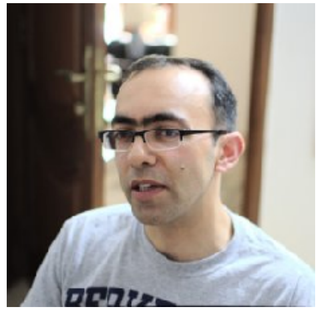
Cross-fluctuation phase transitions reveal sampling dynamics in diffusion models

Sai Niranjan Ramachandran

School of Computation, Information and Technology, Technical University of Munich

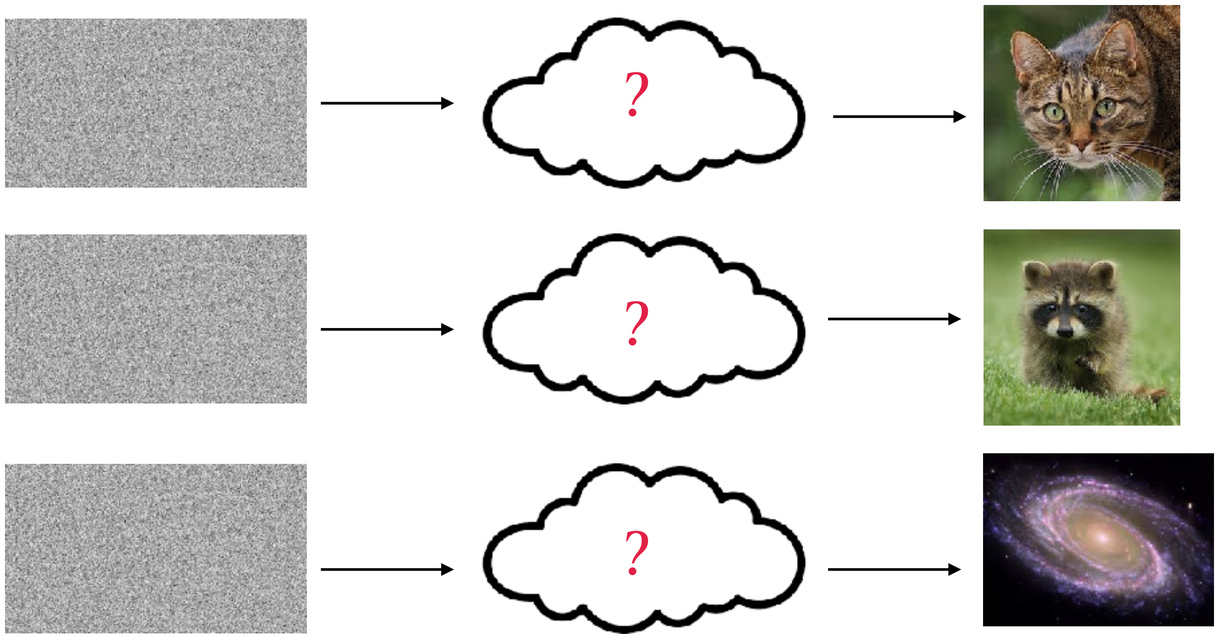
Joint work with

Manish Krishan Lal



Suvrit Sra

Diffusion models are a black box



When does noise become structure?

Which denoising steps drive semantic formation?

