Inductive Bias and Optimization in Deep Learning

Nati Srebro (TTIC)

Based on work with **Behnam Neyshabur** (TTIC→Google), **Suriya Gunasekar** (TTIC→MSR),
Ryota Tomioka (TTIC→MSR), Srinadh Bhojanapalli (TTIC→Google),
Blake Woodworth, Pedro Savarese, David McAllester (TTIC), Greg Ongie, Becca Willett (Chicago),
Daniel Soudry, Elad Hoffer, Mor Shpigel, Itay Sofer (Technion),
Ashia Wilson, Becca Roelofs, Mitchel Stern, Ben Recht (Berkeley),
Russ Salakhutdinov (CMU), **Jason Lee**, Zhiyuan Li (Princeton), Yann LaCun (NYU/Facebook)

Feed Forward Neural Networks

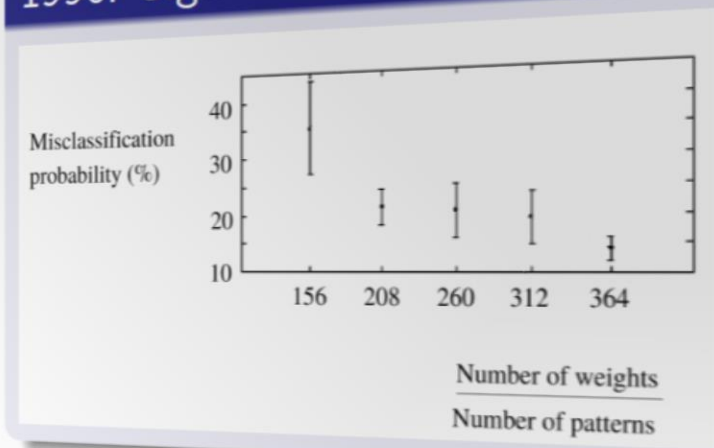
- Fix architecture (connection graph $G(V,E)$, transfer σ)

$$\mathcal{H}_{G(V,E),\sigma} = \{\, f_w(x) = output\ of\ net\ with\ weights\ w \,\}$$

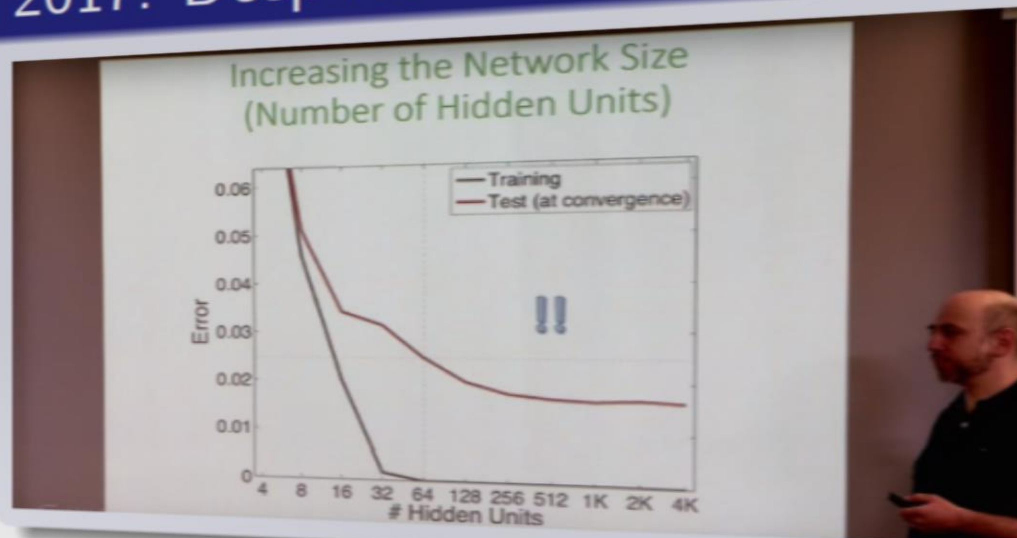
- Capacity / Generalization ability / Sample Complexity
  - $\tilde{\boldsymbol{O}}(|\boldsymbol{E}|)$ (number of edges, i.e. number of weights)
    (with threshold σ, or with RELU and finite precision; RELU with inf precision: $\tilde{\Theta}(|E| \cdot depth)$)
- Expressive Power / Approximation
  - Any continuous function with huge network
  - Lots of interesting things naturally with small networks
  - **Any time T computable function with network of size $\tilde{\boldsymbol{O}}(\boldsymbol{T})$**
- Computation / Optimization
  - NP-hard to find weights even with 2 hidden units
  - Even if function exactly representable with single hidden layer with $\Theta(\log d)$ units, even with no noise, and even if we allow a much larger network when learning: no poly-time algorithm always works
    [Kearns Valiant 94; Klivans Sherstov 06; Daniely Linial Shalev-Shwartz '14]
  - Magic property of reality that makes local search "work"

