Laboratoire de Physique – École Normale Supérieure

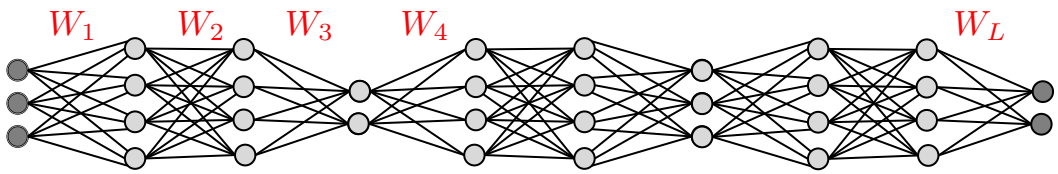
Entropy and mutual information in deep neural networks

Basel – April 15th 2019

Marylou Gabrié (LPS ENS), Andre Manoel (INRIA Saclay, Owkin),
Clément Luneau, Nicolas Macris (EPFL), Jean Barbier (ICTP Trieste),
Lenka Zdeborová (CEA Saclay), Florent Krzakala (LPS ENS) & Léon Bottou (FAIR NY)

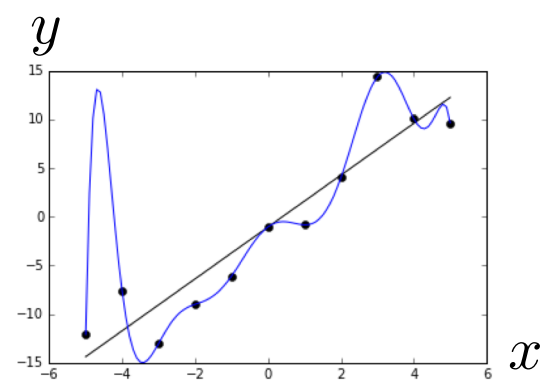
Fundamental question

How come deep neural networks generalize well ?



$$\mathbf{y} = f(\mathbf{W}^{(L)} f(\mathbf{W}^{(L-1)} .. f(\mathbf{W}^{(1)}\mathbf{x} + b^{(1)}) .. + b^{(L-1)}) + b^{(L)})$$

number of parameters >> training data



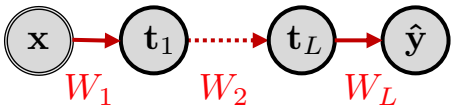
YET no systematic overfitting

Hot question in deep learning,

Recent propositon based on information theory

Information theory on deep networks

(Stochastic) layers variables $\mathbf{t}_\ell = f(W^{(\ell)}\mathbf{t}_{\ell-1};\epsilon)$



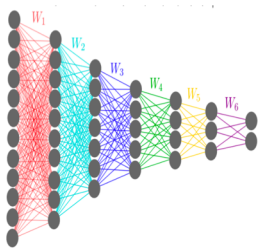
changing with weights learning !

Mutual information with inputs and outputs

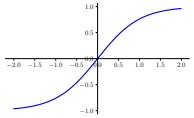
$$\begin{cases} I(\mathbf{X};\mathbf{T}_\ell) = \int \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t}_\ell\, p(x,\mathbf{t}_\ell)\log\dfrac{p(\mathbf{x},\mathbf{t}_\ell)}{p(\mathbf{x})p(\mathbf{t}_\ell)} = KL[p(\mathbf{x},\mathbf{t}_\ell)||p(x)p(\mathbf{t}_\ell)] \\ I(\mathbf{Y};\mathbf{T}_\ell) \end{cases}$$

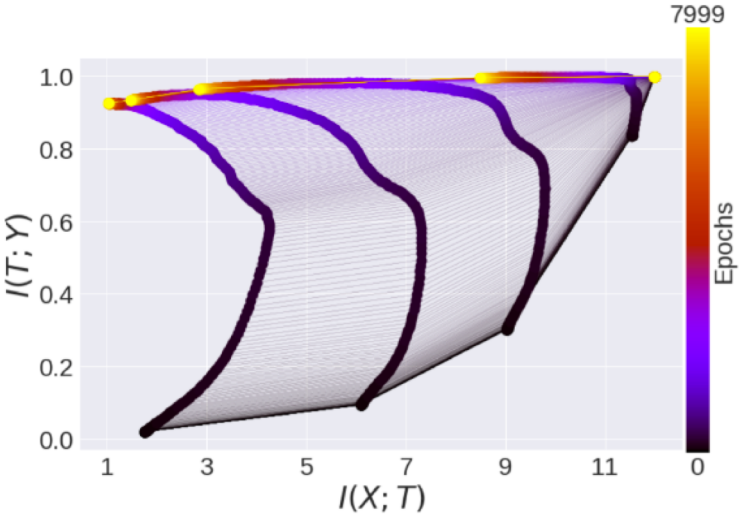
First results training a small network

12 binary input neurons,
bins to discretize activations
(R. Shwartz-Ziv, N. Tishby 2017.)



tanh





Claim: compression of information on inputs linked to generalization

Information bottleneck principle (N. Tishby, F. C. Pereira, and W.Bialek. 1999.)

What are good hidden representations in terms of information ?

$$\mathbf{x} \xrightarrow{\theta} \mathbf{t} \rightarrow \hat{\mathbf{y}}$$

Intuitions :

1. Hidden layer should help to predict output: $\max_{\theta} \; I(\mathbf{Y}, \mathbf{T})$
2. Useful representation get rid of the unnecessary: $I(\mathbf{X}, \mathbf{T}) \leq C$

Learning objective:

$$\boxed{\max_{\theta} \; I(\mathbf{Y}, \mathbf{T}_\ell) - \beta I(\mathbf{X}, \mathbf{T}_\ell)}$$

