Deep Learning, Representation Compression, And Symmetries

A tutorial for particle physicists

CERN, Geneva, April 2019

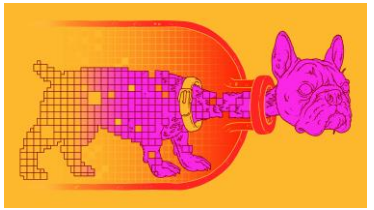
Naftali Tishby

School of Engineering and Computer Science

The Edmond & Lily Safra Center for Brain Sciences

Hebrew University, Jerusalem, Israel

Noga Zaslavsky
Ravid Schwartz-Ziv



We begin to obtain some new understanding…

We combine 3 different ingredients:

- **Rethinking Statistical Learning Theory**
  - Worse case PAC bounds → typical data-dependent model 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
- **The fluctuating 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
  - → Information Bottleneck critical points & interpretation of the layers

Part 1: Rethinking Learning Theory

- **The Deep Learning Revolution**
- **Large scale learning**
  - **The classical PAC bounds**
  - **Distribution dependent bounds**
  - **Thermodynamic limit**
- **Typical patterns, Learning & measure concentration**
  - **Entropy & mutual information**
  - **Information, compression and generalization**
- **The Information Plane Theorems**
  - **Why it's important for Any learning algorithm**
  - **The Information Bottleneck optimal bound**
  - **The role of stochastic learning rules….**

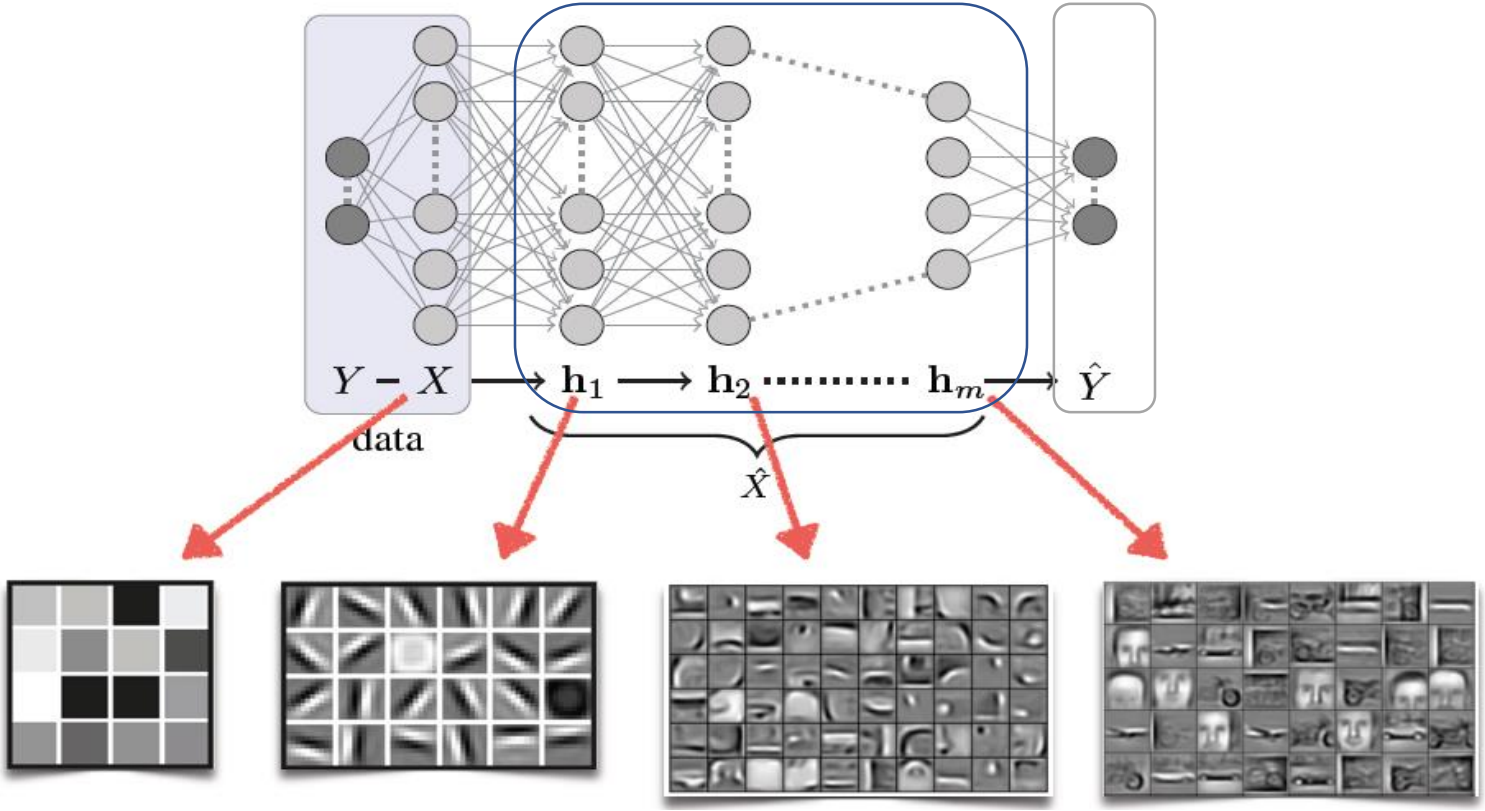
Part 2: the role of Stochastic Gradient Decent

- **The drift and diffusion phases of SGD – flat minima**
  - **The high and low gradient SNR phases**
  - **Scaling of the weights with updates**
  - **A Gaussian bound on the mutual information between layers**
- **The Computational benefit of the layers**
  - **How diffusion compress the layers representation**
  - **Parallel diffusion in independent sub-spaces**
  - **Parallel diffusion with dependencies**
  - **The role of symmetry**
  - **Power law scaling of computation time with number of layers**

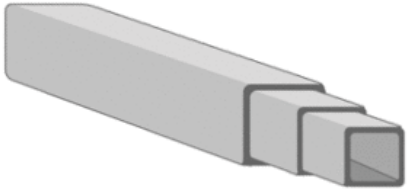
Part 3: Symmetries and IB bifurcations – what do the layers represent?

- **Symmetry and training convergence times**
- **Symmetries and phase IB phase transitions**
  - **Topological transitions in the layers encoders**
- **Phase transitions and diffusion critical-slowing-down**
  - **Water filling on irreducible representations**
  - **The IB bifurcation as codes of the relevant information**
  - **What can it tell us about the weights of DNNs?**

Information Flow in Deep Neural Nets



Analogies: telescopic (leaky) tubes or filter cascade



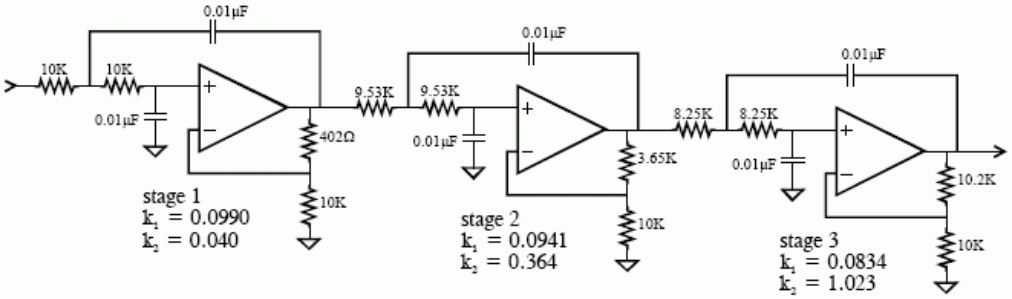


FIGURE 3-9
A six pole Bessel filter formed by cascading three Sallen-Key circuits. This is a low-pass filter with a cutoff frequency of 1 kHz.

