Error and Entropy in the Function Space of Multi-layer Networks

Bo Li (黎勃) and David Saad

22 January 2018

Outline

- Deep learning machines and why entropy in function-space matters
- Statistical mechanics of learning from examples
- Mapping multi-layer networks to dynamical systems - Generating functional analysis
- Continuous and discrete weights – framework and results
- Summary and future work

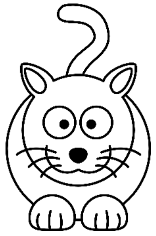
Bo Li and David Saad, arXiv:1708.01422, 2017.

Deep Learning Machines

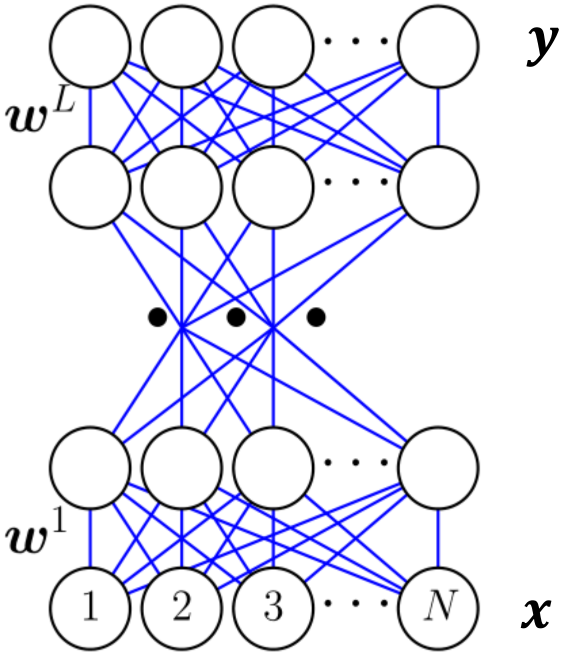
Implement an input-output mapping

$$\boldsymbol{y} = f_{\boldsymbol{w}}(\boldsymbol{x}),$$

where the parameters $\boldsymbol{w}$ are to be estimated based on the training data $\{(\xi^{\mu}, \sigma^{\mu})\}_{\mu=1,2,\ldots P}$ to perform a desired mapping.



$\xrightarrow{f_{\boldsymbol{w}}}$ “cat”



<u>We want to understand:</u>

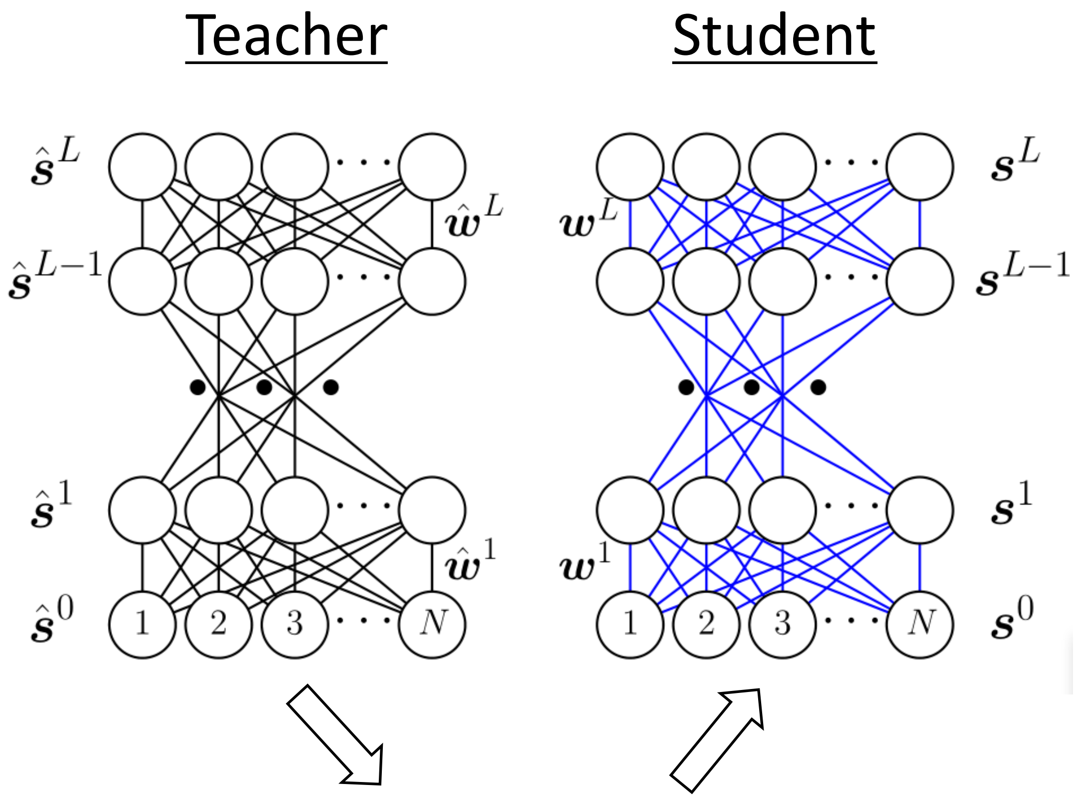
(i) Their **generalization ability** even with numerous parameters
(ii) Nature of the **internal representations**

Macroscopic Analysis –
Typical Behavior of Single Layer Machines

- Mapped to **disordered** systems of **infinite dimension**
- Use **replica analysis**, **cavity method** (batch learning); **dynamical** methods for on-line learning; exploit **high dimensionality**
- **Typical** behavior (in contrast to worse-case) of storage capacity and generalization curves
- Technically quite involved (single or two-layer systems)
- Input data structure and internal representations are rarely addressed

A. Engle and C. Van den Broeck, 2001; D. Saad 1998

Teacher-student Scenario for DLM?



$$\{\boldsymbol{\xi}^1, \sigma^1\}, \{\boldsymbol{\xi}^2, \sigma^2\} \ldots, \{\boldsymbol{\xi}^\mu, \sigma^\mu\}, \ldots, \{\boldsymbol{\xi}^p, \sigma^p\}$$

