Dive into Deep Neural Networks: A Viewpoint from Over-parametrization and Sparsity

Content

Background

Generalization Error => bias + variance + noise

Generalization error / Expected risk: ↓

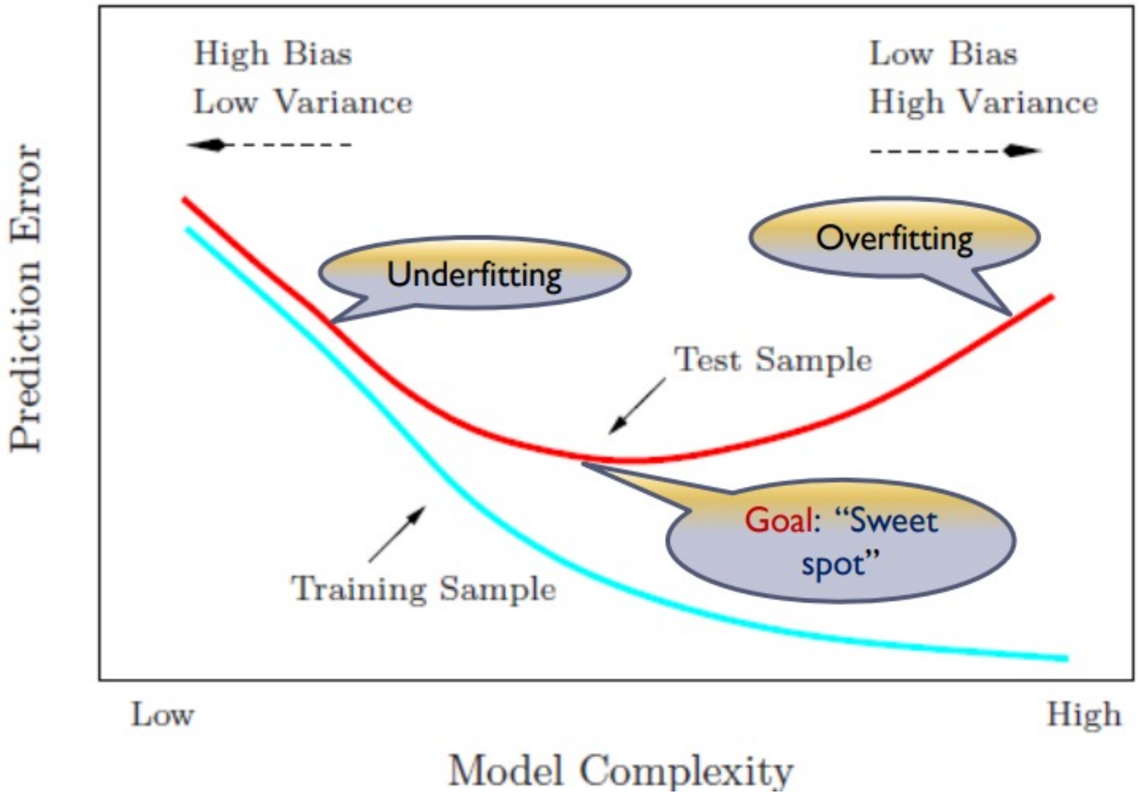
Empirical risk: ↓

$$E\big(L(f^*, y)\big) - \frac{1}{n}\sum L(f^*(x_i), y_i) \le O^*\left(\sqrt{\frac{c}{n}}\right)$$

c: effective model capacity
(VC dimension, Rademacher… - based on model parameters, usually too loose to be useful with neural networks)

n: training samples

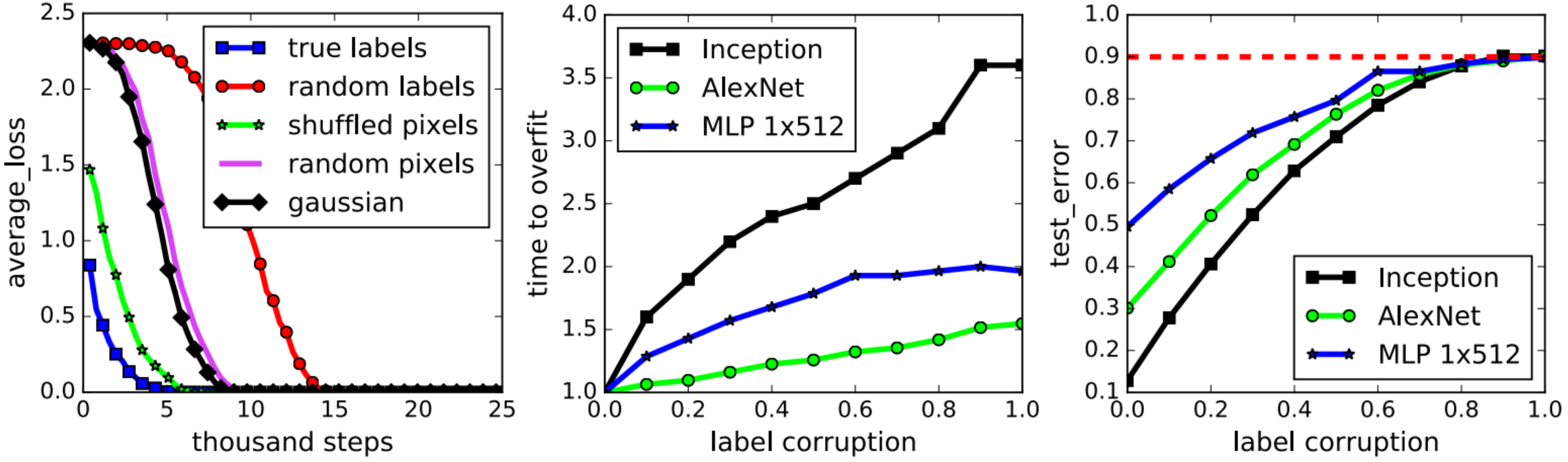
Conventional Wisdom:



U-shaped generalization curve: Bias-Variance Tradeoff

Background

Deep neural networks   number of parameters >> sample size



(a) learning curves   (b) convergence slowdown   (c) generalization error growth

CNNs are able to fit random labels and random pixels on CIFAR10 [1]

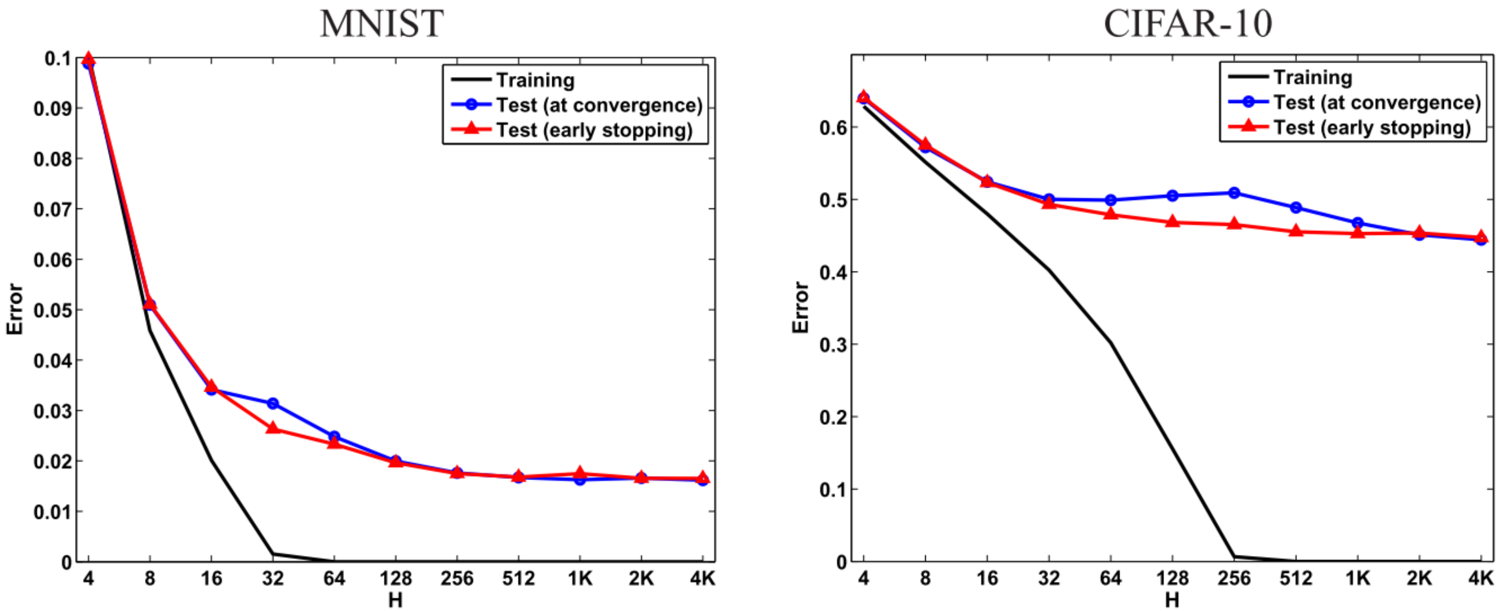
Capacity of deep learning model is excessive!   generalization error – empirical risk $\leq O^*\left(\sqrt{\frac{c}{n}}\right)$

[1] Zhang C, Bengio S, Hardt M, et al. Understanding deep learning requires rethinking generalization[J]. ICLR 2017.