Each layer filters out some of the input pattern entropy and eventually the last layer keeps only the label relevant information

Deep Learning and the Information Bottleneck

Main results:

- **Only 2 numbers matter: Mutual Information of each layer on input & on desired output –**
  - For large typical patterns: expressibility is NOT the issue, rather the dynamics of the optimization.
- **Optimality: The layers converge to the optimal [finite-sample] information-theoretic bound**
  - DL can achieve optimal (model free, rule dependent) sample complexity-accuracy tradeoff
  - Through the diffusion/fluctuations of the Stochastic Gradient Descent dynamics
  - The diffusion compresses the representations by "forgetting" irrelevant details – which dramatically improves generalization
- **Benefit of the Hidden Layers – Computational – boost power depend on the diffusion exponent!**
  - The benefit is mostly computational – boosting the compression with more layers
  - The location of the optimal layers is determined by the problem
- **Interpretability : what do the layers represent?**
  - Full layers can have clear – problem specific – interpretation, NOT single neurons (in general)!
- **Design principles**
  - DL is good for stochastic, compressible rules.
  - Layers final position is determined by critical points of the Information Bottleneck solutions.

Some Information Theory basics

- The KL-distribution divergence, relative/cross Entropy, …:

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:

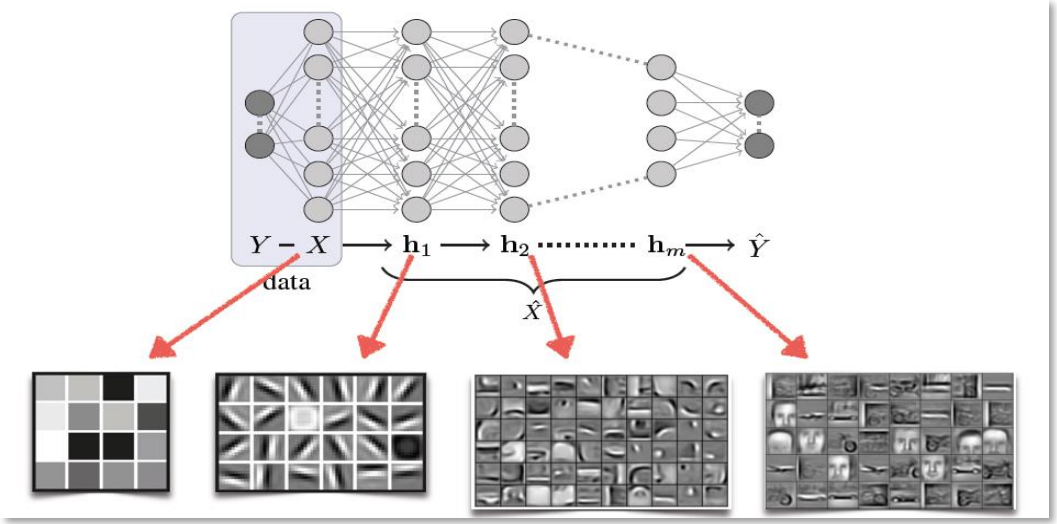
for any Markov chain: $X \to Y \to Z$:

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

Reparametrization Invariance, for invertible f, y :

$$I(X;Y) = I(f(X);y(Y))$$

What do the DNN Layers represent?



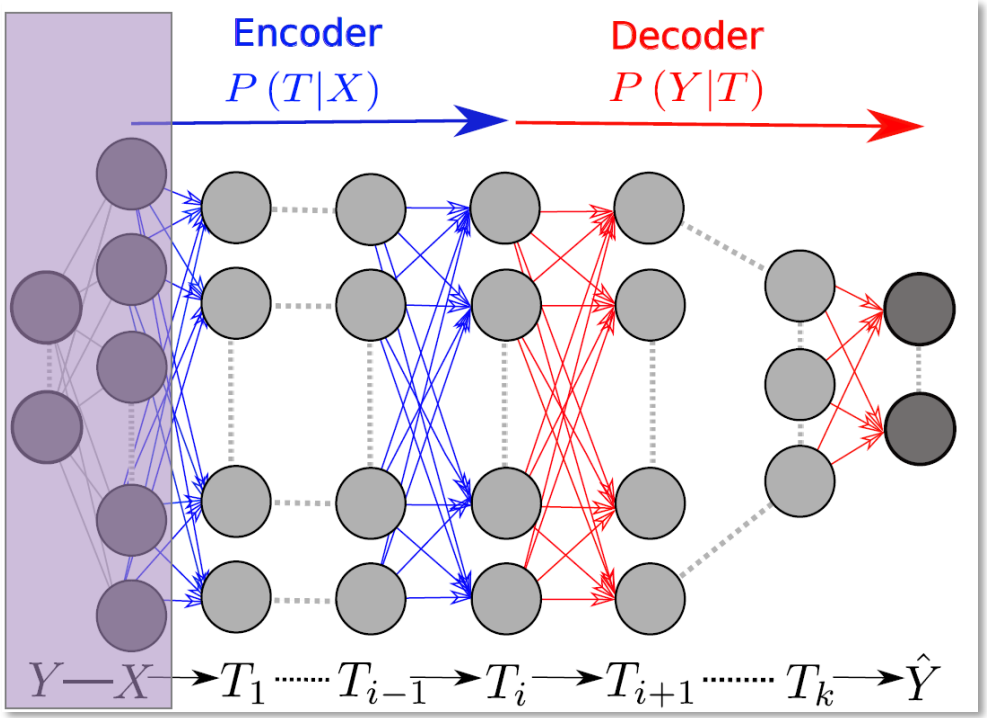
Data Processing Inequalities:

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

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

- A Markov chain of topologically distinct [soft] **partitions (VQ)** 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

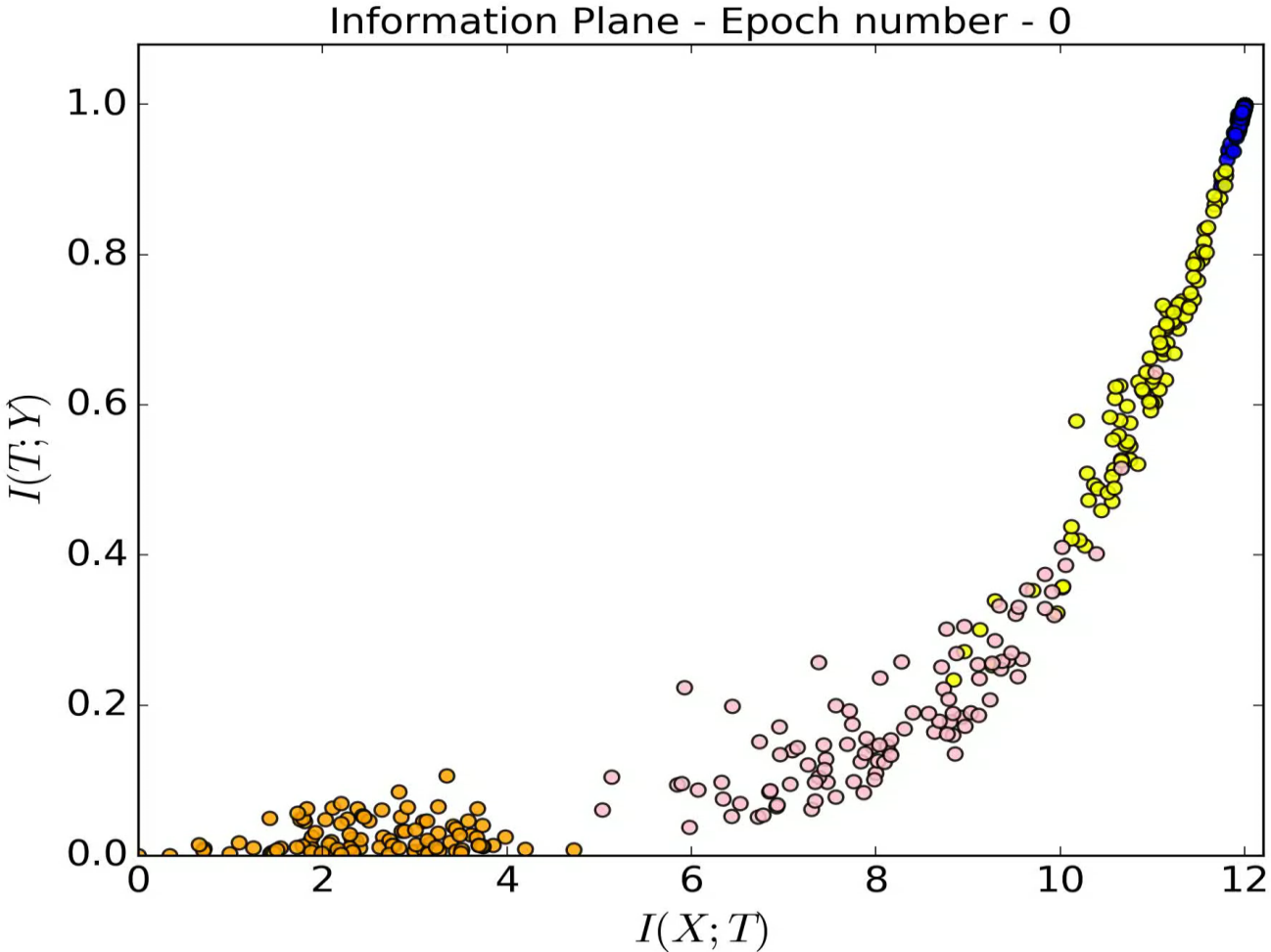


Theorem (Information Plane):