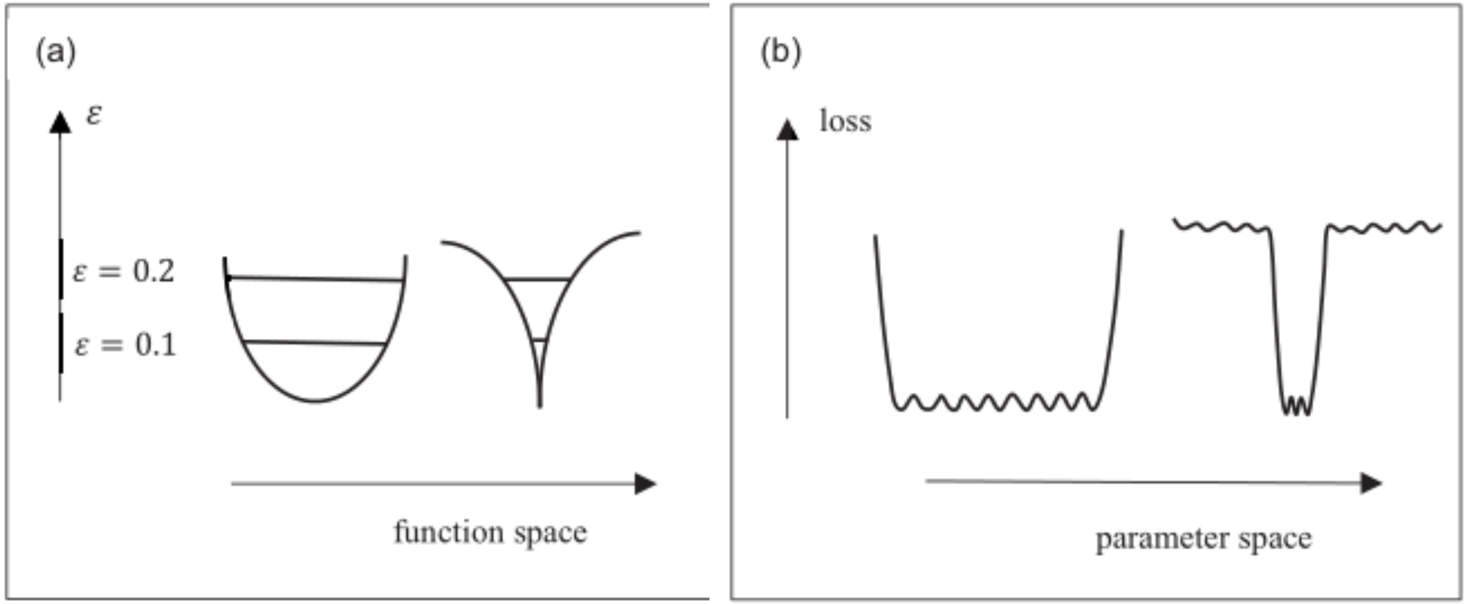
Difficulties:

- Constraints imposed by the examples (input-output pairs) on the hidden units are complex – **recursive nonlinear mapping.**
- Permutation, reflection and other **symmetries/invariances of hidden units**, no simple relation between teacher-student overlap and generalization error.

Especially interesting in the **over-parametrized** regime

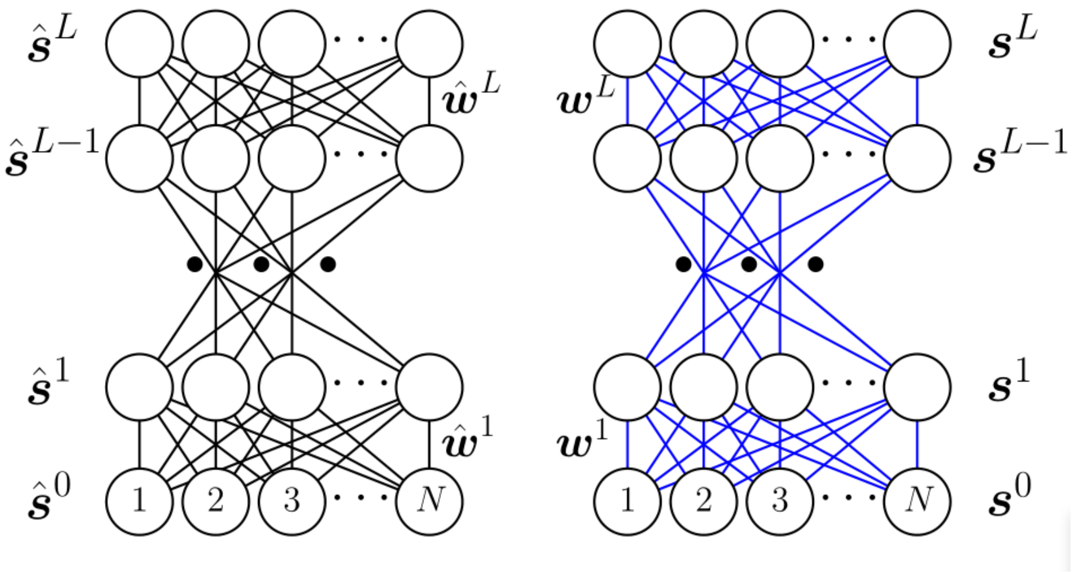
Function Space, Error and Entropy

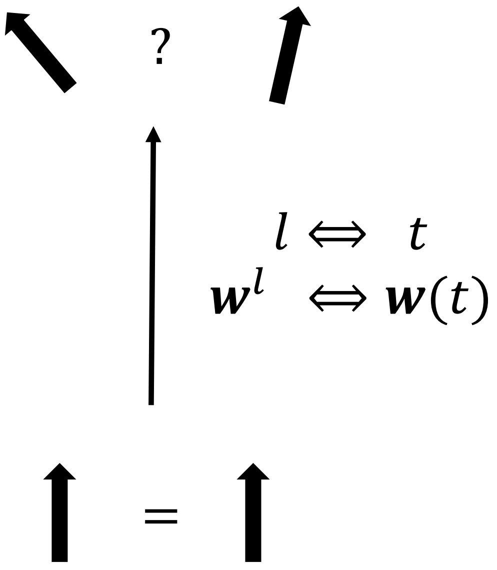
- We would like to approximate a reference/target function $f_{\hat{w}}$, as closely as possible from data.
- Given noisy data, sub-optimal training methods - more relevant to find *good approximations*. **How many such functions exist?**
- The entropy (log-volume) of functions at distance-ε away from $f_{\hat{w}}$ indicates how easy it is obtaining them.



