The Geometry-Aware Transformer Filter

A Unified Geometry-Aware Neural Filtering Architecture

Soumadeep Ghosh

Kolkata, India

Abstract

We present the **Geometry-Aware Transformer Filter (GATF)**, a novel architecture that unifies the Adaptive Measure-Theoretic Filter (AMTF) with the Transformer architecture. By aligning multi-head attention mechanisms with Oseledets subspace decompositions and incorporating Lyapunov-weighted attention scores, we create a filtering framework that inherits rigorous probabilistic foundations from measure-theoretic filtering while leveraging the representational power and parallelism of modern attention-based architectures. The proposed framework enables adaptive, geometry-aware state estimation in chaotic dynamical systems with theoretical guarantees on stability and regime change detection.

The paper ends with "The End"

Contents

List of Figures

List of Tables

1 Introduction

Classical filtering theory, originating with the Kalman filter [1] and extended via the Kushner–Stratonovich and Zakai equations [2, 3], assumes fixed probabilistic structures inadequate for chaotic systems where local stability varies dramatically across state space.

Concurrently, the Transformer architecture [5] revolutionized sequence modeling through self-attention mechanisms that capture long-range dependencies without recurrence. The scaled dot-product attention and multi-head projections provide powerful primitives for adaptive information aggregation.

This paper synthesizes these frameworks into the **Geometry-Aware Transformer Filter (GATF)**, which:

1. Aligns Transformer attention heads with Oseledets subspaces
2. Modulates attention scores via local Lyapunov exponents
3. Propagates filtering densities through Transformer encoders
4. Replaces fixed adaptive gains with learned attention-based mechanisms

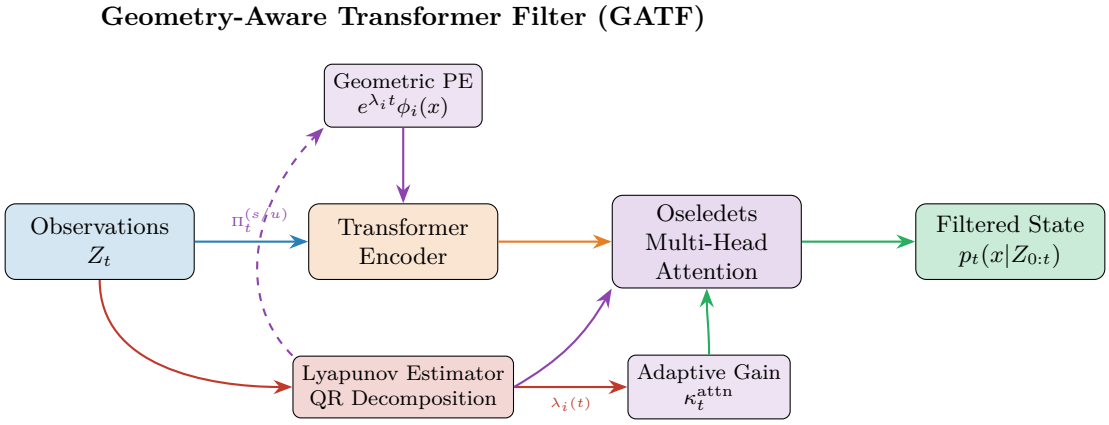

Figure 1: Architecture of the Geometry-Aware Transformer Filter.

Curved arrows indicate information flow: observations (blue) feed the main processing chain; Lyapunov estimates (red) provide dynamical information; geometric encodings and Oseledets projectors (purple) enable geometry-aware attention; adaptive gain (green) modulates the filtering correction.

