Title: The Information Theory of Deep Neural Networks: The statistical physics aspects

Date: Apr 25, 2018 02:00 PM

URL: http://pirsa.org/18040050

Abstract: The surprising success of learning with deep neural networks poses two fundamental challenges: understanding why these networks work so well and what this success tells us about the nature of intelligence and our biological brain. Our recent Information Theory of Deep Learning shows that large deep networks achieve the optimal tradeoff between training size and accuracy, and that this optimality is achieved through the noise in the learning process.

In this talk, I will focus on the statistical physics aspects of our theory and the interaction between the stochastic dynamics of the training algorithm (Stochastic Gradient Descent) and the phase structure of the Information Bottleneck problem. Specifically, I will describe the connections between the phase transition and the final location and representation of the hidden layers, and the role of these phase transitions in determining the weights of the network.

Based partly on joint works with Ravid Shwartz-Ziv, Noga Zaslavsky, and Shlomi Agmon.

The Information Bottleneck Theory of [simple] Deep Learning

Perimeter Institute, Waterloo, April 2018

Naftali Tishby

School of Engineering and Computer Science

The Edmond & Lily Safra Center for Brain Sciences

Hebrew University, Jerusalem, Israel

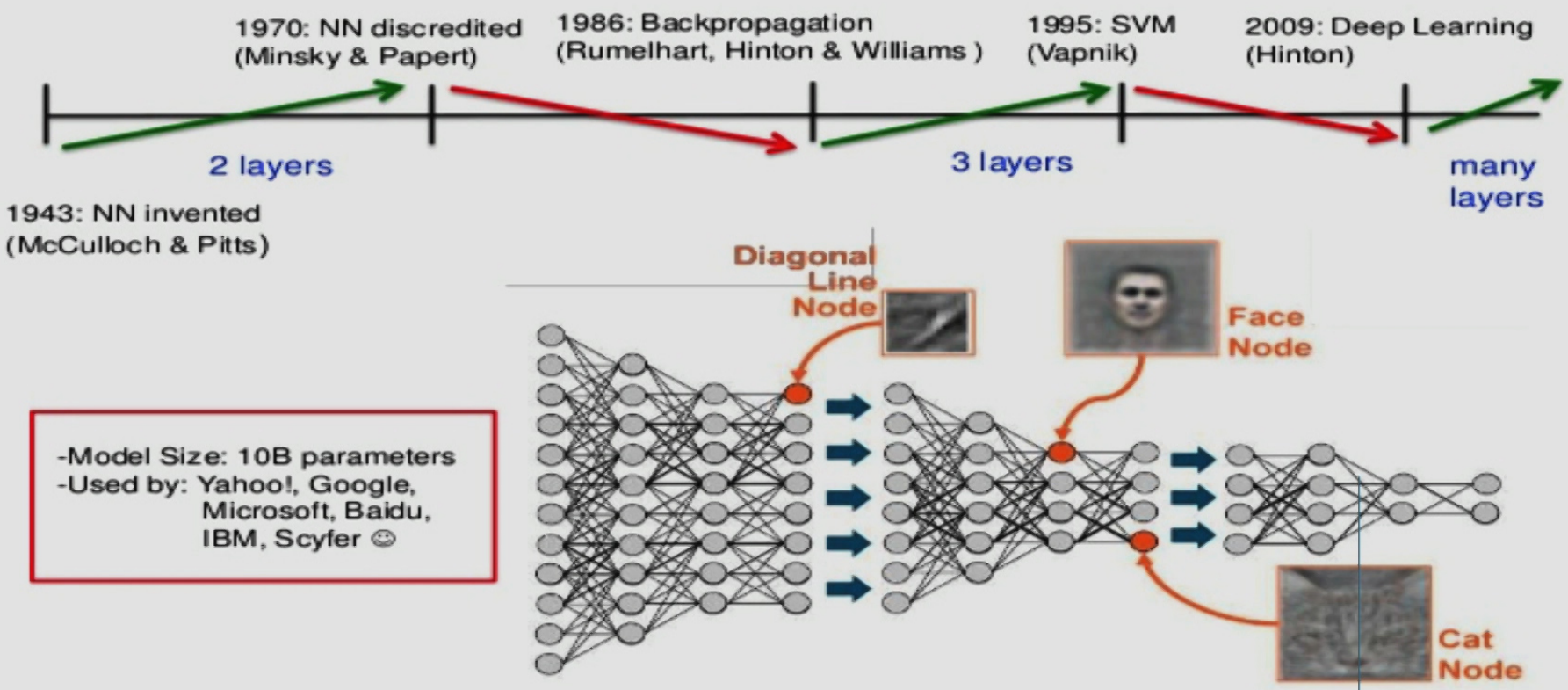
Noga Zaslavsky

Ravid Schwartz-Ziv

The Gatsby Charitable Foundation

ICRI-CI

Deep Learning: Neural-Nets strike back



We begin to obtain some new understanding…

We combine 3 different ingredients:

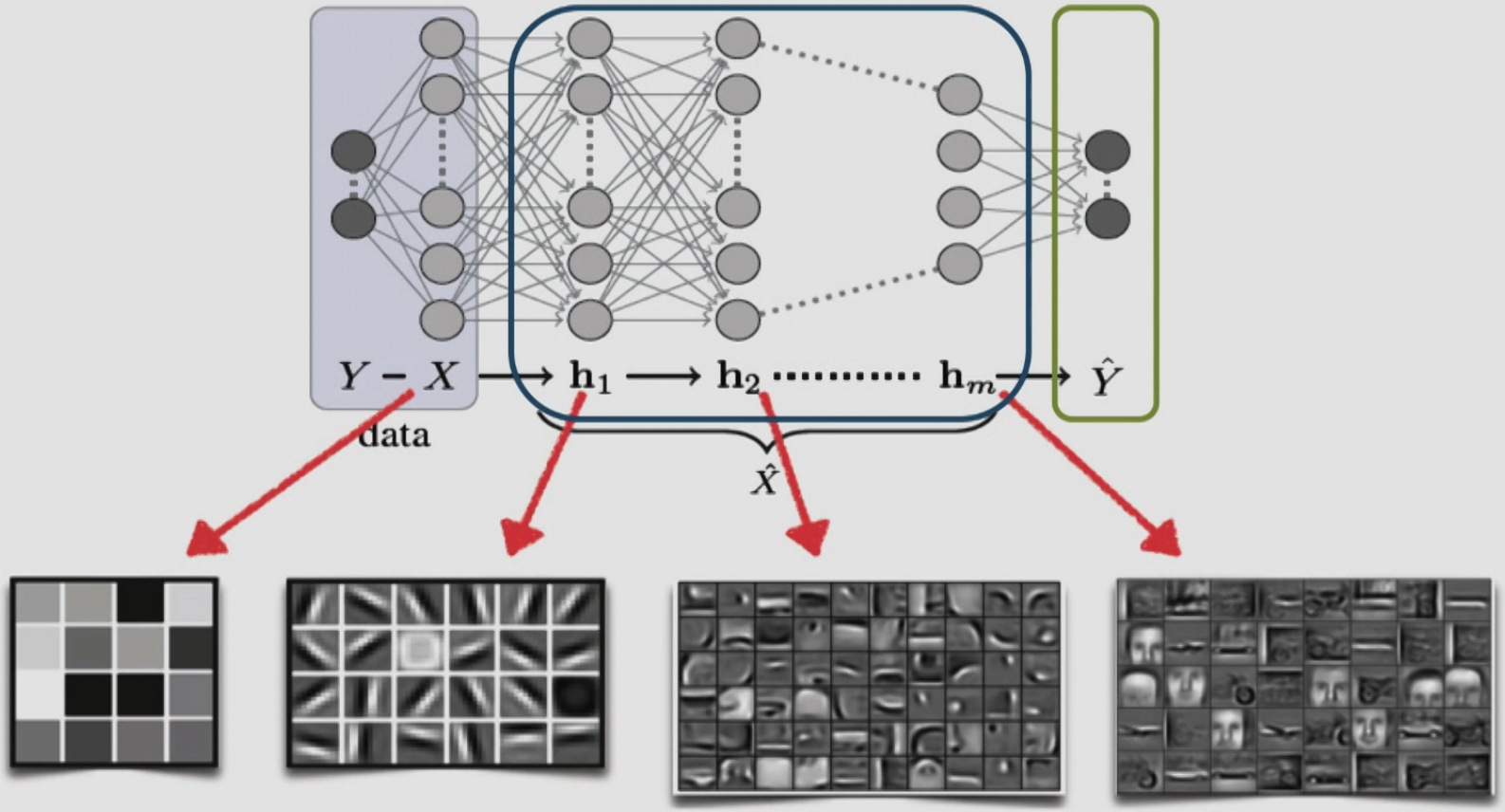
- **Rethinking Statistical Learning Theory**
  - Worse case PAC bounds → typical case architecture free bounds…
  - From expressivity/Hypothesis class → Input Compression bounds
- **Information Theory** (statistical mechanics…)
  - Large scale learning – *Typical* input patterns
  - → Concentration of the Mutual Information values
  - → Huge parameter space - exponentially many optimal solutions
- **Stochastic dynamics of the training process**
  - Convergence of SGD to locally-Gibbs (Max Entropy) weight distribution
  - → The mechanism of representation compression in Deep Learning
  - → Convergence times – explains the benefit of the hidden layers

Known issues & important reservations

Objections to the theory:

- **Information estimation [requires quantization or noise, not scalable? …]**
  - **NOT NEEDED FOR THE THEORY OR TRAINING**, PROVABLE FOR SGD, used only as an illustration!
  - Requires finite precision or quantization – CORRECT!
  - Mutual Information values concentrate & become MORE stable the larger the problem!
- **Compression/Information loss not necessary [ResNets, RevNets,i-RevNets,…]**
  - Compression comes from unit saturation, not seen with RelU's (Saxe 2018) – WRONG!
  - Indeed, good generalization can be achieved without apparent layer compression.
  - Similar to the classical physics paradox of reversible microscopic laws & entropy increase…
  - No "forgetting" of non-informative features (really?)
- **Stochastic Gradients not needed [no convergence to local weight Gibbs distribution]**
  - Good generalization achieved without stochastic gradients in INFINITE TIMES! How?
  - Convergence to Gibbs (MaxEnt) distribution is only local (in each layer).
  - The benefits of the stochasticity is dynamical (computational), but also in saving training data!
  - There is important INFORMATION in the mini-batch fluctuations! Too large batches don't generalize well.
- **Is the IB bound relevant?**
  - It actually gives concreate predictions and interpretation of the layers & weights.
  - May explain biological neural network organization… our ultimate motivation.

Deep Neural Nets and Information Theory ??



