Département de Physique
École normale supérieure

FONDATION CFM POUR LA RECHERCHE

SPHINX

Laboratoire de Physique Statistique – ENS

Entropy and mutual information in models of deep neural networks

Statistical Physics and Machine Learning Back Together
Cargèse, Corsica – August 29th 2018

Marylou Gabrié (LPS ENS), Andre Manoel (INRIA Saclay),
Clément Luneau, Jean Barbier, Nicolas Macris (EPFL),
Lenka Zdeborová (CEA Saclay) & Florent Krzakala (LPS ENS)

Computing information theoretic quantities is a hard problem

Typical setting:

- Random variables of high dimension $\mathbf{X} \in \mathbb{R}^N$, N very large
- Finite number of samples $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(K)}$
- Unknown underlying distribution $p(\mathbf{x})$

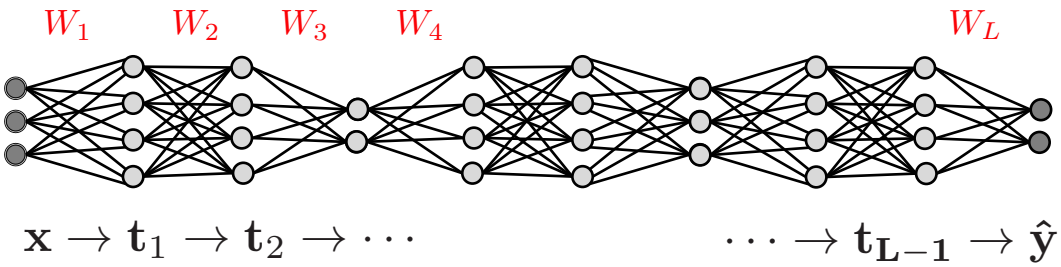
How to compute the entropy ? $H(\mathbf{X}) = -\int \mathrm{d}x_1 \mathrm{d}x_2 \cdots \mathrm{d}x_N \; p(\mathbf{x}) \log p(\mathbf{x})$

- **Small dimension:**
  - discretization (binning)
  - non parametric estimates (based on pairwise distances between samples)
  - numerical integration

- **High dimension:**
  - empirical: estimation from samples unreliable ...
  - known distribution: numerical integration way too slow (exponentially)

"Curse of the dimensionality"

Information theory on stochastic deep networks



Random variables of interest:

- input layer (empirical) $\mathbf{x} \sim p(\mathbf{x}) = \frac{1}{K} \sum_{k=1..K} \delta(\mathbf{x} - \mathbf{x}^k)$
- stochastic layers variables $\mathbf{t}_\ell = f(W^{(\ell)} \mathbf{t}_{\ell-1}; \epsilon)$ changing with weights learning !

Study their relationship through mutual information:

$$I(\mathbf{X}; \mathbf{T}_\ell) = \int \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t}_\ell \, p(x, \mathbf{t}_\ell) \log \frac{p(\mathbf{x}, \mathbf{t}_\ell)}{p(\mathbf{x})p(\mathbf{t}_\ell)} = KL[p(\mathbf{x}, \mathbf{t}_\ell) || p(x)p(\mathbf{t}_\ell)]$$

Are there cases in which we can estimate MI for large networks ?

Analytical formula of mutual information
Linear DNN with Gaussian data and Gaussian noise

Given ...

- Synthetic Gaussian input data $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}\,;\,0, I_N)$

- Linear activations and Gaussian additive noise $\mathbf{t}_\ell = W^{(\ell)} \ldots W^{(1)} \mathbf{x} + \epsilon$
  $\epsilon \sim \mathcal{N}(0, \Sigma)$

Then ...

- Analytical expression of mutual information (Pinsker 1964)

$$I(\mathbf{T}_\ell; \mathbf{X}) = \frac{1}{2} \ln \left[(2\pi e)^{N_\ell} |W^{(\ell)} \ldots W^{(1)} W^{(1)^T} \ldots W^{(\ell)^T} + \Sigma| \right] - \frac{1}{2} \ln \left[(2\pi e)^{N_\ell} |\Sigma| \right]$$

A relevant setting to study deep learning ?

✓ **Large networks**

✗ **Non linear networks**

✗ **Arbitrary dataset (only Gaussian !)**

Outline

1 – Replica formula for multilayer networks

2 – Entropies in neural networks with random weights

3 – Entropies in deep learning

Previous works

- **Single layer Generalized Linear Models (GLMs) with i.i.d measurement matrices**

  $$\mathbf{y} = f(W\mathbf{x}) \text{ with } W_{ij} \text{ i.i.d}$$

  - Statistical physics community from 80's to now (Derrida, Gardner, Mézard, Tanaka, ...)
  - Rigorous works (Reeves et al 2017, Barbier et al 2017 (cf Léo Miolane's talk), ...)

This talk: extension to multilayer

- **Perceptron (single layer) with orthogonally invariant input patterns**
  (Kabashima and Shinzato 2008, 2009)

$$y = \mathrm{sgn}\left(\frac{1}{\sqrt{N}}\vec{x}^{\mathrm{T}}\vec{w}\right) \qquad X = UDV^{\mathrm{T}}$$

- **Multilayer GLMs**
  - ML-AMP algorithm and free energy (i.i.d weight matrices)
    (A. Manoel, F. Krzakala, M. Mézard, L. Zdeborová 2017)
  - ML-VAMP algorithm (orthogonallly invariant weight matrices)
    (A. K. Fletcher , P. Schniter, S. Rangan 2017)
  - Free energy and mutual information (orthogonally invariant weights), heuristic alternative to replicas (G. Reeves 2017)

Extension of the replica formula for entropies in multilayer networks

Setting: For multi-layer neural network with $\mathbf{X} \to \mathbf{T}_1 \to \mathbf{T}_2 \to \cdots \to \mathbf{T}_L$

- factorized input distribution $P_X(\mathbf{x}) = \prod_i P_0(x_i)$
- specified transition probabilities $\mathbf{t}_\ell \sim P_\ell(\mathbf{t}_\ell | W^{(\ell)} \mathbf{t}_{\ell-1})$
- weight matrices 1) orthogonally-invariant
  2) independent of each other
  3) aspect ratio (# cols / # rows) of order 1

$$W^{(\ell)} = U_\ell \, S_\ell \, V_\ell$$

diagonal

orthogonal
Haar distribued

Result: The entropy of each layer in the thermodynamic limit is given by the minimum among all extrema of the replica symmetric potential

$$\lim_{N_0 \to \infty} \frac{1}{N_0} H(\mathbf{T}_\ell) = \min \operatorname*{extr}_{\mathbf{A},\mathbf{V},\tilde{\mathbf{A}},\tilde{\mathbf{V}}} \phi_\ell(\mathbf{A}, \mathbf{V}, \tilde{\mathbf{A}}, \tilde{\mathbf{V}})$$

