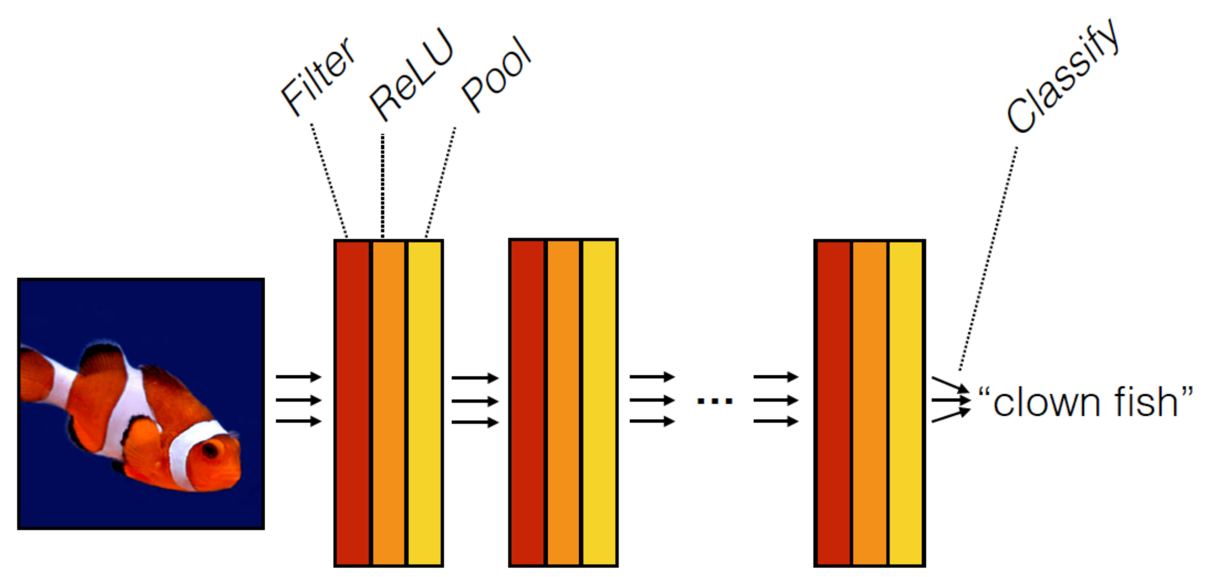
Class 26:
Theory for Deep Nets and future breakthroughs

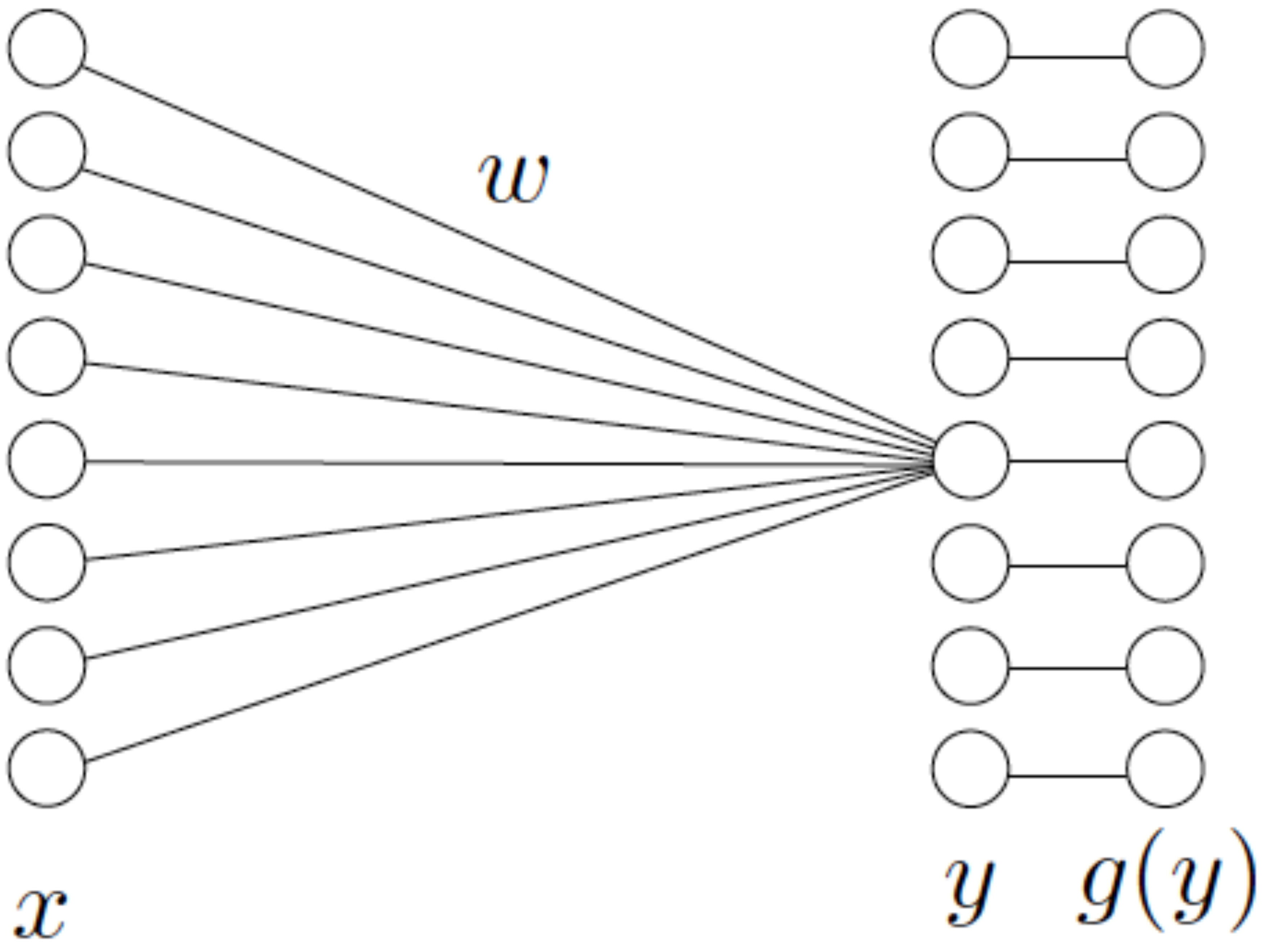
Tomaso Poggio,
9.520/6.860

Training and computation in a neural net

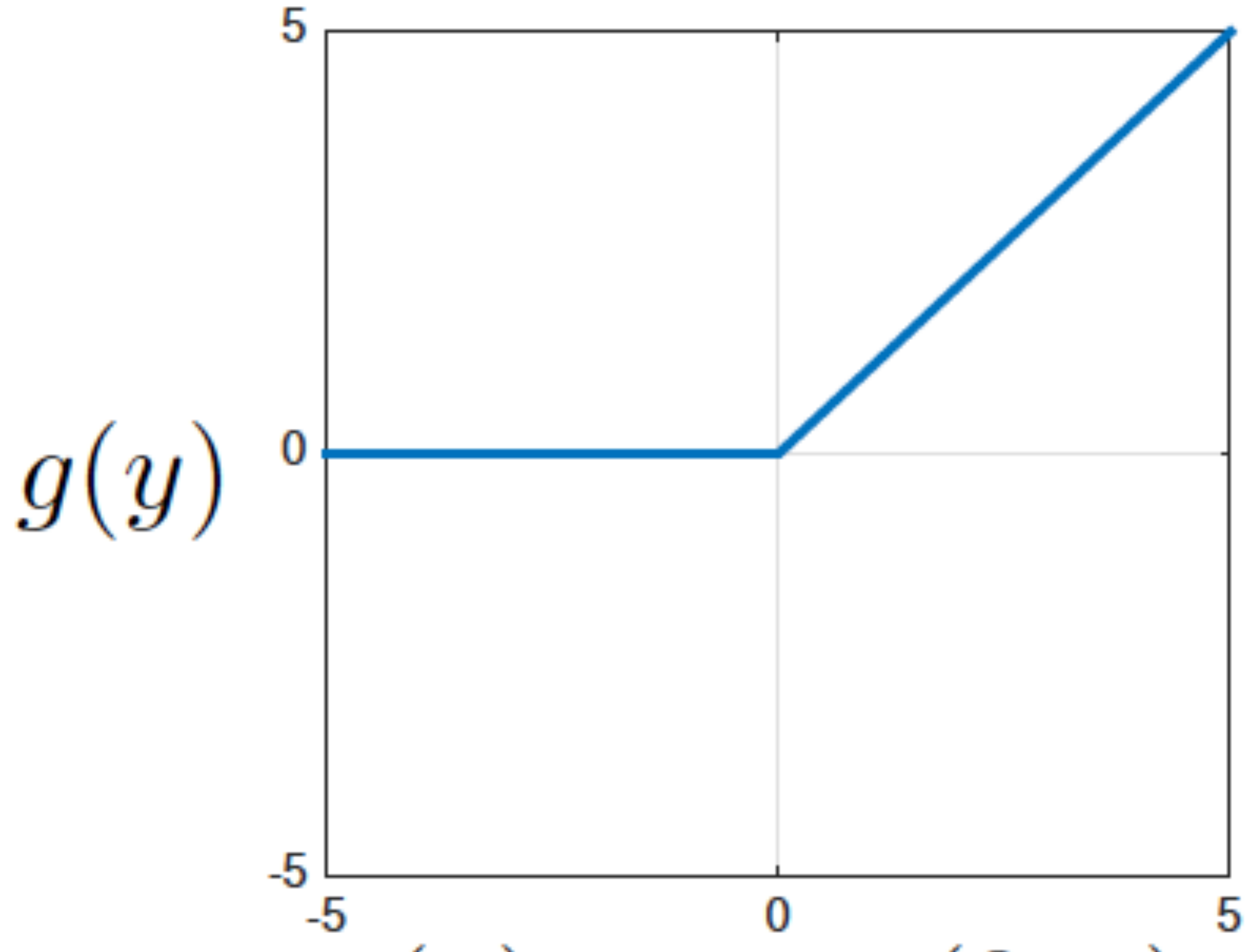
Computation in a neural net



$f(\mathbf{x}) = f_L(\dots f_2(f_1(\mathbf{x})))$



Rectified linear unit (ReLU)



$$g(y) = \max(0, y)$$

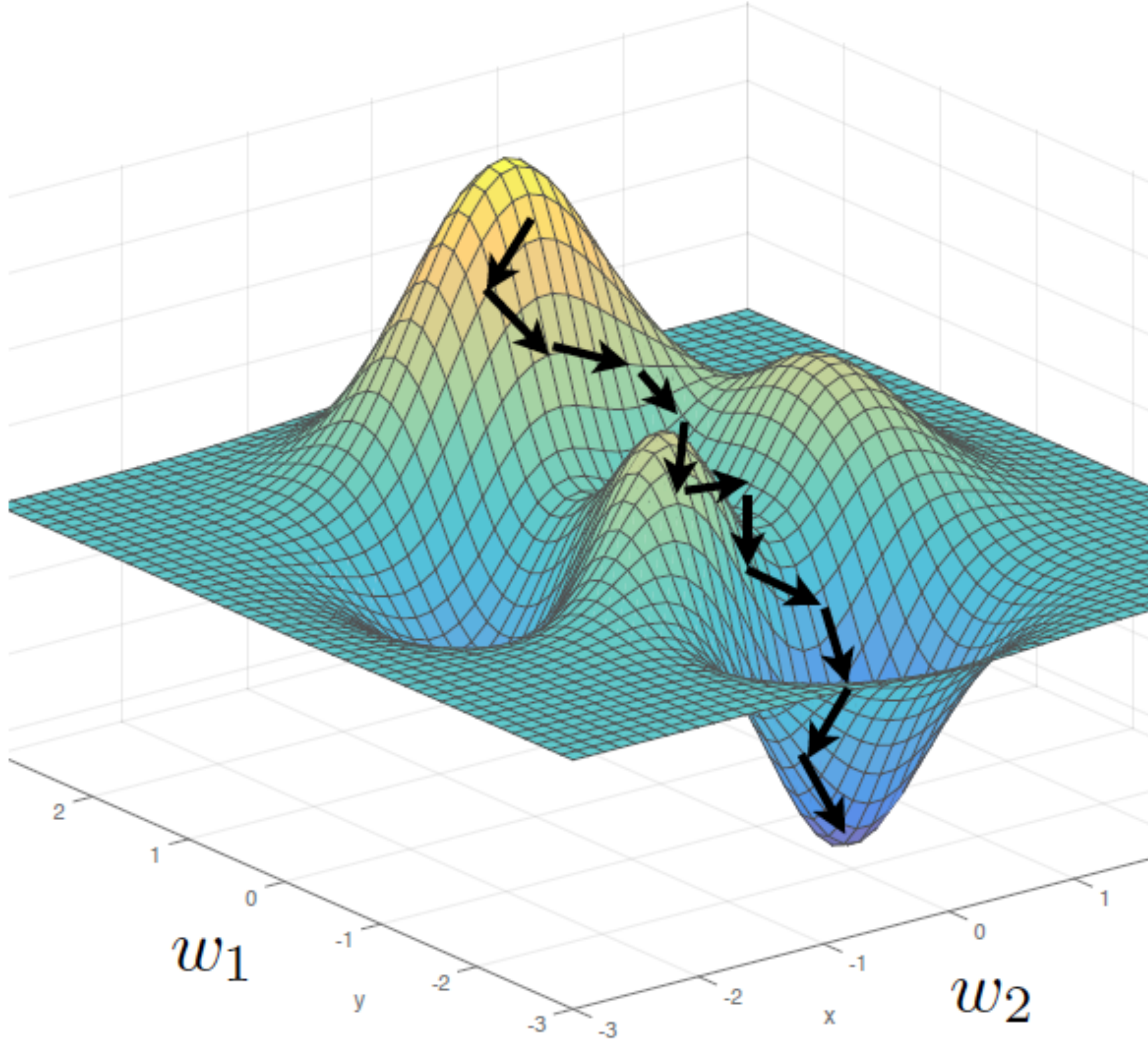
Gradient descent

$$\boxed{\operatorname*{argmin}_{\mathbf{w}} \quad \sum_i \ell(\mathbf{z}_i, f(\mathbf{x}_i; \mathbf{w})) = L(\mathbf{w})}$$

One iteration of gradient descent:

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta_t \frac{\partial L(\mathbf{w^t})}{\partial \mathbf{w}}$$

learning rate



Deep Networks: three theory questions

- *Approximation Theory:* When and why are deep networks better than shallow networks?

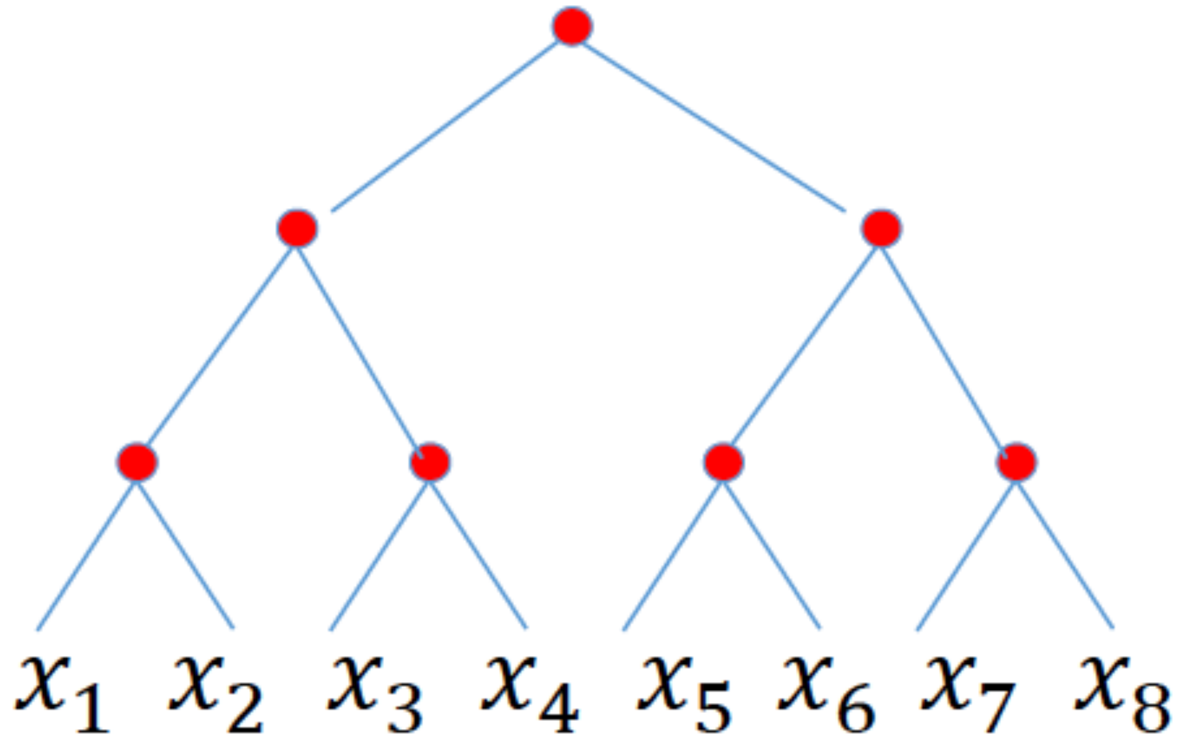
- *Optimization:* What is the landscape of the empirical risk?

- *Dynamical System Approach:* Characterizing SGD solutions. Are they stable? Control of norm? Maximum margin?

- L*earning Theory:* How can deep learning not overfit?