Exploring Function Space in DLM

Reference function $f_{\hat{\boldsymbol{w}}}$ Perturbed function $f_{\boldsymbol{w}}$





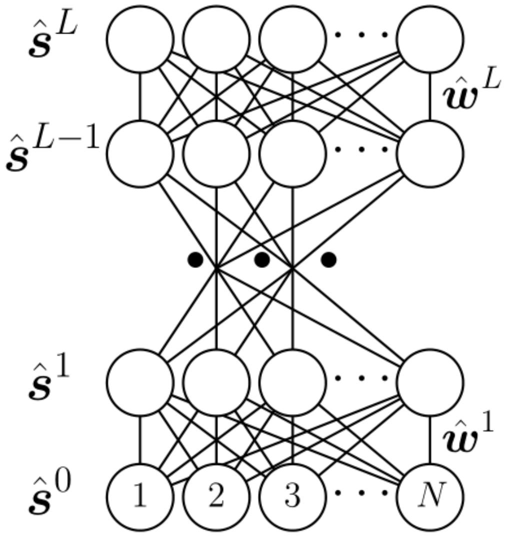
We map the DLM to **disordered spin systems** with discrete dynamics, $\hat{s}_i^l, s_i^l \in \{1, -1\}$, activation function is **sign function** $sgn(x)$.

The framework can be generalized to **real variables** and **other activation functions.**

Investigate the function sensitivity under small perturbations

$$\boldsymbol{w}^l = \mathrm{Perturb}(\hat{\boldsymbol{w}}^l)$$

Deep Learning Machines as Dynamical Systems



$$l \Leftrightarrow t$$
$$\boldsymbol{w}^l \Leftrightarrow \boldsymbol{w}(t)$$

<u>Related work</u>

- Poole et al., NIPS 2016 - Mean field theory to study input sensitivity and expressivity
- Li et al., arXiv:1710.09513, 2017 - Optimal control theory (Pontryagin's maximum principle) to devise new training algorithms

DLM as a Stochastic Dynamical System

- The layer evolution of two coupled DLMs:

$$P\left(\hat{\boldsymbol{s}}^l \middle| \hat{\boldsymbol{w}}^l, \hat{\boldsymbol{s}}^{l-1}, \beta\right) = \prod_i \frac{\exp \beta \hat{s}_i^l \hat{h}_i^l(\hat{\boldsymbol{w}}^l, \hat{\boldsymbol{s}}^{l-1})}{2\cosh \beta \hat{s}_i^l \hat{h}_i^l(\hat{\boldsymbol{w}}^l, \hat{\boldsymbol{s}}^{l-1})}, P\left(\boldsymbol{s}^l \middle| \boldsymbol{w}^l, \boldsymbol{s}^{l-1}, \beta\right) = \cdots,$$

$\hat{h}_i^l(\hat{\boldsymbol{w}}^l, \hat{\boldsymbol{s}}^{l-1}) = \sum_j \hat{w}_{ij}^l \, \hat{s}_j^{l-1}/\sqrt{N},\ \beta$ is the inverse-temperature quantifying the noise level; deterministic rule in the zero-noise limit $\beta \to \infty$.

- Any observable is given by

$$\langle O \rangle := \sum_{\{\hat{\boldsymbol{s}}^l, \boldsymbol{s}^l\}} O \cdot P(\hat{\boldsymbol{s}}^0) \delta_{\hat{\boldsymbol{s}}^0, \boldsymbol{s}^0} \prod_l P\left(\hat{\boldsymbol{s}}^l \middle| \hat{\boldsymbol{w}}^l, \hat{\boldsymbol{s}}^{l-1}, \beta\right) \cdot P(\boldsymbol{s}^l | \boldsymbol{w}^l, \boldsymbol{s}^{l-1}, \beta),$$

summed over all the trajectories subject to the path measure.

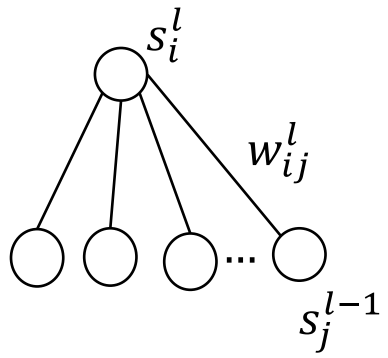
- For discrete spins, the **overlap** between activities of the two systems is of interest

$$q^l(\hat{\boldsymbol{w}}, \boldsymbol{w}, \beta) = \frac{1}{N} \sum_i \langle \hat{s}_i^l s_i^l \rangle$$

Generating Functional Analysis

- Generating functional (characteristic function)

$$\Gamma[\widehat{\boldsymbol{\psi}}, \boldsymbol{\psi}] := \left\langle \exp\left\{-i \sum_{l,i} (\hat{\psi}_i^l \hat{s}_i^l + \psi_i^l s_i^l)\right\}\right\rangle,$$

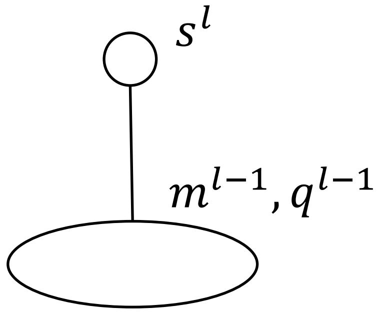


moments such as magnetization $\langle \hat{s}_i^l \rangle$ and overlap $\langle \hat{s}_i^l s_i^l \rangle$ can be obtained by differentiating $\Gamma[\widehat{\boldsymbol{\psi}}, \boldsymbol{\psi}]$; angled brackets – average over all paths.

- Interested in the **typical** behavior of an ensemble of networks $\widehat{\boldsymbol{w}} \sim P(\widehat{\boldsymbol{w}})$, overbar – quenched average

$$\overline{\Gamma[\widehat{\boldsymbol{\psi}}, \boldsymbol{\psi}]} := \sum_{\{\widehat{\boldsymbol{w}}^l, \boldsymbol{w}^l\}} \Gamma[\widehat{\boldsymbol{\psi}}, \boldsymbol{\psi}] P(\widehat{\boldsymbol{w}})\, P(\boldsymbol{w})$$

$$= \int \prod_l \frac{dQ^l dq^l}{2\pi/N} e^{N\Psi[\boldsymbol{q}, \boldsymbol{Q}]} \approx e^{N\Psi[\boldsymbol{q}_e, \boldsymbol{Q}_e]}, \text{ in the limit } N \to \infty,$$



Represented by macroscopic order parameters; the saddle point $\boldsymbol{q}_e, \boldsymbol{Q}_e = \mathrm{extr}_{\boldsymbol{q},\boldsymbol{Q}} \boldsymbol{\Psi}(\boldsymbol{q}, \boldsymbol{Q})$ satisfies certain self-consistent **mean-field** equation.