M. G., A. Manoel, C. Luneau, J. Barbier, N. Macris, F. Krzakala, L. Zdeborová - arXiv:1805.09785

Replica symmetric potential

$$\phi_\ell(\mathbf{A},\mathbf{V},\tilde{\mathbf{A}},\tilde{\mathbf{V}}) = I\left(t_0; t_0+\frac{\xi_0}{\sqrt{\tilde{A}_1}}\right) - \frac{1}{2}\sum_{k=1}^{\ell}\tilde{\alpha}_{k-1}\left[\tilde{A}_k V_k + \alpha_k A_k \tilde{V}_k - F_{W_k}(A_k V_k)\right]$$
$$+ \sum_{k=1}^{\ell-1}\tilde{\alpha}_k\left[H(t_k|\xi_k;\tilde{A}_{k+1},\tilde{V}_k,\tilde{\rho}_k) - \frac{1}{2}\log(2\pi e\tilde{A}_{k+1})\right] + \tilde{\alpha}_\ell H(t_\ell|\xi_\ell;\tilde{V}_\ell,\tilde{\rho}_\ell),$$

with

$$\tilde{\alpha}_k = N_k/N_0 \qquad \rho_k = \int dP_{k-1}(t)\, t^2 \qquad \xi_k \sim \mathcal{N}(0,1)$$

$$\alpha_k = N_k/N_{k-1} \qquad \tilde{\rho}_k = (\mathbb{E}_{\lambda_{W_k}}\lambda_{W_k})\rho_k/\alpha_k \qquad \tilde{\xi},\tilde{z} \sim \mathcal{N}(0,1)$$

where

$$P(t_0) = P_0(t_0)$$
$$P(t_k|\xi_k;A,V,\rho) = \mathbb{E}_{\tilde{\xi},\tilde{z}}\, P_k(t_k+\tilde{\xi}/\sqrt{A}|\sqrt{\rho-V}\xi_k+\sqrt{V}\tilde{z}), \quad k=1,\dots,\ell-1,$$
$$P(t_\ell|\xi_\ell;V,\rho) = \mathbb{E}_{\tilde{z}}\, P_\ell(t_\ell|\sqrt{\rho-V}\xi_\ell+\sqrt{V}\tilde{z})$$

and

$$F_{W_k}(x) = \min_{\theta\in\mathbb{R}}\left\{2\alpha_k\theta + (\alpha_k-1)\ln(1-\theta) + \mathbb{E}_{\lambda_{W_k}}\ln[x\lambda_{W_k}+(1-\theta)(1-\alpha_k\theta)]\right\}$$

M. G., A. Manoel, C. Luneau, J. Barbier, N. Macris, F. Krzakala, L. Zdeborová - arXiv:1805.09785

Extension of the replica formula for entropies in multilayer networks

Setting: For multi-layer neural network with $\mathbf{X} \to \mathbf{T}_1 \to \mathbf{T}_2 \to \cdots \to \mathbf{T}_L$

- factorized input distribution $P_X(\mathbf{x}) = \prod_i P_0(x_i)$
- specified transition probabilities $\mathbf{t}_\ell \sim P_\ell(\mathbf{t}_\ell | W^{(\ell)} \mathbf{t}_{\ell-1})$
- weight matrices 1) orthogonally-invariant
  2) independent of each other
  3) aspect ratio (# cols / # rows) of order 1

diagonal

$$W^{(\ell)} = U_\ell \, S_\ell \, V_\ell$$

orthogonal
Haar distribued

Result: The entropy of each layer in the thermodynamic limit is given by the minimum among all extrema of the replica symmetric potential

$$\lim_{N_0 \to \infty} \frac{1}{N_0} H(\mathbf{T}_\ell) = \min \underset{\mathbf{A},\mathbf{V},\tilde{\mathbf{A}},\tilde{\mathbf{V}}}{\text{extr}} \phi_\ell(\mathbf{A}, \mathbf{V}, \tilde{\mathbf{A}}, \tilde{\mathbf{V}})$$

→ **High dimensional integral replaced by extremization over 4 – L variables**

Algorithm: Efficient implementation Code available in GitHub sphinxteam / dnner

M. G., A. Manoel, C. Luneau, J. Barbier, N. Macris, F. Krzakala, L. Zdeborová - arXiv:1805.09785

Why trust the heuristic of the replica formula ?

$$\lim_{N_0 \to \infty} \frac{1}{N_0} H(\mathbf{T}_\ell) = \min \operatorname*{extr}_{\mathbf{A},\mathbf{V},\tilde{\mathbf{A}},\tilde{\mathbf{V}}} \phi_\ell(\mathbf{A}, \mathbf{V}, \tilde{\mathbf{A}}, \tilde{\mathbf{V}})$$

- **Agreement with other heuristics**
  - Similar formula established from different arguments (G. Reeves 2017)
  - MLAMP-MLVAMP algorithms (D. Donoho et al. 2009, S. Rangan. 2011, L. Zdeborová and F. Krzakala. 2016, A. Manoel et al 2017, A. Fletcher al 2017)
- **Many rigorously proven subcases**

Work / Ref	**# layers**	**separable input distribution**	**smooth arbitrary activations**	**weight matrices ensemble indep**
Conjecture	arbitrary	arbitrary	arbitrary	orthogonally inv

Why trust the heuristic of the replica formula ?

$$\lim_{N_0 \to \infty} \frac{1}{N_0} H(\mathbf{T}_\ell) = \min \operatorname*{extr}_{\mathbf{A},\mathbf{V},\tilde{\mathbf{A}},\tilde{\mathbf{V}}} \phi_\ell(\mathbf{A},\mathbf{V},\tilde{\mathbf{A}},\tilde{\mathbf{V}})$$

- **Agreement with other heuristics**
  - Similar formula established from different arguments (G. Reeves 2017)
  - MLAMP-MLVAMP algorithms (D. Donoho et al. 2009, S. Rangan. 2011, L. Zdeborová and F. Krzakala. 2016, A. Manoel et al 2017, A. Fletcher al 2017)
- **Many rigorously proven subcases**

Work / Ref	**# layers**	**separable input distribution**	**smooth arbitrary activations**	**weight matrices ensemble indep**
Conjecture	arbitrary	arbitrary	arbitrary	orthogonally inv
Pinsker 1964	arbitrary	Gaussian	linear w. awgn	arbitrary
Reeves et al. 2016 Barbier et al. 2017	single	arbitrary	arbitrary	i.i.d entries
Reeves 2017	arbitrary tree	Gaussian	linear w. awgn	orthogonally inv
Barbier et al. 2018	single	arbitrary	linear w. awgn	orthogonally inv*

Why trust the heuristic of the replica formula ?

$$\lim_{N_0 \to \infty} \frac{1}{N_0} H(\mathbf{T}_\ell) = \min \operatorname*{extr}_{\mathbf{A},\mathbf{V},\tilde{\mathbf{A}},\tilde{\mathbf{V}}} \phi_\ell(\mathbf{A}, \mathbf{V}, \tilde{\mathbf{A}}, \tilde{\mathbf{V}})$$

- **Agreement with other heuristics**
  - Similar formula established from different arguments (G. Reeves 2017)
  - MLAMP-MLVAMP algorithms (D. Donoho et al. 2009, S. Rangan. 2011, L. Zdeborová and F. Krzakala. 2016, A. Manoel et al 2017, A. Fletcher al 2017)
- **Many rigorously proven subcases**

Work / Ref	# layers	separable input distribution	smooth arbitrary activations	weight matrices ensemble indep
Conjecture	arbitrary	arbitrary	arbitrary	orthogonally inv
Pinsker 1964	arbitrary	Gaussian	linear w. awgn	arbitrary
Reeves et al. 2016 Barbier et al. 2017	single	arbitrary	arbitrary	i.i.d entries
Reeves 2017	arbitrary tree	Gaussian	linear w. awgn	orthogonally inv
Barbier et al. 2018	single	arbitrary	linear w. awgn	orthogonally inv*
→ Our paper	two	bounded support	arbitrary	i.i.d entries

Rigorous result for 2-layer with Gaussian i.i.d. weights

Theorem:

Suppose

- $(H1)$ the input units distribution P_0 is separable and has bounded support;
- $(H2)$ the activations f_1 and f_2 corresponding to $P_1(t_{1,i}|\mathbf{W}_i^{(1)\intercal}\mathbf{x})$ and $P_2(t_{2,i}|\mathbf{W}_i^{(2)\intercal}\mathbf{t}_1)$ are bounded $\mathcal{C}^2$ with bounded first and second derivatives w.r.t. their first argument;
- $(H3)$ the weight matrices $W^{(1)}$, $W^{(2)}$ have Gaussian i.i.d. entries.