Generalization: Margins and Size of Parameters

1996: Sigmoid networks



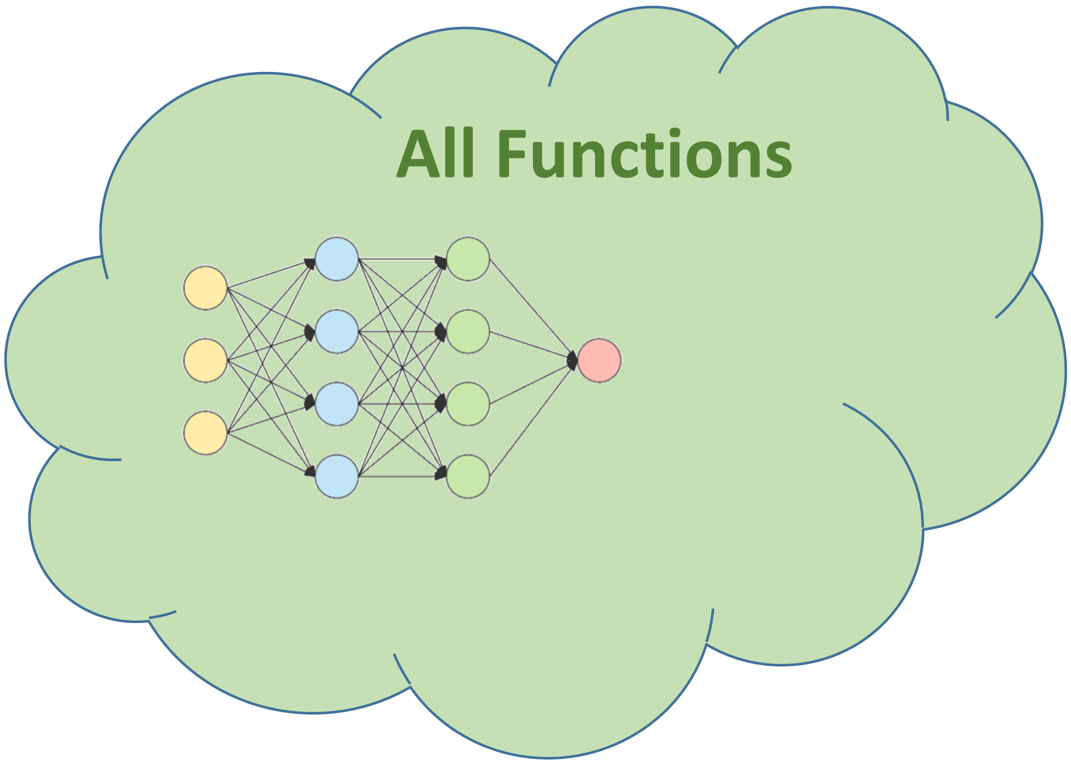
2017: Deep ReLU networks



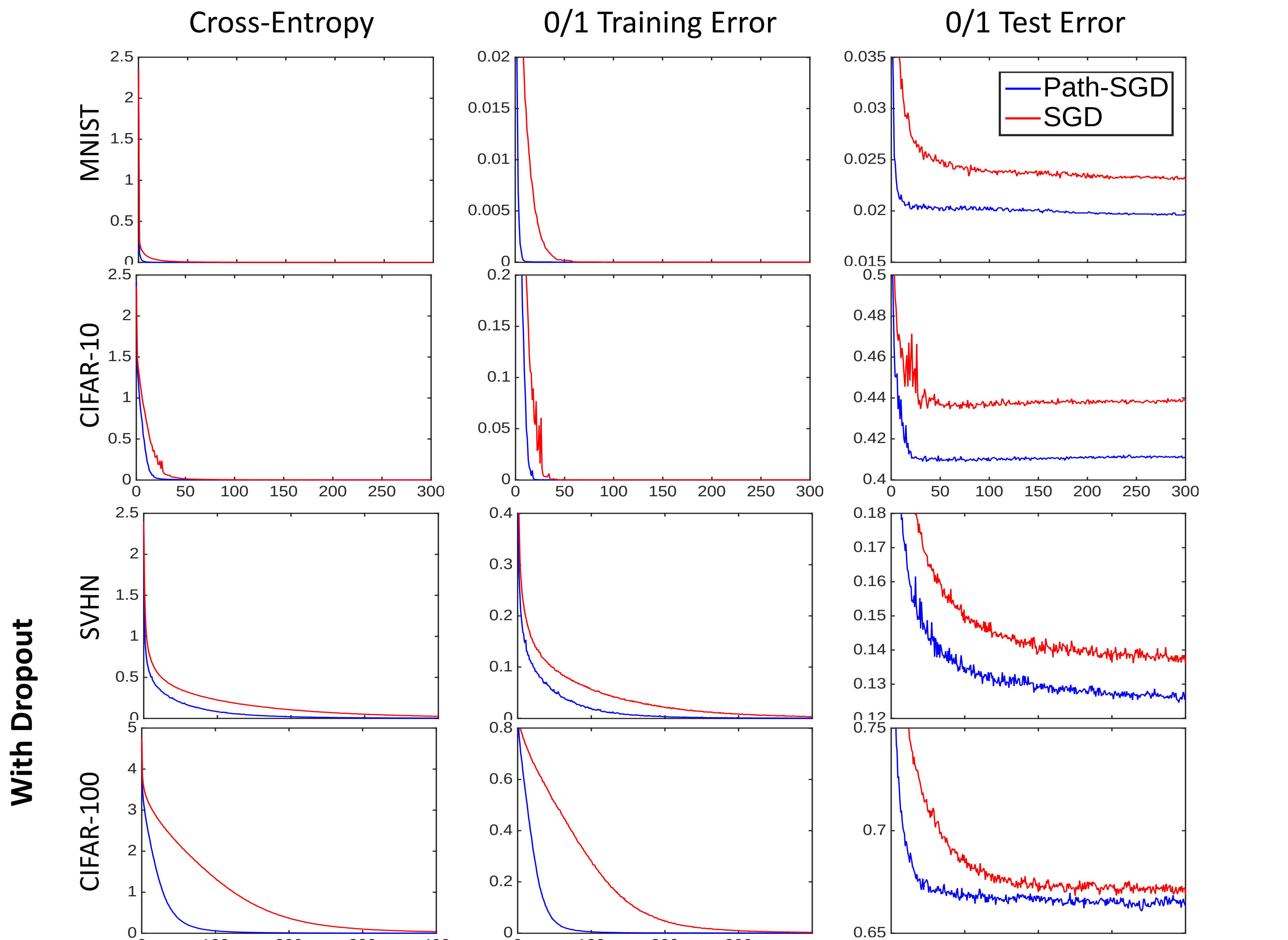
- Qualitative behavior explained
  small weights theorem.

- How to measure the complexity of a ReLU network?

Need to understand optimization alg. not just as reaching ***some*** (global) optimum, but as reaching a ***specific*** optimum

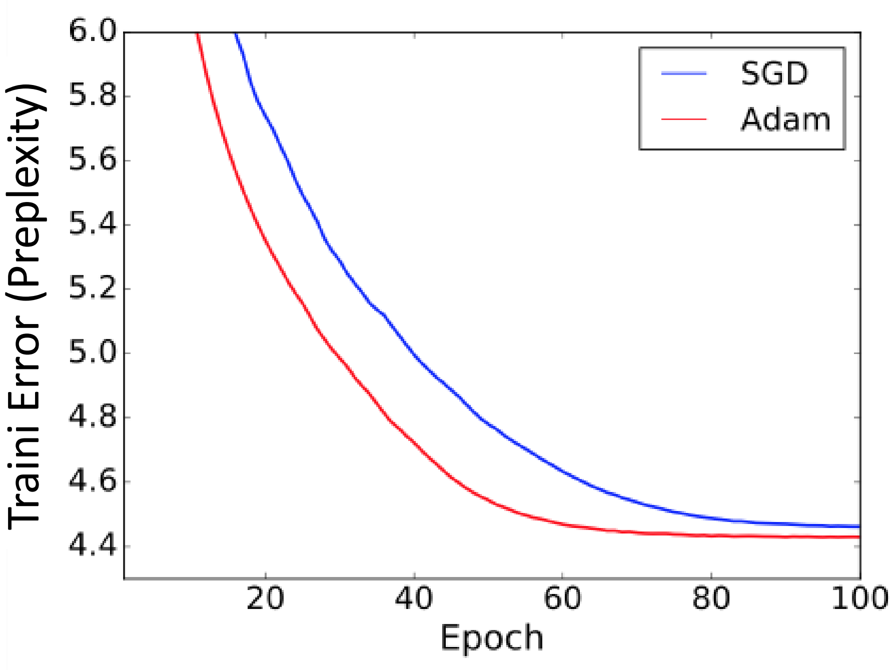


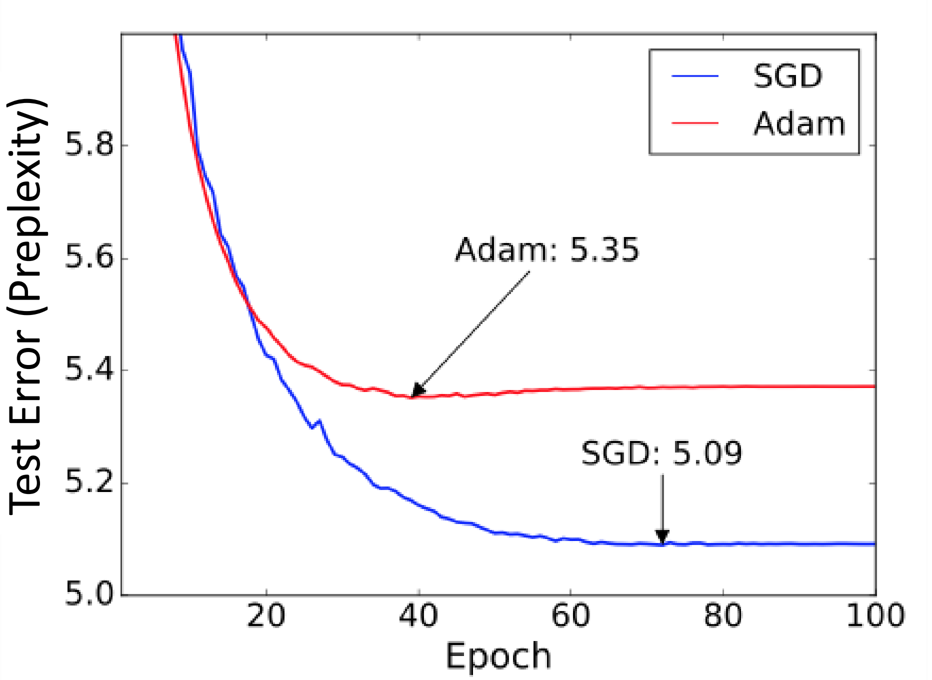
Different optimization algorithm
➔ Different bias in optimum reached
➔ Different Inductive bias
➔ Different generalization properties



[Neyshabur Salakhudtinov S NIPS'15]

SGD vs ADAM





Results on Penn Treebank using 3-layer LSTM

[Wilson Roelofs Stern S Recht, “The Marginal Value of Adaptive Gradient Methods in Machine Learning”, NIPS’17]

The Deep Recurrent Residual Boosting Machine

Joe Flow, DeepFace Labs

Section 1: Introduction

We suggest a new amazing architecture and loss function that is great for learning. All you have to do to learn is fit the model on your training data

Section 2: Learning Contribution: our model

The model class h_w is amazing. **Our learning method is:**

$$\arg\min_{\boldsymbol{w}} \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{loss}(\boldsymbol{h_w}(\boldsymbol{x}); \boldsymbol{y}) \qquad (*)$$

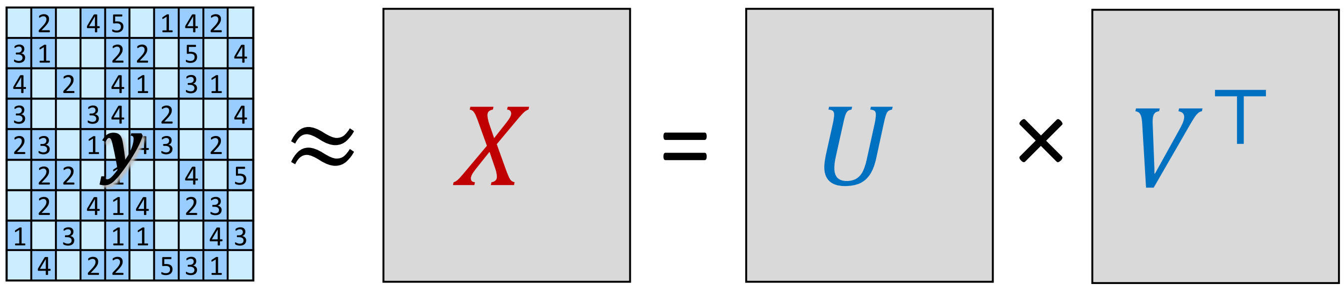
Section 3: Optimization

This is how we solve the optimization problem (*): […]

Section 4: Experiments

It works!