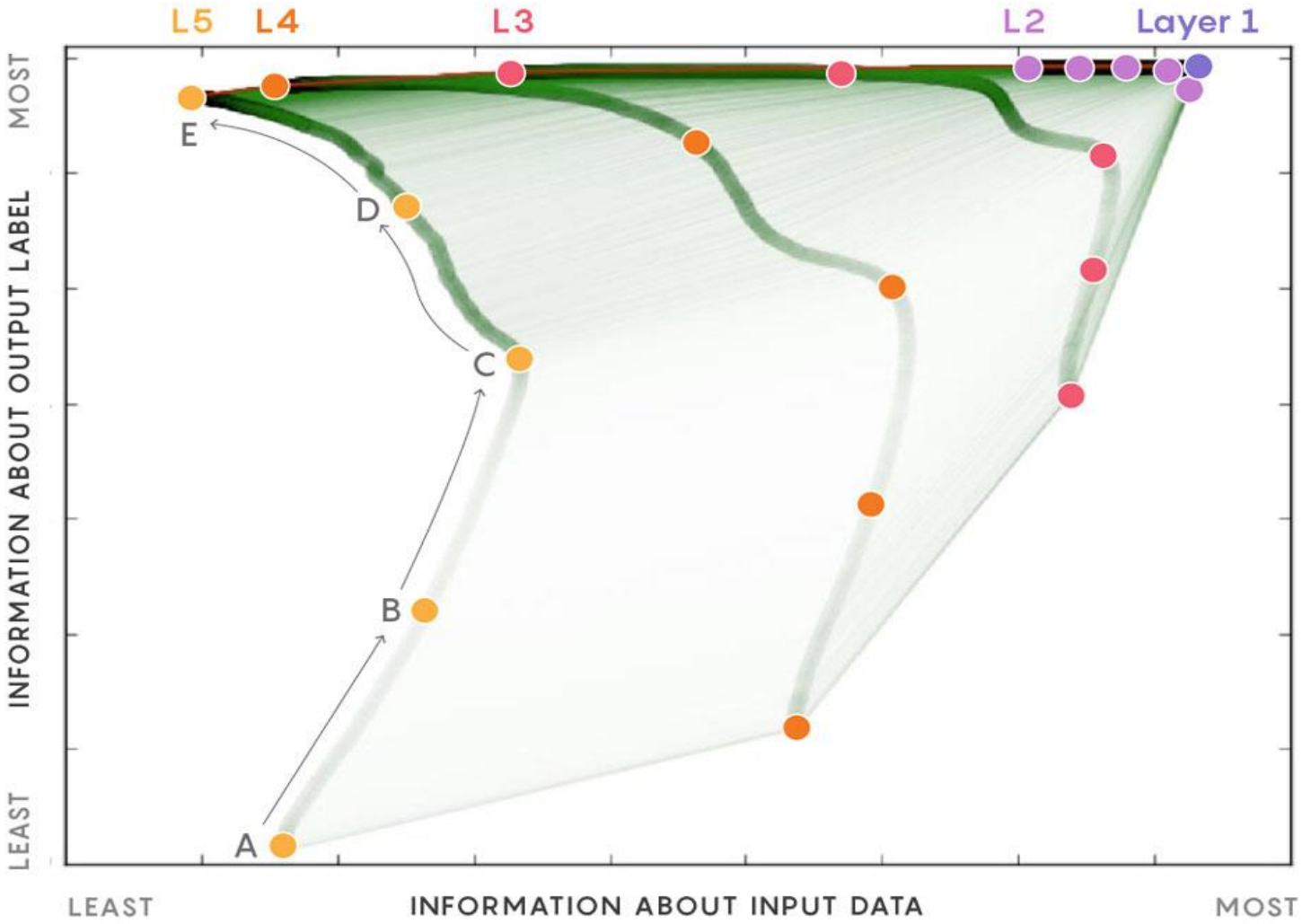
*For large typical **X**, the sample complexity of a DNN is completely determined by the encoder mutual information, **I(X;T)**, of the last hidden layer; the accuracy (generalization error) is determined by the decoder information, **I(T;Y)**, of the last hidden layer.*

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

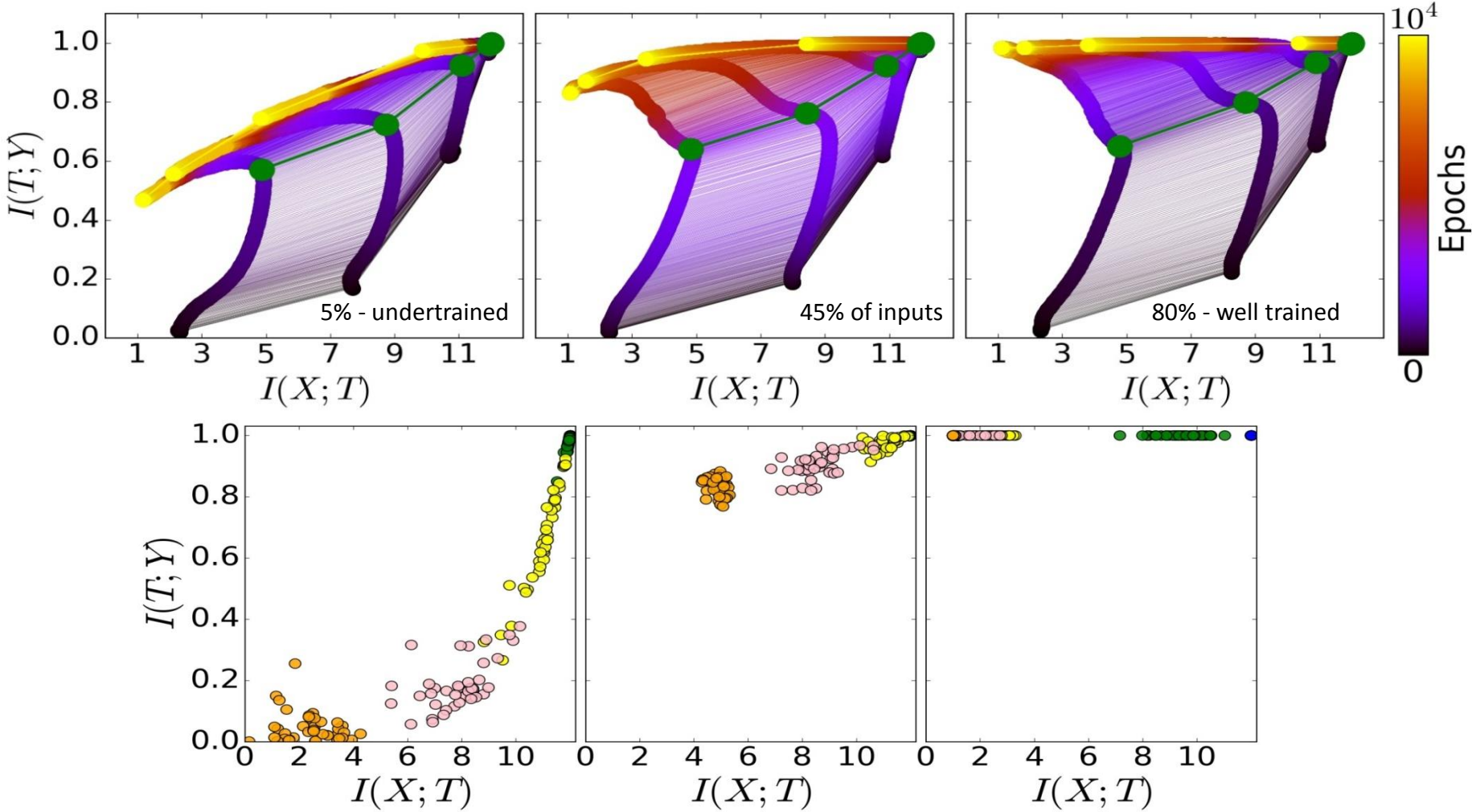
100 DNN Layers in Info-Plane without averaging