Background

Empirical observation:
Over-parametrization helps generalization…

Why do deep neural networks optimize and generalize well?



The training / test error of 2-layer NNs with different number of hidden units (H) [2]

[2] Neyshabur B, Tomioka R, Srebro N. In search of the real inductive bias: On the role of implicit regularization in deep learning[J]. ICLR 2015.

Background

Consider a $m \times n$ linear system:

$$Ax = b, A \in \mathbb{R}^{m \times n}, x \in \mathbb{R}^n$$

Need $rank(A) \leq n$ to get solutions. (at least as many parameters as equations)

Excessive parameters forms a larger hypothesis space that may contain well-generalized solutions.

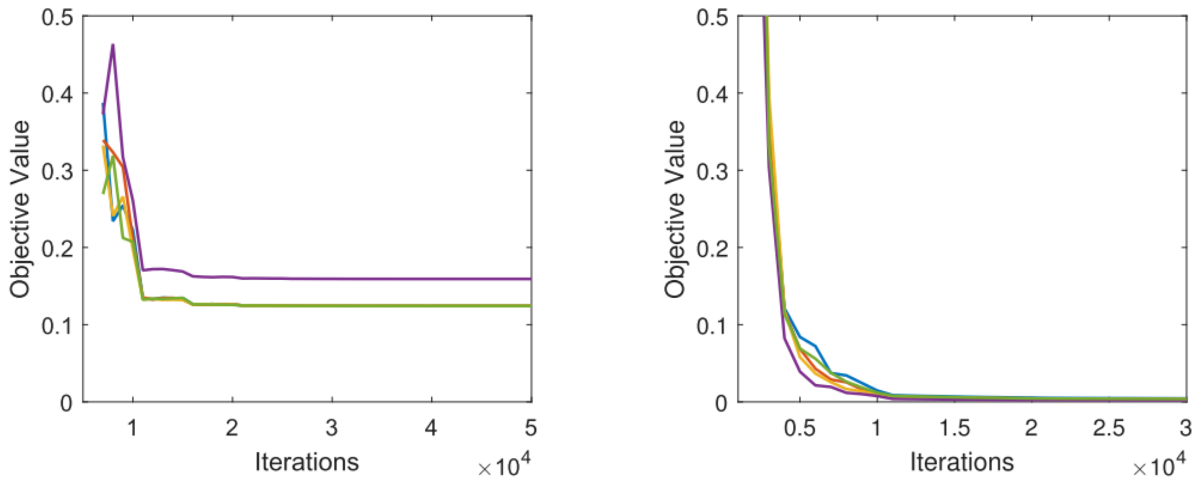
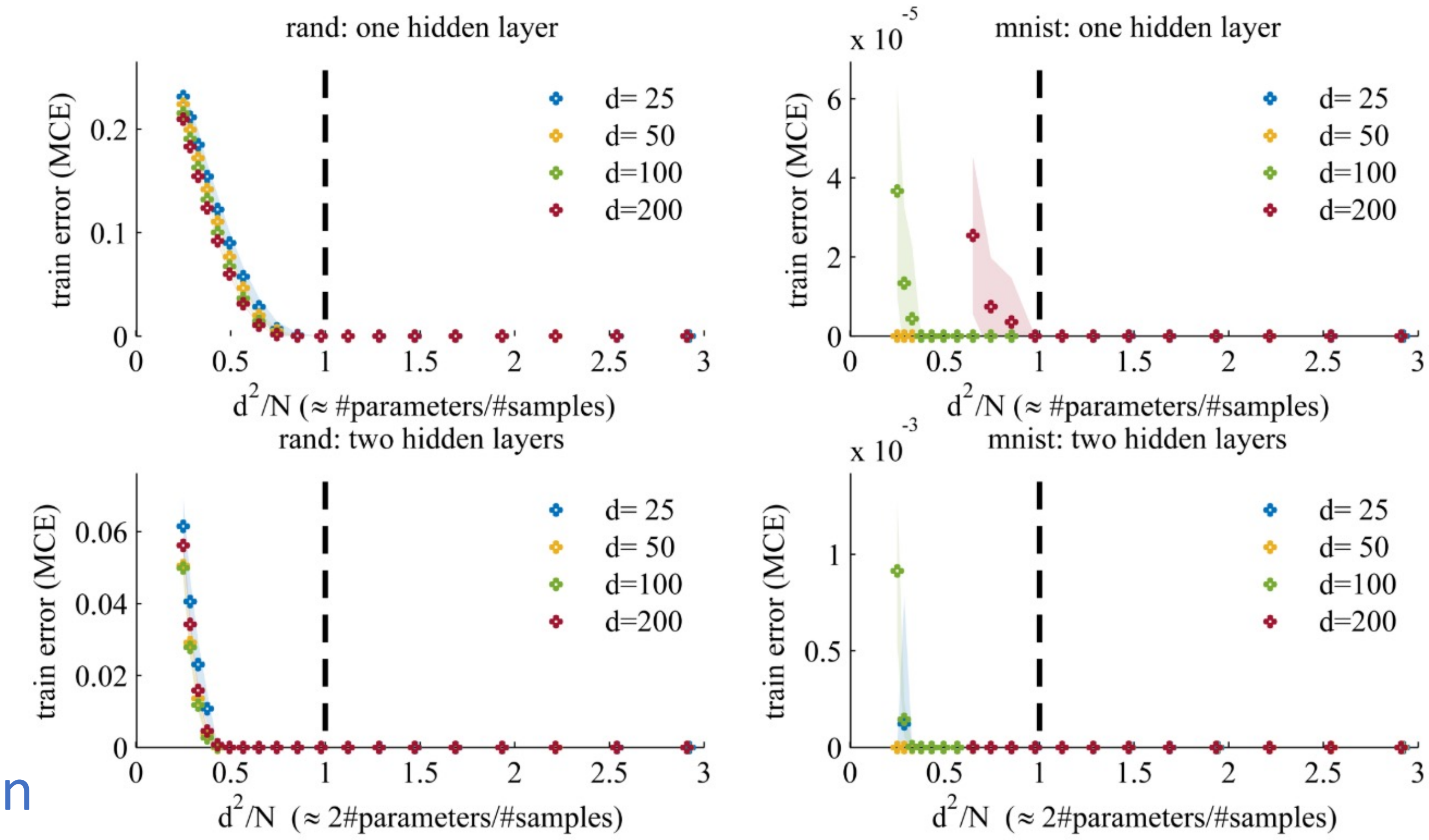


Figure: Experiment first done by Livni-Shalev-Shwartz-Shamir 2014

Background

- Mean Classification Error would be zero at every differentiable local minima[3];
- for deep network: a large class of local minima is globally optimal[4];
- SGD/GD can find global minima in polynomial time for DNNs, CNNs and ResNet[5,6]

In general, over-parametrization networks are easy to optimize



The training error of MNNs with single output, ReLU, MSE loss on two datasets[3]

[3] Soudry D, Carmon Y. No bad local minima: Data independent training error guarantees for multilayer neural networks[J]. arXiv preprint arXiv:1605.08361, 2016.
[4] Nguyen, Q. & Hein, M.. The Loss Surface of Deep and Wide Neural Networks. //ICML,2017:2603-2612
[5] Allen-Zhu Z, Li Y, Song Z. A convergence theory for deep learning via over-parameterization[C]//ICML, 2019: 242-252.
[6] Du S, Lee J, Li H, et al. Gradient descent finds global minima of deep neural networks[C]//ICML, 2019: 1675-1685.