A new way to avoid the curse for compositional functions: *deep networks*

$$f(x_1,x_2,...,x_8)=g_3(g_{21}(g_{11}(x_1,x_2),g_{12}(x_3,x_4))g_{22}(g_{11}(x_5,x_6),g_{12}(x_7,x_8)))$$



Theorem, Mhaskar, Poggio, Liao 2016 (informal statement)

Suppose that a function of d variables is hierarchically locally compositional . Both shallow and deep network can approximate f equally well. The number of parameters of the shallow network depends exponentially on d as $O(\varepsilon^{-d})$ whereas the deep networks show $O(d\varepsilon^{-2})$ vs $O(\varepsilon^{-d})$

Note: *Locality, not weight sharing,* avoids the curse of dimensionality

Deep Networks: three theory questions

- *Approximation Theory:* When and why are deep networks better than shallow networks?

- *Optimization:* What is the landscape of the empirical risk?

- *Dynamics of learning:* What are the solutions? Are they stable? Maximum margin?

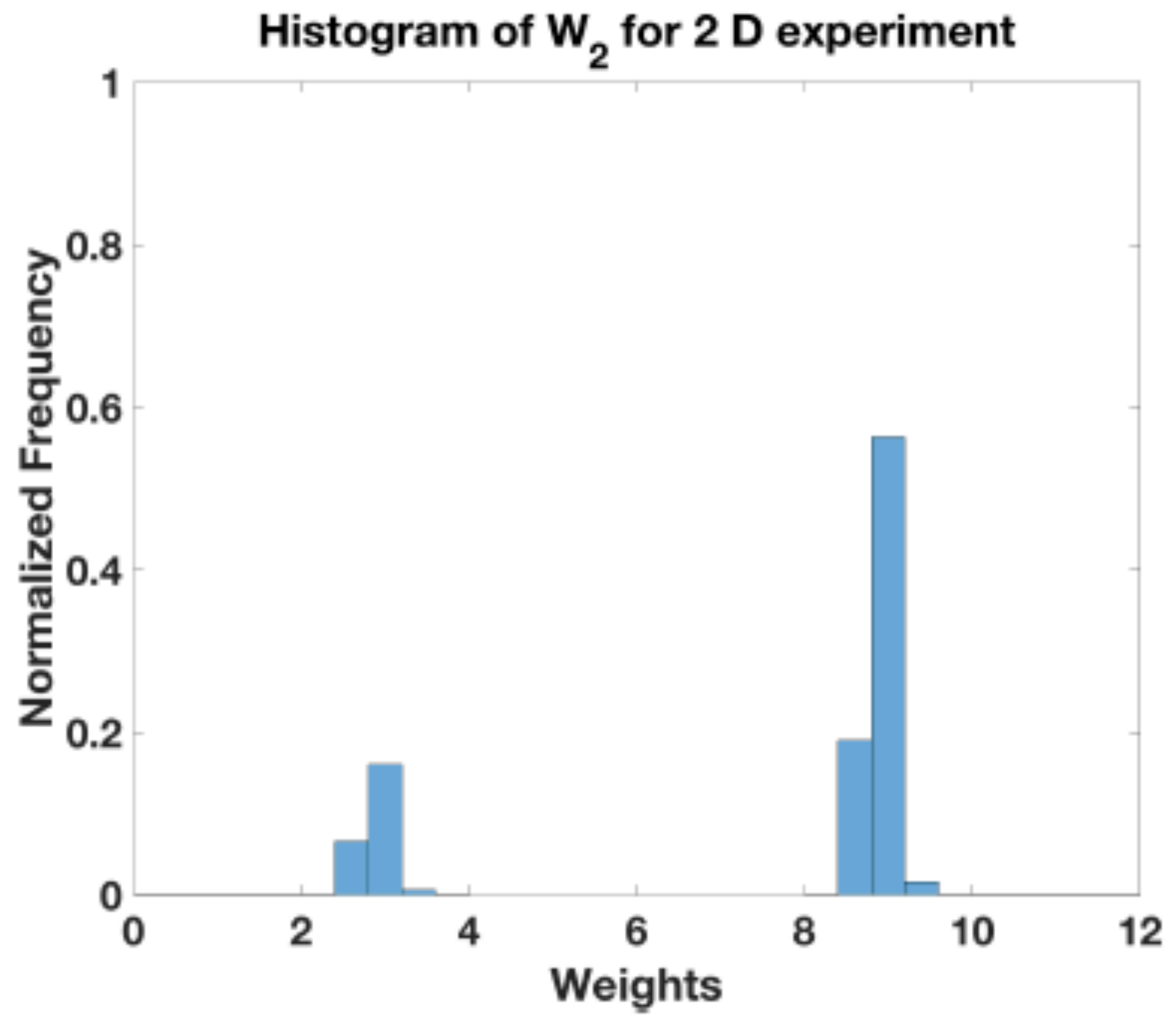
- L*earning Theory:* How can deep learning not overfit?

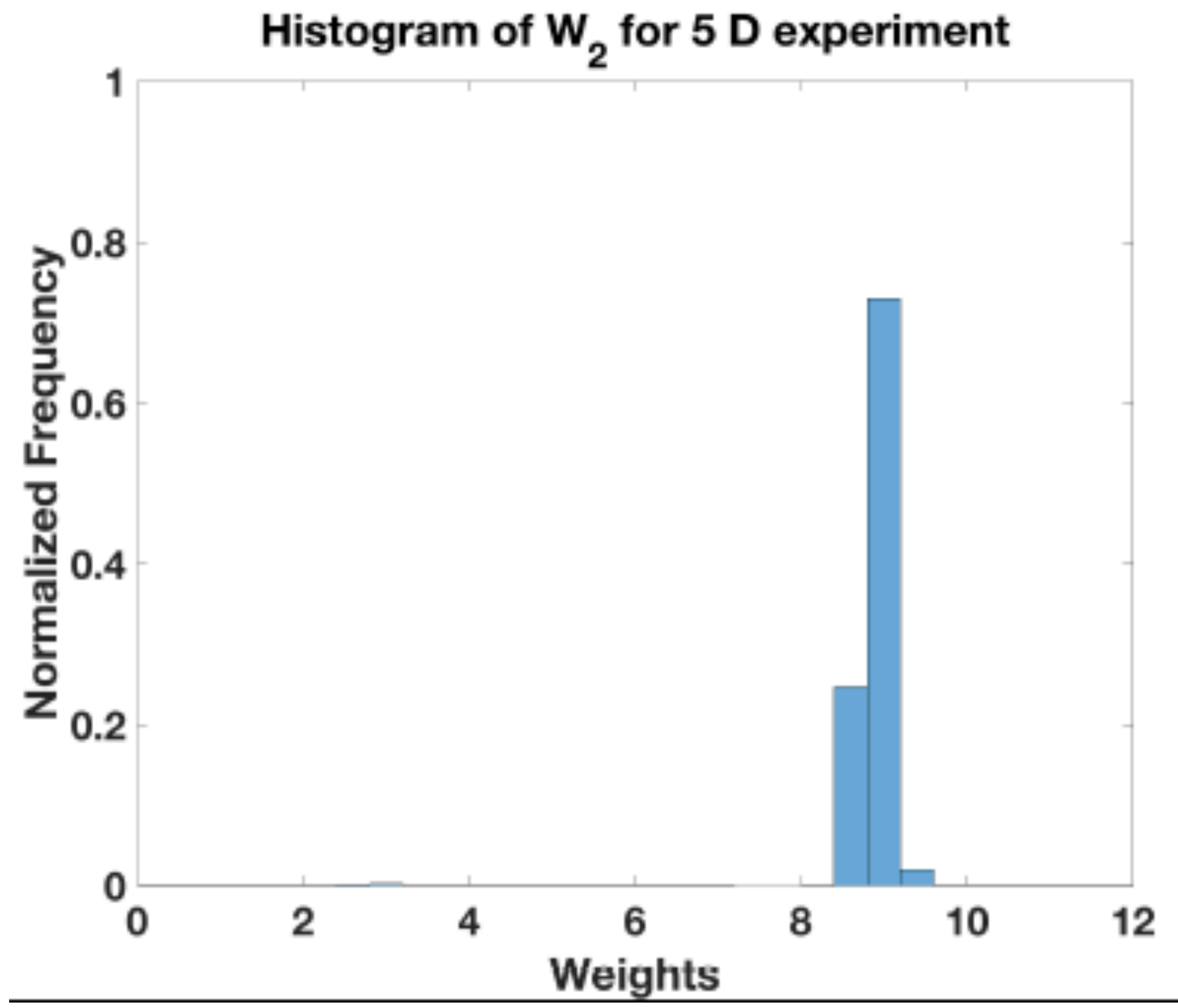
SGDL and SGD observation: summary

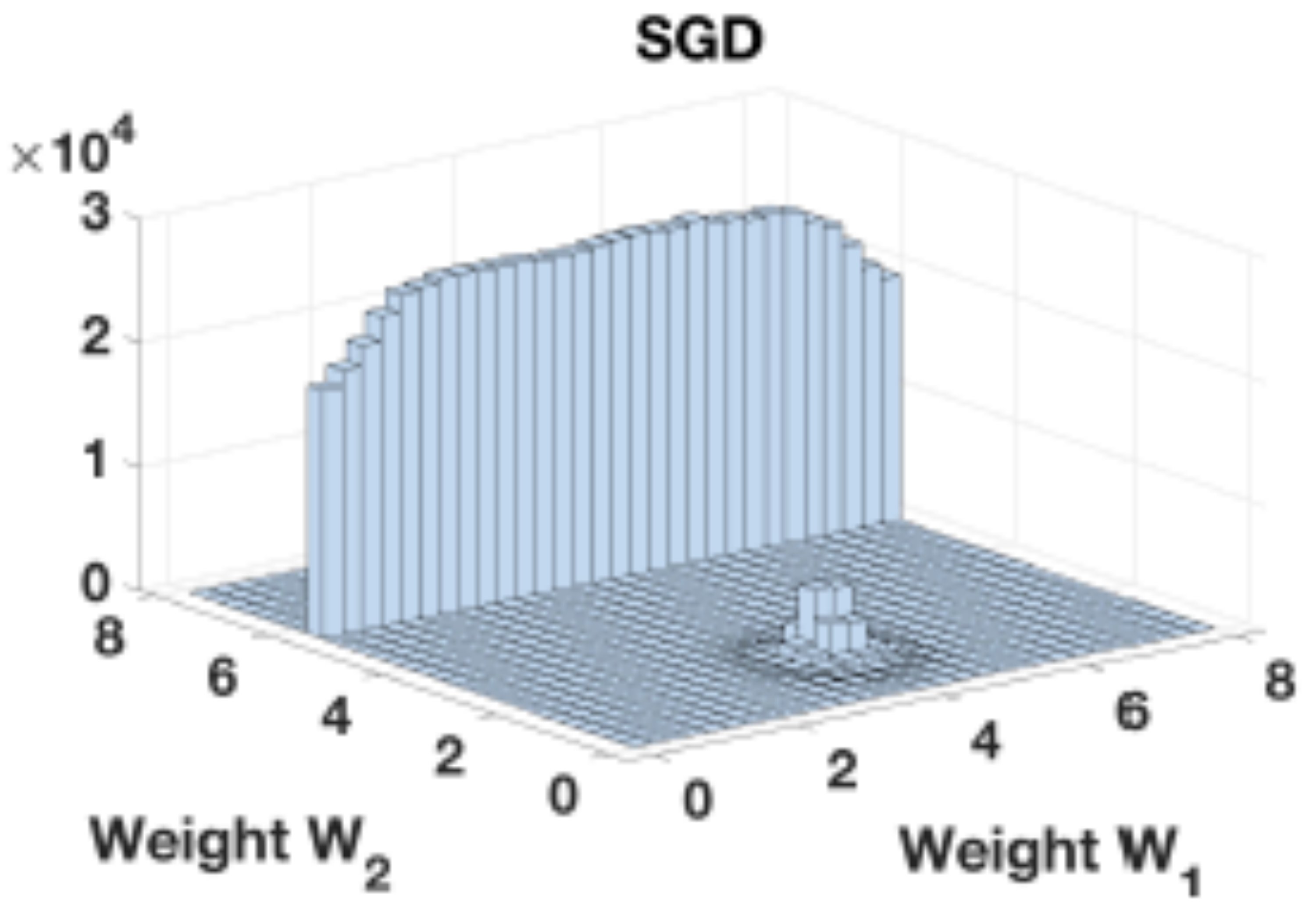
- There are may zero minimizer with overparametrized deep networks because of Bezout theorem

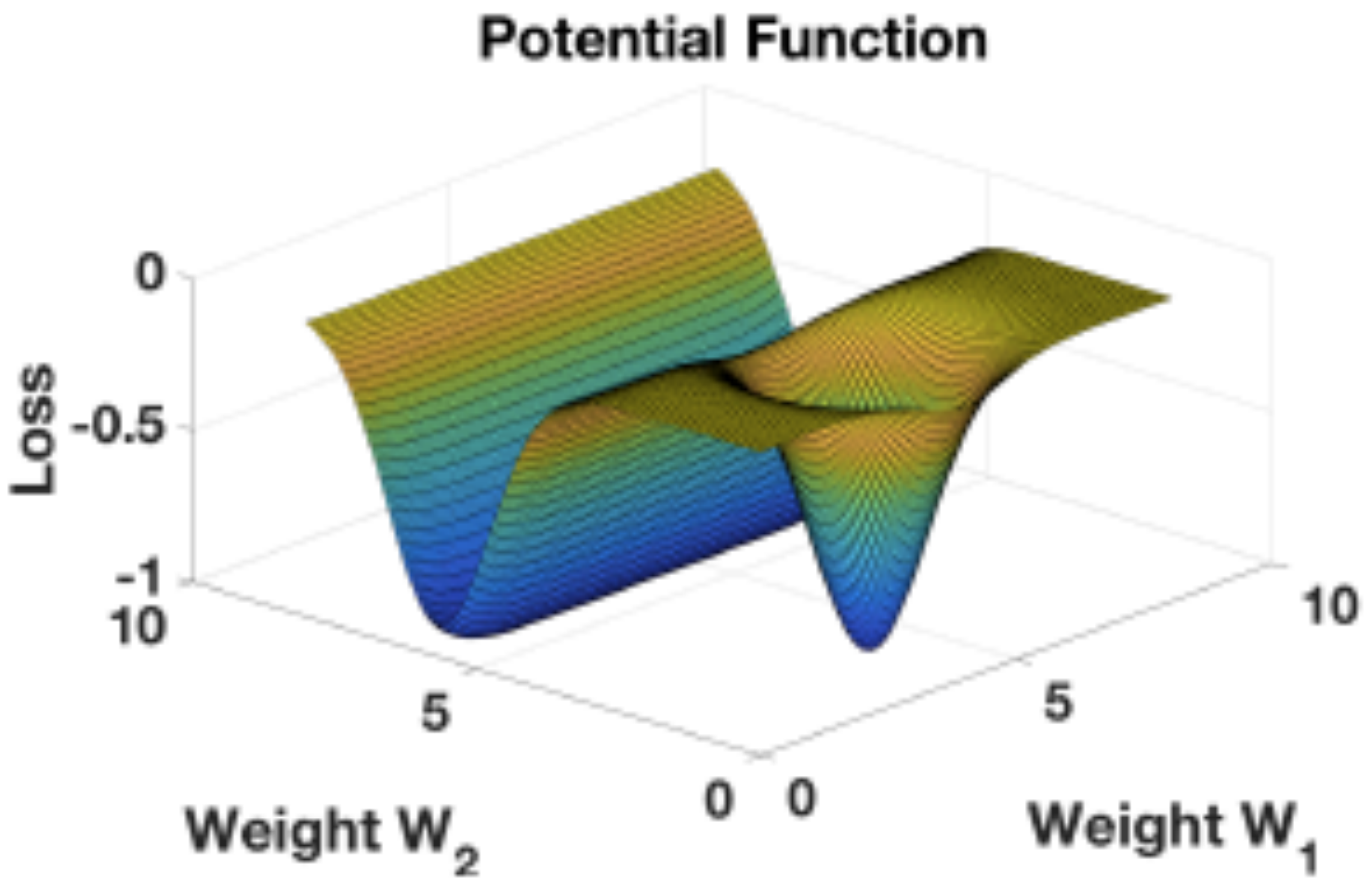
- SGDL finds with very high probability large volume, flat **zero-minimizers**; empirically SGD behaves in a similar way

- Flat minimizers correspond to degenerate zero-minimizers and thus to global minimizers;









Poggio, Rakhlin, Golovitc, Zhang, Liao, 2017

Deep Networks: three theory questions

- *Approximation Theory:* When and why are deep networks better than shallow networks?

- *Optimization:* What is the landscape of the empirical risk?

- *Dynamical System Approach:* Characterizing SGD solutions. Are they stable? Control of norm? Maximum margin?

- *Learning Theory:* How can deep learning not overfit?

Dynamical Systems approach:

degenerate minima in DNNs
represent the main obstacle in extending analysis
from linear to nonlinear networks

What happens in the simplest case, that is linear one layer networks?

Degenerate Hessian, no norm control!

Regularization is needed….