Some Information Theory basics

- The KL-distribution divergence:

for any two distributions p(x) & q(x) over X:

$$D[p(x)\,\|\,q(x)] = \sum_x p(x)\log\frac{p(x)}{q(x)} \geq 0$$

- The Mutual Information:Type equation here.

for any two random variables, X , Y:

$$I(X;Y) = D[p(x,y)\,\|\,p(x)p(y)] = D[p(x\,|\,y)\,\|\,p(x)] = D[p(y\,|\,x)\,\|\,p(y)] = \mathrm{H(X)} - \mathrm{H(X\,|\,Y)}$$

- Data Processing Inequality (DPI) & Invariance:

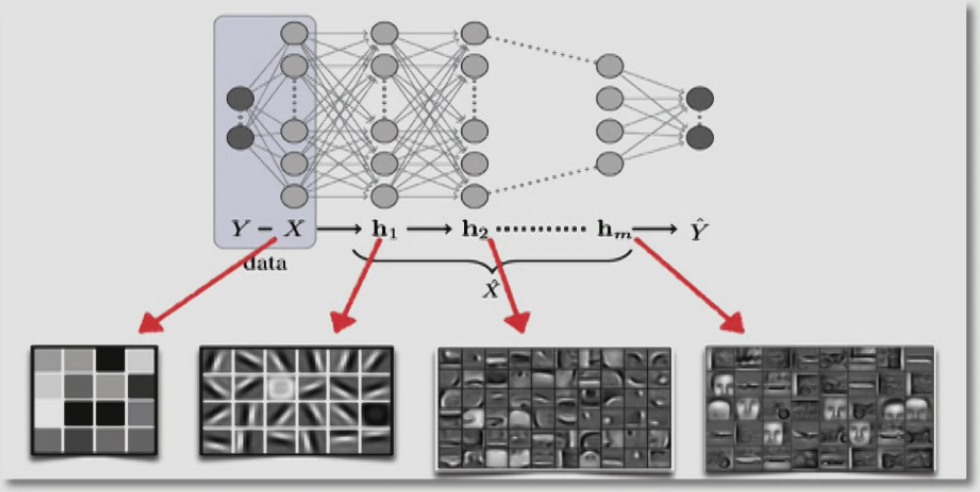
for any Markov chain: $X \rightarrow Y \rightarrow Z$:

$$I(X;Y) \geq I(X;Z)$$

Reparametrization Invariance, for invertible ϕ, ψ:

$$I(X;Y) = I(\phi(X);\psi(Y))$$

What do the DNN Layers represent?

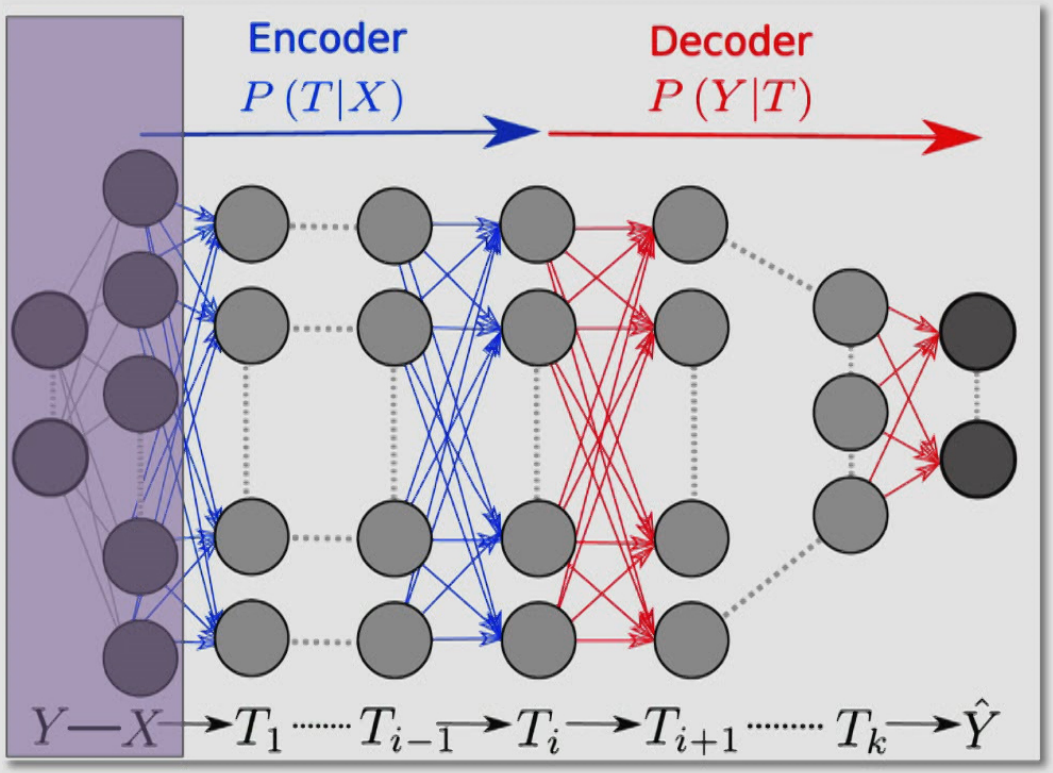


Data Processing Inequalities:

$$H(X) \geq I(X;h_i) \geq I(X;h_{i+1}) \geq I(X;h_{i+2}) \geq \ldots$$
$$I(X;Y) \geq I(h_i;Y) \geq I(h_{i+1};Y) \geq I(h_{i+2};Y) \geq \ldots$$

- A Markov chain of topologically distinct [soft] **partitions** of the input variable X.
- Successive Refinement of Relevant Information
- Individual neurons can be easily "scrambled" within each layer

Each layer is characterized by its Encoder & Decoder Information

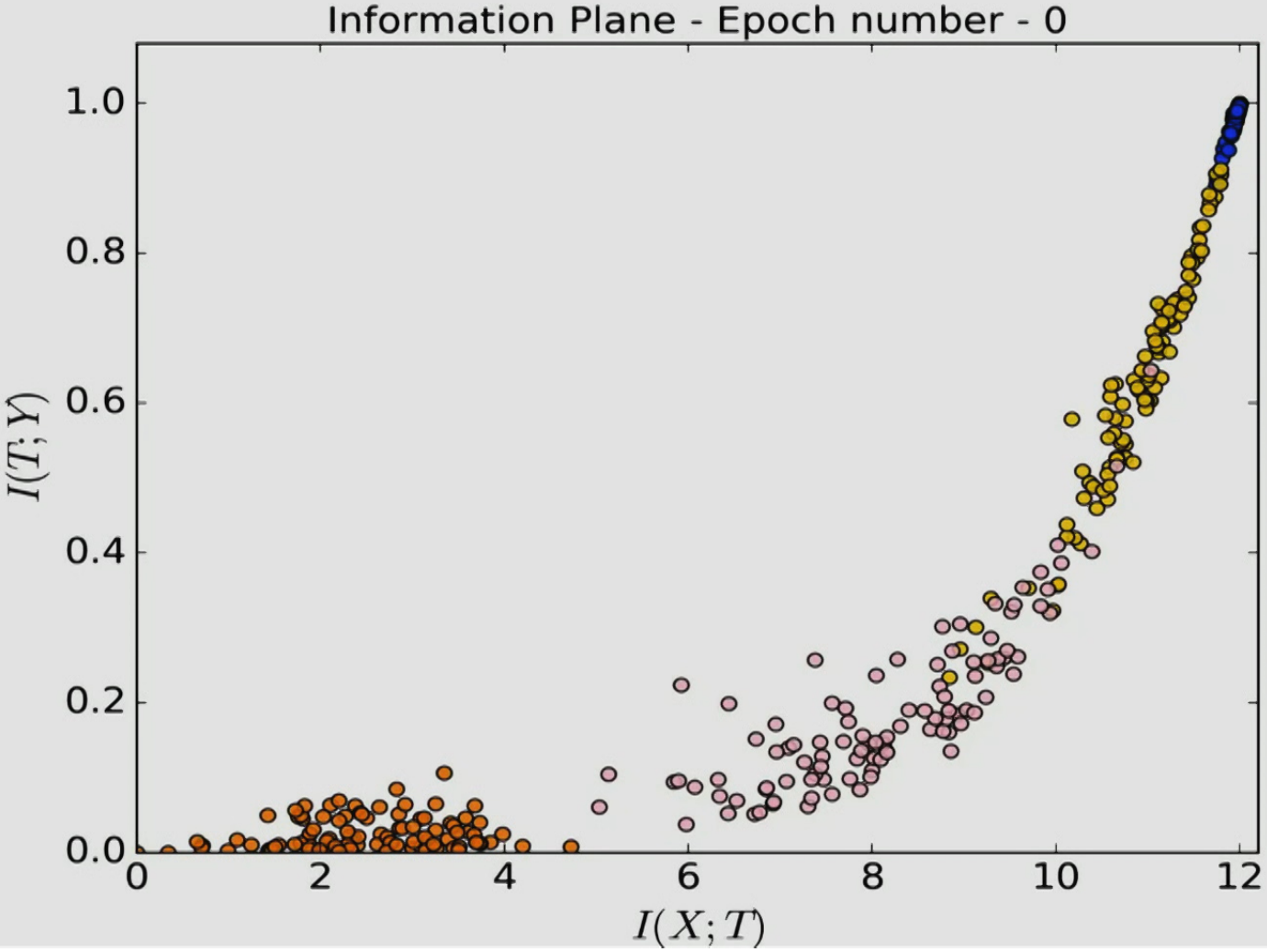


Theorem (Information Plane):

