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

An exact information theory of generalization phase transitions in Bayesian diffusion models

This is the clearest theoretical paper in the batch for anyone tracking why diffusion models generalize at all instead of just memorizing finite datasets. The core novelty is the BIRD construction: an analytically tractable Bayesian reverse-diffusion model with explicit information restriction, which lets the authors derive a sharp memorization-generalization phase boundary in terms of mutual information and dataset size. That is much more interesting than another empirical diffusion scaling paper, because it gives a concrete criterion for when generation should generalize: memorization occurs when restricted noisy observations carry more than log N bits about the training set. The paper also claims that early-training UNets and DiTs are well approximated by these spatially local BIRD models, which, if the approximation is robust, makes the theory relevant to actual modern architectures rather than toy models only. I would read this for the information-theoretic objective itself and for the claim that diffusion generation operates near an edge-of-memorization regime. The main caveat is that the strongest architectural relevance is explicitly to *early training*, so the theory may not yet explain the full trained-model regime.

Dynamics of Gradient Descent with Large Step Size Near a Manifold of Flat Minima

This is a serious optimization theory paper rather than a loose landscape metaphor paper. Its main contribution is extending large-step gradient descent analysis from isolated flat minima and scalar outputs to vector-valued least-squares problems and, more importantly, to *manifolds* of flat minima, which is the geometrically relevant setting for many overparameterized models. The technical centerpiece appears to be a generalized normal form plus new convergence theorems, along with a novel method for solving a singular PDE that arises in the analysis. The application to deep matrix factorization is especially relevant because it yields structural statements about the flat-minima set being a fiber bundle and about sharpness having Morse-Bott structure along that manifold. That combination of optimization dynamics and differential-geometric structure is exactly the kind of non-incremental theory that can transfer beyond the immediate model class. It is narrower in direct practical impact than the diffusion paper above, but stronger in mathematical specificity.

Explaining Near-Zero Hessian Eigenvalues Through Approximate Symmetries in Neural Networks

This paper offers a concrete mechanism for the huge bulk of near-zero Hessian eigenvalues: weakly broken continuous symmetries of the parameterization, producing pseudo-Goldstone modes. That is a more satisfying explanation than generic "flatness" language because it ties spectral structure to identifiable symmetry subspaces and explicitly separates them from high-curvature directions. The authors construct exact zero modes in deep linear networks, then treat ReLU as a weak symmetry-breaking perturbation and show that the Hessian bulk remains almost entirely in the symmetry subspace. If that decomposition holds broadly, it sharpens how we should interpret Hessian spectra in deep nets and may help disentangle optimization-relevant curvature from parameterization artifacts. I would flag this as especially relevant if you care about implicit regularization, landscape geometry, or mechanistic interpretations of overparameterization. It is not a full training-dynamics theory, but it does provide a principled account of a long-standing empirical phenomenon.

PROBABILISTIC AND STATISTICAL METHODS

GradInf: Gradient Estimation as Probabilistic Inference

This is one of the more conceptually fresh papers here because it reframes gradient estimation for probabilistic programs as an inference problem, rather than as a bag of estimator tricks. The key move is a formal reduction using coupling and factorization, after which one can differentiate the solution of the induced inference problem to obtain a gradient estimator. That matters because it opens the door to importing the full machinery of probabilistic inference algorithms into gradient estimation, including for discrete choices and complicated stochastic dependencies where standard estimators are brittle. The paper also seems unusually rigorous on the systems side: the GradInf language is built around source-to-source transformations with denotational soundness proofs and information-flow typing to support partial evaluation and factorization. If the case studies are as strong as claimed, this could become a useful conceptual bridge between probabilistic programming, variational inference, and differentiable computation. This is exactly the kind of paper worth attention if you want new computational primitives rather than another local variance-reduction tweak.

Score Accuracy Along the Forward Diffusion Does Not Certify Numerical Stability in Diffusion Sampling

This paper makes a sharp negative point that many diffusion papers glide past: good score accuracy under forward marginals does *not* imply numerically stable reverse-time sampling after discretization. The authors construct examples where the learned reverse process is close in path-space total variation and nonexplosive, yet Euler-Maruyama samples have diverging positive moments, so weak convergence coexists with Wasserstein failure. That is a genuinely important distinction for anyone who takes score-matching guarantees at face value. The positive result is also useful rather than cosmetic: for compactly supported data, projecting the denoiser onto a known bounded convex support set restores moment control and Wasserstein convergence under mild conditions. I would treat this as a foundational cautionary paper about objective mismatch between training and sampling, with a concrete fix in a restricted setting. It is less broad than the BIRD paper, but arguably more immediately actionable for diffusion theory and evaluation.

BY CATEGORY

NOVEL ARCHITECTURES AND OBJECTIVES

Write-Protected Discrete Bottlenecks for Language-Grounded World Models: A Structural Limitation and Sufficient Fix

A First-Principles Theory of Slow Thinking and Active Perception

Beyond Backpropagation: Monte Carlo Method Can Train Deep Neural Networks

THEORY AND GENERALIZATION

Certified Interventional Fidelity: Anytime-Valid, Adaptive Evaluation of Causal Claims in Mechanistic Interpretability

Contravariance Theory: Strong Alignment for Minimal Solutions to Hard Tasks

Vanilla SGD with Momentum Survives Heavy-Tailed Noise: Convergence Analysis without Gradient Clipping or Normalization

Learning AC^0 under Locally Sampleable Graphical Models

Provably Optimal Learning Algorithms for Assistance Games

PROBABILISTIC AND STATISTICAL METHODS

Bayesian Experimental Design via Score Matching

Expressivity and Statistical Trade-offs in Diffusion Policy Learning

NFTR: From Provable Mode-Averaging to Geodesic Subgoal Selection in Offline Goal-Conditioned RL

Structure Learning on Clustered Data

Selecting Interpretable Circular Coordinates from Data

High-Dimensional Procrustes Matching via Tree Counts

Statistical Efficiency and Inference of Quantile Distributional Reinforcement Learning

Prediction-Powered Active Testing

Joint estimation of high-dimensional spiked covariance matrices via a partially shared subspace

QUANTUM ML AND HYBRID PARADIGMS

A Quantum Reservoir Architecture for Chaotic Forecasting and a Test of Whether Its High Dimension Helps

Back tomorrow with more!

P.S. In the most recent submission window: cs.LG: 136 papers stat.ML: 15 papers