Hi all,

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

Novel Architectures and Objectives

Geometry-Constrained Kolmogorov-Arnold Networks: Learning Edge Geometry via Banach Duality introduces a genuinely new inductive bias for KANs: instead of fixing the edge function basis (splines, Chebyshev, Fourier), the authors derive edge activations from **Banach duality maps** with a learned scalar exponent p per edge. This exponent controls the qualitative geometric response of each edge -- sub-Euclidean values yield LASSO-like thresholding, $p = 2$ recovers linear behavior, and larger values flatten responses near the origin. The key conceptual departure is that the function-space geometry itself becomes a learnable parameter rather than a fixed prior, and the learned exponents provide an interpretable signal revealing target-dependent geometric structure. Empirically, the method is notably more robust to measurement noise than fixed-basis KANs (3.7x degradation vs. 21.6x for unregularized splines), and dominates in the small-sample regime. This matters because it suggests that the choice of representation geometry should be data-driven, not pre-specified.

Theory and Generalization

Beyond Optimal Rates in Stochastic Optimization: Trajectory-Adaptive Stopping Rules by Liviu Aolaritei, Lucas Levy, **Francis Bach**, and **Michael I. Jordan** tackles a fundamental mismatch in SGD theory: guarantees are proven at fixed deterministic horizons, but practitioners stop adaptively by inspecting the trajectory. The authors construct fully observable, trajectory-adaptive upper confidence sequences for the last-iterate distance to the optimizer and weighted-average suboptimality in strongly convex settings. These bounds hold simultaneously over time, achieve the optimal 1/t rate up to iterated-logarithmic factors, and adapt to realized stochastic gradients -- meaning SGD can stop as soon as a prescribed accuracy is certified without sacrificing statistical validity. The technical core involves new recursive confidence-sequence techniques and a time-uniform empirical Bernstein inequality for adapted processes with unbounded predictable ranges. This is a clean theoretical contribution that bridges the gap between how we analyze SGD and how we actually use it.

Canalization Before Generalization: Grokking as a Dynamical Probe uses weight-decay pulses during the grokking plateau as a probe to dissect how solution selection develops during training. The key finding is that early in the plateau, perturbations produce unordered shifts in generalization time, but later a stable dose ordering emerges -- stronger WD increases accelerate generalization, weaker ones delay it -- and this ordering appears *before* visible generalization. Simultaneously, test-loss barriers between perturbed and baseline checkpoints collapse toward zero while the ordered timing effects persist. The author calls this combination "canalization of function selection": increasingly constrained solution selection with persistent dose-ordered timing sensitivity. This matters because it provides a mechanistic window into how overparameterized networks transition from fitting training data to selecting generalizing solutions, and the canalization concept offers a new vocabulary for understanding grokking dynamics beyond simple delayed generalization.

Probabilistic and Statistical Methods

Barycentric Weak Inner-Product Gromov-Wasserstein introduces a weak Gromov-Wasserstein framework that addresses a real limitation of standard GW: pointwise comparison is too sensitive in one-to-many settings where multiple target outcomes refine a single source state. The key idea is to compare source relations with relations between *target conditional laws* induced by a coupling, retaining conditional means for inner-product relations. The resulting barycentric weak inner-product GW satisfies a clean variational formula: it searches for an intermediate target geometry that can be refined into the prescribed target law via mean-preserving spread without changing conditional means. The author proves minimizer existence via martingale gluing, and with ridge regularization derives a min-max problem with an explicit contraction bound for the projected outer iteration under a quantitative ridge condition. This is a principled extension of optimal transport to structured one-to-many settings, with applications demonstrated in cell-type transfer through RNA/ATAC alignment.

Until the next announcement!

P.S. In the most recent submission window: cs.LG: 159 papers stat.ML: 22 papers