Content

Double Descent

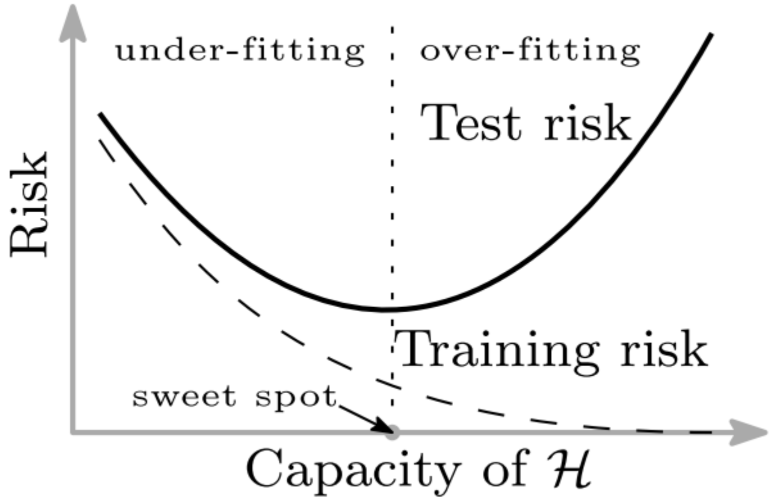
1. Classical and modern regime

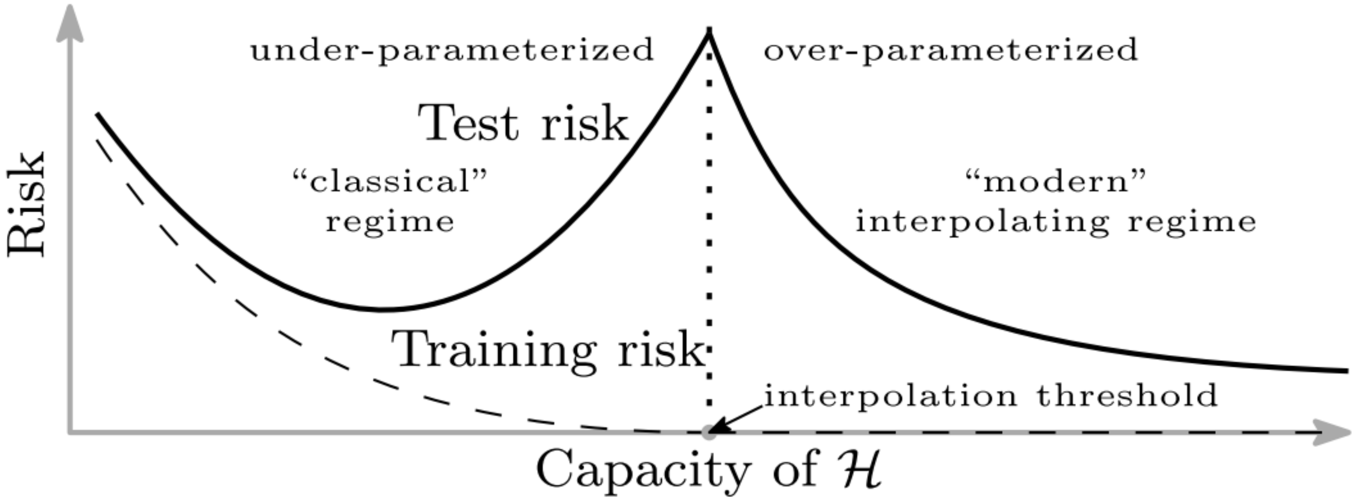
Classical (under-parametrized):

- Many local minima;
- Classical bounds apply;
- SGD (fixed step size) converge slowly.

Modern (interpolation).

- Every local minimum is global, e.g. 0 training error;
- Generalization based on functional smoothness;
- Small batch SGD (fixed step size) converges as fast as GD.

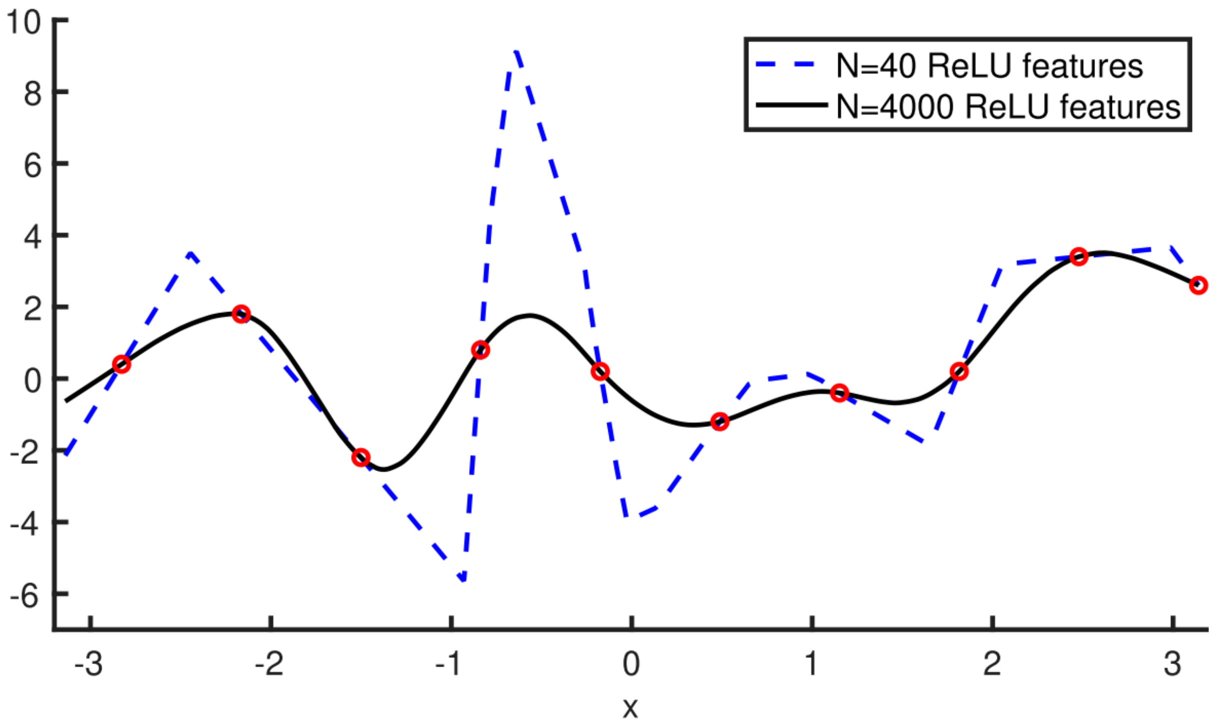




[7] Belkin M, Hsu D, Ma S, et al. Reconciling modern machine-learning practice and the classical bias–variance trade-off[J]. Proceedings of the National Academy of Sciences, 2019, 116(32): 15849-15854.

Double Descent

2. Interpolation



Modern (interpolation).

- Every local minimum is global, e.g. 0 training error;
- Generalization based on functional smoothness;
- Small batch SGD (fixed step size) converges as fast as GD.