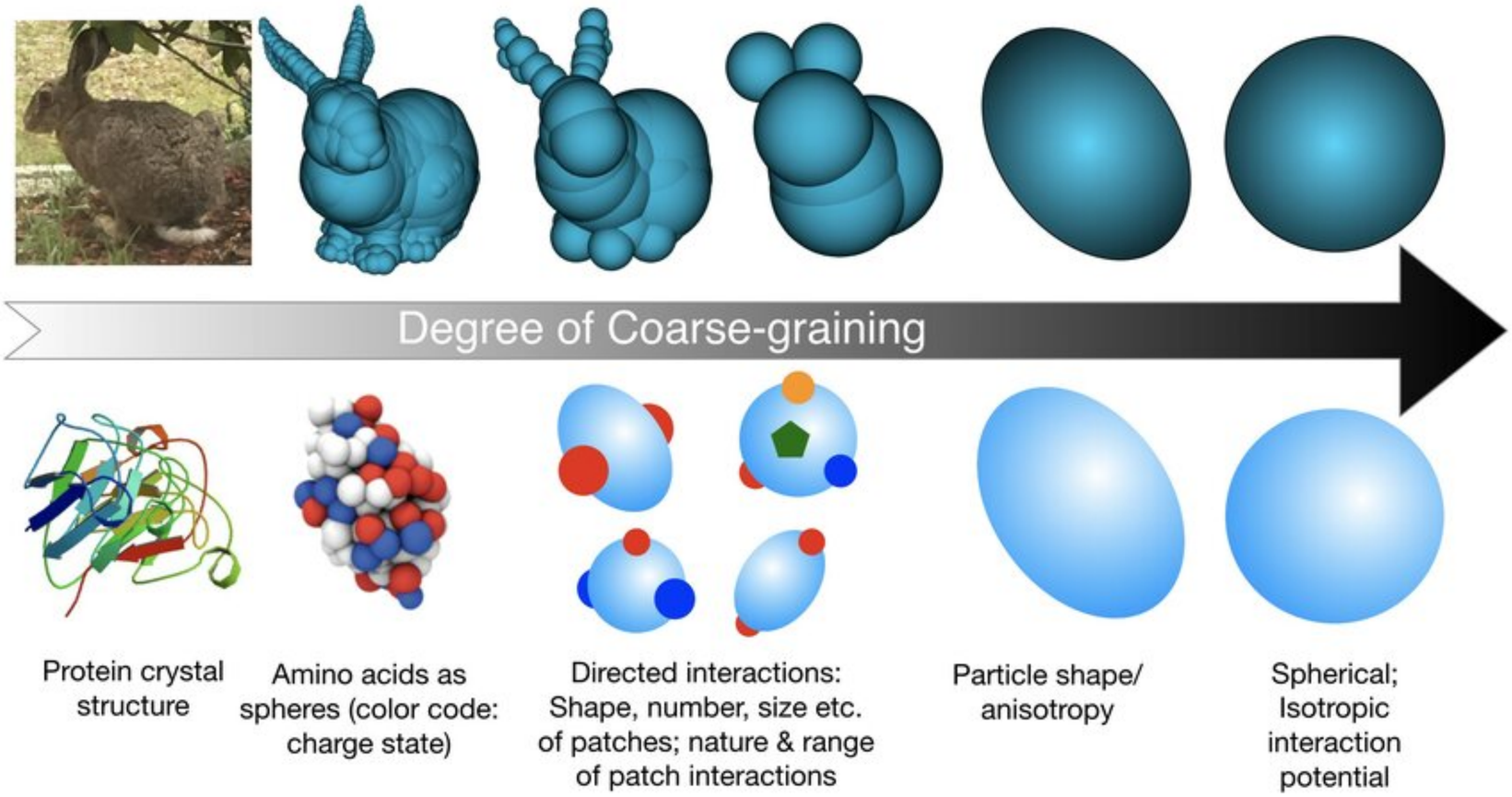
How can we steer generation without retraining?

Diffusion models transform pure noise into complex images through hundreds of denoising steps, but when and how does structure emerge??

We show that leveraging model insight delivers practical gains: zero-cost acceleration and low-overhead state of the art generation and downstream performance.

Key insights on structure emergence from physical systems

Coarse Graining is Universal



We can analyse simple aggregate behaviour while seamlessly bridging scales.

