Wasserstein spaces

We adapt the multiscaling to the space of probability measures

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

endowed with the Wasserstein metric W_p defined by

$$W_p^p(\mu,\nu) = \min_{\sigma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^p d\sigma(x,y)$$

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

Weighted averaging

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

Definition

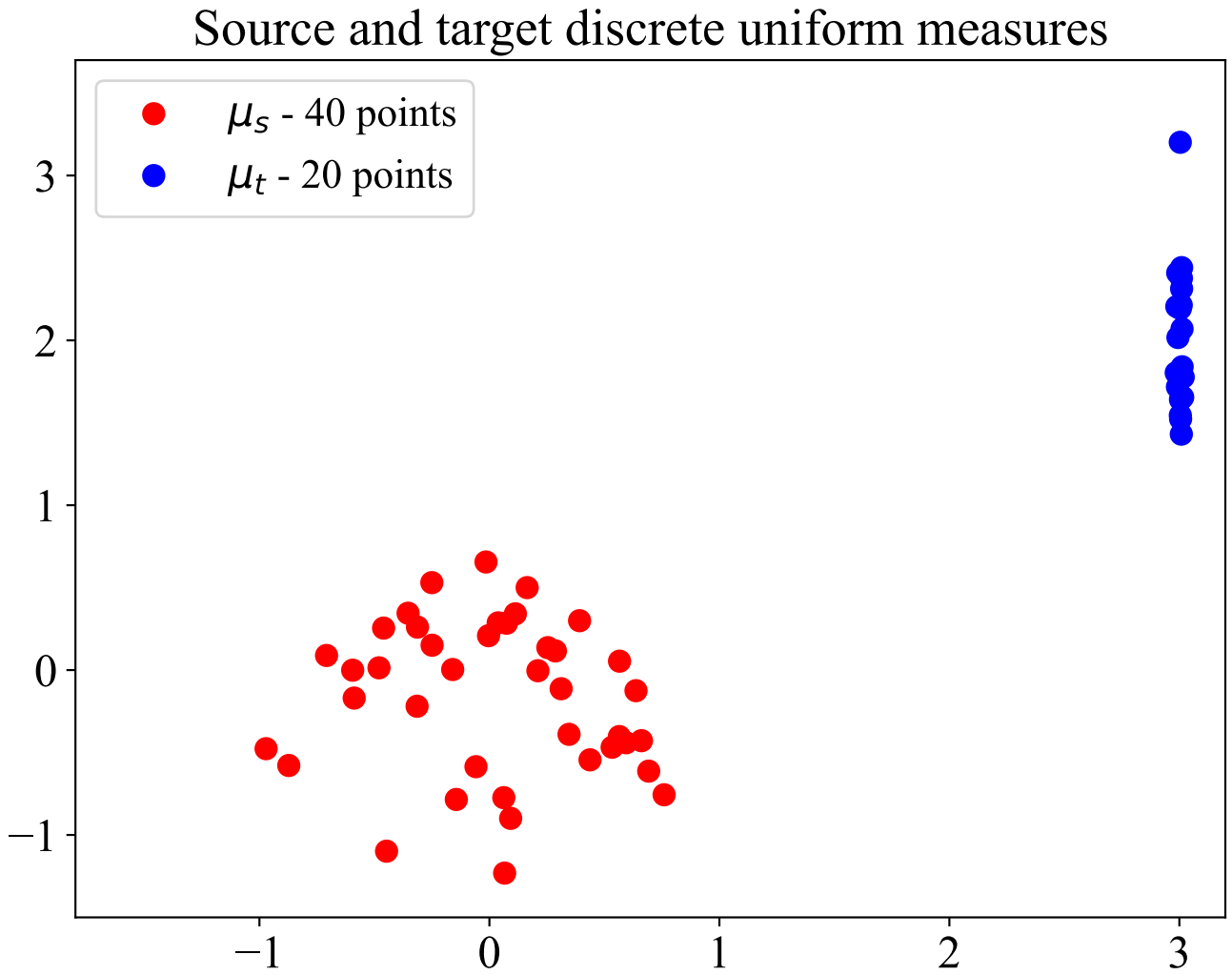
The weighted average between $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ is defined by

$$\mathfrak{M}(\mu, \nu; t) = (\pi^t)_{\#}\sigma, \quad t \in [0, 1],$$

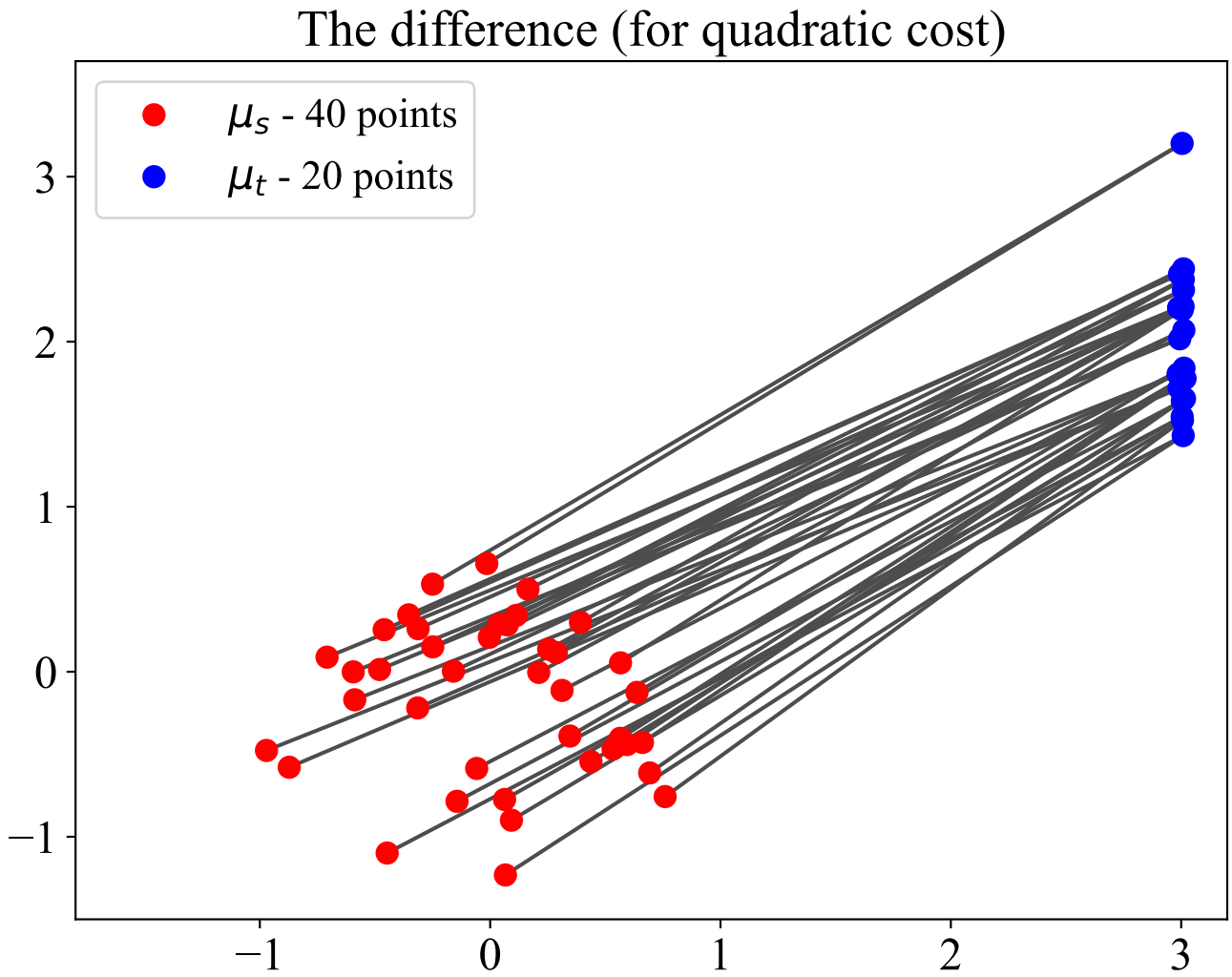
where σ is an optimal transport plan pushing μ onto ν, and the map $\pi^t : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is given by $\pi^t(x, y) = (1 - t)x + ty$.