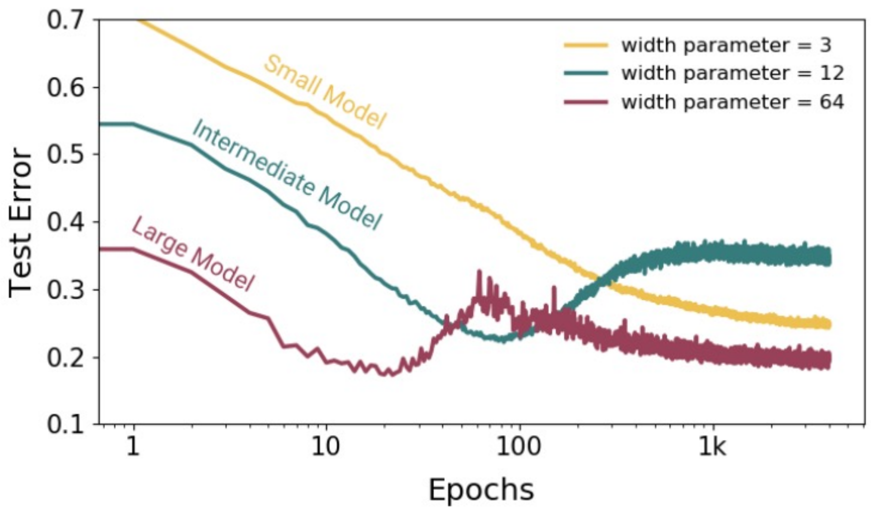
To the right of interpolation of threshold, all function classes are rich enough to achieve zero training risk

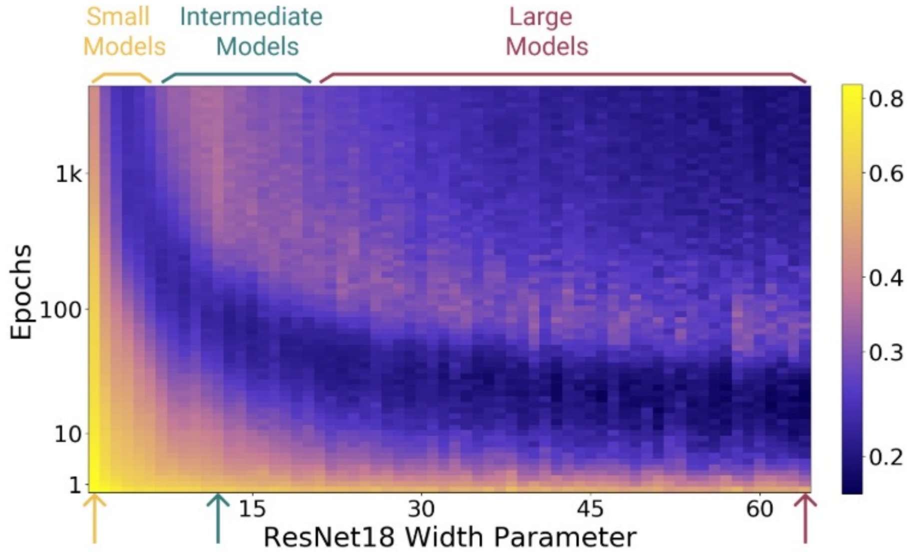
[7] Belkin M, Hsu D, Ma S, et al. Reconciling modern machine-learning practice and the classical bias–variance trade-off[J]. Proceedings of the National Academy of Sciences, 2019, 116(32): 15849-15854.

Double Descent

3. Deep double descent

- Defined effective model complexity (EMC):the maximum number of samples on which it can achieve close to zero training error.
- adding label noise / training samples/ training epochs / data augmentation
  → increase the interpolation threshold (where EMC = training samples)
  → correspondingly shift the peak in test error towards larger models.





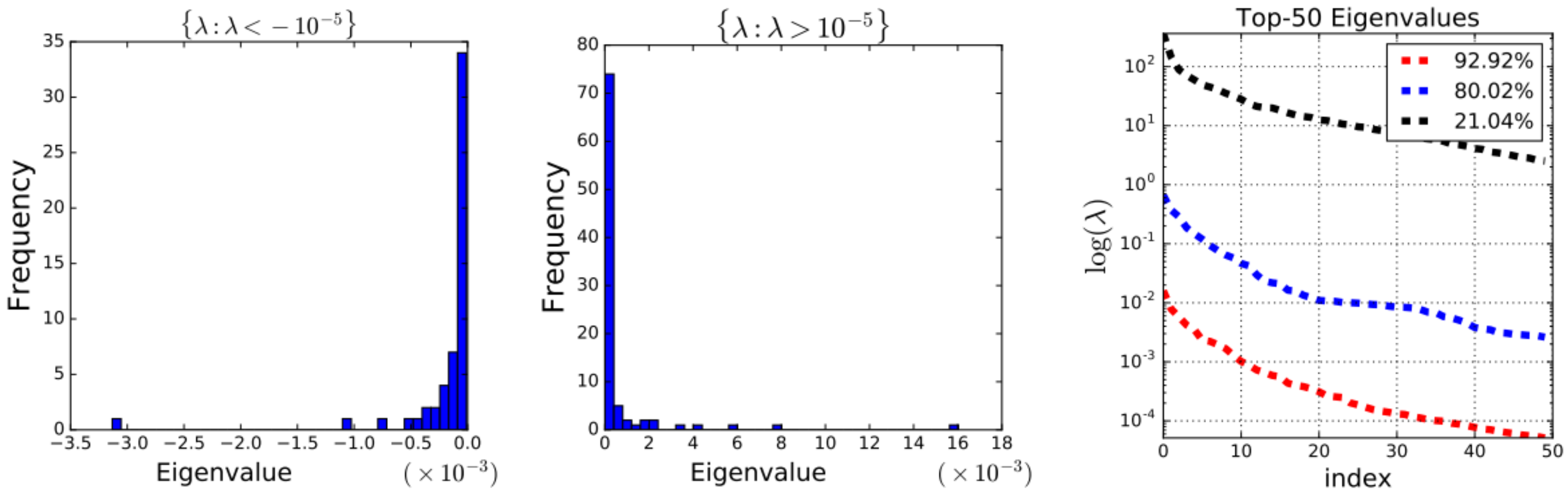
[8] Nakkiran P, Kaplun G, Bansal Y, et al. Deep Double Descent: Where Bigger Models and More Data Hurt[C].ICLR. 2019.

Content

Generalization

1. Inductive Bias -- assumptions on unseen inputs

Occam's razor
deep neural networks guides the optimizers to converge to low-complexity solutions (flat minima)
the volume of basin of good minima dominates over that of poor ones*



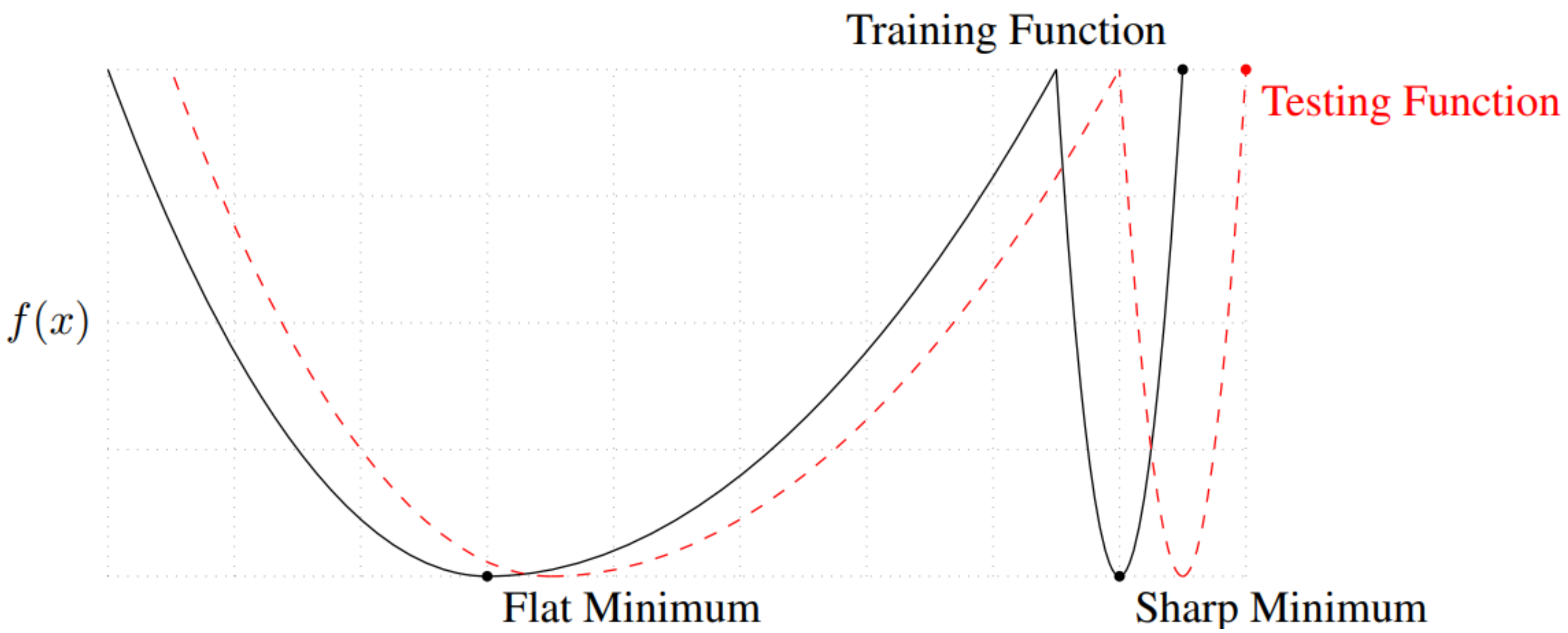
small norm of Hessian matrix this case

Figure 2: **(Left, Middle)** Eigenvalues distribution of a model with 92% test accuracy; **(Right)** Top-k eigenvalues for three solutions, all with training accuracy 100%. The model used here is mLeNet (number of parameter is 3781), and dataset is MNIST. In the experiment, the first 512 training data are selected as our new training set with the rest of training data as attack set. The model is initialized by $\mathcal{N}(0, 2/\text{fan}_{in})$.