Then for model with two layers $L = 2$ the high-dimensional limit of the entropy is rigorously given by the replica formula.

On going work of extending the proof

→ Poster of Clément Luneau later today !

M. G., A. Manoel, C. Luneau, J. Barbier, N. Macris, F. Krzakala, L. Zdeborová - arXiv:1805.09785

Outline

1 – Replica formula for multilayer networks

2 – Entropies in neural networks with random weights

3 – Entropies in deep learning

Random weights networks:
feature extractor, toy model, initialisation of learning ...

Random Features for Large-Scale Kernel Machines

Ali Rahimi and Ben Recht

NIPS, 2007

EEE Trans. Signal Processing, 2016

Deep Neural Networks with Random Gaussian Weights: A Universal Classification Strategy?

Raja Giryes*, Guillermo Sapiro**, and Alex M. Bronstein*

*School of Electrical Engineering, Faculty of Engineering, Tel-Aviv University, Ramat Aviv 69978, Israel.
{raja@tauex, bron@eng}.tau.ac.il
**Department of Electrical and Computer Engineering, Duke University, Durham, North Carolina, 27708, USA.
{guillermo.sapiro}@duke.edu.

The Emergence of Spectral Universality in Deep Networks

Jeffrey Pennington
Google Brain

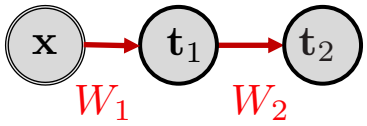
Samuel S. Schoenholz
Google Brain

Surya Ganguli
Google Brain
Applied Physics, Stanford University

AISTAT, 2018

Rigorous entropy with replica
2-layer / random weights / synthetic data

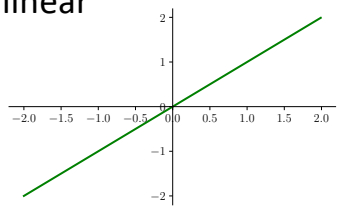
$$x_i \sim \mathcal{N}(0,1) \qquad \mathbf{t_2} = f(W^{(2)} f(W^{(1)}\mathbf{x}) + \epsilon) \qquad \epsilon_i \sim \mathcal{N}(0, 10^{-5})$$



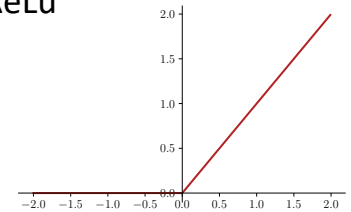
Entropy as a function of weights magnitude

$$W^{(1)}_{i,j} \sim \mathcal{N}(0, \sigma/N_0) \quad W^{(2)}_{i,j} \sim \mathcal{N}(0, \sigma/N_1)$$

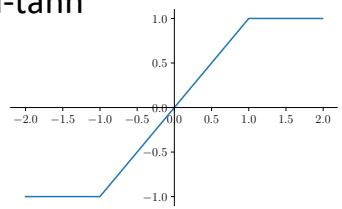
linear

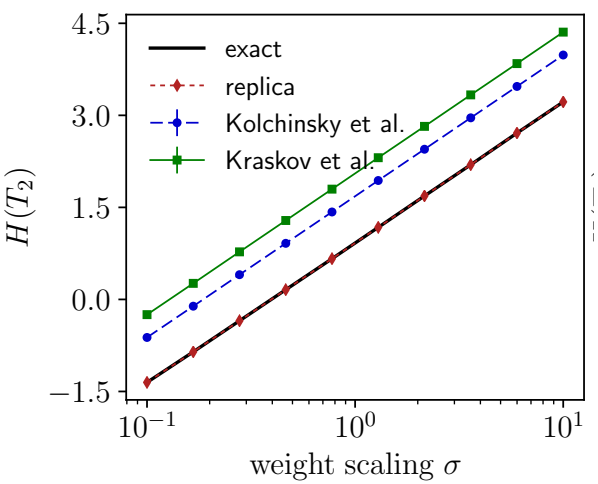


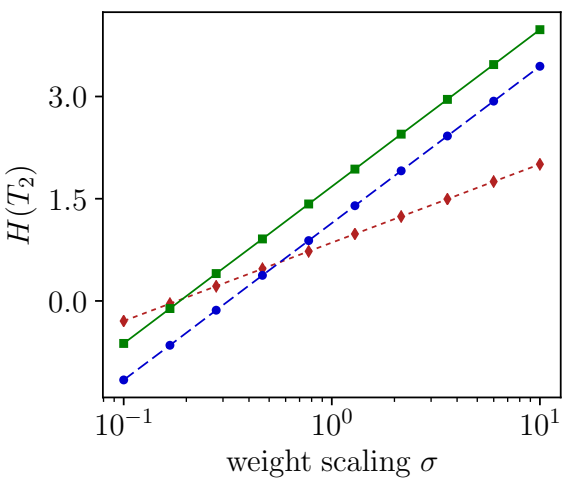
ReLu

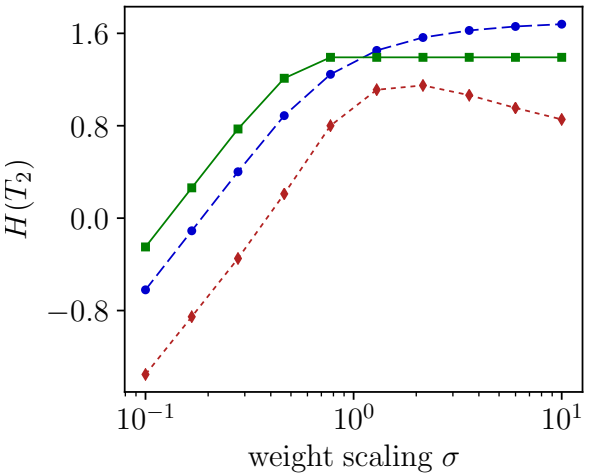


hard-tanh





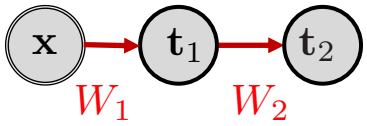




Non parametric estimators: Kraskov et al. 2014, Kolchinsky et al. 2017

Rigorous entropy with replica
2-layer / random weights / synthetic data

$$x_i \sim \mathcal{N}(0,1)$$

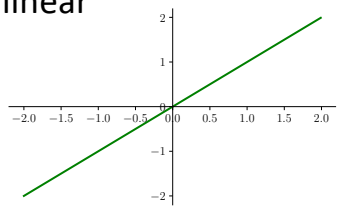


$$\mathbf{t_2} = f(W^{(2)} f(W^{(1)}\mathbf{x}) + \epsilon) \qquad \epsilon_i \sim \mathcal{N}(0, 10^{-5})$$

