Hi all,

Here is your disruptive ML/AI digest for the day.

Here's what you folks might find interesting today.

# TOP PICKS

## Theory and Generalization

### Learning to Reason with Curriculum II: Compositional Generalization

This one looks especially relevant if you care about formal accounts of reasoning rather than just benchmark-level claims. The paper studies compositional generalization through the concrete problem of simulating semiautomata, then proves that an autocurriculum strategy can beat direct learning by a large margin in statistical complexity. The striking claim is that, with intermediate-state feedback, curriculum reduces supervision from an unavoidable linear-in-T regime to a subpolynomial one, which is a real conceptual shift rather than an optimization trick. They also give an RL-with-verifiers style result showing that curriculum weakens the coverage requirements on the reference model from full-horizon competence to short-block competence. That makes it relevant both to chain-of-thought theory and to agentic training setups where verifiable rewards are available. Overall, this is one of the clearest theory papers here on when decomposition genuinely changes learnability.

### Disentangling Continuous-Time Latent Dynamics: Identifiability of Latent SDEs via Diffusion Shifts

This is a strong fit for causal representation learning and identifiability. The key novelty is that it gets latent-coordinate identifiability in continuous-time latent SDEs from environment-dependent shifts in diffusion covariance, while allowing an unknown nonlinear observation diffeomorphism. The result is notable because it does not rely on sparsity of the drift, which is often where identifiability arguments become brittle. They first prove the idea in Ornstein-Uhlenbeck systems and then extend it to general additive-noise latent SDEs, with identifiability of the instantaneous drift-Jacobian graph up to permutation and scaling. That gives a fairly clean bridge between continuous-time stochastic dynamics and causal graph recovery. If you are tracking theory that could matter for time-series representation learning beyond discrete-time SCMs, this is one of the sharper papers in the batch.

### The Curse of Multiple Mediators: Hidden Interaction Effects in Activation Patching

This is one of the more useful mechanistic-interpretability papers here because it questions a core tool rather than proposing another patching variant. The authors re-derive activation patching from causal mediation analysis and show that the usual natural indirect effect estimate inevitably mixes single-component effects with interaction terms from other components. That means a lot of standard patching-based causal attributions can be unstable or misleading exactly when mechanisms are distributed and conditional. They go further by proving when these interaction effects become small, when they scale up, and why greedy component ranking can miss mechanisms that only appear combinatorially. The GPT-2 IOI case study makes the critique concrete rather than purely philosophical. If you care about interpretability grounded in causal theory, this is the kind of paper worth reading closely.

## Probabilistic and Statistical Methods

### Local Fokker-Planck Geometry for Score Estimation: Heat-Ball Mean-Value Representations and Exact High-Dimensional Sampling

This paper is unusually ambitious on the objective side of score modeling. Instead of relying on global denoising-score-matching style averages, it develops a local geometric framework based on parabolic averaging of the Fokker-Planck equation, yielding exact local mean-value representations for the score, density, log-density, and entropy density. The conceptual move is to replace global residual penalties with a one-parameter local family whose small-radius limit recovers the usual pointwise Fokker-Planck residual, which gives a principled generalization rather than a heuristic tweak. The exact factorized sampler for the induced heat-ball integrals is also notable, since high-dimensional Monte Carlo is often where elegant theory dies in practice. This feels relevant if you are interested in diffusion objectives derived from PDE structure and information geometry rather than empirical recipe tuning. Among today's generative-model papers, this is probably the most mathematically distinctive.

**Difference of Convex Programming in the Wasserstein Space with Applications to MMD Optimization**

This is a good pick if you like optimization over probability measures and geometry-aware learning objectives. The main contribution is to lift classical difference-of-convex programming, specifically CCCP, into Wasserstein space for nonconvex functionals that admit a DC decomposition. The paper then works out explicit decompositions for MMD and Energy Distance, proves almost-stationarity and local convergence under assumptions on the convex pieces, and shows empirically that the resulting scheme can be more stable than plain Wasserstein gradient descent. What makes it stand out is that it is not just "use MMD for X", but a structural optimization framework for measure-valued objectives. That is directly aligned with interests in optimal transport, f-divergence style geometry, and principled alternatives to standard first-order training. It is also one of the cleaner examples today of importing a classical optimization idea into modern probabilistic ML in a nontrivial way.

# BY CATEGORY

### Novel Architectures and Objectives

Sampling the Schwinger Model with Gauge-Equivariant Diffusion

A Unified Framework for Vision Transformers Equivariant to Discrete Subgroups of O(2)

Masked Language Flow Models

Scalable and Differentiable Point-Cloud Registration Using Maximum Mean Discrepancy

Physics-constrained neural networks for surrogate modeling of lossless periodic structures

Accelerating Hierarchical Sparse Predictive Coding with Hybrid Amortized Inference

### Theory and Generalization

PAC-Bayesian Certificates for Quadratic Closed-Loop Control

How Width and Data Shape Generalization Scaling Laws in Quadratic Neural Networks

On the Inseparability of Instructions and Data in Shared-Embedding Sequence Models

Dangerous Liaisons of Convex Learning and Non-Affine Aggregation

Surprises in Proper Positive-Only Learning

## Probabilistic and Statistical Methods

[VGB for Masked Diffusion Model: Efficient Test-time Scaling for Reward Satisfaction and Sample Editing]

[Directed Graph Topology Inference via Graph Filter Identification]

[Adversarial Contamination Meets Hard Thresholding: An Iterative Algorithm with Signal Adaptivity and Minimax Optimality]

[The Decision Geometry of Covariance Estimation for the Global Minimum-Variance Portfolio under Heavy Tails]

## Transferable Methods from Adjacent Fields

[Reduction of Probabilistic Chemical Reaction Networks]

## Deprioritized / Lower Priority for This Profile

[Odyssey: Constructing Verifiable Local Truth-Preserving Foundation Models]

[Benchmarking on Tasks That Matter: Dataset Selection for Preserving Model Rankings]

Back tomorrow with more!

P.S. In the most recent submission window: cs.LG: 124 papers stat.ML: 10 papers