Optimistic Rates: A Unifying Theory for Interpolation Learning and Regularization in Linear Regression

LIJIA ZHOU*, Department of Statistics, University of Chicago, USA
FREDERIC KOEHLER*, Department of Computer Science, Stanford University, USA
DANICA J. SUTHERLAND, University of British Columbia; Alberta Machine Intelligence Institute, Canada
NATHAN SREBRO, Toyota Technological Institute at Chicago, USA

PROBLEM STATEMENT

Generalization theory proposes to explain the ability of machine learning models to generalize to fresh examples by bounding the gap between the test error (error on new examples) and training error (error on the data they were trained upon). Most generalization bounds are too loose to explain the performance of overfit models which nevertheless generalize well. Recently proposed bounds using "uniform convergence of interpolators" can explain this phenomena in the context of linear models, but such bounds are specific to heavily overfit models. Are there generalization bounds which can provide a unified picture for both overfit and heavily regularized models?

METHODS

We study this problem in the setting of linear models with Gaussian covariates and the squared loss. At a technical level, this allows us to analyze the (nonconvex) generalization landscape using tools from Gaussian processes, in particular Gordon's theorem, which has proven very powerful in the analysis of M-estimation in previous works.

RESULTS

We establish a sharp generalization bound which fits cleanly into an existing framework in the statistical learning theory literature: optimistic rates theory. Our bound is nonasymptotic and controls the generalization gap by a function of the training error and Rademacher complexity of a class of functions. Unlike previous optimistic rates bounds, our bound has sharp constants and we show it can explain the ability of heavily overfit predictors to generalize well ("benign overfitting"), in particular recovering results "uniform convergence of interpolators" as a special case. We show that our result can cleanly recover many other results concerning the performance of M-estimators like the LASSO and ridge regression.

SIGNIFICANCE

This result gives insights into what optimal generalization bounds look like and how they reconcile modern phenomena like overfitting and double descent with the mathematical theory of statistical learning. It also establishes connections between tools like Rademacher complexity from learning theory and exact proportional asymptotics studied in high-dimensional statistics and other areas.

Keywords: Generalization theory, High-dimensional Statistics, Interpolation, Regularization, Uniform Convergence

Authors' addresses: Lijia Zhou, Department of Statistics, University of Chicago, Illinois, USA, zlj@uchicago.edu; Frederic Koehler, Department of Computer Science, Stanford University, California, USA, fkoehler@stanford.edu; Danica J. Sutherland, University of British Columbia; Alberta Machine Intelligence Institute, Canada, dsuth@cs.ubc.ca; Nathan Srebro, Toyota Technological Institute at Chicago, Illinois, USA, nati@ttic.edu.

Research Article

*These authors contributed equally.