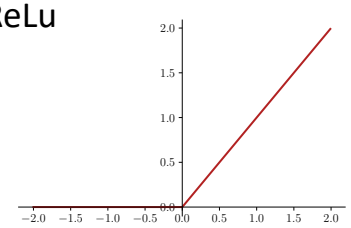
Mutual information as a function of weights magnitude / signal-to-noise ratio

$$W^{(1)}_{i,j} \sim \mathcal{N}(0, \sigma/N_0) \quad W^{(2)}_{i,j} \sim \mathcal{N}(0, \sigma/N_1)$$

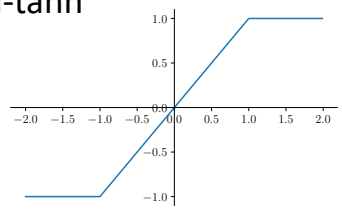
linear

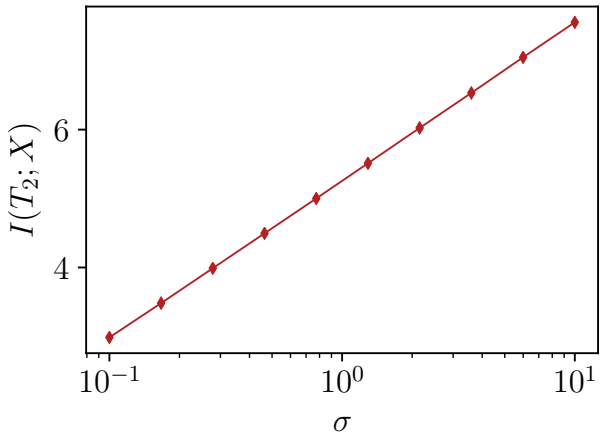


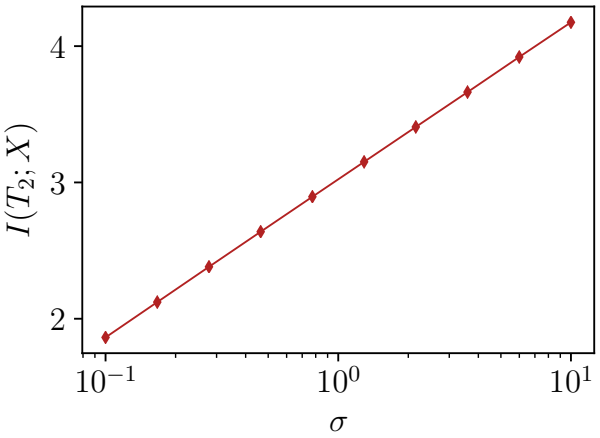
ReLu

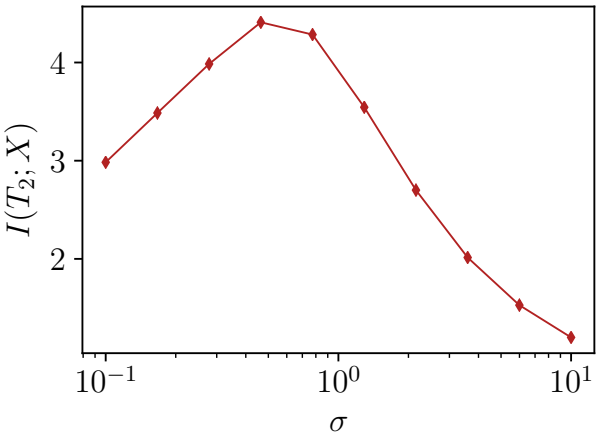


hard-tanh









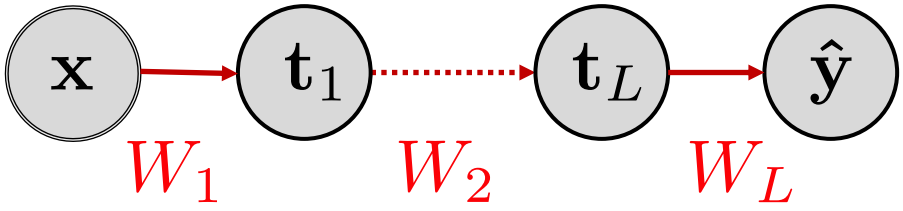
Outline

1 – Replica formula for multilayer networks

2 – Entropies in neural networks with random weights

3 – Entropies in deep learning

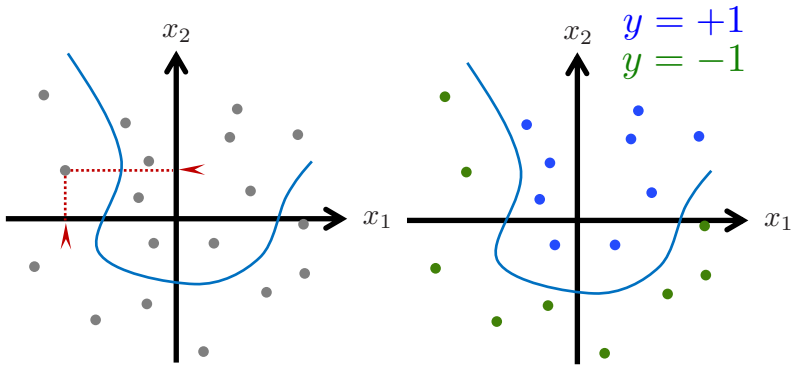
Can we follow information during learning (with SGD) ?

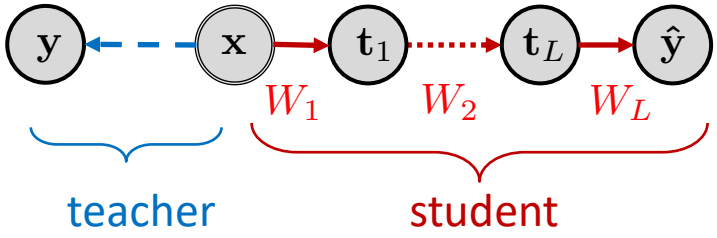


$$I(\mathbf{X};\mathbf{T}_\ell)\ ?$$

Synthetic data framework for supervised learning

- **Scenario 1: Input data with separable prior** $P_X(\mathbf{x}) = \prod_i P_0(x_i)$

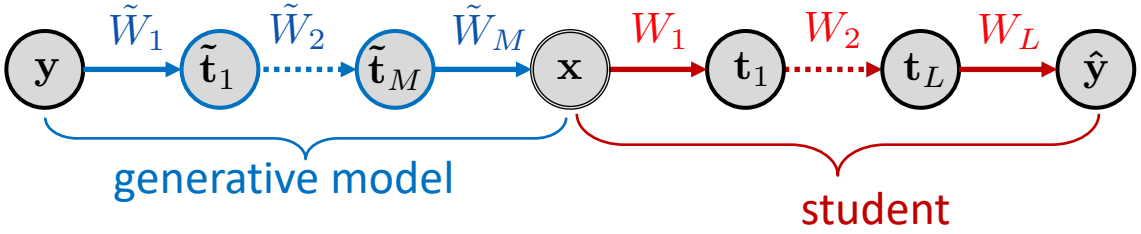




- **Scenario 2: Separable target outputs + multilayer generative model for inputs**