Decoding the structure of Diffusion Models

Coarse Graining (What): *Partition data space into meaningful groups ($\sqcup_i \Omega_i = \Omega$) eg semantic clusters / classes etc.*

Property to track (How) : *Statistical distinctiveness of events throughout the generation process through some metric $\mathscr{M}$.*

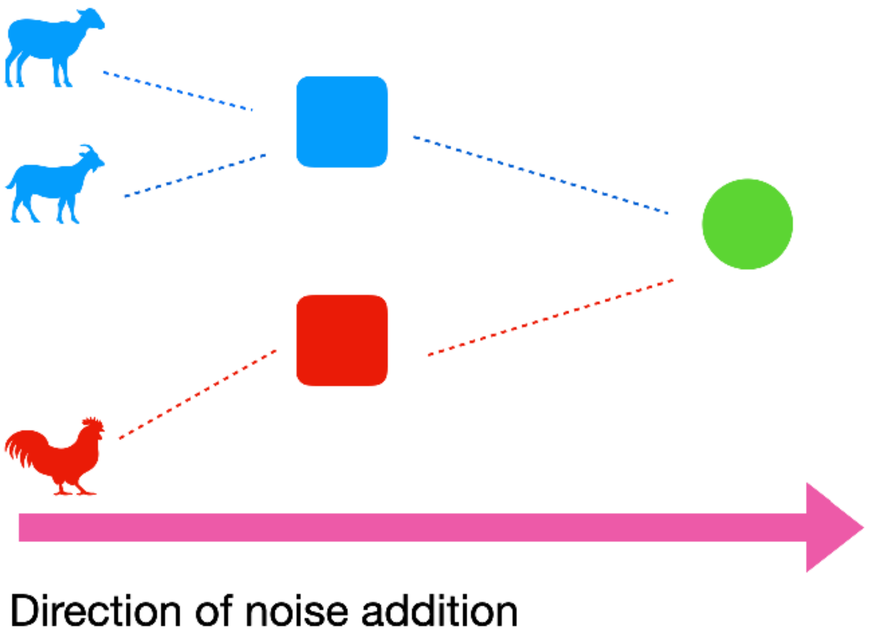
Scale-Bridging (Why) : *Achievable through good choices of $\mathscr{M}$. We show that cross-moments satisfy this constraint.*

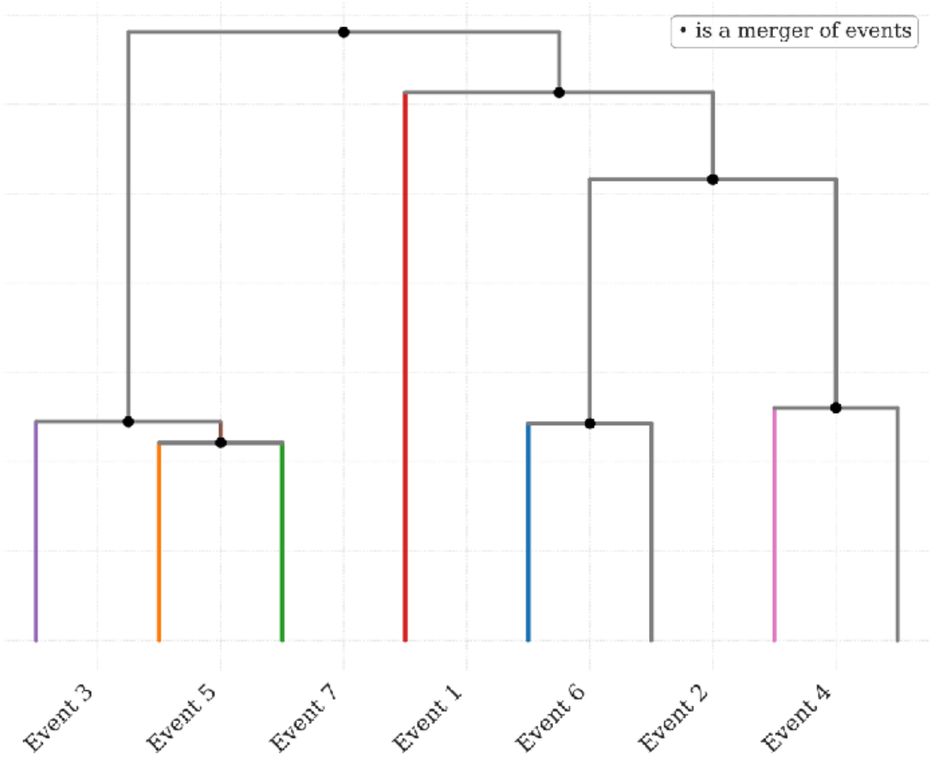
Computationally efficient once we ....