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

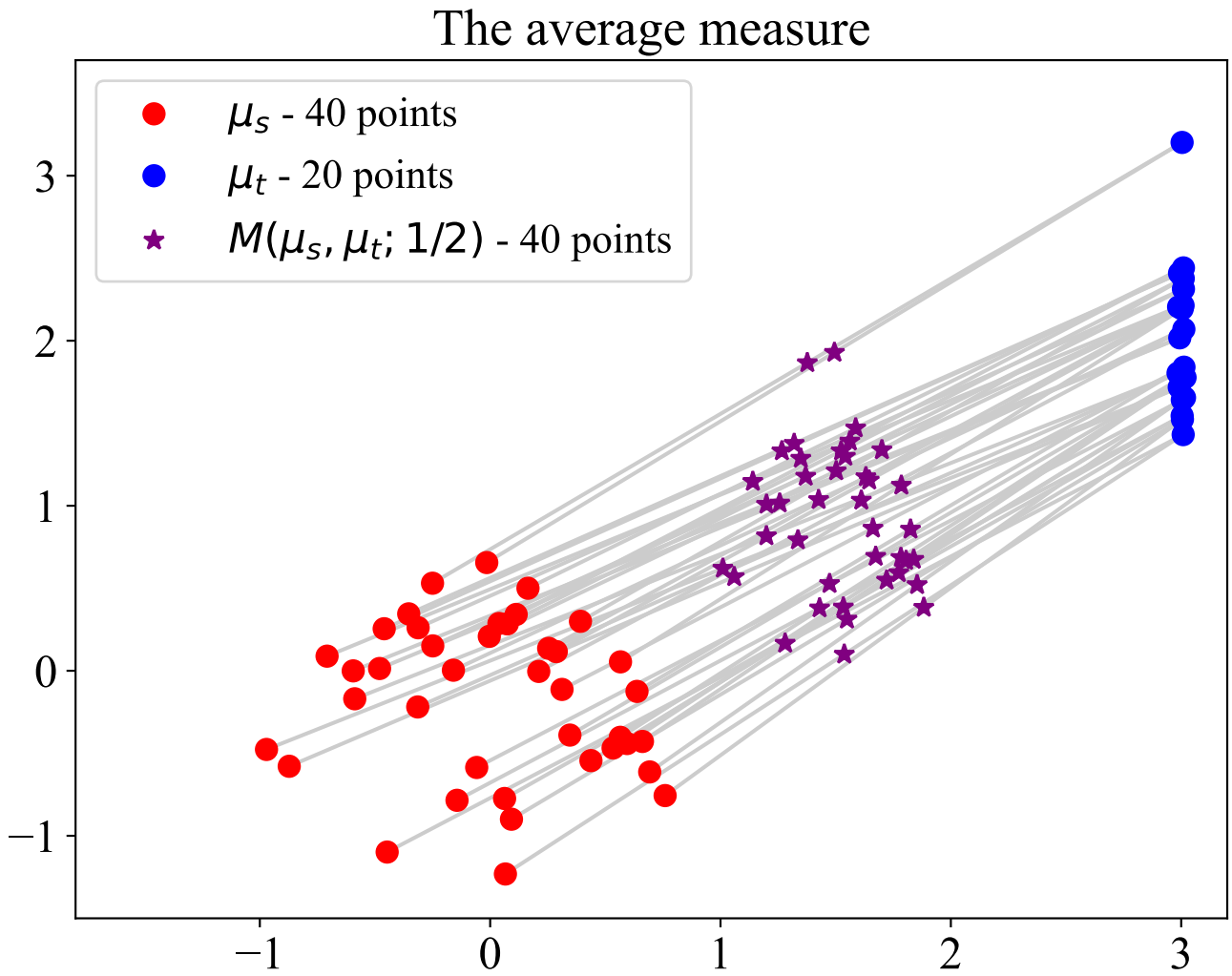
Operators on discrete measures (illustrations)



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 $\mathcal{S}$ that is given by the rules

$$\begin{cases} (\mathcal{S}\boldsymbol{c})_{2k} = c_k, \\ (\mathcal{S}\boldsymbol{c})_{2k+1} = \frac{1}{2}c_k + \frac{1}{2}c_{k+1}, \end{cases} \qquad k \in \mathbb{Z},$$

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

$$\begin{cases} (\mathcal{S}\boldsymbol{\mu})_{2k} = \mu_k, \\ (\mathcal{S}\boldsymbol{\mu})_{2k+1} = \mathfrak{M}(\mu_k, \mu_{k+1}; 1/2), \end{cases} \qquad k \in \mathbb{Z}.$$

The $\ominus$ and $\oplus$ operators

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

$$\nu \ominus \mu = T_\mu^\nu - I,$$

where T_μ^ν is the optimal transport map, and I is the identity map. With this definition, we have that $\mu \ominus \mu = 0$ the zero map, and

$$\|\nu \ominus \mu\|^p_{L^p(\mathbb{R}^d;\mu)} = \int_{\mathbb{R}^d} \|T_\mu^\nu(x) - x\|^p d\mu(x) = W_p^p(\mu, \nu).$$

Moreover, the compatible $\oplus$ operator is defined via

$$\mu \oplus \psi = (\psi + I)_\# \mu,$$

for any Borel measurable map $\psi : \mathbb{R}^d \to \mathbb{R}^d$.

The $\ominus$ and $\oplus$ operators (for discrete measures)

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

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

the difference is defined via

$$\nu \ominus \mu = \left(\left[x_j^\nu - x_i^\mu \right]_{i=1,\dots,m}^{j=1,\dots,n}, \ \Lambda_\mu^\nu \right),$$

where Λ_μ^ν is the coupling matrix between μ and ν. In this case, the compatible $\oplus$ operator is defined via

$$\mu \oplus \psi = \sum_{i=1}^{m} \sum_{j=1}^{k} \lambda_{i,j}^\psi \delta_{x_i^\mu + x_{i,j}^\psi},$$

for any tensor pair $\psi = (x^\psi, \Lambda^\psi)$ where $x \in \mathbb{R}^{m \times k \times d}$ and $\Lambda^\psi \in \mathbb{R}^{m \times k}$.

Theoretical results

Multiscaling a sequence $\boldsymbol{\mu}^{(J)}$ in $\mathcal{P}_p(\mathbb{R}^d)$ with the above notations via the interpolating elementary $\mathcal{S}$ becomes straightforward.

