ECE826 Lecture 11:

The success of Deep Learning: Is it all about SGD?

Contents

- On the (lack of) Implicit Bias of SGD
- Bad Local Minima Exist
- SGD Can Reach them

Last time: How fast we can approximate ERM

- The empirical cost function that we have access to

$$\min_{h \in \mathscr{H}} \left(R_S[h] = \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i); y_i) \right)$$

- Question: Can we approximate the solution to this minimization? If so how fast?

- The answer must depend on:
  1) n, the sample size
  2) $\mathscr{H}$, the hypothesis class and loss function
  3) $\mathscr{D}$, the data distribution
  4) the optimization algorithm that outputs our classifier

Loss landscapes and optimization in over-parameterized non-linear systems and neural networks

Chaoyue Liu[a], Libin Zhu[b,c], and Mikhail Belkin[c]

[a]*Department of Computer Science and Engineering, The Ohio State University*
[b]*Department of Computer Science and Engineering, University of California, San Diego*
[c]*Halicioğlu Data Science Institute, University of California, San Diego*

May 28, 2021

Overparameterized Nonlinear Learning: Gradient Descent Takes the Shortest Path?

Samet Oymak* and Mahdi Soltanolkotabi†

A Convergence Theory for Deep Learning via Over-Parameterization

Zeyuan Allen-Zhu
zeyuan@csail.mit.edu

Yuanzhi Li
yuanzhil@stanford.edu

Zhao Song
zhaos@utexas.e
UT-Austin
ity of Washington

On the Convergence Rate of Training Recurrent Neural Networks

Zeyuan Allen-Zhu
zeyuan@csail.mit.edu
Microsoft Research AI

Yuanzhi Li
yuanzhil@stanford.edu
Stanford University
Princeton University

Zhao Song
zhaos@utexas.edu
UT-Austin
University of Washington
Harvard University

October 28, 2018

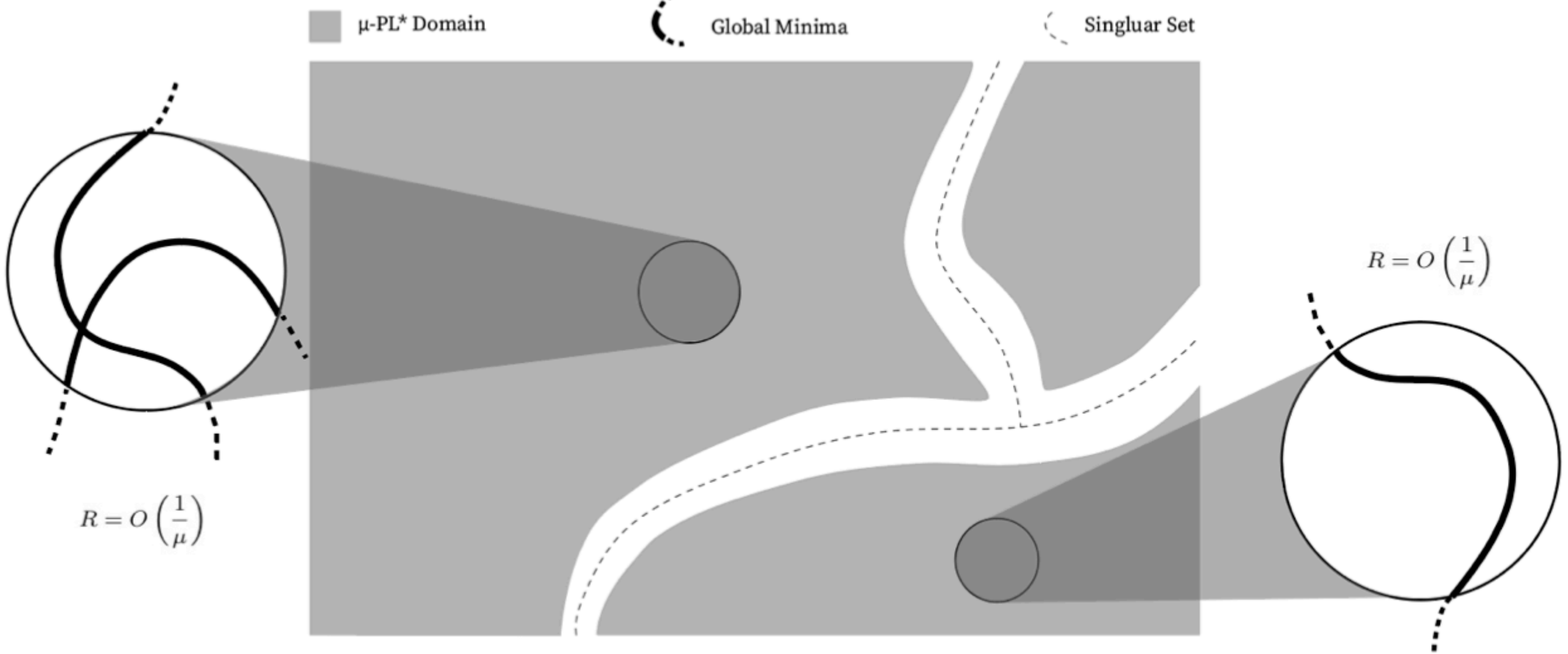
No bad local minima: Data independent training error guarantees for multilayer neural networks

Daniel Soudry
Department of Statistics
Columbia University
New York, NY 10027, USA
daniel.soudry@gmail.com

Yair Carmon
Department of Electrical Engineering
Stanford University
Stanford, CA 94305, USA
yairc@stanford.edu

Gradient Descent Finds Global Minima of Deep Neural Networks

Simon S. Du [*1] **Jason D. Lee** [*2] **Haochuan Li** [*3 4] **Liwei Wang** [*5 4] **Xiyu Zhai** [*6]



via Over-Parameterization

Zeyuan Allen-Zhu
zeyuan@csail.mit.edu
Microsoft Research AI

Yuanzhi Li
yuanzhil@stanford.edu
Stanford University
Princeton University

Zhao Song
zhaos@utexas.edu
UT-Austin
University of Washington
Harvard University

PL-like conditions hold in neighborhoods around initialization/optima.

Current theoretical SOTA

Subquadratic Overparameterization for Shallow Neural Networks

Chaehwan Song[1*] Ali Ramezani-Kebrya[1*]

Thomas Pethick[1] Armin Eftekhari[2†] Volkan Cevher[1]

something odd..

Table 1: Scaling with the number of training data in the overparameterization regime. QL=quadratic loss, CLL=convex and Lipschitz loss, SD=separable data.

Depth	Algorithm	Setting	Activation	Scaling	Reference
2	GD on layer 1	QL	ReLU	$\tilde{\Omega}(n^2)$	Oymak and Soltanolkotabi [38]
L	GD on layer L	CLL	ReLU	$\tilde{\Omega}(n)$	Kawaguchi and Huang [21]
2	GD	SD	ReLU	$\tilde{\Omega}(n^2)$	Song and Yang [39]
2	GD	SD and QL	ReLU	$\tilde{\Omega}(n^6)$	Du et al. [12]
L	GD	SD and QL	ReLU	$\Omega(n^8 L^{12})$	Zou and Gu [44]
2	GD	QL	Smooth	$\tilde{\Omega}(n^{\frac{3}{2}})$	**This paper**

A curious observation on fitting the data

Small ReLU networks are powerful memorizers: a tight analysis of memorization capacity

Chulhee Yun
MIT
Cambridge, MA 02139
chulheey@mit.edu

Suvrit Sra
MIT
Cambridge, MA 02139
suvrit@mit.edu

Ali Jadbabaie
MIT
Cambridge, MA 02139
jadbabai@mit.edu

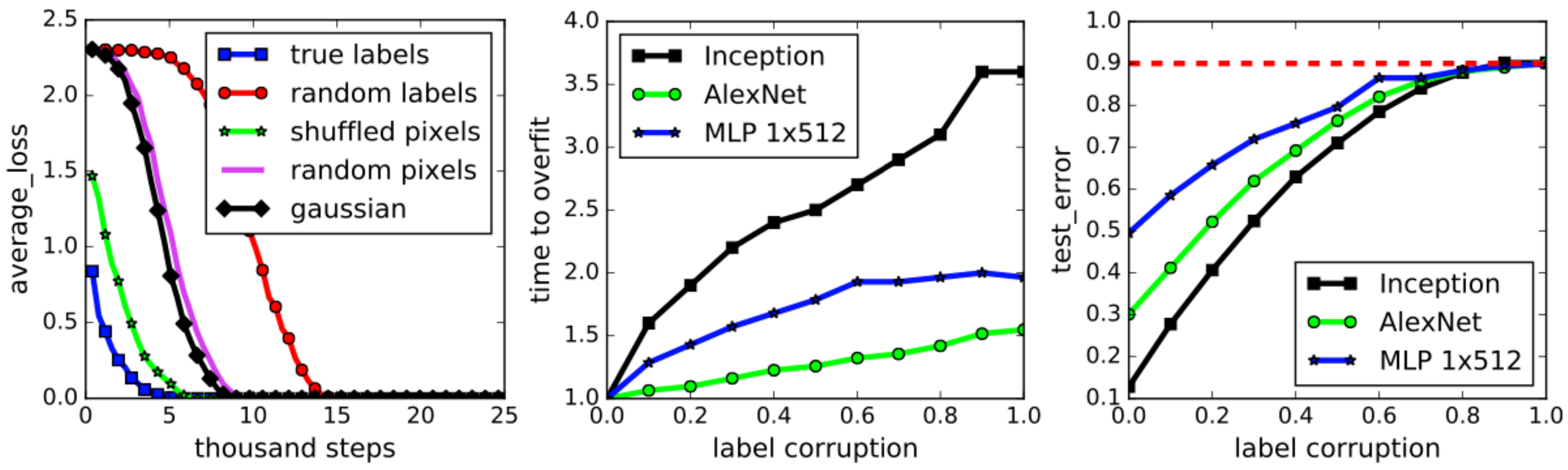
Theorem:

Any data set of size n can be memorized by a 3-layer ReLU neural network with $O(n)$ weights.

These constructions can be made in linear time. Yet SGD on the same arch needs so much more larger overarm. Why??

But somehow SGD does more than just that..

Rethinking Generalization [Zhang et al. ICLR17]

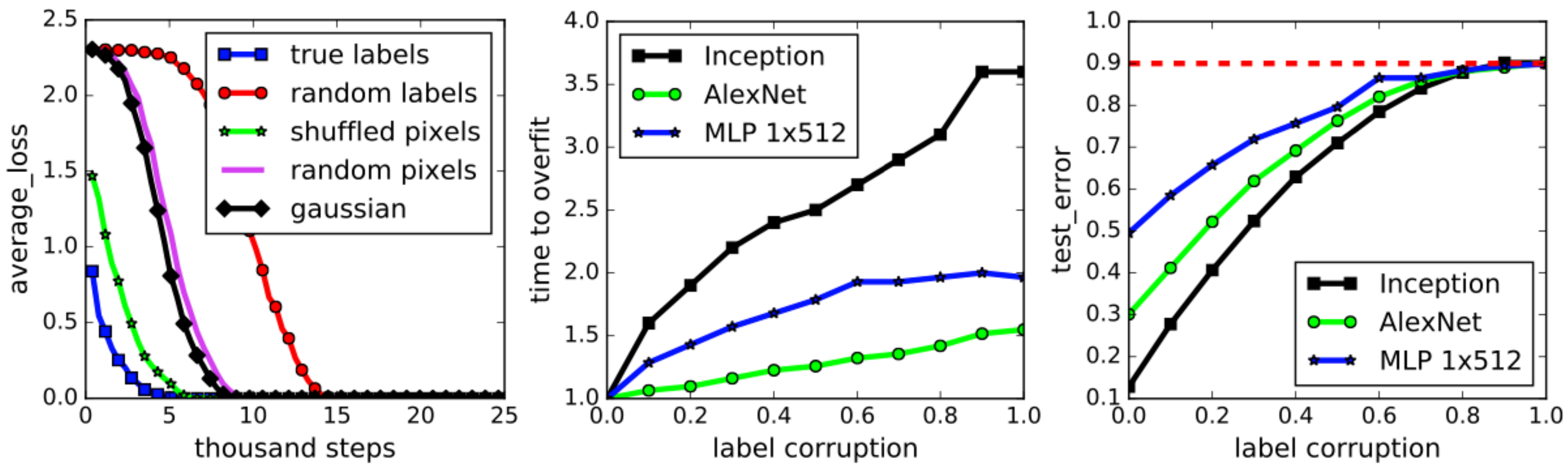


(a) learning curves (b) convergence slowdown (c) generalization error growth

Figure 1: Fitting random labels and random pixels on CIFAR10. (a) shows the training loss of various experiment settings decaying with the training steps. (b) shows the relative convergence time with different label corruption ratio. (c) shows the test error (also the generalization error since training error is 0) under different label corruptions.