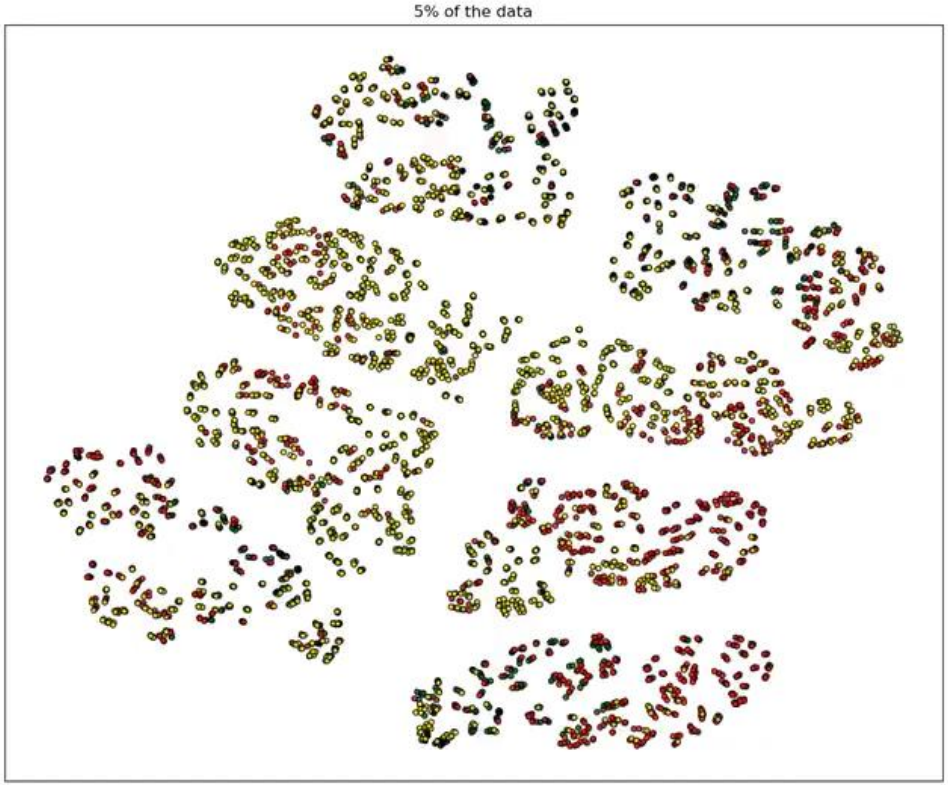
- Is this the general picture?
- Why do the MI values concentrate?
- Only 2 numbers per layer are relevant?
- What do they mean?
- What governs their dynamics?