Theorem (decay of detail coefficients)

Assume $\boldsymbol{\mu}^{(J)}$ is sampled over the dyadic grid parametrization $2^{-J}\mathbb{Z}$ from an absolutely continuous curve $\boldsymbol{\mu}$ in $\mathcal{P}_p(\mathbb{R}^d)$ with a finite metric derivative $\Gamma = \sup_{t\in\mathbb{R}} |\mu'|_t < \infty$. Then, the resulting detail coefficients $\boldsymbol{\psi}^{(\ell)}$ satisfy

$$\|\boldsymbol{\psi}^{(\ell)}\|_\infty \leq \Gamma 2^{1-\ell}, \quad \ell = 1, \dots, J,$$

where

$$\|\boldsymbol{\psi}^{(\ell)}\|_\infty = \sup_{k\in\mathbb{Z}} \|\psi_k^{(\ell)}\|_{L^p(\mu_k^{(\ell)})}.$$

Theoretical results

The refinement $\mathcal{S}$ is called stable if there exists $K > 0$ such that

$$\mathcal{W}_p(\mathcal{S}\boldsymbol{\mu}, \mathcal{S}\boldsymbol{\nu}) \leq K\, \mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu})$$

for any sequences $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ where $\mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \sup_{k\in\mathbb{Z}} W_p(\mu_k, \nu_k)$.

Theorem (stability of reconstruction)

Let $\{\boldsymbol{\mu}^{(0)}; \boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(J)}\}$ and $\{\widetilde{\boldsymbol{\mu}}^{(0)}; \widetilde{\boldsymbol{\psi}}^{(1)}, \ldots, \widetilde{\boldsymbol{\psi}}^{(J)}\}$ be two pyramid representations of $\boldsymbol{\mu}^{(J)}$ and $\widetilde{\boldsymbol{\mu}}^{(J)}$, respectively. Assume that $\|\psi_k^{(\ell)}\|_{\mathrm{Lip}} \leq C$ for all $\ell = 1, \ldots, J$ and $k \in \mathbb{Z}$. If $\mathcal{S}$ is stable with constant K, then for $L = \max\{1, (KC)^J\}$ we have

$$\mathcal{W}_p(\boldsymbol{\mu}^{(J)}, \widetilde{\boldsymbol{\mu}}^{(J)}) \leq L\left(\mathcal{W}_p(\boldsymbol{\mu}^{(0)}, \widetilde{\boldsymbol{\mu}}^{(0)}) + \sum_{\ell=1}^{J} \|\boldsymbol{\psi}^{(\ell)} - \widetilde{\boldsymbol{\psi}}^{(\ell)}\|_\infty\right).$$

Outline

Application 1: Denoising

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

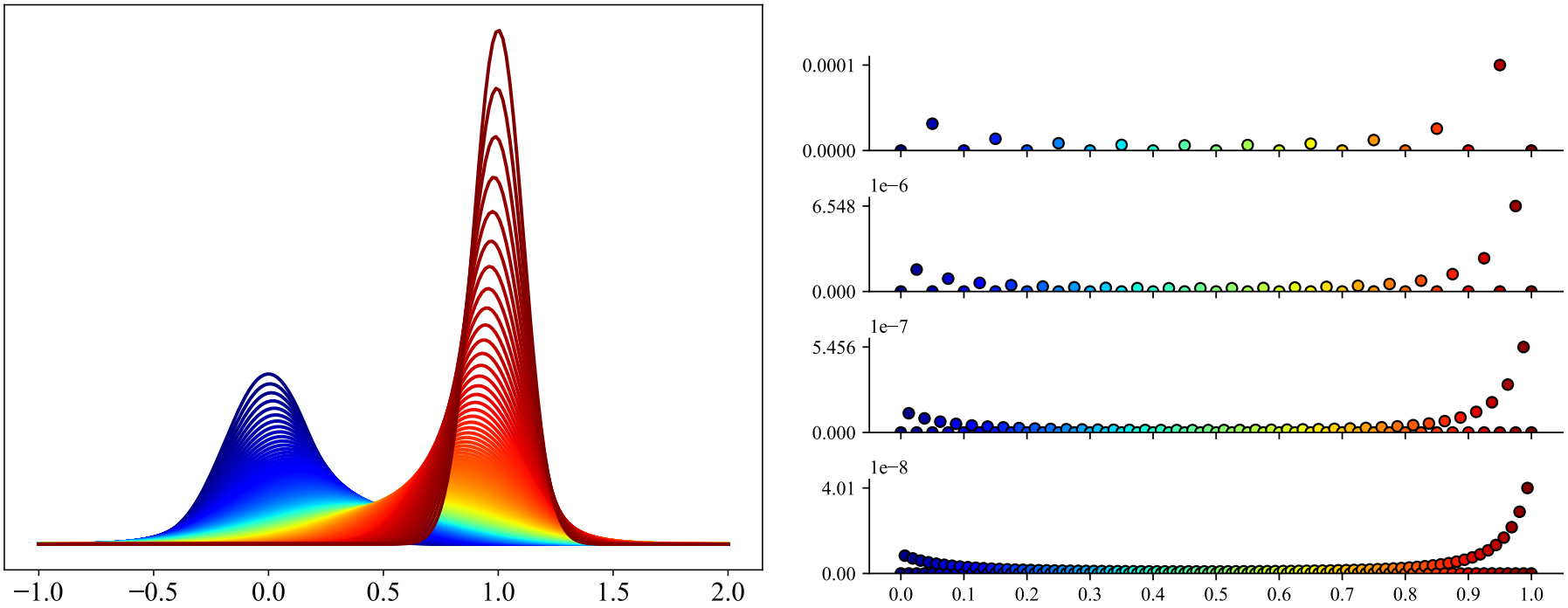


Figure: Sequence of Gaussian measures and its multiscale transform.