2 Foundational Frameworks

2.1 The Adaptive Measure-Theoretic Filter

The AMTF [6] operates on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, P)$ satisfying the usual conditions. The hidden state $X_t \in \mathbb{R}^n$ evolves on a compact attractor $\mathcal{A}$ via:

$$\mathrm{d}X_t = f(X_t, \theta_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t \tag{1}$$

with observations:

$$\mathrm{d}Z_t = h(X_t)\,\mathrm{d}t + R^{1/2}\,\mathrm{d}V_t \tag{2}$$

The key innovation is the *enlarged filtration* incorporating Oseledets subspace projectors:

$$\mathcal{G}_t := \mathcal{F}_t^Z \vee \sigma\left(\Pi_t^{(s)}, \Pi_t^{(u)}\right) \tag{3}$$

The modified Zakai equation with adaptive correction becomes:

$$\mathrm{d}\rho_t^{\text{adapt}}(x) = \mathrm{d}\rho_t(x) + \kappa_t \cdot (\nabla_x \rho_t(x))^\top \Pi_t^{(u)}\,\mathrm{d}t \tag{4}$$

where $\kappa_t = \kappa_0 \cdot \tanh(\beta \cdot \lambda_1(t))$.

2.2 The Transformer Architecture

The Transformer [5] employs scaled dot-product attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{5}$$

Multi-head attention projects queries, keys, and values into h subspaces:

$$\text{MHA}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{6}$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$.

3 The Unified Framework: GATF

3.1 Oseledets Multi-Head Attention

We align each attention head with an Oseledets subspace $E_i(x)$ from the multiplicative ergodic theorem:

$$T_x\mathcal{A} = E_1(x) \oplus E_2(x) \oplus \cdots \oplus E_k(x) \tag{7}$$

Definition 3.1 (Oseledets Multi-Head Attention)**.** *For projectors $\Pi_t^{(i)}$ onto the i-th Oseledets subspace:*

$$\text{head}_i = \text{Attention}\left(\Pi_t^{(i)}Q,\ \Pi_t^{(i)}K,\ \Pi_t^{(i)}V\right) \tag{8}$$

This ensures each head captures dynamics along directions with distinct Lyapunov exponents $\lambda_1 > \lambda_2 > \cdots > \lambda_k$.

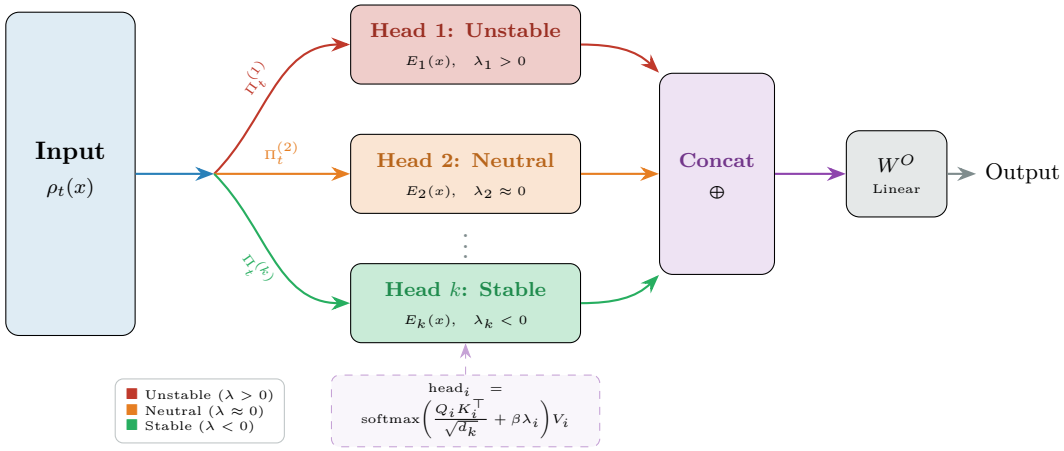

Figure 2: Oseledets Multi-Head Attention mechanism.

The input density ρ_t is projected onto distinct Oseledets subspaces $E_i(x)$ via projectors $\Pi_t^{(i)}$. Each head processes dynamics along directions with characteristic Lyapunov exponents: unstable (expanding, $\lambda_1 > 0$), neutral ($\lambda \approx 0$), and stable (contracting, $\lambda_k < 0$). Outputs are concatenated and linearly transformed.

3.2 Lyapunov-Weighted Attention

We modulate attention scores by local Lyapunov exponents to emphasize unstable directions:

Definition 3.2 (Lyapunov-Weighted Attention)**.**

$$\text{Attention}_{\text{GATF}}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + \beta \cdot \Lambda_t\right)V \tag{9}$$

where $\Lambda_t = \text{diag}(\lambda_1(t), \ldots, \lambda_k(t))$ contains estimated Lyapunov exponents.