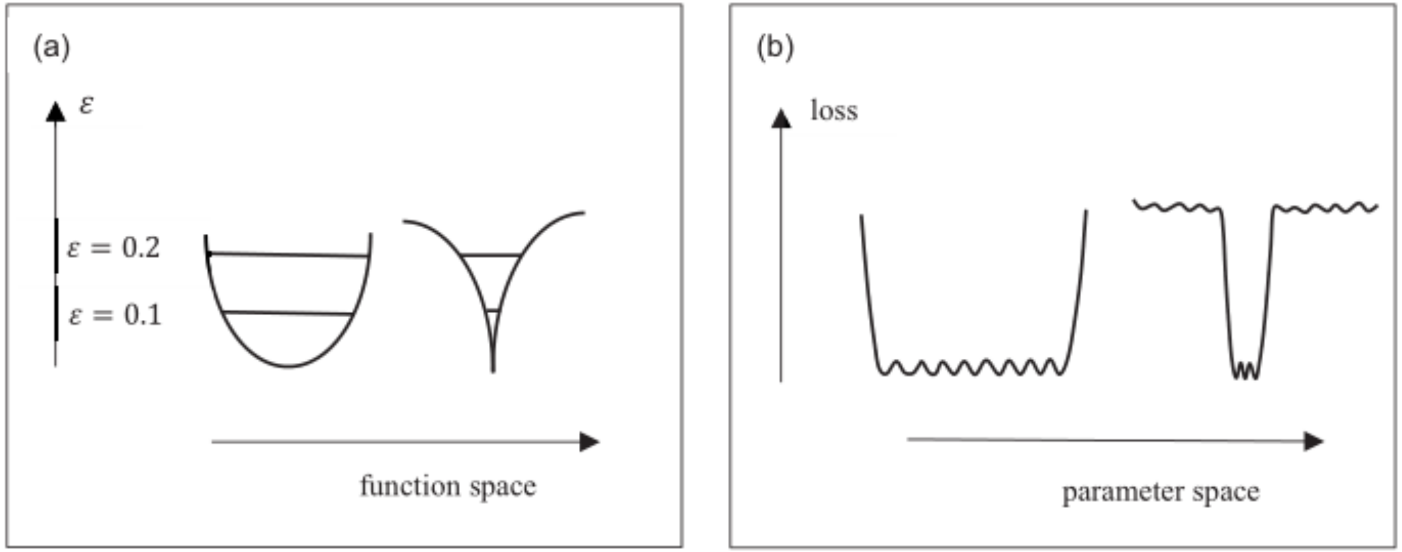
Function Error and Entropy

- Function error is defined as the **expected Hamming distance** of output layers between $f_{\hat{\mathbf{w}}}$ and $f_{\mathbf{w}}$

$$\varepsilon := \frac{1}{2N}\sum_{i=1}^{N}\overline{\langle|\hat{s}_i^L - s_i^L|\rangle} = \frac{1}{2}(1 - q^L),$$

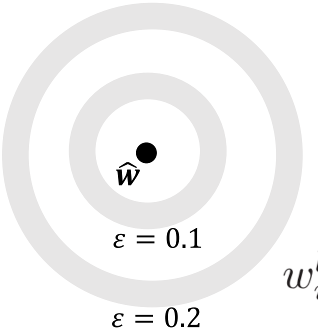
which provides a distance measure between $f_{\hat{\mathbf{w}}}$ and $f_{\mathbf{w}}$.

- Also interested in the entropy (log-volume) of $f_{\mathbf{w}}$ at distance-ε away from the reference function $f_{\hat{\mathbf{w}}}$, e.g.,



Fully-connected Networks – Continuous/Binary weights

Consider fully-connected networks, with $P(\hat{\boldsymbol{w}}_i^l) = \prod_j P(\hat{w}_{ij}^l)$



Continuous weights:

$$\hat{w}_{ij}^l \sim \mathcal{N}(0, \sigma^2)$$

Perturbation strength at layer l

$$w_{ij}^l = \sqrt{1-(\eta^l)^2}\,\hat{w}_{ij}^l + \eta^l \delta w_{ij}^l$$

$$\delta w_{ij}^l \sim \mathcal{N}(0, \sigma^2)$$

Binary weights:

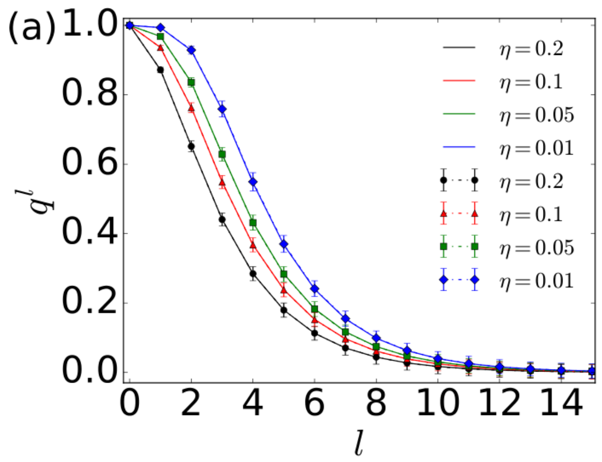
$$P(\hat{w}_{ij}^l) = \frac{1}{2}\delta_{\hat{w}_{ij}^l,1} + \frac{1}{2}\delta_{\hat{w}_{ij}^l,-1}$$

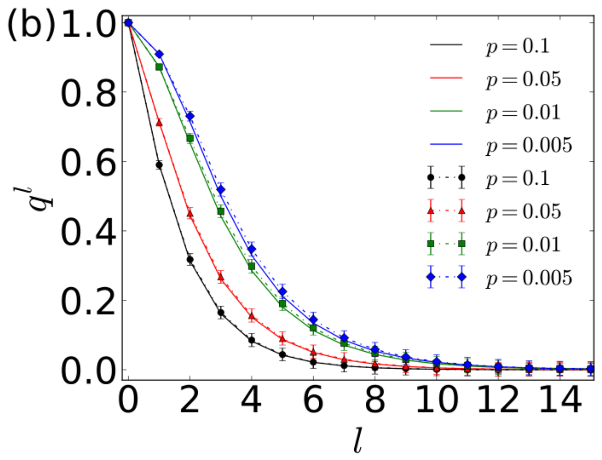
$$P(w_{ij}^l) = (1-p^l)\delta_{w_{ij}^l,\hat{w}_{ij}^l} + p^l \delta_{w_{ij}^l,-\hat{w}_{ij}^l}$$