1. *Switch to the forward process to track "mergers" instead.*

2. *Track eigenspectrum of covariances instead of complete arbitrary cross-moments.*

Unveiling an unsupervised semantic hierarchy





We can pinpoint exactly when the model loses the ability to distinguish between different semantic classes.

This reveals the model's own implicit hierarchy over the data - discovered in an unsupervised manner.

Applications

Choice of events	Revealed Significance	Downstream Task
$\Omega_{1,0} = p_{\text{data}}, \Omega_{2,0} = \mathcal{N}(\mathbf{0}, \mathbf{I})$	Convergence of data to the stationary distribution.	Accelerating sampling by relying on the data's properties
$\Omega_{k,0} = k$-th class of p_{data}	Emergence of class-level structure.	Optimising guidance. Intervals for class-conditional sampling.
$\Omega_{k,0} = k$-th class of p_{data}	Emergence of class-level structure.	Zero shot classification.
$\Omega_{k,0} = k$-th *rare* class of p_{data}	Emergence of class-level structure for *rare classes*.	Improved rare-class generation.
$\Omega_{k,0} = k$-th class of p_{source}	Emergence of class-level structure for source/domain content.	Improved style transfer.

Accelerated Inference

Model / Dataset	FID(↓)	Steps(↓)	GFLOPs(↓)
DiT-XL/2 (ImageNet, full)	3.42 ± 0.21	250	4100
DiT-XL/2 (ImageNet, ours)	$\mathbf{3.37} \pm 0.31$	**175**	**2870**
DDPM (MNIST, full)	2.27 ± 0.19	1000	2000
DDPM (MNIST, ours)	2.29 ± 0.17	**600**	**1200**
DDPM (CIFAR-10, full)	3.62 ± 0.35	500	6000
DDPM (CIFAR-10, ours)	$\mathbf{3.47} \pm 0.34$	**300**	**3600**

Our data-centric approach accelerates sampling by starting the reverse process at the data-noise merger time ($i^\star$), achieving comparable FID scores with significantly fewer steps and lower computational cost.

Class conditional generation

Model/Dataset	FID (↓)	Precision (↑)	Recall (↑)	Density (↑)	Coverage (↑)
DiT-XL/2 (Imagenet, IG Baseline)	3.22 ± 0.16	0.78 ± 0.01	0.23 ± 0.05	0.83 ± 0.01	0.35 ± 0.02
DiT-XL/2 (Imagenet, IG Ours)	$\mathbf{2.86 \pm 0.15}$	$\mathbf{0.83 \pm 0.02}$	$\mathbf{0.26 \pm 0.04}$	$\mathbf{0.85 \pm 0.01}$	$\mathbf{0.39 \pm 0.02}$
DDPM (MNIST, IG Baseline)	2.15 ± 0.06	0.80 ± 0.02	0.25 ± 0.01	0.85 ± 0.01	0.36 ± 0.02
DDPM (MNIST, IG Ours)	$\mathbf{1.99 \pm 0.11}$	$\mathbf{0.85 \pm 0.01}$	$\mathbf{0.28 \pm 0.03}$	$\mathbf{0.89 \pm 0.02}$	$\mathbf{0.40 \pm 0.01}$
DDPM (CIFAR10, IG Baseline)	3.32 ± 0.25	0.77 ± 0.01	0.19 ± 0.14	0.81 ± 0.03	0.32 ± 0.02
DDPM (CIFAR10, IG Ours)	$\mathbf{3.01 \pm 0.14}$	$\mathbf{0.79 \pm 0.02}$	$\mathbf{0.22 \pm 0.01}$	$\mathbf{0.84 \pm 0.01}$	$\mathbf{0.35 \pm 0.04}$