Unconstrained Matrix Completion

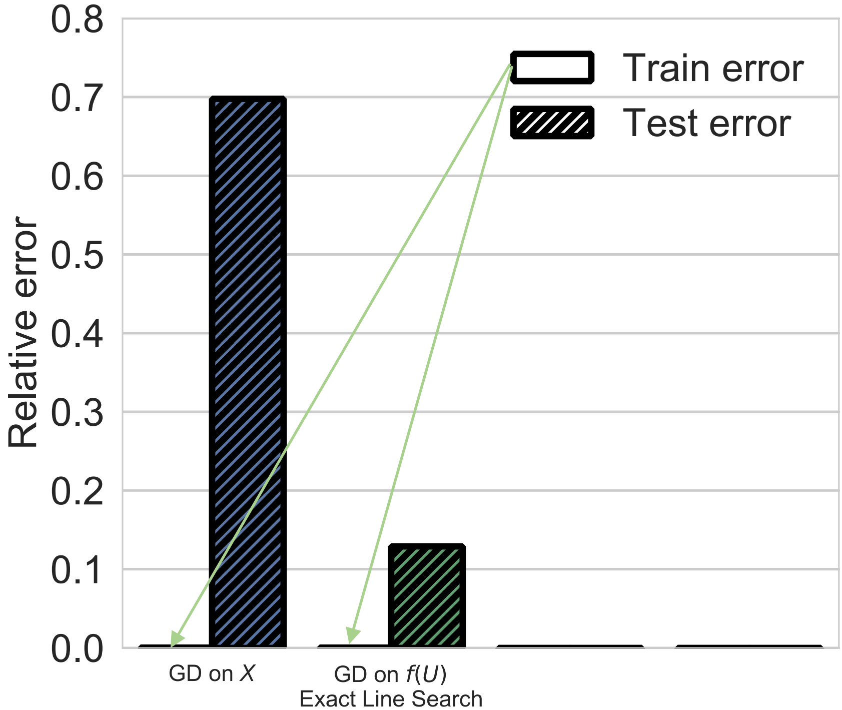


$$\min_{X\in\mathbb{R}^{n\times n}} \|observed(X) - y\|_2^2 \equiv \min_{U,V\in\mathbb{R}^{n\times n}} \left\|observed(UV^\top) - y\right\|_2^2$$

- Underdetermined non-sensical problem, lots of useless global min
- Since U, V full dim, no constraint on X, all the same non-sense global min

What happens when we optimize by gradient descent on U, V ?

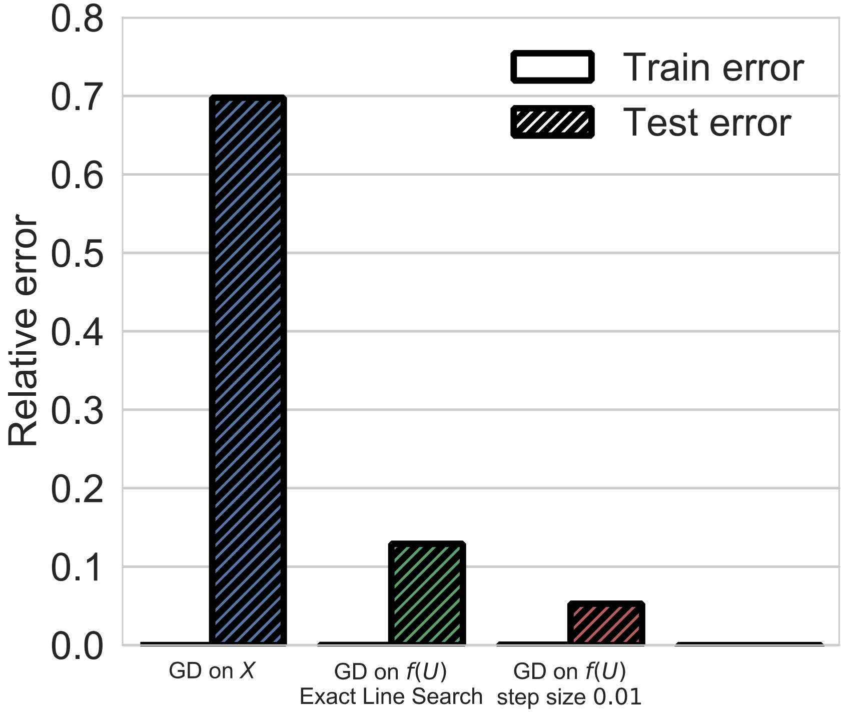
[Gunasekar Woodworth Bhojanapalli Neyshabur S 2017]



$n = 50,\ m = 300,\ A_i$ iid Gaussian, X^* rank-2 ground truth
$y = \mathcal{A}(X^*) + \mathcal{N}(0, 10^{-3}),\ y_{\text{test}} = \mathcal{A}_{\text{test}}(X^*) + \mathcal{N}(0, 10^{-3})$

Gradient descent on $f(U, V)$ gets to "good" global minima

[Gunasekar Woodworth Bhojanapalli Neyshabur S 2017]

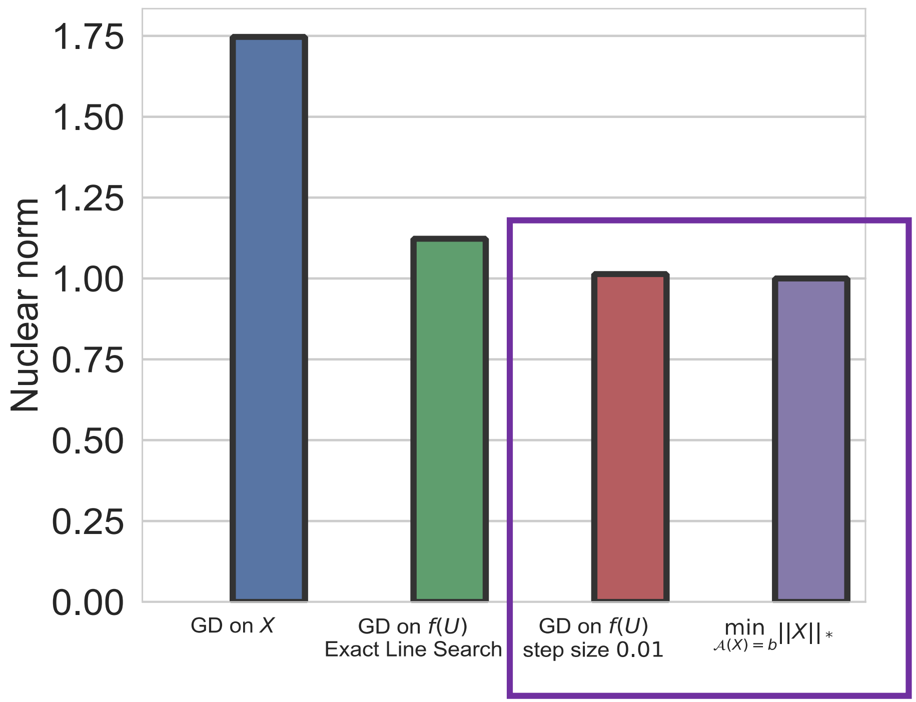


$n = 50,\ m = 300,\ A_i$ iid Gaussian, X^* rank-2 ground truth
$y = \mathcal{A}(X^*) + \mathcal{N}(0, 10^{-3}),\ y_{\text{test}} = \mathcal{A}_{\text{test}}(X^*) + \mathcal{N}(0, 10^{-3})$

Gradient descent on $f(U,V)$ gets to "good" global minima

Gradient descent on $f(U,V)$ generalizes better with smaller step size

[Gunasekar Woodworth Bhojanapalli Neyshabur S 2017]



Grad Descent on U, V with inf. small stepsize and initialization

→ **min nuclear norm solution**

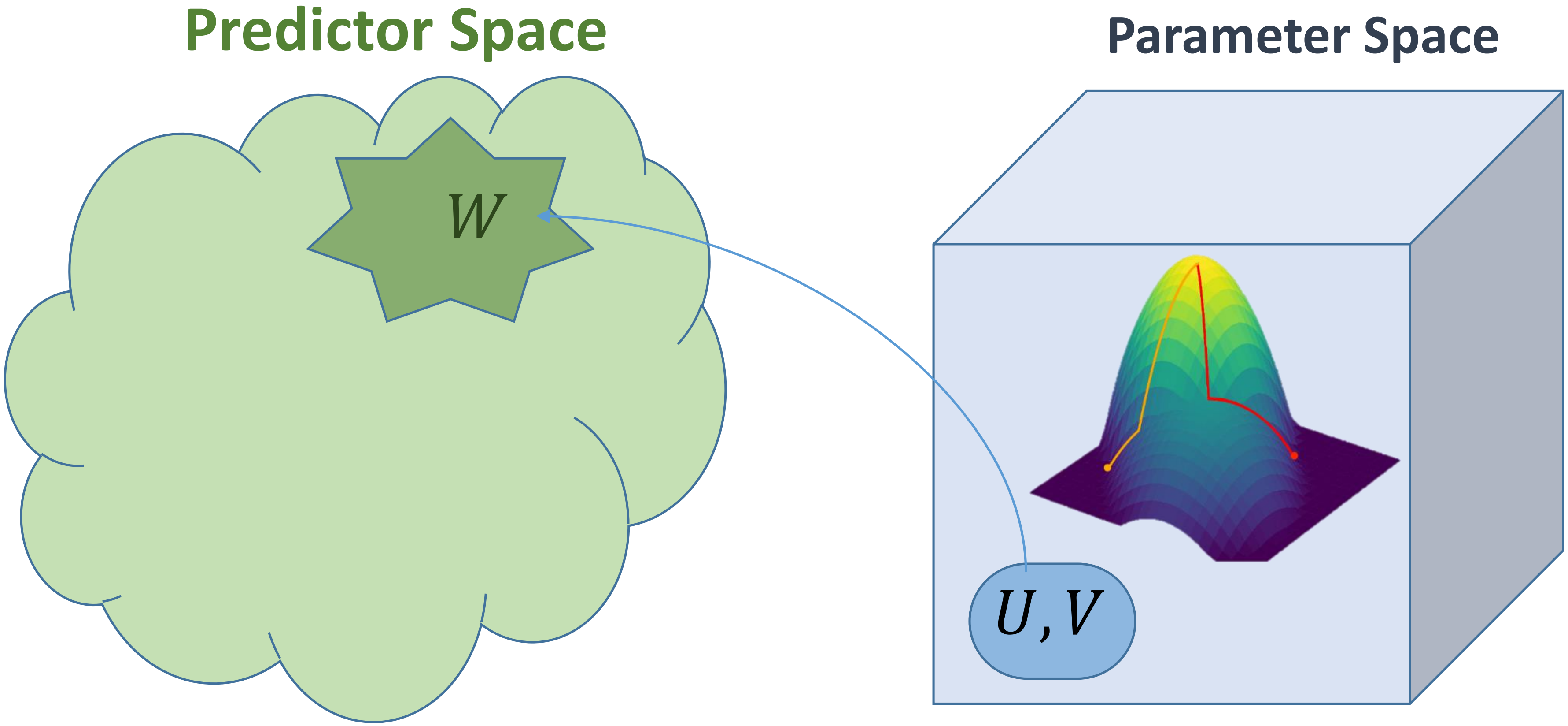
$$\arg\min \|X\|_* \;\; s.t. \;\; obs(X) = y$$