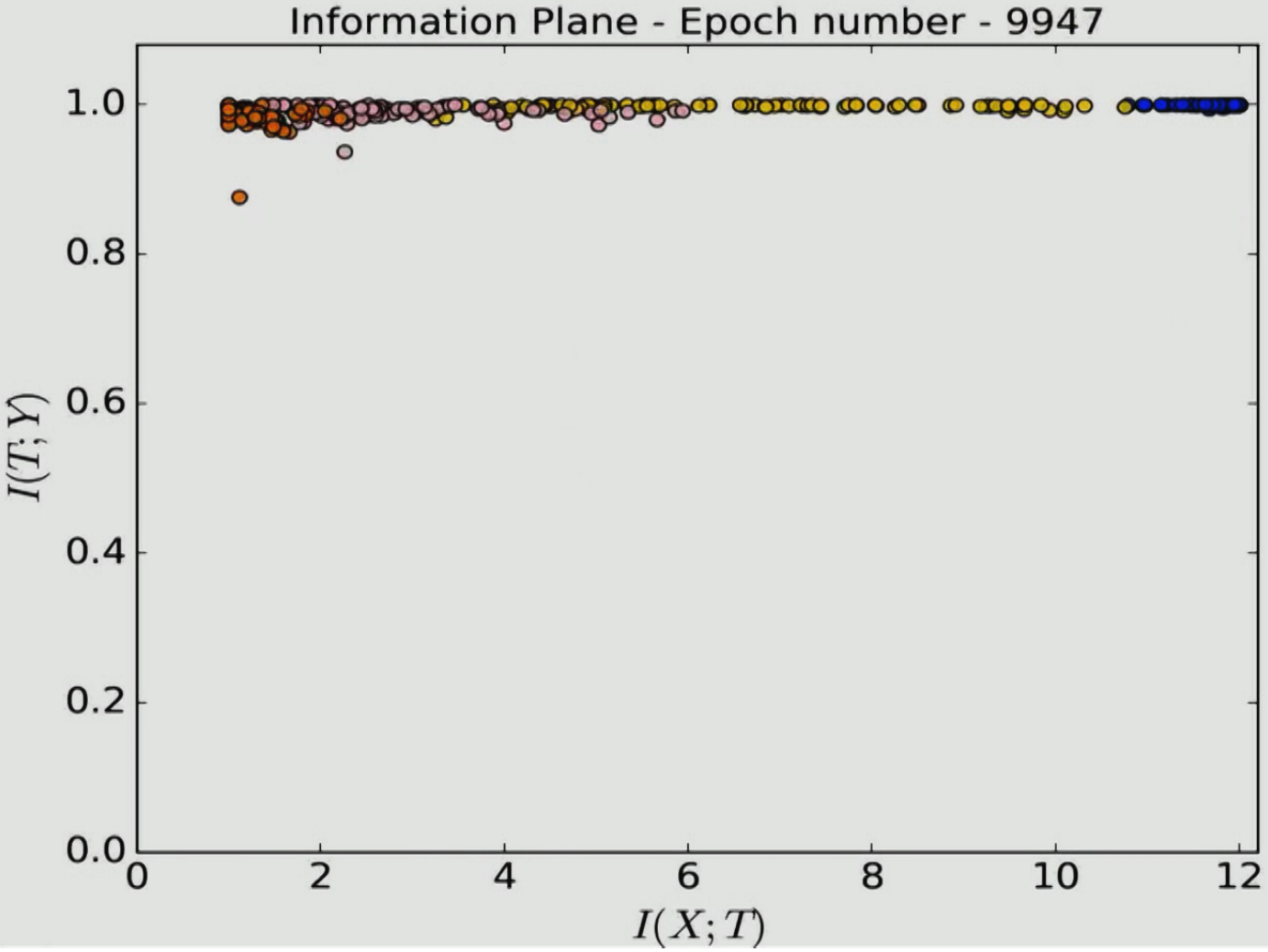
*For large typical **X**, the sample complexity of a DNN is completely determined by the encoder mutual information, **I(X;T)**, of the last hidden layer; the accuracy (generalization error) is determined by the decoder information, **I(T;Y)**, of the last hidden layer.*

The complexity of the problem shifts from the decoder to the encoder, across the layers…

100 DNN Layers in Info-Plane without averaging



100 DNN Layers in Info-Plane without averaging



- Is this the general picture?
- Why do the MI values concentrate?
- What do they mean?
- What governs their dynamics?

Rethinking Learning Theory…

What are "large typical" patterns?

Typicality emerges when the underling pattern distribution can be asymptotically expressed as a long product of localized conditional probabilities.

E.g. Markov Random Fields, Hidden Markov Models, *pairwise* interaction Hamiltonians in physics, all common Graphical models, etc.

In our case it includes images, speech & text, long molecular sequences, signals generated by localized dynamic systems, etc.

Then, the Shannon-McMillen limit for the entropy exists:

$$\lim_{n\to\infty} -\frac{1}{n}\log p(x_1,\ldots,x_n) = H(X)$$

and *almost all* patterns are *typical* with probability:

$$p(x_1,\ldots,x_n) \simeq 2^{-nH(X)}$$

and also for large enough typical partitions, T:

$$p(x_1,\ldots,x_n \mid T) \simeq 2^{-nH(X|T)}$$

Concentration of Mutual Information

$$I(X;T)=\left\langle \log \frac{p(x|t)}{p(x)} \right\rangle_{X,T} = \left\langle \log \prod_i \frac{p(x_i|Pa(x_i),t)}{p(x_i|Pa(x_i))} \right\rangle_{X,T} = \left\langle \sum_i \log \frac{p(x_i|Pa(x_i),t)}{p(x_i|Pa(x_i))} \right\rangle_{X,T}$$

$$I(T;Y)=\left\langle \log \sum_x p(y|x)p(x|t) - \log p(y) \right\rangle_{Y,T} = \left\langle \log \left[\sum_x p(y|x) \prod_i p(x_i|Pa(x_i),t) \right] - \log p(y) \right\rangle_{Y,T}$$

Proposition:

1. Both I(T;X) and I(T;Y), as defined, concentrate, uniformly, under the partition typicality assumption.
2. Both can be estimated uniformly well (over the partitions) from a sample of p(X,Y).

Rethinking Learning Theory

"Old" Generalization bounds:

$$\varepsilon^2 < \frac{\log|H_\varepsilon| + \log 1/\delta}{2m}$$

ε - generalization error

δ - confidence

m - number of training examples

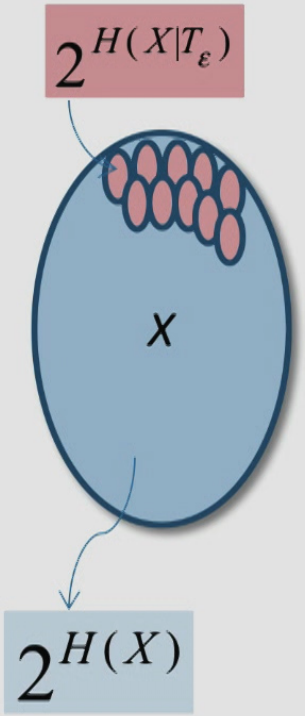
H_ε - ε-cover of the Hypothesis class

typically we assume: $|H_\varepsilon| \sim \left(\frac{1}{\varepsilon}\right)^d$

d - the class (VC,...) dimension

... Don't work for Deep Learning!

Higher expressivity - worse bound!



New: Input Compression bound:

$$|H_\varepsilon| \sim 2^{|X|} \to 2^{|T_\varepsilon|}$$

T_ε - ε-partition of the input variable X

Information Theory: $|T_\varepsilon| \sim 2^{I(T_\varepsilon;X)}$

$$\varepsilon^2 < \frac{2^{I(T_\varepsilon;X)} + \log 1/\delta}{2m}$$

... K bits of compression of X are like a factor of 2^K training examples!

The Information Bottleneck (IB) Method

(Tishby, Pereira, Bialek, 1999)

(1) Approximate Minimal Sufficient Statistics:

Markov chain: $Y \to X \to S(X) \to \hat{X}$

$$\hat{X} = \arg\min_{S(X):I(S(X);Y)=I(X;Y)} I(S(X);X)$$

Relaxation - given $p(X,Y)$:

$$\hat{X} = \arg\min_{p(\hat{x}|x)} I(\hat{X};X) - \beta I(\hat{X};Y)\ ,\ \beta > 0$$

(Shamir, Sabato,T., TCS 2010)

Free Energy like

$I(\hat{X};X) \sim$ -Entropy

$I(\hat{X};Y) \sim$ Energy

(2) A Rate-Distortion problem with KL- divergence distortion:

$$d_{IB}(x,\hat{x}) = D[p(y|x) \,\|\, p(y|\hat{x})]$$

(Bachrach, Navot,T., COLT 2006)

(3) The ONLY distributional quantization measure which satisfy both DPI (f-divergences) and Statistical Consistency (Bregman divergences)

(Harremoes-T., ISIT 2008)

The Information Bottleneck optimality bound

(Tishby, Pereira, Bialek, 1999)

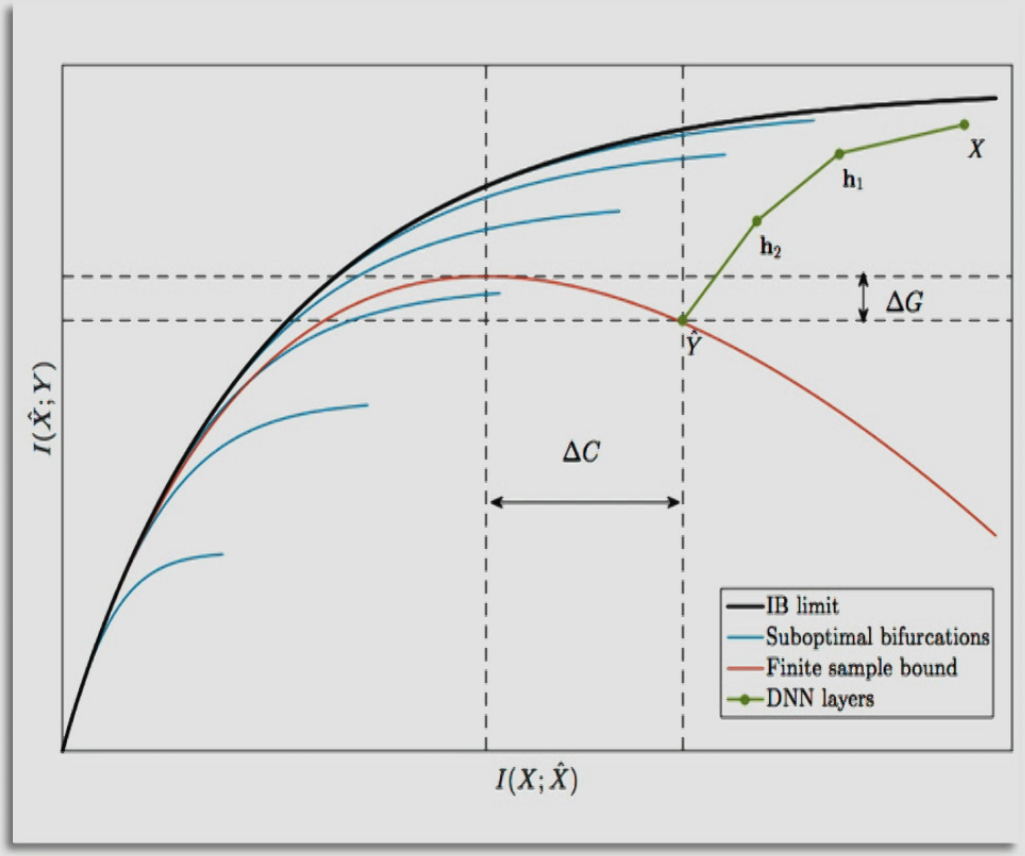
The IB bound optimality equations:

$$\min_{p(\hat{x}|x):\,Y\to X\to\hat{X}} I(\hat{X};X)-\beta I(\hat{X};Y)\ ,\ \ \beta>0$$

$$\begin{cases} p(x\,|\,\hat{x})=\dfrac{p(x)}{Z(x,\beta)}\exp(-\beta D[p(y\,|\,x)\,\|\,p(y\,|\,\hat{x}]) \\ Z(x,\beta)=\sum_{\hat{x}} p(\hat{x})\exp(-\beta D[p(y\,|\,x)\,\|\,p(y\,|\,\hat{x}]) \\ p(\hat{x})=\sum_{x} p(\hat{x}\,|\,x)p(x) \\ p(y\,|\,\hat{x})=\sum_{x} p(y\,|\,x)p(x\,|\,\hat{x}) \end{cases}$$