Application 1: Denoising

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

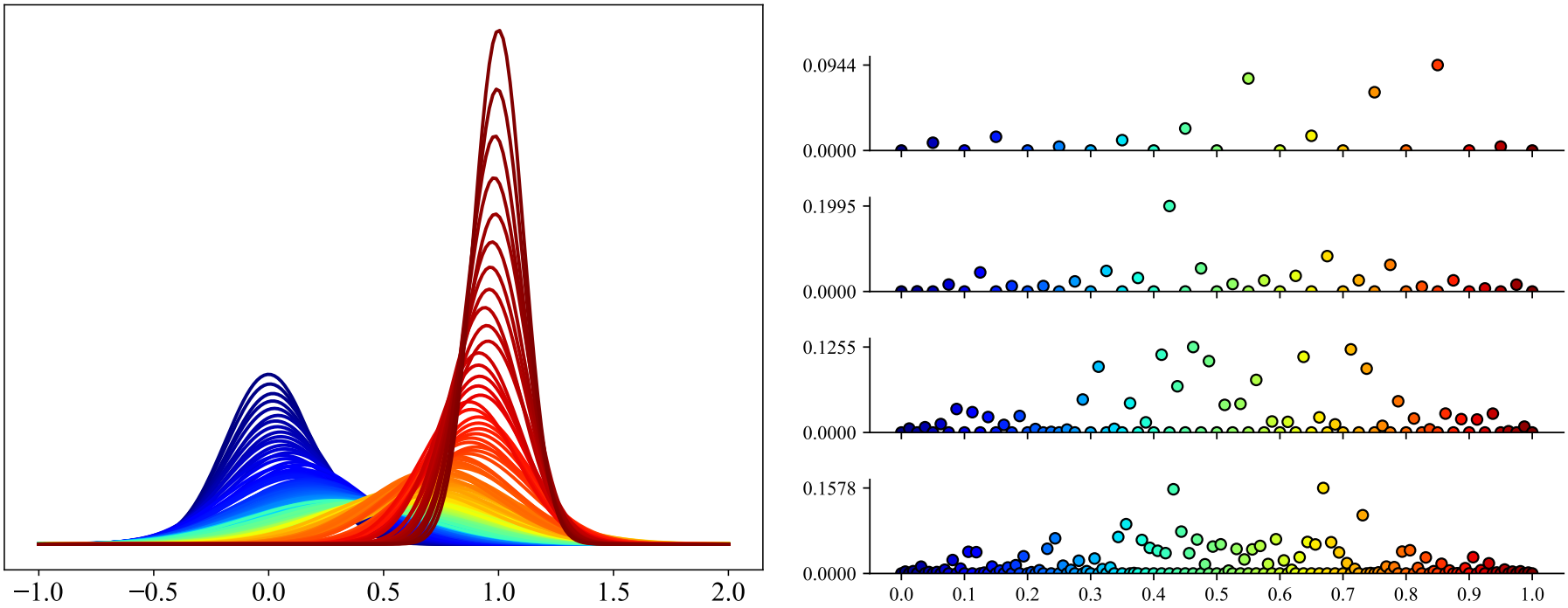


Figure: Contamination with noise.

Application 1: Denoising

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

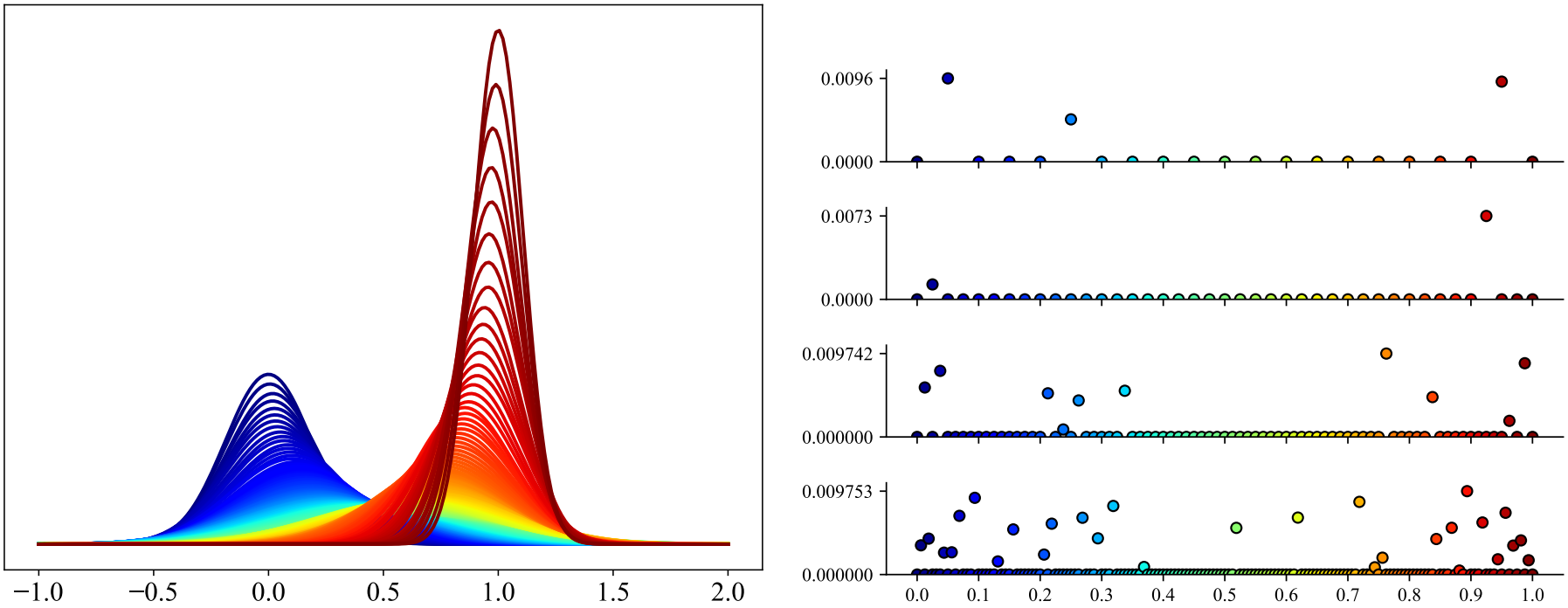


Figure: Denoising result obtained by thresholding with 0.01.

Application 2: Anomaly detection

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

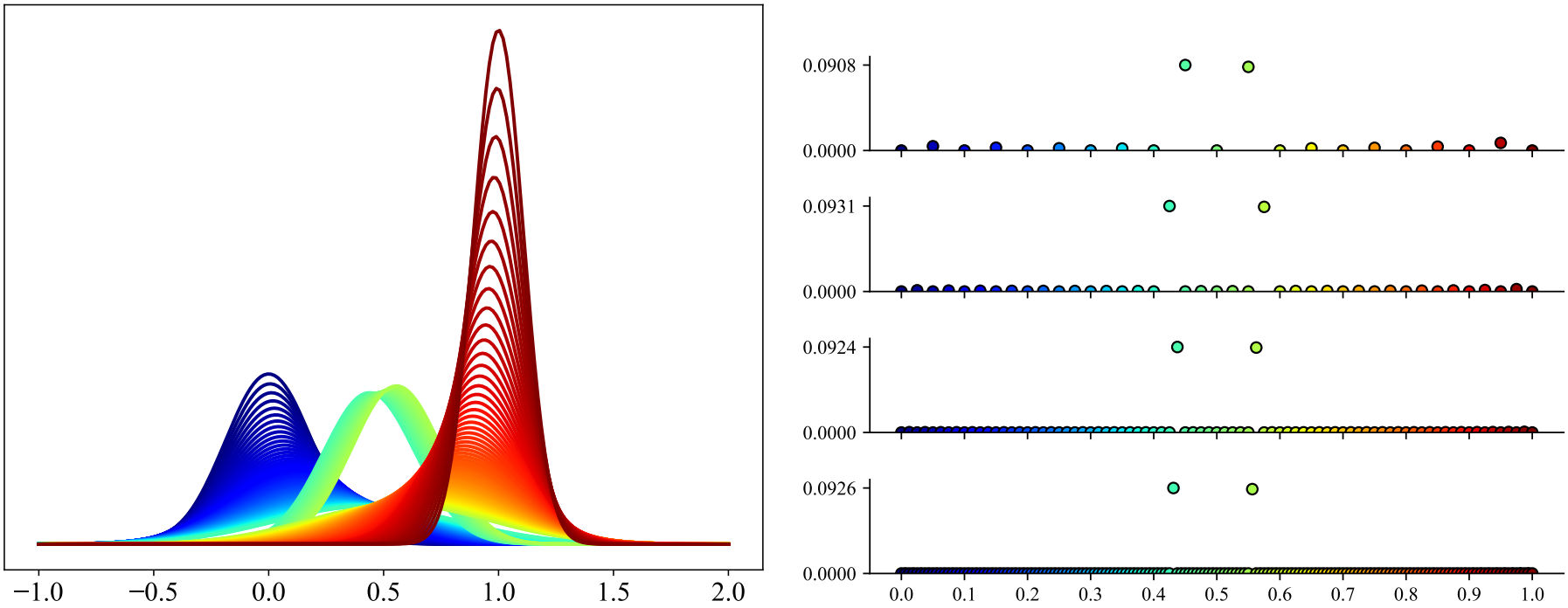


Figure: Detecting jump discontinuities via multiscaling.