Information bottleneck Lagrangian

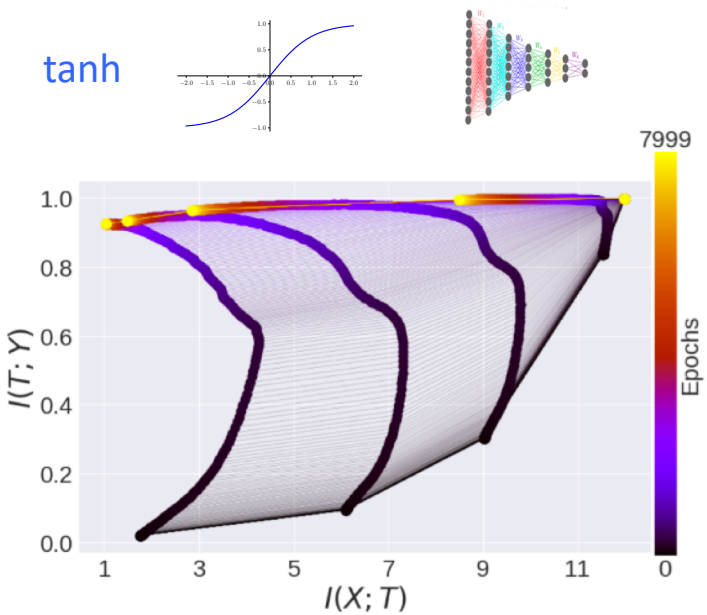
Questions raised for deep learning theory :

1. **Can we efficiently train neural networks with this alternative criteria ?**
2. **Trade-off already optimized by commonly used training algorithms (SGD) ?**

How general is this observation ?

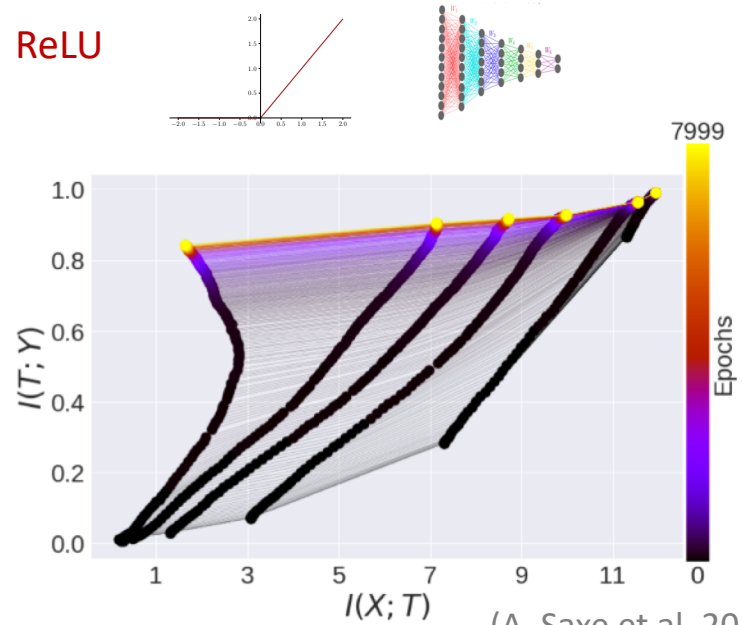
First small network tested

tanh



(R. Shwartz-Ziv, N. Tishby 2017.)

ReLU



(A. Saxe et al, 2018)

Same experiment with ReLUs without compression

How general is this observation: Challenges to test the hypothesis

1 – Definition of mutual information between layers of DNNs

2 – Precise evaluation of the mutual information

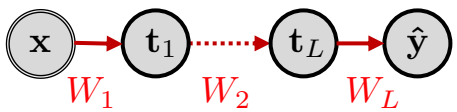
How general is this observation: Challenges to test the hypothesis

1 – Definition of mutual information between layers of DNNs

2 – Precise evaluation of the mutual information

Definition of the mutual information

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



Deterministic neural networks

$$p(\mathbf{T}_\ell|\mathbf{X}) = \delta\left(\mathbf{T}_\ell - F_W(\mathbf{X})\right)$$

$$I(\mathbf{X};\mathbf{T}_\ell) = H(\mathbf{T}_\ell) - \underbrace{H(\mathbf{T}_\ell|\mathbf{X})}_{\to -\infty}$$

Shannon entropy for continuous variables can diverge

Regularization strategies to keep finite mutual infromation

1. Discretize values

   (R. Shwartz-Ziv, N. Tishby 2017.)

   (A. Saxe et al, 2018)

$$p_i = \frac{1}{K}\sum_{k=1}^{K}\delta(b_i \le t_k < b_{i+1})$$

$$H = -\sum_{i=1}^{N_{\text{bins}}} p_i \ln p_i$$

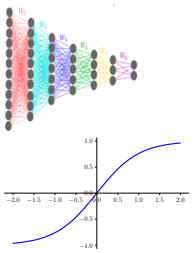
2. Add noise in the network

   (A. Saxe et al, 2018)

$$p(\mathbf{t}_\ell|\mathbf{t}_{\ell-1}) = p_{noise}\left(\mathbf{t}_\ell - f(W_\ell \mathbf{t}_{\ell-1})\right)$$

$$H(\mathbf{T}_\ell|\mathbf{X}) = -\int \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t}_\ell\, p(x,\mathbf{t}_\ell)\log p(\mathbf{t}_\ell|\mathbf{x}) \le \infty$$

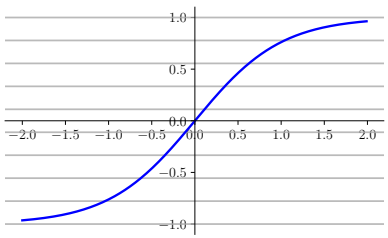
Regularization 1: Discretization a posteriori

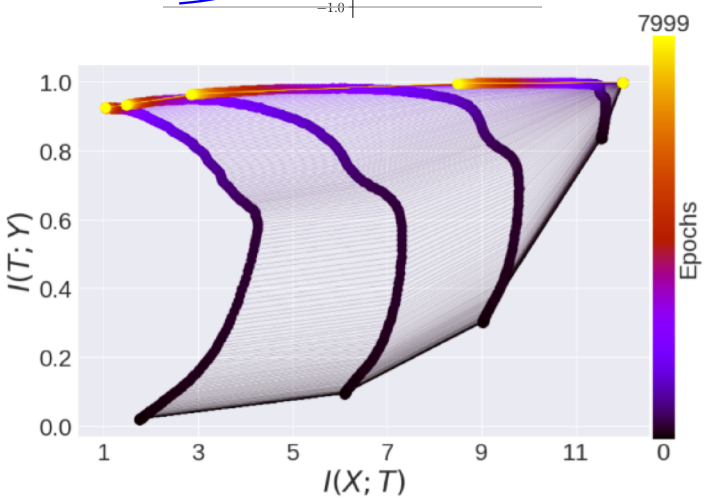


tanh

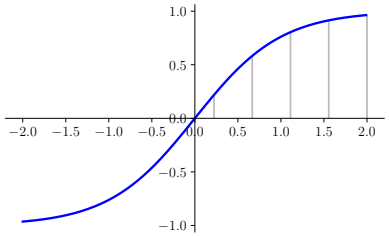
First small network tested

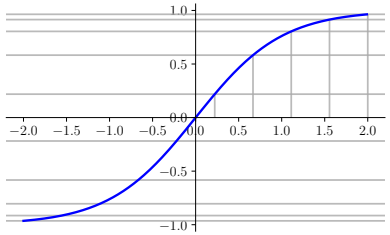
Even binning of activations **VS** **Even binning of pre - activations**