Solved by Arimoto-Blahut like iterations,

but with possibly sub-optimal solutions, bifurcations (!),



Rethinking Learning Theory…

… but we need to guarantee the label homogeneity of the ϵ -partition with finite samples. Without additional structural information on the inputs (stability, robustness, topology), we must use the stochasticity of the rule and the IB distortion measure:

The ε – partition, T_ε, is with the empirical distortion

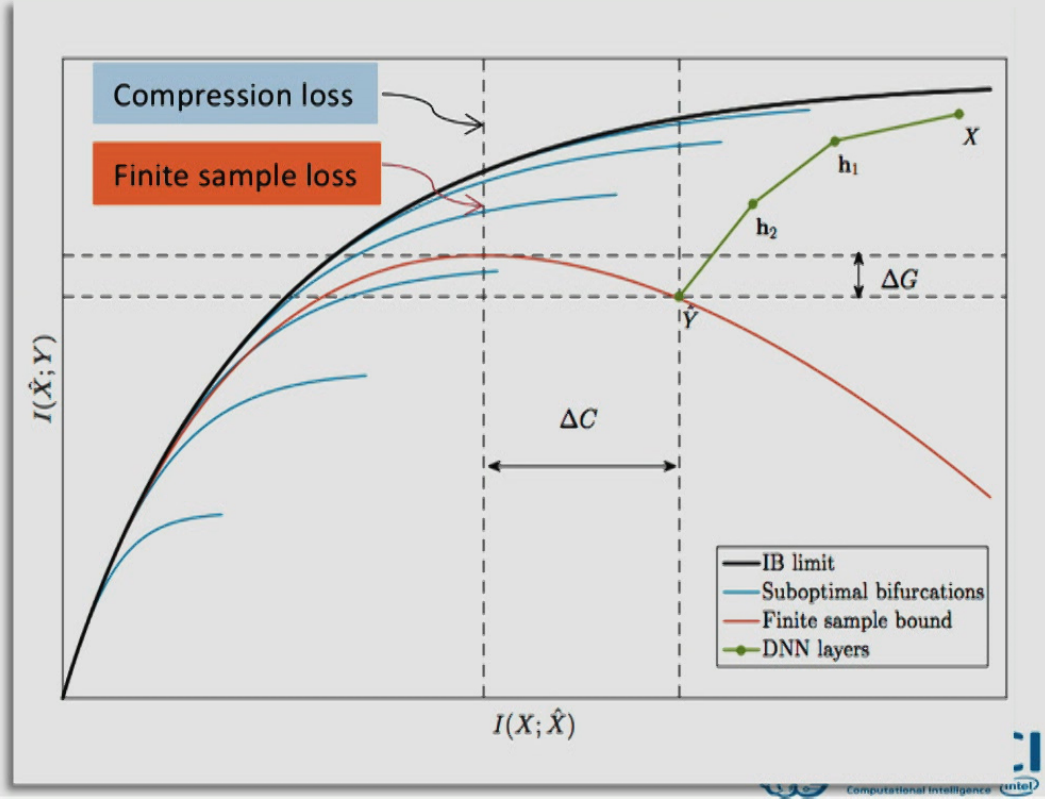
$$d_{IB}(x,t) = D[p_{emp}(y|x) \| p(y|t)]$$

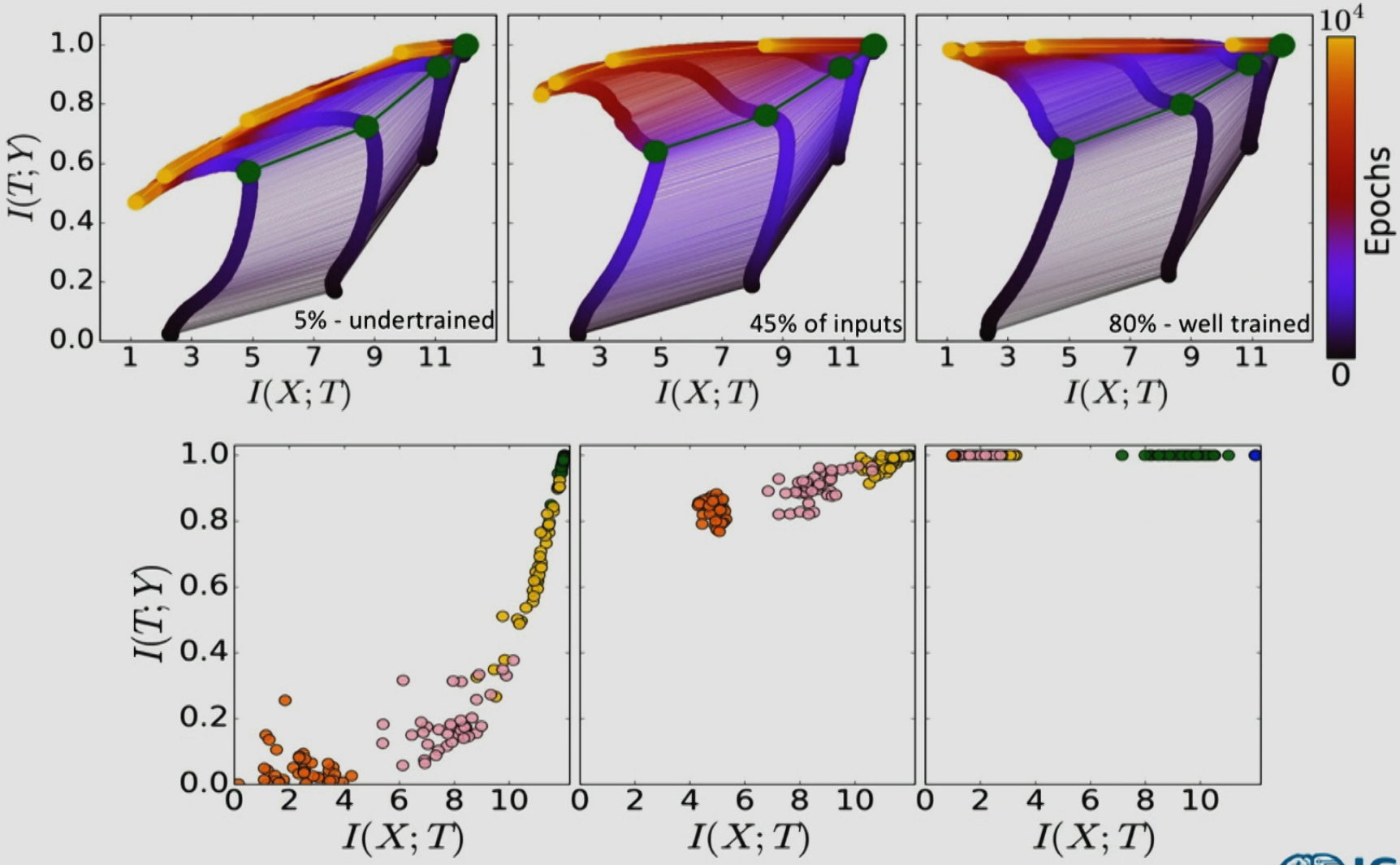
$$\text{as } \langle d_{IB} \rangle_{emp} = I(X;Y) - \hat{I}_{emp}(T;Y)$$

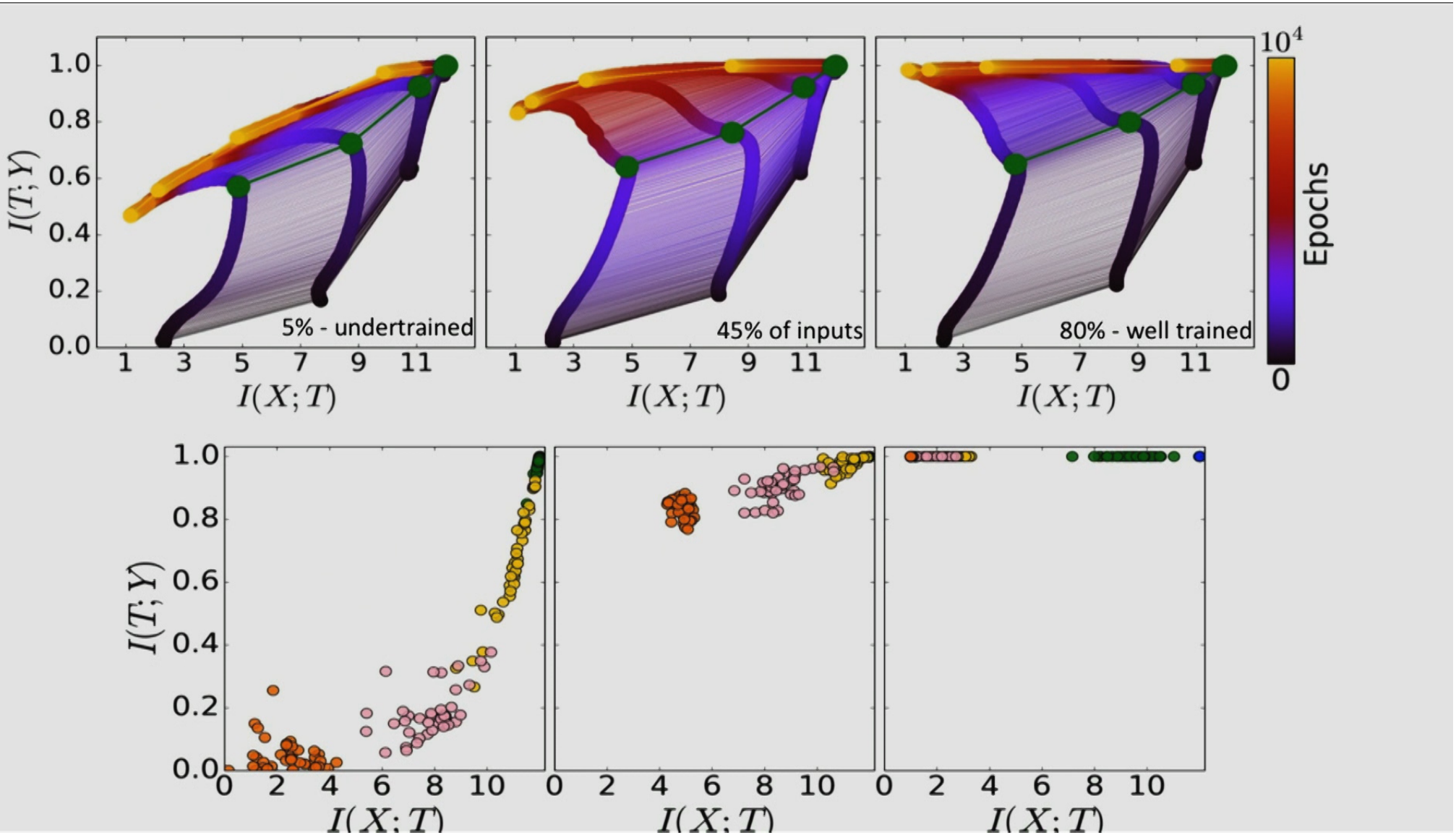
with a finite sample there is another information loss:

$$I(T;Y) \le \hat{I}_{emp}(T;Y) + O\sqrt{\left(\frac{2^{I(T;X)}|Y|}{m}\right)},$$