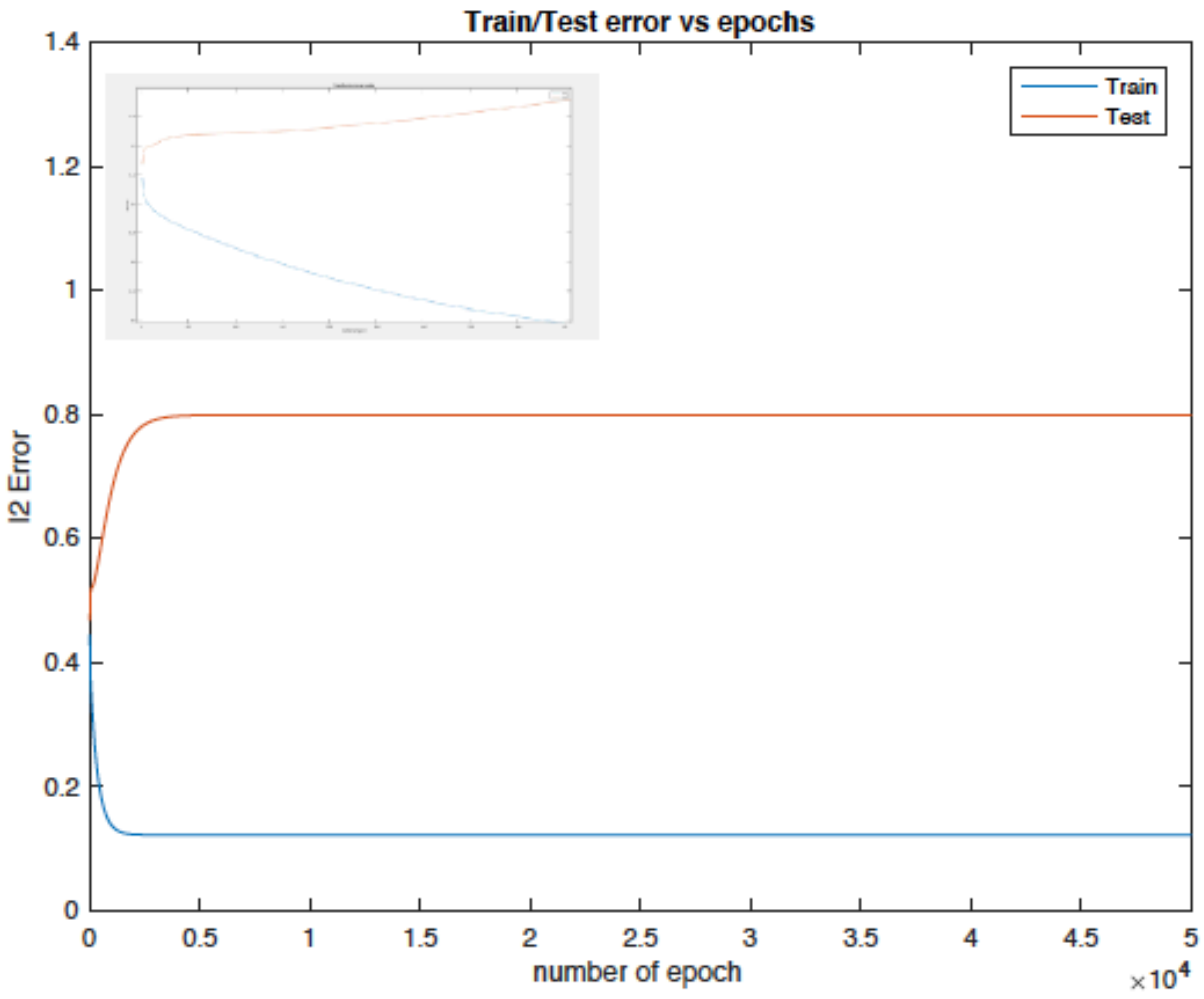
Explicit regularization is OK in the degenerate linear one-layer case under square loss

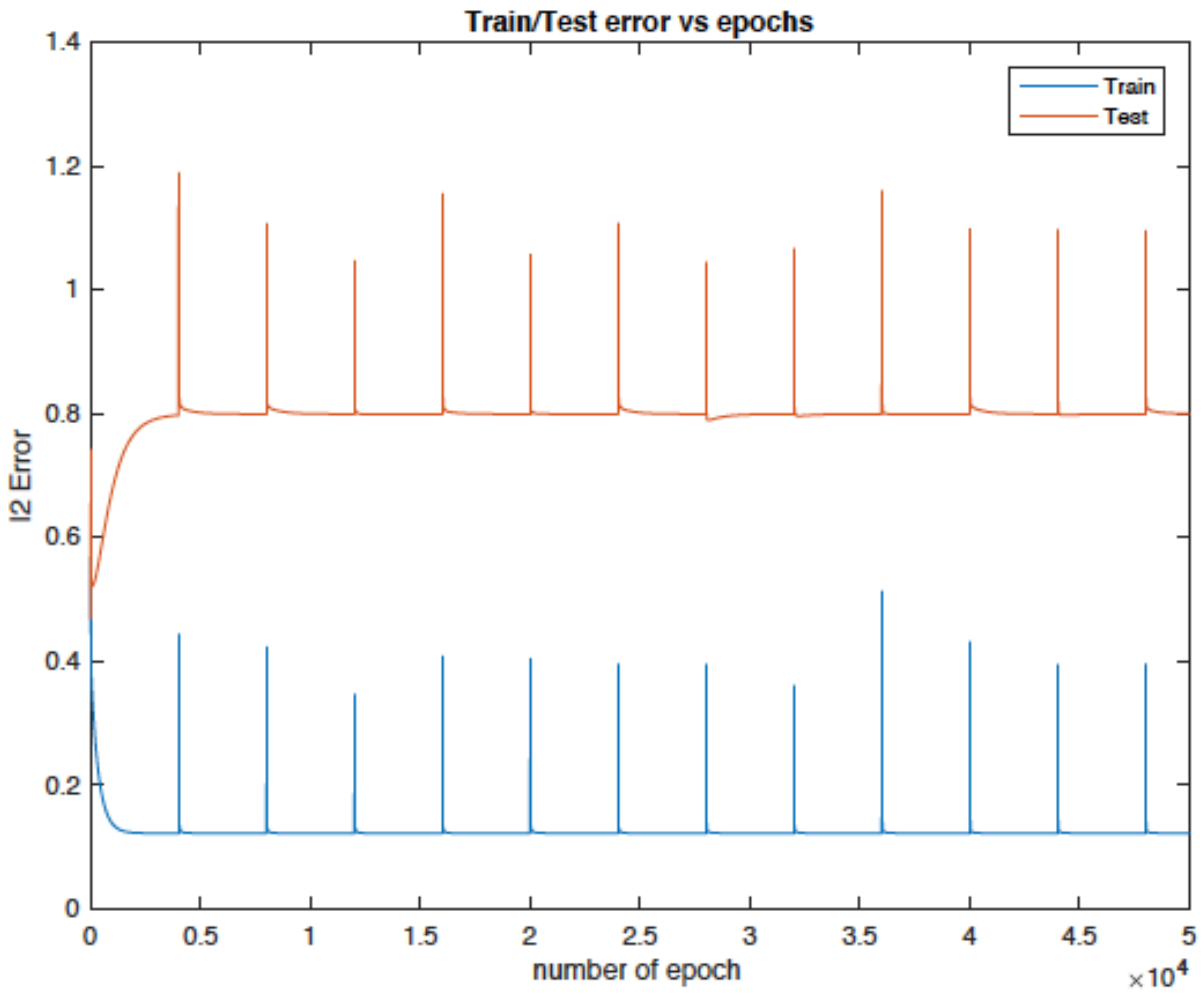
$$L = \sum_{n}^{N} (y_n - w_1{}^{1,j} x^j{}_n))^2 + \lambda \| w_1 \|^2$$

$$\dot{w}_1{}^{1,j} = 2 \sum_{n}^{N} E_n x^j{}_n + \lambda w^{1,j}{}_1$$

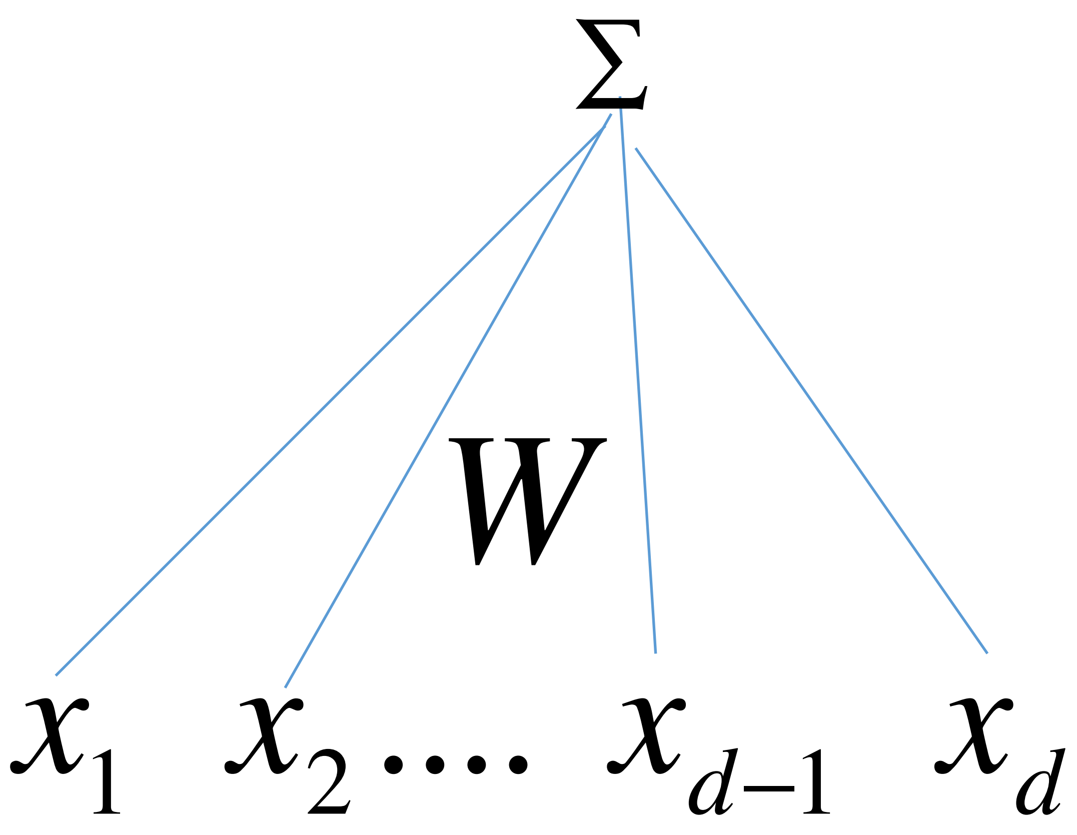
$H_{i,j=} \sum_{n}^{N} x_n^i x_n^j + \lambda I$ is positive definite for arbitrarily small λ!

Hyperbolic minimum, non degenerate Hessian, does not depend on initial conditions/perturbations





Even just implicit regularization by GD+SGD “works” in the degenerate linear one-layer case under square loss



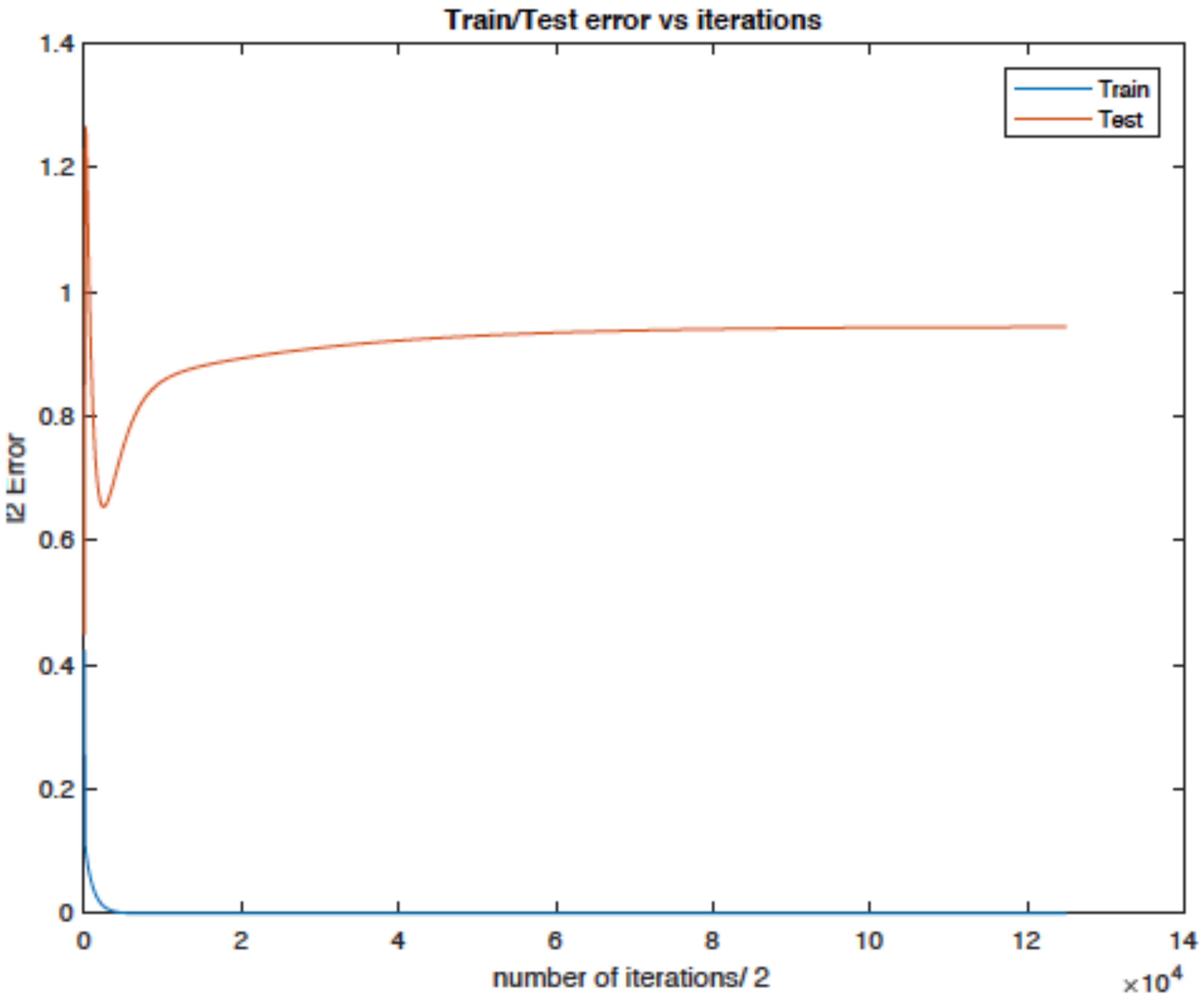
$$W = YX^{\dagger}$$

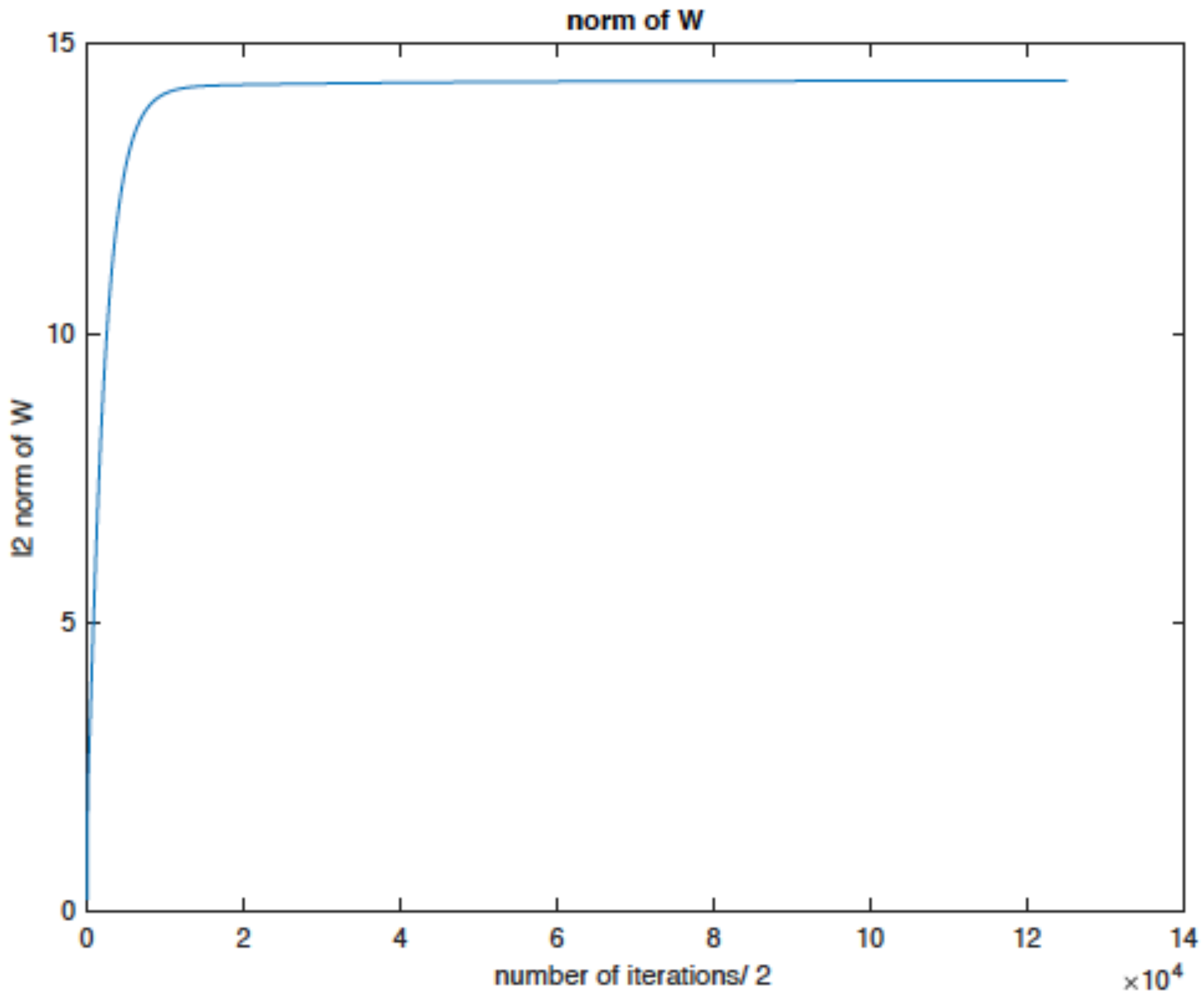
Corollary 1. *When initialized with zero, both GD and SGD converges to the minimum-norm solution.*

