Empirical Evidence of the Hidden Costs of Overparameterization:

Prediction Instability in Statistical & Machine Learning Models

The emergence of overparameterized models—models where the number of parameters (*p)* far exceeds the training sample size (*n*)—has been accompanied by a near-exclusive focus on model summaries of prediction performance (e.g., log-loss, AUC, accuracy). Such summaries mask individual-level prediction instability, i.e., how much individual predictions vary across independent training instances. We show that such instability is propagated not only by data properties (e.g., *n*, noise, nuisance features) but also by design choices, such as the fitting routine, optimization target, architecture, effective degrees of freedom, and computational settings. While overparameterization provides added flexibility, it incurs significant costs: greater variance and prediction instability. Indeed, we show that this type of instability can persist even when increasing the training sample size (*n*). We present empirical results applied to simulated and real data.

We report three model-agnostic diagnostics: (i) prediction-interval width across training instances [1] (ii) δ - exceedance rate—the proportion of individual predictions with at least one training instance deviating from the individual prediction mean by more than the margin δ and (iii) decision-flip rate—the proportion of individual prediction whose binary decision changes across training instances. These diagnostics show that although overparameterized models can match or exceed underparameterized baselines on aggregate metrics, they exhibit substantially higher variability across training instances at the individual level. In contrast, simpler models (e.g., underparameterized logistic regression) stabilize more rapidly as *n* increases and can approach near-zero instability for a sufficient training sample size. We also find that even with the training and test data held fixed, overparameterized models continue to display individual-level instability across training instances.

Prediction instability is more pervasive than previously recognized, particularly when machine learning algorithms are applied in data-deficient situations. Analysts should not assume that individual-level prediction performance is stable when models are retrained and/or achieve near equivalent loss-optimality. Our study underscores the importance of assessing and minimizing prediction stability before putting a model into production.

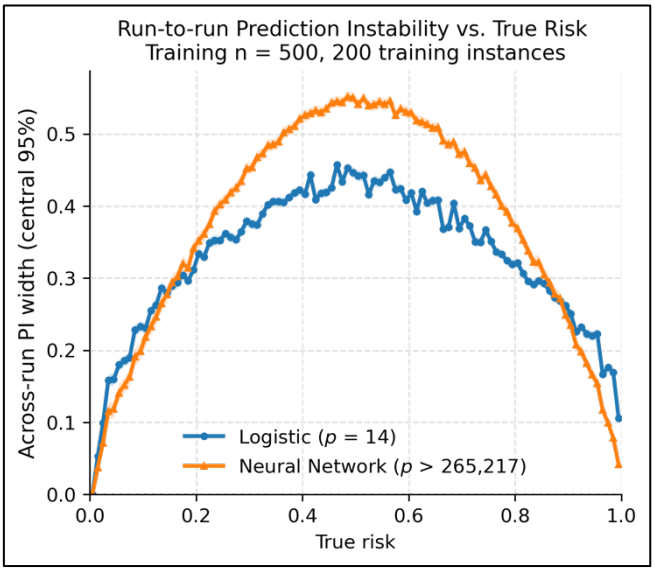

Figure 1. Prediction instability vs. true risk across 200 training instances. Curves show the mean per-individual central 95% across-instance interval width; the overparameterized NN shows greater variability, specifically near the decision threshold.

[1] Riley, R. D., & Collins, G. S. (2023). Stability of clinical prediction models developed using statistical or machine learning methods. Biometrical Journal, 65, 2200302. https://doi.org/10.1002/bimj.202200302