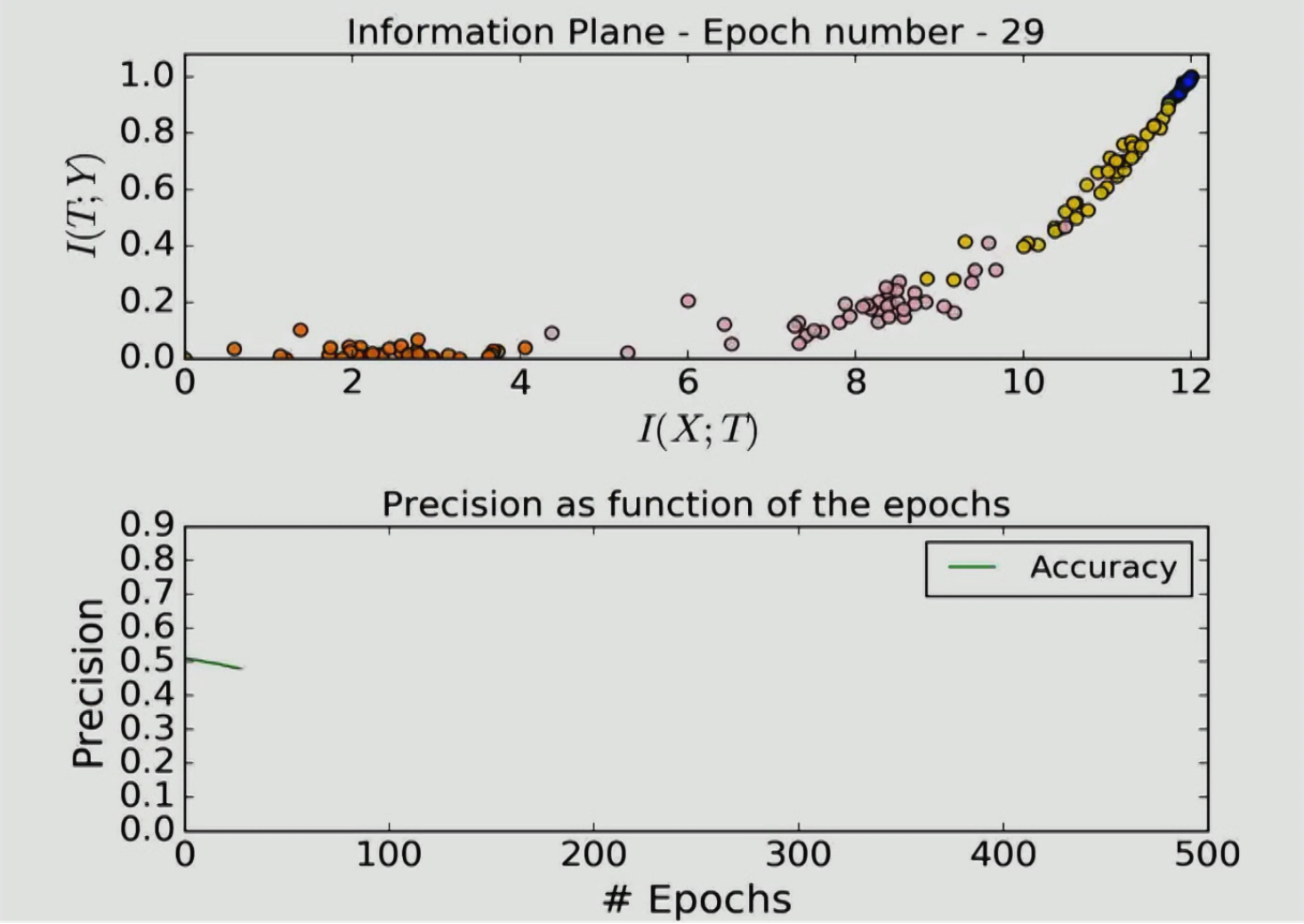
both should remain small for good generlaization!



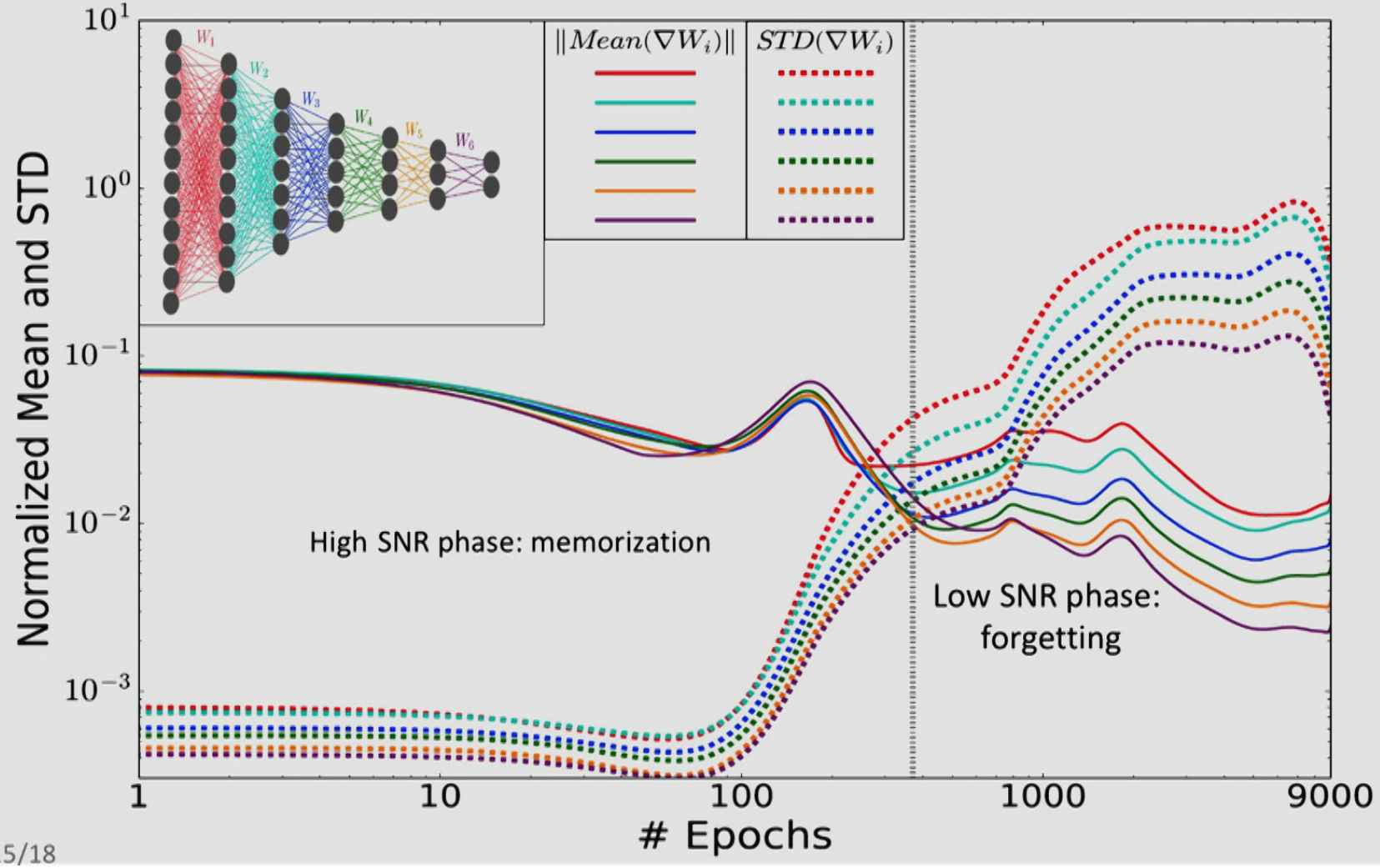

5% - undertrained
45% of inputs
80% - well trained
$I(T;Y)$
$I(X;T)$
Epochs
10^4
0
1.0
0.8
0.6
0.4
0.2
0.0
1 3 5 7 9 11
0 2 4 6 8 10


1.0
0.8
0.6
0.4
0.2
0.0
$I(T;Y)$
1 3 5 7 9 11
$I(X;T)$
5% - undertrained
45% of inputs
80% - well trained
10^4
Epochs
0
0 2 4 6 8 10

Layers paths with training/generalization error



ICRI-CI Intel Collaborative Research Institute Computational Intelligence



ICRI-CI Intel Collaborative Research Institute Computational Intelligence (intel)

