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

Safety from Honesty in a Disinterested AI Predictor

This is one of the clearest attempts in today's batch to make a *formal* safety case for a non-agentic AI objective rather than just proposing another alignment heuristic. The core idea is to train a "Scientist AI" Predictor to approximate a Bayesian posterior over epistemically contextualized statements, so goals expressed in text are treated as evidence to model rather than objectives to internalize. What makes it especially relevant is that the paper ties safety to the training objective itself: downstream deployment effects never become reward signals, and the argument is that this structurally suppresses emergent agency and coordinated deception. The abstract also claims an actual probabilistic guarantee under assumptions on training dynamics and sparsity of dangerous predictors, which puts it squarely in the theory-first alignment bucket rather than empirical safety evals. **Yoshua Bengio** is on this paper, which is worth noting given your interest in conceptually new safety paradigms with formal grounding. **Relevance score: 9.7/10**

Dead-Direction Conditioners: Gauge-Equivariant Preconditioning for Deep Networks

This paper looks unusually aligned with your preference for optimization methods derived from geometry rather than empirical optimizer tinkering. The main contribution is a preconditioning framework that explicitly respects parameter-space gauge symmetries such as ReLU rescaling, LayerNorm scale, logit shift, and per-head attention rotations, so optimization happens on the symmetry quotient rather than drifting along redundant directions. That is a real conceptual departure from standard Adam-style coordinatewise conditioning: the optimizer state is decomposed according to a G-invariant metric, and the paper claims exact equivariance on top of Adam plus a compatible construction for Muon. What makes it especially compelling is that the geometry is not just decorative; the authors argue it changes both the minima reached and what can be measured there, including "dead-direction rates" tied to singular learning dynamics. If the claims hold, this could matter both for practical training stability and for a more principled theory of optimization in overparameterized networks. **Relevance score: 9.4/10**

A Stochastic-Geometric Theory of Scaling Laws in Grokking

This is a strong fit for the "training dynamics with real theory" category. The paper proposes a shell-core topology of reachable solutions induced by Adam plus weight shrinkage, where memorization and generalization occupy distinct geometric regions in parameter space, and grokking corresponds to a stopping-time transition between them. That is more mechanistic than the usual empirical grokking papers because it tries to explain the delayed phase transition through optimization geometry and stochastic process arguments, not just curve-fitting observations. The payoff is a derivation of scaling laws for learning rate, batch size, and L2 regularization, with experiments positioned as validation rather than the main contribution. If you're tracking work that turns emergent training phenomena into analyzable geometric objects, this is one of the better candidates today. **Relevance score: 8.9/10**

SGD Provably Prioritizes a Shortcut Spurious Feature in the XOR Model

This is a clean theory paper on shortcut learning that seems more valuable than the many empirical spurious-correlation papers because it gives an end-to-end optimization account. In a two-layer ReLU network trained by online minibatch SGD, the authors show that a linear spurious feature is learned first and exponentially fast, and that this actively suppresses learning

of the true XOR signal. The interesting part is the phase-transition analysis: the mechanism is not just "SGD likes easy features", but a specific coupling between first-layer feature growth, second-layer sign alignment, and majority-group margin effects. That makes it relevant if you're interested in implicit bias and why optimization dynamics produce brittle representations even in simple settings. It is narrow in setup, but the mechanistic clarity is exactly the kind of thing that can transfer to broader theory. **Relevance score: 8.6/10**

Probabilistic and Statistical Methods

The Fundamental Limits of Valid Transport Map Estimation

This is probably the most theoretically useful generative-model paper in the list. Rather than proposing yet another transport-based method, it asks a more foundational question: if we only care about learning *some* valid transport map rather than the optimal transport map, when does that actually buy us statistical efficiency? The answer, in the abstract, is surprisingly sharp: under standard OT stability assumptions, estimating any valid transport map is minimax-equivalent to estimating the OT map itself, which gives lower bounds that apply broadly to diffusion, flow matching, and related methods. That is exactly the kind of abstraction that can clarify whether current generative modeling objectives are solving an intrinsically hard problem or just hiding it behind different parameterizations. The paper also identifies the regime where suboptimal maps really can be easier, so it is not just a negative result. **Relevance score: 9.1/10**

BY CATEGORY

Novel Architectures and Objectives

Geometric Algebra Meets Cartesian Tensors: Higher-Order Equivariance for Interatomic Potentials

Residual-Guided Dictionary Learning for Spectrally Accurate Koopman Approximation

Predictive Objectives Discard Exogenous Control-Relevant Features: A Controlled Mechanistic Study

ScaleAware-JEPA: Latent Representation for Discovery in Multiscale Physical Fields

Theory and Generalization

Mechanistically Eliciting Latent Behaviors in Language Models

Sample Complexity of Scientific Discovery: PAC Learnability of Compositional Function Trees

I-BBS: Coordinate-Free Inference of Latent Sub-Manifolds Using Random Distance Matrix Theory

Non-parametric recovery of causal diffusion mechanisms from steady-state observations

Optimization Dynamics Imprint Semantic Specificity in Contrastive Embedding Norms

Convergence of Continual Learning in Homogeneous Deep Networks

Optimizer Memory Makes Shuffle Order a First-Order Source of Fine-Tuning Noise

What LLMs explain is not what they believe: Evaluating explanation sufficiency under models' own input beliefs

Probabilistic and Statistical Methods

ITSPACE: Monotone Gaussian Optimal Transport Updates

Few-Step Boltzmann Generators via Scalable Likelihood Flow Maps

Variance Reduction for Stochastic Gradient Generalized Non-reversible Langevin Monte Carlo Algorithms

Quantum ML and Hybrid Paradigms

Learning the structure of open quantum systems

A Machine-Verified Proof of a Quantum-Optimization Conjecture

Emergent Behavior and Agentic AI

How AI settled the complexity of the oldest SGD algorithm

Back tomorrow with more!

P.S. In the most recent submission window: cs.LG: 250 papers stat.ML: 49 papers