Deep Learning Theory and Applications

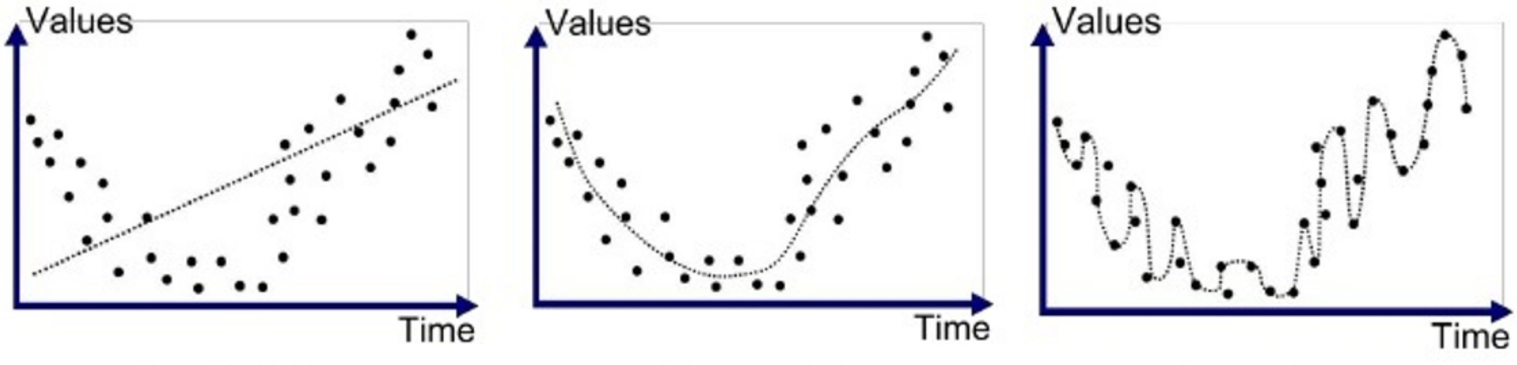
Learning, Memorization and Generalization

CPSC/AMTH 663

Outline

1. Effective capacity
2. Generalization
3. Inductive Bias
4. Norm-based regularization
5. Implicit regularization

Classic notion of over/underfit



Underfitted

Good Fit/Robust

Overfitted

Connects generalization of a fit to model capacity

Oversimple model underfits
Correct complexity of model fits well
Overly complex (over parameterized model) leads to overfitting

Generalization

- Underfit: training error high, test error high
- Good fit: training error low, test error low
- Overfit: training error low, test error high

- Generalization is the ability for a model to perform well on test data, i.e. generalize its results from training to test

- Thus classically people have linked generalization to model capacity
- Ex: regularizations motivated as ability to reduce model capacity and therefore increase generalizations

But do generalization and model capacity go hand in hand?

NO !