Relevant and Irrelevant local dimensions

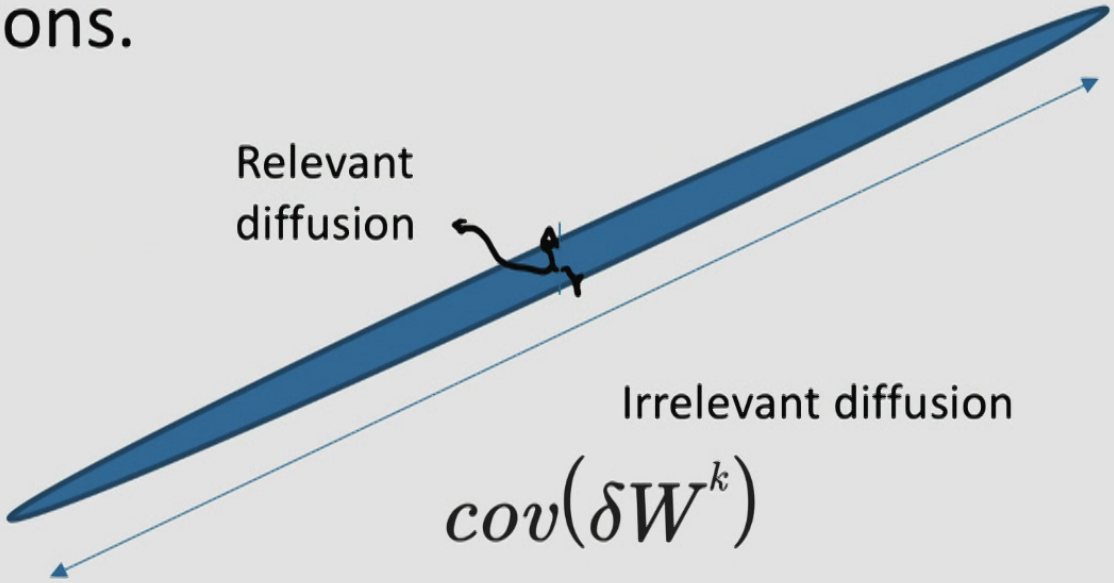
- The covariance matrix of the gradients is very narrow in the relevant local dimensions and very wide in the many other dimensions.

$$W^k \rightarrow W^k_{cca} + \delta W^k$$

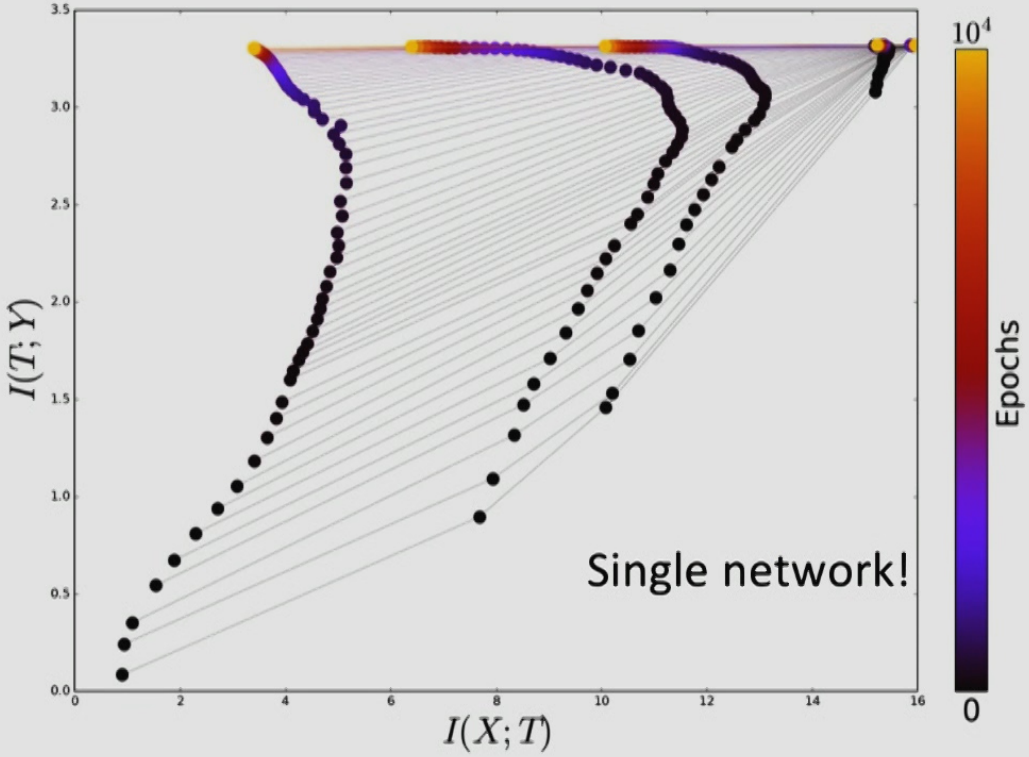
$$T^{k+1} = \sigma\left(W^k_{cca} T^k + \xi^k\right)$$

$$\xi^k = \delta W^k T^k \sim N\left(0, cov\left(\delta W^k\right)\right)$$

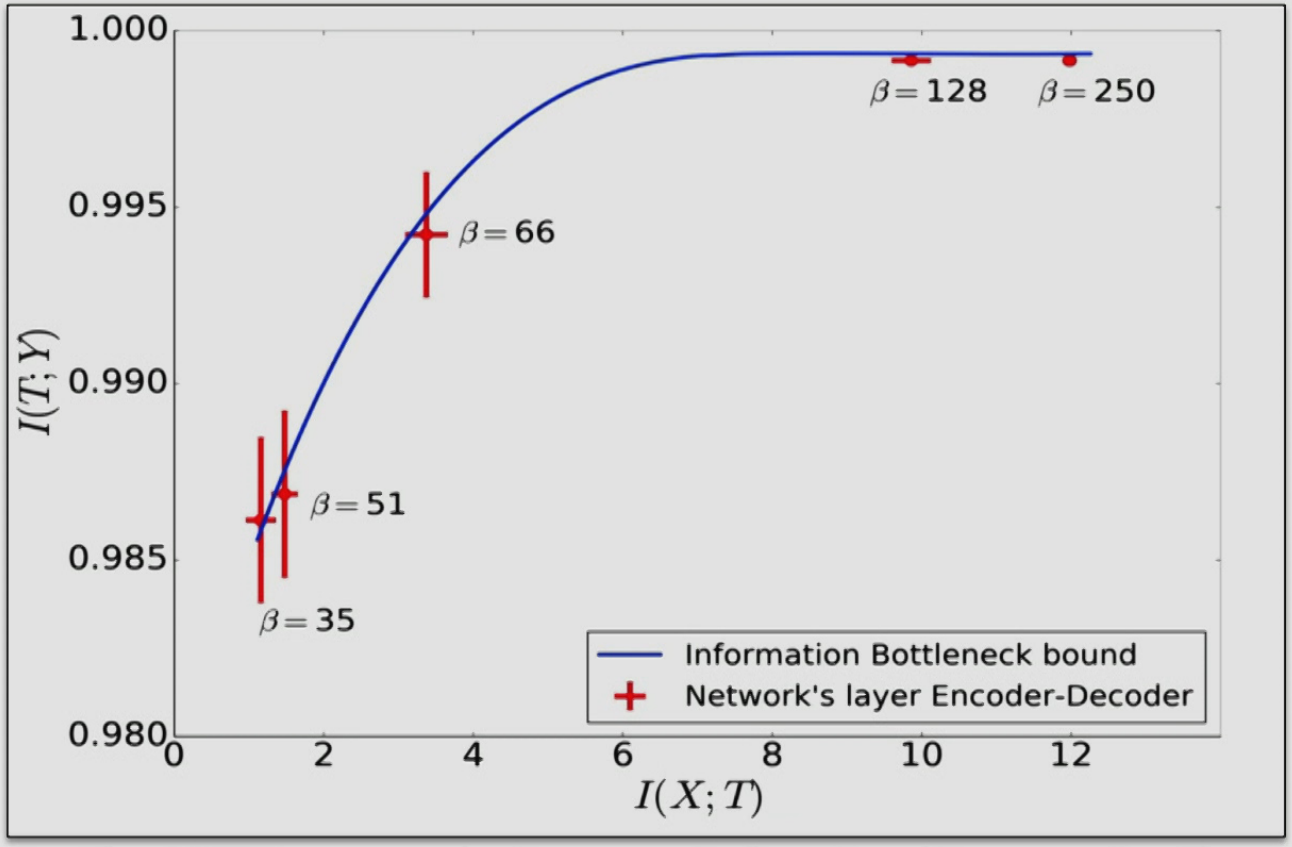
$$I\left(T^k; T^{k+1}\right) \leq \frac{1}{2}\log\left(1 + \frac{\|W^k_{cca}\|}{\|\delta W^k\|}\right)$$



… and for “Real-world” problems? Yes!



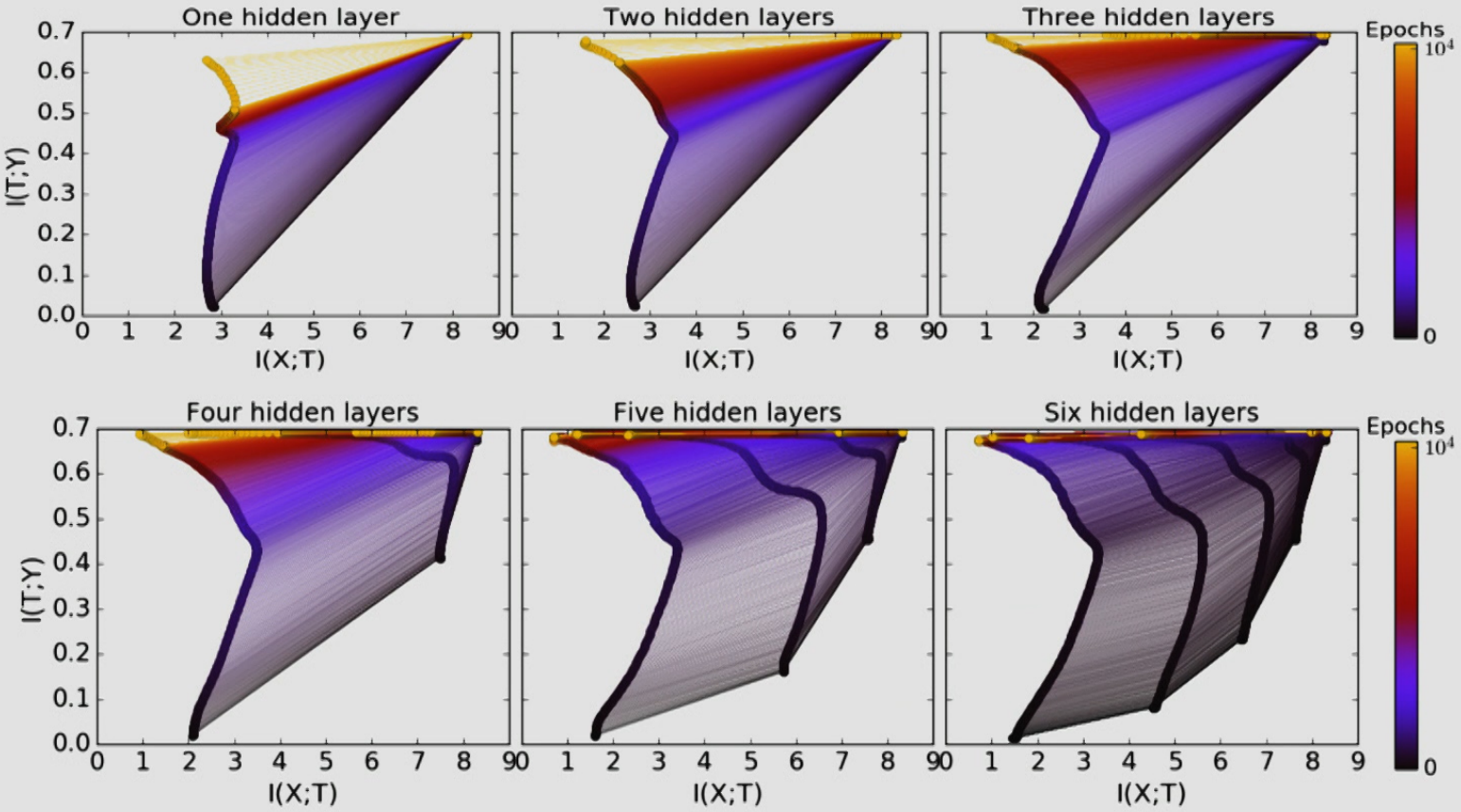
MNIST handwriting digit recognition with RelU’s a CNN architecture



- Layers of optimal DNN converge to [a successively refineable approximation of] the optimal finite-sample IB limit information-curve
- Layers must be in "different topological phases" of the IB solutions
- The DNN encoder & decoder for each layer satisfy the IB self-consistent equations

ICRI-CI Intel Collaborative Research Institute Computational Intelligence

The benefit of the hidden layers



More layers take much FEWER training epochs for good generalization.