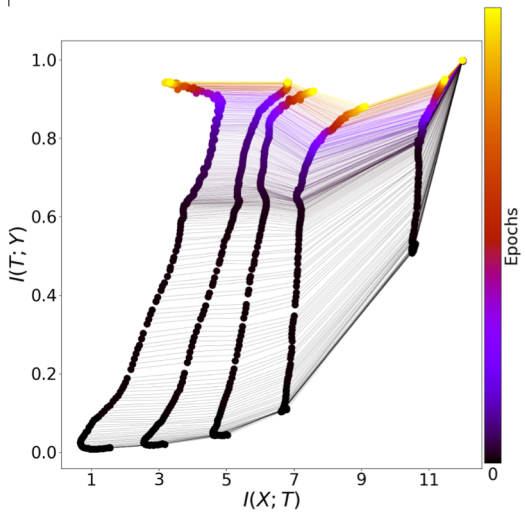




(R. Shwartz-Ziv, N. Tishby 2017.)







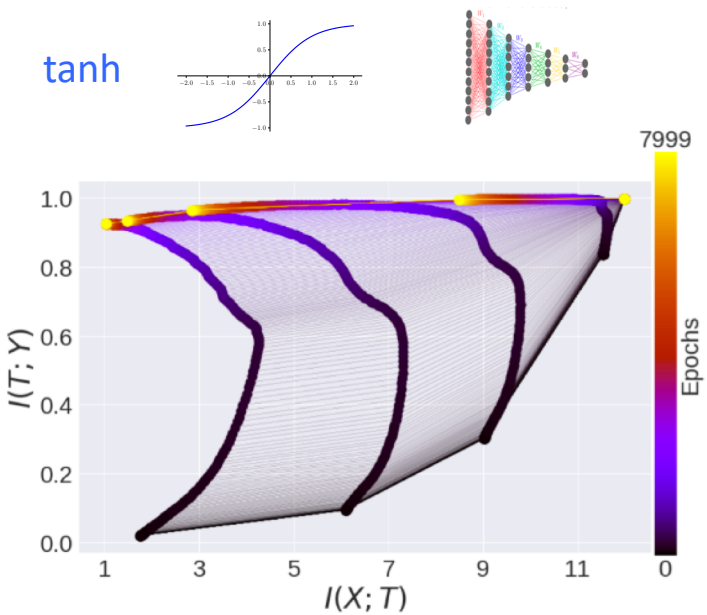
(A. Saxe et al, 2018)

Interpretation: Compression due to binning of saturating tanh activations

Regularization 1: Discretization a posteriori

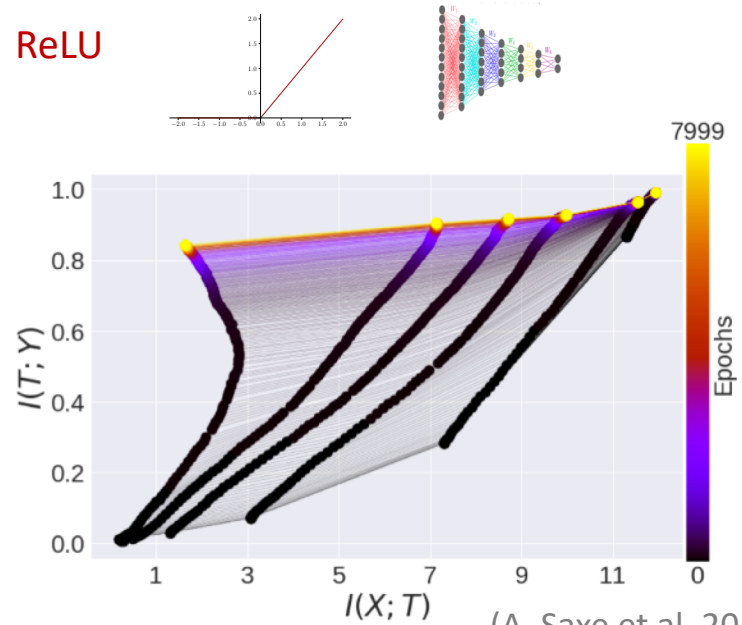
First small network tested

tanh



(R. Shwartz-Ziv, N. Tishby 2017.)

ReLU

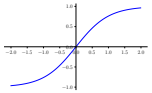


(A. Saxe et al, 2018)

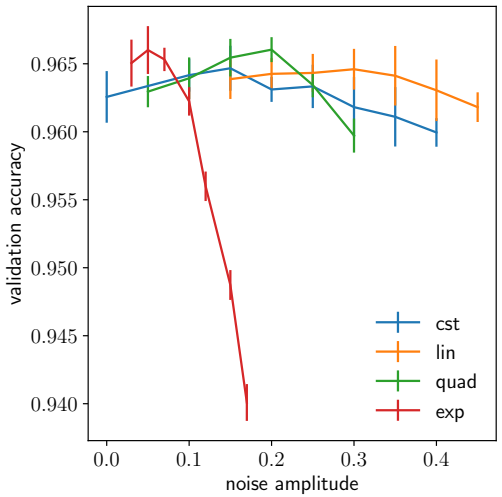
Interpretation: Compression due to binning of saturating tanh activations ?