Application 3: Analyzing NN learning dynamics

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

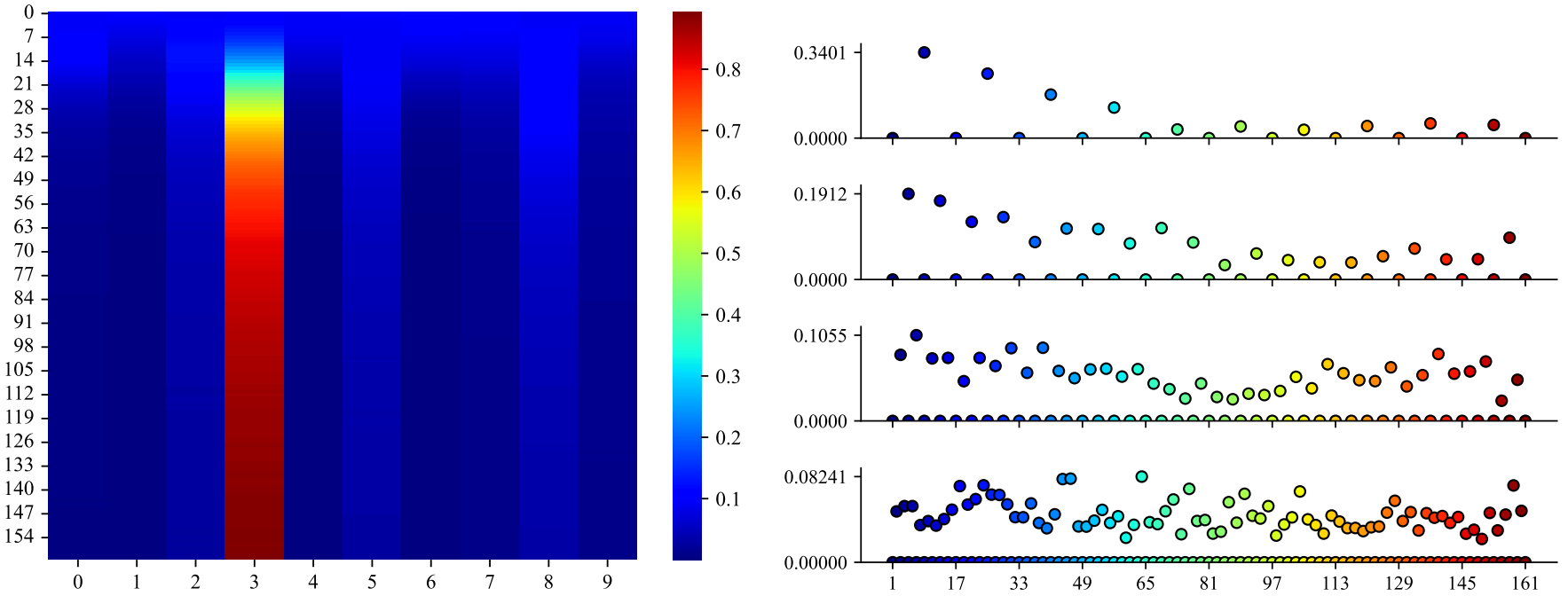


Figure: Analyzing the learning dynamics of a simple neural network on MNIST dataset. On the left, the prediction of the digit "3" across epoch iterations. On the right, the multiscale transform of the resulted measure sequence. The convergence is clear on coarse scales.

Application 4: Contrast enhancement

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

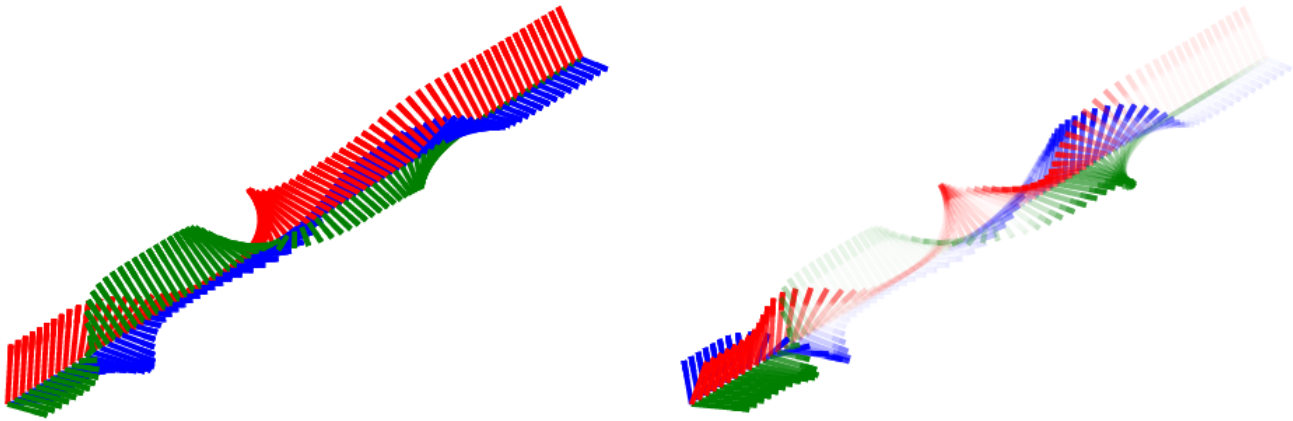


Figure: Contrast enhancement of a sequence of rotation matrices. On the left, illustration of the original sequence applied to the standard basis of $\mathbb{R}^3$. On the right, the enhanced sequence obtained by scaling the top 20% of detail coefficients with a factor of 40%.

Application 5: Data compression

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

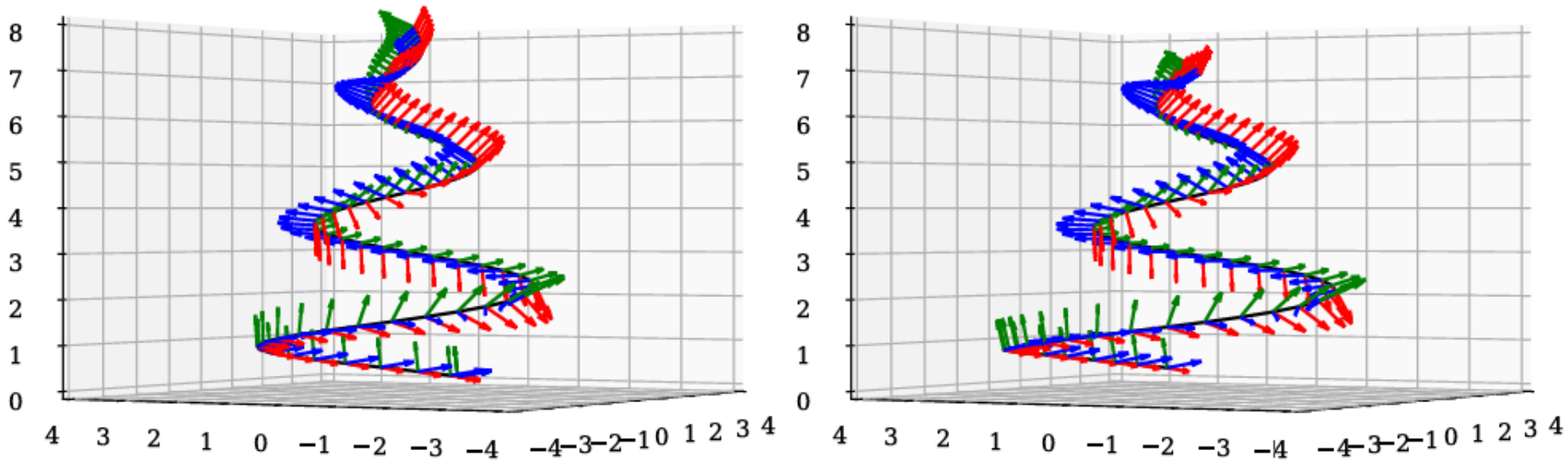


Figure: Data compression of a rigid body motion. On the left, illustration of 641 matrices representing the special Euclidean sequence. On the right, the decompressed trajectory that is obtained by 41 matrices in addition to 12 detail coefficients.

