Sharp *and* Flat Minima Generalize Better

2020-01-10

Introduction — Prior work has varyingly found that *sharp* minima generalize better (after all, l^2 regularization increases curvature) or that *flat* minima generalize better (after all, flat minima are more robust to small displacements in weight space). We reconcile these competing intuitions by showing how the relationship of generalization and curvature depends on the learning task's noise structure. We emphasize the role of a metric in controlling this distinction. In doing so, we offer a modern derivation of the Takeuchi Information Criterion (TIC), a generalization of the Akaike Information Criterion (AIC) that to our knowledge has not been derived in the English language literature. Because the TIC estimates a smooth hypothesis class's generalization gap, it is tempting to use it as an additive regularization term. However, the TIC is singular where the hessian is, and as such gives insensible results for over-parameterized models. We explain these singularities and explain how the implicit regularization of gradient descent both demands and enables a singularity-removing correction to the TIC. The resulting *Stabilized TIC* (STIC) uses the metric implicit in gradient descent to threshold flat from sharp minima. It thus offers a principled method for optimizer-aware model selection easily compatible with automatic differentiation systems. By descending on STIC, we may tune smooth hyperparameters such as l_2 coefficients. We give bounds on generalization performance and validate on classic datasets.

The Metric in Generalization —

The Takeuchi Information Criterion —

Descent's Implicit Regularization Removes TIC's Singularities —

Stabilized TIC for Smooth Model Selection —

Related Work —

Conclusion —