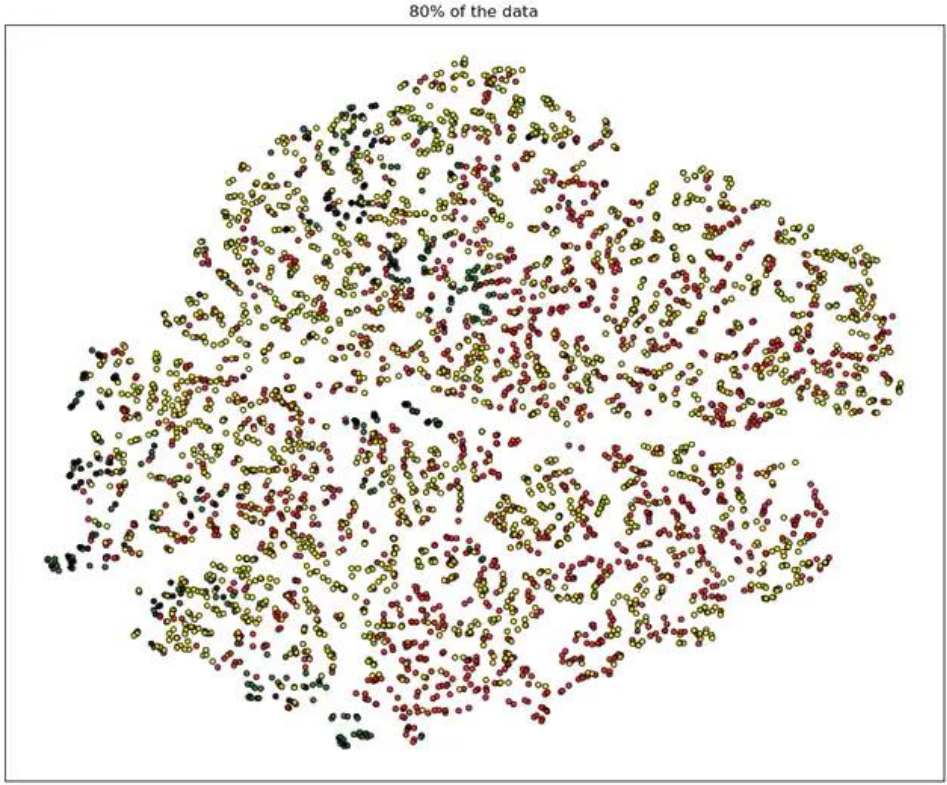

L5
L4
L3
L2
Layer 1
MOST
LEAST
INFORMATION ABOUT OUTPUT LABEL
E
D
C
B
A
LEAST
INFORMATION ABOUT INPUT DATA
MOST

Finite sample/data limits

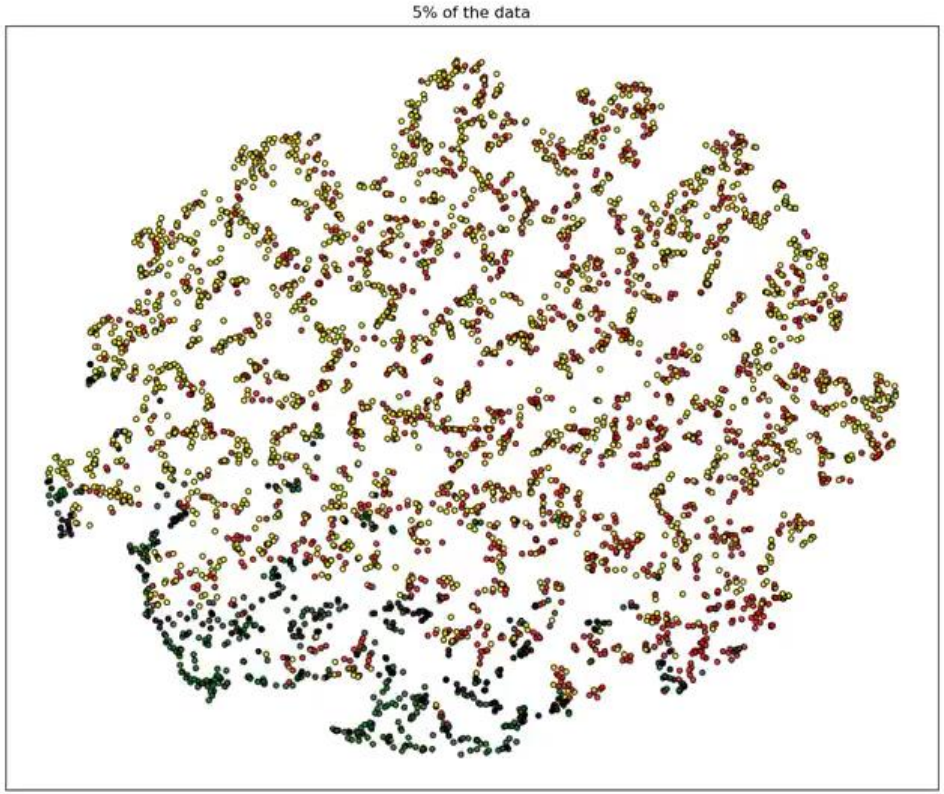


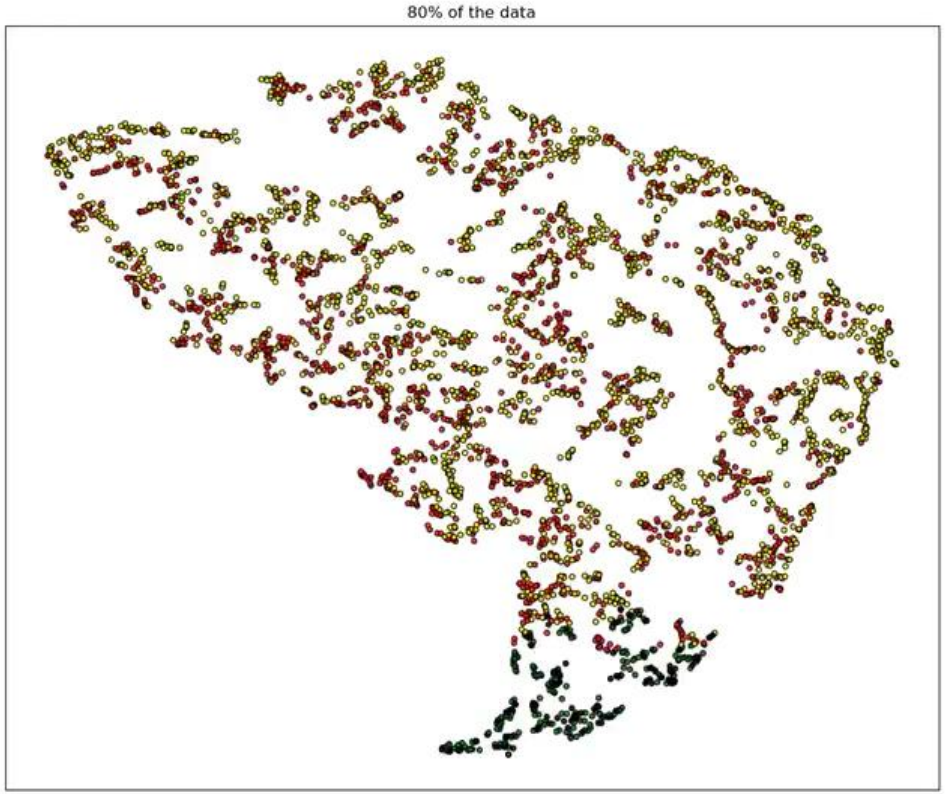
Iteration number 1, Layer number 0





Iteration number 1, Layer number 4





Information plane theorems:

Generalization error:

$$\varepsilon_g < I(X;Y) - I(T_\varepsilon;Y)$$

Effective class dimensionality:

$$d_\varepsilon = |T_\varepsilon| \simeq 2^{I(X;T_\varepsilon)}$$

Input compression generalization bound:

$$\varepsilon_{gap}^2 < \frac{2^{I(X;T)} + \log 1/\delta}{m}$$

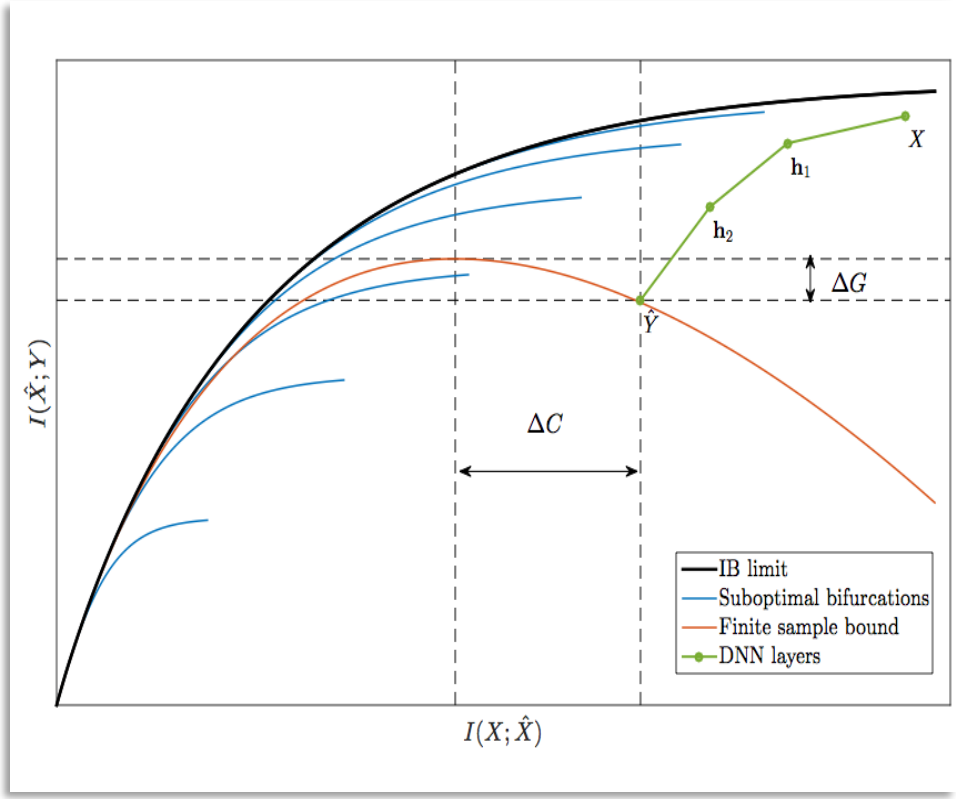
The Information Bottleneck optimality bound

(Tishby, Pereira, Bialek, 1999)

The IB bound optimality equations:

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

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



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 e- partition, T_e, is with the empirical distortion

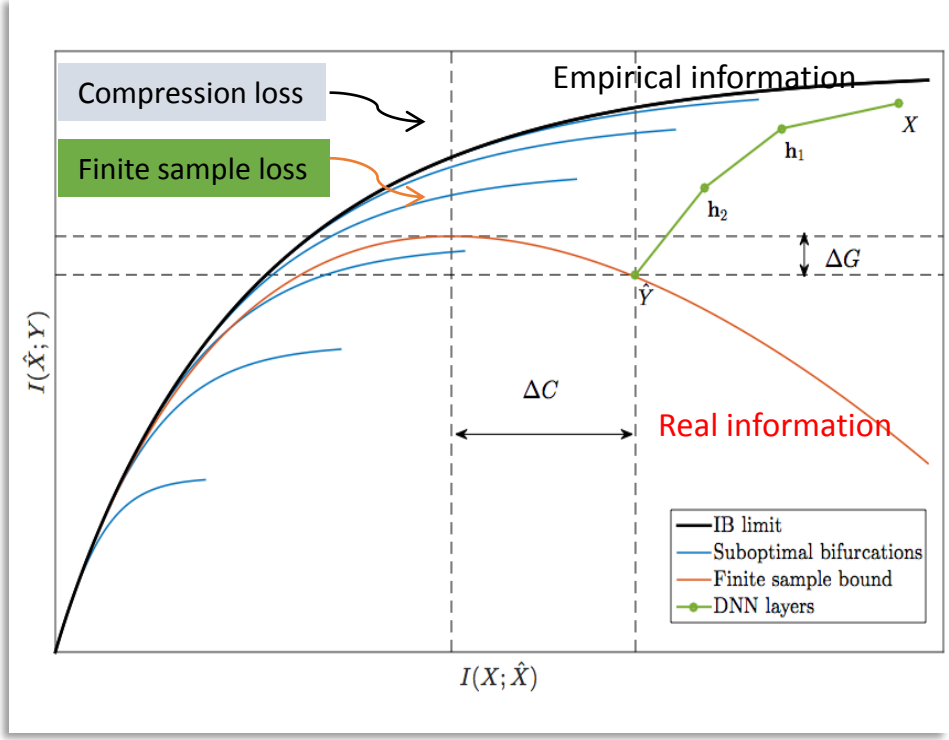
$$d_{IB}(x,t) = D[p_{emp}(y|x) \,\|\, p(y|t)]$$