Min norm solution is the limit for $\lambda \to 0$ *of regularized solution*

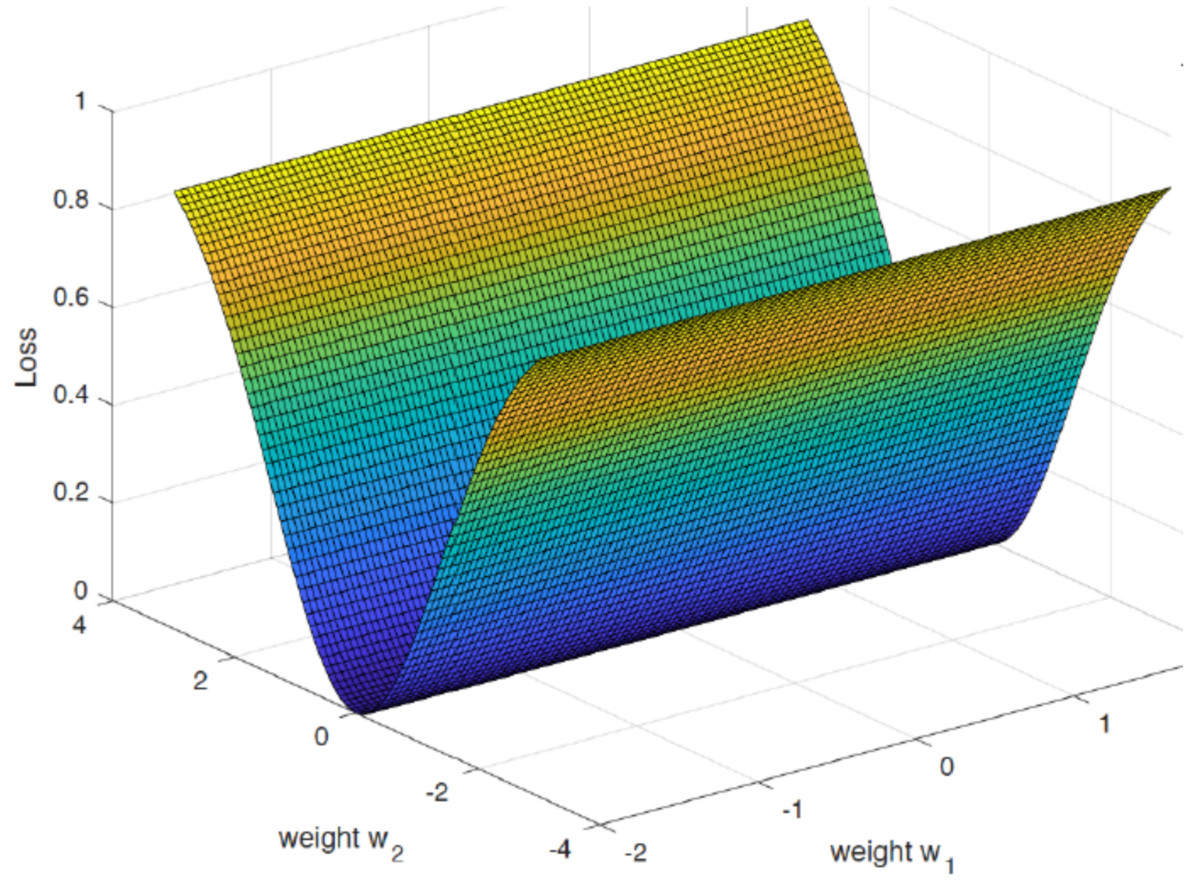
CENTER FOR
Brains
Minds+
Machines

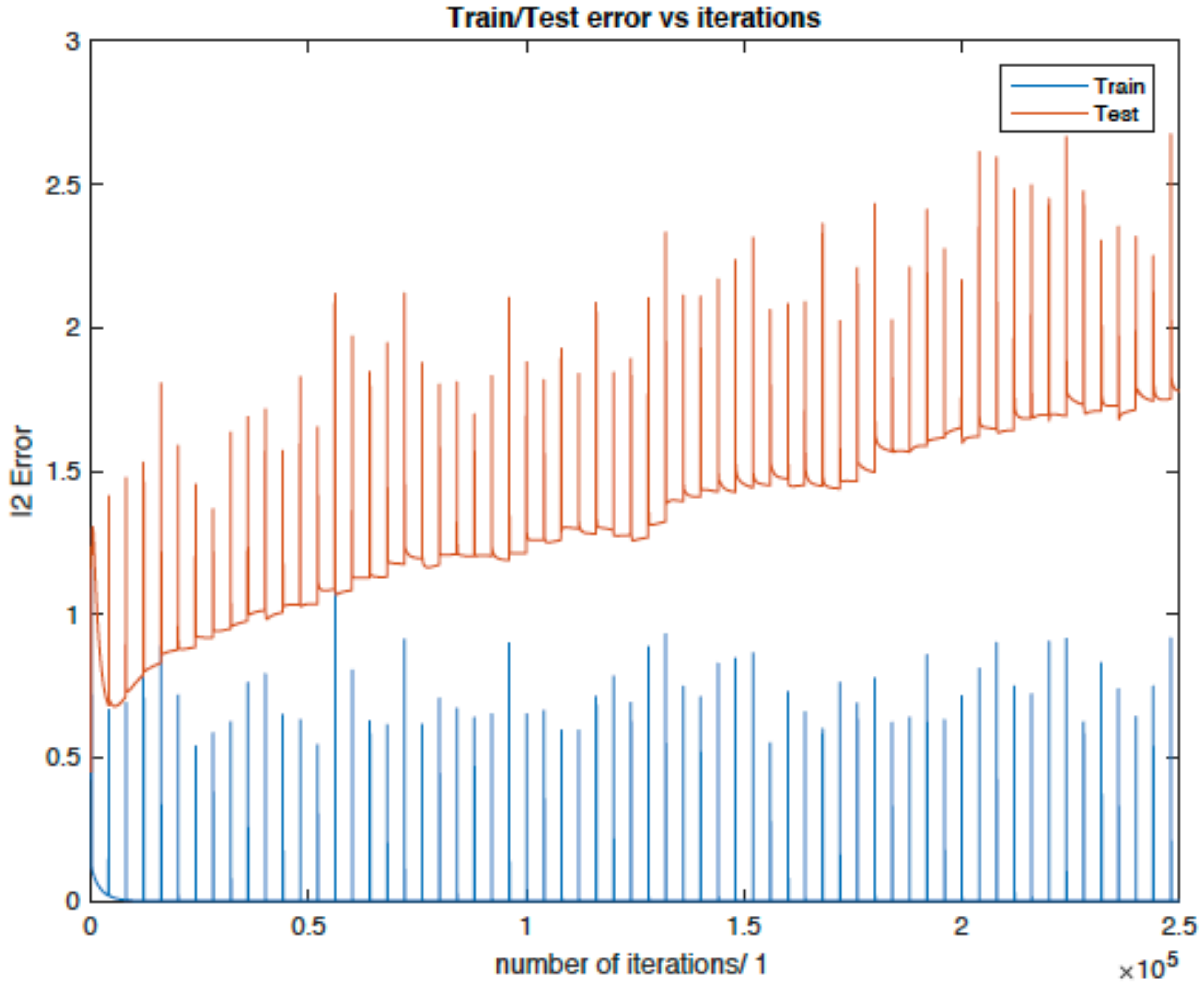
Degenerate Hessian

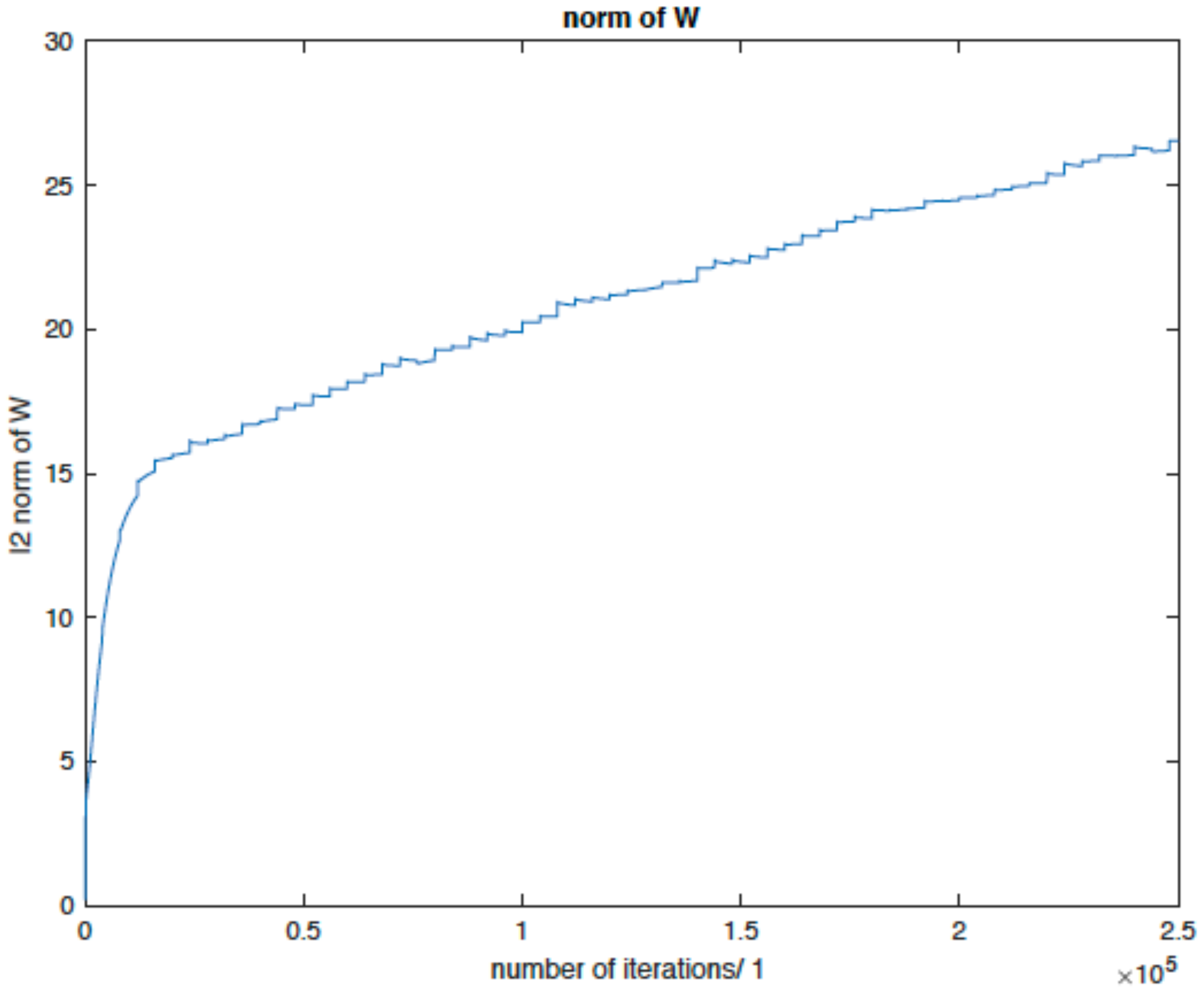




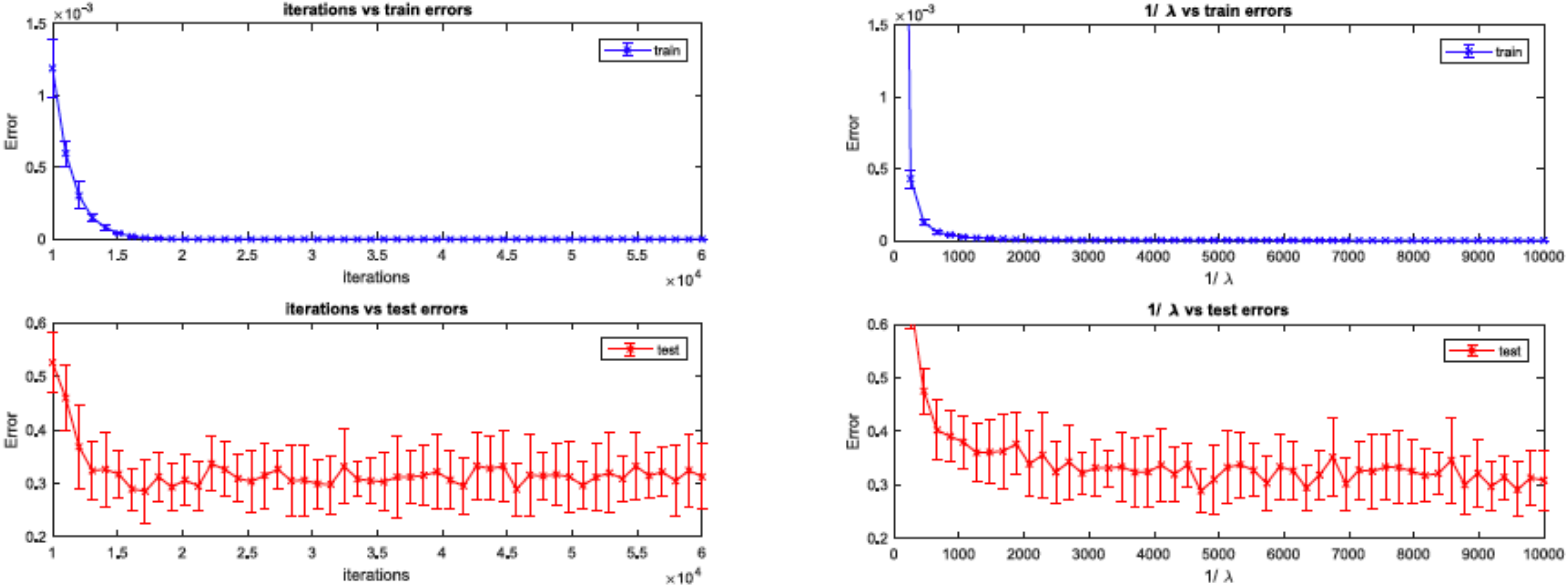
Degenerate Hessian, depends on initial conditions/perturbations which affect minimum norm property and may decrease test performance





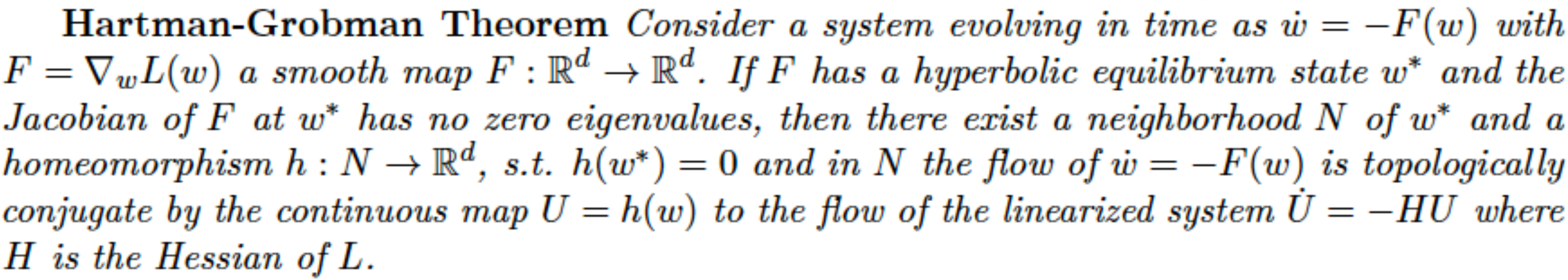


Deep linear network: GD as implicit regularizer



GD regularizes implicitly deep linear networks as it does for linear networks

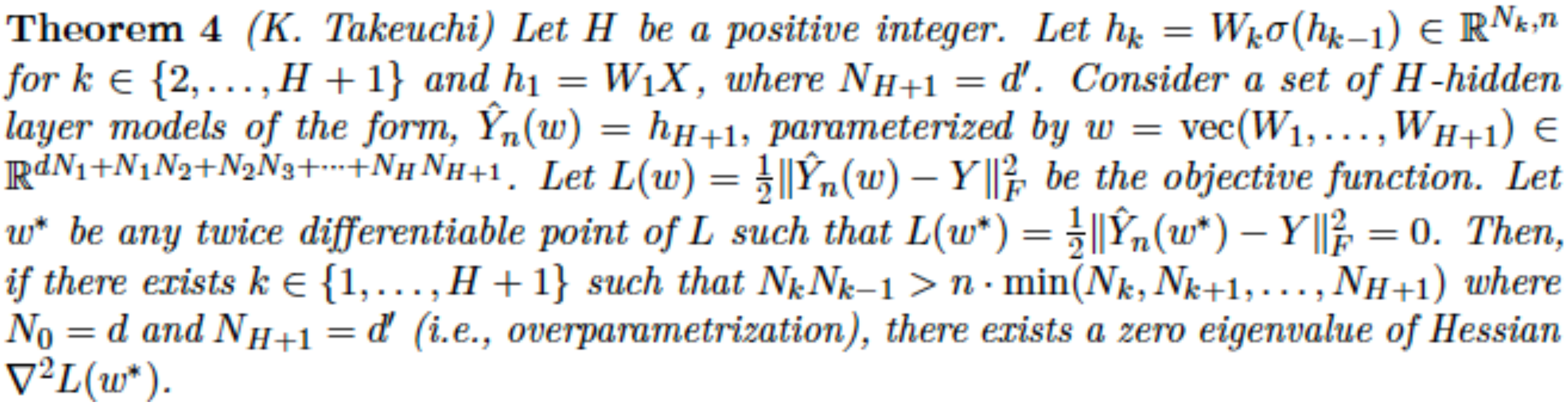
Linearization yields dynamics of nonlinear RELU network around a minimum if Hessian is non degenerate



Hartman-Grobman Theorem *Consider a system evolving in time as* $\dot{w} = -F(w)$ *with* $F = \nabla_w L(w)$ *a smooth map* $F : \mathbb{R}^d \to \mathbb{R}^d$. *If* F *has a hyperbolic equilibrium state* w^* *and the Jacobian of* F *at* w^* *has no zero eigenvalues, then there exist a neighborhood* N *of* w^* *and a homeomorphism* $h : N \to \mathbb{R}^d$, *s.t.* $h(w^*) = 0$ *and in* N *the flow of* $\dot{w} = -F(w)$ *is topologically conjugate by the continuous map* $U = h(w)$ *to the flow of the linearized system* $\dot{U} = -HU$ *where* H *is the Hessian of* L.