- Overparameterized, SGD-trained models :
  1. Can fit even completely random labels (i.e., huge capacity)
  2. Yet, generalize well

Rethinking Generalization [Zhang et al. ICLR17]



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

Figure 1: Fitting random labels and random pixels on CIFAR10. (a) shows the training loss of various experiment settings decaying with the training steps. (b) shows the relative convergence time with different label corruption ratio. (c) shows the test error (also the generalization error since training error is 0) under different label corruptions.

- Overparameterized, SGD-trained models :
  1. Can fit even completely random labels (i.e., huge capacity)
  2. Yet, generalize well

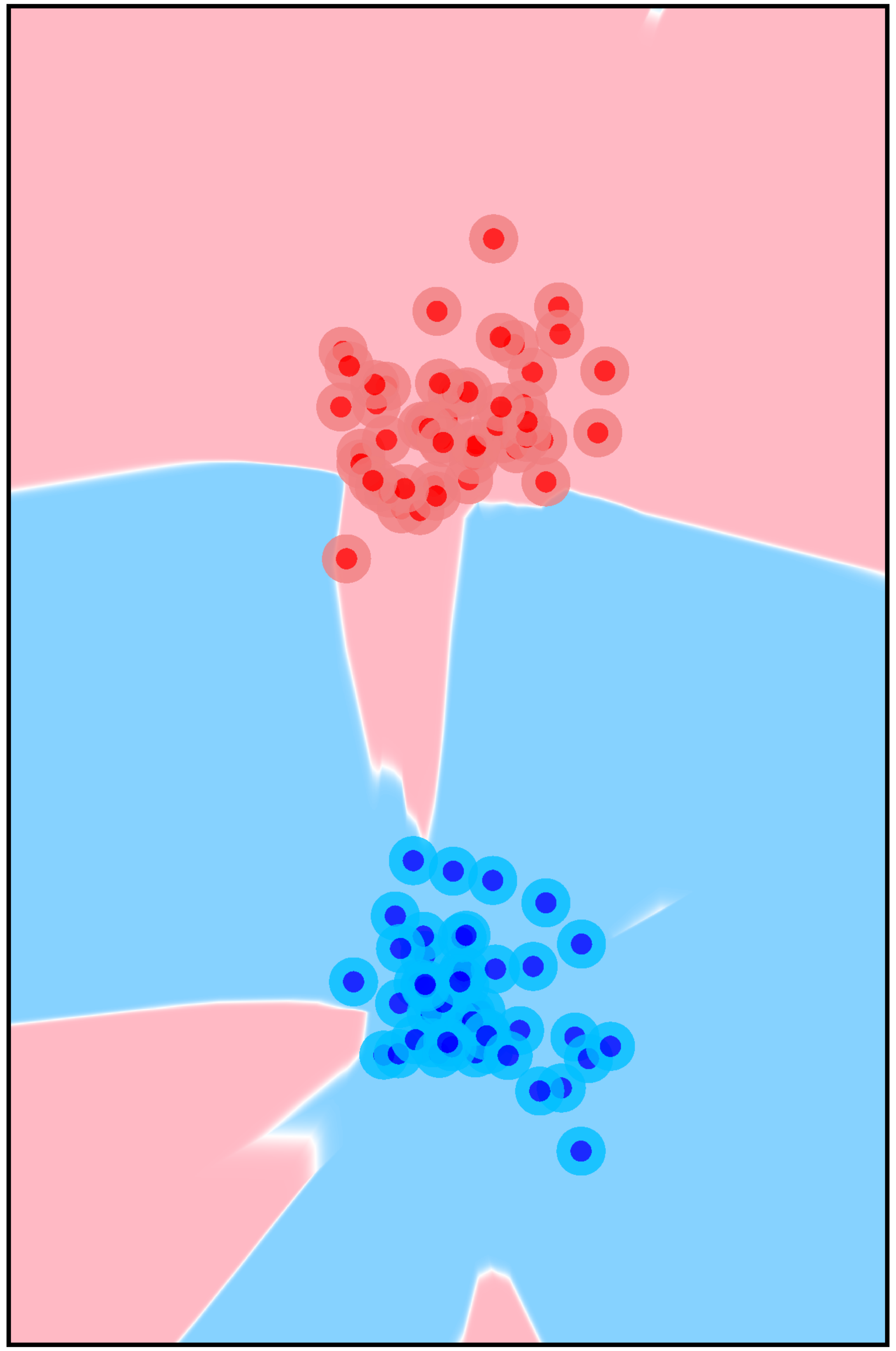
Open Question: How can this be?

Possible Explanations of Generalization

- Maybe every model that fits the training data generalizes (no bad global minima)
- Maybe SGD is special "can avoid" bad global minima (implicit regularization)?
- Maybe the data distribution is what allows everything to fall into place?

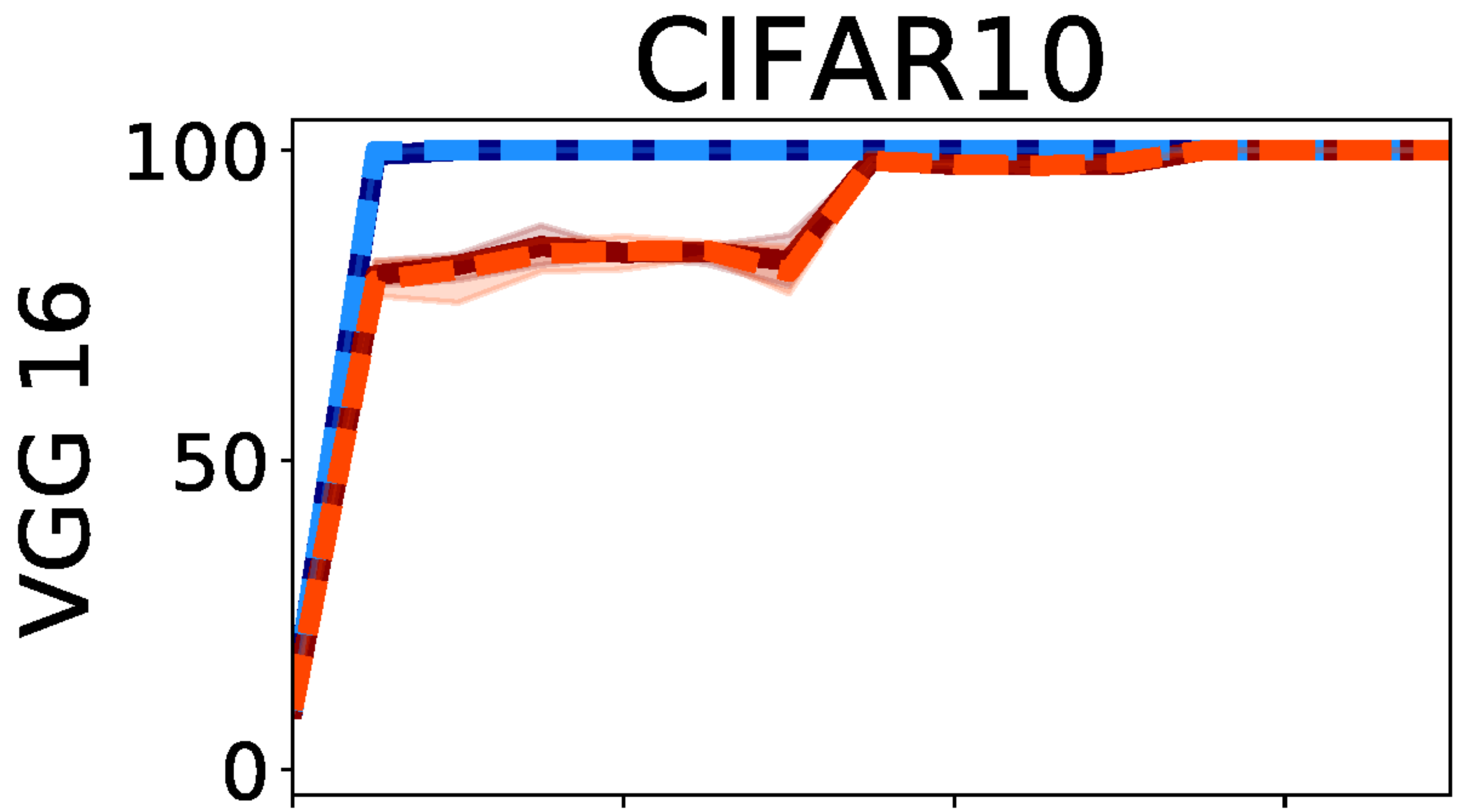
Maybe all interpolating points generalize!

What is a bad global minimum?