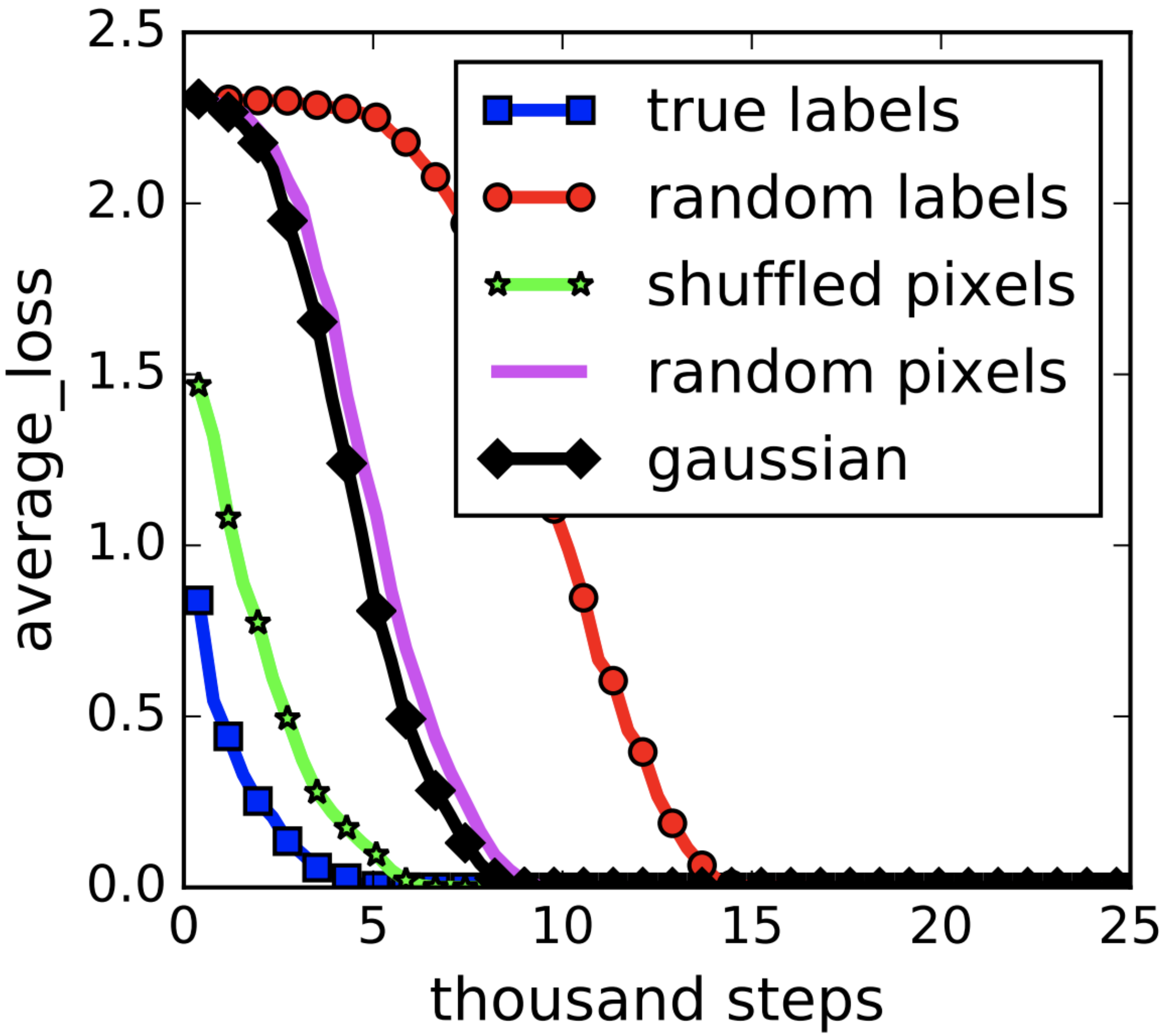
Admission: I’ve lied to you! ☹

Effective capacity

- The effective capacity of neural networks is sufficient to "memorize" the entire dataset
- Deep networks do fine at reducing training error on random samples
    - Reshuffled labels
    - Random images
- They form a model that is "over-parameterized", i.e. the number of parameters approaches the size of the data
- Yet they generalize!

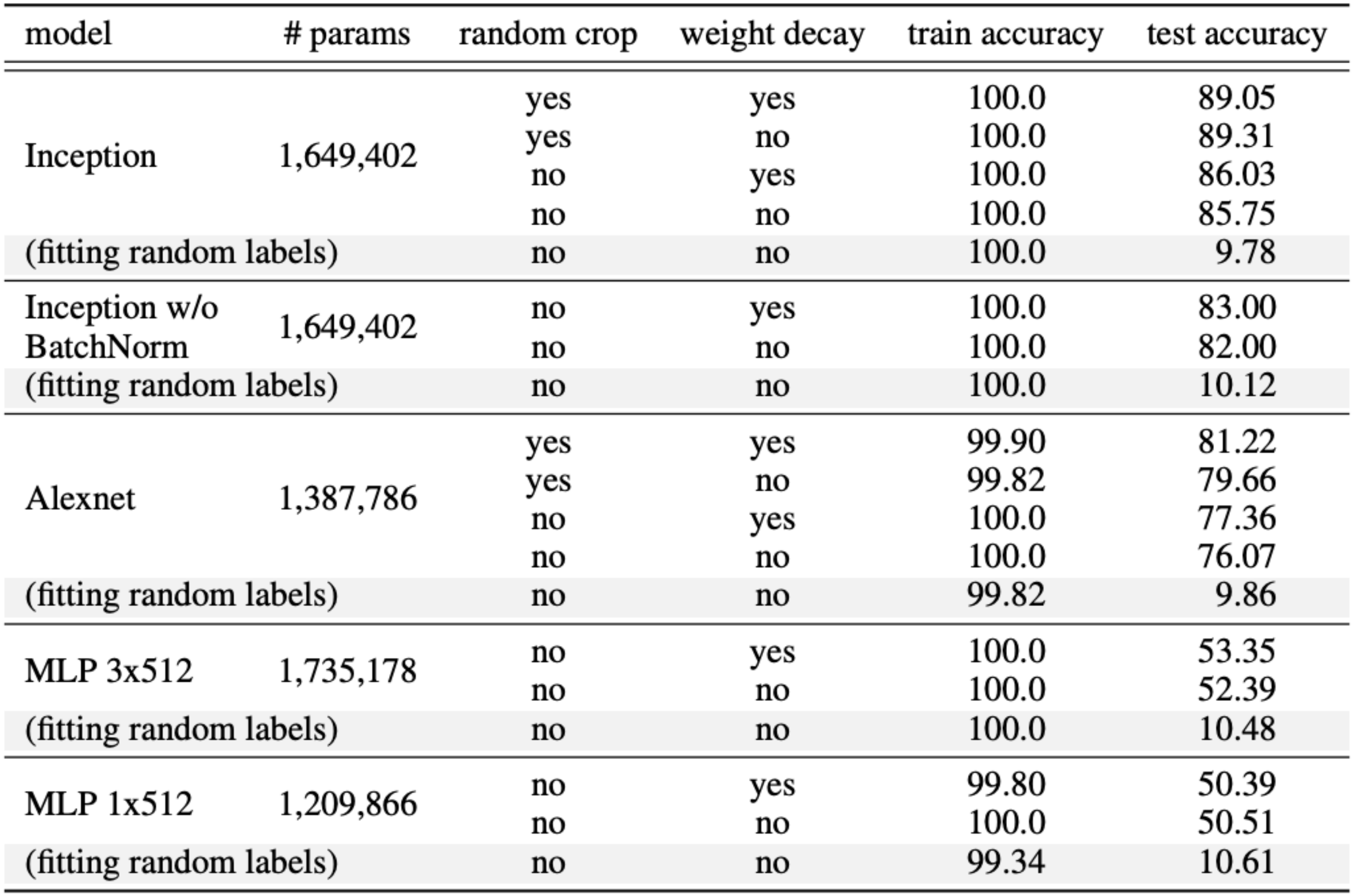
[Zhang et al ICLR 2017]