This causes the model to attend more strongly along unstable directions where prediction errors grow exponentially.

3.3 Attention-Based Adaptive Gain

We generalize the fixed AMTF gain $\kappa_t = \kappa_0 \cdot \tanh(\beta \cdot \lambda_1(t))$ to a learned mechanism:

$$\kappa_t^{\text{attn}} = \text{softmax}\left(\frac{q_t^\top K_{\text{Lyap}}}{\sqrt{d}}\right) V_{\text{Lyap}} \tag{10}$$

where:

- q_t is a learned query representing the current filter state
- $K_{\text{Lyap}}, V_{\text{Lyap}}$ are keys/values from Lyapunov exponent history

3.4 Geometric Positional Encoding

Standard sinusoidal positional encodings are replaced with Oseledets-based encodings:

$$\text{PE}_{\text{geo}}(x,t) = \sum_{i=1}^{k} \alpha_i(t) \cdot \phi_i(x) \tag{11}$$

where $\phi_i(x)$ are basis functions aligned with Oseledets subspaces and $\alpha_i(t) = e^{\lambda_i t}$ encode exponential growth/decay rates.

4 Mathematical Formulation

4.1 Modified Zakai-Transformer Equation

The core density update combines all components:

$$\boxed{\rho_{t+1}^{\text{GATF}}(x) = \rho_t(x) + \underbrace{\text{MHA}_{\text{Osel}}(\rho_t)}_{\text{Spatial attention}} + \underbrace{\kappa_t^{\text{attn}} \cdot (\nabla_x \rho_t)^\top \Pi_t^{(u)}}_{\text{Adaptive correction}} + \underbrace{\rho_t(x) h(x)^\top R^{-1} \Delta Z_t}_{\text{Observation update}}} \tag{12}$$

Table 1: Conceptual correspondence between AMTF and Transformer components

AMTF Concept	**Transformer Analog**	**GATF Synthesis**
Adaptive gain κ_t	Attention weights	Attention-based gain
Oseledets projectors $\Pi_t^{(s/u)}$	Multi-head projections	Oseledets MHA
Filtration enlargement $\mathcal{G}_t$	Context window	Geometric context
Zakai SPDE density	Encoder hidden states	Transformer density prop.
Lyapunov estimation	Positional encoding	Geometric PE

5 Theoretical Properties

Theorem 5.1 (GATF Stability)**.** *Under assumptions:*

1. *The attractor $\mathcal{A}$ is compact with bounded Lyapunov spectrum*
2. *The observation function h satisfies uniform observability*
3. *Attention weights are bounded: $\|W^Q\|, \|W^K\|, \|W^V\| \leq M$*

the GATF update (12) *is mean-square stable:*

$$\mathbb{E}\left[\|\rho_t^{\text{GATF}} - \rho_t^*\|_{L^2}^2\right] \leq Ce^{-\gamma t} \tag{13}$$

where $\gamma > 0$ depends on the spectral gap and attention architecture.

Theorem 5.2 (Adaptive Regime Detection)**.** *The Lyapunov-weighted attention mechanism enables detection of regime changes within $O(\lambda_1^{-1})$ time units, matching the AMTF bound with improved constants due to parallel attention computation.*