$P_Y(\mathbf{y}) = \prod_i P_0(y_i)$

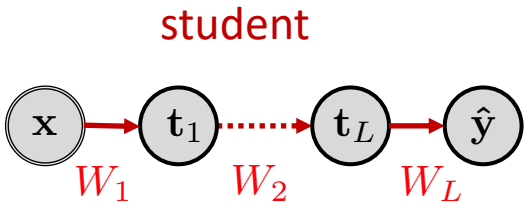
$P_X(\mathbf{x}) \neq \prod_i P_0(x_i)$ structured data !



e.g. Variational Auto Encoders Kingma et. al 2014, Rezende et al 2014

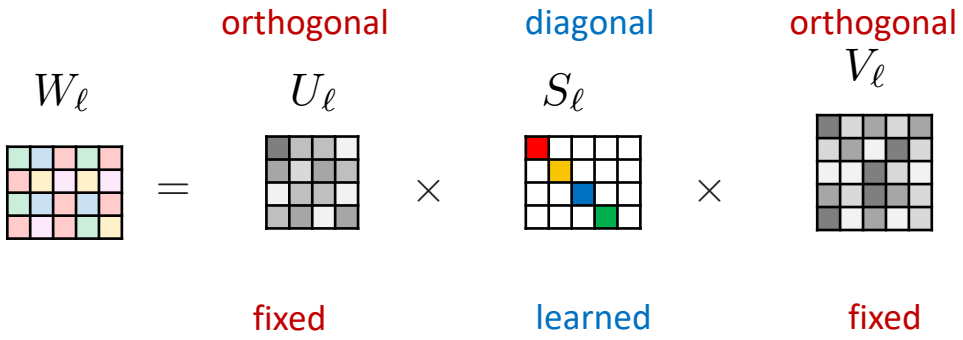
e.g. Generative Adversarial models Goodfellow et al. 2014

Learning orthogonally invariant weight matrices "USV-layers"

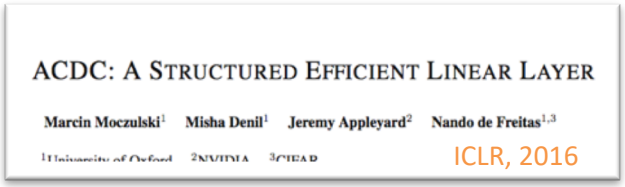


How to guarantee orthogonal invariance during learning ?

- Initialize Gaussian i.i.d W matrices
- Perform singular value decomposition
- Only learn spectrum (N degrees of freedom instead of N^2)



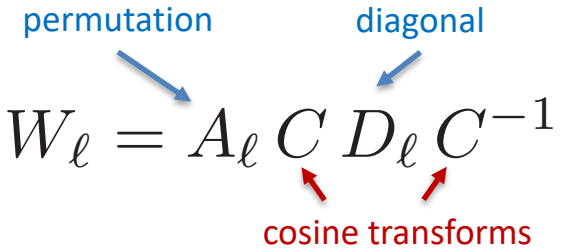
Side note: Related weight constraints for speed concerns


ACDC: A Structured Efficient Linear Layer

Marcin Moczulski[1] Misha Denil[1] Jeremy Appleyard[2] Nando de Freitas[1,3]

[1]University of Oxford [2]NVIDIA [3]CIFAR

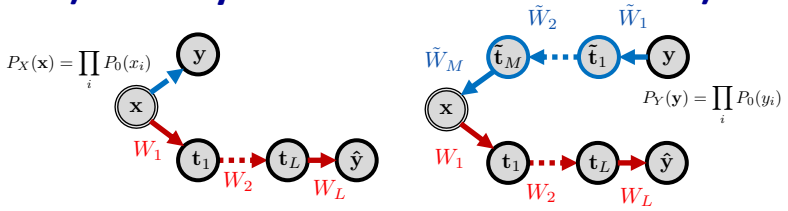
ICLR, 2016

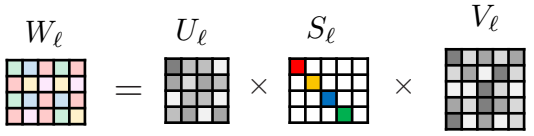


Learning experiments framework

- **Framework for learning experiments: Be careful with**

  1) Data: synthetic data sets **2) Weights: preserving orthogonal invariance**





Code available in GitHub marylou-gabrie / **learning-synthetic-data**

- **Method to compute entropies in large deep neural networks throughout training**

$$\lim_{N_0 \to \infty} \frac{1}{N_0} H(\mathbf{T}_\ell) = \min_{\mathbf{A},\mathbf{V},\tilde{\mathbf{A}},\tilde{\mathbf{V}}} \operatorname{extr} \phi_\ell(\mathbf{A}, \mathbf{V}, \tilde{\mathbf{A}}, \tilde{\mathbf{V}})$$

Code available in GitHub sphinxteam / **dnner**

- **Next: Numerical results**
  1) Check against the linear exact case
  2) Run exploratory non-linear training experiments

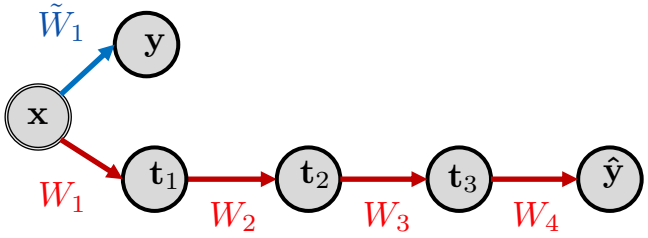
Safety-check on linear networks

Model: Linear teacher – linear student:

$$\mathbf{t}_\ell = \mathbf{W}^{(\ell)} \cdots \mathbf{W}^{(1)} \mathbf{x} + \epsilon$$