The optimization time depend super-linearly (exponentially?) on the compressed information, delta Ix, for each layer.

Relaxation times and the benefit of the hidden layers

Noisy relaxation (SGD): $\frac{\partial W_k}{\partial t} = -\nabla E(\mathrm{W}_k) + \beta_k^{-1}\xi(t)$, layer k, $\xi \sim N(0,1)$

$\Rightarrow$ Maximum Entropy (via Focker-Planck): $P_{Gibbs}(W_k) \propto \exp\left(-\beta_k E(W_k)\right)$,

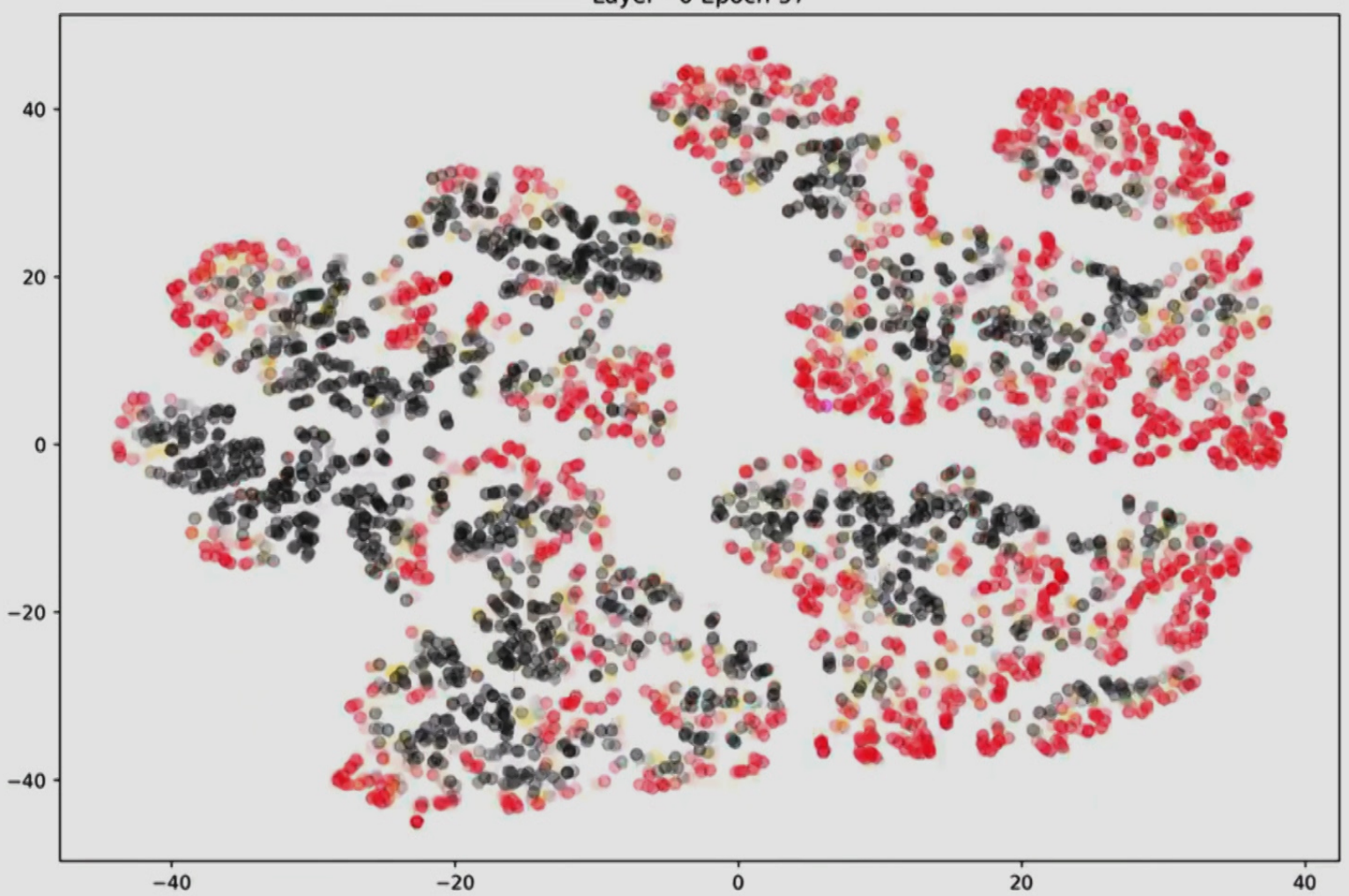
Relaxation time for non-strongly convex error: $\Delta t_k \sim \exp(\Delta S_k)$

Denote the layer compression be: $\Delta S_k = I(X;T_k) - I(X;T_{k-1})$

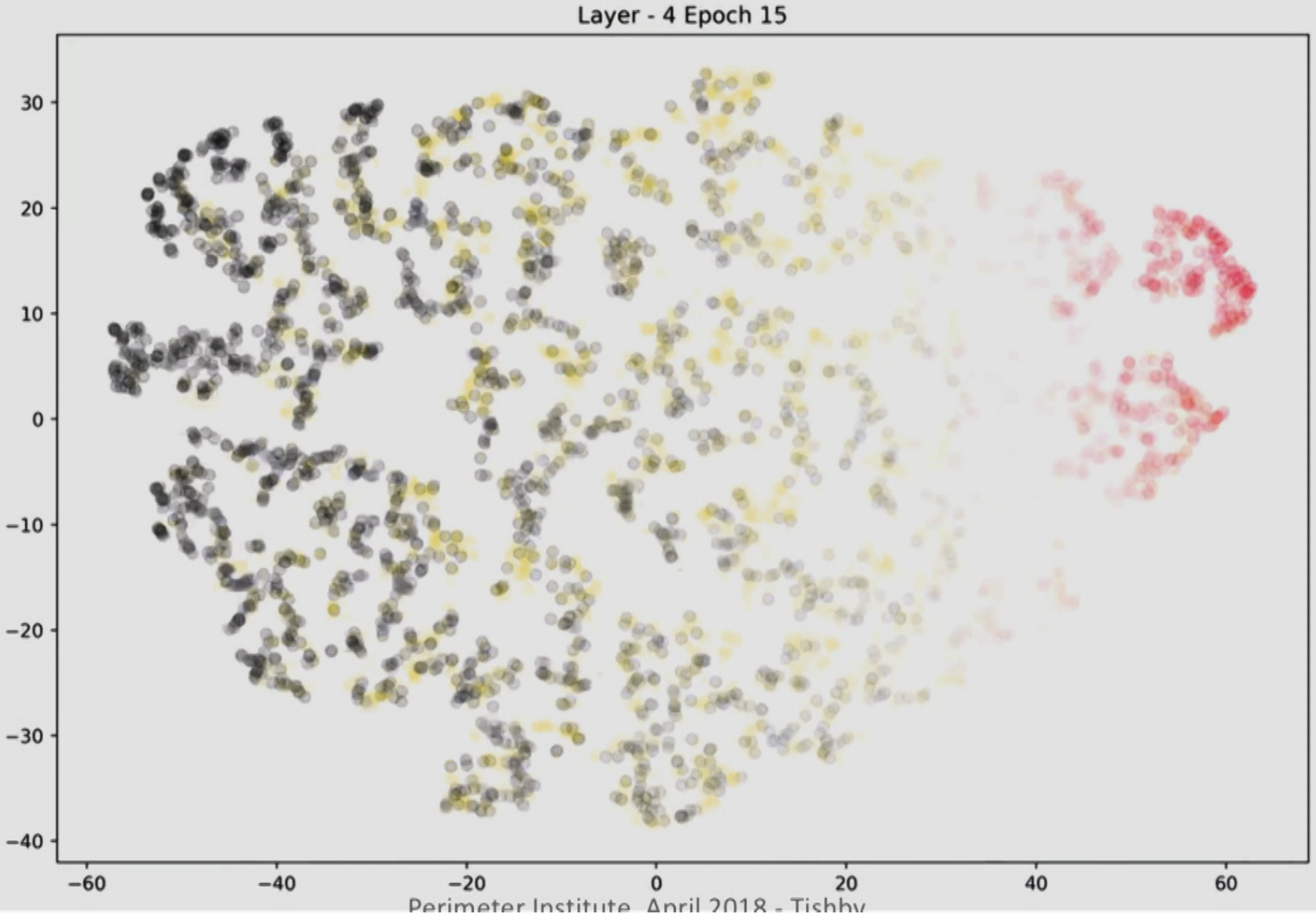
Since $\exp\left(\sum_k \Delta S_k\right) \gg \sum_k \exp\left(\Delta S_k\right) > \max_k \exp\left(\Delta S_k\right) \Rightarrow$

Exponential boost in the relaxation time with K layers!