Memorization = fitting to random data



[Zhang et al. ICLR 2017]

Random noise is fit with or without regularization !



model	# params	random crop	weight decay	train accuracy	test accuracy
Inception	1,649,402	yes	yes	100.0	89.05
		yes	no	100.0	89.31
		no	yes	100.0	86.03
		no	no	100.0	85.75
(fitting random labels)		no	no	100.0	9.78
Inception w/o BatchNorm	1,649,402	no	yes	100.0	83.00
		no	no	100.0	82.00
(fitting random labels)		no	no	100.0	10.12
Alexnet	1,387,786	yes	yes	99.90	81.22
		yes	no	99.82	79.66
		no	yes	100.0	77.36
		no	no	100.0	76.07
(fitting random labels)		no	no	99.82	9.86
MLP 3x512	1,735,178	no	yes	100.0	53.35
		no	no	100.0	52.39
(fitting random labels)		no	no	100.0	10.48
MLP 1x512	1,209,866	no	yes	99.80	50.39
		no	no	100.0	50.51
(fitting random labels)		no	no	99.34	10.61

[Zhang et al 2017]

Regularization is not affecting effective capacity much!

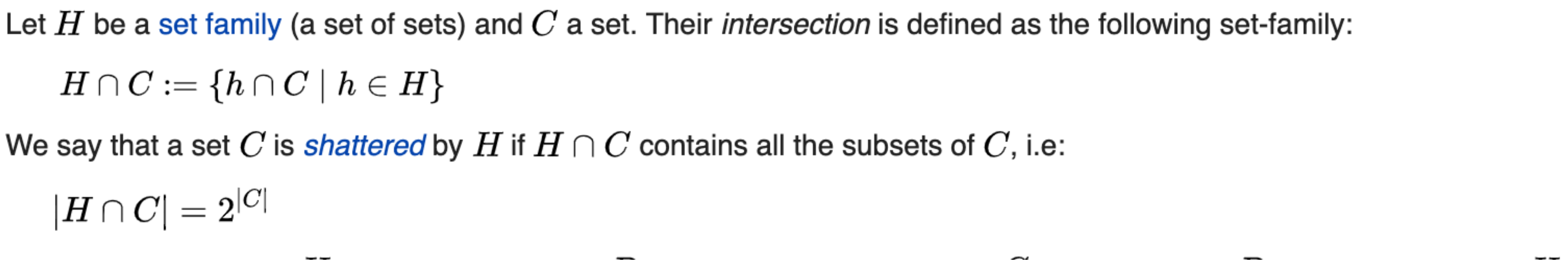
Rademacher complexity

- The supremum measures, for a given set S and Rademacher vector σ, the maximum correlation between $f(z_i)$ and σ_i over all $f \in F$.
- Taking the expectation over σ, we can then say that the empirical Rademacher complexity of F measures the ability of functions from F to fit random noise.

$$\operatorname{Rad}(A) := \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{a\in A}\sum_{i=1}^{m}\sigma_i a_i\right]$$

$\sigma_1, \ldots, \sigma_m$ are independent random variables uniformly chosen from $\{-1, 1\}$

VC Dimension

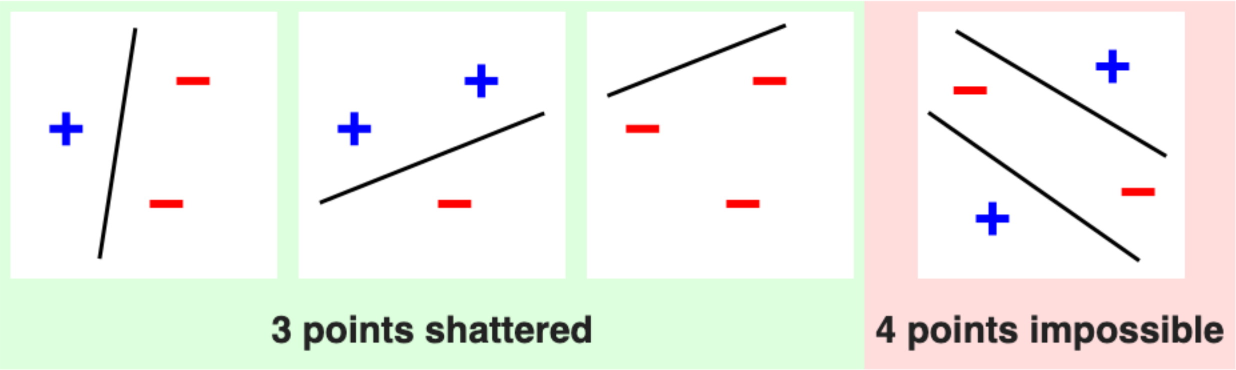


Let H be a set family (a set of sets) and C a set. Their *intersection* is defined as the following set-family:

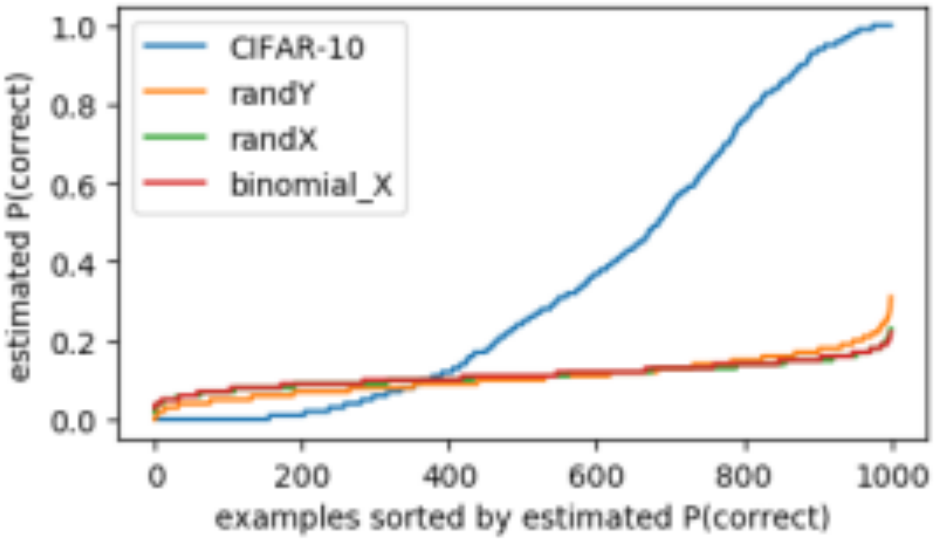
$$H \cap C := \{h \cap C \mid h \in H\}$$

We say that a set C is *shattered* by H if $H \cap C$ contains all the subsets of C, i.e:

$$|H \cap C| = 2^{|C|}$$

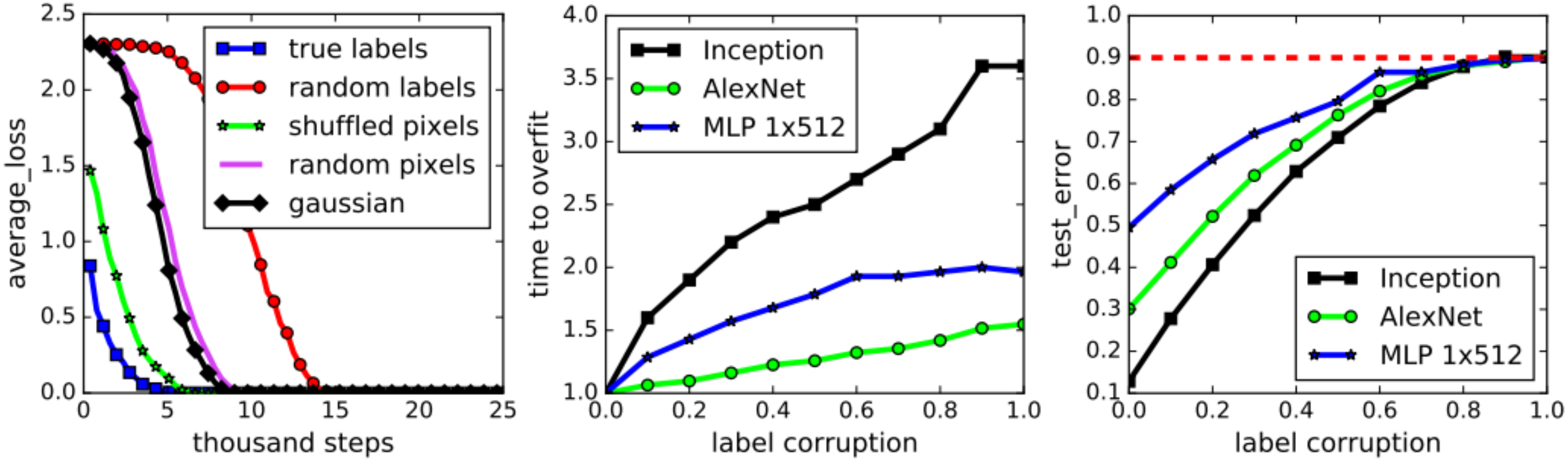


However real data generalizes better



[Arpit et al 2017]