Typical overlaps:

$$q^l = \frac{2}{\pi}\sin^{-1}\left(\sqrt{1-(\eta^l)^2}\,q^{l-1}\right)$$

$$q^l = \frac{2}{\pi}\sin^{-1}\left((1-2p^l)q^{l-1}\right)$$

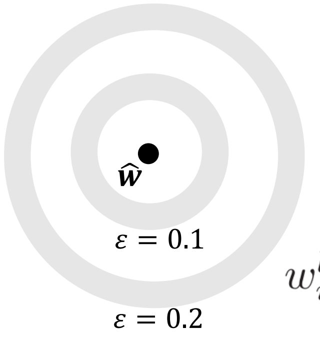




[B. Poole et al.,NIPS 2016]

Entropy of Perturbed Functions

Consider fully-connected networks, with $P(\hat{\boldsymbol{w}}_i^l) = \prod_j P(\hat{w}_{ij}^l)$



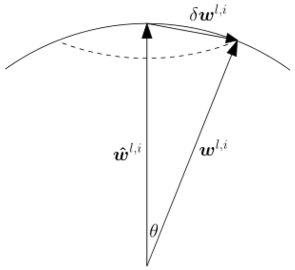
<u>Continuous weights:</u>

$$\hat{w}_{ij}^l \sim \mathcal{N}(0, \sigma^2)$$

Perturbation strength at layer l

$$w_{ij}^l = \sqrt{1 - (\eta^l)^2}\,\hat{w}_{ij}^l + \eta^l \delta w_{ij}^l$$

$$\delta w_{ij}^l \sim \mathcal{N}(0, \sigma^2)$$



<u>Binary weights:</u>

$$P(\hat{w}_{ij}^l) = \tfrac{1}{2}\delta_{\hat{w}_{ij}^l,1} + \tfrac{1}{2}\delta_{\hat{w}_{ij}^l,-1}$$

$$P(w_{ij}^l) = (1 - p^l)\delta_{w_{ij}^l,\hat{w}_{ij}^l} + p^l \delta_{w_{ij}^l,-\hat{w}_{ij}^l}$$

Entropy of $f_{\boldsymbol{w}}$: $\quad \mathcal{S}_{\mathrm{con}}(\{\eta^l\}) = \frac{1}{L}\sum_{l=1}^{L} \log \eta^l \qquad \mathcal{S}_{\mathrm{bin}}(\{p^l\}) = \frac{1}{L}\sum_{l=1}^{L} -p^l \log p^l - (1 - p^l)\log(1 - p^l)$

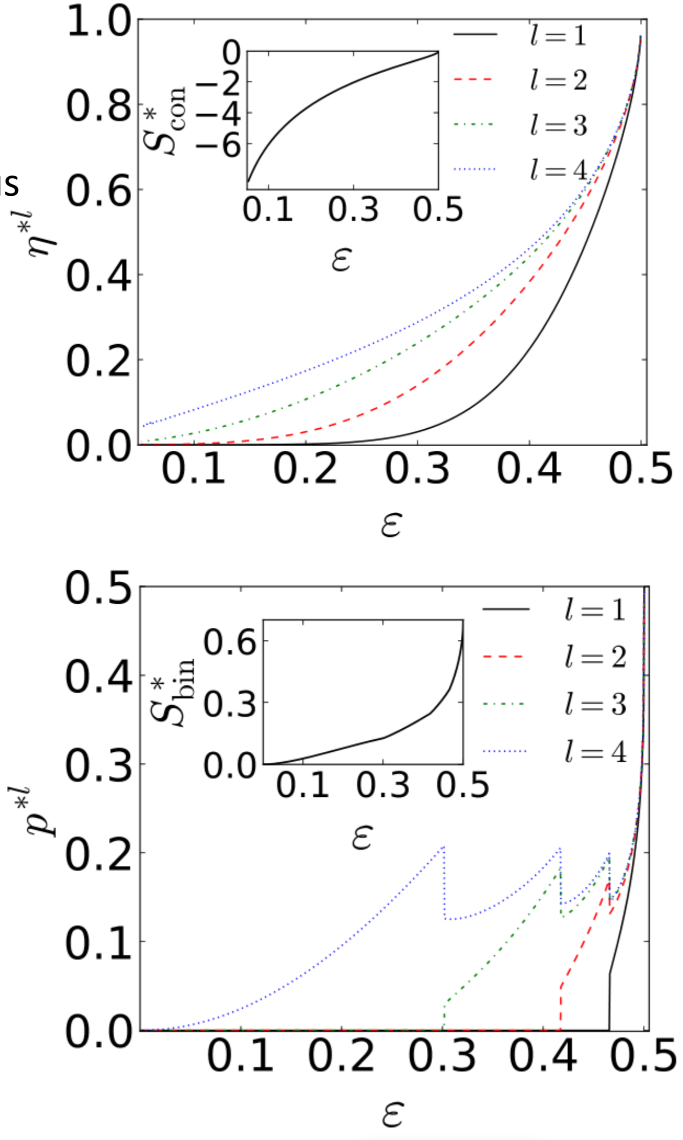
The distance-ε surface of $f_{\boldsymbol{w}}$ with volume $\Omega(\{\eta^l\}) = \exp N_p\, S_{\mathrm{con}}(\{\eta^l\})$, is **exponentially** dominated by the maximum-entropy solutions when $N_p \to \infty$:

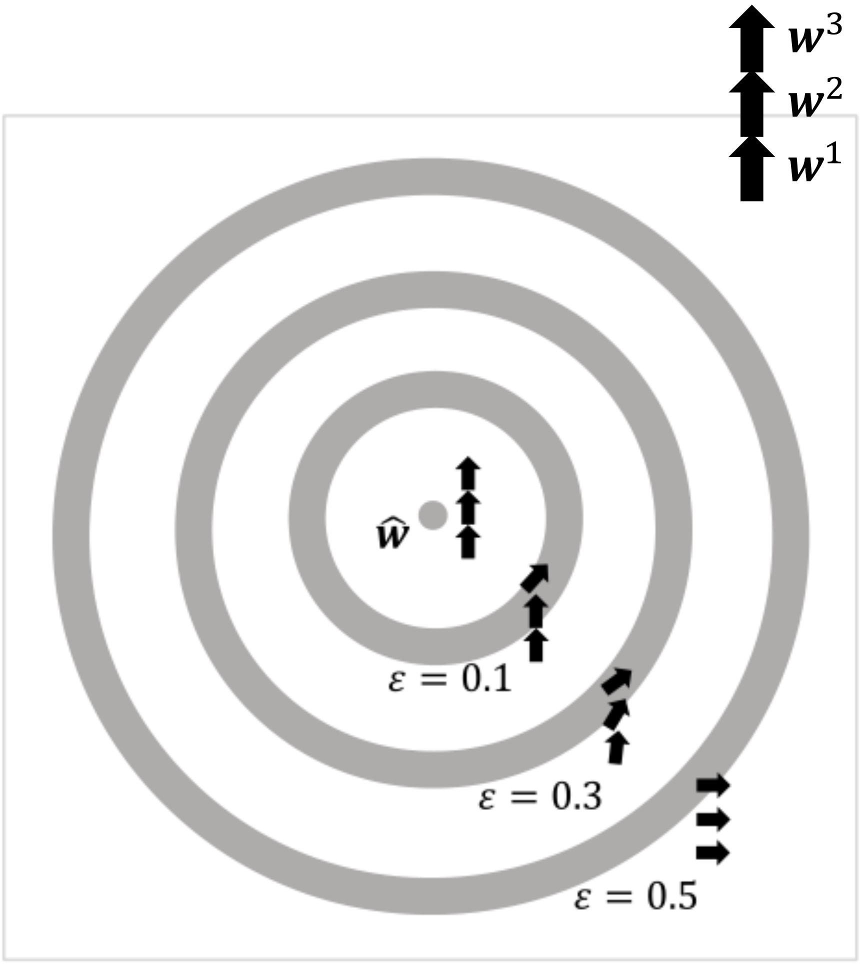
$$\eta^{*l} = \arg\max_{\eta^l} S_{\mathrm{con}}(\{\eta^l\}), \quad \text{s.t. } q^L(\{\eta^l\}) = 1 - 2\varepsilon$$

Earlier Layers Converge First When Decreasing ε

Continuous weights

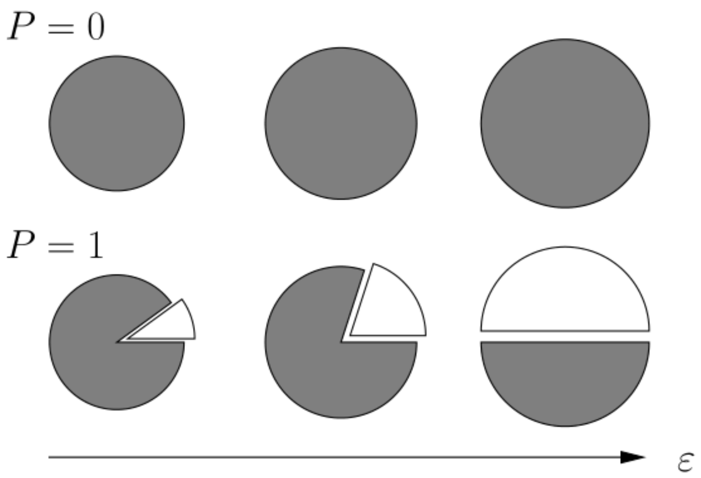
Binary weights





M. Raghu et al, *ICML* 2017

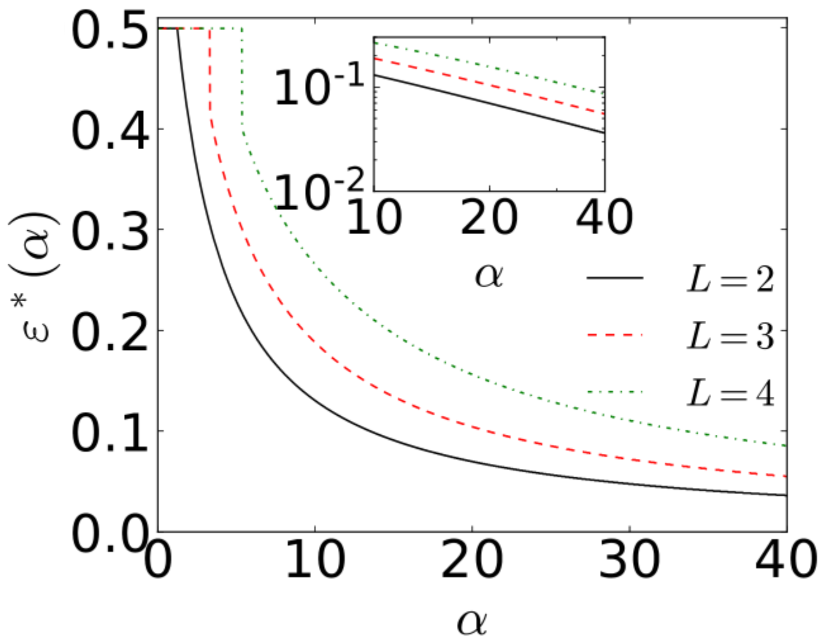
Approximated Generalization Curve (dense DLM with continuous weights)



$$\Omega_\alpha(\varepsilon) = \Omega_{\rm tot}(\{\eta^{*l}\}, \varepsilon)(1-\varepsilon)^P$$

$$P = \alpha L N^2 \qquad \varepsilon^*(\alpha) = \mathrm{argmax}_\varepsilon \Omega_\alpha(\varepsilon)$$

Annealed theory of learning



Relevant in small ε (large α) limit.

A. Engle and C. Van den Broeck, 2001

Sparsely Connected Binary Networks

- Same architecture as before, except that each node is **randomly** connected to k units in the previous layer and $\hat{w}_{ij}^{l} = 1$.

  Such layered networks can implement a large class of Boolean functions.

  [A. Mozeika and D. Saad PRL 2009]

- In addition to the overlap q^l the magnetization $m^l := 1/N \sum_i \langle s_i^l \rangle$ order parameter characterizes the macroscopic dynamics.