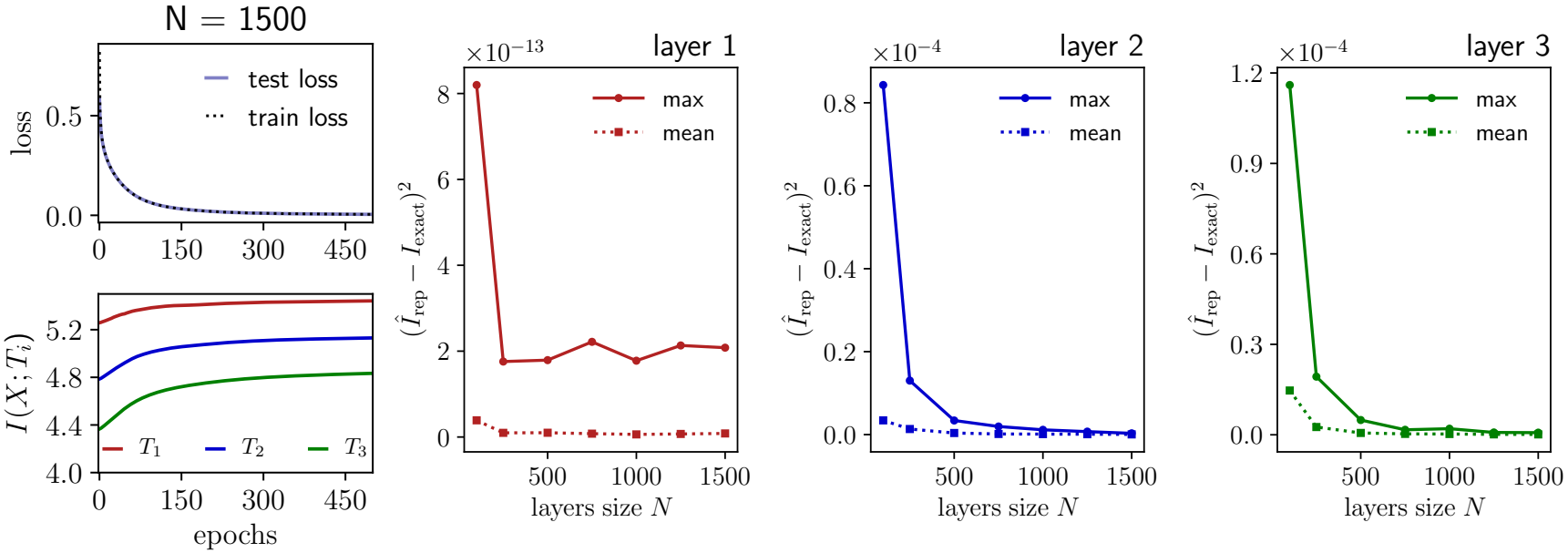
$$\epsilon_i \sim \mathcal{N}(0, 10^{-5})$$

$x_i \sim \mathcal{N}(0,1)$



Task: regression $\min \; ||\mathbf{y} - \hat{\mathbf{y}}||^2$

Test: deviation of replica estimator from analytical $(\hat{I}_{\text{rep}} - I_{\text{exact}})^2$



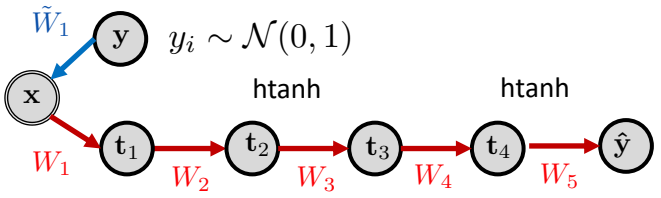
Exploratory experiment: A binary classification example

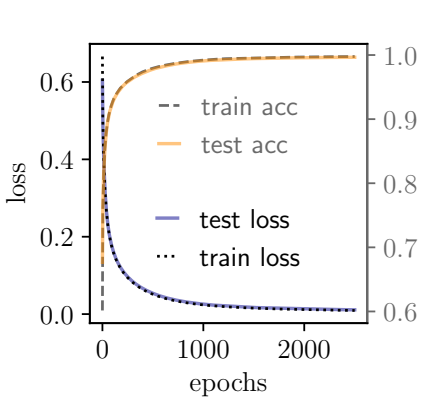
Model: linear generative model – non-linear student

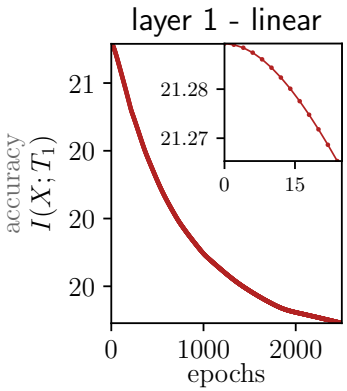
Task: predict $Y = \text{sign}(y_0)$ for each $\mathbf{x}$

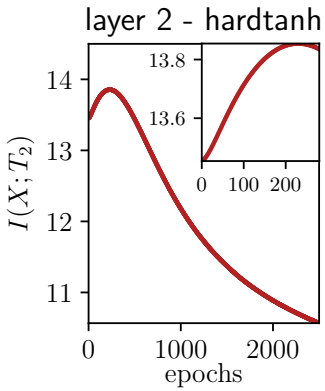
Noise: $\mathbf{t}_\ell = f(\mathbf{W}^{(\ell)}..f(\mathbf{W}^{(1)}\mathbf{x})..+\epsilon)$

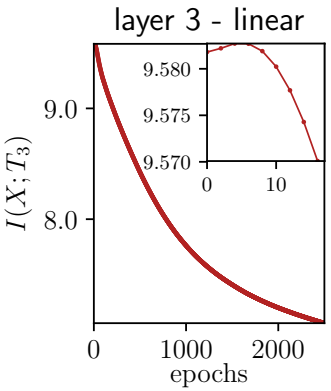
$\epsilon_i \sim \mathcal{N}(0, 10^{-5})$

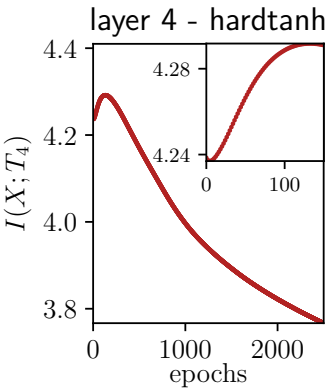


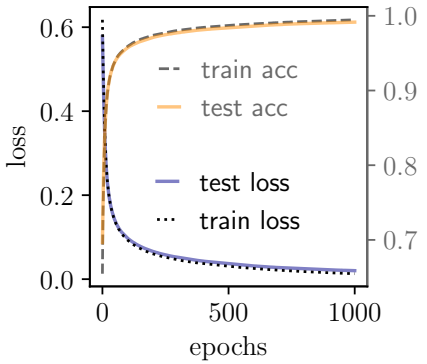


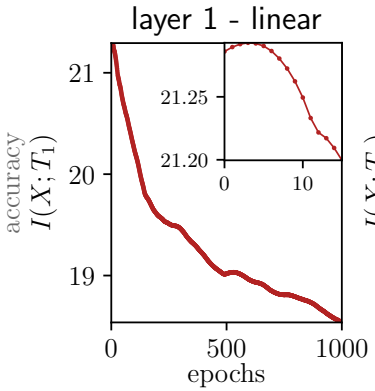


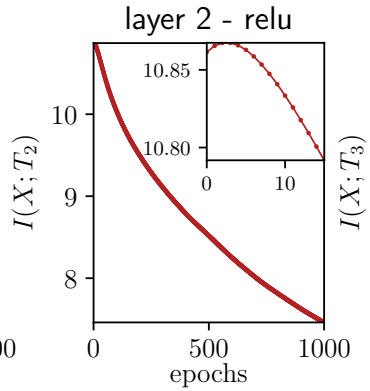


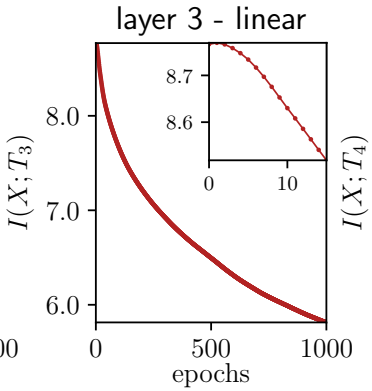


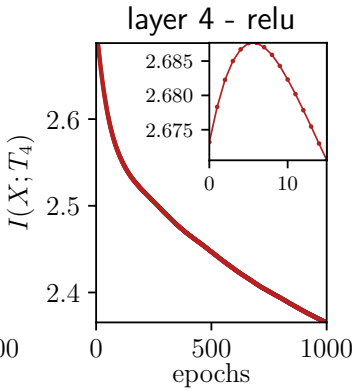












Compression as well with non-saturated activation functions

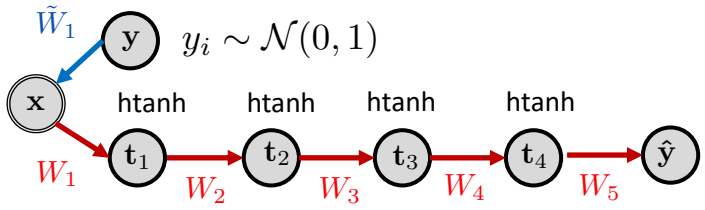
Exploratory: A second binary classification example