(exact and rigorous only under additional conditions!)

→ good generalization if Y (aprox) low rank

[Gunasekar Woodworth Bhojanapalli Neyshabur S 2017]

[Yuanzhi Li, Hongyang Zhang, Tengyu Ma 2018][Sanjeev Arora, Nadav Cohen, Wei Hu, Yuping Luo 2019]



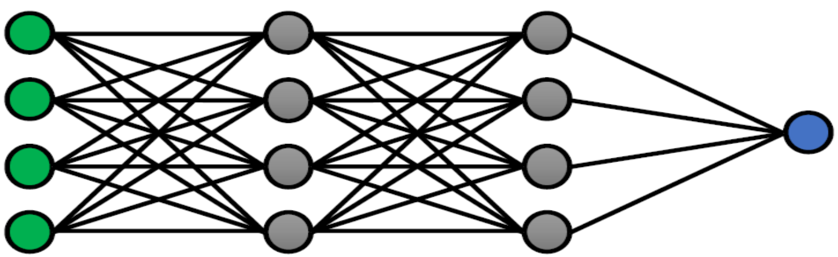
Optimization Geometry and hence Inductive Bias effected by:

- Geometry of local search in parameter space
- Choice of parameterization

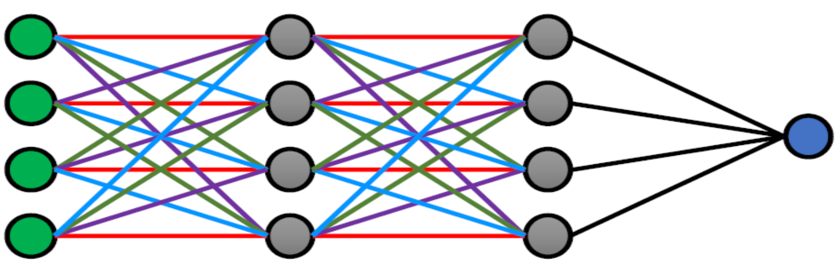
- **Matrix completion** (also: reconstruction from linear measurements)
  - $W = UV$ is over-parametrization of all matrices $W \in \mathbb{R}^{n \times m}$
  - GD on U, V ➔ implicitly minimize $\|\boldsymbol{W}\|_*$

- **Linear Convolutional Network**:
  - Complex over-parametrization of linear predictors β
  - GD on weight ➔ implicitly minimize $\|\boldsymbol{DFT(\beta)}\|_p$ for $p = \frac{2}{depth}$.
  (sparsity in frequency domain)

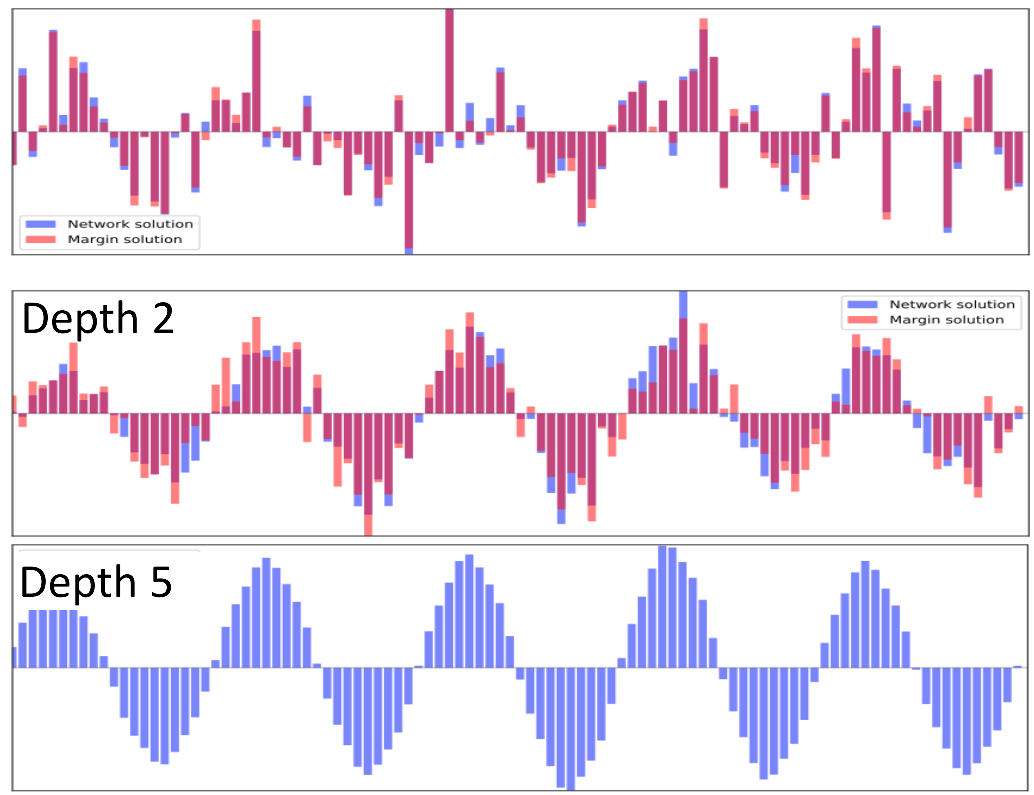
[Gunasekar Lee Soudry S 2018]



$$\min \|\boldsymbol{\beta}\|_2 \; s.t. \forall_i y_i \langle \beta, x_i \rangle \geq 1$$



$$\min \|\boldsymbol{DFT(\beta)}\|_{2/L} \; s.t. \forall_i y_i \langle \beta, x_i \rangle \geq 1$$



- **Matrix completion** (also: reconstruction from linear measurements)
  - $W = UV$ is over-parametrization of all matrices $W \in \mathbb{R}^{n\times m}$
  - GD on U, V ➔ implicitly minimize $\|\boldsymbol{W}\|_*$