Bad Minima = zero margin/complex boundary
=> 100% train error + poor test

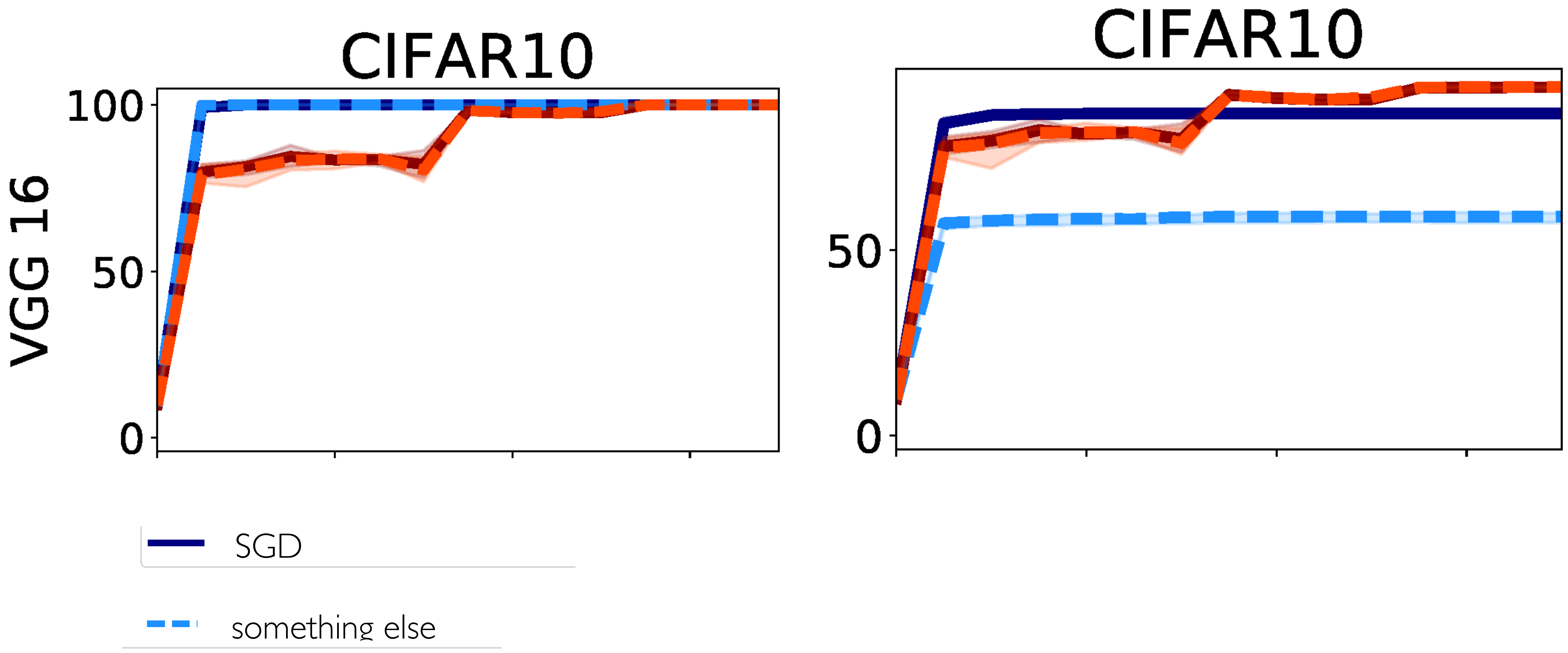
Bad Global Minima Exist



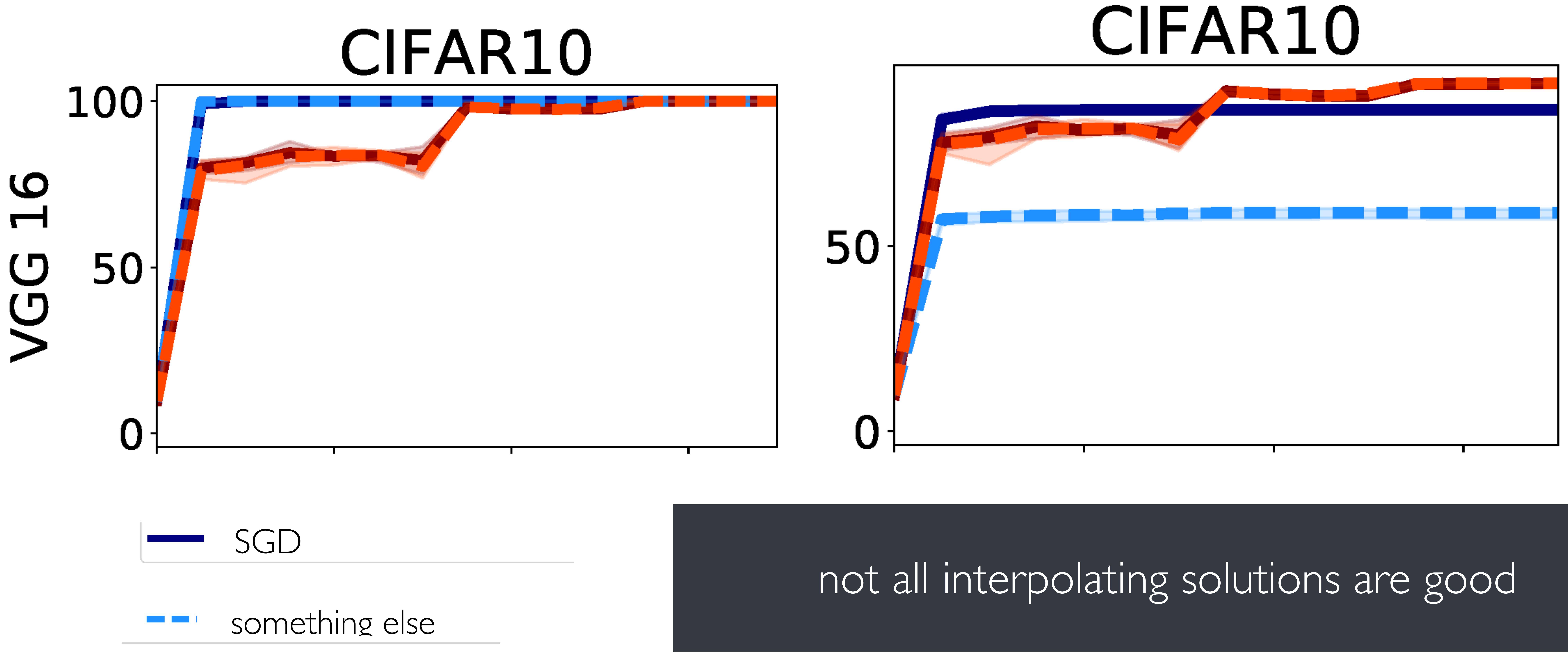
SGD

something else

Bad Global Minima Exist



Bad Global Minima Exist



not all interpolating solutions are good

Possible Explanations of Generalization

nope

- Maybe every model that fits the training data generalizes (no bad global minima)
- Maybe SGD is special "can avoid" bad global minima (implicit regularization)?
- Maybe the data distribution is what allows everything to fall into place?

Maybe (S)GD is special?

GD + LS = a red herring

- Let's say we want to solve a least squares problem $\min_w \|X^T w - y\|^2$ with GD

GD + LS = a red herring

- Let's say we want to solve a least squares problem $\min_w \|X^T w - y\|^2$ with GD
- The iterates of GD look like

$$\begin{aligned} w_{k+1} &= w_k - \frac{\gamma}{2} \nabla L(w_k) \\ &= w_k - \gamma X(X^T w_k - y) \end{aligned}$$

GD + LS = a red herring

- Let's say we want to solve a least squares problem $\min_{w} \|X^T w - y\|^2$ with GD
- The iterates of GD look like

$$
\begin{aligned}
w_{k+1} &= w_k - \frac{\gamma}{2} \nabla L(w_k) \\
&= w_k - \gamma X(X^T w_k - y) \\
&= (I_d - \gamma XX^T) w_k + \gamma Xy \\
&= (I_d - \gamma XX^T)^2 w_{k-1} + (I_d - \gamma XX^T)\gamma Xy + \gamma Xy
\end{aligned}
$$

GD + LS = a red herring

- Let's say we want to solve a least squares problem $\min_{w} \|X^T w - y\|^2$ with GD
- The iterates of GD look like

$$
\begin{aligned}
w_{k+1} &= w_k - \frac{\gamma}{2} \nabla L(w_k) \\
&= w_k - \gamma X(X^T w_k - y) \\
&= (I_d - \gamma X X^T) w_k + \gamma X y \\
&= (I_d - \gamma X X^T)^2 w_{k-1} + (I_d - \gamma X X^T)\gamma X y + \gamma X y \\
&= (I_d - \gamma X X^T)^2 w_{k-1} + \gamma \left(\sum_{i=0}^{1} (I_d - \gamma X X^T)^i \right) X y
\end{aligned}
$$

GD + LS = a red herring

- Let's say we want to solve a least squares problem $\min_w \|X^T w - y\|^2$ with GD
- The iterates of GD look like

$$
\begin{aligned}
w_{k+1} &= w_k - \frac{\gamma}{2} \nabla L(w_k) \\
&= w_k - \gamma X(X^T w_k - y) \\
&= (I_d - \gamma X X^T) w_k + \gamma X y \\
&= (I_d - \gamma X X^T)^2 w_{k-1} + (I_d - \gamma X X^T)\gamma X y + \gamma X y \\
&= (I_d - \gamma X X^T)^2 w_{k-1} + \gamma \left(\sum_{i=0}^{1} (I_d - \gamma X X^T)^i \right) X y \\
&= (I_d - \gamma X X^T)^k w_0 + \gamma \left(\sum_{i=0}^{k-1} (I_d - \gamma X X^T)^i \right) X y
\end{aligned}
$$

GD + LS = a red herring

- Let's say we want to solve a least squares problem $\min_w \|X^T w - y\|^2$ with GD
- The iterates look like $w_{k+1}(I_d - \gamma XX^T)^k w_0 + \gamma \left(\sum_{i=0}^{k-1} (I_d - \gamma XX^T)^i \right) Xy$

GD + LS = a red herring

- Let's say we want to solve a least squares problem $\min_w \|X^T w - y\|^2$ with GD
- The iterates look like $w_{k+1}(I_d - \gamma XX^T)^k w_0 + \gamma \left(\sum_{i=0}^{k-1} (I_d - \gamma XX^T)^i \right) Xy$
- Assuming we start at zero, the iterates of GD look like

GD + LS = a red herring

- Let's say we want to solve a least squares problem $\min_{w} \|X^T w - y\|^2$ with GD
- The iterates look like $w_{k+1}(I_d - \gamma XX^T)^k w_0 + \gamma \left(\sum_{i=0}^{k-1} (I_d - \gamma XX^T)^i \right) Xy$
- Assuming we start at zero, the iterates of GD look like

$$w_{k+1} = \gamma \left(\sum_{i=0}^{k-1} (I_d - \gamma XX^T)^i \right) Xy$$

GD + LS = a red herring

- Let's say we want to solve a least squares problem $\min_w \|X^T w - y\|^2$ with GD
- The iterates look like $w_{k+1}(I_d - \gamma XX^T)^k w_0 + \gamma \left(\sum_{i=0}^{k-1} (I_d - \gamma XX^T)^i \right) Xy$
- Assuming we start at zero, the iterates of GD look like

$$w_{k+1} = \gamma \left(\sum_{i=0}^{k-1} (I_d - \gamma XX^T)^i \right) Xy$$

- What does that imply? Let's take GD to infinity

GD + LS = a red herring

- Let's say we want to solve a least squares problem $\min_w \|X^T w - y\|^2$ with GD
- The iterates look like $w_{k+1}(I_d - \gamma XX^T)^k w_0 + \gamma \left(\sum_{i=0}^{k-1} (I_d - \gamma XX^T)^i \right) Xy$
- Assuming we start at zero, the iterates of GD look like

$$w_{k+1} = \gamma \left(\sum_{i=0}^{k-1} (I_d - \gamma XX^T)^i \right) Xy$$

- What does that imply? Let's take GD to infinity

$$w_\infty = \gamma \left(\sum_{i=0}^{\infty} (I_d - \gamma XX^T)^i \right) Xy$$

GD + LS = a red herring

- Let's say we want to solve a least squares problem $\min_w \|X^T w - y\|^2$ with GD
- The iterates look like $w_{k+1}(I_d - \gamma XX^T)^k w_0 + \gamma \left(\sum_{i=0}^{k-1} (I_d - \gamma XX^T)^i \right) Xy$
- Assuming we start at zero, the iterates of GD look like

$$w_{k+1} = \gamma \left(\sum_{i=0}^{k-1} (I_d - \gamma XX^T)^i \right) Xy$$

- What does that imply? Let's take GD to infinity

$$w_\infty = \gamma \left(\sum_{i=0}^{\infty} (I_d - \gamma XX^T)^i \right) Xy$$

- Do you remember what this infinite sum converges to?

GD + LS = a red herring

- Let's say we want to solve a least squares problem $\min_w \|X^T w - y\|^2$ with GD
- The iterates look like $w_{k+1}(I_d - \gamma X X^T)^k w_0 + \gamma \left(\sum_{i=0}^{k-1} (I_d - \gamma X X^T)^i \right) Xy$
- Assuming we start at zero, the iterates of GD look like

$$w_{k+1} = \gamma \left(\sum_{i=0}^{k-1} (I_d - \gamma X X^T)^i \right) Xy$$

- What does that imply? Let's take GD to infinity

$$w_\infty = \gamma \left(\sum_{i=0}^{\infty} (I_d - \gamma X X^T)^i \right) Xy$$

- Do you remember what this infinite sum converges to?

$$\sum_{i=0}^{\infty} (I_d - \gamma X X^T)^i = (X X^T)^{-1}$$

GD + LS = a red herring

- Let's say we want to solve a least squares problem $\min_w \|X^T w - y\|^2$ with GD
- Let's take GD to infinity

$$w_\infty = (X^T X)^{-1} X y$$

GD + LS = a red herring

- Let's say we want to solve a least squares problem $\min_w \|X^T w - y\|^2$ with GD
- Let's take GD to infinity

$$w_\infty = (X^T X)^{-1} X y$$

- Do you remember what this is called?

GD + LS = a red herring

- Let's say we want to solve a least squares problem $\min_{w} \|X^T w - y\|^2$ with GD
- Let's take GD to infinity

$$w_\infty = (X^T X)^{-1} X y$$

- Do you remember what this is called?
- The minimum Euclidean norm solution of squares solution to $X^T w = y$

$$\arg\min_{w} \|w\|_2, \text{ s.t. } W^T x = y$$

IMPLICIT BIAS/Regularization??!!!

- Let's say we want to solve a least squares problem $\min_{w} \|X^T w - y\|^2$ with GD
- Let's take GD to infinity

$$w_\infty = (X^T X)^{-1} X y$$

- Do you remember what this is called?
- The minimum Euclidean norm solution of squares solution to $X^T w = y$

$$\arg\min_{w} \|w\|_2, \text{ s.t. } W^T x = y$$

out of all the linear functions that interpolate the training data, (S)GD selects the minimal Euclidean norm one.
Wow.

GD + LS = a red herring

Theorem

For linear least squares GD converges to the minimum norm solution of $X^T w = y$

GD is IMPLICITLY regularizing against large norm solutions? It's Implicitly biased towards GENERALIZABLE solutions?

Well, linear LS is what's special

Theorem

ANY algorithm that converges to 0-error and whose iterates converge to $w_\infty = \sum_i a_i x_i$ returns a min norm solution to the LS problem.

- Proof:

All interpolating solutions in the data span are min norm

Theorem

ANY algorithm that converges to 0-error and whose iterates converge to $w_\infty = \sum_i a_i x_i$ returns a min norm solution to the LS problem.