Regularization 2: Stochastic neural networks MNIST experiment

tanh



5 hidden layers NN trained on MNIST



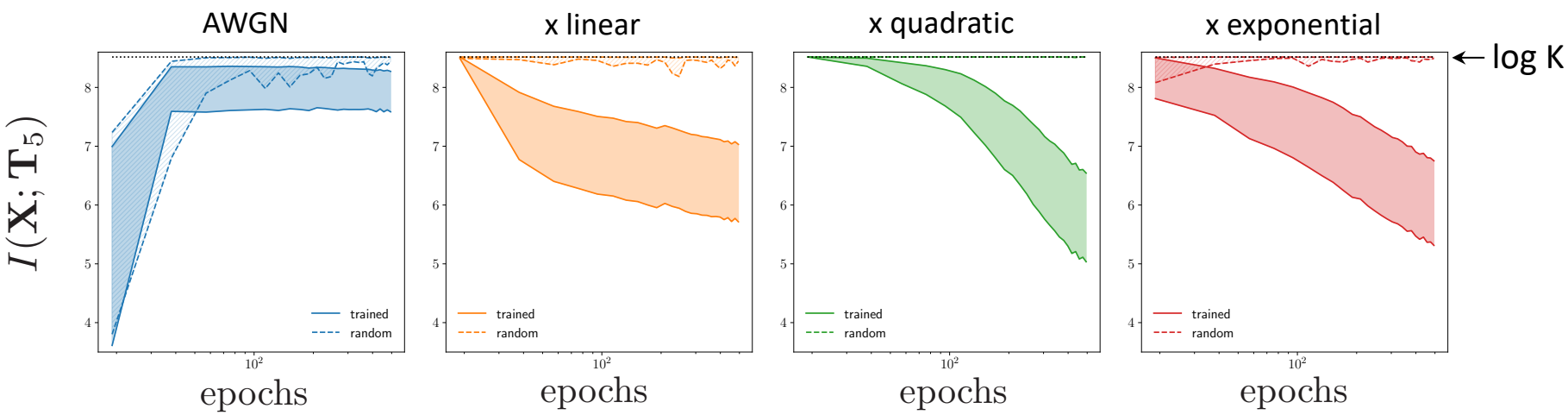
Add noise $\mathbf{t}_\ell = f(\mathbf{z}_\ell) = f(W^{(\ell)}\mathbf{t}_{\ell-1} + \epsilon)$

Pick optimal noise promoting regularization

Estimate finite mutual info $I(\mathbf{X};\mathbf{T}_\ell) = H(\mathbf{T}_\ell) - H(\mathbf{T}_\ell|\mathbf{X})$

using Kolchinsky et al. 2017 non parametric method:

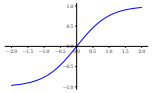
- Kernel density estimation from **K samples** - **O(K²)** runtime
- Upper and lower bound on MI – up to **log K**



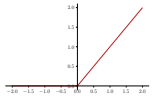
Loss of information due to organized saturation tanh activations

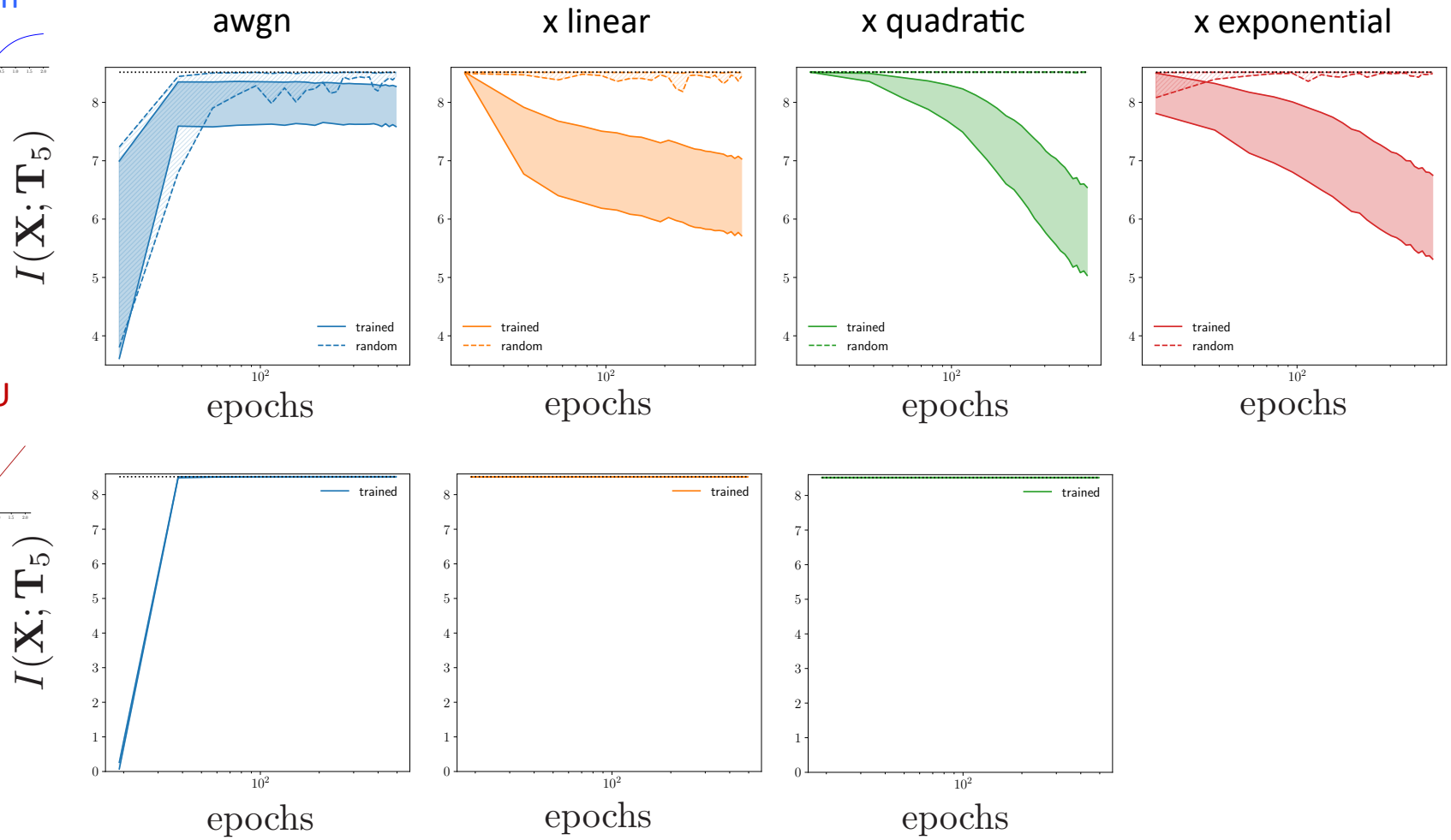
Regularization 2: Stochastic neural networks
MNIST experiment

tanh



ReLU





Method not precise enough to conclude for ReLUs

How general is this observation: Challenges to test the hypothesis

1 – Definition of the mutual information between layers of DNNs

- Mutual information in deterministic networks is diverging
- Need to be careful about how the computation is regularized
- Continuous stochastic networks appears more principled
- But is hard …

2 – Precise evaluation of the mutual information

- A formula inspired by statistical physics
- Application to follow mutual information during learning

Computing information theoretic quantities for large networks is a hard problem

How to compute the mutual information ?