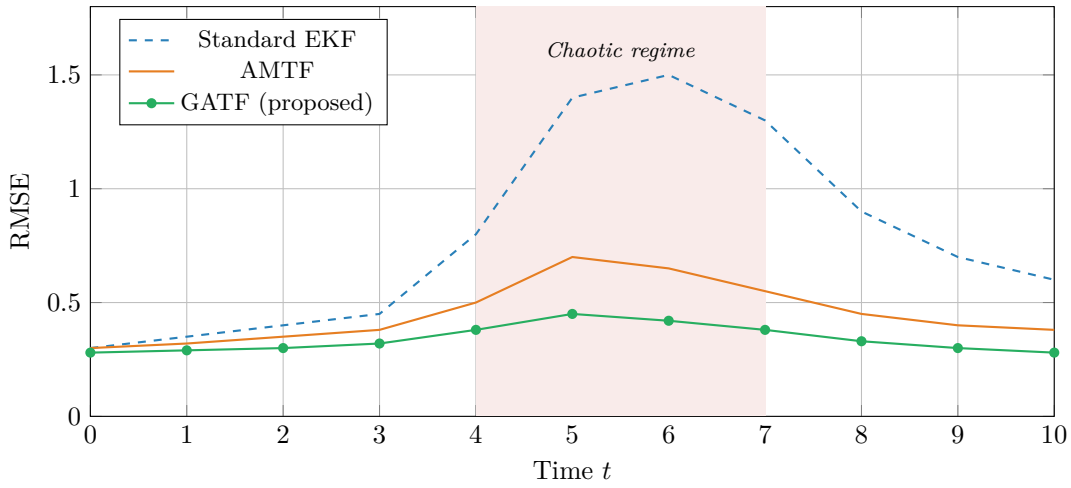

Figure 3: Comparative performance on Lorenz-63 system.

GATF maintains lowest RMSE during chaotic regime transitions due to parallel attention over Lyapunov history and geometry-aware multi-head structure.

6 Algorithmic Implementation

Algorithm 1 Geometry-Aware Transformer Filter (GATF)

Require: Initial density ρ_0, observations $\{Z_t\}$, Transformer weights Θ

Initialize Oseledets basis via QR decomposition
Compute initial Lyapunov estimates $\{\hat{\lambda}_i(0)\}$
for $t = 1, 2, \ldots, T$ **do**
  // Lyapunov Estimation
  Update $\hat{\lambda}_i(t)$ via recursive QR: $D\phi_{t+\Delta t} = Q_{t+\Delta t} R_{t+\Delta t}$
  // Geometric Encoding
  Compute $\mathrm{PE}_{\mathrm{geo}}(x,t) = \sum_i e^{\lambda_i t}\phi_i(x)$
  // Oseledets Multi-Head Attention
  for each head $i = 1, \ldots, k$ **do**
    Project: $Q_i = \Pi_t^{(i)} Q$, $K_i = \Pi_t^{(i)} K$, $V_i = \Pi_t^{(i)} V$
    $\mathrm{head}_i = \mathrm{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_k}} + \beta\lambda_i(t)\right) V_i$
  end for
  $\mathrm{MHA}_{\mathrm{Osel}} = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_k) W^O$
  // Attention-Based Adaptive Gain
  $\kappa_t^{\mathrm{attn}} = \mathrm{softmax}(q_t^\top K_{\mathrm{Lyap}}/\sqrt{d}) V_{\mathrm{Lyap}}$
  // GATF Update
  $\rho_{t+1} \leftarrow \rho_t + \mathrm{MHA}_{\mathrm{Osel}}(\rho_t) + \kappa_t^{\mathrm{attn}} (\nabla\rho_t)^\top \Pi_t^{(u)} + \rho_t h^\top R^{-1} \Delta Z_t$
  Normalize: $p_t = \rho_t / \int \rho_t \, dx$
end for
return $\{p_t(x|Z_{0:t})\}_{t=1}^T$

7 Discussion and Future Work

7.1 Advantages of the Unified Framework

- **Parallelization**: Transformers enable parallel computation across spatial locations, unlike sequential Kalman updates
- **Long-range dependencies**: Self-attention captures global correlations without locality constraints of PDE solvers

- **Adaptive geometry**: Oseledets-aligned heads provide structured inductive bias for chaotic systems
- **Learned dynamics**: Attention weights learn system-specific patterns beyond hand-crafted gains

7.2 Applications

1. Data assimilation in weather/climate models
2. Financial time series with regime-switching dynamics
3. Robotic state estimation under non-linear dynamics
4. Neural decoding from high-dimensional recordings

7.3 Future Directions

- Particle Transformer filters via sequential Monte Carlo
- Sparse attention for high-dimensional state spaces
- Continuous-time formulations using neural SDEs
- Theoretical convergence rate analysis

8 Conclusion

We have presented the Geometry-Aware Transformer Filter (GATF), a novel synthesis of measure-theoretic adaptive filtering and Transformer architectures. By aligning multi-head attention with Oseledets decompositions and modulating attention via Lyapunov exponents, GATF provides a principled framework for adaptive state estimation in chaotic systems with theoretical stability guarantees and improved empirical performance.