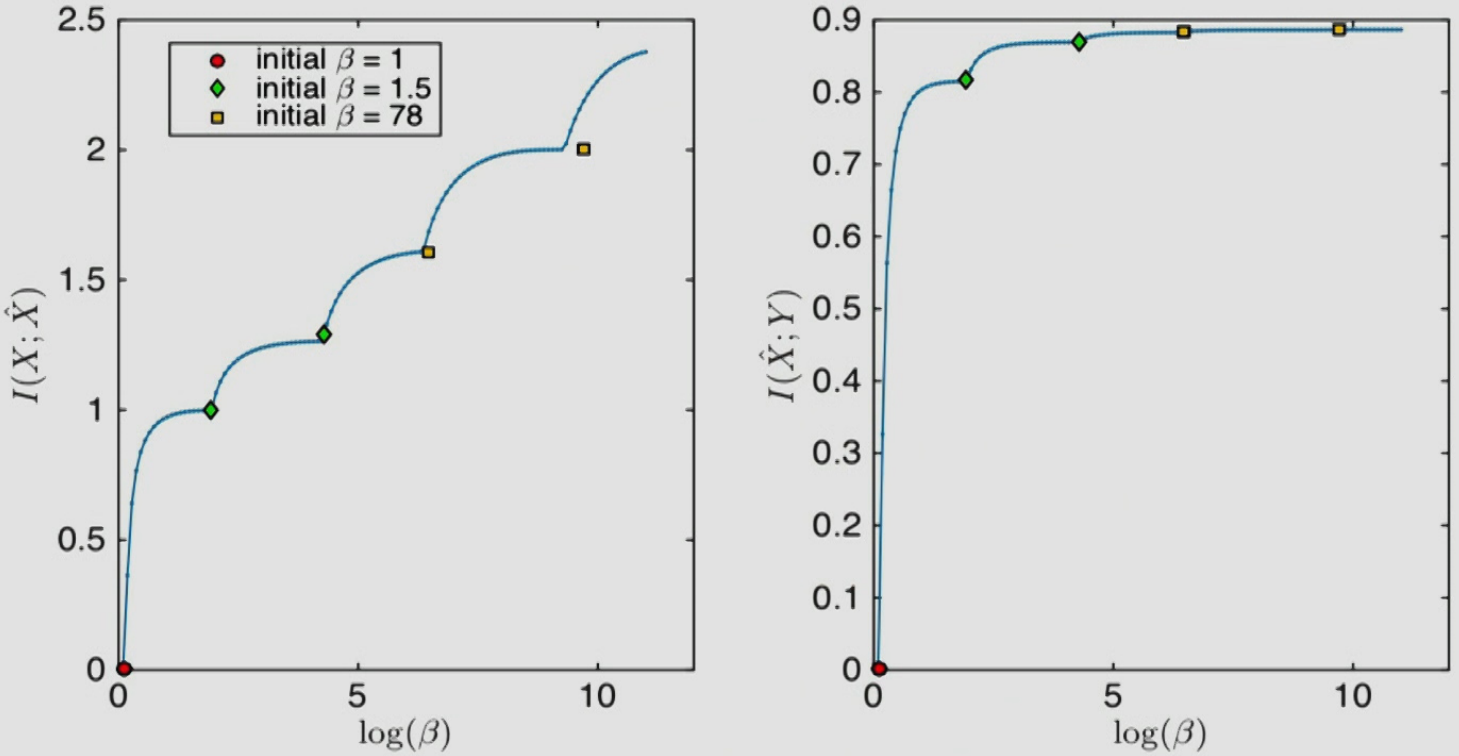
Topological phase transitions in the layers



Topological phase transitions in the layers



Second order phase transitions on the IB curve



Perimeter Institute, April 2018 - Tishby

ICRI-CI

The IB bifurcation (phase-transitions) points

The IB bifurcation points can be found as follows:

$$p_\beta(x\,|\,\hat{x}) = \frac{p(x)}{Z(x,\beta)}\exp(-\beta D[p(y\,|\,x)\,\|\,p_\beta(y\,|\,\hat{x}])$$

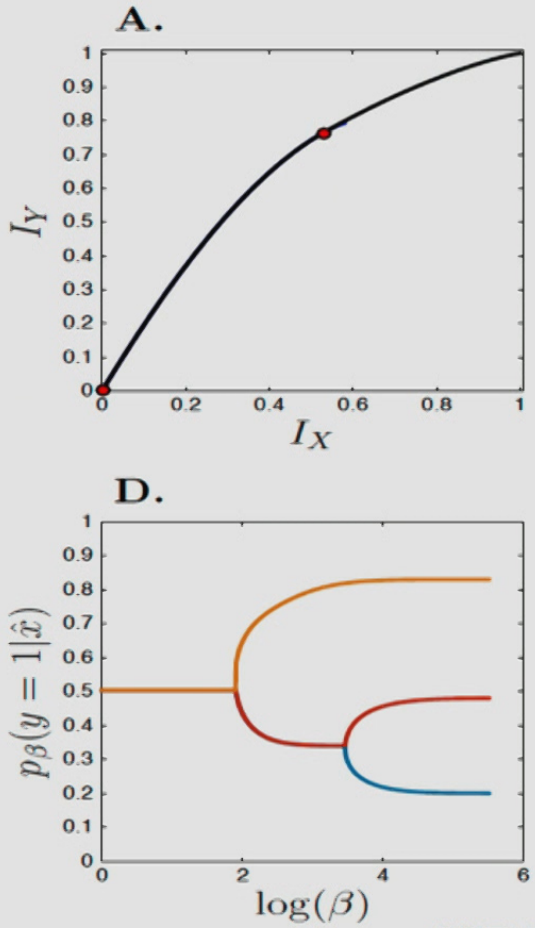
or $$\ln p_\beta(x\,|\,\hat{x}) = \ln\frac{p(x)}{Z(x,\beta)} - \beta D[p(y\,|\,x)\,\|\,p_\beta(y\,|\,\hat{x})]$$

then:

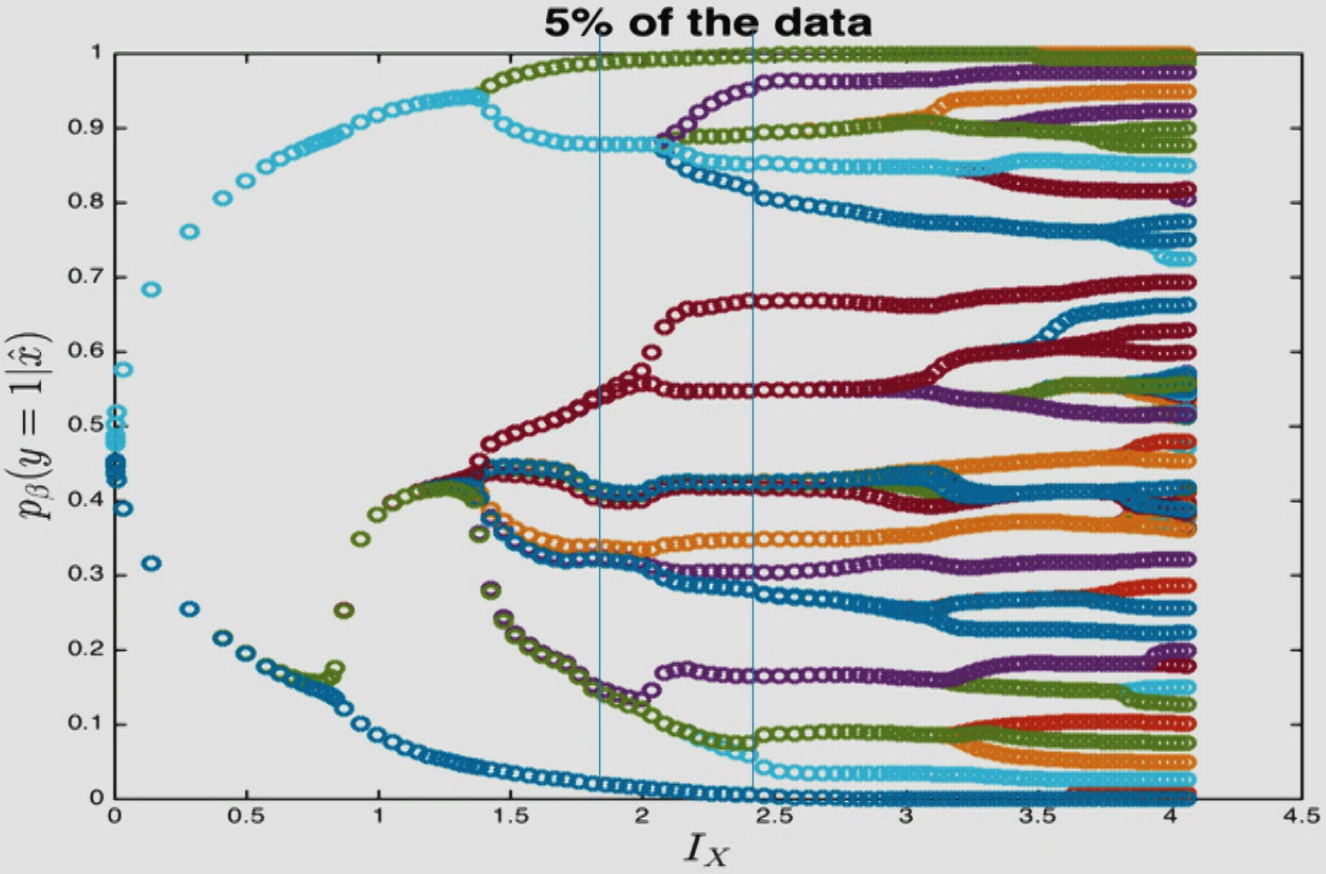
$$\frac{\partial \ln p_\beta(x\,|\,\hat{x})}{\partial \hat{x}} = \beta\sum\nolimits_y p(y\,|\,x)\frac{\partial \ln p_\beta(y\,|\,\hat{x})}{\partial \hat{x}}$$

similarily: $p_\beta(y\,|\,\hat{x}) = \sum_x p(y\,|\,x)p_\beta(x\,|\,\hat{x})$

$$\frac{\partial \ln p_\beta(y\,|\,\hat{x})}{\partial \hat{x}} = \frac{1}{p_\beta(y\,|\,\hat{x})}\sum\nolimits_x p(y\,|\,x)p_\beta(x\,|\,\hat{x})\frac{\partial \ln p_\beta(x\,|\,\hat{x})}{\partial \hat{x}}$$



Bifurcation diagrams in symmetric rule: layers diffusion slows down at phase transitions



$$W^k \approx \sum_{splits} \frac{\partial \log p\left(x|t_s^{k-1}\right)}{\partial t_s^{k-1}}$$

ICRI-CI Intel Collaborative Research Institute Computational Intelligence (intel)

Summary

- **The Information Plane provides a unique visualization of DL**
  - Most of the learning time goes to compression
  - Layers are learnt bottom up – and "help" each other
  - The layers converge to special (critical?) points on the IB bound
- **The advantage of the layers is mostly computational**
  - Relaxation times are super-linear (exponential?) in the Entropy gap
  - Hidden layers provide intermediate steps and boost convergence time
  - Hidden layers help in avoiding critical slowing down
- **Further directions**
  - Exactly solvable DNN models (through symmetry & group theory)
  - New/better learning algorithms & design principles
  - Predictions on the organization of biological layered networks …