[9] Wu L, Zhu Z. Towards understanding generalization of deep learning: Perspective of loss landscapes[J]. arXiv preprint arXiv:1706.10239, 2017.

Generalization

1. Inductive Bias -- assumptions on unseen inputs



[10] Keskar N S, Nocedal J, Tang P T P, et al. On large-batch training for deep learning: Generalization gap and sharp minima[C], ICLR 2017

Generalization

2. Regularization

   Explicit regularization:
   weight decay (l2 regularization)..

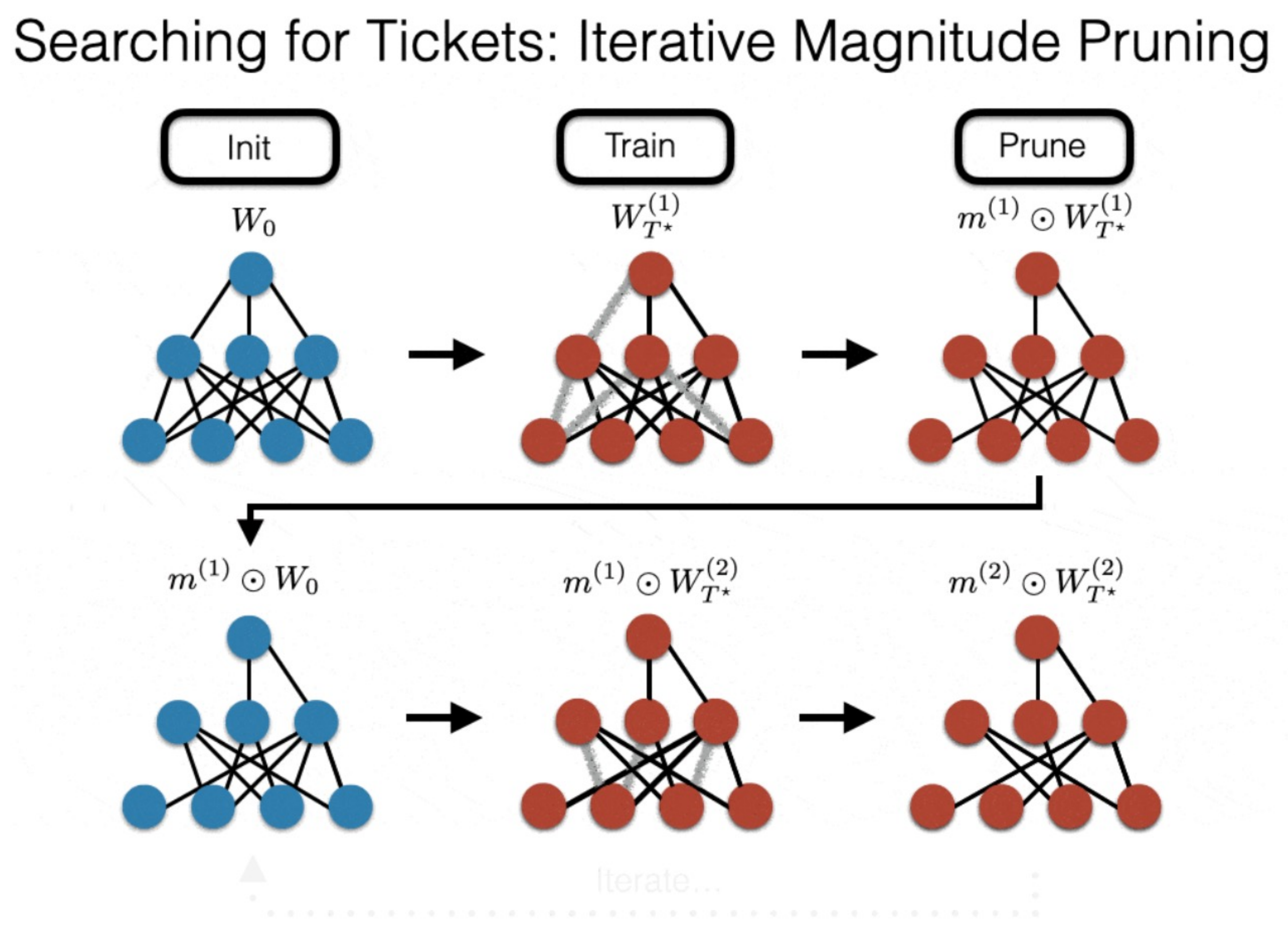
   Implicit regularization:
   SGD, dropout, batch normalization…

   SGD can filter out global minima with large non-uniformity;
   [How SGD Selects the Global Minima in Over-parameterized Learning: A Dynamical Stability Perspective '18]

Content

Sparsity

1. Lottery Tickets Hypothesis
   - Randomly initialization
   - Training to convergence
   - Iterative pruning
   - Late resetting



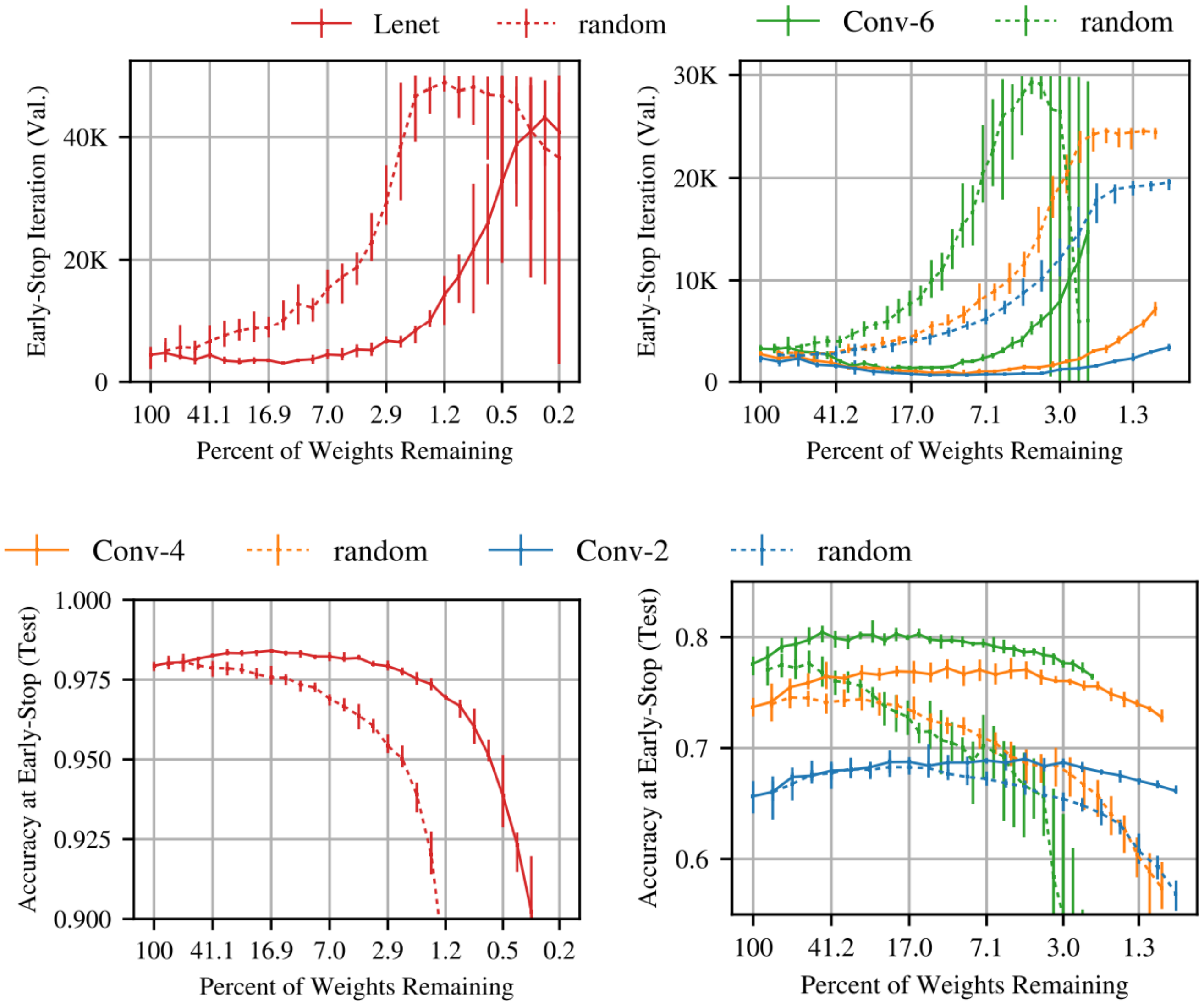
[11] Frankle J, Carbin M. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks[C]//ICLR. 2019.