Outline

Bivariate multiscale transform

Multivariate multiscale transforms can be constructed by applying tensor products to α and γ.

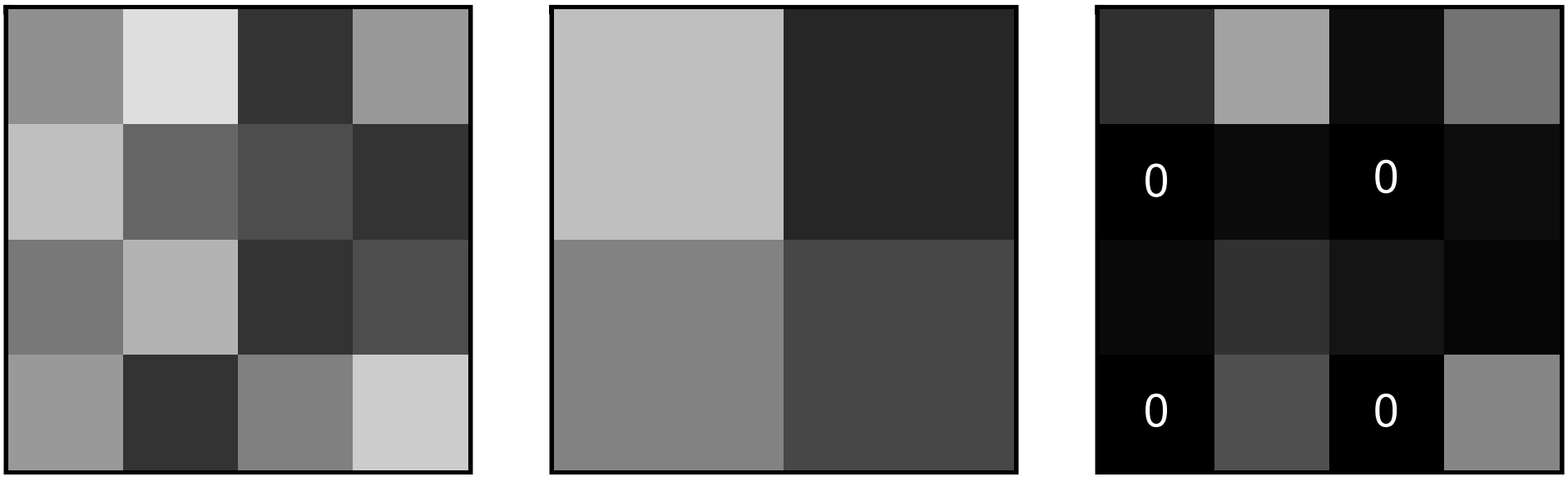


Figure: Sketch for decomposing a 4×4 grayscale image.

Multiscaling an image

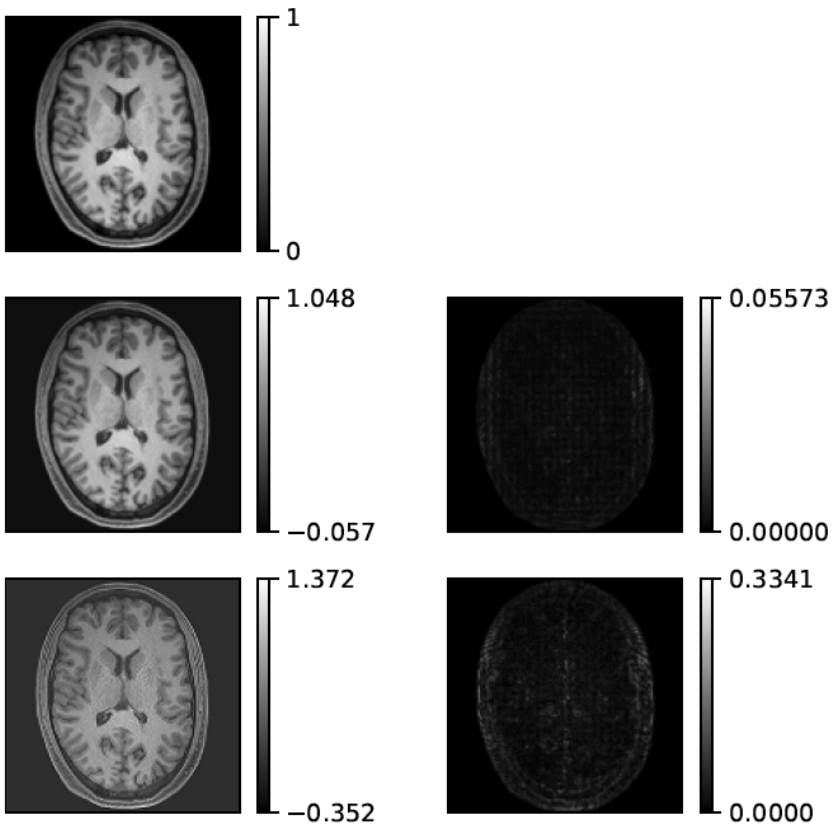
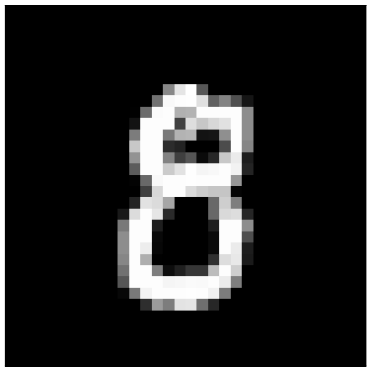


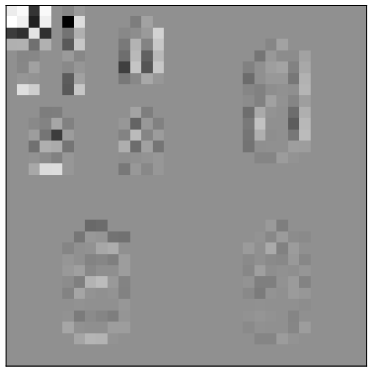
Figure: Multiscaling an image of the author's own brain (MRI).

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.