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

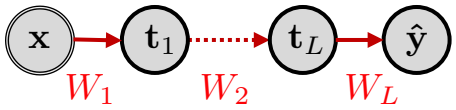
Unknown distributions

- Non parametric estimates unreliable in high dimensions
  (based on pairwise distances between samples)

Known distributions

- Numerical integration exponentially hard
- Particular subcases ? Closed forms or reliable approximations ?

Analytical formula for linear DNN



Given

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

Analytical expression of mutual information (Pinsker 1964)

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

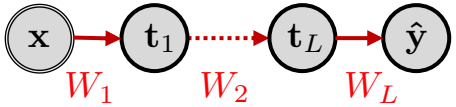
A relevant setting to study deep learning ?

✓ **Large networks**

✗ **Non linear networks**

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

Replica formula for entropies: model of stochastic multilayer networks

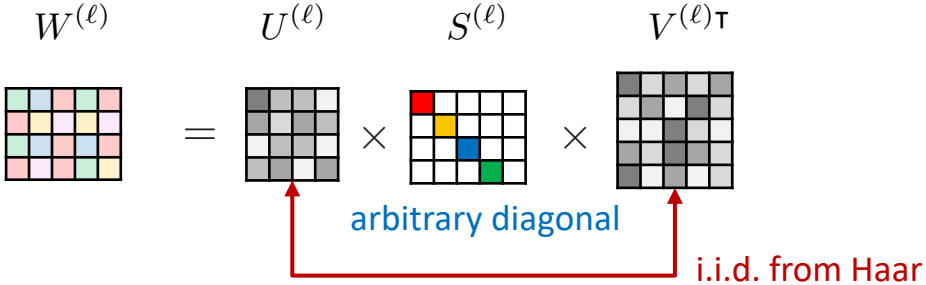


Given

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

- Kwon factorized input distribution $P_X(\mathbf{x}) = \prod_i P_0(x_i)$

- Orthogonally invariant random weight matrices



Result: The entropy of each layer in the thermodynamic (=large size) limit is

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

M. G., A. Manoel, C. Luneau, J. Barbier, N. Macris, F. Krzakala, L. Zdeborová - arXiv:1805.09785 – NIPS 2018

Replica formula for entropies: a glance at the replica symmetric potential

Formula:

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

Potential:

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

- Asymptotic prediction in the size of the network
- Additive over the layers
- Expressed via scalar entropies and mutual informations
- Depending on the weights through their spectrum

Efficient implementation of extremization: sphinxteam / dnner

M. G., A. Manoel, C. Luneau, J. Barbier, N. Macris, F. Krzakala, L. Zdeborová - arXiv:1805.09785 – NIPS 2018

Where does the replica formula come from ?

- **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, ...)

This talk: extension to multilayer

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

$$\mathbf{y} = \text{sign}(W\mathbf{x}) \text{ with } W = USV^T$$

- **<u>Multilayer</u> 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)

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

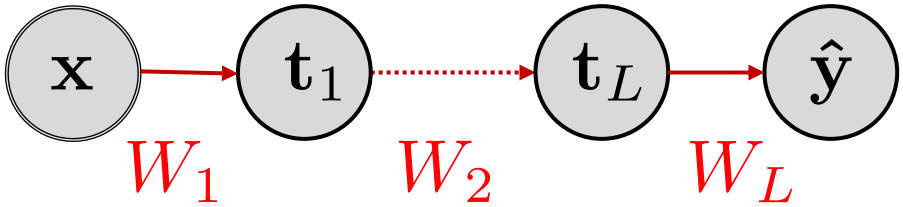
Theorem:

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.

Can we follow information during **learning** with the replica formula?

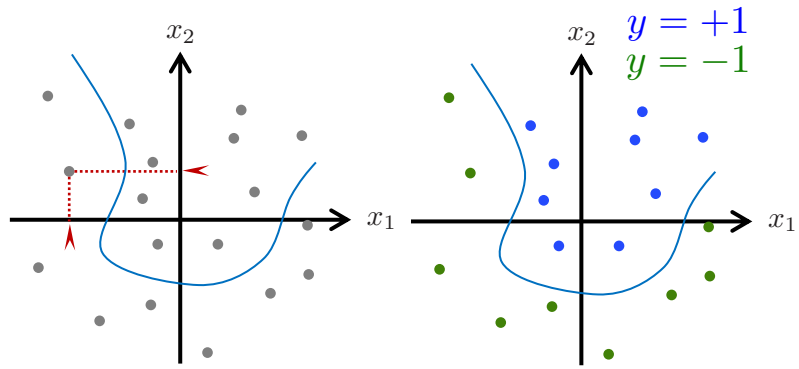


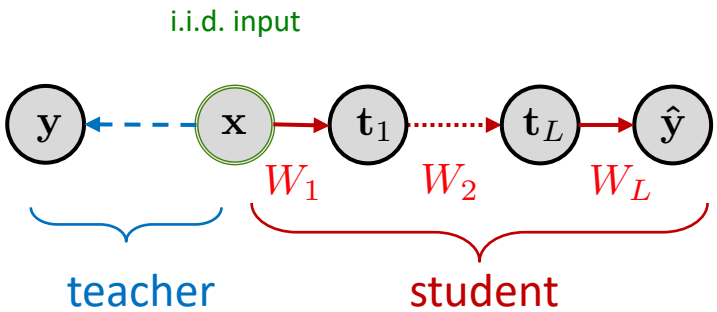
$$\lim_{N_0 \to \infty} \frac{1}{N_0} H(\mathbf{T}_\ell) = \min \operatorname*{extr}_{\mathbf{A},\mathbf{V},\tilde{\mathbf{A}},\tilde{\mathbf{V}}} \phi_\ell(\mathbf{A}, \mathbf{V}, \tilde{\mathbf{A}}, \tilde{\mathbf{V}})$$

