A PAC-Bayesian Perspective on the Interpolating Information Criterion

Liam Hodgkinson, Chris van der Heide, Robert Salomone,
Fred Roosta & Michael W. Mahoney

Interpolators

Parameter estimates minimize empirical risk L_n

$$\theta^\star \in \mathcal{M} := \arg\min_{\theta\in\Theta\subseteq\mathbb{R}^d} L_n(\theta) := \sum_{i=1}^n \ell(f(x_i,\theta),y_i)$$

If $d > n$, $\mathcal{M}$ is a **submanifold**, so infinitely many solutions. Instead, consider an **interpolator** $\theta^\star$ using a **regularizer** R:

$$\theta^\star = \arg\min_{\theta\in\mathcal{M}} R(\theta) \text{ vs. } \theta_0 = \arg\min_{\theta\in\Theta} R(\theta).$$

e.g. $R(\theta) = \|\theta\|^2$ for linear regression.

Interpolating Information Criterion

The **interpolating information criterion** is a measure of ***model quality*** for interpolators derived as a free energy approximation in

Hodgkinson, L., van der Heide, C., Salomone, R., Roosta, F., & Mahoney, M. W. (2023). The Interpolating Information Criterion for Overparameterized Models.

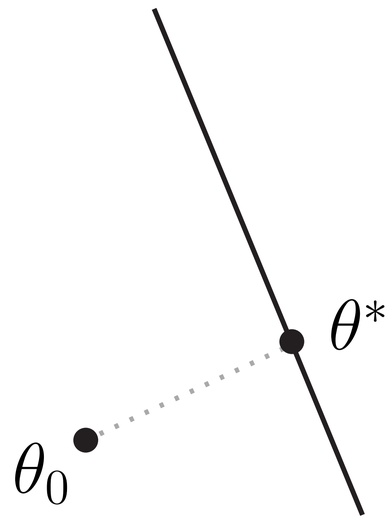
The **IIC** $=$ **log-regularizer** $+$

$$\frac{\textbf{sharpness} + \textbf{curvature}}{n} - \log(n).$$

Key Factors

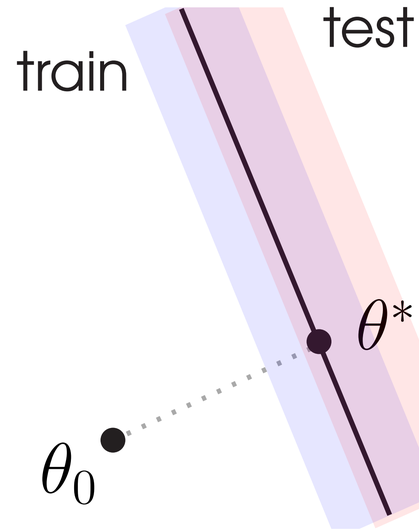
log-regularizer

$$\log[R(\theta^\star) - R(\theta_0)]$$

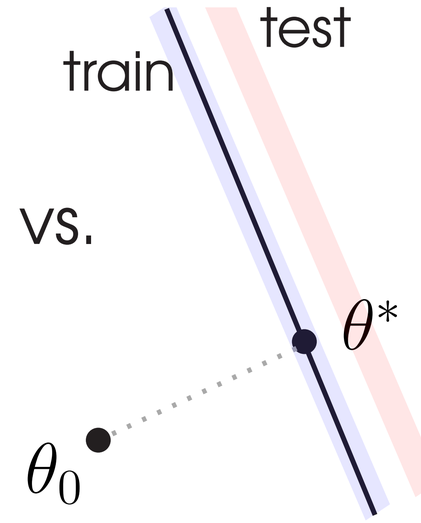


sharpness[a]

$$\log\det(DF(\theta^\star)DF(\theta^\star)^\top)$$

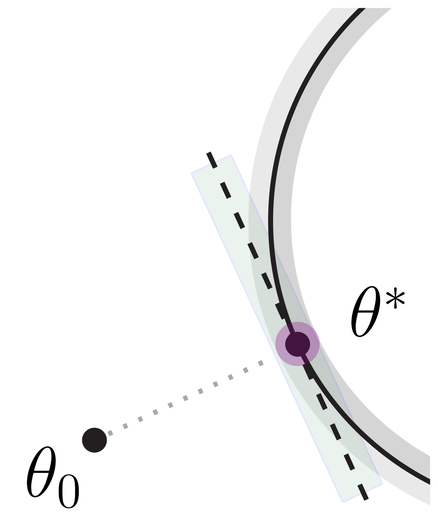


vs.



curvature[b]

$$\log\frac{\det_+\nabla^2_{\mathcal{M}}R(\theta^\star)}{\det\nabla^2R(\theta_0)}$$



dispersion[c]: $2n^{-1}\log\mathbb{E}_{(x,y)\sim\mathcal{D}}e^{L(\theta_0)-L_n(\theta_0)}$ ($\approx$ variance due to data noise)

[a]The matrix $DF(\theta)\in\mathbb{R}^{mn\times d}$ is the Jacobian of the function $\theta\in\mathbb{R}^d\mapsto(f(x_i,\theta))_{i=1}^n\in\mathbb{R}^{mn}$.
[b]Here, $\det_+$ is the pseudodeterminant (product of nonzero eigenvalues) and $\nabla^2_{\mathcal{M}}$ is the manifold Hessian on $\mathcal{M}$.
[c]The test loss $L(\theta)=\mathbb{E}_{(x,y)\sim\mathcal{D}}L_n(\theta)$ where $\mathcal{D}$ is the data distribution.

PAC-Bayes Bound

A general bound on the test error for interpolators in terms of the cold posterior ρ_γ. The IIC is the most significant contributing factor of the bound.

Theorem. *Under mild assumptions, there exists a family of posterior distributions ρ_γ concentrating on the interpolator $\theta^\star$ as $\gamma\to0^+$ such that the expected test loss satisfies*

$$2\mathbb{E}_{\theta\sim\rho_\gamma}L(\theta) \le m\cdot\Big(\textbf{IIC} + 1 + \log(2\pi)\Big) + \textbf{dispersion} + n^{-1}\log(\delta^{-2}) + \mathcal{O}(n^{-2}) + \mathcal{O}(\gamma).$$

Mathematics of Modern Machine Learning (M3L) Workshop
NEURAL INFORMATION PROCESSING SYSTEMS