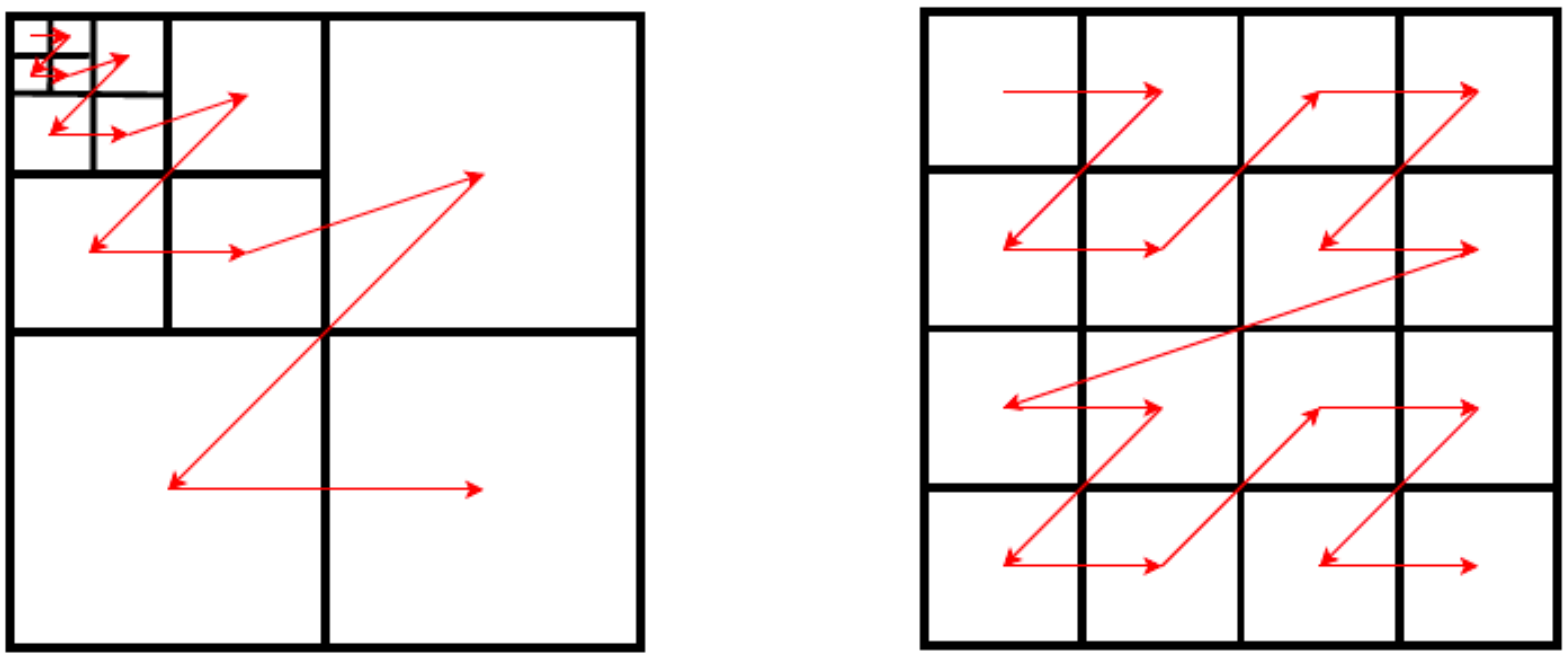


Wavelets are all you need for autoregressive image generation.

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



(a) Subband scanning order.

(b) Scanning order of 4×4 blocks.

Figure 13: A sketch illustrating the outer and inner scanning orders.

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 ω its MRA decomposition. Namely,

$$\omega_{(i_1,i_2)} = \begin{cases} \langle f, \widetilde{\varphi}_{m,(i_1,i_2)} \rangle, & 1 \le i_1, i_2 \le 2, \\ \langle f, \widetilde{\psi}^e_{j,k} \rangle, & 3 \le i_1, i_2 \le M, \quad 1 \le j \le m, \quad e = 1,2,3. \end{cases}$$

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

$$\max_{(i_1,i_2)} |\omega_{(i_1,i_2)}| \le 2^{m-1}.$$

Now, compute $\widetilde{m} = \lceil \log_2 \max_{(i_1,i_2)} |\omega_{(i_1,i_2)}| \rceil$ and initialize the first threshold $T = 2^{\widetilde{m}-1}$ for the beginning of the tokenization.

First bit-plane

Beginning with the initial bit-plane $[T/2, T]$:

- If $|\omega_{(i_1,i_2)}| \geq T/2 \quad \Rightarrow \quad$ the coefficient will be reported with 'NowSignificantPos' or 'NowSignificantNeg' and gets the approximation value $\pm 3T/4$ depending on its sign.

- If $|\omega_{(i_1,i_2)}| < T/2 \quad \Rightarrow \quad$ 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 $\pm\, T/8$.

- If $T/4 \leq |\omega_{(i_1,i_2)}| \leq T/2 \quad \Rightarrow \quad$ the coefficient will be reported with 'NowSignificantPos' or 'NowSignificantNeg' and gets the approximation value $\pm\, 3T/8$ depending on its sign.

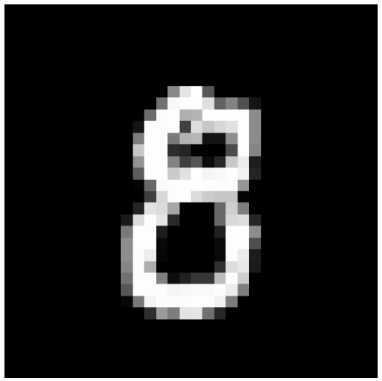
Zero blocks

We introduce two additional tokens to shorten the tokenization:

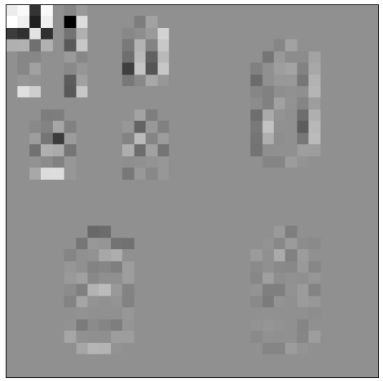
- The token 'Group2x2' replaces each square block of 4 'Insignificant' tokens falling in the same subband if the top left position indices are divisible by 2.
- The token 'Group4x4' replaces each square block of 16 'Insignificant' tokens falling in the same subband if the top left position indices are divisible by 4.

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

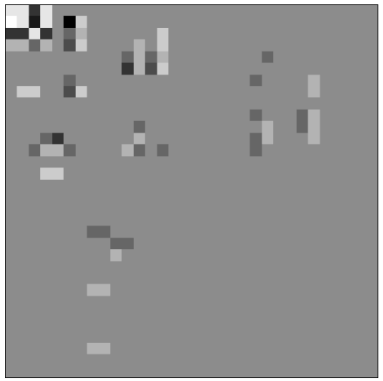
Tokenization process (illustration)



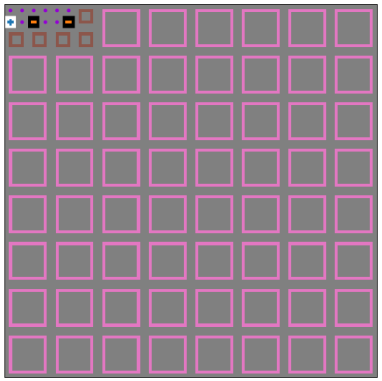
(A) 32×32 padded MNIST image.



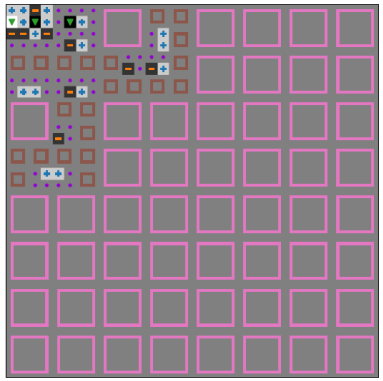
(B) Wavelet Transform.