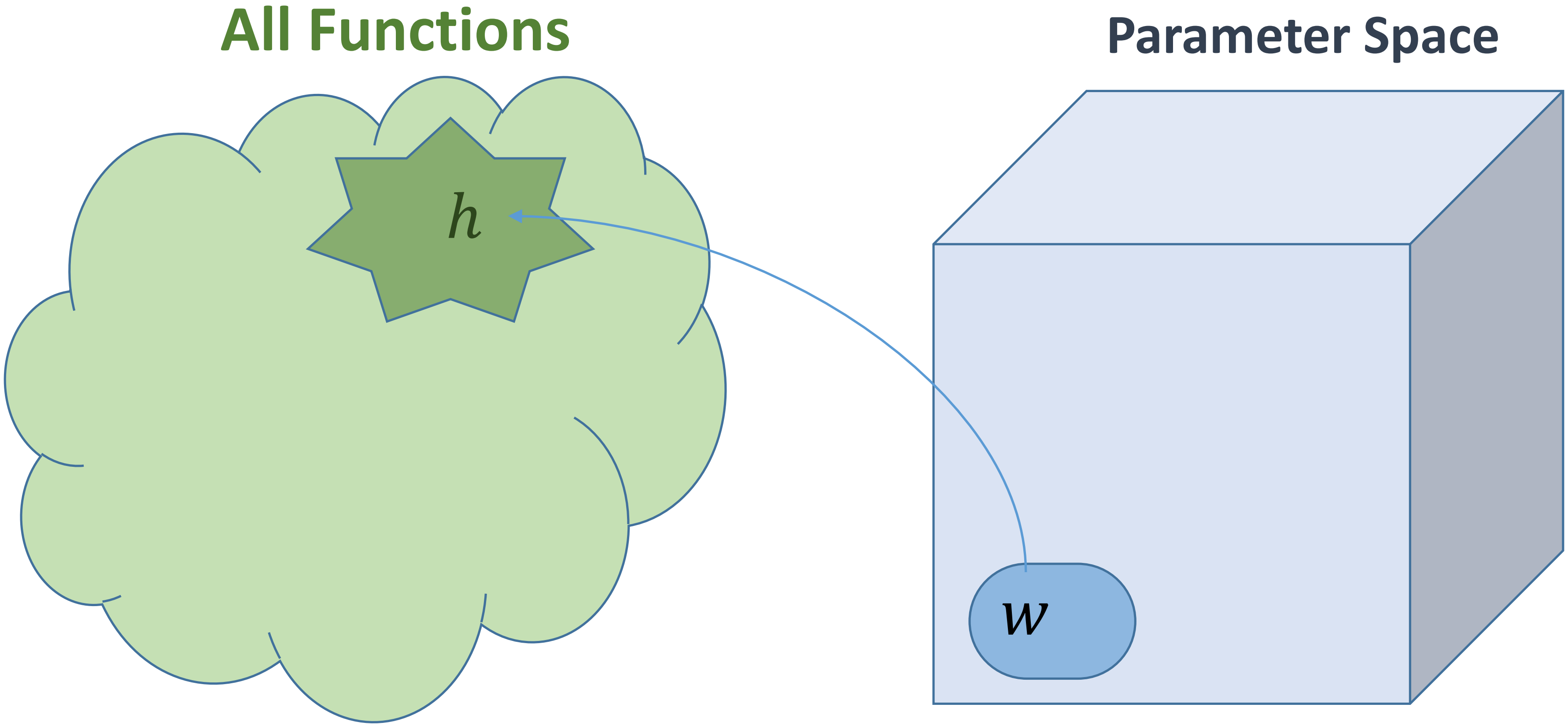
- **Linear Convolutional Network**:
  - Complex over-parametrization of linear predictors β
  - GD on weight ➔ implicitly minimize $\|\boldsymbol{DFT(\beta)}\|_{\boldsymbol{p}}$ for $p = \frac{2}{depth}$.
    (sparsity in frequency domain) [Gunasekar Lee Soudry S 2018]

- **Infinite Width ReLU Net**:
  - Parametrization of essentially all functions $h: \mathbb{R}^d \to \mathbb{R}$
  - Weight decay ➔ for $d = 1$, implicitly minimize

$$\max\left(\int |\boldsymbol{h''}| \boldsymbol{dx} \;, |h'(-\infty) + h'(+\infty)| \right)$$

[Savarese Evron Soudry S 2019]

  - For $d > 1$, implicitly minimize $\int \left|\boldsymbol{\partial_b^{d+1} Radon(h)}\right|$
    (roughly speaking; need to define more carefully to handle non-smoothness, extra correction term for linear part) [Ongie Willett Soudry S 2019]



Optimization Geometry and hence Inductive Bias effected by:

- Geometry of local search in parameter space
- Choice of parameterization

Doesn’t it all boil down to the NTK?

Is it all just a Kernel?

$$f(w,x) \approx f(w^{(0)},x) + \langle w, \phi_{w^{(0)}}(x) \rangle$$

focus on "unbiased initialization": $f(w^{(0)},x) = 0$

$\phi_w(x) = \nabla_w f(w,x)$

Corresponding to a kernelized linear model with kernel:

$$K_w(x,x') = \langle \nabla_w f(w,x), \nabla_w f(w,x') \rangle$$

Kernel regime: 1st order approx about $w^{(0)}$ remains valid throughout optimization

$$K_{w^{(t)}} \approx K_{w^{(0)}} = K_0$$

➔ GD on squared loss converges to $\mathbf{arg\,min} \|h\|_{K_0}$ $s.t.\ h(x_i) = y_i$

Kernel Regime and Scale of Init

- For D-homogenous model, $f(cw,x) = c^D f(w,x)$, consider gradient flow with:

  $\dot{w}_\alpha = -\nabla L_S(w)$ and $w_\alpha(0) = \alpha w_0$ with unbiased $f(w_0,x) = 0$

  We are interested in $w_\alpha(\infty) = \lim_{t\to\infty} w_\alpha(t)$

- For squared loss, under some conditions [Chizat and Bach 18]:

$$\lim_{\alpha\to\infty} \sup_t \left\| w_\alpha\left(\frac{1}{\alpha^{D-1}} t\right) - w_K(t) \right\| = 0$$

Gradient flow of linear least squares w.r.t tangent kernel K_0 at initialization

and so $\boldsymbol{f(w_\alpha(\infty), x) \xrightarrow{\alpha\to\infty} \hat{h}_K(x)}$ where $\boldsymbol{\hat{h}_K = \arg\min \|h\|_{K_0} \ \text{s.t.}\ h(x_i) = y_i}$

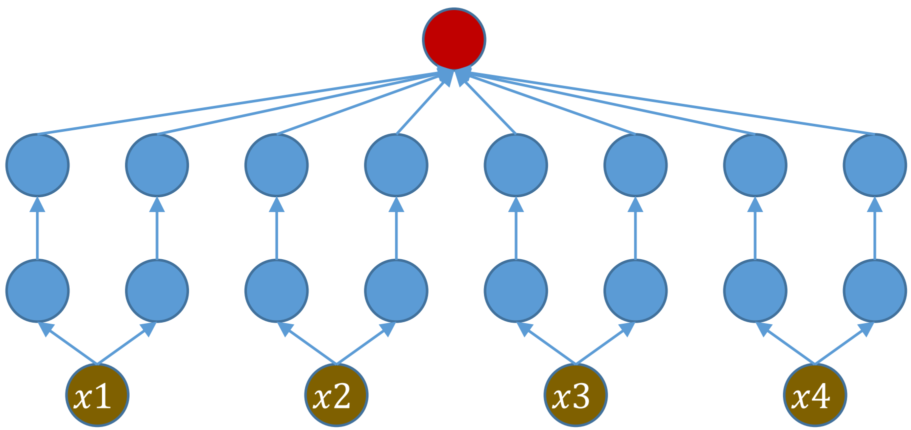
- But: when $\alpha \to 0$, we got interesting, non-RKHS inductive bias (e.g. nuclear norm, sparsity)

Scale of Init: Kernel vs Rich

Consider linear regression with squared parametrization:

$$f(w,x) = \sum_j \left(w_+[j]^2 - w_-[j]^2\right)x[j] = \langle \beta(w), x\rangle \qquad \text{with } \beta(w) = w_+^2 - w_-^2$$

And unbiased initialization $w_\alpha(0) = \alpha\mathbf{1}$ (so that $\beta\left(w_\alpha(0)\right) = 0$).



$$f((U,V),x) = (UV^\top).\begin{bmatrix} diag(x) & 0 \\ 0 & -diag(x) \end{bmatrix}$$

$$U(0) = V(0) = \alpha I$$

Scale of Init: Kernel vs Rich

Consider linear regression with squared parametrization:

$$f(w,x) = \sum_j \left(w_+[j]^2 - w_-[j]^2\right) x[j] = \langle \beta(w), x\rangle \qquad \text{with } \beta(w) = w_+^2 - w_-^2$$

And unbiased initialization $w_\alpha(0) = \alpha\mathbf{1}$ (so that $\beta\left(w_\alpha(0)\right) = 0$).

What's the implicit bias of grad flow w.r.t square loss $L_S(w) = \sum_i (f(w,x_i) - y_i)^2$?

$$\beta_\alpha(\infty) = \lim_{t\to\infty} \beta(w_\alpha(t))$$

In Kernel Regime $\alpha \to \infty$: $K_0(x,x') = 4\langle x, x'\rangle$ and so

$$\beta_\alpha(\infty) \xrightarrow{\alpha\to\infty} \hat{\beta}_{L2} = \arg\min_{X\beta=y} \|\beta\|_{\mathbf{2}}$$

In Rich Regime $\alpha \to 0$: special case of MF with commutative measurements

$$\beta_\alpha(\infty) \xrightarrow{\alpha\to 0} \hat{\beta}_{L1} = \arg\min_{X\beta=y} \|\beta\|_{\mathbf{1}}$$