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

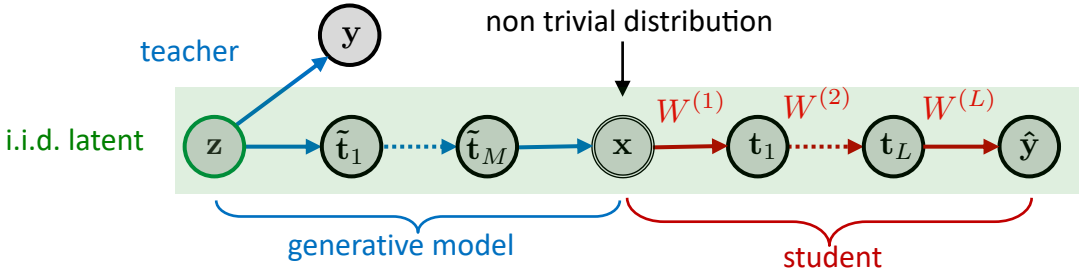
Synthetic data framework for supervised learning

- **Teacher – student scenario: Input data with separable prior** $P_X(\mathbf{x}) = \prod_i P_0(x_i)$





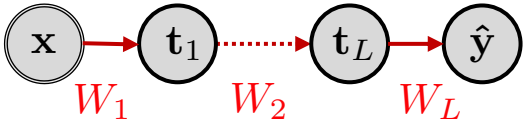
- **Generalization: including generative models**



e.g. Variational Auto Encoders Kingma et. al 2014, Rezende et al 2014
e.g. Generative Adversarial models Goodfellow et al. 2014

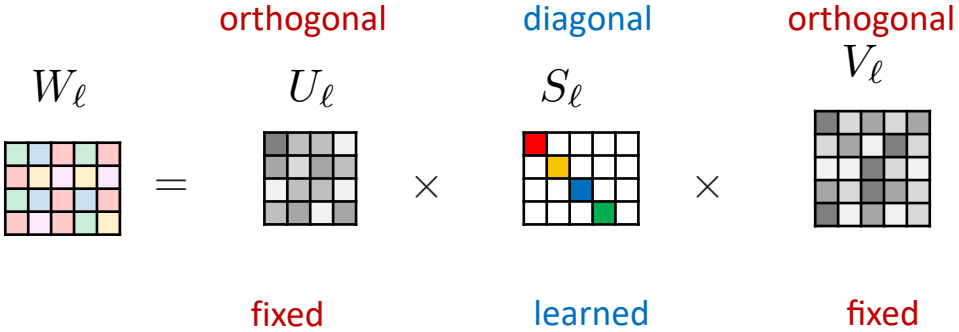
Learning orthogonally invariant weight matrices "USV-layers"

student

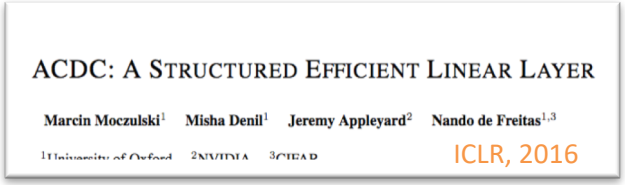


How to guarantee orthogonal invariance during learning ?

- 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]

ICLR, 2016

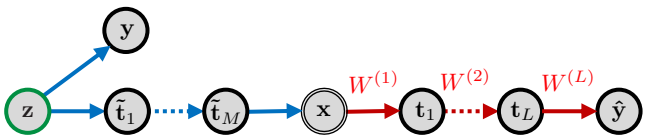
permutation diagonal

$$W_\ell = A_\ell \, C \, D_\ell \, C^{-1}$$

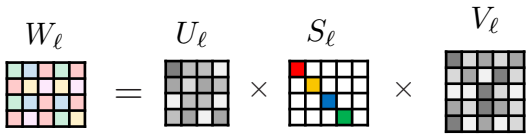
cosine transforms

Learning experiments framework

- **Be careful with**
  - **1) Data: synthetic data sets**



  - **2) Weights: preserving orthogonal invariance**



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

- **Compute entropies in large DNNs throughout training with replicas**

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

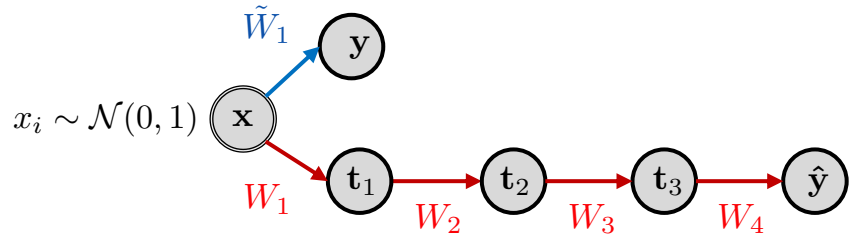
Code available in GitHub sphinxteam / **dnner**

- **Next: Numerical experiments**

Safety-check on linear networks

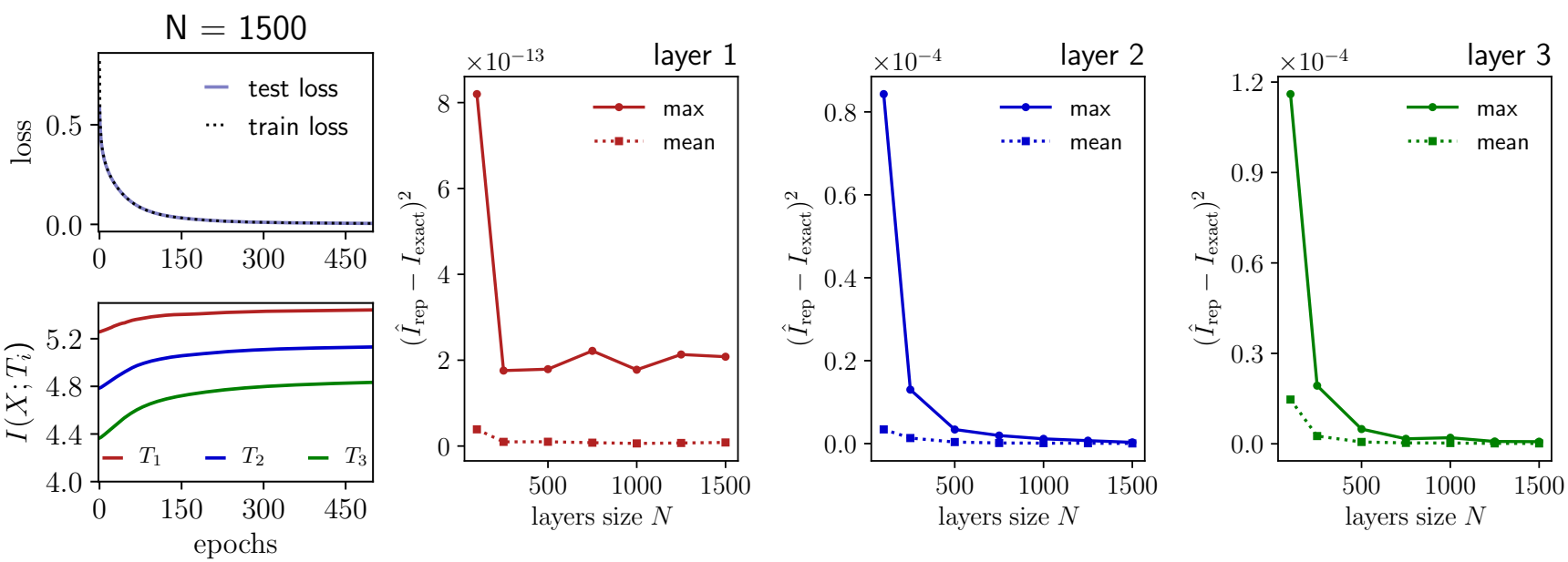
Model: $\mathbf{t}_\ell = \mathbf{W}^{(\ell)} \cdots \mathbf{W}^{(1)} \mathbf{x} + \epsilon$