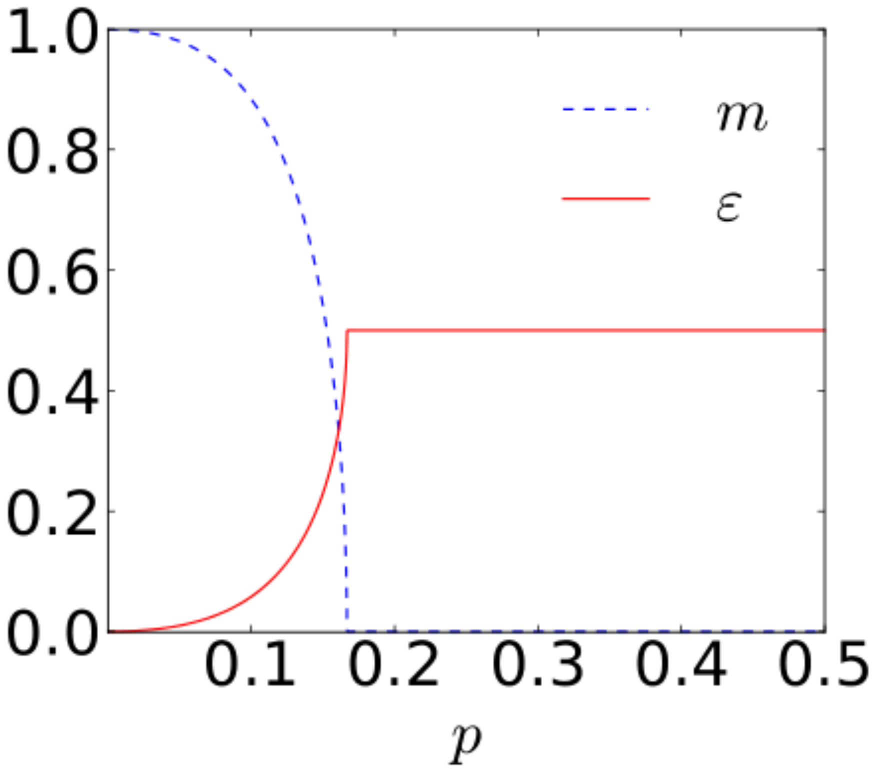
$$m^l = \sum_{\{s_j\}} \prod_{j=1}^{k} \frac{1}{2} \left[1 + s_j m^{l-1}(1-2p)\right] \operatorname{sgn}\left[\sum_{j=1}^{k} s_j\right]$$

$$q^l = \sum_{\{s_j, \hat{s}_j\}} \prod_{j=1}^{k} \frac{1}{4} \left[1 + \hat{s}_j \hat{m}^{l-1} + s_j m^{l-1}(1-2p)\right.$$

$$\left. + s_j \hat{s}_j q^{l-1}(1-2p)\right] \operatorname{sgn}\left[\sum_{j=1}^{k} \hat{s}_j\right] \operatorname{sgn}\left[\sum_{j=1}^{k} s_j\right]$$

Sparsely Connected Binary Networks

For the same for perturbation at every layer $p^l = p$ (weight flipping probability) and at the infinite depth limit $L \to \infty$



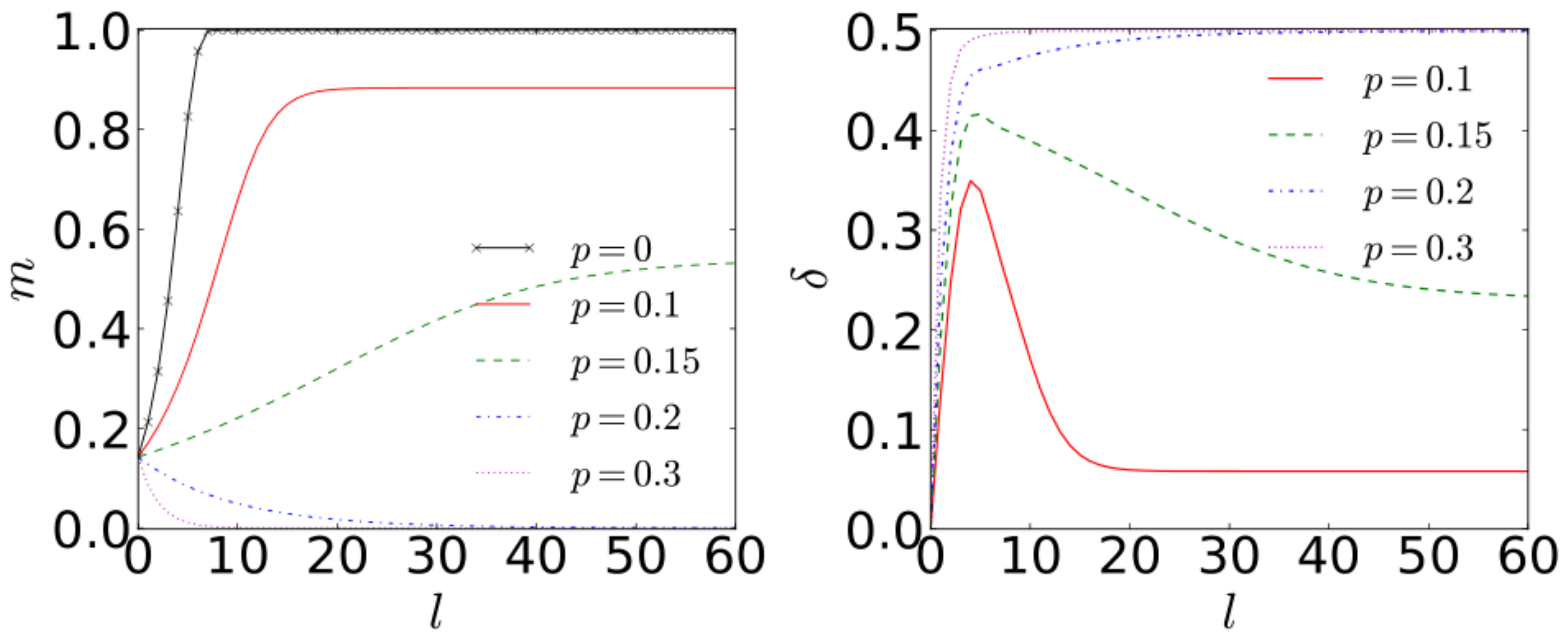
Reference function $f_{\hat{\mathbf{w}}} = sgn(\sum \hat{s}_i^0)$ is majority vote of input.

Phase transition of **stationary states** as $L \to \infty;\ k = 3, m^0 > 0.$
($m^\infty < 0$ if $m^0 < 0$)

Similar to the phase transition of varying thermal noise β in the noisy computation setting.