For any α:

$$\beta_\alpha(\infty) = ???$$

$$\beta(t) = w_+(t)^2 - w_-(t)^2 \qquad L = \|X\beta - y\|_2^2$$

$$\dot{w}_+(t) = -\nabla L(t) = -2X^\top r(t) \circ 2w_+(t) \qquad w_+(t) = w_+(0) \circ \exp\left(-2X^\top \int_0^t r(\tau)\, d\tau\right)$$

$$\dot{w}_-(t) = -\nabla L(t) = +2X^\top r(t) \circ 2w_-(t) \qquad w_-(t) = w_-(0) \circ \exp\left(+2X^\top \int_0^t r(\tau)\, d\tau\right)$$

$$\boxed{\beta(t) = \alpha^2\left(e^{-4X^\top \int_0^t r(\tau)\, d\tau} - e^{4X^\top \int_0^t r(\tau)\, d\tau}\right) \qquad r(t) = X\beta(t) - y}$$

$$s = 4\int_0^\infty r(\tau) d\tau \in \mathbb{R}^m$$

$$\beta(\infty) = \alpha^2\left(e^{-X^T s} - e^{X^\top s}\right) = 2\alpha^2 \sinh X^\top s$$

$$X\beta(\infty) = y$$

$$\min Q(\beta) \quad s.t.\ X\beta = y$$

$$\nabla Q(\beta^*) = X^\top \nu$$

$$X\beta^* = y$$

$$\beta(\infty) = \alpha^2 \left(e^{-X^T s} - e^{X^\top s} \right) = 2\alpha^2 \sinh X^\top s$$

$$X\beta(\infty) = y$$

$$\nabla Q(\beta) = \sinh^{-1}\frac{\beta}{2\alpha^2}$$

$$Q(\beta) = \sum_i \int \sinh^{-1}\frac{\beta[i]}{2\alpha^2} = \alpha^2 \sum_i \left(\frac{\beta[i]}{\alpha^2}\sinh^{-1}\frac{\beta[i]}{2\alpha^2} - \sqrt{4 + \left(\frac{\beta[i]}{\alpha^2}\right)^2} \right)$$

$$\min Q(\beta) \quad s.t.\ X\beta = y$$

$$\nabla Q(\beta^*) = X^\top \nu \qquad \sinh^{-1}\frac{\beta(\infty)}{2\alpha^2} = X^\top s$$

$$X\beta^* = y \qquad X\beta(\infty) = y$$

Scale of Init: Kernel vs Rich

Consider linear regression with squared parametrization:

$$f(w,x) = \sum_j \left(w_+[j]^2 - w_-[j]^2\right) x[j] = \langle \beta(w), x\rangle \qquad \text{with } \beta(w) = w_+^2 - w_-^2$$

And unbiased initialization $w_\alpha(0) = \alpha \mathbf{1}$ (so that $\beta\left(w_\alpha(0)\right) = 0$).

What's the implicit bias of grad flow w.r.t square loss $L_S(w) = \sum_i (f(w,x_i) - y_i)^2$?

$$\beta_\alpha(\infty) = \lim_{t\to\infty} \beta(w_\alpha(t))$$

In Kernel Regime $\alpha \to \infty$: $K_0(x,x') = 4\langle x, x'\rangle$ and so

$$\beta_\alpha(\infty) \xrightarrow{\alpha\to\infty} \hat{\beta}_{L2} = \arg\min_{X\beta=y} \|\beta\|_{\mathbf{2}}$$

In Rich Regime $\alpha \to 0$: special case of MF with commutative measurements

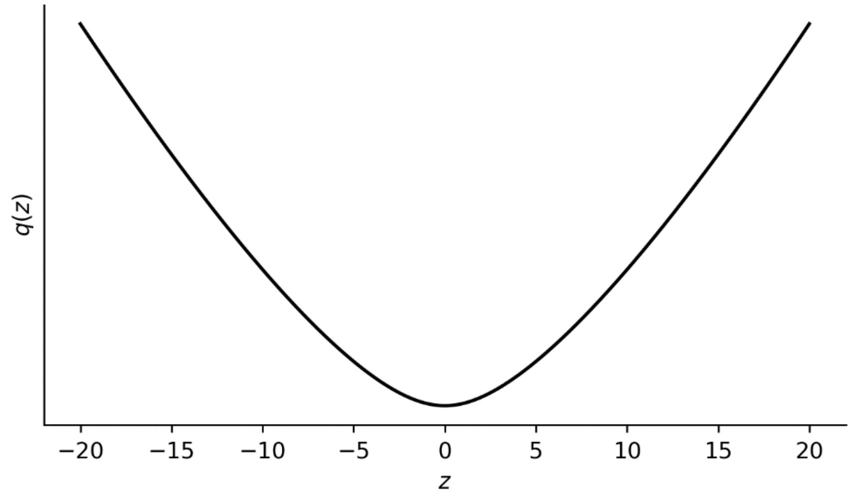
$$\beta_\alpha(\infty) \xrightarrow{\alpha\to 0} \hat{\beta}_{L1} = \arg\min_{X\beta=y} \|\beta\|_{\mathbf{1}}$$

For any $\boldsymbol{\alpha}$: $$\beta_\alpha(\infty) = \boldsymbol{arg}\,\boldsymbol{\min_{X\beta=y}}\, \boldsymbol{Q_\alpha(\beta)}$$

where $\boldsymbol{Q_\alpha(\beta) = \sum_j q\left(\frac{\beta[j]}{\alpha^2}\right)}$ and $\boldsymbol{q(b) = 2 - \sqrt{4+b^2} + b\sinh^{-1}\left(\frac{b}{2}\right)}$

$$\beta_{\alpha}(\infty) = \arg\min_{X\beta=y} Q_{\alpha}(\beta)$$

where $Q_{\alpha}(\beta) = \sum_j q\left(\frac{\beta[j]}{\alpha^2}\right)$ and $q(b) = 2 - \sqrt{4+b^2} + b\sinh^{-1}\left(\frac{b}{2}\right)$



Induced dynamics: $\dot{\beta}_{\alpha} = -\sqrt{\beta_{\alpha}^2 + 4\alpha^4} \odot \nabla L_s(\beta_{\alpha})$

Theorem 2. *For any* $0 < \epsilon < d$,

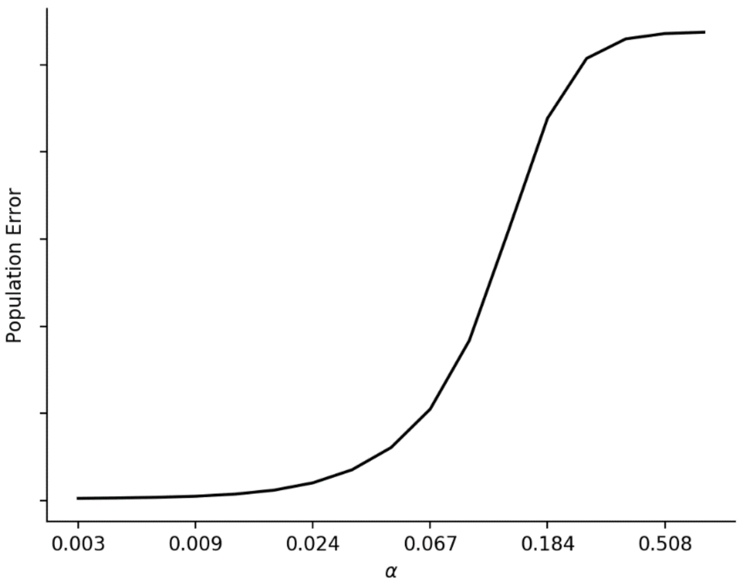
$$\alpha \leq \min\left\{ \left(2(1+\epsilon)\left\|\boldsymbol{\beta}^*_{L1}\right\|_1\right)^{-\frac{2+\epsilon}{2\epsilon}}, \exp\left(-\frac{d}{\epsilon\left\|\boldsymbol{\beta}^*_{L1}\right\|_1}\right)\right\} \implies \left\|\hat{\boldsymbol{\beta}}_{\alpha}\right\|_1 \leq (1+\epsilon)\left\|\boldsymbol{\beta}^*_{L1}\right\|_1$$

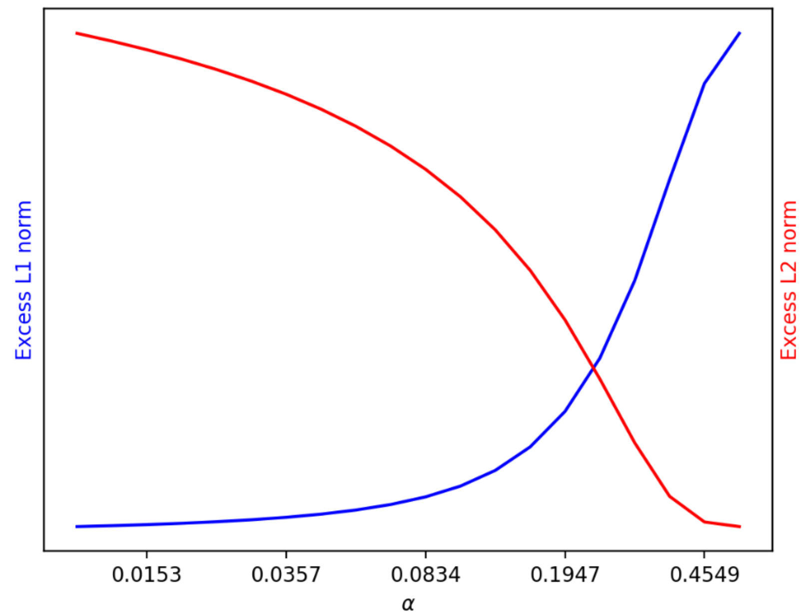
Theorem 3. *For any* $\epsilon > 0$

$$\alpha \geq \sqrt{2(1+\epsilon)\left(1+\frac{2}{\epsilon}\right)\left\|\boldsymbol{\beta}^*_{L2}\right\|_2} \implies \left\|\hat{\boldsymbol{\beta}}_{\alpha}\right\|_2^2 \leq (1+\epsilon)\left\|\boldsymbol{\beta}^*_{L2}\right\|_2^2$$

Sparse Learning

$$y_i = \langle \beta^*, x_i \rangle + N(0{,}0.01)$$
$$d = 1000, \qquad \|\beta^*\|_0 = 5, \qquad m = 100$$