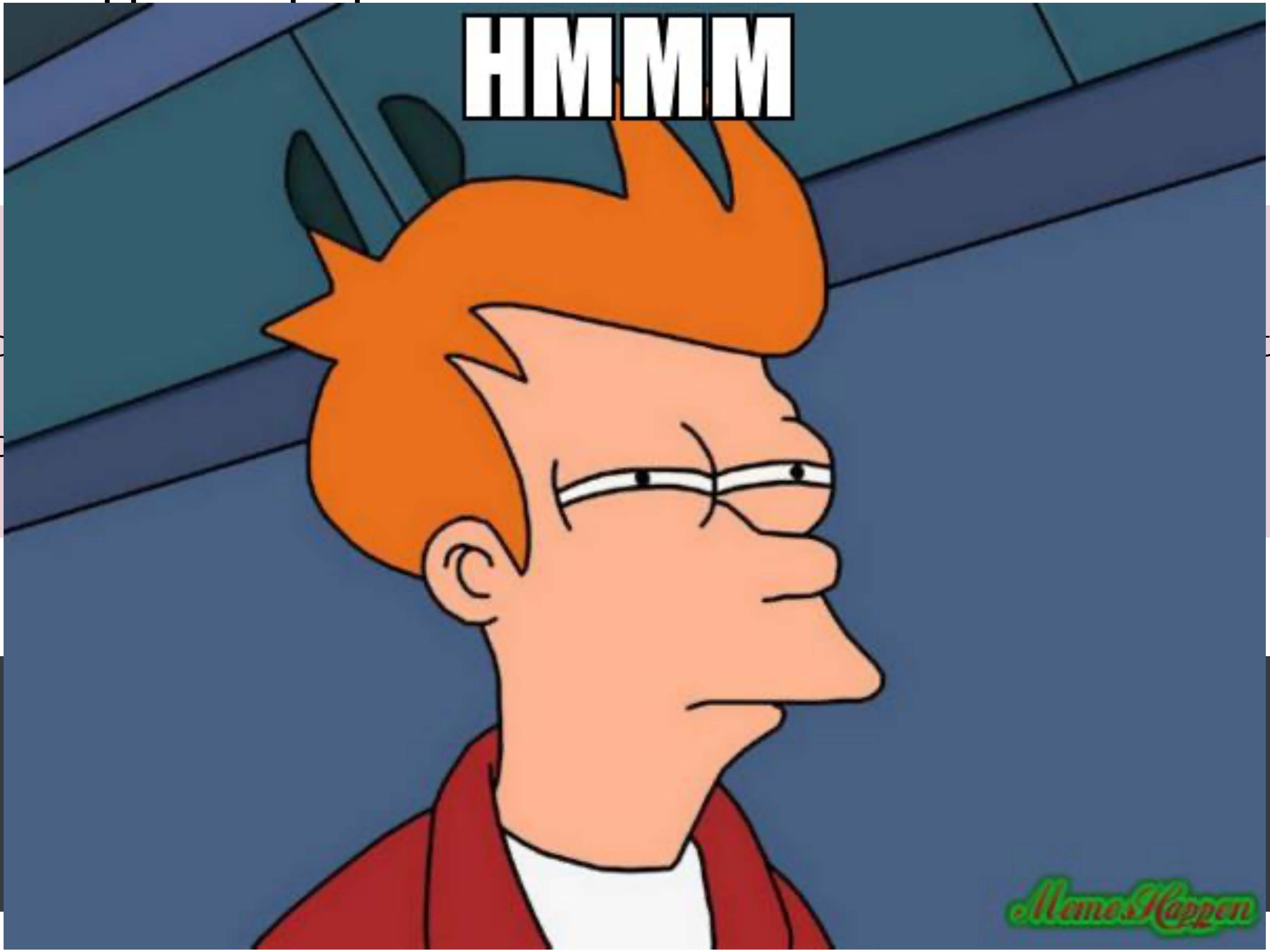
OK so maybe GD is … not that special???

All interpolating solutions

Theorem

ANY algorithm that c turns a min norm

solution to the LS pro



Implicit Bias of Gradient Descent on Linear Convolutional Networks

Suriya Gunasekar TTI at Chicago, USA suriya@ttic.edu

Jason D. Lee USC Los Angeles, USA jasonlee@marshall.usc.edu

Daniel Soudry Technion, Israel daniel.soudry@gmail.com

Nathan Srebro TTI at Chicago, USA nati@ttic.edu

Characterizing Implicit Bias in Terms of Optimization Geometry

Suriya Gunasekar [1] Jason Lee [2] Daniel Soudry [3] Nathan Srebro [1]

Implicit Regularization in Nonconvex Statistical Estimation: Gradient Descent Converges Linearly for Phase Retrieval and Matrix Completion

Cong Ma [1] Kaizheng Wang [1] Yuejie Chi [2] Yuxin Chen [3]

On the Spectral Bias of Neural Networks

Nasim Rahaman [*1 2] Aristide Baratin [*1] Devansh Arpit [1] Felix Draxler [2] Min Lin [1] Fred A. Hamprec
Yoshua Bengio [1] Aaron Courville [1]

Implicit Bias of Gradient Descent for Wide Two-layer Neural Networks Trained with the Logistic Loss

Lénaïc Chizat LENAIC.CHIZAT@UNIVERSITE-PARIS-SACLAY.FR
Laboratoire de Mathématiques d'Orsay, CNRS, Université Paris-Saclay, France

Francis Bach FRANCIS.BACH@INRIA.FR
INRIA, ENS, PSL Research University, Paris, France

Implicit Regularization in Deep Matrix Factorization

Princeton University and Institute for Advanced Study arora@cs.princeton.edu

Tel Aviv University cohennadav@cs.tau.ac.il

for some problems, GD does look like it's converging on solutions that seem to be regularized (small norm), in some sense

Implicit Regularization in Matrix Factorization

Suriya Gunasekar TTI at Chicago suriya@ttic.edu

Blake Woodworth TTI at Chicago blake@ttic.edu

Srinadh Bhojanapalli TTI at Chicago srinadh@ttic.edu

Behnam Neyshabur TTI at Chicago behnam@ttic.edu

Nathan Srebro TTI at Chicago nati@ttic.edu

In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Dee[p] Learning

Behnam Neyshabur, Ryota Tomioka & Nathan Srebro
Toyota Technological Institute at Chicago
Chicago, IL 60637, USA
{bneyshabur,tomioka,nati}@ttic.edu

The Implicit Bias of Gradient Descent on Separable Data

Daniel Soudry DANIEL.SOUDRY@GMAIL.COM
Elad Hoffer ELAD.HOFFER@GMAIL.COM
Mor Shpigel Nacson MOR.SHPIGEL@GMAIL.COM
Department of Electrical Engineering,Technion
Haifa, 320003, Israel

Suriya Gunasekar SURIYA@TTIC.EDU
Nathan Srebro NATI@TTIC.EDU
Toyota Technological Institute at Chicago
Chicago, Illinois 60637, USA

Theorem 3 *For any dataset which is linearly separable (Assumption 1), any β-smooth decreasing loss function (Assumption 2) with an exponential tail (Assumption 3), any stepsize $\eta < 2\beta^{-1}\sigma_{\max}^{-2}(\mathbf{X})$ and any starting point $\mathbf{w}(0)$, the gradient descent iterates (as in eq. 2) will behave as:*

$$\mathbf{w}(t) = \hat{\mathbf{w}} \log t + \boldsymbol{\rho}(t) \ , \tag{3}$$

where $\hat{\mathbf{w}}$ is the L_2 max margin vector (the solution to the hard margin SVM):

$$\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}} \|\mathbf{w}\|^2 \ \text{ s.t. } \mathbf{w}^\top \mathbf{x}_n \geq 1, \tag{4}$$

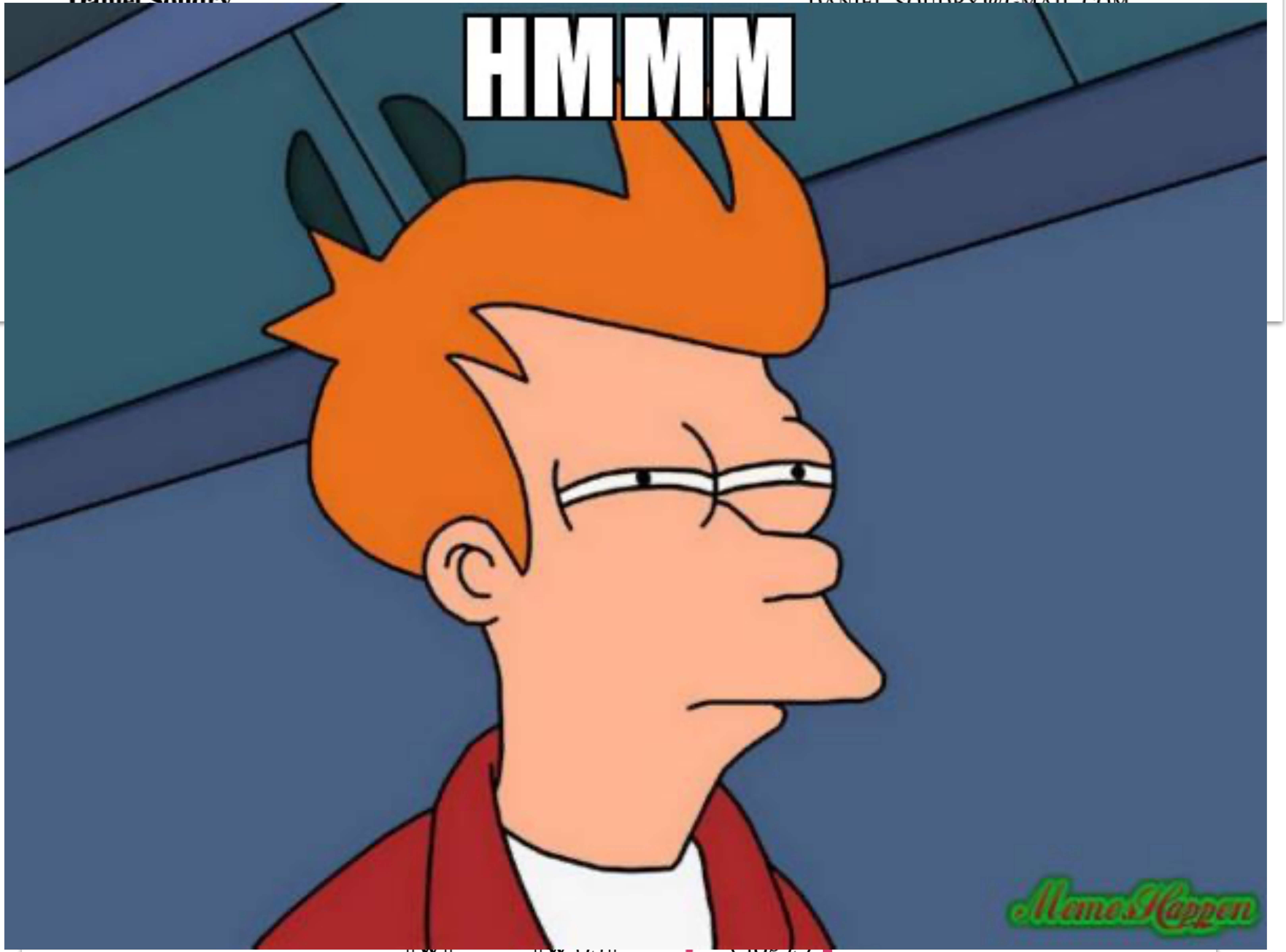
and the residual grows at most as $\|\boldsymbol{\rho}(t)\| = O(\log\log(t))$, and so

$$\lim_{t\to\infty} \frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|} = \frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|}.$$

Furthermore, for almost all data sets (all except measure zero), the residual $\rho(t)$ is bounded.

The Implicit Bias of Gradient Descent on Separable Data

Daniel Soudry DANIEL.SOUDRY@GMAIL.COM



Does SGD really regularize??