Sparsity

1. Lottery Tickets Hypothesis

   Dashed lines: randomly sampled sparse networks
   Solid lines: winning tickets

   ➢ Compared with nicely pruned networks, randomly pruned networks seem to optimize and generalize difficultly



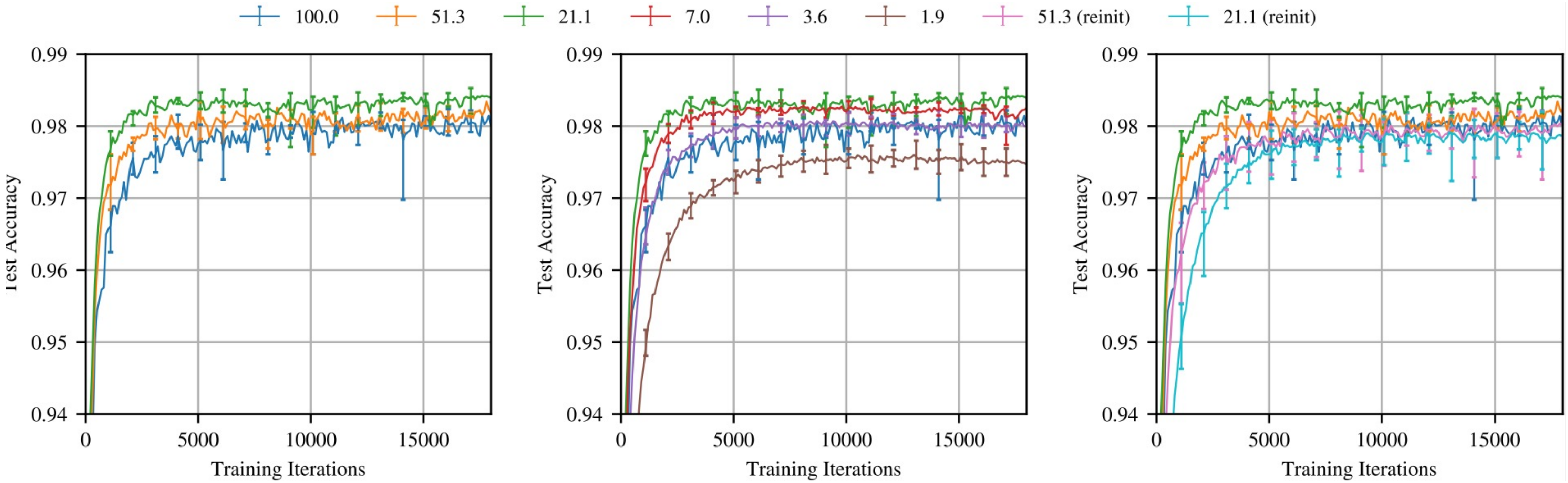
[11] Frankle J, Carbin M. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks[C]//ICLR. 2019.

Sparsity

1. Lottery Tickets Hypothesis

Solid lines: reset the remaining parameters to their values in θ_0, creating the winning ticket
Dashed lines: random initialization

➢ Compared with winning tickets, random initialization makes networks learn slower.

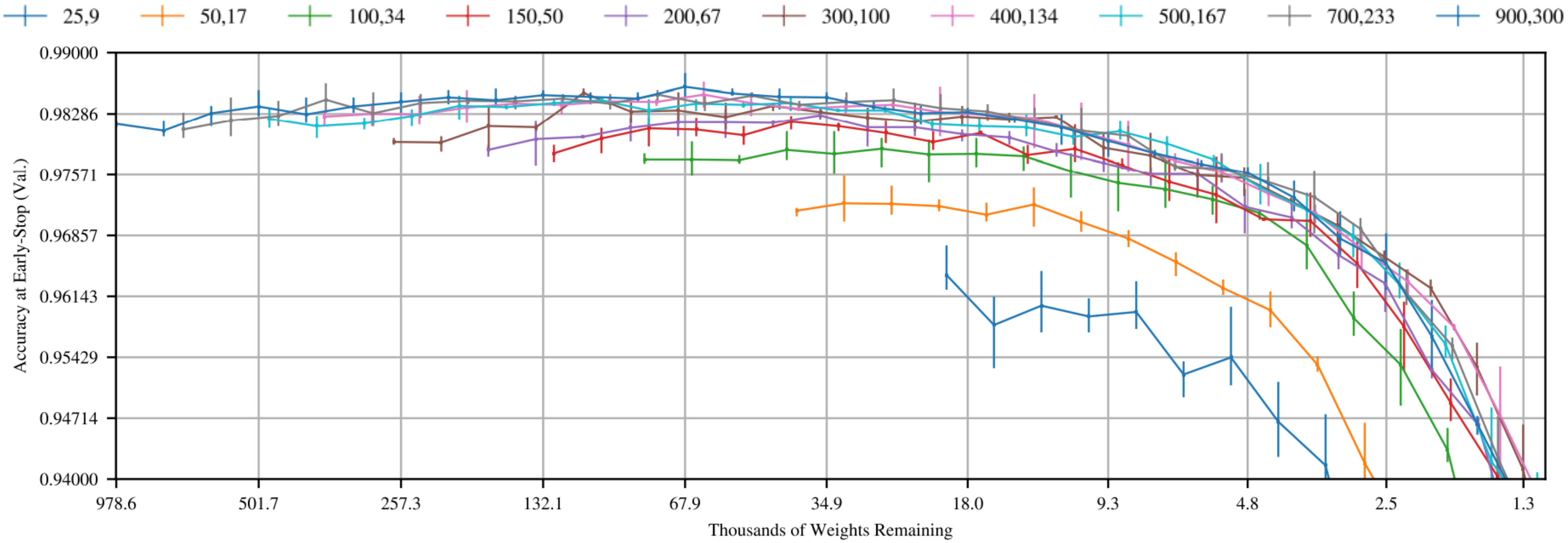


[11] Frankle J, Carbin M. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks[C]//ICLR. 2019.

Sparsity

1. Lottery Tickets Hypothesis

   ➢ winning tickets derived from initially larger networks reach higher accuracy.