Memorization does not generalize

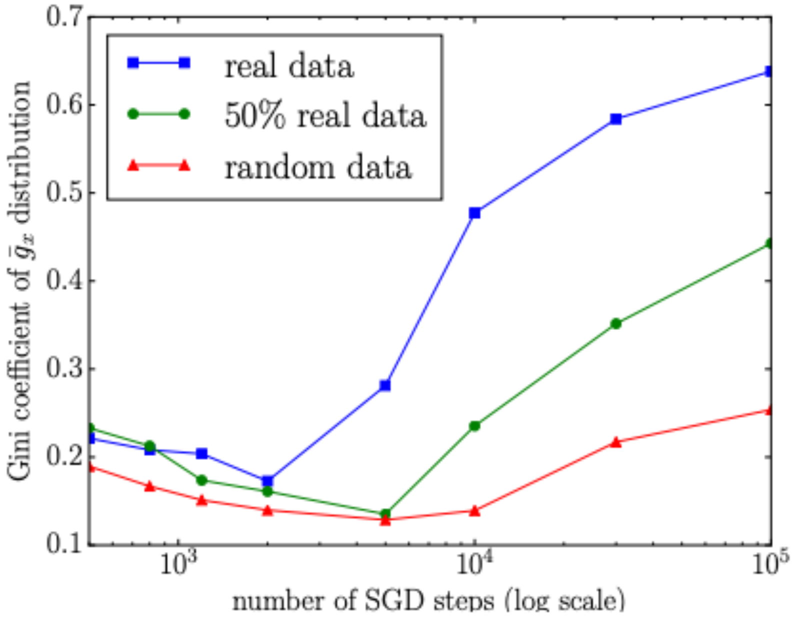


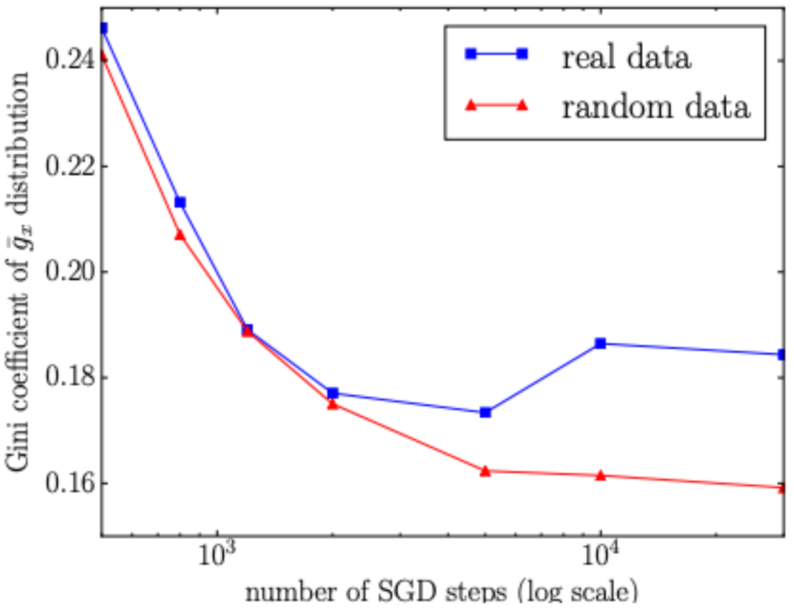
(a) learning curves

(b) convergence slowdown

(c) generalization error growth

What exactly are neural networks doing that’s different?

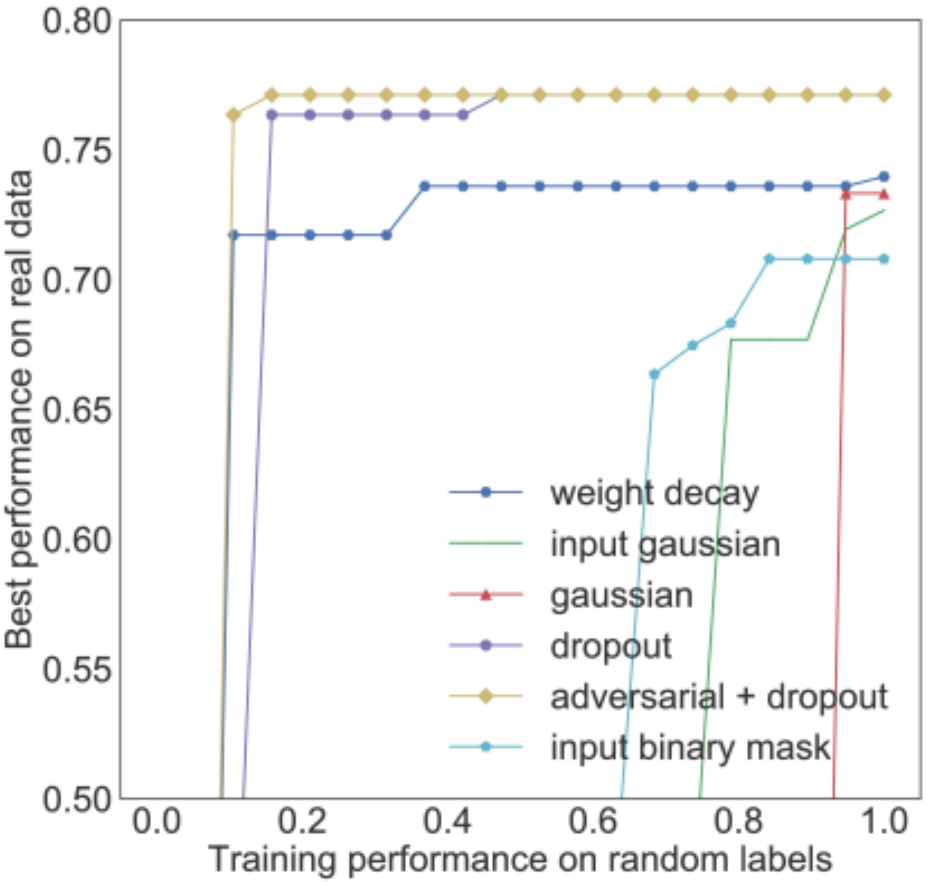




[Arpit et al 2017]