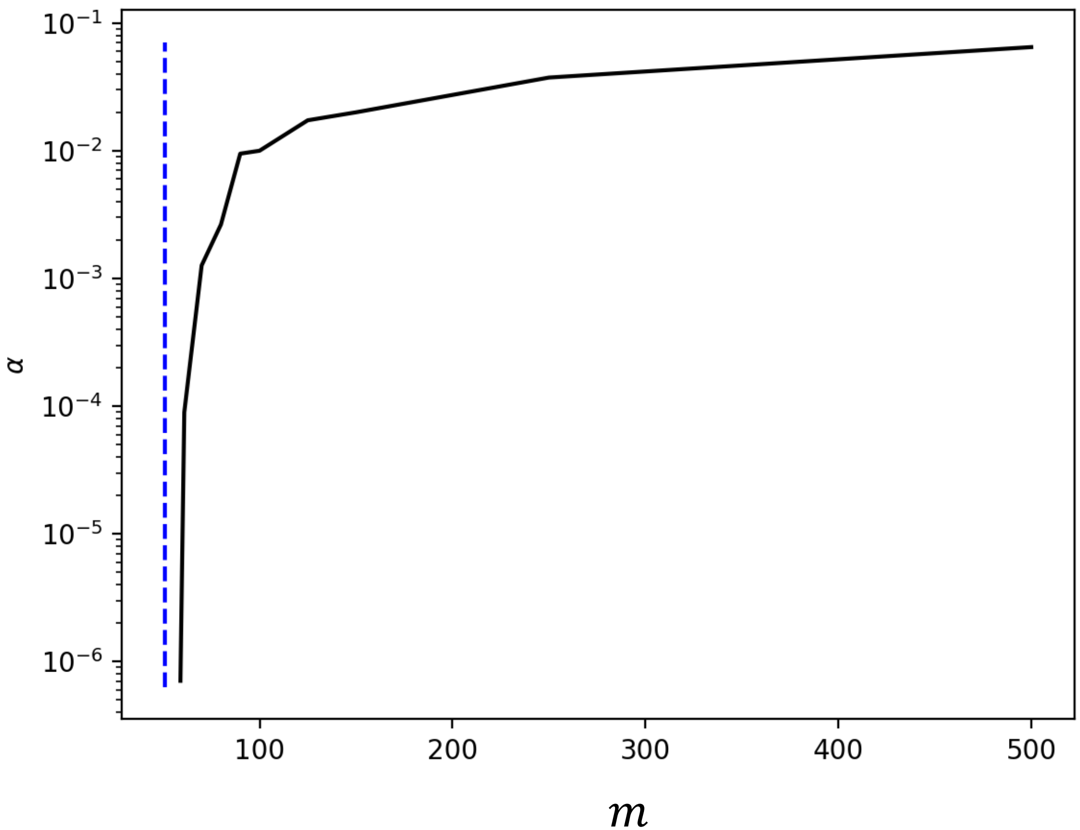




Sparse Learning

$$y_i = \langle \beta^*, x_i \rangle + N(0{,}0.01)$$
$$d = 1000, \qquad \|\beta^*\|_0 = k$$

How small does α need to be to get $L\big(\beta_\alpha(\infty)\big) < 0.025$

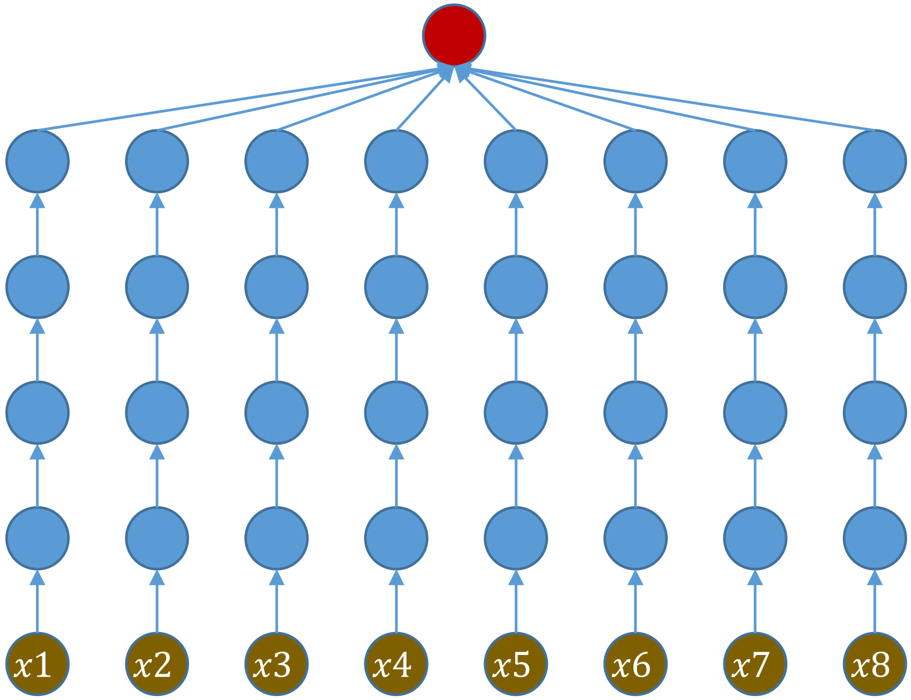


More controlling parameters

- Depth
- Width
- Optimization accuracy
- Stepsize, batchsize, ??

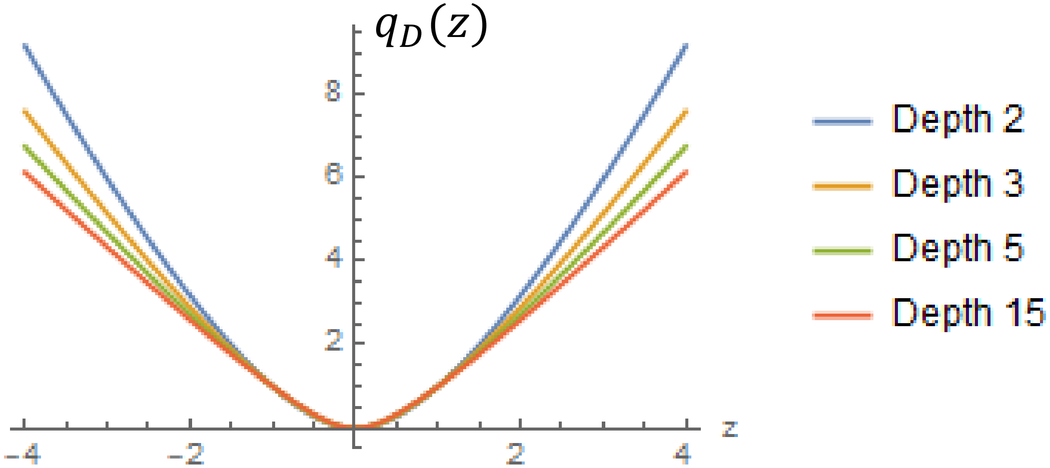
Depth

$$\beta(t) = w_+(t)^D - w_-(t)^D$$



Depth

$\beta(t) = w_+(t)^D - w_-(t)^D \qquad \beta(\infty) = \arg\min Q_D\left({}^{\beta}/_{\alpha^D}\right)\ s.t.\ X\beta = y$



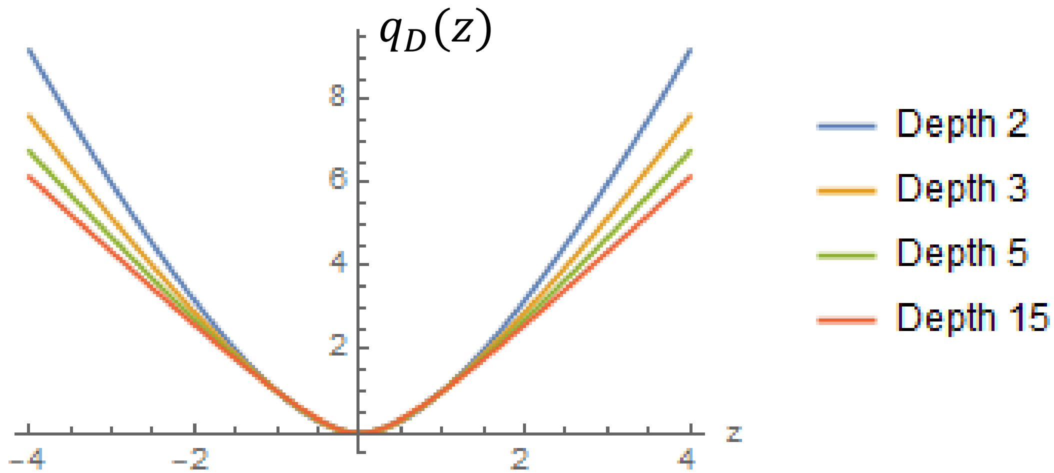
$$h_D(z) = \alpha^D\left(\left(1+\alpha^{D-2}D(D-2)z\right)^{\frac{-1}{D-2}} - \left(1-\alpha^{D-2}D(D-2)z\right)^{\frac{-1}{D-2}}\right)$$

$$q_D = \int h_D^{-1}$$

$$Q_D(\beta) = \sum_i q_D\left(\frac{\beta[i]}{\alpha^D}\right)$$

Depth

$$\beta(t) = w_+(t)^D - w_-(t)^D \qquad \beta(\infty) = \arg\min Q_D\left({}^{\beta}/_{\alpha^D}\right) \;\; s.t.\; X\beta = y$$