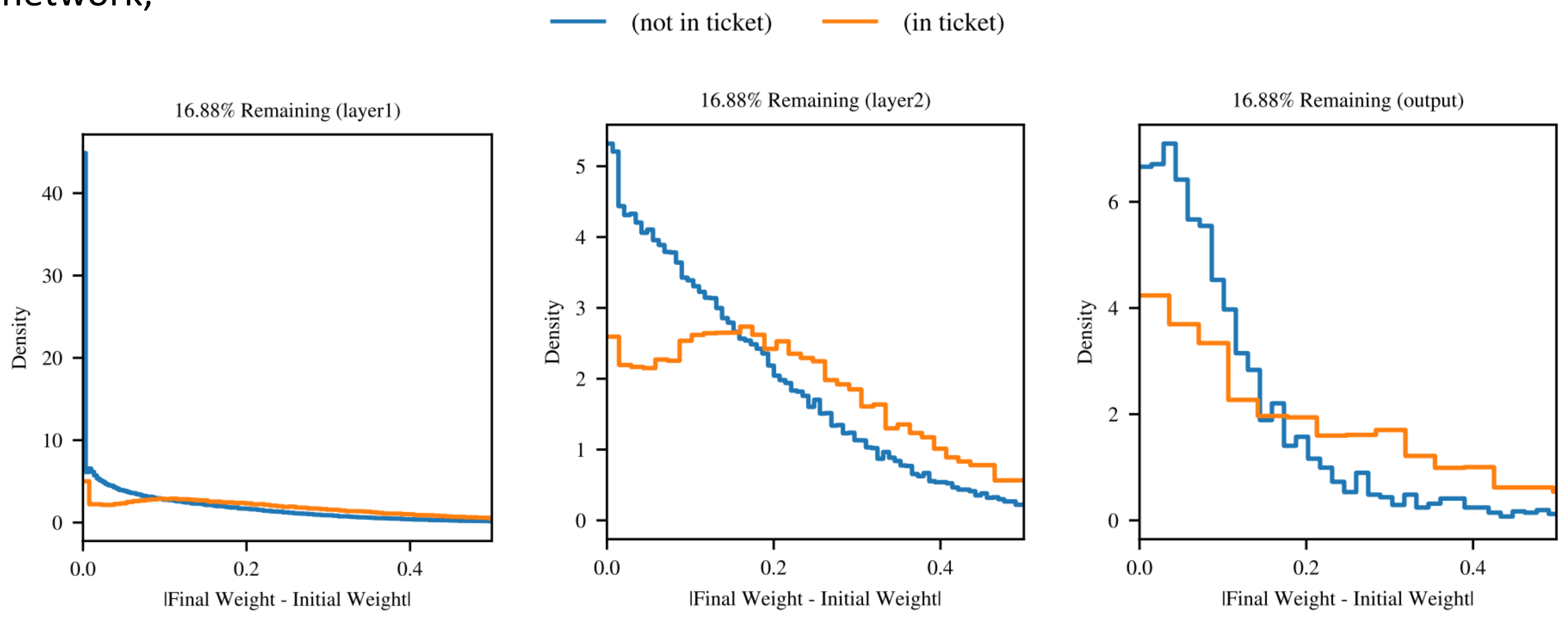


[11] Frankle J, Carbin M. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks[C]//ICLR. 2019.

Sparsity

1. Lottery Tickets Hypothesis

   ➢ winning ticket weights tend to change by a larger amount then weights in the rest of the network,

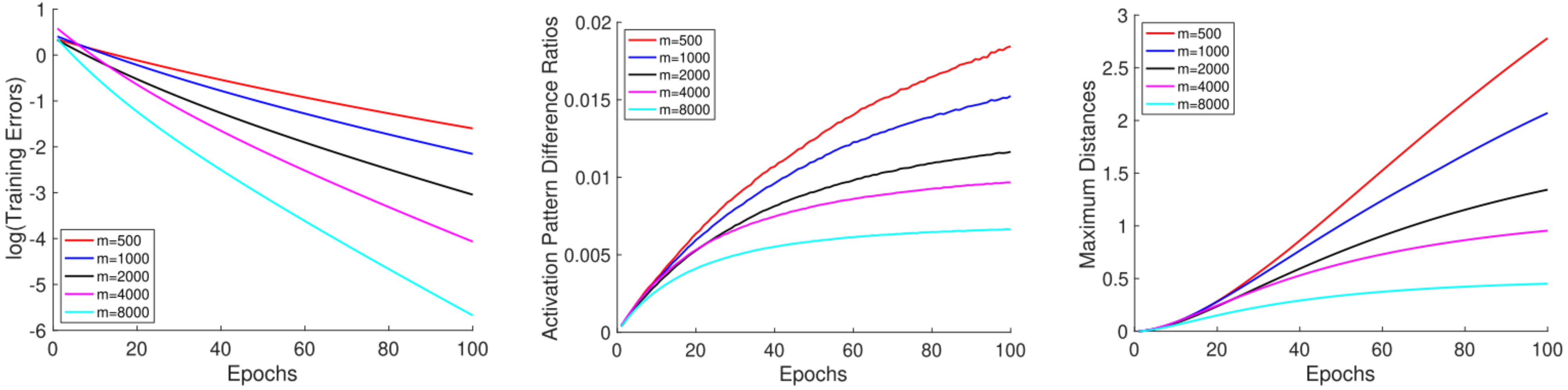


[11] Frankle J, Carbin M. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks[C]//ICLR. 2019.

Sparsity

1. Lottery Tickets Hypothesis

   Consider: over-parameterization, random initialization, and the linear convergence jointly restrict every weight vector wr to be close to its initialization



(a) Convergence rates. (b) Percentiles of pattern changes. (c) Maximum distances from initialization.

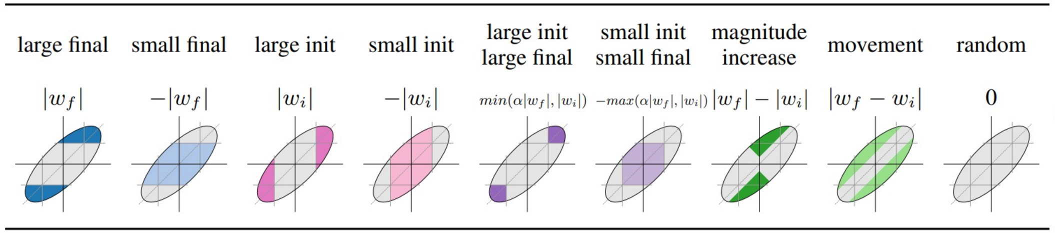
Figure 1: Results on synthetic data.

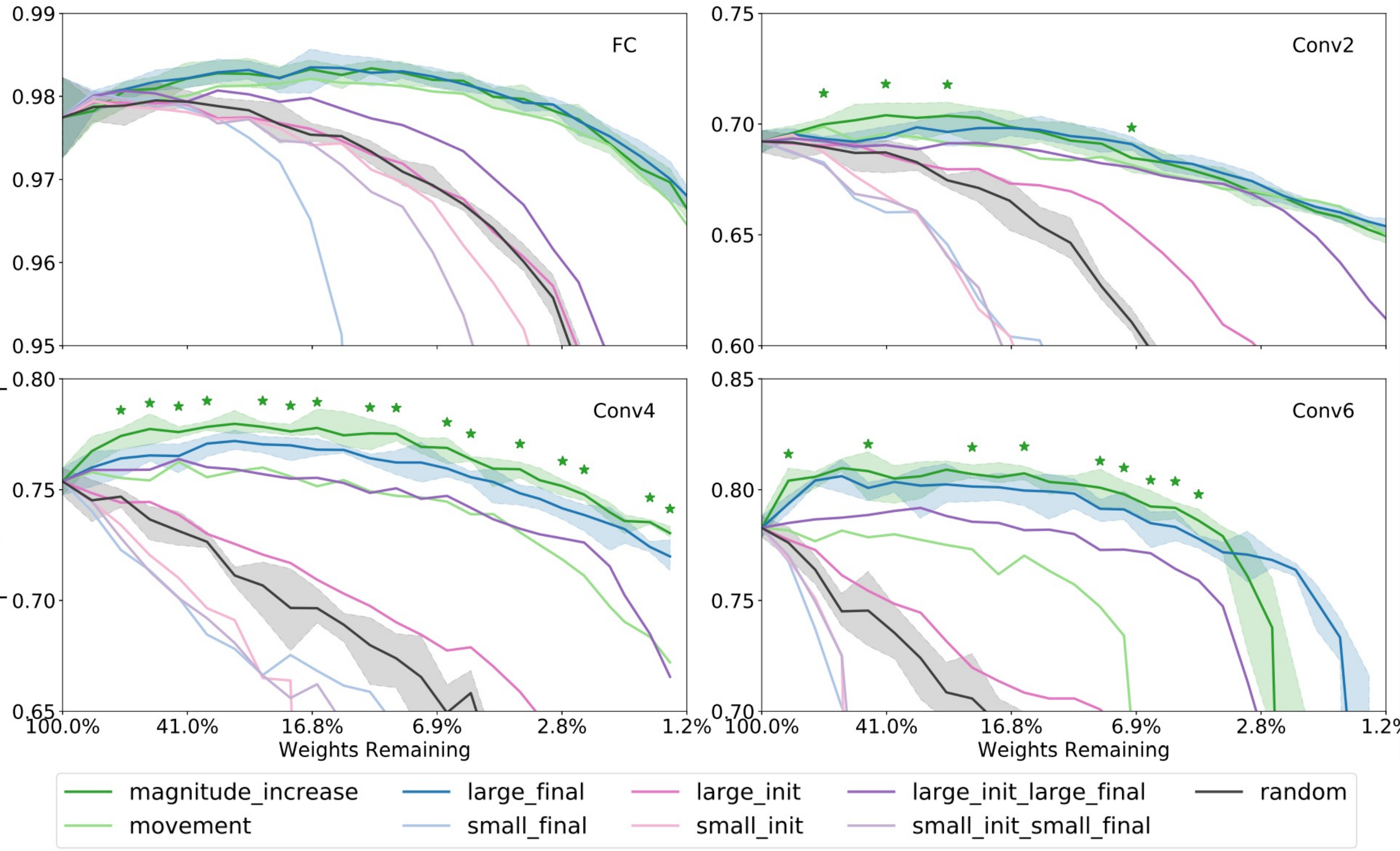
* Du S S, Zhai X, Poczos B, et al. Gradient Descent Provably Optimizes Over-parameterized Neural Networks[C]// ICLR. 2018.