Table 2: Class conditional generation

Model/Dataset	CLIP Similarity (↑)	Precision (↑)	Recall (↑)	Density (↑)	Coverage (↑)
SD (iNaturalist, IG baseline)	0.21 ± 0.03	0.70 ± 0.01	0.15 ± 0.01	0.75 ± 0.01	0.27 ± 0.02
SD (iNaturalist, IG Ours)	0.24 ± 0.02	0.73 ± 0.01	0.18 ± 0.01	0.78 ± 0.02	0.30 ± 0.04
SD (iNaturalist, IG-ILVR Ours)	$\mathbf{0.27 \pm 0.01}$	$\mathbf{0.76 \pm 0.01}$	$\mathbf{0.19 \pm 0.02}$	$\mathbf{0.81 \pm 0.01}$	$\mathbf{0.31 \pm 0.03}$
SD (CUB200, IG baseline)	0.24 ± 0.05	0.75 ± 0.01	0.14 ± 0.01	0.79 ± 0.02	0.30 ± 0.02
SD (CUB200, IG Ours)	0.26 ± 0.01	0.78 ± 0.05	0.17 ± 0.01	0.82 ± 0.02	$\mathbf{0.33 \pm 0.01}$
SD (CUB200, IG-ILVR Ours)	$\mathbf{0.27 \pm 0.02}$	$\mathbf{0.82 \pm 0.02}$	$\mathbf{0.18 \pm 0.02}$	$\mathbf{0.85 \pm 0.01}$	0.31 ± 0.02

Table 3: Rare class generation using StableDiffusion

Baseline results from a single, grid-searched guidance interval.

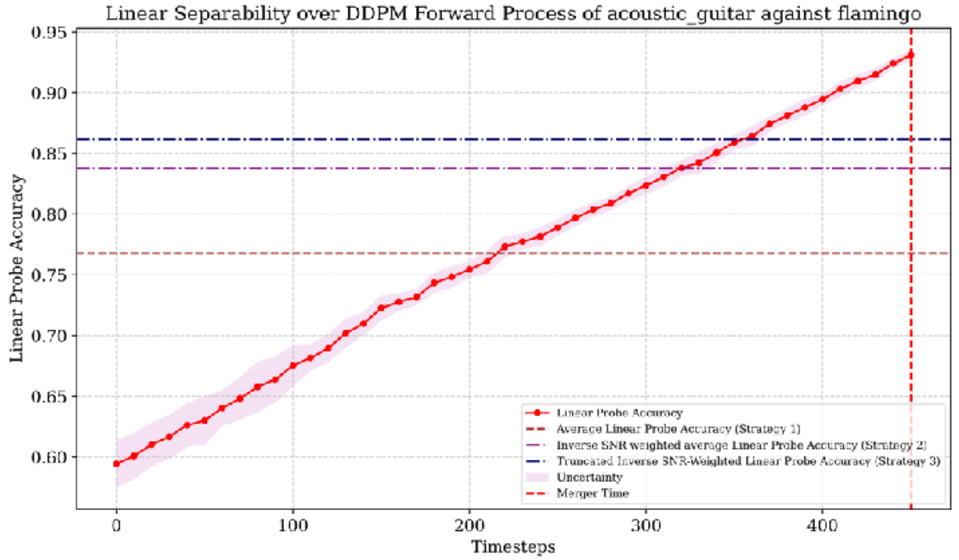
Our results using efficient merger based guidance.

Our data-centric approach provides fine-grained, class-specific guidance intervals, achieving comparable or superior visual quality at a fraction of the computational cost.