Loss gradient with each input looks different, i.e., the impact that an input has on future loss looks very different

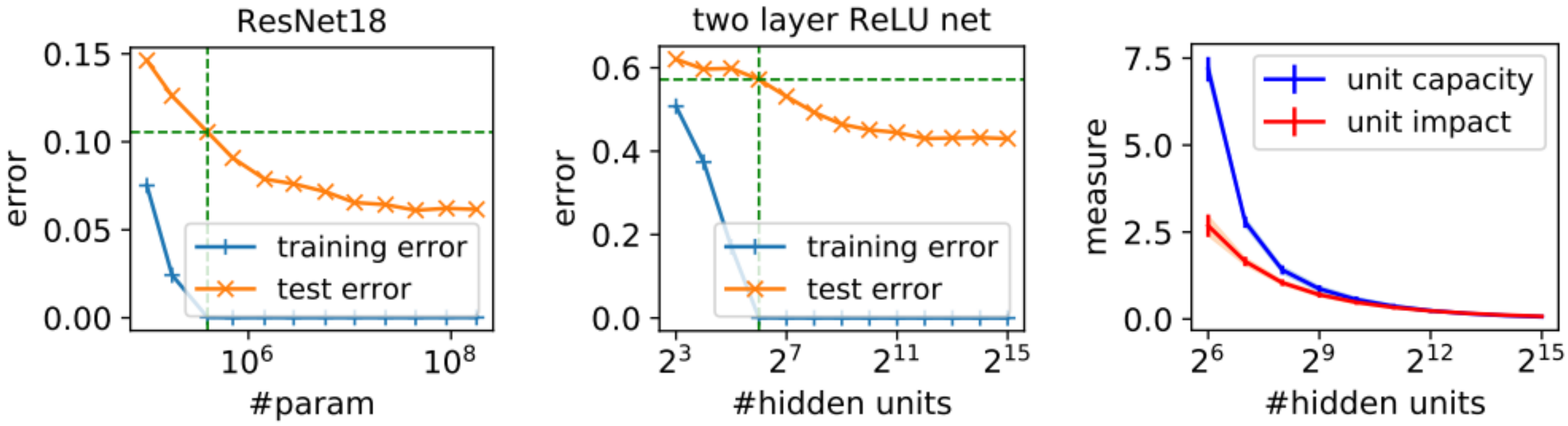
What are regularizations doing?



[Arpit et al. 2017[

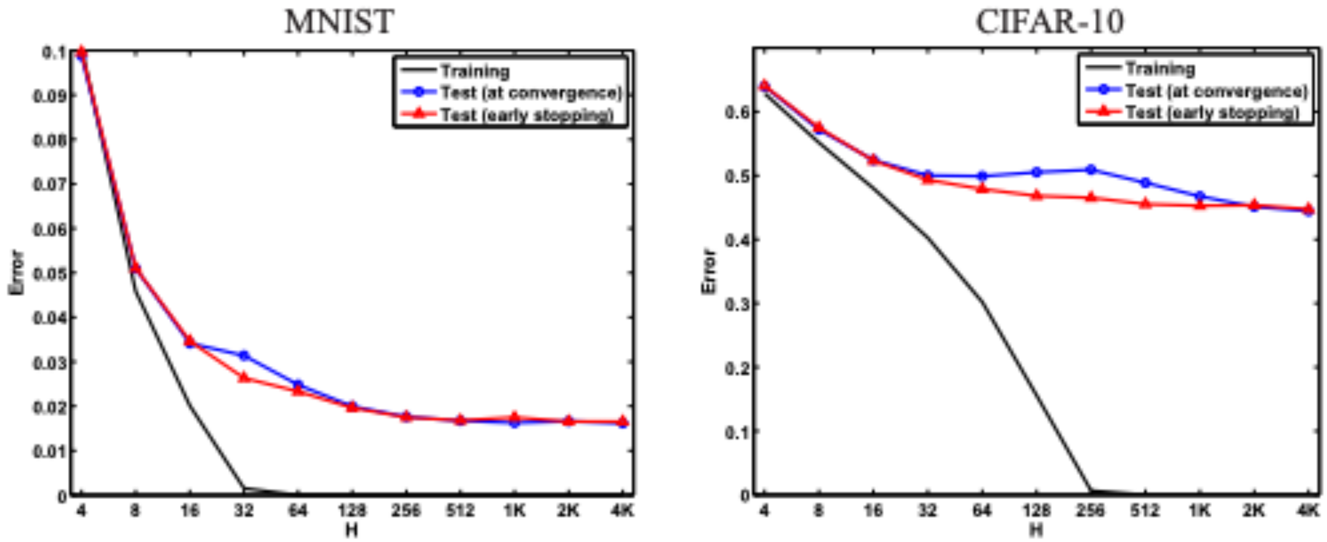
They affect the ability to generalize but not on the ability model capacity

Increasing model size helps Generalization in Nnets!