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



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

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

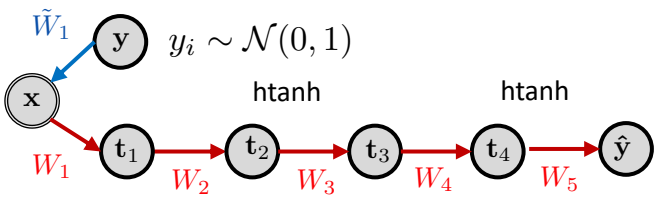


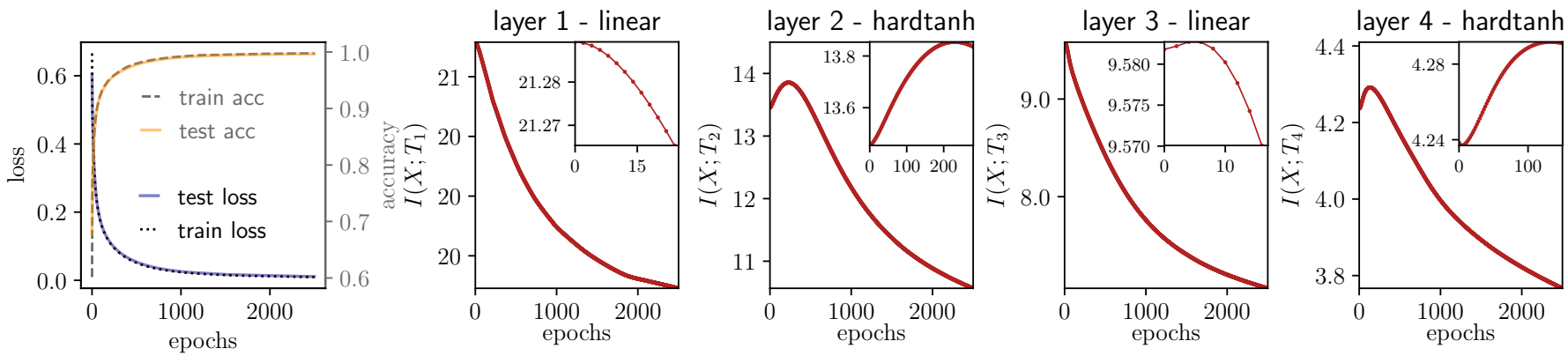
Exploratory experiment: A binary classification example

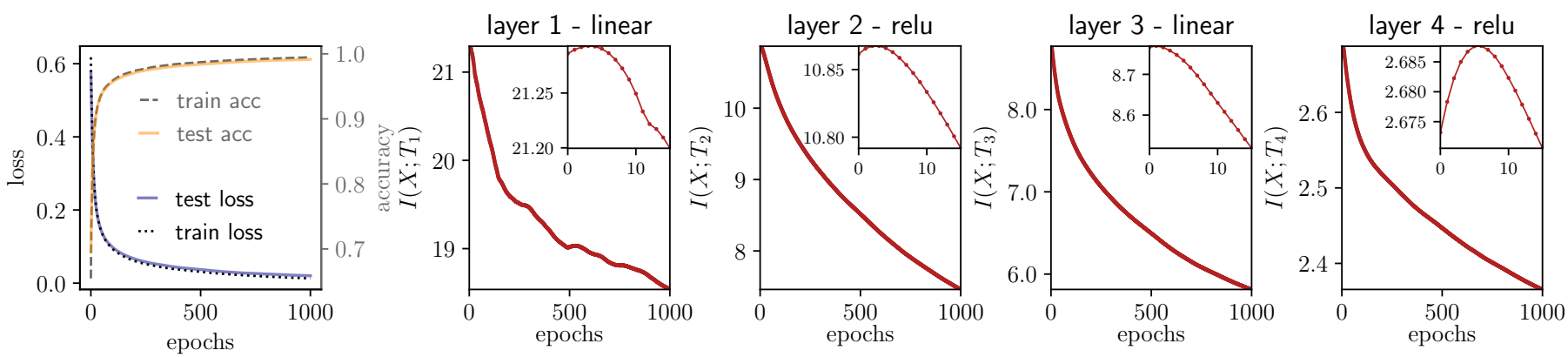
Model: linear generative model – non-linear student

$$\mathbf{t}_\ell = f(\mathbf{W}^{(\ell)}..f(\mathbf{W}^{(1)}\mathbf{x})..+\epsilon) \quad \epsilon_i \sim \mathcal{N}(0, 10^{-5})$$

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







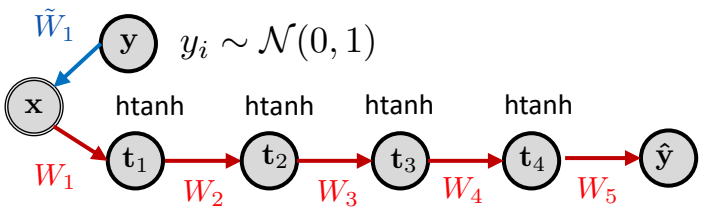
Compression as well with non-saturated activation functions

