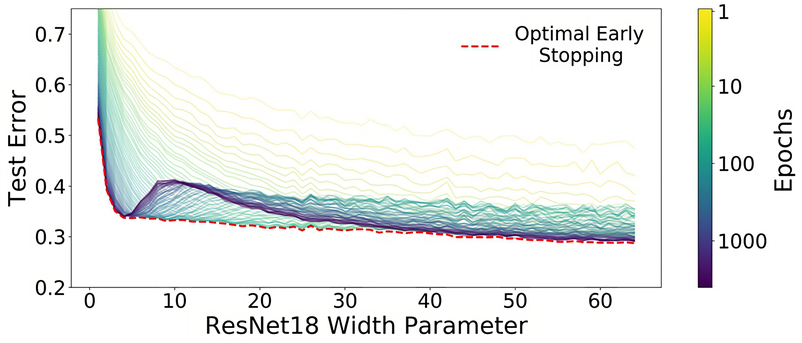
Strong inductive biases provably prevent harmless interpolation

Michael Aerni*, Marco Milanta*, Konstantin Donhauser, Fanny Yang

Harmless interpolation

Surprisingly, **some** highly **overparameterized** models

- generalize well, despite fitting the entire training data, including noise [1]
- do **not** require early stopping for optimal performance

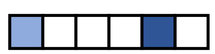
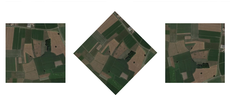


[1] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, "Deep double descent: Where bigger models and more data hurt," *JSTAT*, 2021.

Which models exhibit **harmless interpolation**, and which require **early stopping**? **Inductive bias** is the key!

Inductive bias

The strength of an inductive bias determines how heavily an estimator favors solutions with a certain structure.

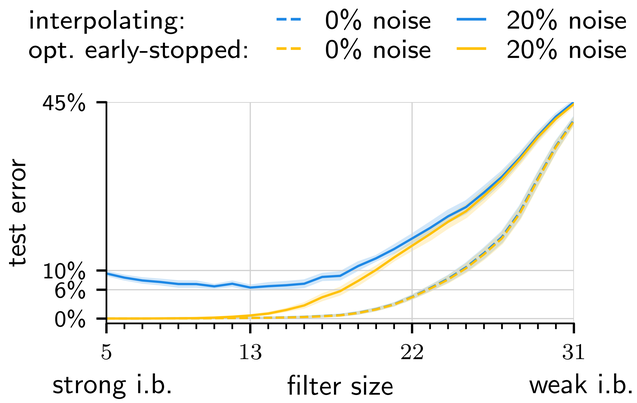
Structure		Inductive bias
sparsity		→ ℓ_1 regularization
locality		→ convolution filters
symmetry		→ rotational invariance

Main takeaways

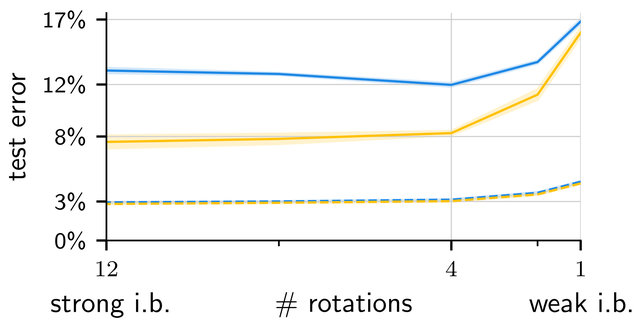
- **weak** inductive bias → **harmless** interpolation: optimal performance at convergence, i.e., after interpolating noisy training data
- **strong** inductive bias → **harmful** interpolation: early stopping is required

Empirical evidence

Wide CNNs on $n = 200$ synthetic images where varying filter size controls inductive bias:



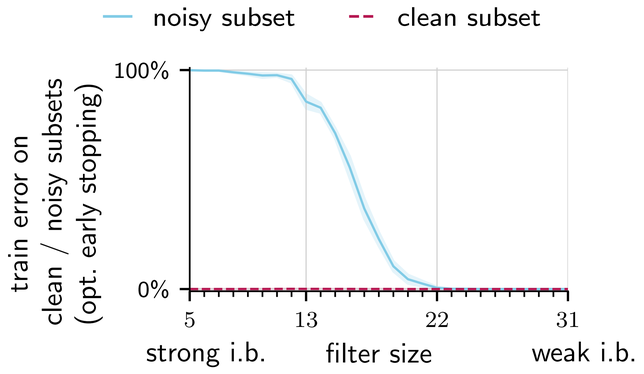
Deep WRNs on $n = 7680$ satelite images where varying #rotations for data augmentation controls inductive bias:



Is this just double descent? No! The phenomenon persists even as width increases.

Interpolation may even be necessary!

Training error of the optimally early-stopped model for noisy and clean subsets of the training data:



- **strong** inductive bias → only fits clean samples
- **weak** inductive bias → interpolates all samples

Theoretical evidence

Proof for high-dimensional Kernel Ridge Regression with a convolutional kernel:

- features $x_1, \dots, x_n \overset{i.i.d.}{\sim} \mathcal{U}(\{-1, 1\}^d)$ with $n \in \Theta(d^\ell)$, $\ell > 0$
- observations $y_i = f^\star(x_i) + \epsilon_i$ were the ground truth $f^\star$ is a "localized" polynomial of constant degree
- convolutional kernel $\mathcal{K}(x, x') = \frac{1}{d}\sum_{k=1}^{d} \kappa(\langle x_{(k,q)}, x'_{(k,q)}\rangle / q)$ with filter size $q \in \Theta(d^\beta)$, $\beta \in (0, 1)$

Main Theorem: tight non-asymptotic matching upper and lower bounds for the prediction error rate $\Theta(n^\alpha)$.