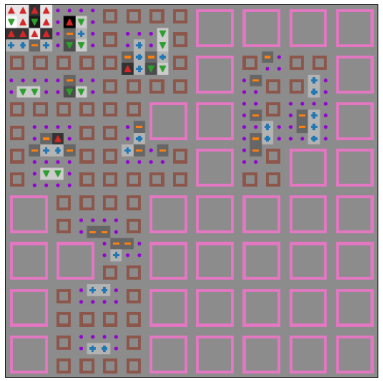
(C) Three bit planes of coefficients.



(D) First bit plane.



(E) Second bit plane.



(F) Third bit plane.

Dataset structure

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

	sequence	w_position_id	class_id	max_threshold	bit_plane
0	4x4_zero_block 4x4_zero_blo...	(1, 1, 1) (1, 5, 1) (5, 1, 1) ...	5	4	5
1	insignificant now_significant_p...	(1, 1, 1) (1, 2, 1) (2, 1, 1) ...	0	4	5
2	insignificant now_significant_p...	(1, 1, 1) (1, 2, 1) (2, 1, 1) ...	4	2	4
3	insignificant insignificant now_...	(1, 1, 1) (1, 2, 1) (2, 1, 1) ...	1	4	5
4	4x4_zero_block 2x2_zero_blo...	(1, 1, 1) (1, 5, 1) (1, 7, 1) ...	9	4	5
5	insignificant insignificant now_...	(1, 1, 1) (1, 2, 1) (2, 1, 1) ...	2	4	5
6	4x4_zero_block 2x2_zero_blo...	(1, 1, 1) (1, 5, 1) (1, 7, 1) ...	1	4	5
7	insignificant now_significant_p...	(1, 1, 1) (1, 2, 1) (2, 1, 1) ...	3	4	5
8	insignificant insignificant insig...	(1, 1, 1) (1, 2, 1) (2, 1, 1) ...	1	2	4
9	insignificant now_significant_p...	(1, 1, 1) (1, 2, 1) (2, 1, 1) ...	4	2	4
10	insignificant now_significant_p...	(1, 1, 1) (1, 2, 1) (2, 1, 1) ...	3	4	5
11	insignificant now_significant_p...	(1, 1, 1) (1, 2, 1) (2, 1, 1) ...	5	2	4
12	now_significant_positive insig...	(1, 1, 1) (1, 2, 1) (2, 1, 1) ...	3	4	5
13	insignificant insignificant now_...	(1, 1, 1) (1, 2, 1) (2, 1, 1) ...	6	4	5
14	4x4_zero_block insignificant i...	(1, 1, 1) (1, 5, 1) (1, 6, 1) ...	1	2	4
15	insignificant now_significant_p...	(1, 1, 1) (1, 2, 1) (2, 1, 1) ...	7	4	5

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.

$$\Big(\underbrace{0,\ldots,1,\ldots,0}_{\text{token}},\ \underbrace{i_1,i_2,k}_{\text{position \& BP}}\ ,\underbrace{0,\ldots,1,\ldots,0}_{\text{threshold}},\underbrace{0,\ldots,1,\ldots,0}_{\text{class}}\Big)$$

Conditional inference

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

- 'NowSignificant' tokens can only be preceded by 'Insignificant' or 'Group' tokens and followed by 'NextAccuracy' tokens.
- 'NextAccuracy' tokens can only be preceded by 'NextAccuracy' or 'NowSignificant' tokens.
- 'Insignificant' can only be preceded by either 'Insignificant' 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.

Figure 16: Digits generated with Top-$p = 0.6$ along with a depiction of the generated wavelet coefficients.

Empirical results

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

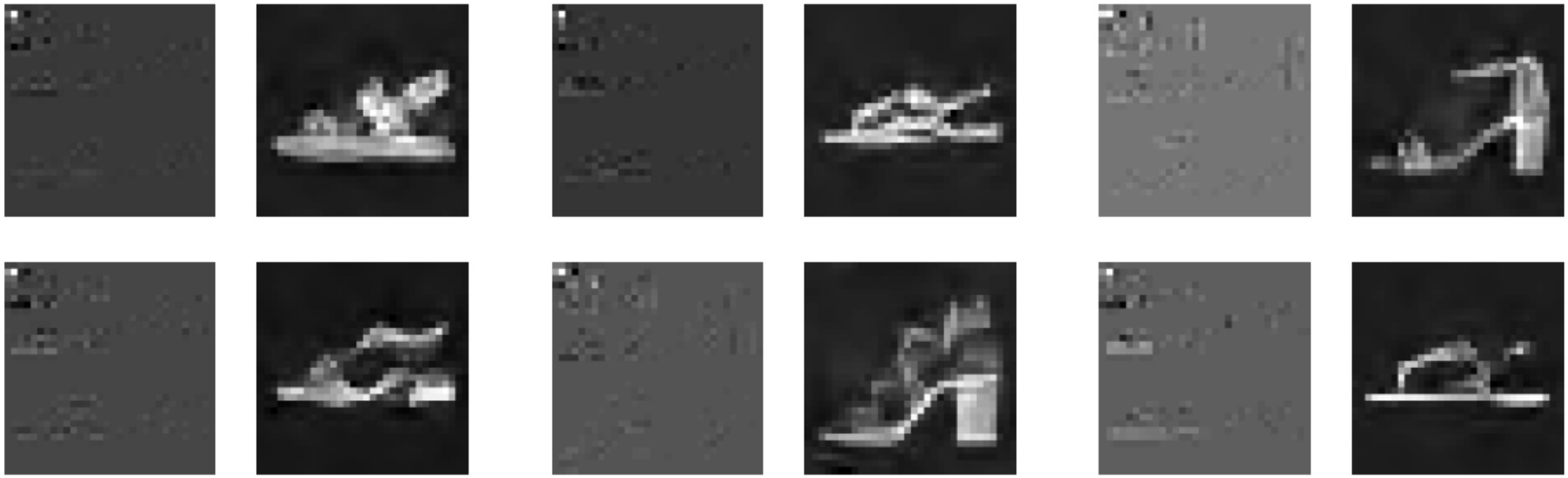


Figure 17: Sandals generated with Top-$k = 2$ along with a depiction of the generated wavelet coefficients.

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.

Figure 18: More MNIST results.

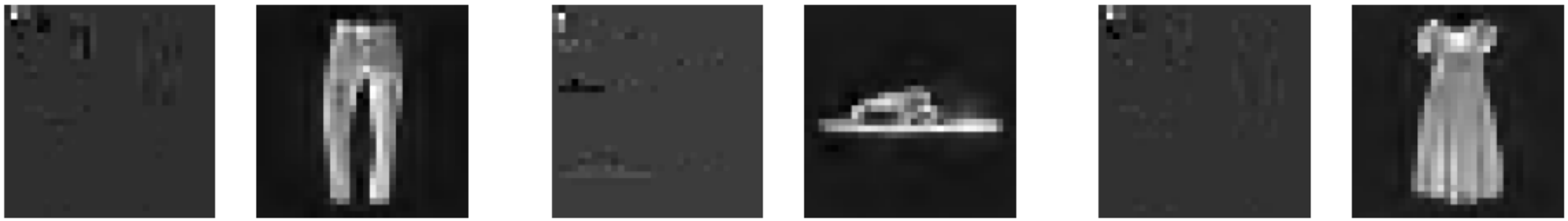


Figure 19: More FashionMNIST results.

Thank you for listening!
Questions?

Image generated by Google's Gemini