Zero-shot classification

Method	ImageNet ↑	CIFAR-10 ↑	Oxford-IIIT Pets ↑
SD, uniform (Li et al.)	54.96±0.67	84.67±1.23	82.87±0.39
SD, uniform (ours)	57.91±0.53	85.17±0.17	86.17±0.26
SD, inverse-SNR	64.17±0.33	87.26±0.67	88.17±0.29
SD, trunc. inverse-SNR	**65.28±0.46**	**88.38±0.43**	**89.15±0.26**
CLIP RN-50	58.41±0.35	75.42±0.26	85.61±0.29
OpenCLIP ViT-H/14	76.91±0.75	96.87±0.59	94.61±0.37

Table 4: Zero-shot multi-class accuracy (%); ± 95% CI over five runs.

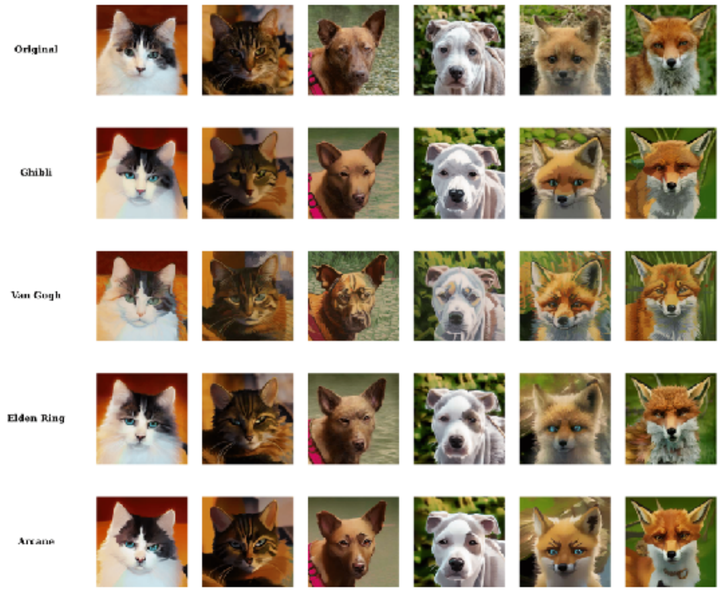


Diffusion process acts as a low-pass filter, simplifying representations and removing non-essential features, making them most discriminative right before core structures merge.

Zero-shot style transfer

Style	Ghibli		van-Gogh	
Models/Metrics	PSNR (↑)	MSE (↓)	PSNR (↑)	MSE (↓)
SD Edit (OxfordIIITPets)	25.67 ± 1.14	0.09 ± 0.001	26.16 ± 0.72	0.09 ± 0.003
Ours (OxfordIIITPets)	$\mathbf{28.71 \pm 0.86}$	$\mathbf{0.03 \pm 0.005}$	$\mathbf{28.65 \pm 0.49}$	$\mathbf{0.03 \pm 0.002}$
SD Edit (AFHQv2)	27.12 ± 0.59	0.05 ± 0.006	27.49 ± 0.27	0.04 ± 0.004
Ours (AFHQ v2)	$\mathbf{27.65 \pm 0.57}$	$\mathbf{0.04 \pm 0.004}$	$\mathbf{28.07 \pm 0.32}$	$\mathbf{0.03 \pm 0.006}$

Table 5: Style Transfer results for Ghibli and van-Gogh styles



We obtain effective style transfer by transferring the merger time computed only on the source content, eliminating costly grid searches.

Appendix : Fluctuations and Mergers

Let ρ represent a given behaviour (**state**) of a system. We analyse the expectation of the n-th **Fluctuation Tensor,**

$$\mathbb{E}[\mathscr{F}_{\rho}^{(n)}(\omega)] := \mathbb{E}[\bigotimes_{k=1}^{n} \big(\rho(\omega) - \mathbb{E}[\rho]\big)] \,. \qquad n = 2 \text{ reduces to the Covariance matrix.}$$

To understand how structures form or dissolve, we compare the dynamics of **conditional** fluctuations of two distinct events in the data space using the **absolute normalised cross-fluctuation**.