Model: linear generative model – non-linear student

Task: predict $Y = \text{sign}(y_0)$ for each $\mathbf{x}$

Noise: $\mathbf{t}_\ell = f(\mathbf{W}^{(\ell)}..f(\mathbf{W}^{(1)}\mathbf{x})..+\epsilon)$

$\epsilon_i \sim \mathcal{N}(0, 10^{-5})$

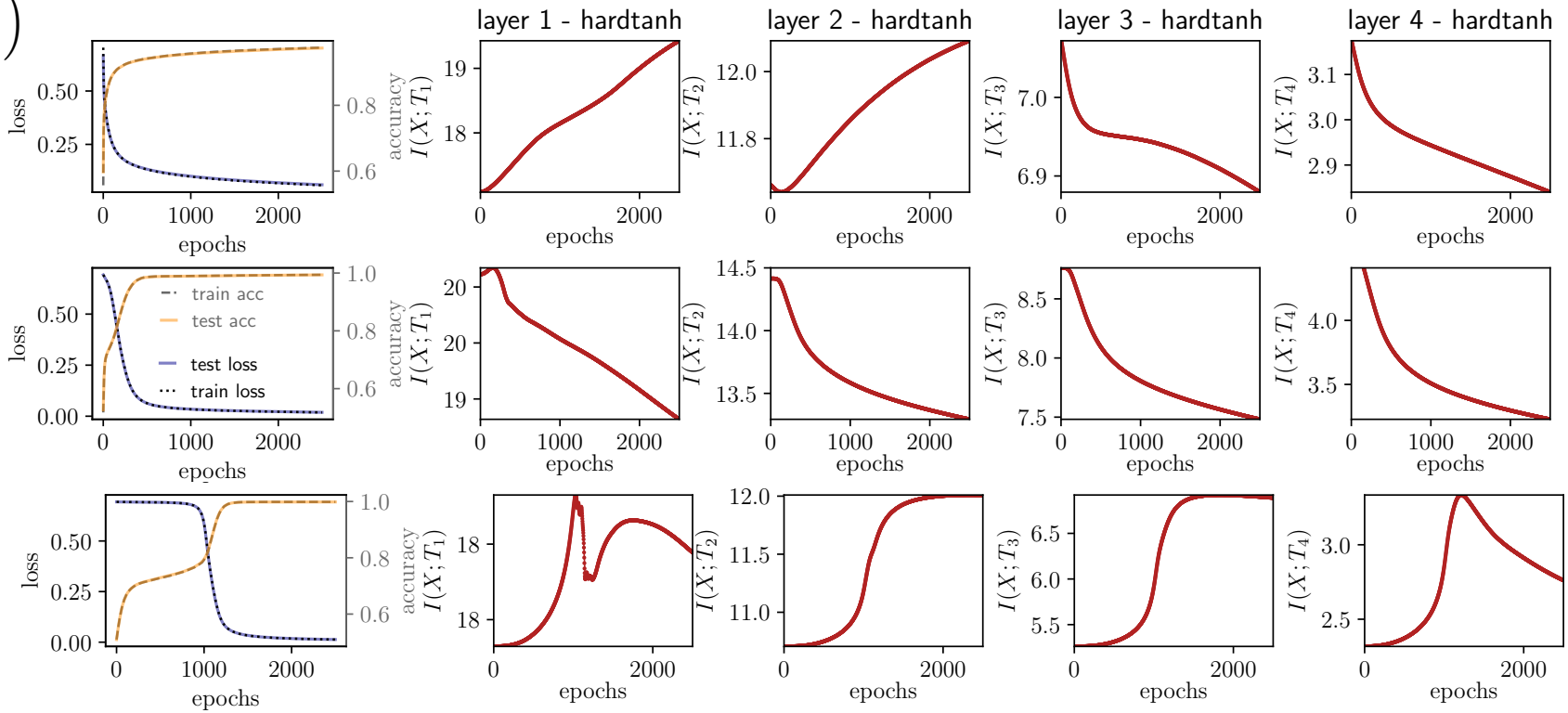


initial $\text{var}\left(\mathbf{W}^{(\ell)}_{\mathbf{i,j}}\right)$

$\frac{2.}{N_{\ell-1}}$

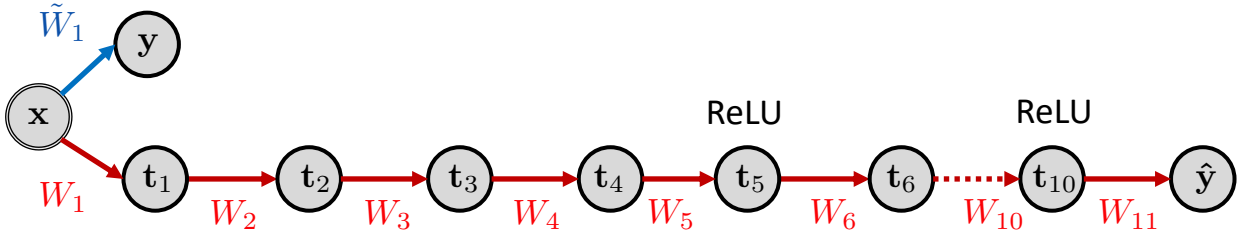
$\frac{1.}{N_{\ell-1}}$

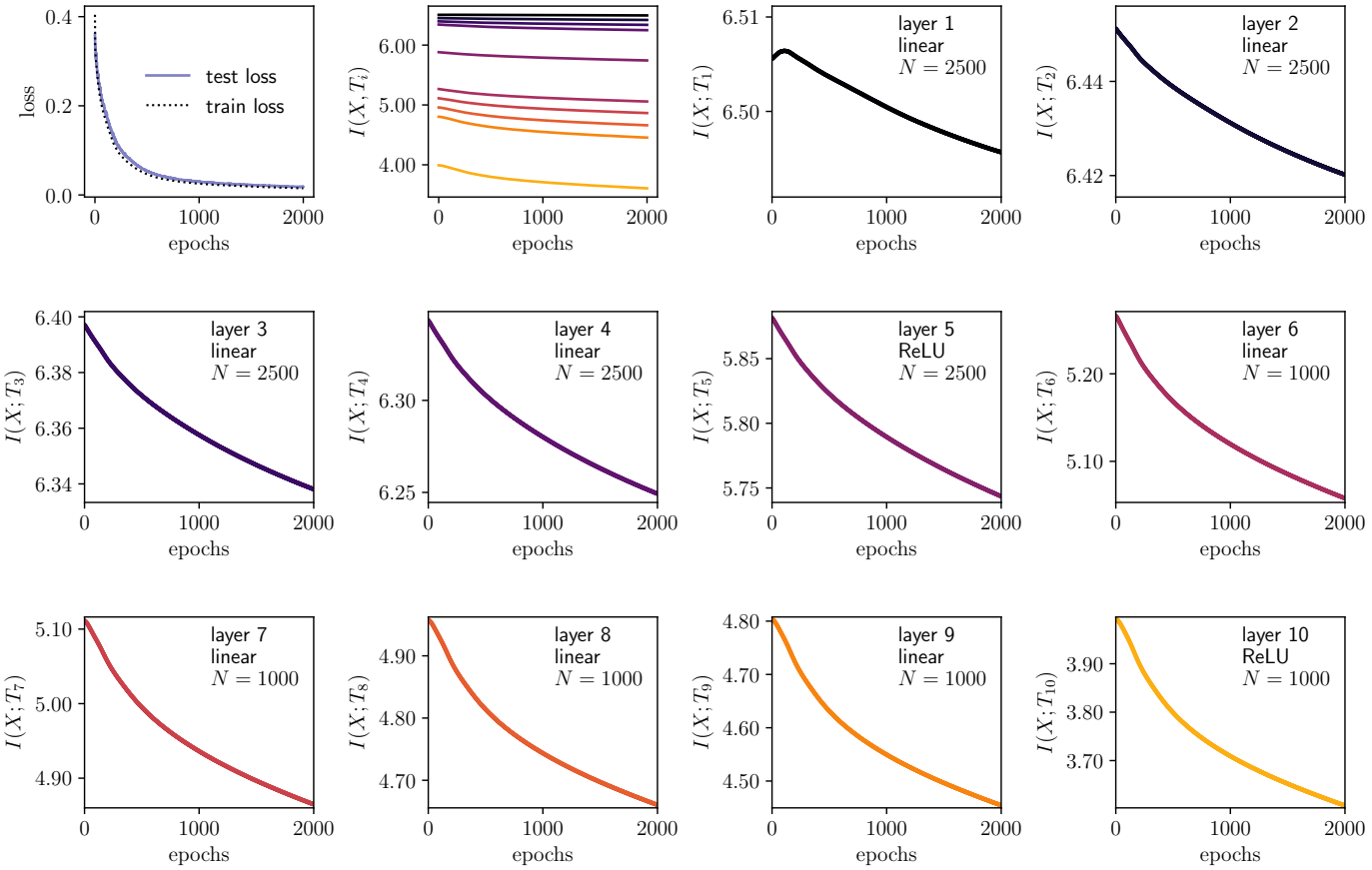
$\frac{0.25}{N_{\ell-1}}$



Slight changes in hyper parameters seem to drastically modify behaviors

Exploratory experiment: A regression example





Compression increasing slightly with depth

Recap

A relevant setting to study deep learning ?

- ✓ **Non linear networks**
- ✓ **Large networks (but only USV layers with N parameters)**
- ✗ **Real data (but sophisticated synthetic data)**

Perspectives

- **Add biases in the theory**
- **Improve the number of parameters one can learn per layer ? (towards over parametrized regime)**
- **Learn the generative model to get almost real data**

THANK YOU !

References

1. Barbier, J., Krzakala, F., Macris, N., Miolane, L. & Zdeborová, L. Phase Transitions, Optimal Errors and Optimality of Message-Passing in Generalized Linear Models. 1–59 (2017). at <http://arxiv.org/abs/1708.03395>
2. Barbier, J., Macris, N., Maillard, A. & Krzakala, F. The Mutual Information in Random Linear Estimation Beyond i.i.d. Matrices. (2018). at <http://arxiv.org/abs/1802.08963>
3. Donoho, D. L., Maleki, A. & Montanari, A. Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci.* 106, 18914–18919 (2009).
4. Fletcher, A. K., Rangan, S. & Schniter, P. Inference in Deep Networks in High Dimensions. in *2018 IEEE International Symposium on Information Theory (ISIT)* 1, 1884–1888 (IEEE, 2018).
5. Giryes, R., Sapiro, G. & Bronstein, A. M. Deep Neural Networks with Random Weights. *IEEE Trans. Signal Process.* 1–14 (2016).
6. Goodfellow, I. J. *et al.* Generative Adversarial Networks. in *Advances in Neural Information Processing Systems 27* 1–9 (2014).
7. Kabashima, Y. Inference from correlated patterns: a unified theory for perceptron learning and linear vector channels. *J. Phys. Conf. Ser.* 95, 012001 (2008).
8. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. in *International Conference on Learning Representations (ICLR)* 1–14 (2014). at
9. Kolchinsky, A., Tracey, B. D. & Wolpert, D. H. Nonlinear Information Bottleneck. 1–11 (2017). at <http://arxiv.org/abs/1705.02436>
10. Kraskov, A., Stögbauer, H. & Grassberger, P. Estimating mutual information. *Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top.* 69, 16 (2004).
11. Manoel, A., Krzakala, F., Mezard, M. & Zdeborova, L. Multi-layer generalized linear estimation. in *2017 IEEE International Symposium on Information Theory (ISIT)* 2098–2102 (IEEE, 2017). doi:10.1109/ISIT.2017.8006899
12. Moczulski, M., Denil, M., Appleyard, J. & de Freitas, N. ACDC: A Structured Efficient Linear Layer. 1–12 (2015). at <http://arxiv.org/abs/1511.05946>
13. Pennington, J., Schoenholz, S. S. & Ganguli, S. The Emergence of Spectral Universality in Deep Networks. *Int. Conf. Artif. Intell. Stat.* 1924–1932. (2018). at <http://arxiv.org/abs/1802.09979>
14. Pinsker, M. S. *Information and Information Stability of Random Variables and Processes*. (Holden-Day, 1964).
15. Rahimi, A. & Recht, B. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Adv. Neural Inf. Process. Syst.* 1, 1–8 (2009).
16. Rangan, S. Generalized approximate message passing for estimation with random linear mixing. in *2011 IEEE International Symposium on Information Theory Proceedings* 2168–2172 (IEEE, 2011). doi:10.1109/ISIT.2011.6033942
17. Reeves, G. Additivity of information in multilayer networks via additive Gaussian noise transforms. *55th Annu. Allert. Conf. Commun. Control. Comput. Allert. 2017* 2018–January, 1064–1070 (2018).
18. Reeves, G. & Pfister, H. D. The replica-symmetric prediction for compressed sensing with Gaussian matrices is exact. in *2016 IEEE International Symposium on Information Theory (ISIT)* 665–669 (IEEE, 2016). doi:10.1109/ISIT.2016.7541382
19. Rezende, D. J., Mohamed, S. & Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *Proc. 31st …* 32, 1278–1286 (2014).
20. Shinzato, T. & Kabashima, Y. Perceptron capacity revisited: Classification ability for correlated patterns. *J. Phys. A Math. Theor.* 41, (2008).
21. Shinzato, T. & Kabashima, Y. Learning from correlated patterns by simple perceptrons. *J. Phys. A Math. Theor.* 42, (2009).