Qianli Liao, Andrzej Barbuski

Deep RELU networks under square loss are usually degenerate



Theorem 4 *(K. Takeuchi) Let H be a positive integer. Let $h_k = W_k\sigma(h_{k-1}) \in \mathbb{R}^{N_k,n}$ for $k \in \{2, \ldots, H+1\}$ and $h_1 = W_1 X$, where $N_{H+1} = d'$. Consider a set of H-hidden layer models of the form, $\hat{Y}_n(w) = h_{H+1}$, parameterized by $w = \mathrm{vec}(W_1, \ldots, W_{H+1}) \in \mathbb{R}^{dN_1+N_1N_2+N_2N_3+\cdots+N_HN_{H+1}}$. Let $L(w) = \frac{1}{2}\|\hat{Y}_n(w) - Y\|_F^2$ be the objective function. Let w^* be any twice differentiable point of L such that $L(w^*) = \frac{1}{2}\|\hat{Y}_n(w^*) - Y\|_F^2 = 0$. Then, if there exists $k \in \{1, \ldots, H+1\}$ such that $N_kN_{k-1} > n \cdot \min(N_k, N_{k+1}, \ldots, N_{H+1})$ where $N_0 = d$ and $N_{H+1} = d'$ (i.e., overparametrization), there exists a zero eigenvalue of Hessian $\nabla^2 L(w^*)$.*

Degenerate minima of Deep RELU networks under square loss can (?) be regularized (dynamically)

Usual regularization is not sufficient because regularization shifts the minimum

$$H = -2\sum_{n=1}^{N}(\nabla_{W^k} f(W;x_n))(\nabla_{W^{k'}} f(W;x_n)) - \lambda_k \delta_{kk'} \mathbb{I},$$

However regularization “switched on” centered at w^* where the Hessian is degenerate and loss is very small, should enforce positive definiteness of the Hessian. The minimum in this dynamically regularized version of GD controls now the Frobenius norm at each layer k. The argument can be applied to DNNs because of Hartman-Grobman.

Qianli Liao, Andrzej Barbuski

Dynamical Systems Approach:

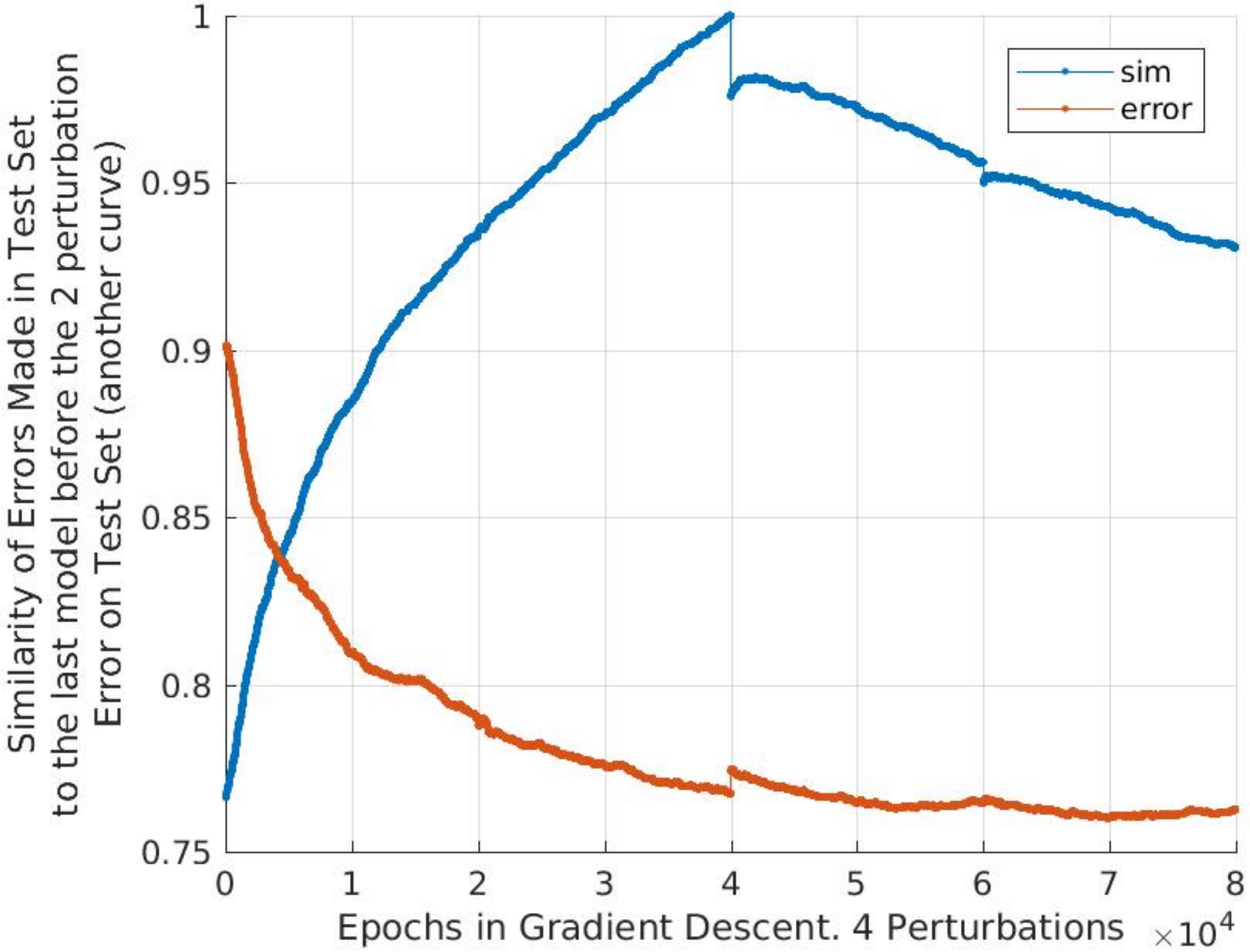
linear networks and exponential loss

$$L = \sum_{n}^{N} e^{-y_n w^T x_n}$$

$$\dot{w} = \sum_{n}^{N} e^{-y_n w^T x_n} x_n$$

There is no minimum, weights grow to infinity. In analogy with the linear case but for completely different reasons Srebro et al. proved (2017) that there can be convergence to the minimum L_2 solution *independently of initial conditions* at infinite time of $\frac{w}{\|w\|}$. It is still unclear whether experiments (next slide) support this.

Exponential loss linear net, these simulations show no convergence at finite time, dependence on perturbations



Qianli Liao, Andrzej Barbuski

Dynamical Systems Approach:

linear networks, exponential loss and regularization

$$L = \sum_{n}^{N} e^{-y_n w^T x_n} + \lambda \ ||w||^2$$

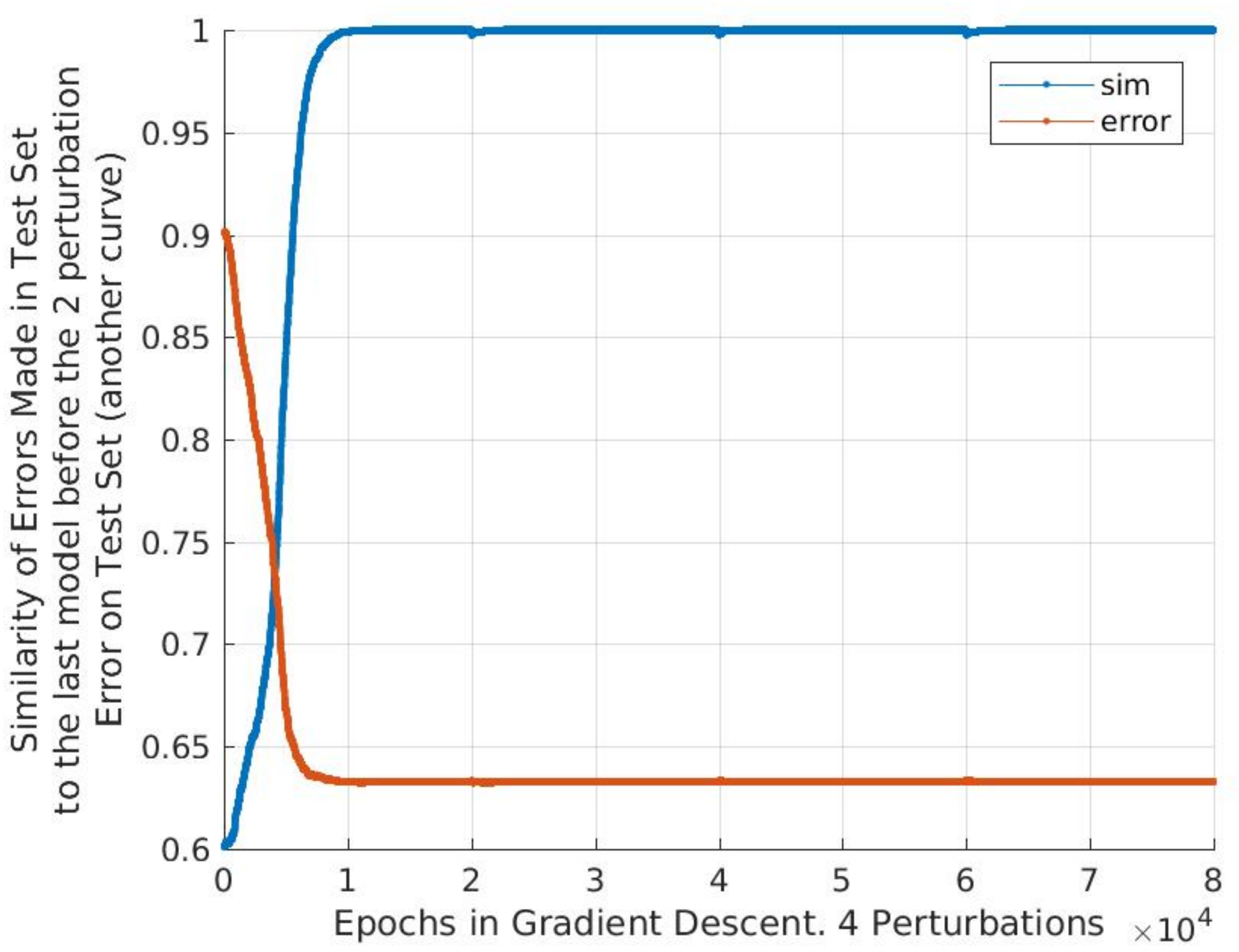
$$\dot{w} = \sum_{n}^{N} e^{-y_n w^T x_n} x_n - \lambda w$$

There is a stable minimum, also *independent of initial conditions,* and the Hessian of L is positive definite

$$H = -\sum_{n}^{N} e^{-y_n w^T x_n} x_n x_n^T - \lambda I$$

Qianli Liao, Andrzej Barbuski

Exponential loss linear net, regularization, does not depend on perturbations



Qianli Liao, Andrzej Barbuski

Homogeneity properties of RELU networks

For RELUs $\sigma(\mathrm{z}) = \mathrm{z}\,\frac{\partial \sigma}{\partial z}$ and

$$\mathrm{f}\,(\mathrm{W};\,\mathrm{x}) = \prod \rho_k \tilde{f}(V;x) \text{ where } \rho_k \mathrm{v}_k^{i,j} = w_k^{i,j},\;\; \rho_k^2 = \sum_{i,j} (w_k^{i,j})^2$$

Also (Rakhlin et al., 2017)

$$w_h^{ij} \frac{\partial f}{\partial w_k^{ij}} = f$$

Dynamical Systems Approach:

Deep Networks, exponential loss

$$L = \sum_{n}^{N} e^{-y_n f(x_n)}$$

$$\dot{w}_k^{i,j} = \sum_{n}^{N} e^{-y_n f(x_n)} \frac{\partial f}{\partial w_k^{i,j}}$$

$$H_{abk'}^{ijk} = \sum_{n}^{N} e^{-y_n f(x_n)} \left(- \frac{\partial f}{\partial w_k^{i,j}} \frac{\partial f}{\partial w_{k'}^{a,b}} + \frac{\partial^2 f}{\partial w_k^{i,j} \partial w_{k'}^{a,b}}\right)$$

The minimum is usually degenerate (first term has correct sign but is rank one).

Qianli Liao, Andrzej Barbuski

Dynamical Systems Approach

A (slightly) different algorithm: time-dependent regularization

$$\dot{w}_k^{i,j} = \sum_n^N e^{-y_n f(x_n)} \frac{\partial f}{\partial w_k^{i,j}} - \lambda \sum_k \| w_k - w_k^0 \|^2$$

$$H_{abk'}^{ijk} = \sum_n^N e^{-y_n f(x_n)} \left(-\frac{\partial f}{\partial w_k^{i,j}} \frac{\partial f}{\partial w_{k'}^{a,b}} + \frac{\partial^2 f}{\partial w_k^{i,j} \partial w_{k'}^{a,b}}\right) - \lambda(t) I$$

Switching on a regularization term at large time creates a regularized minimum where L is close to zero and H is degenerate. That minimum is stable with <u>arbitrarily small</u> λ . Hartman-Grobman can now be used to guarantee norm control at each layer of a deep network for this algorithm.

Qianli Liao, Andrzej Barbuski

Dynamical Systems Approach:

Deep Networks, exponential loss, weight normalization

Weight normalization (different forms of it) induce a regularization-like term and maximize $\tilde{f}$ - to be precise $\max \ \min_n y_n \tilde{f}(x_n)$ - subject to $\|V_k\| = 1$.
 This does not guarantees stable minima: there is in fact a counterexample for the linear network case. However, from the point of view of learning theory adding the constraint $\|V_k\| = 1$ to the dynamical system does not change it: this shows that normalization is enforced!.

- Define $\frac{w}{||w||} = \tilde{w}$; thus $w = ||w||\tilde{w}$ with $||\tilde{w}|| = 1$
- We assume $f(w) = f(||w||, \tilde{w}) = ||w|| f(1, \tilde{w}) = ||w||\tilde{f}$

1. $\frac{\partial ||w||}{\partial w} = \tilde{w}$
2. $\frac{\partial \tilde{w}}{\partial w} = \frac{I - \tilde{w}\tilde{w}^T}{||w||} = S.$

$$\dot{\tilde{w}} = \sum_{n=1}^{N} e^{-\rho \tilde{f}(x_n)} (\frac{1}{\rho}(\frac{\partial \tilde{f}(x_n)}{\partial \tilde{w}} - \tilde{w}\tilde{w}^T \frac{\partial \tilde{f}(x_n)}{\partial \tilde{w}})$$

Qianli Liao, Andrzej Barbuski

Dynamical Systems Approach: ***conjecture***

The vanilla SGD algorithms *do not converge to a stable minimum*. They minimize the margin, $\gamma = \rho\tilde{f}$ — avoiding the classification errors by growing the norm. It seems that *separating the data may be good enough in most cases*. Empirically the convergence seems to be to a degenerate minimum as indicated by Theory II and the linear network case. The weight normalization algorithms satisfy the key requirement of maximizing $\tilde{f}$ *while strictly enforcing the normalization constraint*. However they converge to degenerate minima.

Recall $\mathbb{E}\mathbf{L}_{01}(\widehat{\mathrm{f}}_n) \le \mathbb{E}\widehat{\mathbf{L}}_{\psi}(\widehat{\mathrm{f}}_n) + \frac{2}{\gamma}\mathscr{R}(\mathcal{F})$ and that the goal is to minimize

$\frac{\Re(f)}{\gamma} = \frac{\rho}{y\rho\tilde{f}} = \frac{1}{\tilde{f}}$ that is to maximize $\tilde{f}$ subject to $\left\|V_k\right\| = 1$

Dynamical Systems Approach:

New SGD algorithm for maximizing $\tilde{f}$ with non degenerate minimum..

We define $W_k = \rho_k V_k$ where we split each weight matrix into a matrix of unit Frobenius norm and the scalar norm. Then we use a penalty term in the loss to control the Frobenius norm

$$L = \sum_n^N e^{-f(x_n)y_n} + \sum_k^K \lambda(\|V_k\|^2 - 1)$$

The resulting GD equations give the dynamical system

$$\dot{\rho} = \rho \sum_n^N e^{-\rho \tilde{f}(x_n)} \tilde{f}(x_n) \qquad \dot{v}_k^{i,j} = \rho \sum_n^N e^{-\rho \tilde{f}(x_n)} \frac{\partial \tilde{f}(x_n)}{\partial v_k^{i,j}} - 2\lambda v_k^{i,j}$$

with Hessian

$$\left[-\rho^2 \frac{\partial \tilde{f}(x_*)}{\partial v_k^{i,j}} \frac{\partial \tilde{f}(x_*)}{\partial v_{k'}^{a,b}} + \rho \frac{\partial^2 \tilde{f}(x_*)}{\partial v_k^{i,j} \partial v_{k'}^{a,b}} \right] e^{-\rho \tilde{f}(x_*)} - 2\lambda I$$

Qianli Liao, Andrzej Barbuski

Dynamical Systems Approach:

new SGD algorithm for maximizing margin with “early stopping”

Consider the dynamical system

$$\dot{\rho} = \sum_n^N e^{-\rho \tilde{f}(x_n)} \tilde{f}(x_n) \qquad \dot{v}_k^{i,j} = \rho \sum_n^N e^{-\rho \tilde{f}(x_n)} \frac{\partial \tilde{f}(x_n)}{\partial v_k^{i,j}} - 2\lambda v_k^{i,j}$$

with $\frac{\partial}{\partial t} \| V_k \|^2 = 2\rho \sum_n^N e^{-\rho \tilde{f}(x_n)} \tilde{f}(x_n) - 2\lambda \| V_k \|^2 = 2(\rho\dot{\rho} - \lambda \| V_k \|^2)$ implying that for $\lambda = \rho \sum_n^N e^{-\rho \tilde{f}(x_n)} \tilde{f}(x_n)$ then $\| V_k \| = 1$, the constraint is enforced, SGD should be stopped while the loss is being minimized and the margin maximized under the constraint

Qianli Liao, Andrzej Barbuski

Remarks

Set $yf(x) = y\rho\tilde{f}(x)$ where $\rho = \prod \rho_k$ and $\tilde{f}(x)$ contains the "to-be-normalized" weight matrices V_k with components $v_k^{i,j}$.