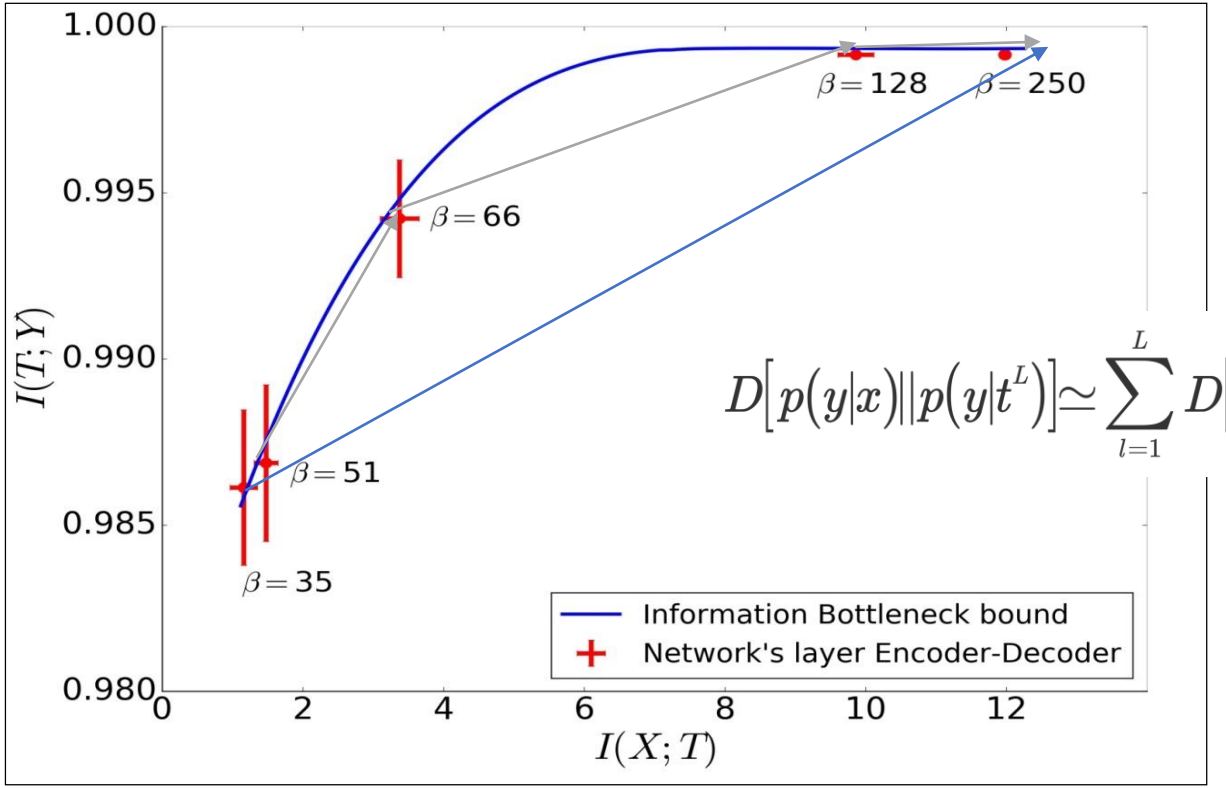
as $\langle d_{IB} \rangle_{emp} = I(X;Y) - \hat{I}_{emp}(T;Y)$

with a finite sample there is another information loss:

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

both should remain small for good generlaization!





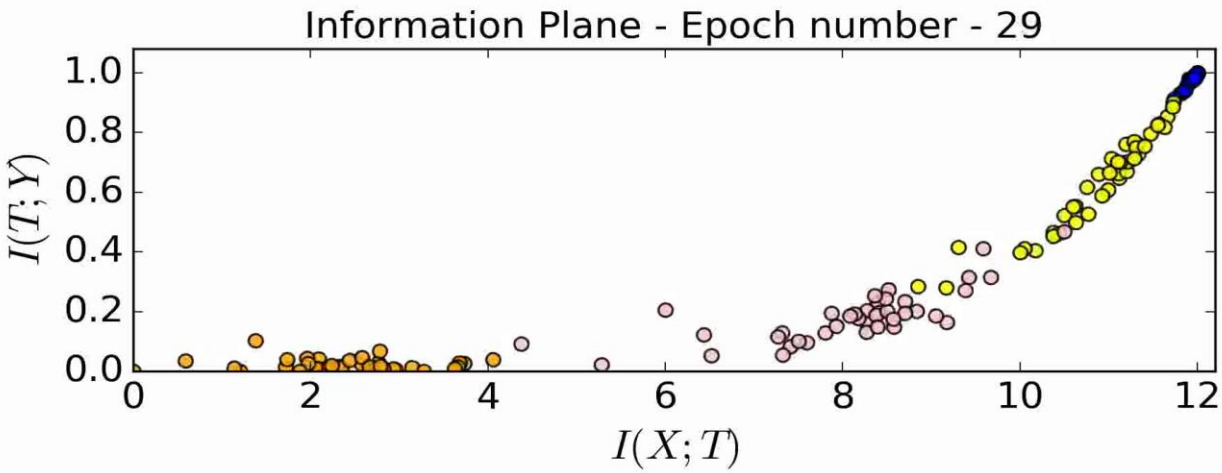
$$D\left[p(y|x)\|p(y|t^L)\right] \simeq \sum_{l=1}^{L} D\left[p(y|t^{l-1})\|p(y|t^l)\right]$$

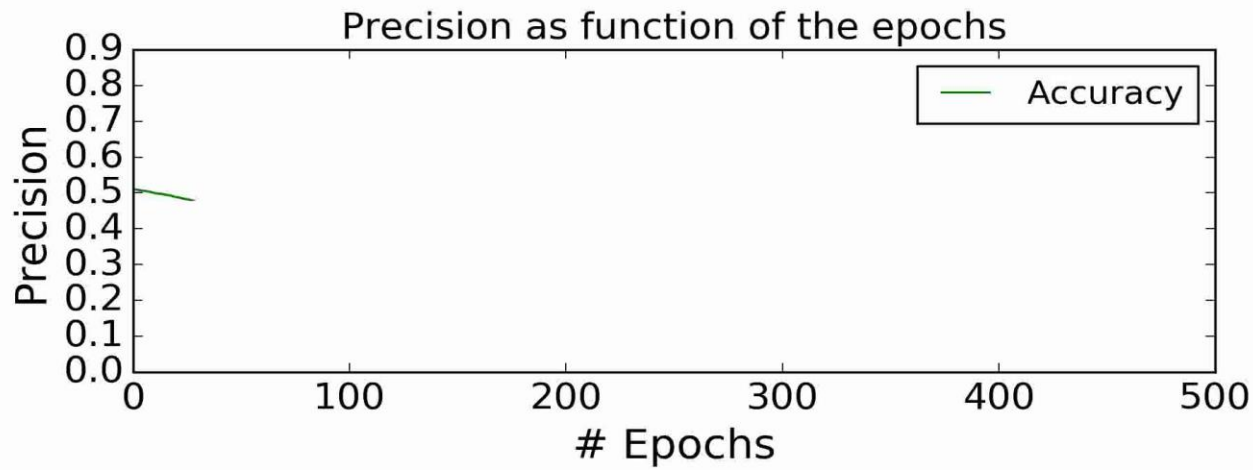
- Layers of optimal DNN converge to [a successively refineable approximation of] the optimal finite-sample IB limit information-curve
- Layers must be in “different topological phases” of the IB solutions
- The DNN encoder & decoder for each layer satisfy the IB self-consistent equations