[A. Mozeika and D. Saad PRL 2009]

Deep Layers For Reliable Computation in Sparse Binary Deep Learning Machines

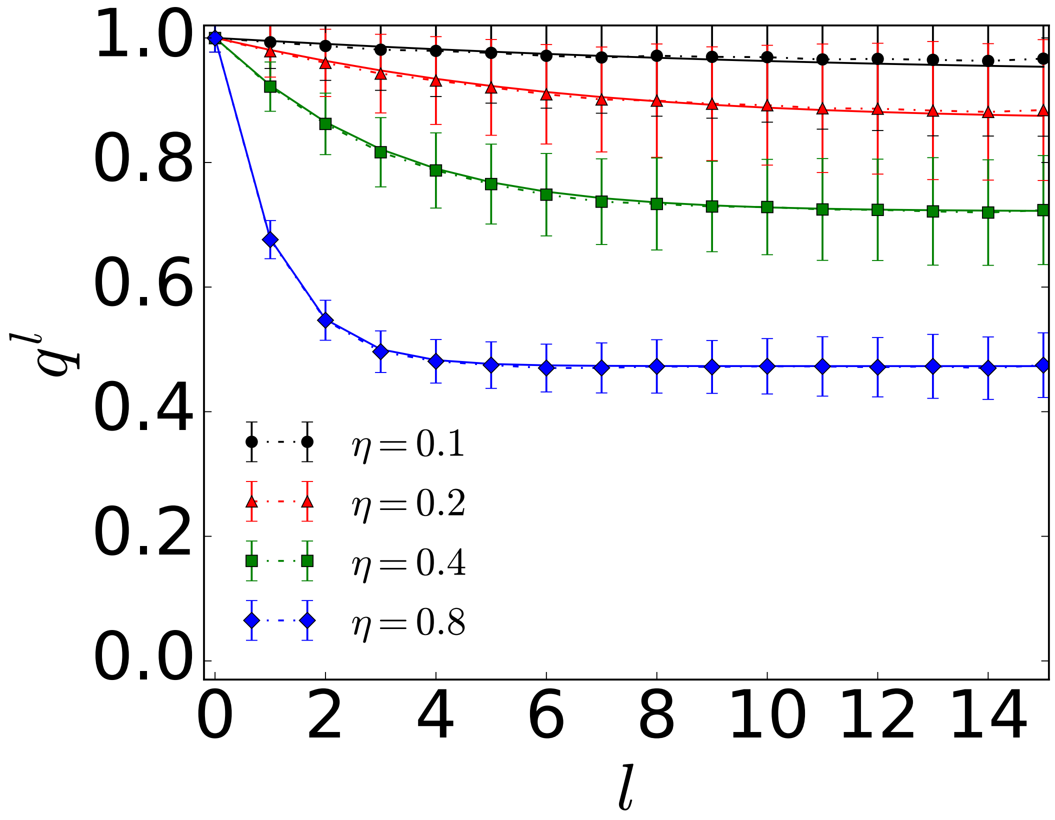


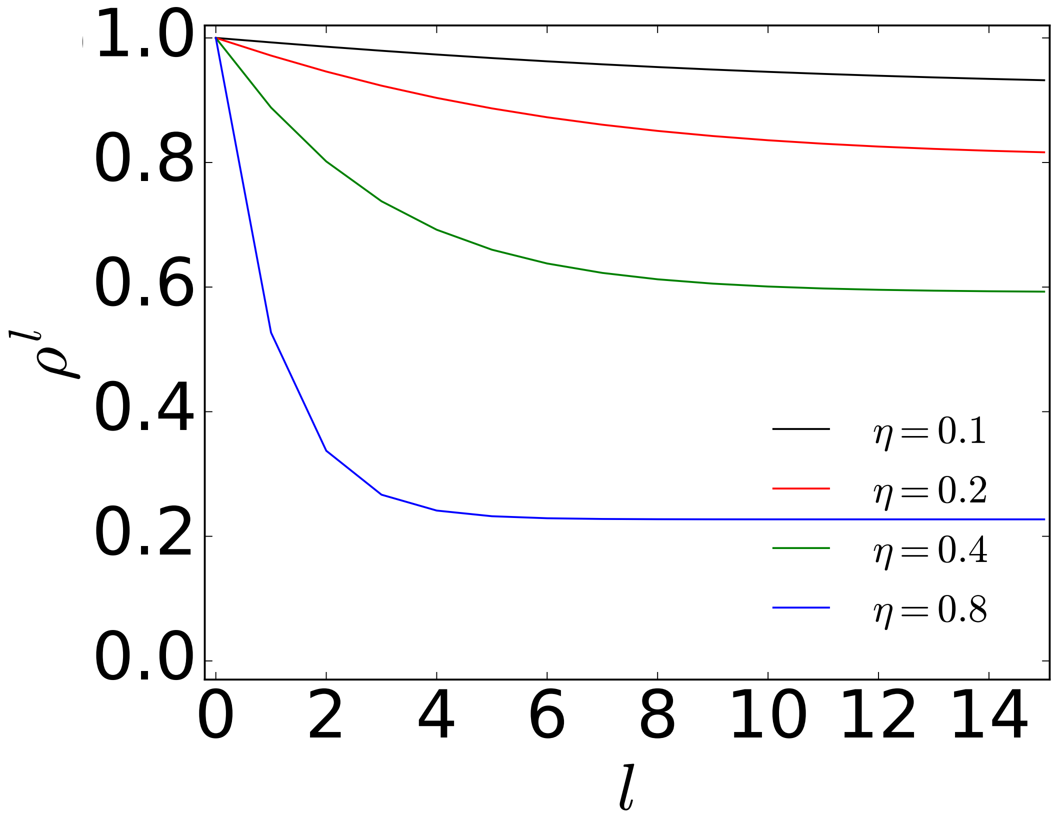
$$\delta^l := \frac{1}{2}(1 - q^l)$$

Internal error of activations

Continuous Variables and Activation Functions

We extended the framework for the ReLu function $\phi(x) = \begin{cases} x & x > 0 \\ 0 & x \leq 0 \end{cases}$





Summary and Future Work

Summary:

- The **generating functional analysis** is a principle method for deriving the **typical** behaviors of DLMs; it allows for non-trivial extensions.
- **Layer-by-layer matching** of weights is observed when getting closer to the reference function in densely-connected networks.
- Sparsely connected networks favor **deep layers** for a reliable representation.

Future work:

- A model learned from data $p(\boldsymbol{w}|D)$ – Typically many **redundant** weights.
- Exploring the function space for **correlated inputs**.
- The role of **over-parametrization** in the function landscape, error and generalization.
- Other models? Optimizing **variable hidden layer size**.
- **Noisy** training/computation for better generalization.