Then minimization of $L = \sum_n^N e^{-y_n f(x_n)}$ with the max effect in the exponential corresponds to maximizing the margin $\min_n y_n f(x_n) = \min_n y_n \rho \tilde{f}(x_n)$

The complexity of the solution attained by GD can be measured in terms of its Rademacher averages which are bounded by the product of the Frobenius norms of the weight matrices of the network and normalized by the margin. Thus

- the *margin normalized complexity* which is relevant for bounding the expected error in *classification* is proportional to $\tilde{f}$

- the *complexity* which is relevant for bounding the expected loss in terms of the *cross-entropy loss* is given by the Rademacher average and thus here by ρ

Remarks

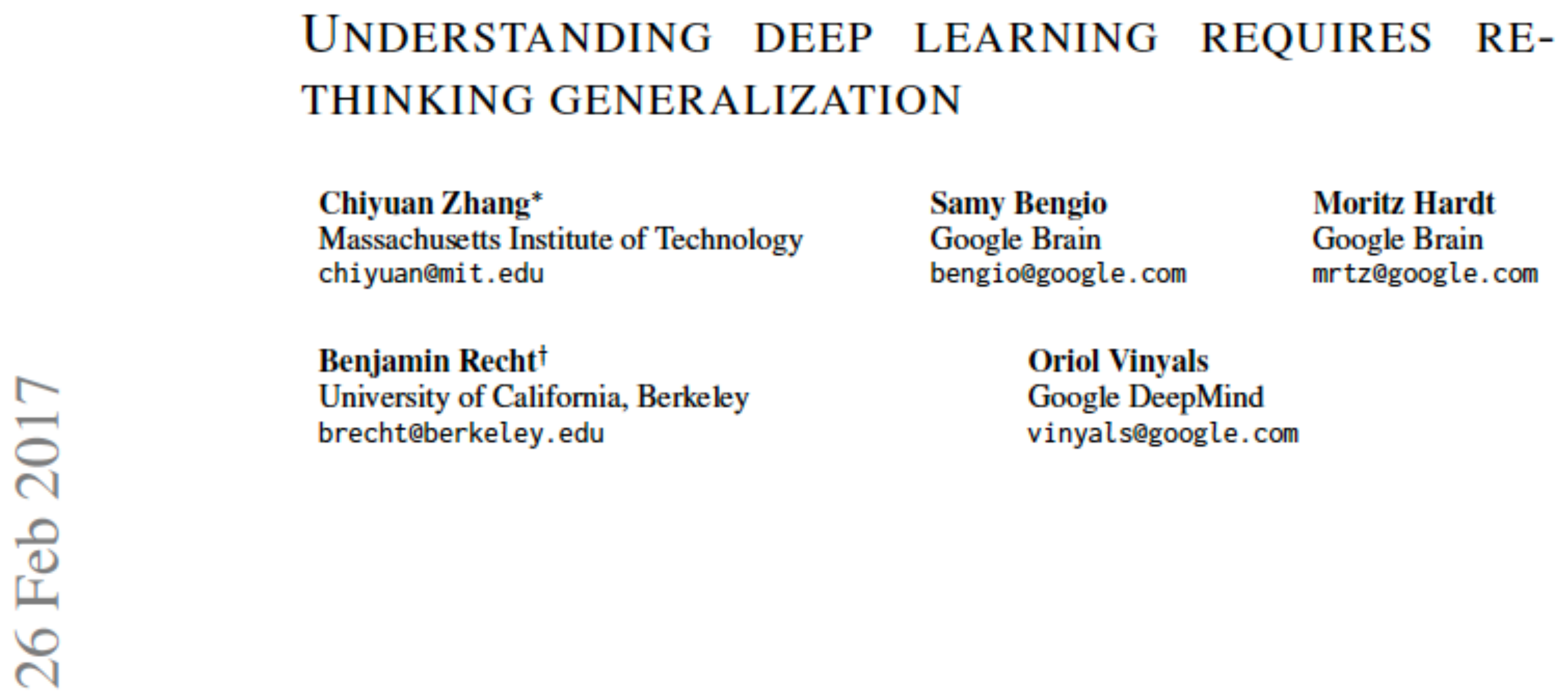
These properties

- the *margin normalized complexity* which is relevant for bounding the expected error in *classification* is proportional to $\tilde{f}$

- the *complexity* which is relevant for bounding the expected loss in terms of the *cross-entropy loss* is given by the Rademacher average and thus here by ρ

explain a few little puzzles:

- we can dial down complexity of f by setting $\rho = 1$: then this (scaled down) empirical minimizer predict well its expected *loss*…it can even predict expected from empirical error on *randomly labeled examples* (disproving claims in Zhang et al paper)!
- one of the next slides shows that test loss can increase while the classification error does not: this is because ρ increases affecting the bound on regression error but this does not affect the margin normalized complexity which is the key for bounding classification error.



UNDERSTANDING DEEP LEARNING REQUIRES RETHINKING GENERALIZATION

Chiyuan Zhang[*]
Massachusetts Institute of Technology
chiyuan@mit.edu

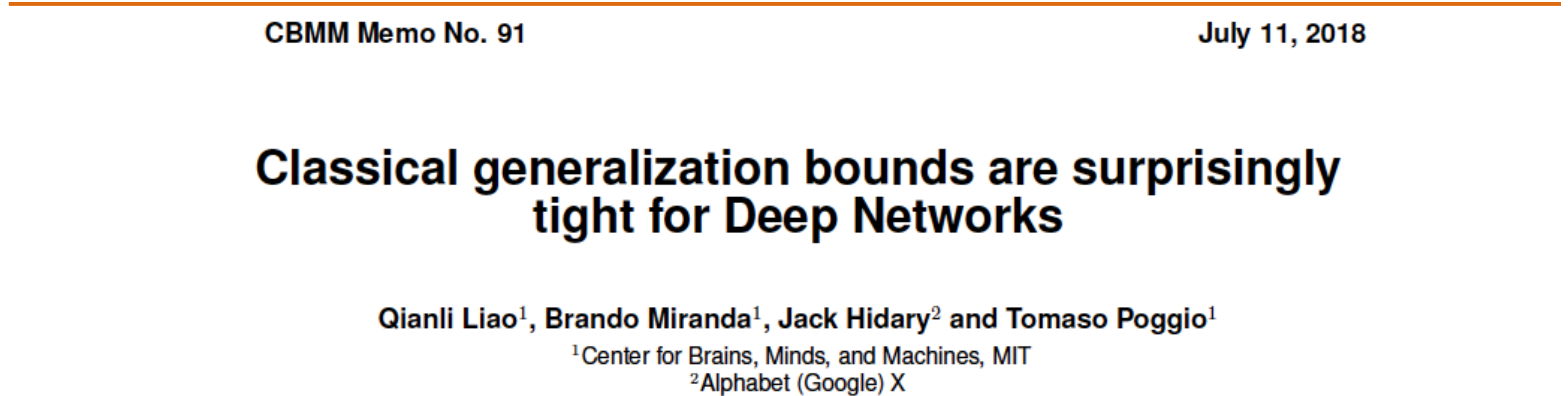
Samy Bengio
Google Brain
bengio@google.com

Moritz Hardt
Google Brain
mrtz@google.com

Benjamin Recht[†]
University of California, Berkeley
brecht@berkeley.edu

Oriol Vinyals
Google DeepMind
vinyals@google.com

26 Feb 2017



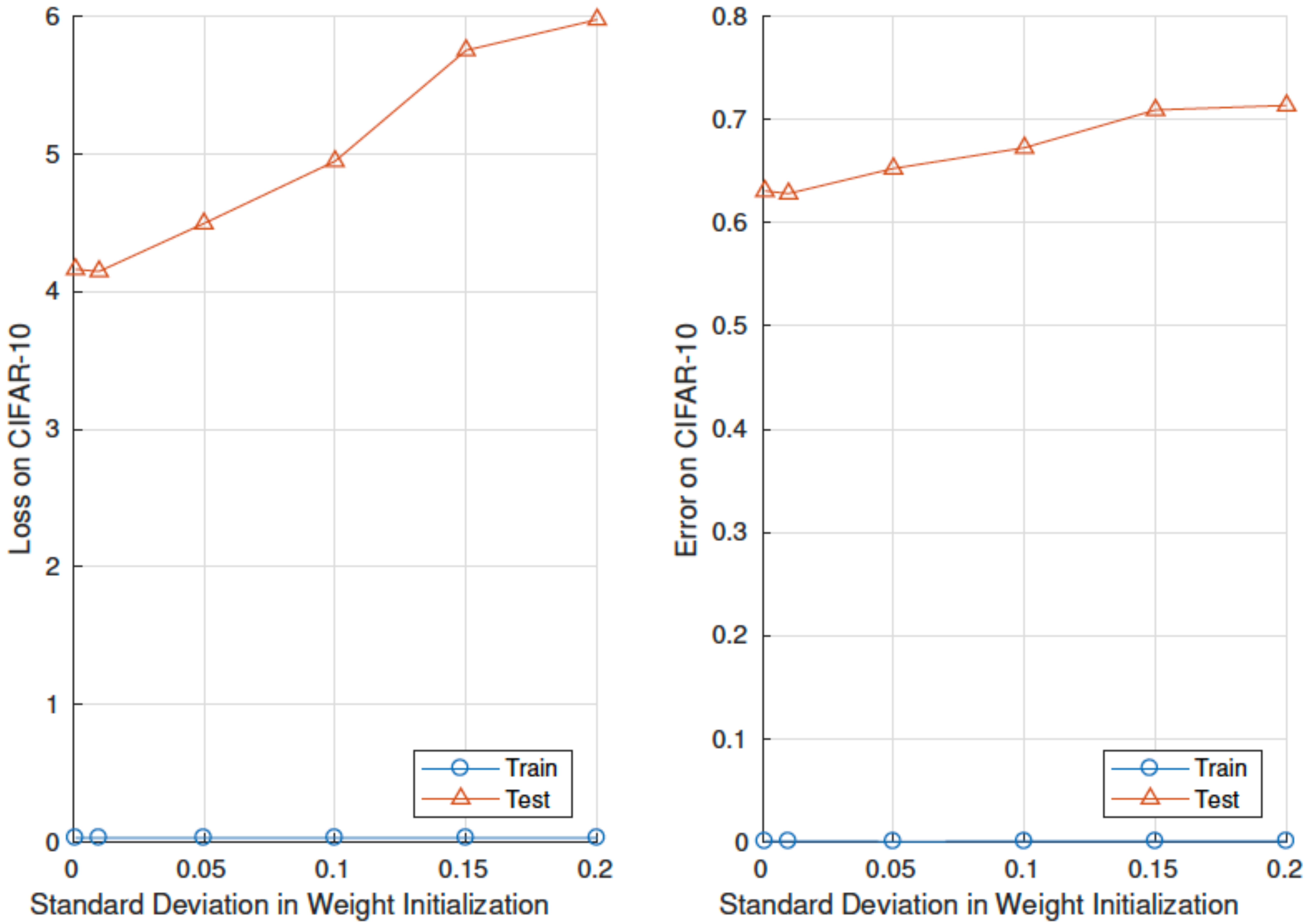
CBMM Memo No. 91

July 11, 2018

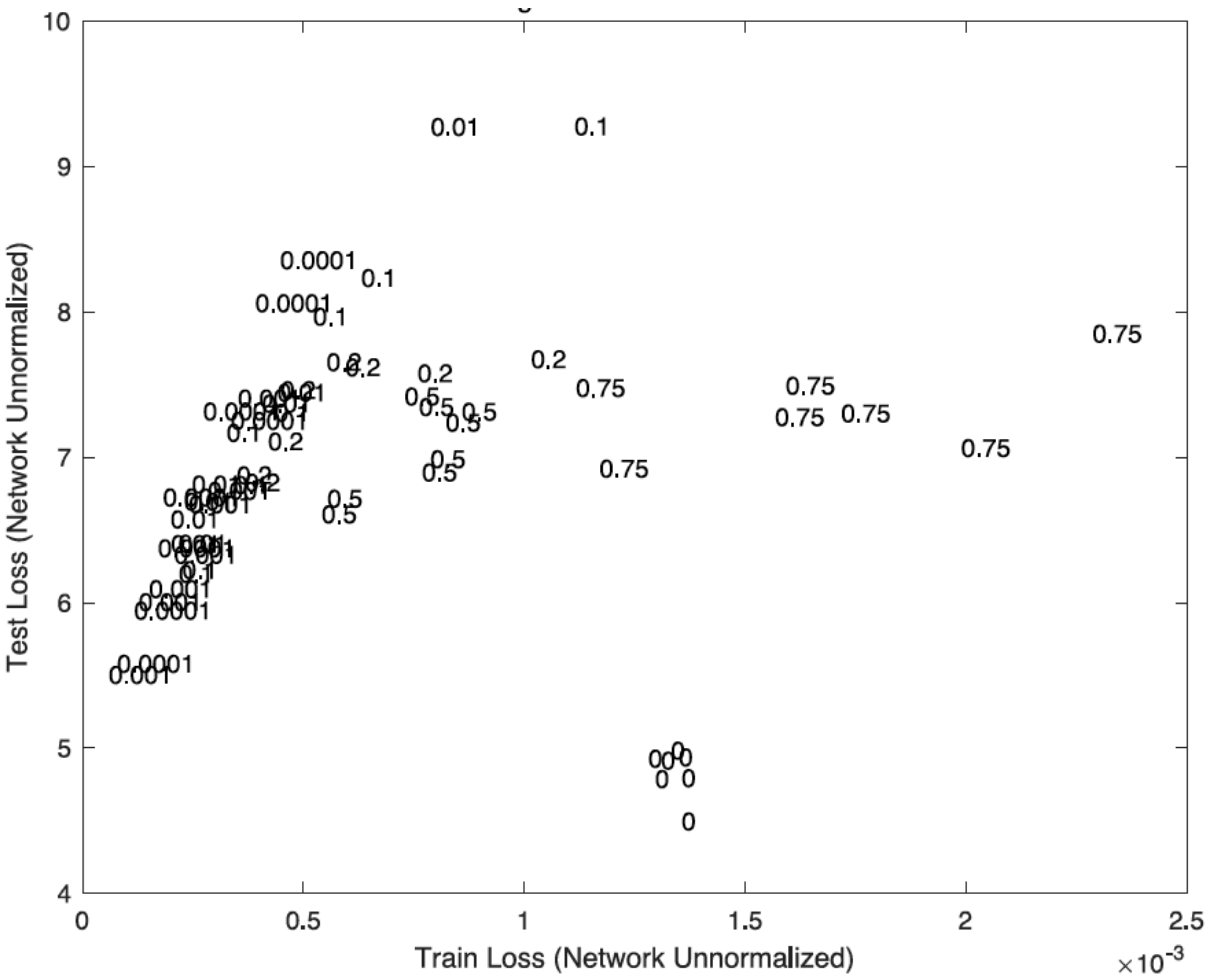
Classical generalization bounds are surprisingly tight for Deep Networks

Qianli Liao[1], **Brando Miranda**[1], **Jack Hidary**[2] **and Tomaso Poggio**[1]
[1]Center for Brains, Minds, and Machines, MIT
[2]Alphabet (Google) X