[Neyshabur. et al. ICLR 2019]

Inductive bias



[Neyshabur et al 2015]

An inductive bias that induces some sort of capacity control (i.e. restricts or encourages predictors to be “simple” in some way), which in turn allows for generalization

The success of learning then depends on how well the inductive bias captures reality

The inductive bias is not the size of the neural network, what could it be?

An analogy to matrix factorization

- Consider a neural network with a single hidden layer consisting of linear activations

$$y[j] = \sum_{h=1}^{H} v_{hj} \left\langle \boldsymbol{u}_h, \boldsymbol{x} \right\rangle .$$

- y=Wx where W = UV^T
- Limiting the number of hidden units corresponds to a low rank factorization (zeros in columns of V)
- However recently there has been a lot of success in low *norm factorizations*
- *Low Frobenius norm:* $$\|\boldsymbol{W}\|_{tr} = \min_{\boldsymbol{W}=\boldsymbol{V}\boldsymbol{U}^\top} \frac{1}{2}(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2).$$

Frobenius norm of a matrix

$$||\mathsf{A}||_F \equiv \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{i\,j}|^2}$$

L2 norm of all matrix entries, corresponds to “volume of matrix”

Low rank vs Low norm

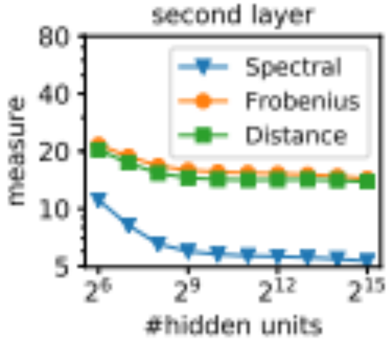
- Unlike the rank, the trace-norm (as well as other factorization norms) is convex, and leads to tractable learning problems
- In fact, even if learning is done by a local search over weights of the network, if the dimensionality is high enough and the norm is regularized, we can ensure convergence to a global minima [Burer, Choi 2006]
- This is in stark contrast to the dimensionality-constrained low-rank situation, where the limiting factor is the number of hidden units, and local minima are abundant [Sreebo, Jakkola, 2003]

Sensible Inductive Bias?

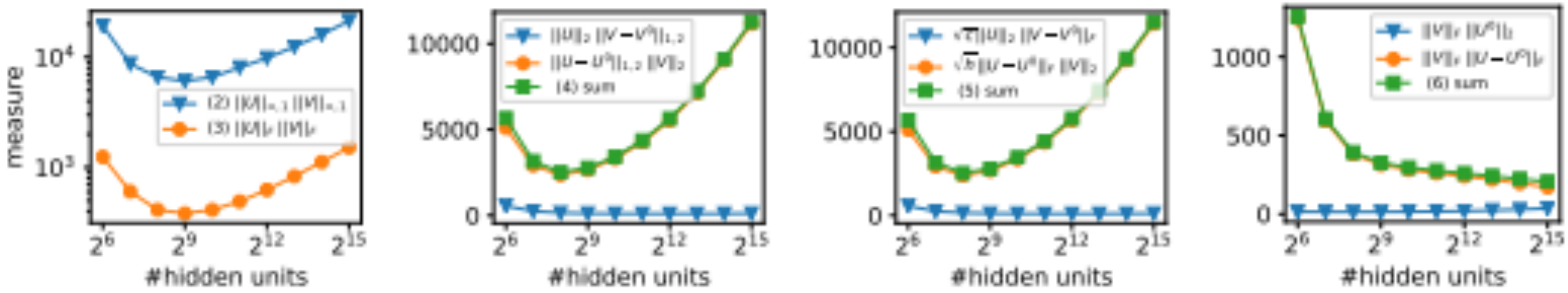
- A. low-trace norm model corresponds to a realistic factor model with many factors of limited overall influence
- What situations is this realistic in?
- The norm of the factorization may be a better inductive bias than the number of weights
- Perhaps learning is succeeding because there is a good representation with small overall norm, and the optimization is implicitly biasing us toward low-norm models

New Notion of Complexity that's "norm-based"

- Authors of Neyshabur. et al. ICLR 2019 define a new notion of complexity that increases with the number of hidden units in a restricted setting (2-layer Rel U Network)
- The idea is that with a low-norm initialization hidden units only step away from the initialization by a limited amount
- Tdistance Frobenius norm, measured w.r.t. initialization ||kU – U0k ||$_F$ decreases with width of hidden layer h



Complexity decreases with increasing hidden units



Gradient Descent itself a regularization?