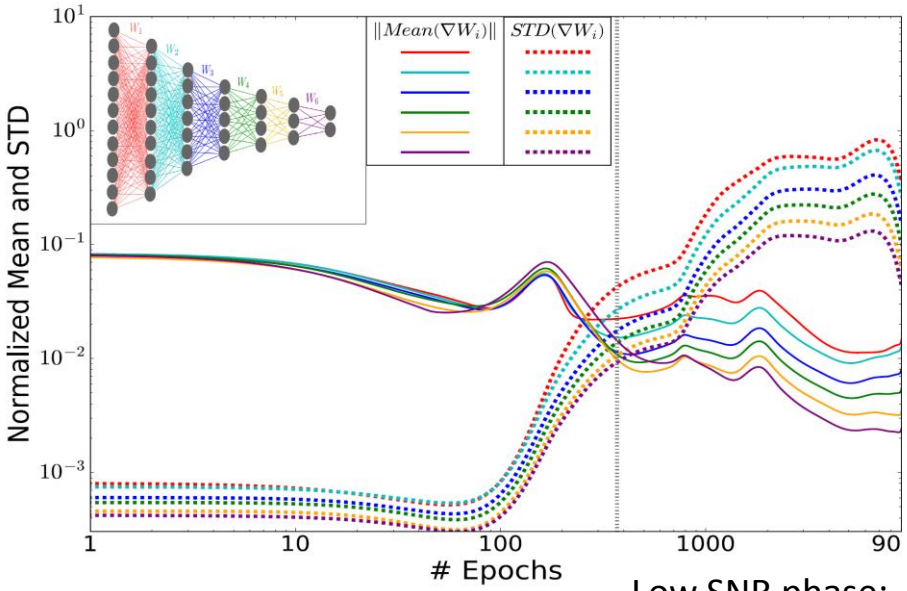
Part 2: the role of Stochastic Gradient Decent

- The drift and diffusion phases of SGD – flat minima
  - The high and low gradient SNR phases
  - Scaling of the weights with updates
  - A Gaussian bound on the mutual information between layers
- The Computational benefit of the layers
  - How diffusion compress the layers representation
  - Parallel diffusion in independent sub-spaces
  - Parallel diffusion with dependencies
  - The role of symmetry
  - Power law scaling of computation time with layers

Layers paths with training/generalization error

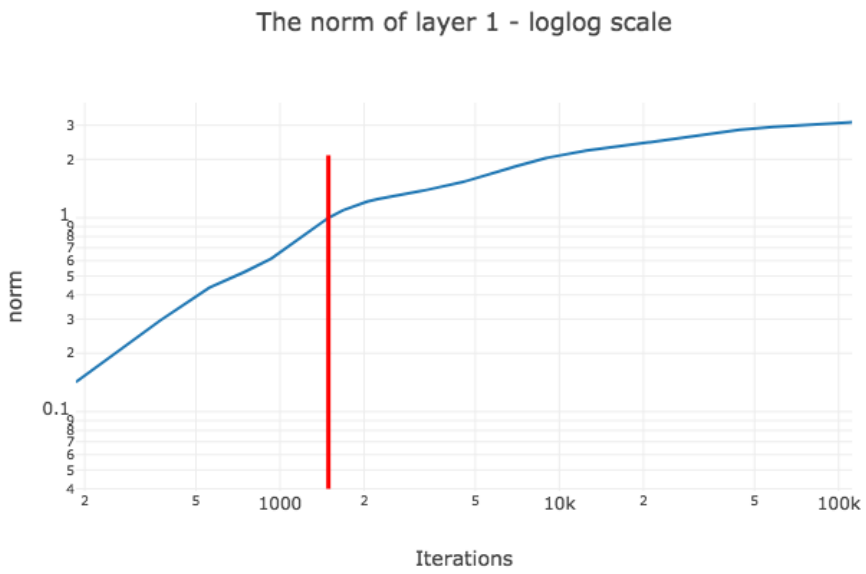






High SNR phase: memorization

Low SNR phase: forgetting

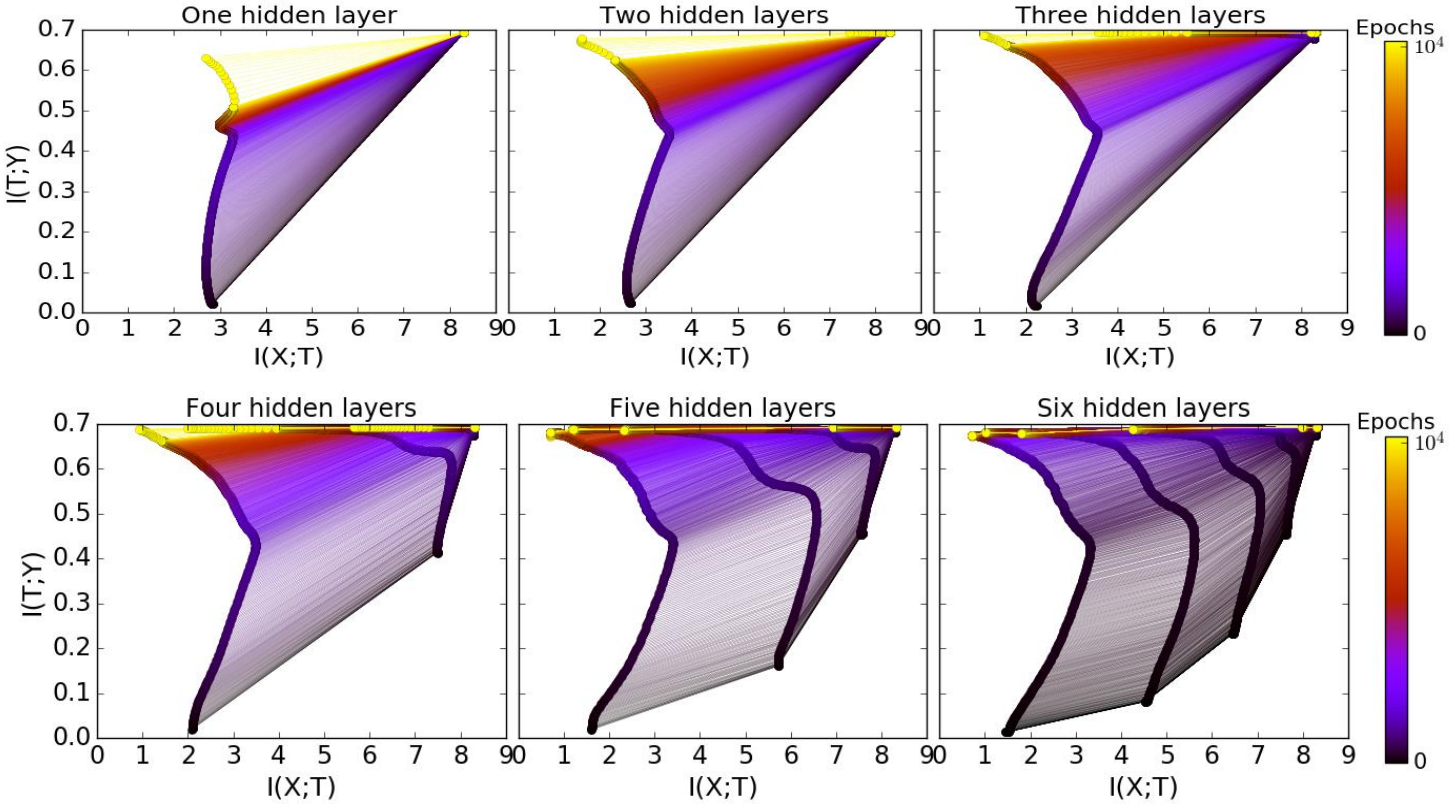


High SNR phase: Linear drift

Low SNR phase: Diffusion

In the noisy phase the weights diffuse and grow like $O(\sqrt{t})$

The benefit of the hidden layers is computational!



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.