For all depth $D \geq 2$, $\beta(\infty) \xrightarrow{\alpha\to 0} \arg\min_{X\beta=y} \|\beta\|_{\mathbf{1}}$

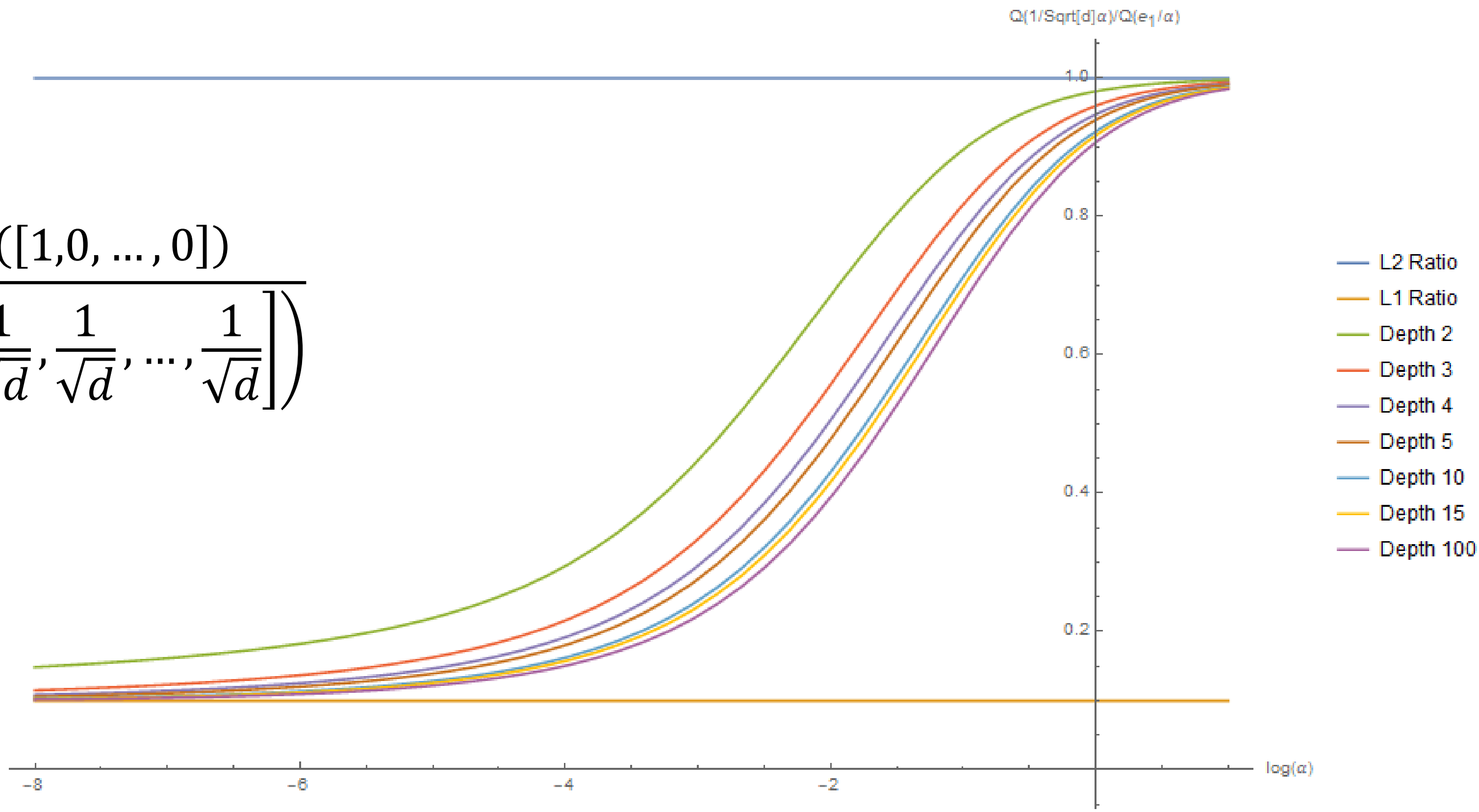
- Contrast with explicit reg: For $R_\alpha(\beta) = \min_{\beta = w_+^D - w_-^D} \|w - \alpha\mathbf{1}\|_2^2$, $\quad R_\alpha(\beta) \xrightarrow{\alpha\to 0} \|\beta\|_{2/_D}$

  also observed by [Arora Cohen Hu Luo 2019]
- Also with logistic loss, $\beta(\infty) \xrightarrow{\alpha\to 0} \propto SOSP\ of\ \|\beta\|_{2/_D}$ [Gunasekar Lee Soudry Srebro 2018]
- With sq loss, always $\|\cdot\|_{\mathbf{1}}$, but for deep D, we get there quicker

Depth

$$\beta(t) = w_+(t)^D - w_-(t)^D \qquad \beta(\infty) = \arg\min Q_D\left(\beta/_{\alpha^D}\right) \;\; s.t.\; X\beta = y$$

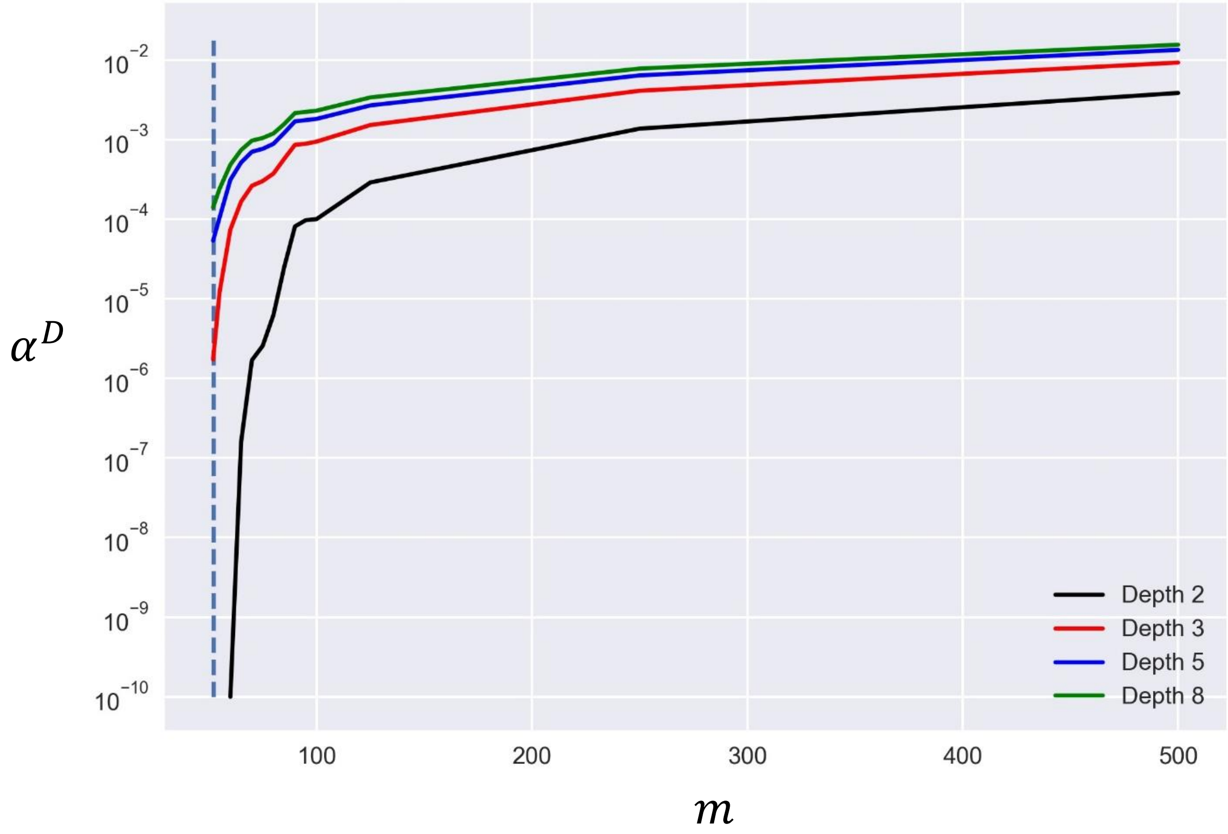
$$\frac{Q([1,0,\dots,0])}{Q\left(\left[\frac{1}{\sqrt{d}},\frac{1}{\sqrt{d}},\dots,\frac{1}{\sqrt{d}}\right]\right)}$$



Sparse Learning with Depth

$$y_i = \langle \beta^*, x_i \rangle + N(0,0.01)$$
$$d = 1000, \qquad \|\beta^*\|_0 = k$$

How small does α need to be to get $L\big(\beta_\alpha(\infty)\big) < 0.025$



Deep Learning

- Expressive Power
  - We are searching over the space of all functions…
    … but with what inductive bias?
  - How does this bias look in function space?
  - Is it reasonable/sensible?
- Capacity / Generalization ability / Sample Complexity
  - What's the true complexity measure (inductive bias)?
  - How does it control generalization?
- Computation / Optimization
  - How and where does optimization bias us? Under what conditions?
  - Magic property of reality under which deep learning "works"