- [Hardt et. Al. 2016] Any model trained with stochastic gradient method in *a reasonable* amount of time attains small generalization error
- A randomized algorithm **A** is *uniformly stable* if for all data sets differing in only one element, the learned models produce nearly the same predictions
- Stochastic gradient descent is uniformly stable

- Results rest on an important theorem that ***uniform stability implies generalization in expectation [***O. Bousquet and A. Elisseeff. 2002 Stability and generalization]

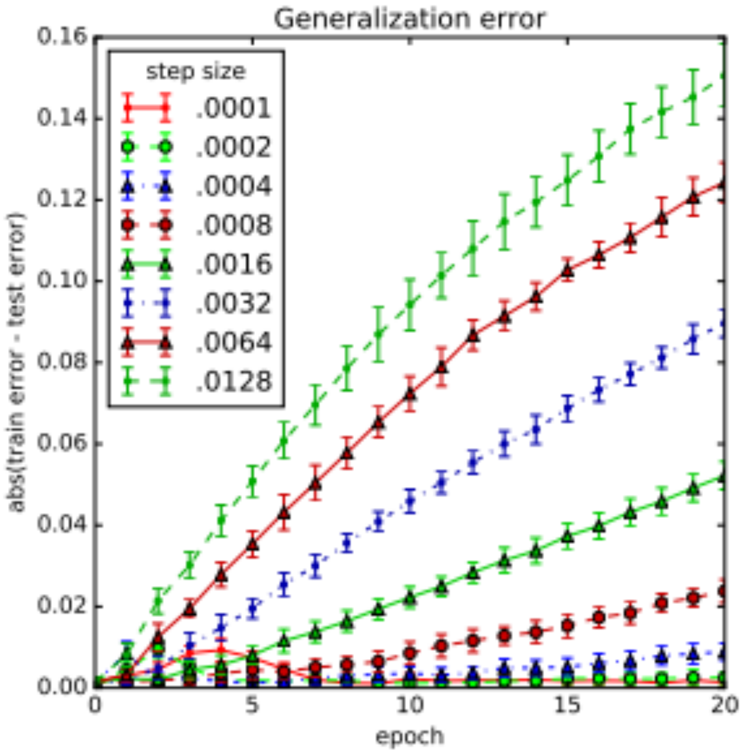
SGD is uniformly stable for a ”reasonable” number of steps

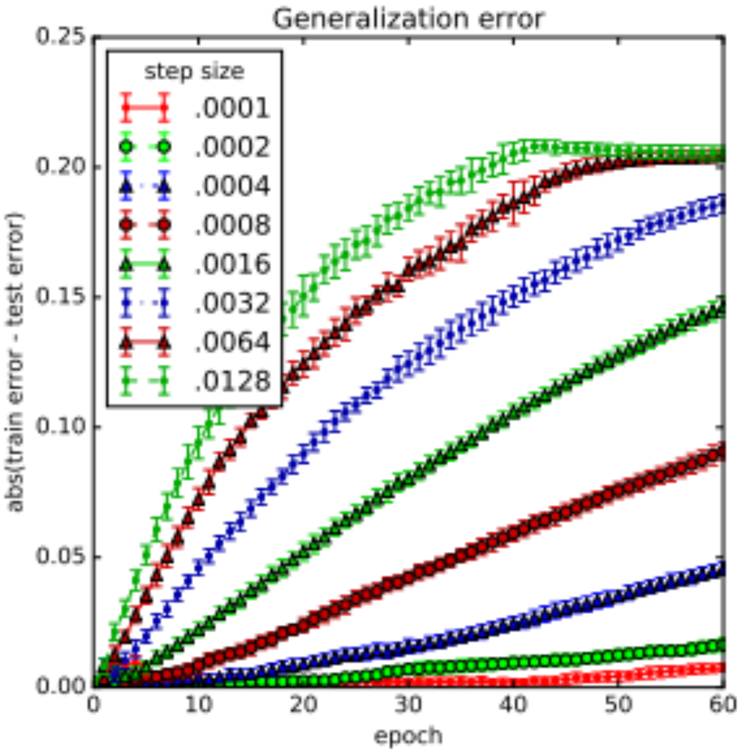
- Proof idea: analyze the output of the algorithm on two data sets that differ in precisely one point. Note that if the loss function is L-Lipschitz for every example z.

$$\mathbb{E}\,|f(w;z) - f(w';z)| \le L\,\mathbb{E}\,\|w - w'\| \text{ for all } w \text{ and } w'$$

- Halving the step size roughly halves the generalization error
- This behavior is fairly consistent for both generalization error defined with respect to classification accuracy and cross entropy. [Hard et al. 2016]

Stepsize vs generalization error





Infinite sized norm-regularized networks?

- Global weight decay, i.e. adding a regularization term that penalizes the sum of squares of all weights in the network
- Weight decay can often improves generalization but it also improves stability! [Krogh et al. NIPS 1992]

Further reading

- Understanding Deep Learning requires rethinking generalization, Zhang et al ICLR 2017 https://arxiv.org/pdf/1611.03530.pdf
- A closer look at memorization in deep networks [Arpit et al 2017] https://arxiv.org/pdf/1706.05394.pdf
- In search of the real inductive bias: on the role of implicit regularization in deep learning [Neyshabur et al, 2015] https://arxiv.org/pdf/1412.6614.pdf
- The role of over-parametrization in generalization of neural networks [Neyshabur et al, 2019] https://openreview.net/pdf?id=BygfghAcYX
- Train faster, generalize better: stability of stochastic gradient descent [Hardt et al, 2016] https://arxiv.org/pdf/1509.01240.pdf
- Stability and generalization [Bousquet and A. Elisseeff 2002] https://www.academia.edu/13743279/Stability_and_generalization
- Rademacher complexity: http://www.cs.cmu.edu/~ninamf/ML11/lect1117.pdf