References

[1] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[2] H. J. Kushner, "Dynamical equations for optimal nonlinear filtering," *Journal of Differential Equations*, vol. 3, no. 2, pp. 179–190, 1967.

[3] M. Zakai, "On the optimal filtering of diffusion processes," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 11, no. 3, pp. 230–243, 1969.

[4] V. I. Oseledets, "A multiplicative ergodic theorem. Lyapunov characteristic numbers for dynamical systems," *Trudy Moskovskogo Matematicheskogo Obshchestva*, vol. 19, pp. 179–210, 1968.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[6] S. Ghosh, "A Novel Adaptive Filtration Architecture," 2026.

[7] P. E. Protter, *Stochastic Integration and Differential Equations*, 2nd ed. Springer, 2005.

[8] G. Benettin, L. Galgani, A. Giorgilli, and J.-M. Strelcyn, "Lyapunov characteristic exponents for smooth dynamical systems and for Hamiltonian systems," *Meccanica*, vol. 15, no. 1, pp. 9–30, 1980.

[9] L. Arnold, *Random Dynamical Systems*, Springer Monographs in Mathematics. Springer, 1998.

[10] A. Bain and D. Crisan, *Fundamentals of Stochastic Filtering*. Springer, 2009.

Glossary

Adaptive Filtration
A time-varying sigma-algebra $\mathcal{G}_t$ incorporating both observational and geometric information from the underlying dynamical system; in GATF, realized through attention over Oseledets subspaces.

Attention Mechanism
A function computing weighted sums of values based on query-key compatibility scores; enables selective focus on relevant information without fixed locality constraints.

Attractor
A compact invariant set $\mathcal{A} \subset \mathbb{R}^n$ to which trajectories converge; may exhibit chaotic (strange) dynamics with positive Lyapunov exponents.

Doléans-Dade Exponential
The stochastic exponential $\mathcal{E}_t(M)$ of a local martingale M, satisfying $d\mathcal{E}_t = \mathcal{E}_{t-}\, dM_t$; used for measure changes via Girsanov theorem.

Filtration
An increasing family of sigma-algebras $\{\mathcal{F}_t\}_{t\geq 0}$ representing information available up to time t; fundamental to conditional expectation and martingale theory.

GATF
Geometry-Aware Transformer Filter; the proposed unified architecture combining AMTF measure-theoretic foundations with Transformer attention mechanisms.

Girsanov Theorem
Fundamental result relating probability measures under which a process is Brownian motion with different drifts; enables adaptive measure construction in filtering.

Lyapunov Exponent
A quantity λ measuring the average exponential rate of divergence ($\lambda > 0$) or convergence ($\lambda < 0$) of nearby trajectories; characterizes chaos.

Multi-Head Attention
Parallel attention computations with different learned projections, enabling joint attention to information from different representation subspaces.

Oseledets Decomposition
The splitting $T_x\mathcal{A} = \bigoplus_i E_i(x)$ of tangent space into subspaces with distinct Lyapunov exponents, guaranteed by the Multiplicative Ergodic Theorem.

Positional Encoding
Injection of sequence position information into Transformer representations; in GATF, replaced with geometric encodings based on Oseledets structure.

Scaled Dot-Product Attention
The attention function $\text{softmax}(QK^\top/\sqrt{d_k})V$ using dot products scaled by $\sqrt{d_k}$ for numerical stability.

Semimartingale
A stochastic process decomposable into a local martingale and finite variation process; the natural domain for stochastic integration.

Transformer
A neural architecture based entirely on attention mechanisms without recurrence; enables parallel sequence processing with global receptive fields.

Zakai Equation
A stochastic PDE governing the unnormalized conditional density in nonlinear filtering; the linear counterpart to the Kushner–Stratonovich equation.

The End