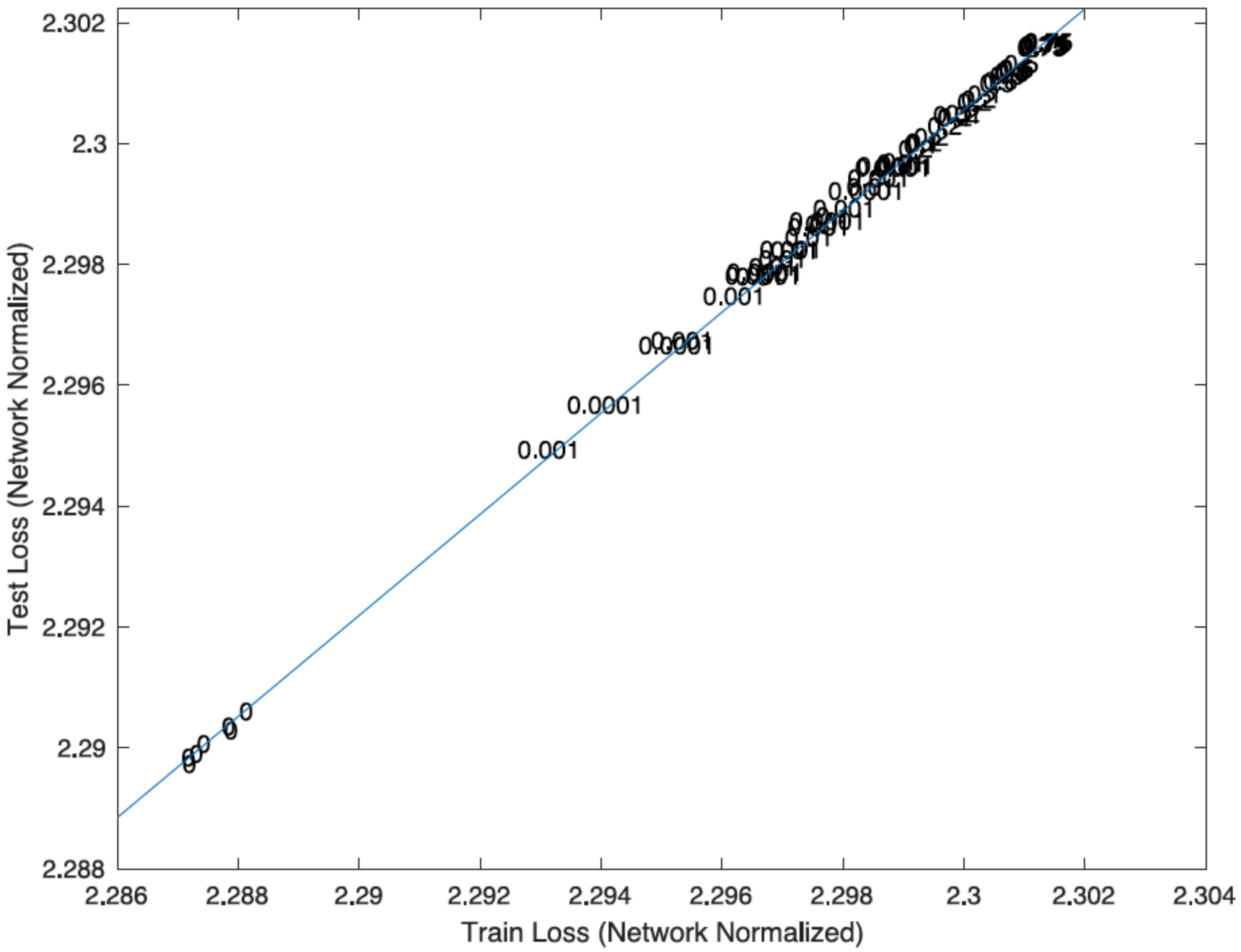
How to generate minima with zero training error but different test errors



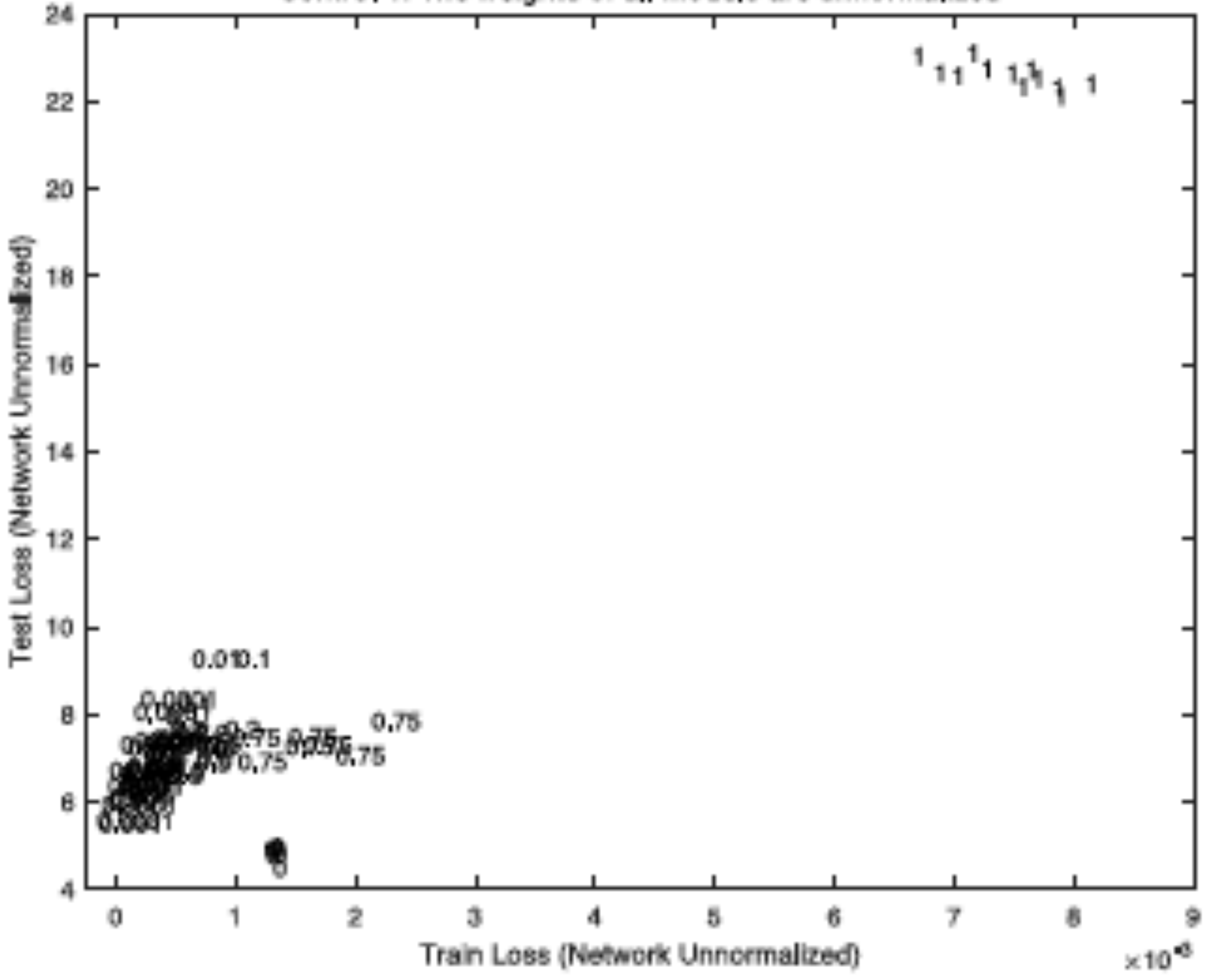
Unnormalized mess

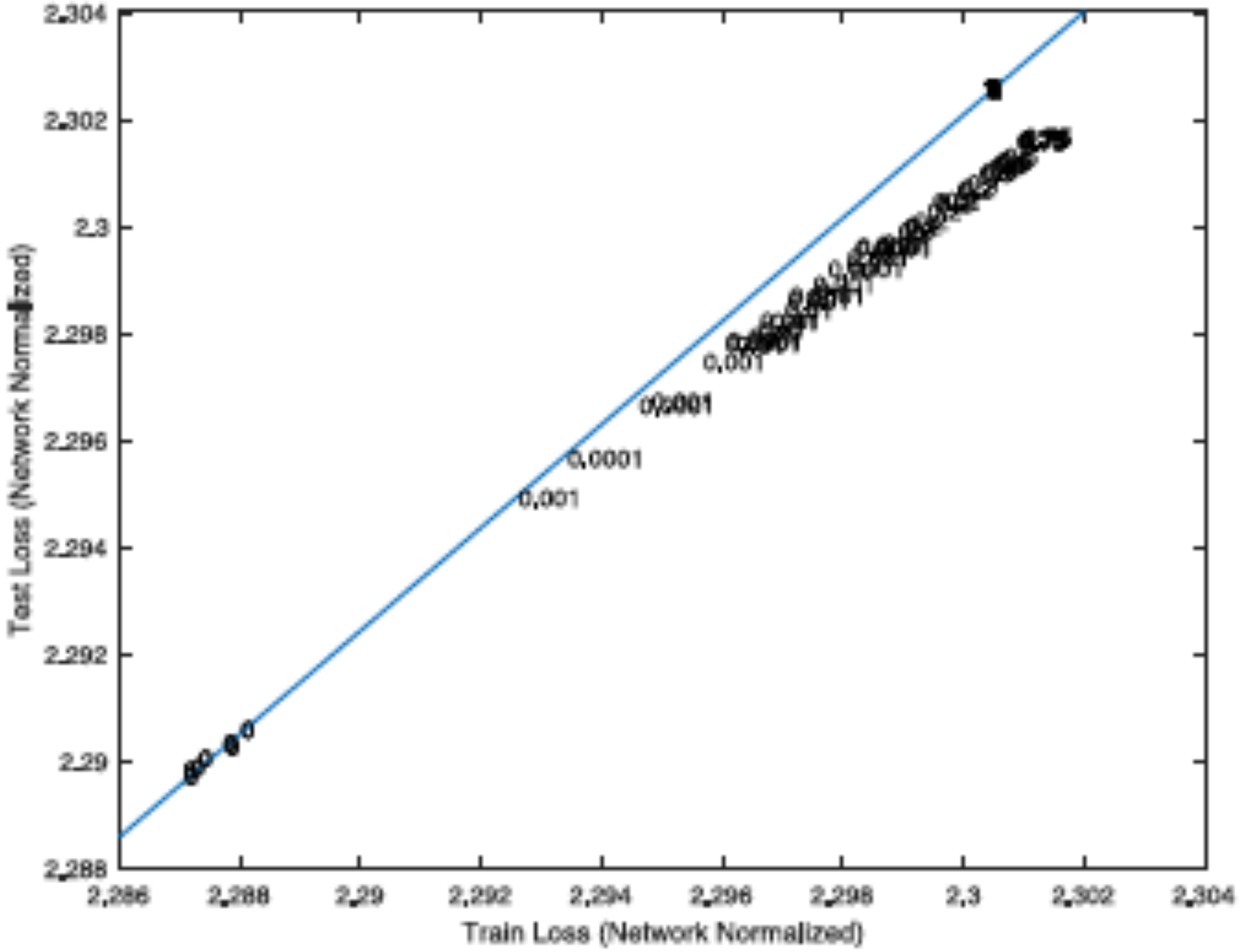


The magic of layer-wise normalization



Control 1: The weights of all models are unnormalized





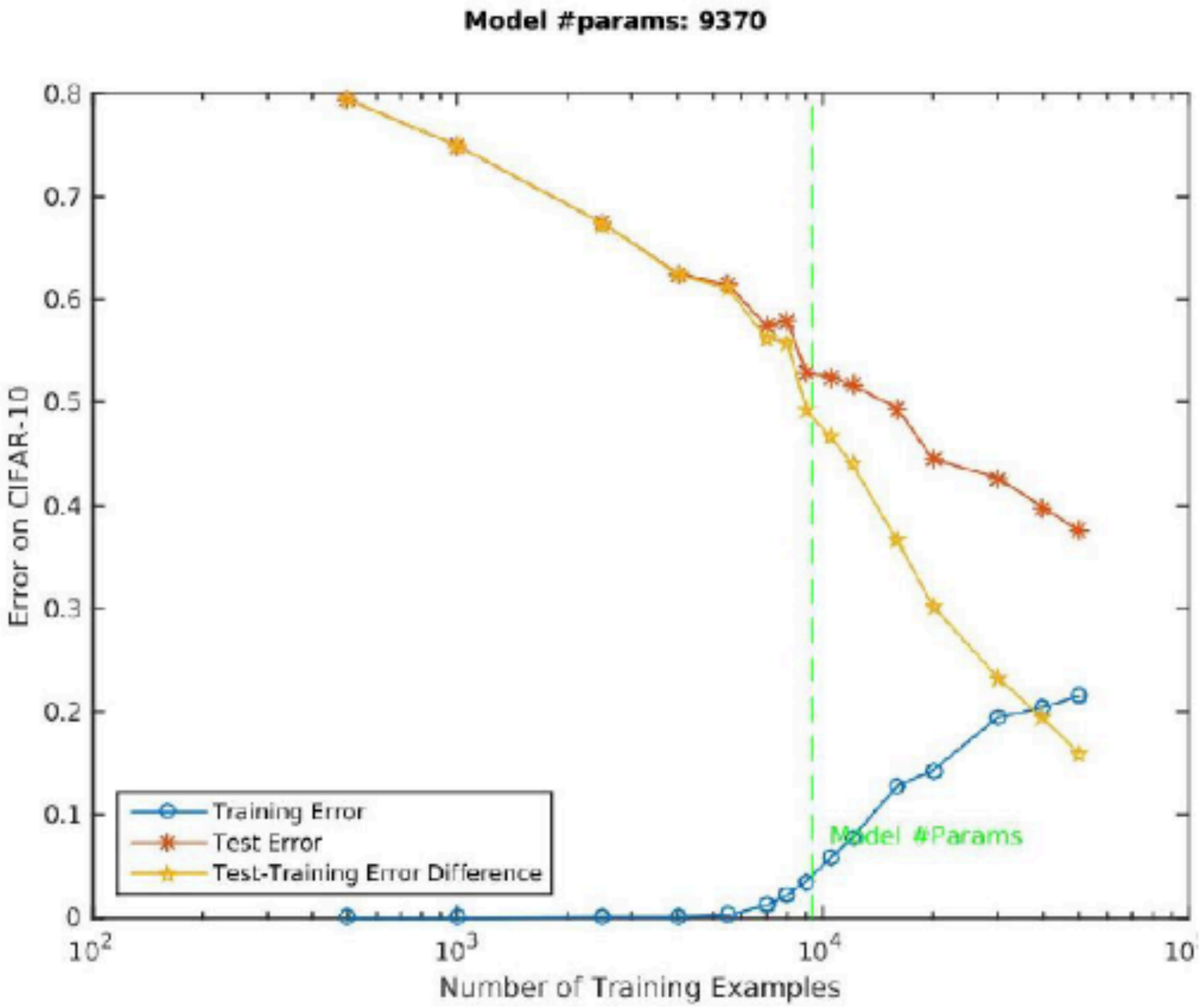
The reason is that we are looking at generalization in regression (not classification)

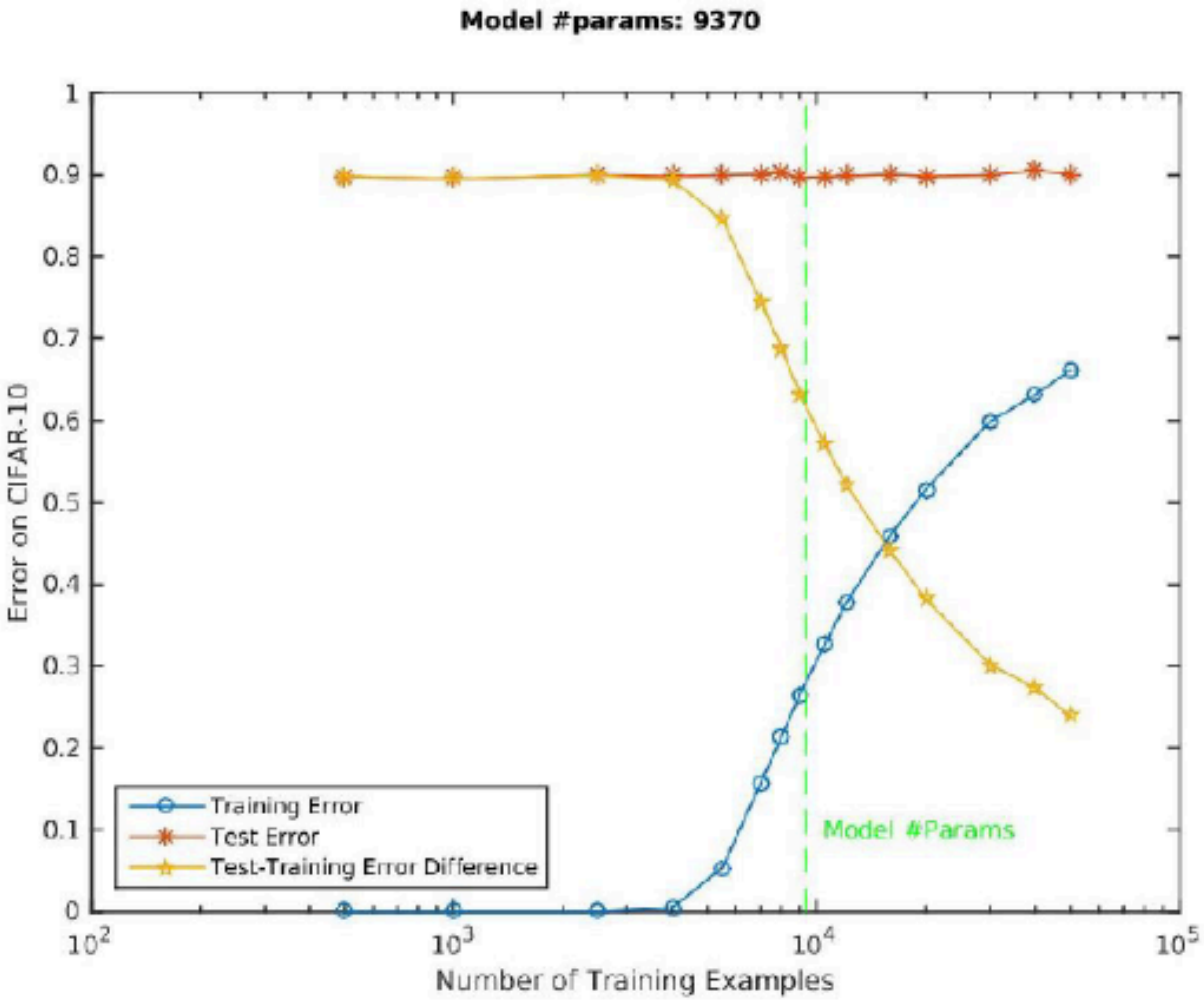
Consider typical generalization bounds for regression: they have the following form:

$$\textit{W}\text{ith probability } \geq (1\text{-}\delta)$$

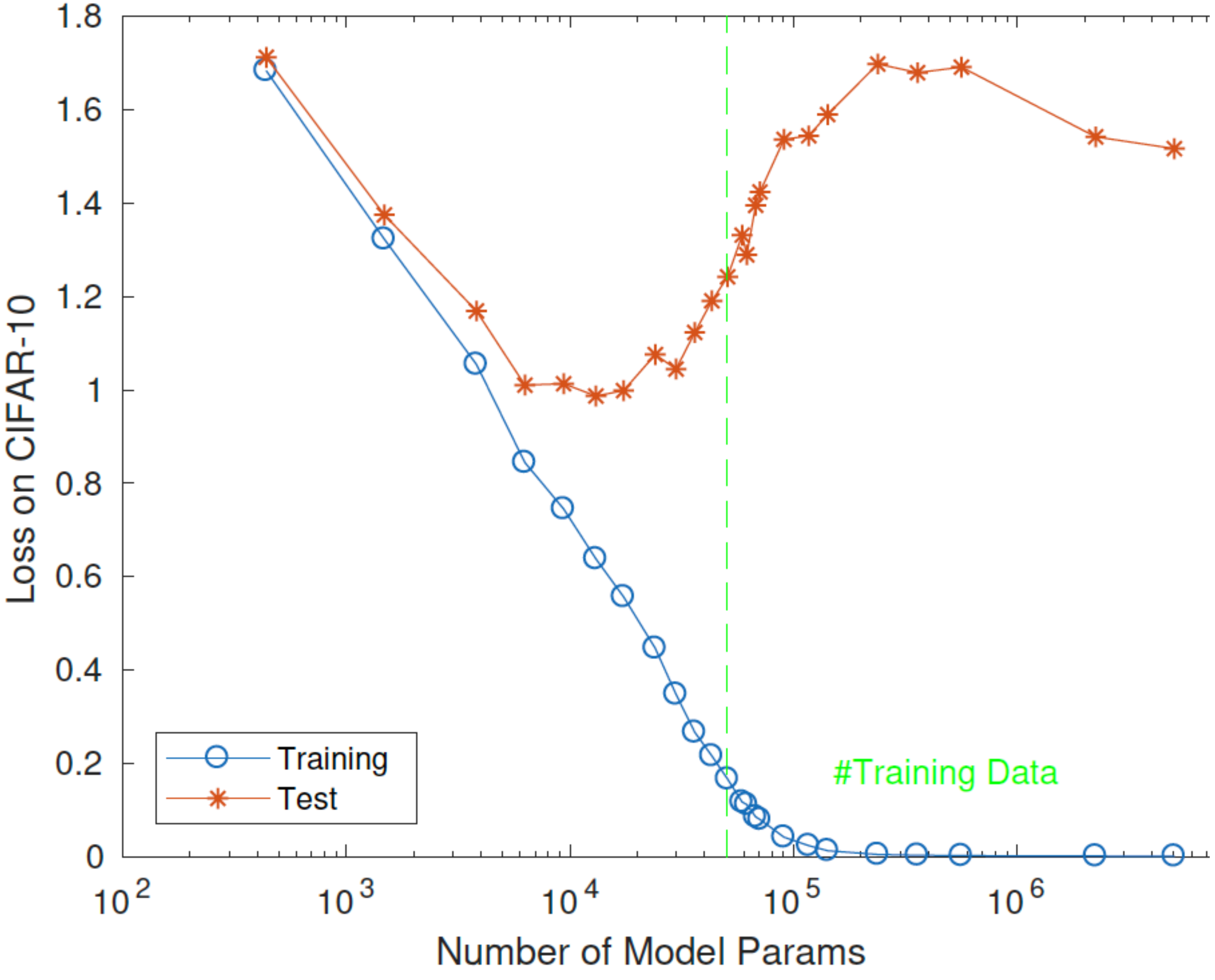
$$\left|E(f)-E_S(f)\right| \leq 2\mathbb{R}_N(\mathbb{F})+\sqrt{\frac{\ln\frac{1}{\delta}}{2N}}$$

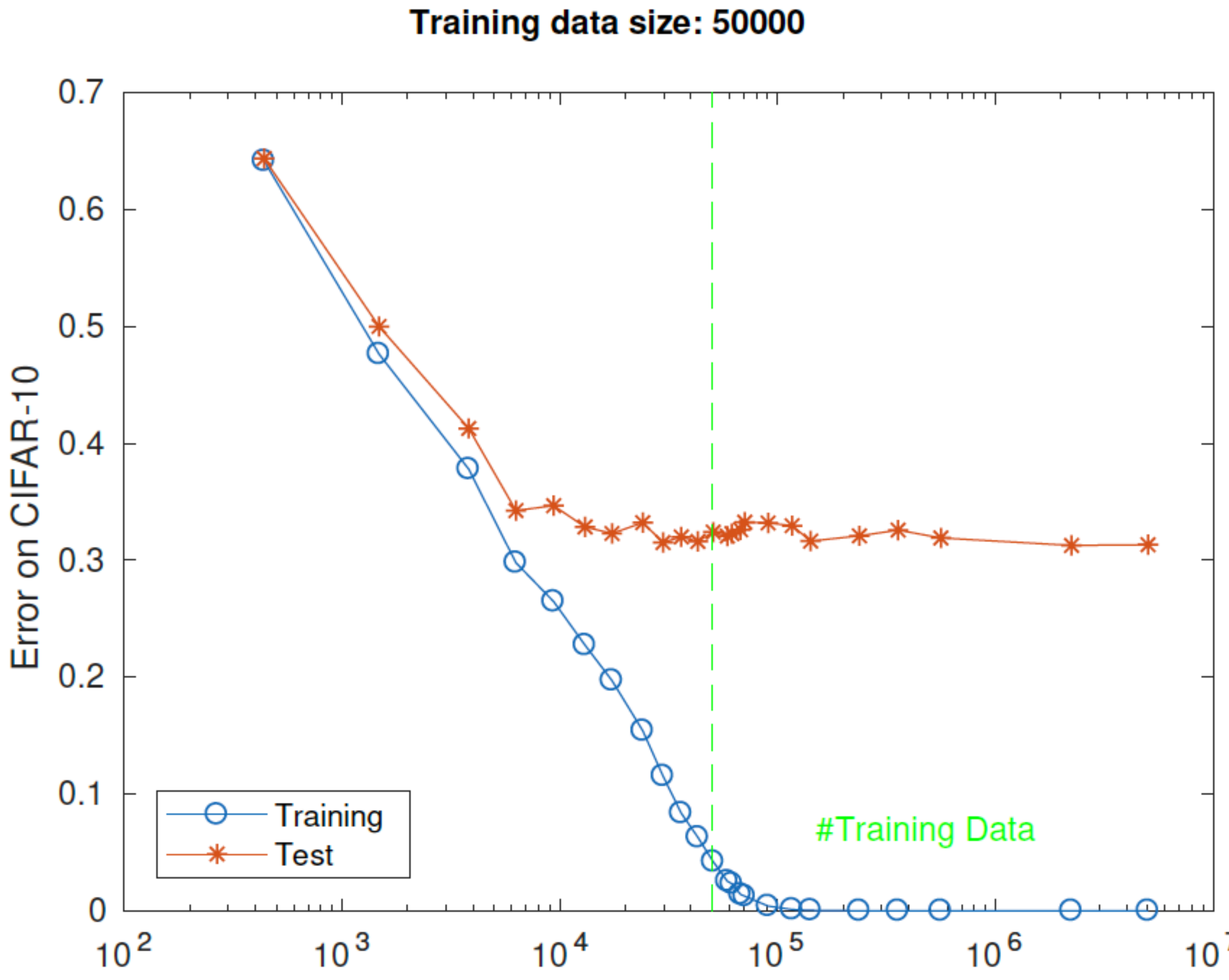
Good generalization with less data than # weights
Large capacity: fitting randomly labeled data





Deep nets puzzle: ρ grows and $\tilde{f}$ does not





General musings

The evolution of computer science

- there were programmers
- there are now labelers
- there may be schools for bots…

Today's science, tomorrow's engineering: learn like children learn

The first phase (and successes) of ML:

supervised learning, big data: $n \to \infty$

Siri

Use your voice to send messages, set reminders, search for information, and more.

Autonomous Driving

Google's modified Toyota Prius uses an array of sensors to navigate public roads without a human driver. Other components, not shown, include a GPS receiver and an inertial motion sensor.

LIDAR
A rotating sensor on the roof scans more than 200 feet in all directions to generate a precise three-dimensional map of the car's surroundings.

POSITION ESTIMATOR
A sensor mounted on the left rear wheel measures small movements made by the car and helps to accurately locate its position on the map.

VIDEO CAMERA
A camera mounted near the rear-view mirror detects traffic lights and helps the car's onboard computers recognize moving obstacles like pedestrians and bicyclists.

RADAR
Four standard automotive radar sensors, three in front and one in the rear, help determine the positions of distant objects.

Source: Google

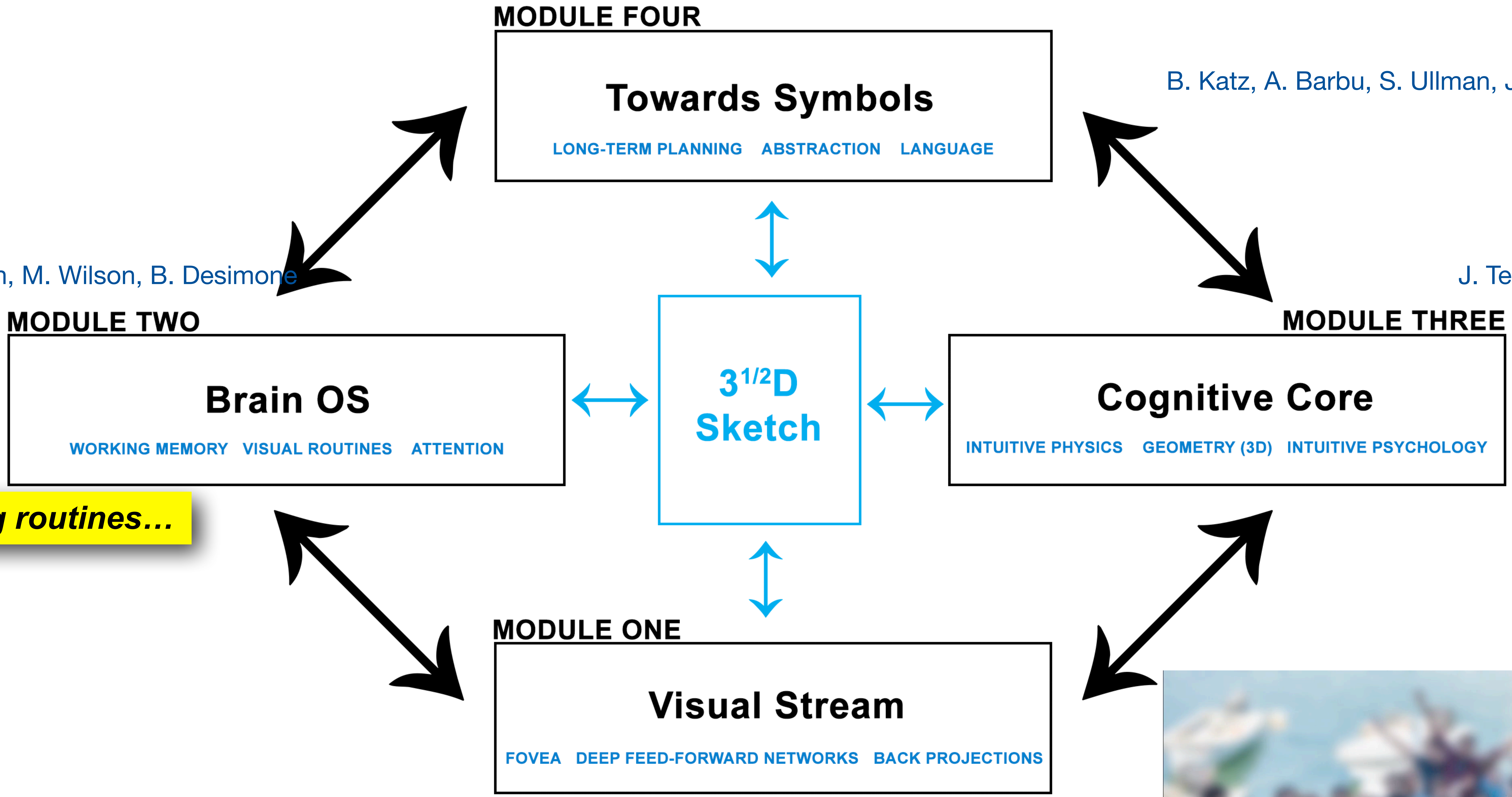
from programmers…

…to labelers…

…to computers that learn like children…

The next phase of ML: implicitly supervised learning,

learning like children do, small data: $n \to 1$



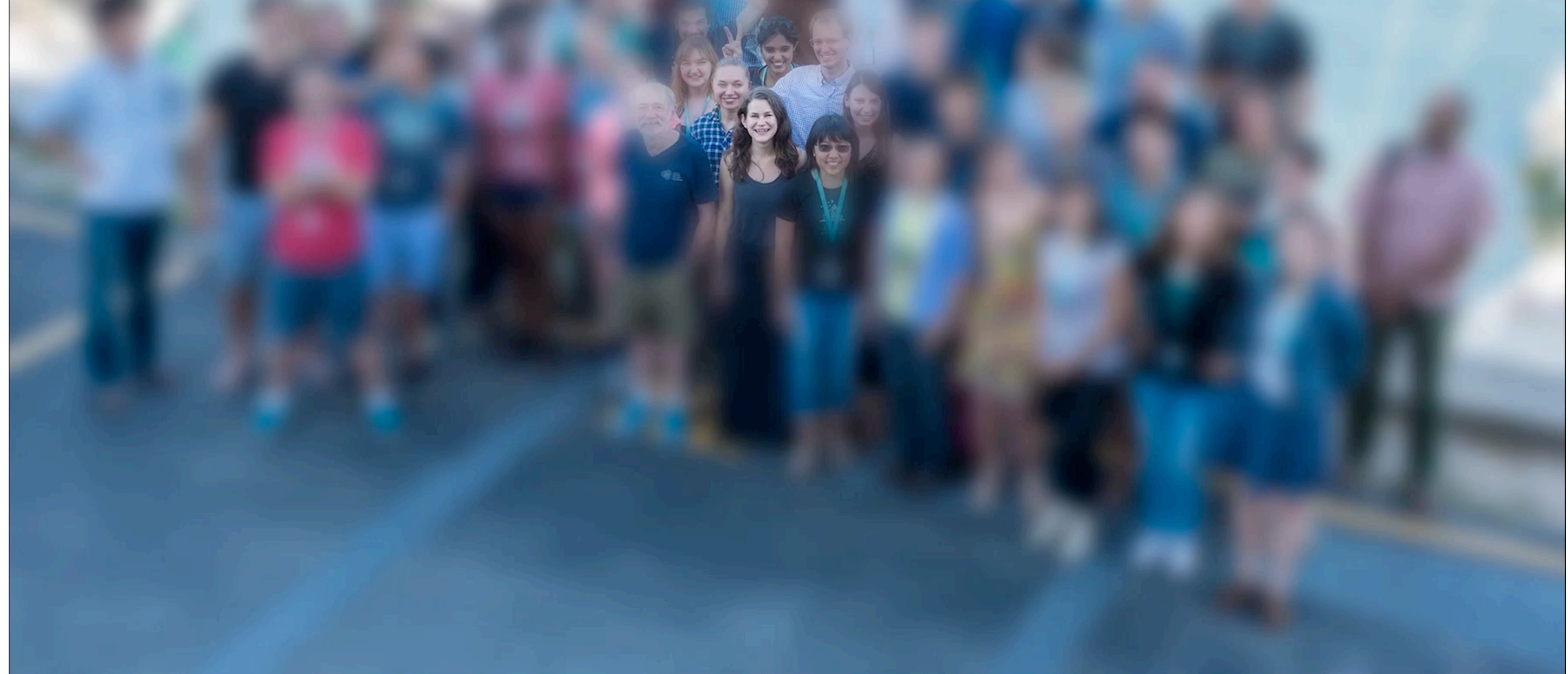
B. Katz, A. Barbu, S. Ullman, J. Tenenbaum

G, Kreiman, M. Wilson, B. Desimone

J. Tenenbaum, N. Kanwisher, Spelke

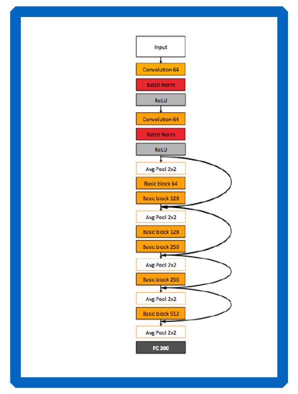
Running routines…

Depository of vision routines… synthesizing routines as needed



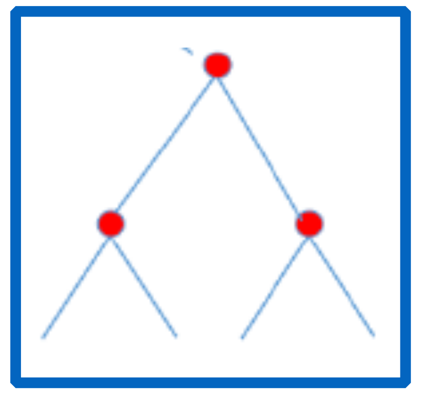
Musings on Near Future Breakthroughs

- new architectures/class of applications from basic DCN block (example GAN + RL/DL + …)



- Implicit labeling: evolution is opportunistic…few bits…face area…motion machinery…bootstrapping…predicting next “frame”…

- Learning and representing *symbols…with networks of neurons …*abstract concepts, relations, routines…new circuit motif in addition to DCN?



?

- New learning algorithm — more biologically plausible than SGD …