Sparsity

2. Deconstructing lottery tickets

the magnitude_increase criterion turns out to work just as well as the large_final criterion, and in some cases significantly better





[13] Zhou H, Lan J, Liu R, et al. Deconstructing lottery tickets: Zeros, signs, and the supermask[C]. NIPS, 2019.

Sparsity

2. Deconstructing lottery tickets

- training the mask, instead of training network weights can get competitive performance[13].

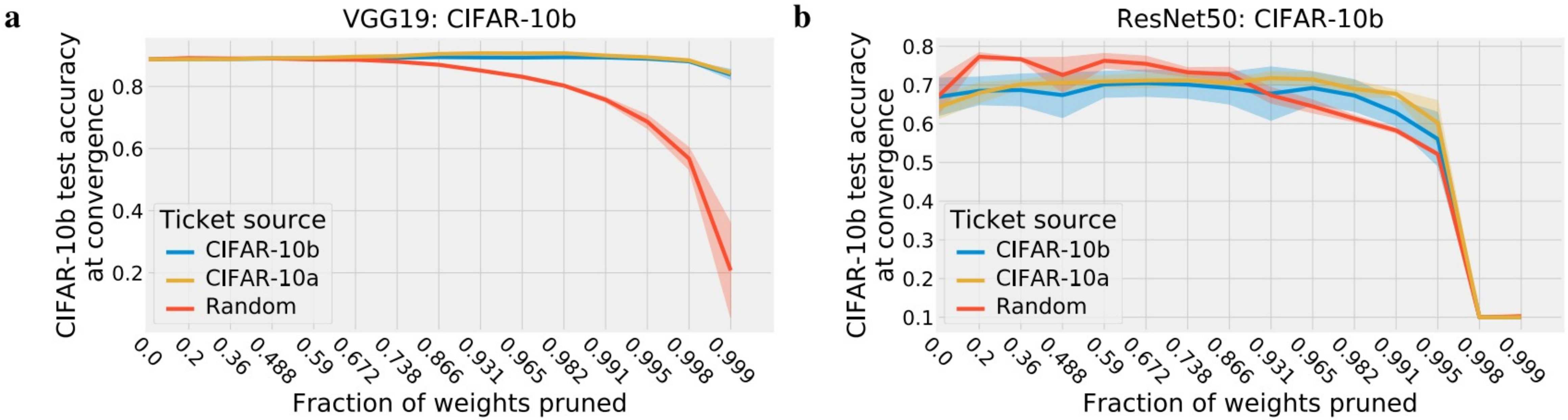
- Proved by [14], a ReLU network of arbitrary depth L can be approximated by pruning weight of a random initialized network of depth 2L and sufficient width. (But computationally hard!)

Network	mask $\odot$ init	mask $\odot$ S.C.	learned mask $\odot$ init	learned mask $\odot$ S.C.	DWR learned mask $\odot$ init	DWR learned mask $\odot$ S.C.	trained weights
MNIST FC	79.3	86.3	95.3	96.4	97.8	**98.0**	97.7
CIFAR Conv2	22.3	37.4	64.4	**66.3**	65.0	66.0	69.2
CIFAR Conv4	23.	39.7	65.4	66.2	71.7	**72.5**	75.4
CIFAR Conv6	24.0	41.0	65.3	65.4	76.3	**76.5**	78.3

[13] Zhou H, Lan J, Liu R, et al. Deconstructing lottery tickets: Zeros, signs, and the supermask[C]. NIPS, 2019.
[14] Malach E, Yehudai G, Shalev-Schwartz S, et al. Proving the lottery ticket hypothesis: Pruning is all you need[C]//ICML, 2020: 6682-6691.

3. Generalizing lottery tickets

winning tickets provide beneficial inductive bias



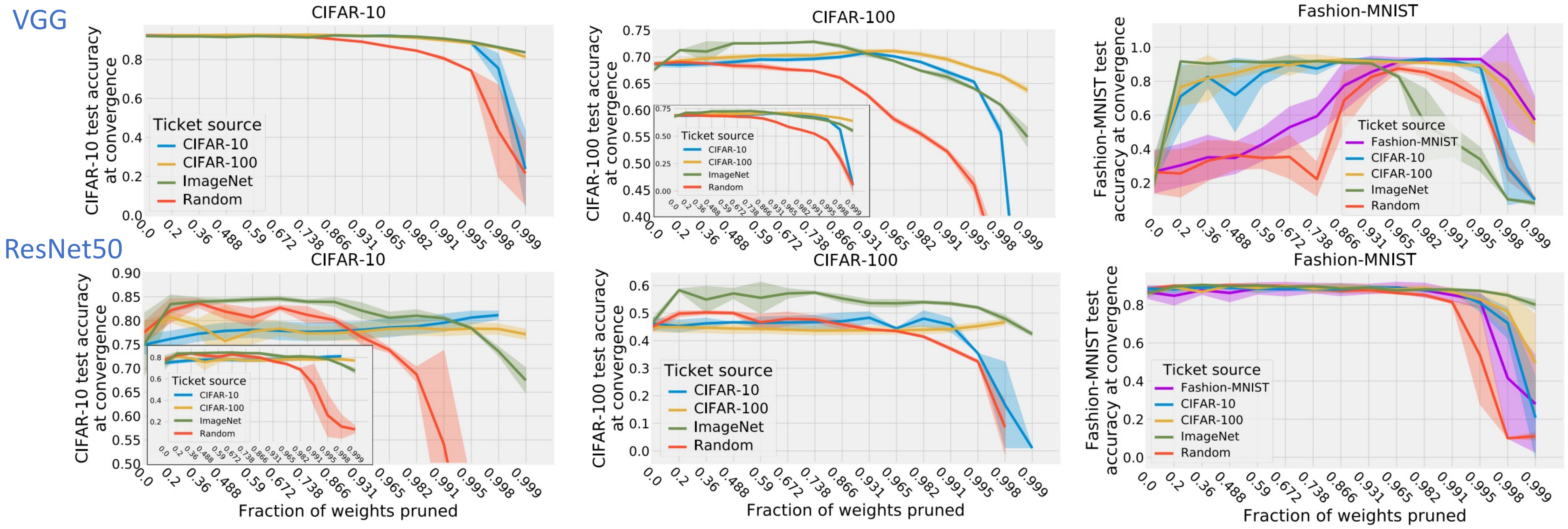
within the same data distribution

[15] Morcos A S, Yu H, Paganini M, et al. One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers[C]. NIPS, 2019..

Sparsity

3. Generalizing lottery tickets

winning tickets provide beneficial inductive bias



across datasets

[15] Morcos A S, Yu H, Paganini M, et al. One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers[C]. NIPS, 2019..

Remaining Questions and Possible Directions

1. How to reduce computational cost, increase learning stability/robustness?
2. What make the winning tickets special? How to balance between over-parametrization and sparsity, and enhance generalization?
3. Is there room to improve the initialization methods?

- ➢ Advance in pruning algorithms…
  [What’s hidden in a randomly weighted neural network? CVPR 2020]
  [Picking Winning Tickets Before Training by Preserving Gradient Flow. ICLR 2020]
- ➢ Exploit the optimization (or generalization) properties…
  [One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers. NIPS 2019]
  [Linear Mode Connectivity and the Lottery Ticket Hypothesis. ICML 2020]
- ➢ Investigate early learning…
  [The Early Phase Of Neural Network Training. ICLR 2020]
  [Robust Early-learning: Hindering The Memorization Of Noisy Labels. ICLR 2021]

Thanks!