Implicit Regularization in ReLU Networks with the Square Loss

Gal Vardi GAL.VARDI@WEIZMANN.AC.IL and **Ohad Shamir** OHAD.SHAMIR@WEIZMANN.AC.IL
Weizmann Institute of Science

Editors: Mikhail Belkin and Samory Kpotufe

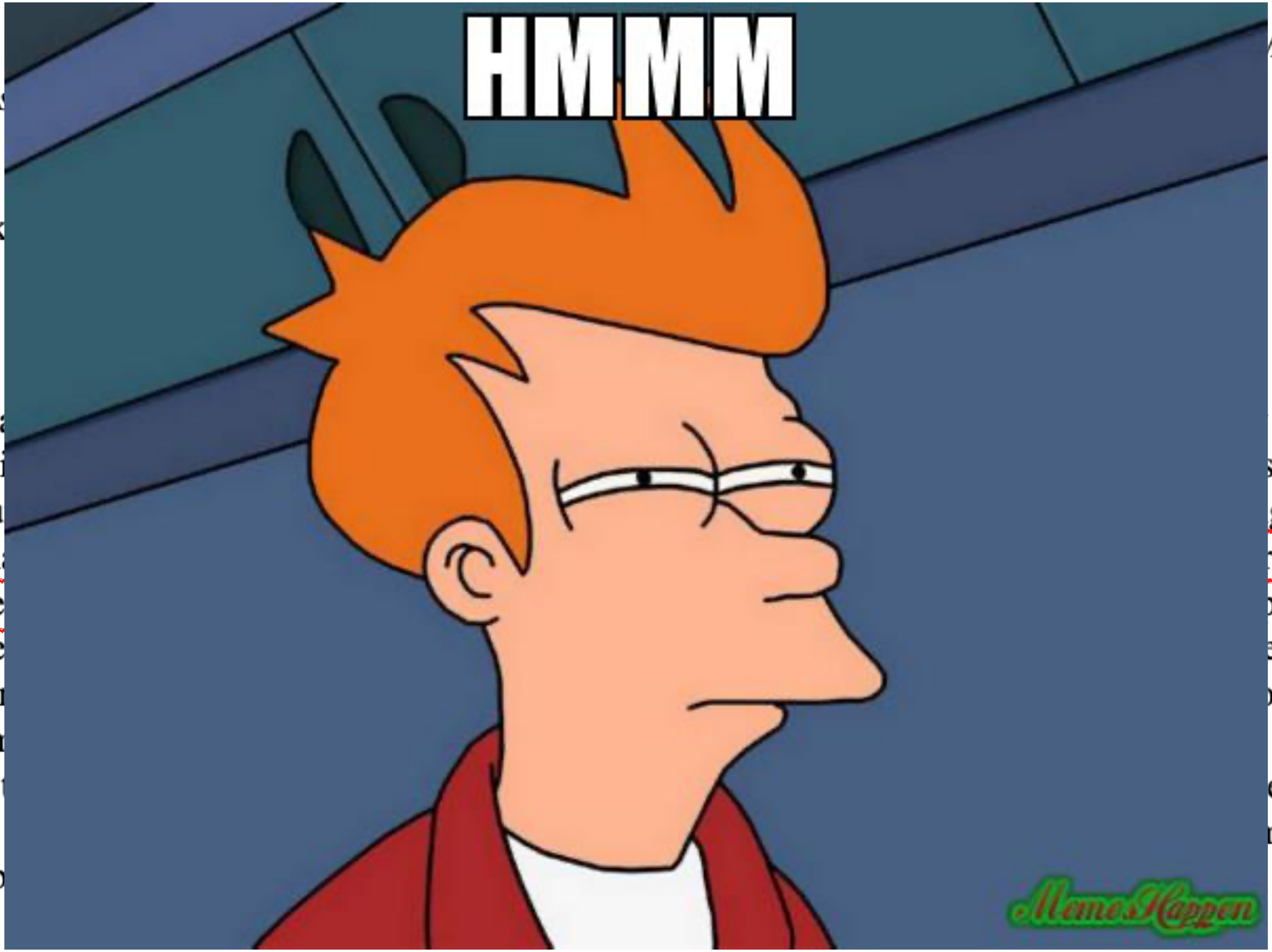
Abstract

Understanding the implicit regularization (or implicit bias) of gradient descent has recently been a very active research area. However, the implicit regularization in nonlinear neural networks is still poorly understood, especially for regression losses such as the square loss. Perhaps surprisingly, we prove that even for a single ReLU neuron, it is impossible to characterize the implicit regularization with the square loss by any explicit function of the model parameters (although on the positive side, we show it can be characterized approximately). For one hidden-layer networks, we prove a similar result, where in general it is impossible to characterize implicit regularization properties in this manner, except for the "balancedness" property identified in Du et al. (2018). Our results suggest that a more general framework than the one considered so far may be needed to understand implicit regularization for nonlinear predictors, and provides some clues on what this framework should be.

Implicit Regularization in ReLU Networks with the Square Loss

Gal Vardi ... MANN.AC.IL
Weizmann Ins

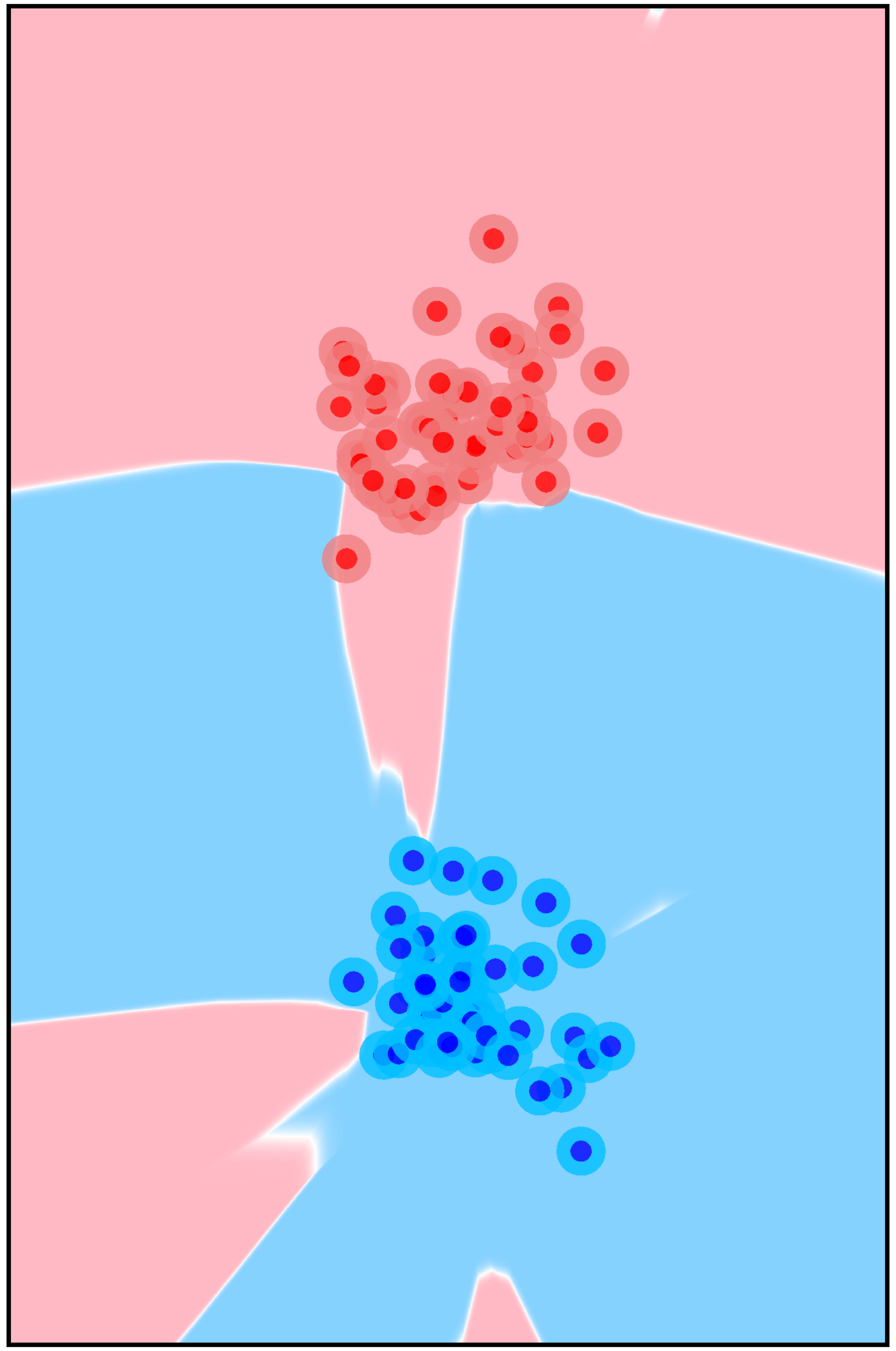
Editors: Mik



Understa ... been a
very acti ... is still
poorly u ... gly, we
prove th ... ization
with the ... ositive
side, we ... prove
a simila ... perties
in this n ... results
suggest ... erstand
implicit ... nework
should b

But maybe some form of hard to describe regularization is happening???

Can SGD reach bad global minima?

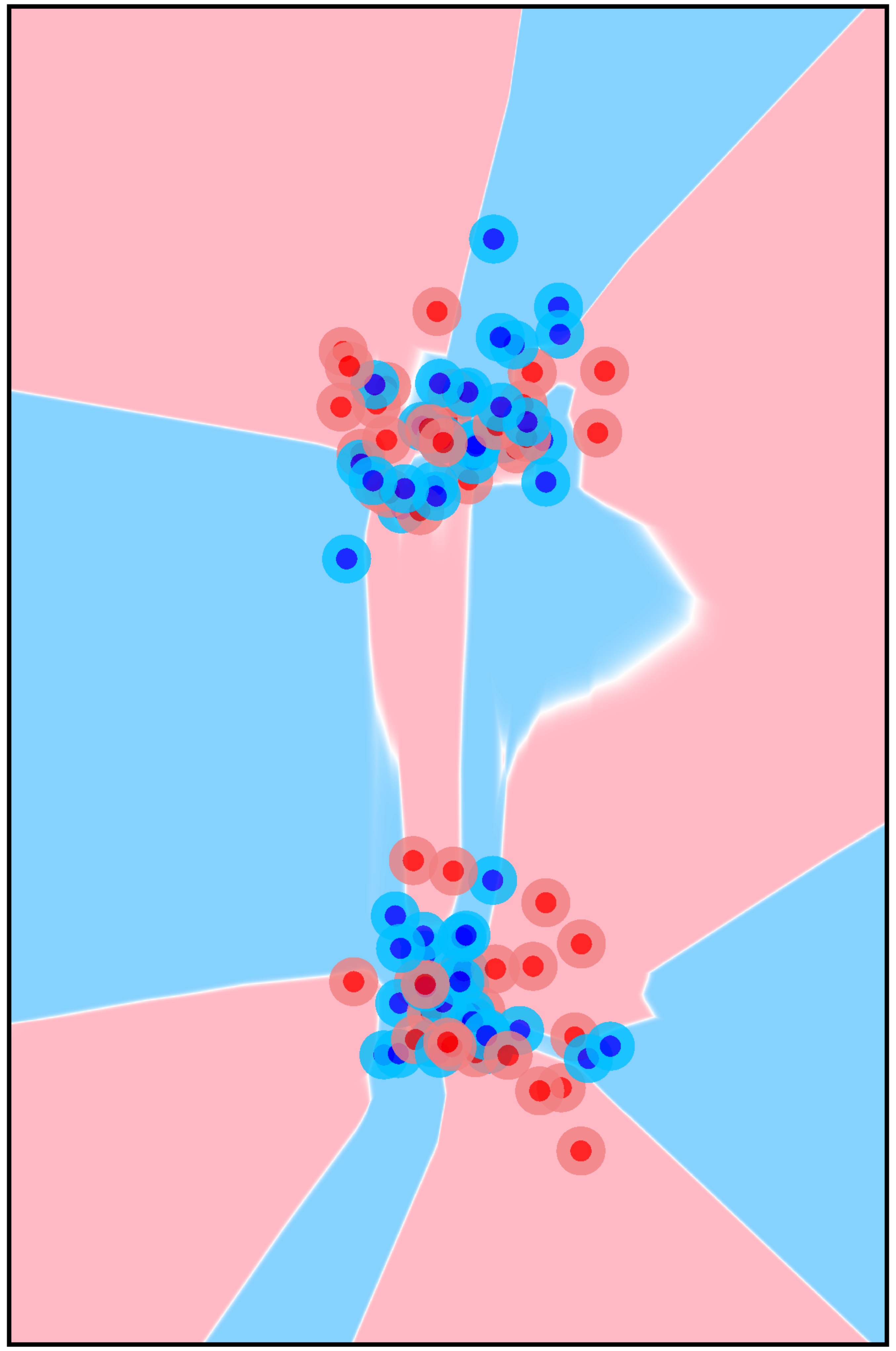


- Of course … if you initialize at a bad global min.