$$\mathscr{M}_{\rho}^{(n)}(\Omega_1, \Omega_2) := \frac{\left| \langle \mathbb{E}_1[\mathscr{F}_{\rho}^{(n)}], \mathbb{E}_2[\mathscr{F}_{\rho}^{(n)}] \rangle_{\mathscr{H}_n} \right|}{\| \mathbb{E}_1[\mathscr{F}_{\rho}^{(n)}] \|_{\mathscr{H}_n} \| \mathbb{E}_2[\mathscr{F}_{\rho}^{(n)}] \|_{\mathscr{H}_n}} \,. \qquad \mathbb{E}_k \text{ is expectation conditioned on } \Omega_k$$

For our analysis, we set $\rho(\omega) = \omega$, reducing our metric to comparing standard cross-moments. If we consider the dynamics of cross-moments in forward time, then theory suggests these *merge* only at $t \to \infty$, but due to numerical effects we observe a finite time merger which has interesting practical consequences.

Appendix: Mergers in Diffusion

We analyse the tractable forward process. Time-reversal symmetry and proven model fidelity (Chen et al., 2022a) guarantee our findings on merger times of cross-moments of events transfer directly to the learned sampler.

$$\widetilde{\mathscr{M}}_{\rho}^{(n)}(i) = \begin{cases} \mathscr{M}_{\rho}^{(n)}\big(\Omega_{1,i}, \Omega_{2,i}\big), & d_n\Big(\mathbb{E}_1[\mathscr{F}_{\rho}^{(n)}(\Omega_{1,i})], \mathbb{E}_2[\mathscr{F}_{\rho}^{(n)}(\Omega_{2,i})]\Big) > \varepsilon, \\ 1, & \text{otherwise,} \end{cases}$$

d_n is induced from $\mathscr{H}_n$. The merger time $i^\star$ is the earliest discrete step where distance falls below the fixed threshold $\epsilon > 0$.

As the process is isotropic, for any bijective ρ, $i^\star$ is **unique**.

Appendix: The geometry of mergers

Main Theorem: Let $p_{k,i}$, $k \in \{1,2\}$be the conditional laws of $\Omega_{k,i}$ at forward step i. Under the existence of all moments upto order $n+1$, we have that,

$$\widetilde{\mathscr{M}}_{\rho}^{(n)}(i) \to 1 \implies d_{TV}(p_{1,i}, p_{2,i}) \to 0.$$

Here $\rho(\omega) = \omega$, and d_{TV} is the **Total-Variation Distance.**

Proof Sketch: We show that topologically $\widetilde{\mathscr{M}}_{\rho}^{(n)}(i) \to 1 \iff \|\mathbb{E}_1[\mathscr{F}_{\rho}^{(n)}(\Omega_{1,i})] - \mathbb{E}_2[\mathscr{F}_{\rho}^{(n)}(\Omega_{2,i})]\|_{\mathscr{H}_n} \to 0.$ Thus, the central moments converge if a merger happens. By the properties of characteristic functions and Essen's Inequality (Ibragimov 1975), assuming the existence of all moments upto order $n+1$, the desired convergence follows by considering the centred laws directly.

Appendix: Scale-Bridging

We also show a stronger result: the distance between moment structures **contracts exponentially** over time.

$$\|\mathbb{E}_i[\mathscr{F}_\rho^{(n)}(\Omega_i(t))] - \mathbb{E}_j[\mathscr{F}_\rho^{(n)}(\Omega_j(t))]\|_{\mathscr{H}_n} \leq C_{ij} e^{-\lambda t} \, \|\mathscr{F}_\rho^{(n)}\|_{L^2(\mu, \mathscr{H}_n)}$$

Here ρ is any Lipschitz random variable well defined for all time steps. μ is the gaussian measure, C_{ij} is a constant depending on the initial densities and λ is a parameter related to the diffusion process called the spectral gap.

This predictable collapse provides a consistency that allows for bridging across scales by the implicit construction of a neighbourhood based graph.

Our method is thus a temporal analogue to manifold learning, just as neighbourhood clustering methods like t-SNE/UMAP connect points in space to reveal structure, we can track when distributions become "nearby" in time!