Hey there,

Here is your disruptive ML/AI digest for the day. Here's what you folks might find interesting today.

TOP PICKS

Theory and Generalization

PAC-Bayes Bounds on Quotient Parameter Spaces: Geometry-induced Implicit-Bias Priors (Aladrah, Anselmi)

This paper addresses a fundamental gap in PAC-Bayesian analysis of overparameterized models: when different parameter configurations yield the same predictor, the standard KL term in PAC-Bayes bounds is inflated by purely parametrization-level differences that carry no information about generalization. The authors show that performing the analysis on the quotient predictor space removes this spurious KL contribution. The key conceptual contribution is constructing a canonical prior that accounts for the geometric volume of equivalent parametrizations, effectively turning the model's implicit bias into a data-independent prior. This is not just a tightening of bounds for its own sake -- it provides a principled bridge between the geometry of parameter spaces and generalization certificates, with empirical validation showing 21% tighter PAC-Bayes certificates in Fourier-Hadamard regression and conditional improvements in Query-Key attention trained with vanilla SGD.

On the Diverse Dynamical Behaviors Arising in Deep Linear Transformers (Li, Maranzato, Peszek, Teolis, Akkoc, Riedl, Ulukus, Garcia Trillos)

This work recasts inference-time token interactions in deep linear transformers as a generalized Kuramoto-type interacting particle system with second-harmonic coupling, in embedding dimension two. The connection to Watanabe-Strogatz theory reveals that the dynamics are intrinsically low-dimensional regardless of the key/query/value matrices, and the authors uncover a hidden Hamiltonian structure that governs oscillations and bifurcations. This is a genuine conceptual departure from the usual "attention as soft retrieval" framing: it provides a dynamical-systems lens through which long-time transformer behavior -- clustering, oscillation, bifurcation -- can be characterized and predicted. The structural stability result showing persistence of these behaviors near the Ott-Antonsen manifold, combined with numerical evidence in higher dimensions, suggests this framework may scale beyond the 2D setting where it is provable.

Novel Architectures and Objectives

1-Lipschitz Neural Networks on Hadamard Manifolds (Murari, Ghirardelli, Adcock, Celledoni, Owren, Schoenlieb)

The authors construct neural network layers that are exactly 1-Lipschitz on Hadamard (nonpositively curved) manifolds, using Busemann functions as the core building block. This is architecturally distinct from existing Lipschitz-constrained networks, which operate in Euclidean space and typically rely on spectral normalization or gradient penalty tricks. The Busemann gradient flow construction yields layers that are quasi-alpha-firmly nonexpansive, a property from monotone operator theory that provides stronger guarantees than mere Lipschitz continuity. Explicit constructions for hyperbolic space and the SPD manifold are given, with applications to robust classification on the Poincare disk and Plug-and-Play priors for covariance reconstruction. The geometric grounding matters: robustness on manifold-valued data (graphs, covariance matrices, hierarchical embeddings) requires stability guarantees that respect the ambient geometry, not Euclidean approximations.

Probabilistic and Statistical Methods

Signed Rectified Flow: Negativity-Controlled Generation (Liao, Su, Chen, Liu)

Signed Rectified Flow generalizes Rectified Flow to target a signed measure -- a convex combination of a distribution to promote and a distribution to suppress. While direct sampling from a signed measure is ill-defined, the authors show that the flow dynamics induce a valid generative process that concentrates mass where the signed measure is positive and creates exclusion barriers where it is negative. The charged-particle interpretation of the signed continuity equation provides the physical intuition: negative mass repels, forming boundaries the generative process cannot cross. This is a principled, theory-driven framework for incorporating negative information into flow-based generation, going beyond heuristic classifier guidance or negative prompting. The applications -- anti-memorization, adversarial prompt defense in Stable Diffusion, fidelity-diversity tradeoffs on ImageNet -- demonstrate that the theory translates to practical gains.

Emergent Behavior and Agentic AI

Elicitation without Backpropagation: Steering Model Behavior by Optimizing the Latent Posterior (Baker, Pathak, Murfet, Wei)

This paper introduces Posterior Prefix Tuning (PPT), a method for steering transformer behavior that requires no forward passes and no backpropagation through the transformer itself. The approach exploits the latent posterior model of transformer behavior -- where the next-token distribution arises from a posterior over latent predictive models -- in the specific setting of Bayes-filtered transformers (BFTs) meta-learned on hierarchical priors. Because the elicitation objective factors through the latent posterior, its gradient can be estimated entirely from prior samples via importance sampling, with the samples being utility-independent. This means a single set of samples drives elicitation for arbitrarily many utility functions at negligible marginal cost. The conceptual novelty is significant: it reframes prompt optimization as a problem of posterior manipulation rather than gradient-based search through the transformer's computational graph, and while currently limited to the BFT setting, it points toward a fundamentally different paradigm for model steering.

BY CATEGORY

cs.LG

- Attractor Geometry Determines the Identifiability Limits of System Discovery
- Relative Positions Generalize, Absolute Positions Memorize: An Implicit-Bias Account of Length Generalization in Attention
- Decafs: Disentangled Conditional adversarial Flows
- Enhanced NQS via Annealed Gradient Descent
- Conditioned Direct Feedback Alignment via Activity and Error Geometry
- Algebraic Signatures for Structural Learning in Probability Tensors
- The Tractability Landscape of Sampling with Inexact Scores
- Neural Kolmogorov Equations: Parallelizable Learning of Stochastic Dynamics under General Noise
- Riemannian Deep Learning: Modules, Networks, and Geometries
- Spectral Higher-Order Neural Networks Have Sharp Expressivity Bounds

stat.ML

- The Price of Hidden Curvature: An Omega(d^{5/4} sqrt(T)) Lower Bound for Bandit Convex Optimization
- Provable diffusion-based posterior sampling for linear inverse problems via DDIM
- Mixing-Free and Signal-Optimal Learning of Gaussian Graphical Models from Glauber Dynamics

- Boundary-Adapted PINNs for Elliptic Dirichlet Problems: H^2(Omega) A Priori Error Bounds with Application to Mean Escape Time Computation
- Fundamental limits of distributed multiclass classification from simple binary decisions
- Optimizing Regret
- Uncertainty quantification in mechanics: A unified Bayesian perspective
- On the sensitivity of machine-learned probabilistic weather forecast models to scale-aware scoring rules
- Deep learning-based prediction of time-resolved adhesive forces in viscoelastic Hertzian contacts
- Some cautionary tales about Bayesian predictive inference
- Using binary silver labels in electronic health records-based computable phenotyping algorithms

---

Back tomorrow with more!

P.S. In the most recent submission window: cs.LG: 126 papers stat.ML: 15 papers