- But, SGD still converges to them, if adversarially initialized <u>even without loss-landscape knowledge</u>

We can construct initializers
using only <u>unlabeled data</u>
From which SGD is attracted to bad global minima

Adversarial Initialization



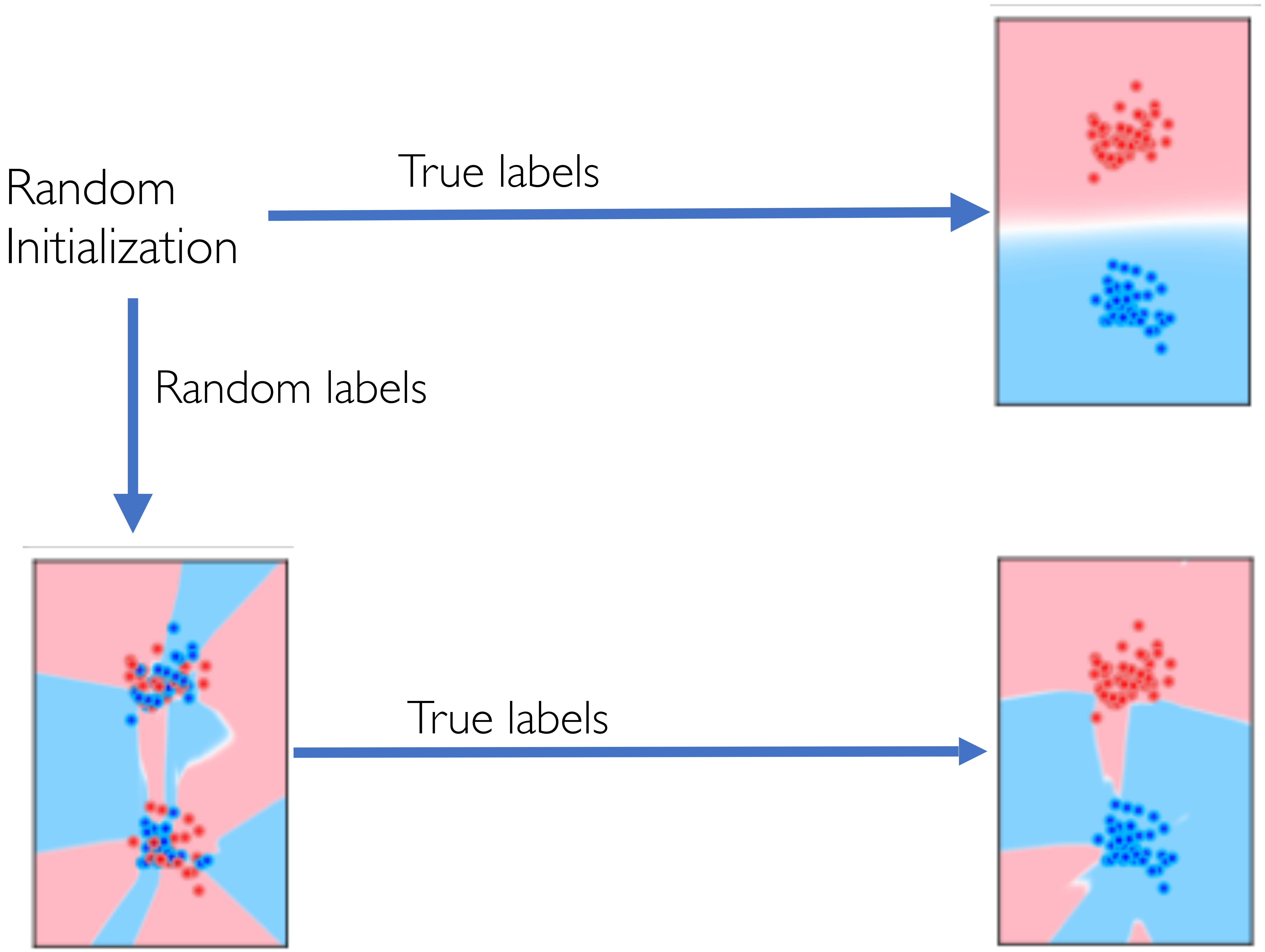
Input: Training dataset *S*; Replication factor *R*; Noise factor *N*

for every image x ∈ S repeat *R times*

zero-out a random subset of N % pixels in x
give it a random label
Add it to set C

Train to 100% accuracy on C, from a random init using vanilla SGD

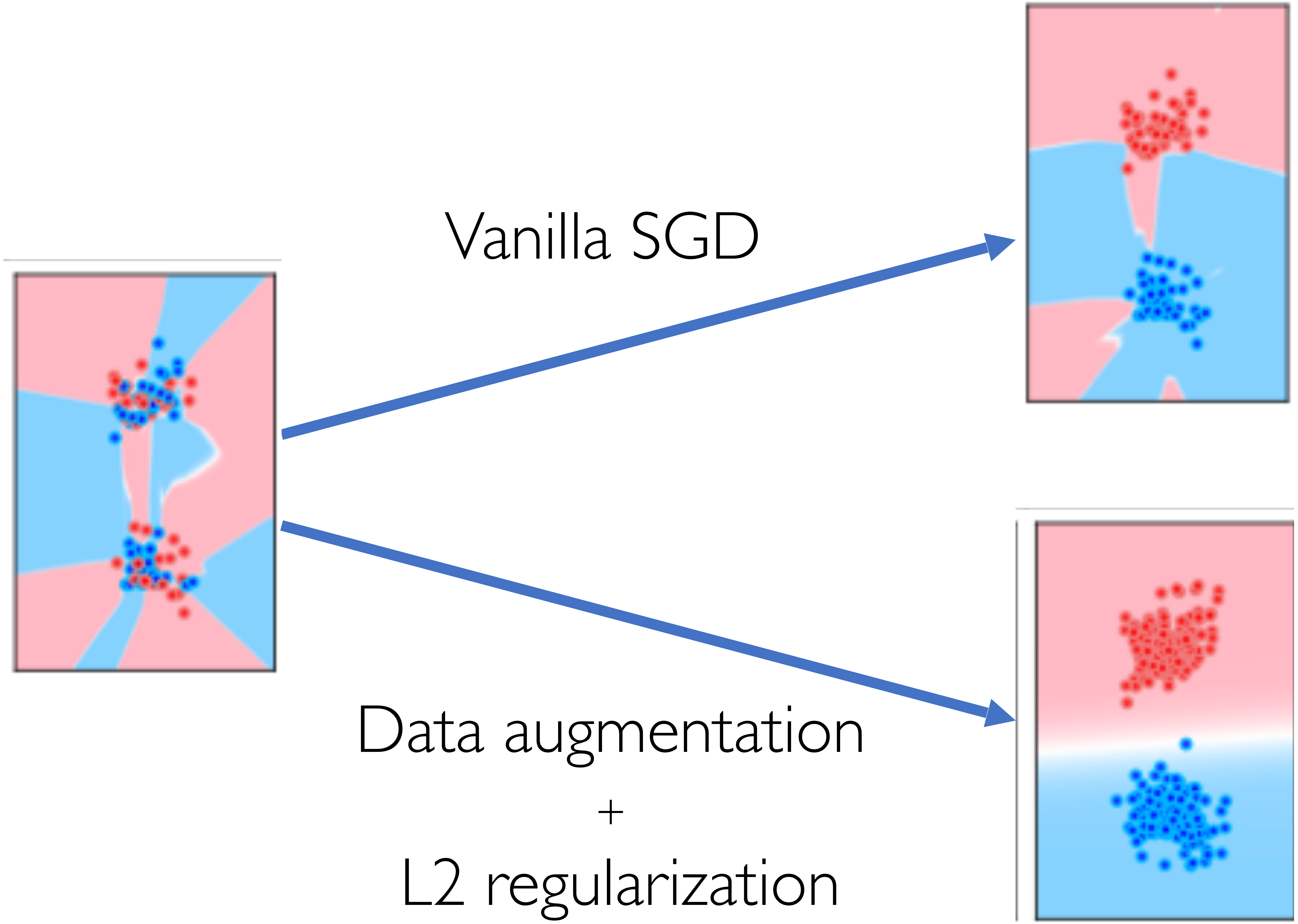
How Vanilla SGD gets in bad global minima