Exploratory: A second binary classification example

Model: linear generative model – non-linear student

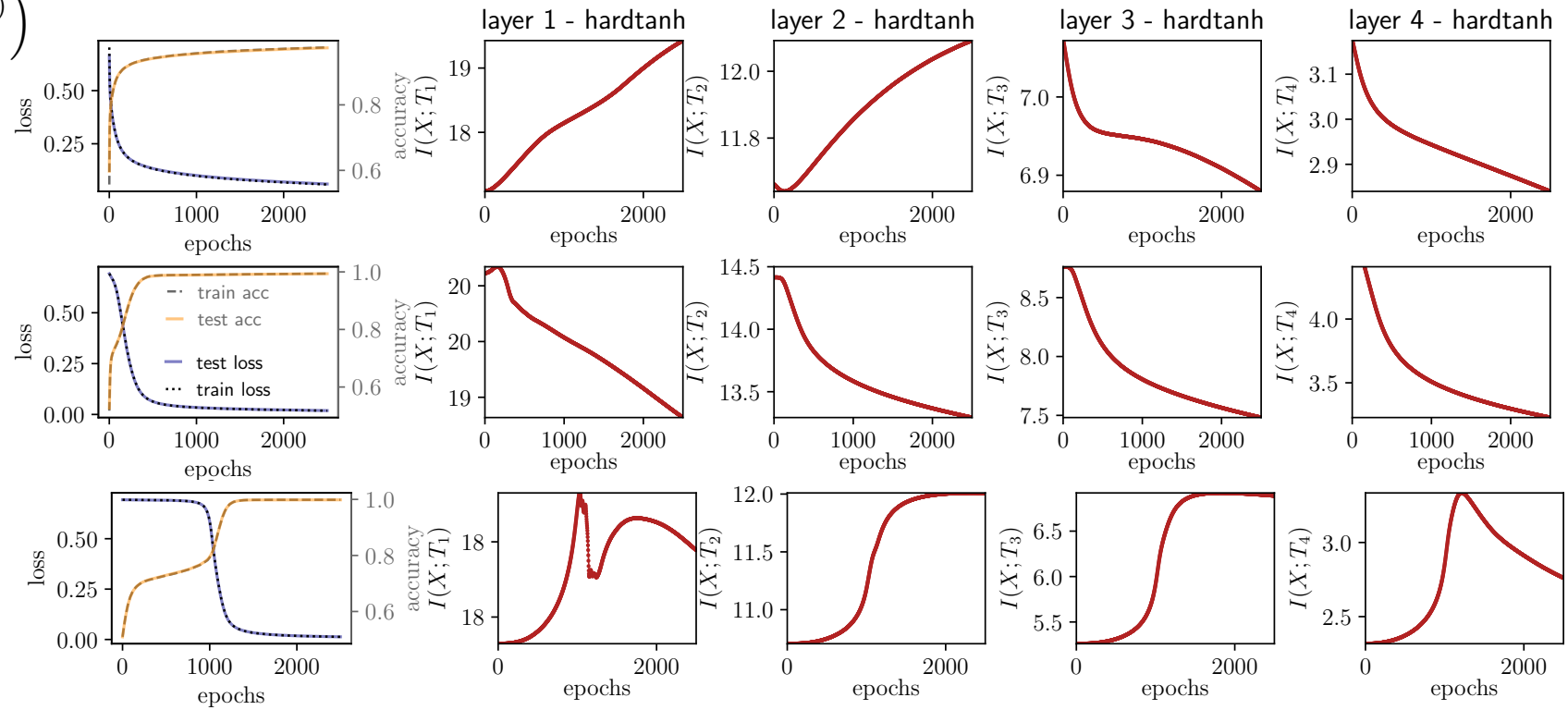
$$\mathbf{t}_\ell = f(\mathbf{W}^{(\ell)}..f(\mathbf{W}^{(1)}\mathbf{x})..+\epsilon) \quad \epsilon_i \sim \mathcal{N}(0, 10^{-5})$$

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



initial

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



Picture still not simple

How general are compressions ?
Recap from the replica formula strategy

With the replica method we have access to

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

Our experiments

- **Random weights:**
  - different non-linearities behaving differently
- **Learning experiments:**
  - Compressions even with ReLU
  - Yet not systematic

Conclusion & Perspectives

- Validate hypothesis that SGD optimizes the information bottleneck Lagrangian for stochastic NN is very hard in general

- Cases accessible to experiments appear not to follow a clear trend

- Interpretation for deterministic networks remains an open question

Yet, still a lot of directions under scrutiny by the community

- Many training algorithms for supervised and unsupervised learning using information related concepts

  M. Chalk et al.. Relevant sparse codes with variational information bottleneck 2016. A. Achille et al. Information Dropout: Learning Optimal Representations Through Noisy Computation 2018. - A. Alemi et al. Deep variational information bottleneck. 2017. - A. Achille et al. Emergence of Invariance and Disentangling in Deep Representations. In ICML 2017 Workshop on Principled Approaches to Deep Learning, 2017.- A. Kolchinsky, et al. Nonlinear Information Bottleneck. arXiv:1705.02436, 2017. - M.I. Belghazi et al. MINE: Mutual Information Neural Estimation. 2018 - S. Zhao et al . InfoVAE: Information Maximizing Variational Autoencoders. arXiv:1706.02262, 2017. etc.

- Interesting discussions about other forms of information or interpretation of the information

  Goldfeld, Z. et al. Estimating Information Flow in Neural Networks. (2018), Achille, A et al, S. Critial learning periods in deep networks. (2018), etc.

THANK YOU !

And to collaborators

Andre Manoel (INRIA Saclay, Owkin),
Clément Luneau, Jean Barbier, Nicolas Macris (EPFL),
Lenka Zdeborová (CEA Saclay),
Florent Krzakala (LPS ENS) & Léon Bottou (FAIR NY)

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. A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox, “on the Information Bottleneck,” Iclr 2018, pp. 1–27, 2018.
21. Shinzato, T. & Kabashima, Y. Perceptron capacity revisited: Classification ability for correlated patterns. *J. Phys. A Math. Theor.* 41, (2008).
22. Shinzato, T. & Kabashima, Y. Learning from correlated patterns by simple perceptrons. *J. Phys. A Math. Theor.* 42, (2009).