Hello folks,

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

Theory and Generalization

- A Defense of the Quadratic Model This paper stress-tests the simplest optimization model, the quadratic model, and demonstrates that it can accurately predict optimization dynamics in LLMs with 150M parameters over windows lasting up to 10% of training. The authors use Lanczos quadrature to estimate the Hessian spectrum deep into the tail, revealing surprising structure in eigenvalues and eigenvectors that depends on batch size and preconditioner. They also empirically verify local linear stability, showing that LLM optimization typically occurs at a stochastic edge of stability. This work provides a theoretically tractable proxy for understanding pretraining dynamics, which is crucial for formal analysis of training dynamics and optimization landscapes.

- From Score Approximation to Distribution Approximation in Score-Based Diffusion Models This work establishes a rigorous quantitative connection between score function approximation and probability distribution approximation in score-based diffusion models. The author proves that if a neural network approximates the true score function sufficiently well, the generated distribution is close to the target in KL divergence, with an explicit upper bound on the error derived using Girsanov's theorem and the data processing inequality. This bridges a gap in approximation-theoretic foundations by linking classical neural network approximation theory directly to the quality of generative models, moving beyond empirical success to formal guarantees.

- On the Identifiability of Controlled World Models This paper tackles the fundamental question of when controlled latent prediction in Joint-Embedding Predictive Architectures (JEPAs) identifies both the underlying state and controlled dynamics. The authors establish a joint identifiability theory for controlled world models with Gaussian latent states, identifying two policy-dependent conditions for representation and transition identifiability. They prove that under these conditions, every global minimizer of the JEPA objective identifies the latent state and controlled transition up to an orthogonal transformation. This is a significant contribution to causal representation learning, providing formal bounds and insights into counterfactual prediction under limited action coverage.

Emergent Behavior and Agentic AI

- CausalForge: A Formally Grounded, Self-Improving Agentic Framework for Automated Research in Causal Inference This paper presents CausalForge, a framework for automated theoretical research in causal inference that is grounded in the Lean proof assistant. It combines Causalean, a foundational Lean library with over 7,000 machine-checked declarations, with CausalSmith, an agentic pipeline that selects topics, proposes results, and constructs proofs. Crucially, the framework includes a statement audit to ensure formal theorems faithfully capture informal scientific claims, addressing a key limitation of LLM-based automated research. This represents a strong step toward neurosymbolic approaches with formal grounding for agentic AI.

Until the next announcement!

P.S. In the most recent submission window: cs.LG: 106 papers stat.ML: 24 papers