SGD "repairs" the boundary just enough to fit the data

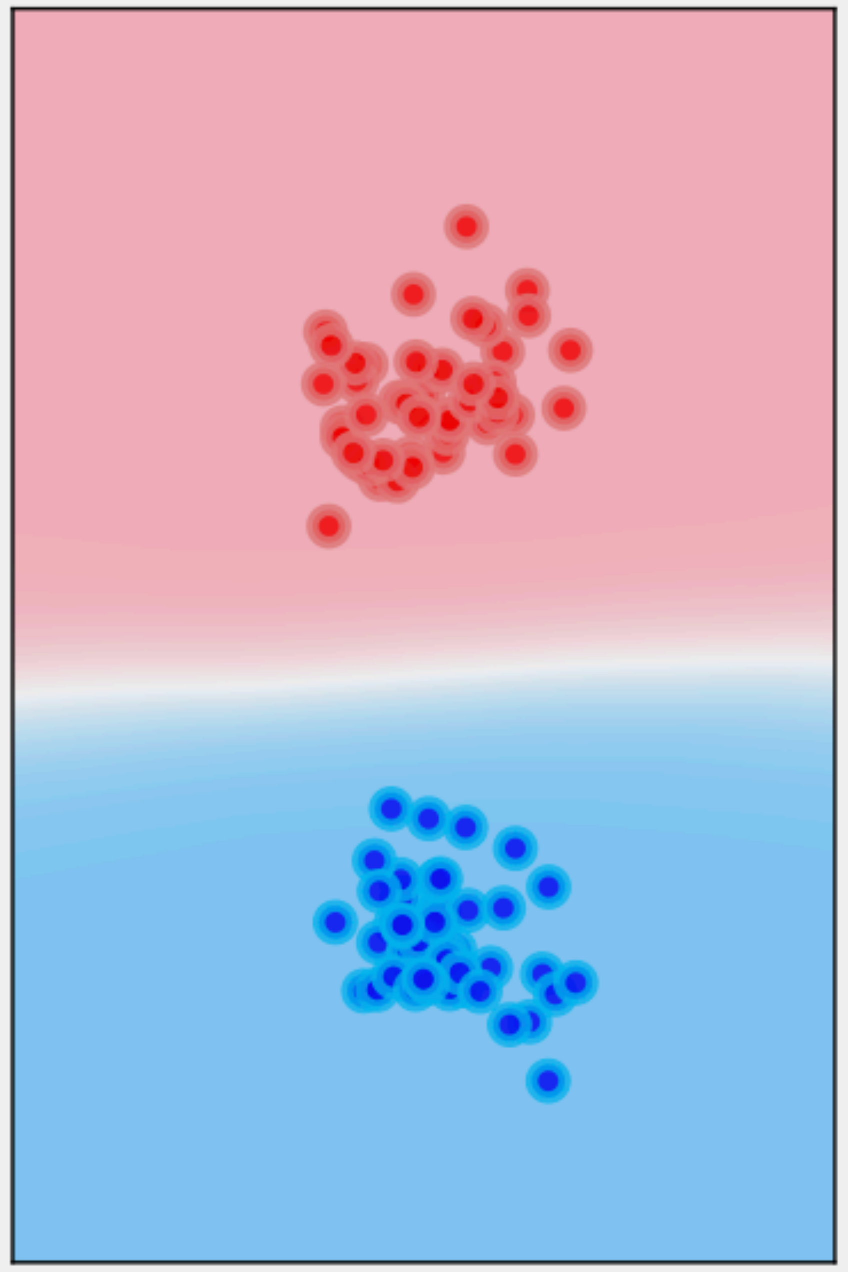
Can't "forget" the bad initialization

Regularization saves the day

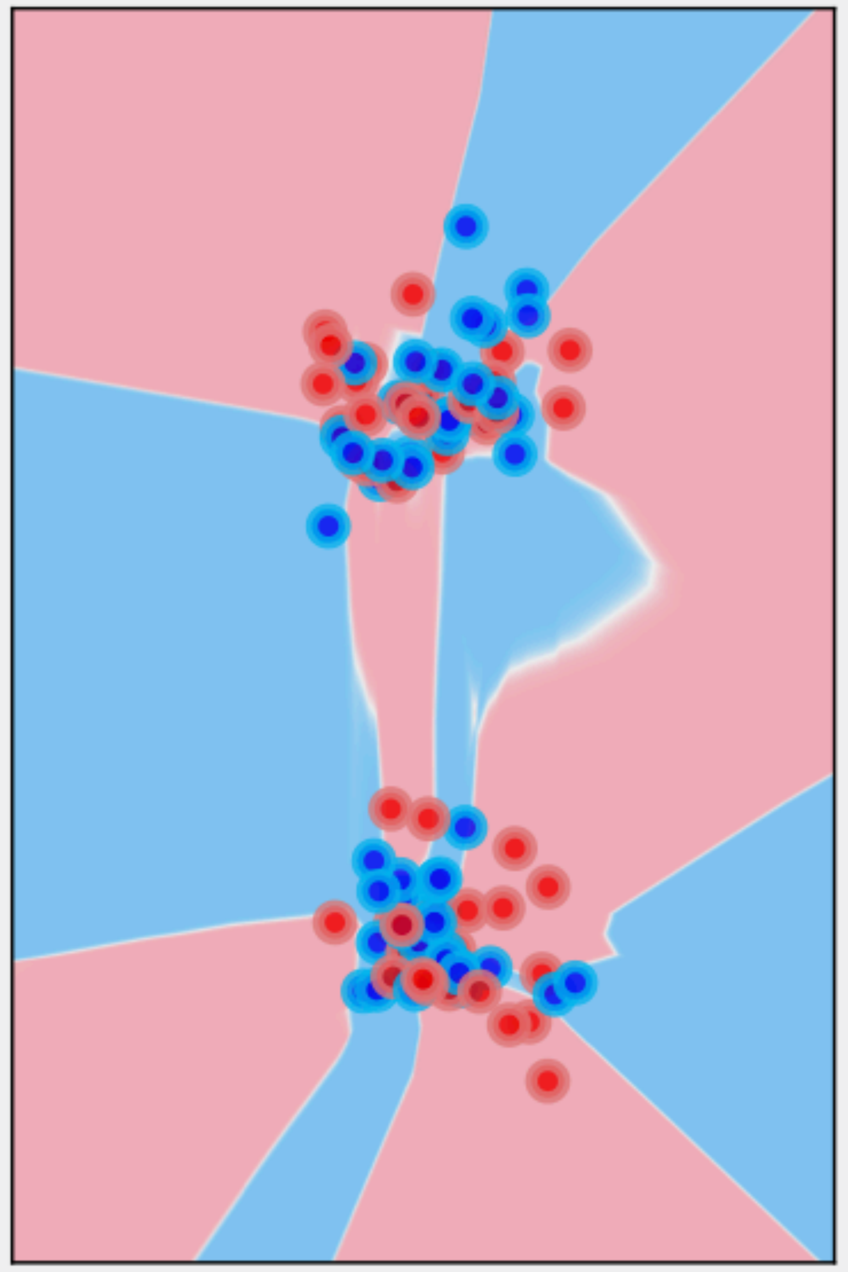


Regularization allows SGD to escape adversarial initializations

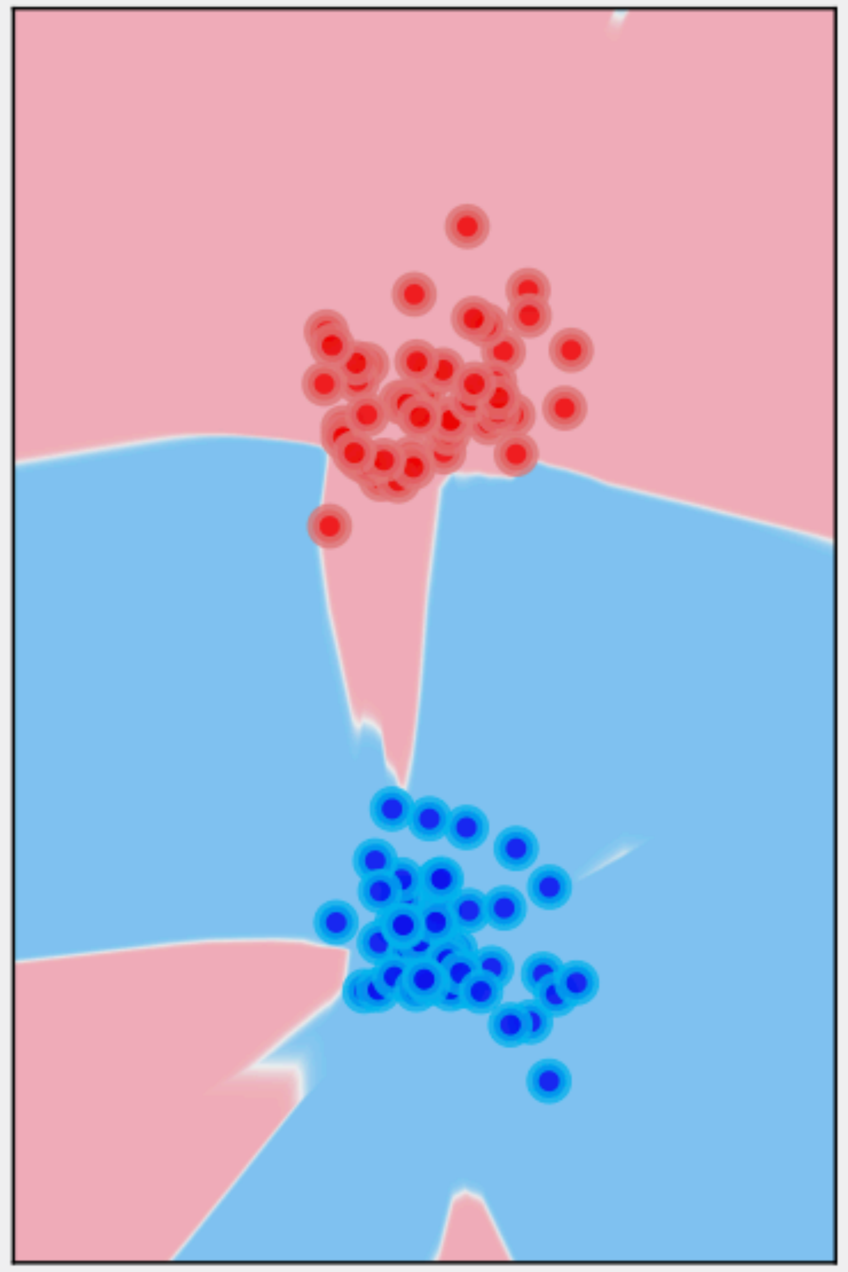
A recap of the setups



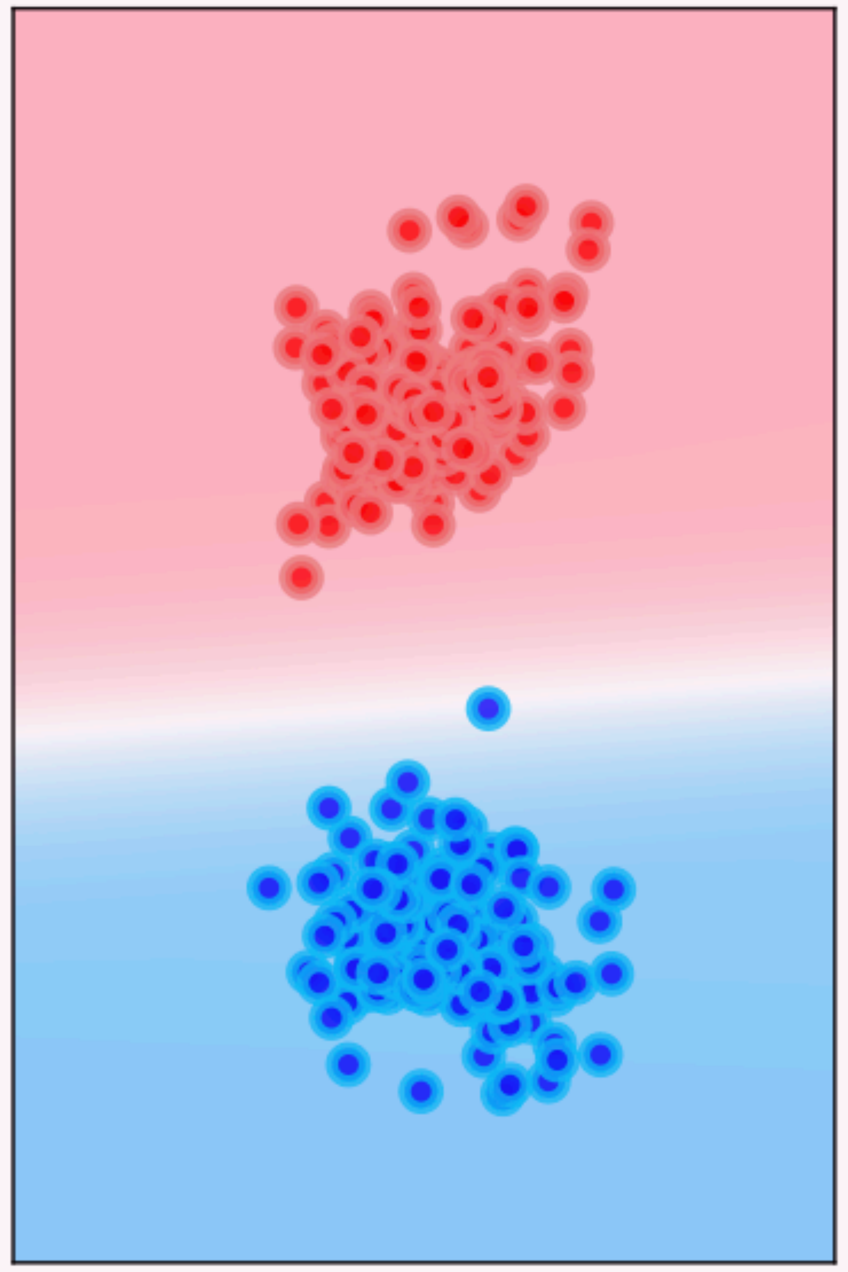
True labels
Random Init



Random labels
Random Init



True labels,
Adversarial Init



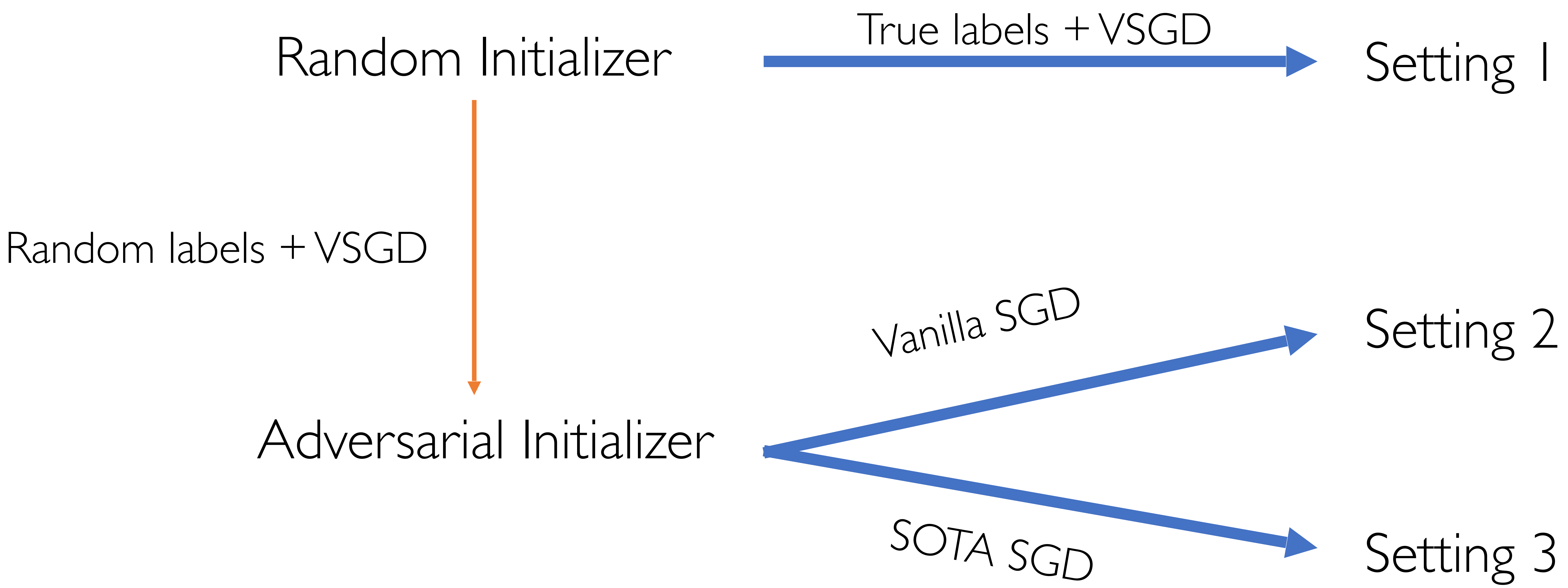
True labels labels,
Adversarial init

Vanilla SGD

SOTA SGD

Experiments

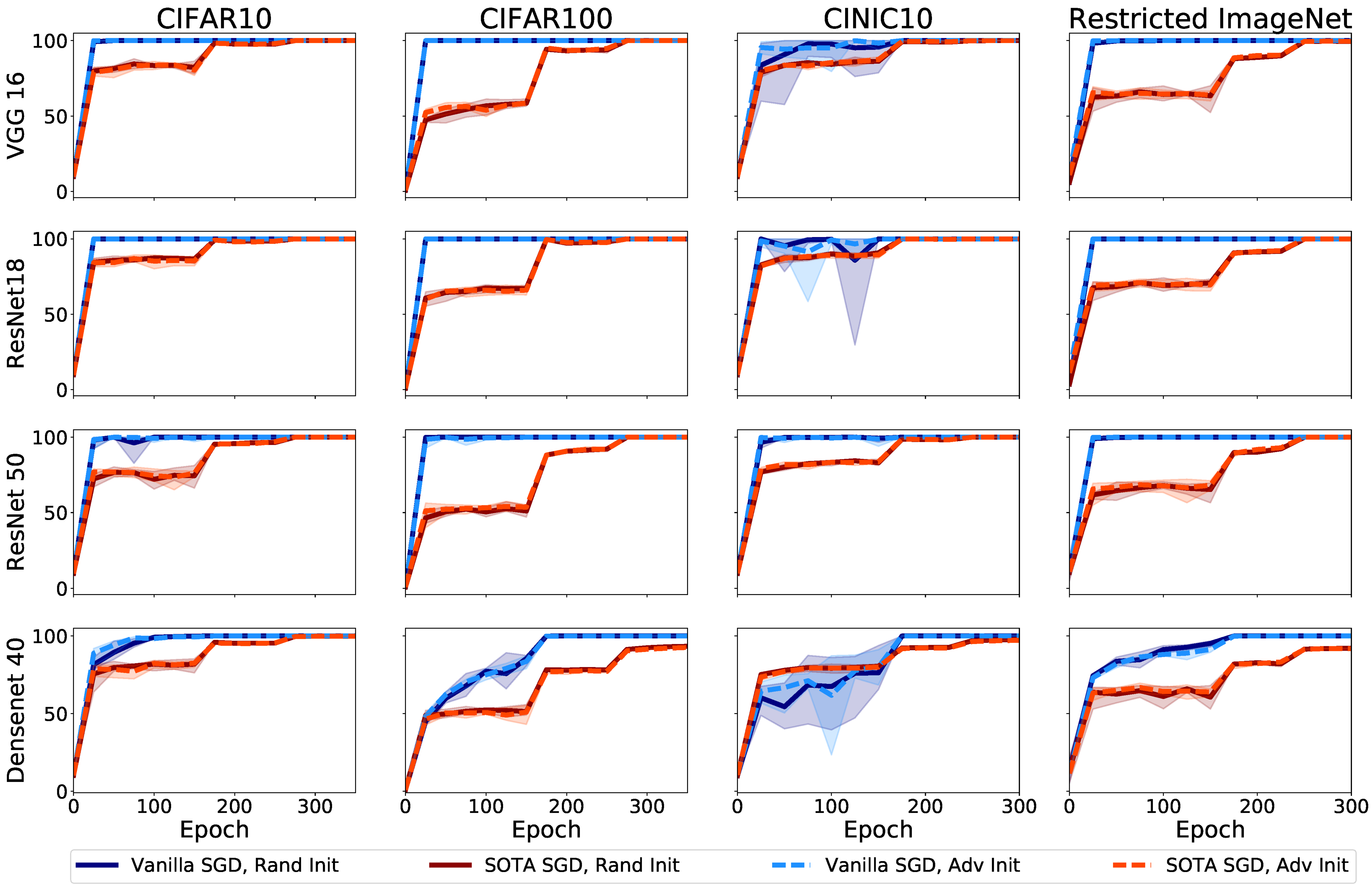
- **Data Sets**: Cifar10/100, CINIC10, Restricted Imagenet
- **Architectures:** VGG16, Resnet18/50, DenseNet40
- Hyperparameters tuned for faster convergence on train



Main Findings

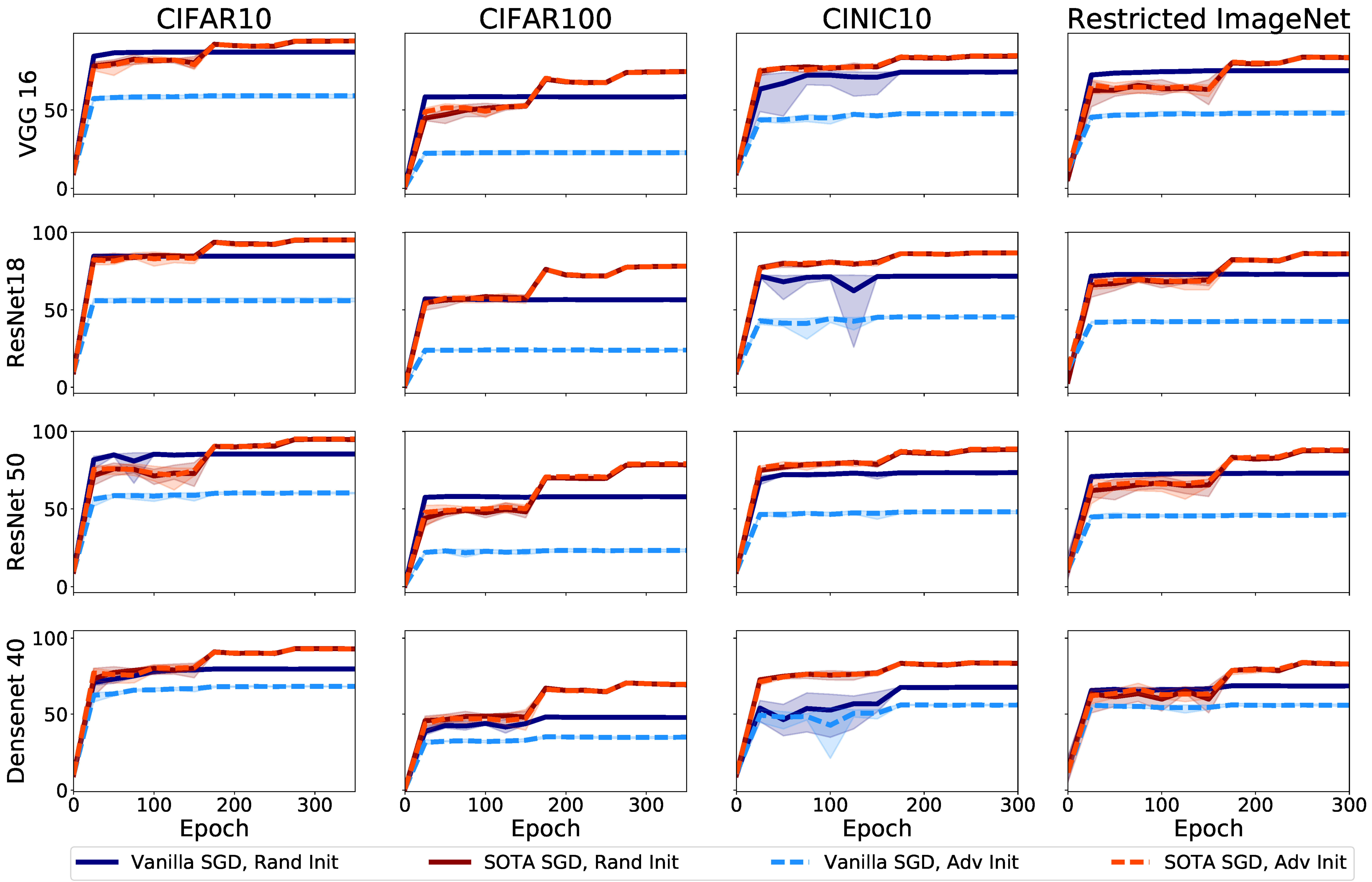
- Adversarial initialization causes VSGD up to 40% drop in test accuracy
- The model found is close to the adversarial initialization.
- Data augmentation, momentum, and L2 regularization all contribute to SGD escaping adversarial initialization.
- Any two of {DA, M, L2} are enough.

Train Accuracy



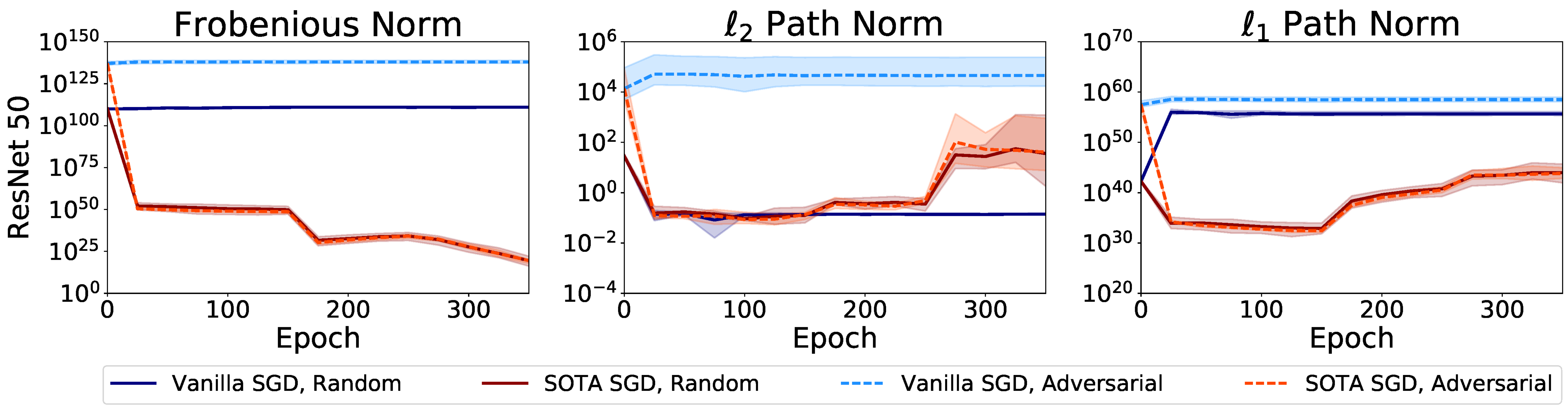
TL;DR:
Everything converges to 100% train accuracy

Test Accuracy



TL;DR:
Test error deteriorates for Vanilla SGD and Adv initialization

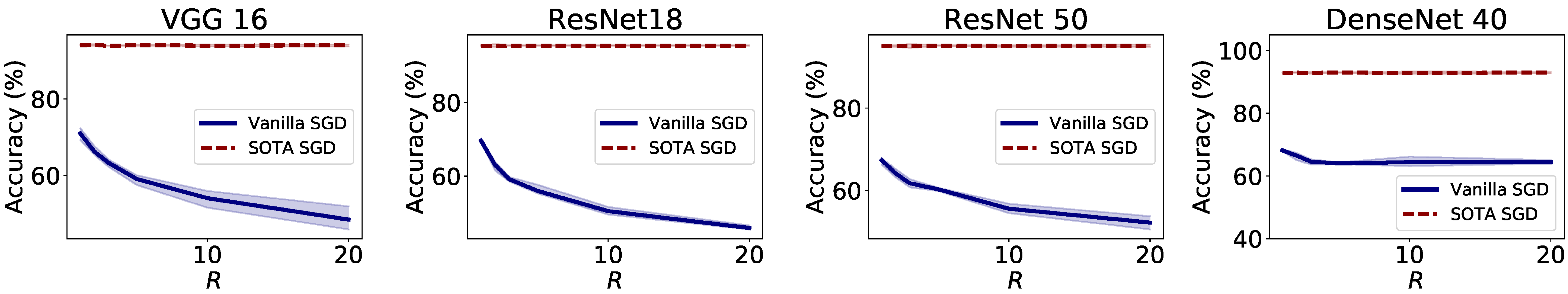
Model Complexity



ResNet 50 trained on CIFAR10.

TL;DR:
SGD on adv init has higher complexity measures compared to all other models

Effect of Replication Factor



TL;DR:
The more you augment the randomly labeled set,
the worse test error becomes

What is the point of all this

Implicit bias is likely weak in comparison to explicit regularization

Regularization affects the entire search dynamics, not just around global minima

The importance of regularization even very far away from the minima of the loss landscape.

Possible Explanations of Generalization

nope

- Maybe every model that fits the training data generalizes (no bad global minima)

Current implicit bias studies can't capture such a strong effect

- Maybe SGD is special "can avoid" bad global minima (implicit regularization)?
- Maybe the data distribution is what allows everything to fall into place?

Possible Explanations of Generalization

nope

- Maybe every model that fits the training data generalizes (no bad global minima)

Current implicit bias studies can't capture such a strong effect

- Maybe SGD is special "can avoid" bad global minima (implicit regularization)?

- Maybe the data distribution is what allows everything to fall into place?

Nobody knows

reading list

Soudry, D., Hoffer, E., Nacson, M.S., Gunasekar, S. and Srebro, N., 2018. The implicit bias of gradient descent on separable data. The Journal of Machine Learning Research, 19(1), pp.2822-2878.
Vancouver
https://www.jmlr.org/papers/volume19/18-188/18-188.pdf

Gunasekar, S., Lee, J., Soudry, D. and Srebro, N., 2018, July. Characterizing implicit bias in terms of optimization geometry. In International Conference on Machine Learning (pp. 1832-1841). PMLR.
Vancouver
http://proceedings.mlr.press/v80/gunasekar18a/gunasekar18a.pdf

Neyshabur, B., Tomioka, R. and Srebro, N., 2014. In search of the real inductive bias: On the role of implicit regularization in deep learning. arXiv preprint arXiv:1412.6614.
Vancouver
https://arxiv.org/pdf/1412.6614.pdf

Vardi, G. and Shamir, O., 2021, July. Implicit regularization in relu networks with the square loss. In Conference on Learning Theory (pp. 4224-4258). PMLR.
http://proceedings.mlr.press/v134/vardi21b/vardi21b.pdf

Liu, S., Papailiopoulos, D. and Achlioptas, D., 2020. Bad global minima exist and sgd can reach them. Advances in Neural Information Processing Systems, 33, pp.8543-8552.
Vancouver
https://arxiv.org/abs/1906.02613