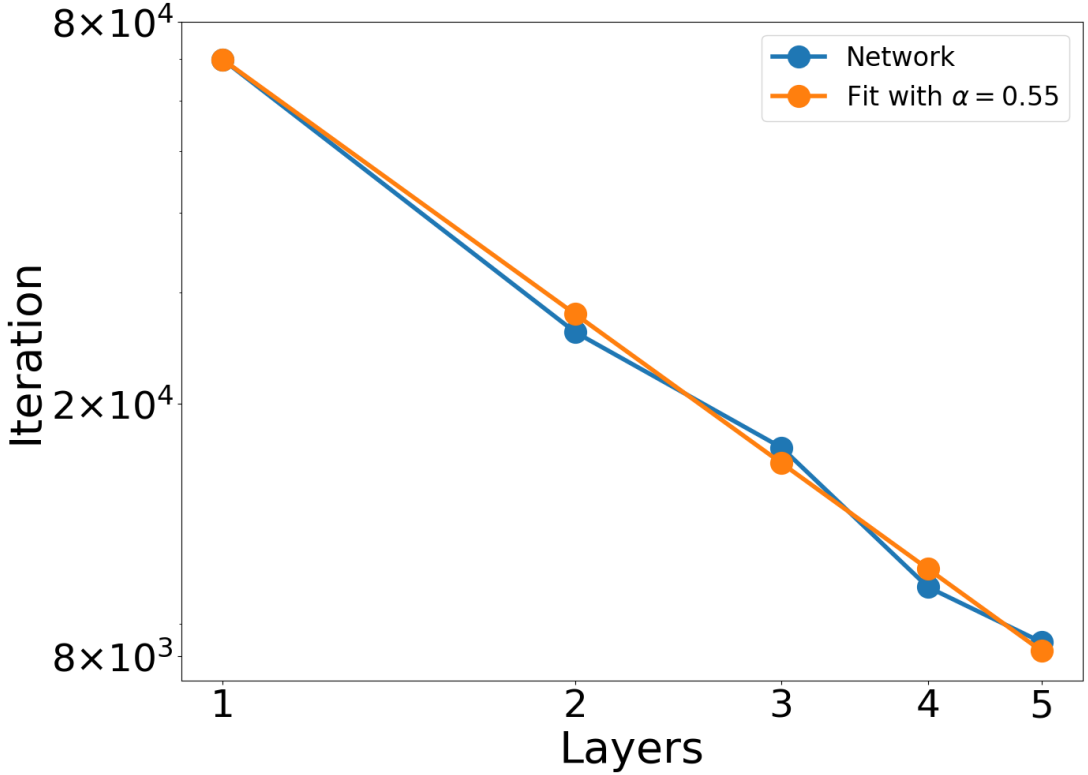
The Compression Theorem

Theorem: During the diffusion phase, the Mutual Information between consecutive layers decrease with the number of SGD iterations, t, like:

$$I(T_k;T_{k+1}) \leq C_k + t^{-\alpha} R$$

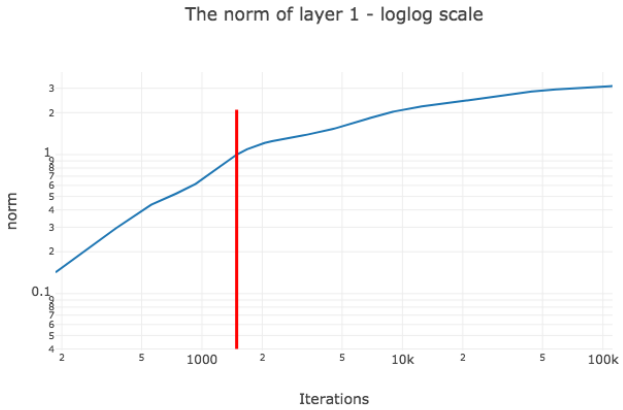
where c_k depends only on the relevant dimension at the *k* layer and α is the diffusion exponent. R - the "power" of the target, depends only on the problem - not on the architecture.

The Computational benefit of the layers

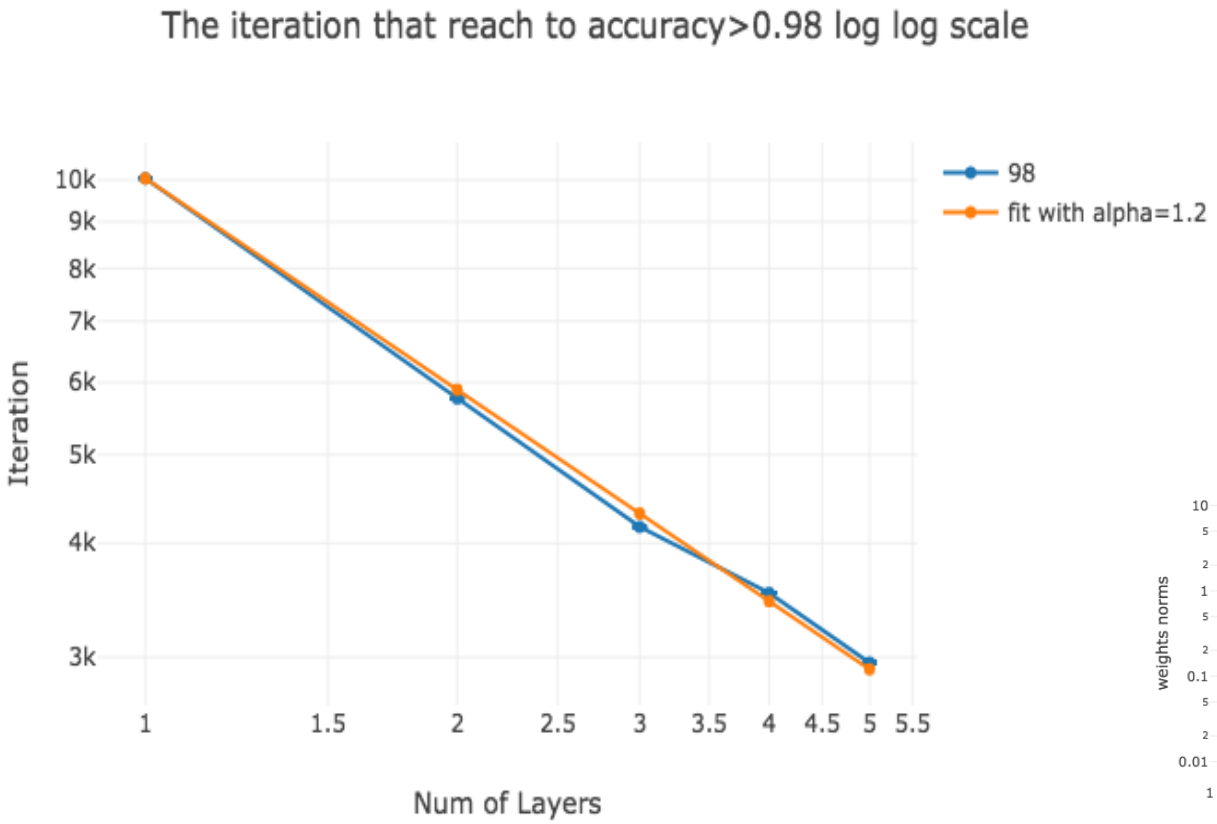


Time to reach 0.98 bits on Y, as function of the number of hidden layers.

The slope is very close to 1/(diffusion exponent) as our theory predicts.

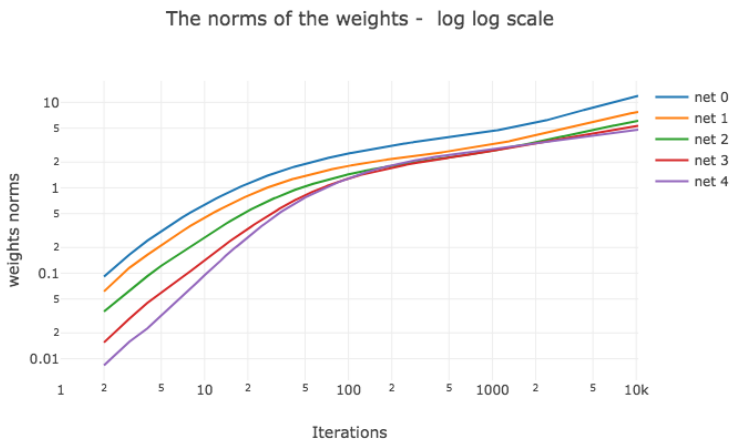


The Computational benefit of the layers



Time to reach 0.98 bits on Y, as function of the number of hidden layers, for MNIST, CNN, RelU's.

The slope is lower than 1/(diffusion exponent) as our bound predicts. Why?



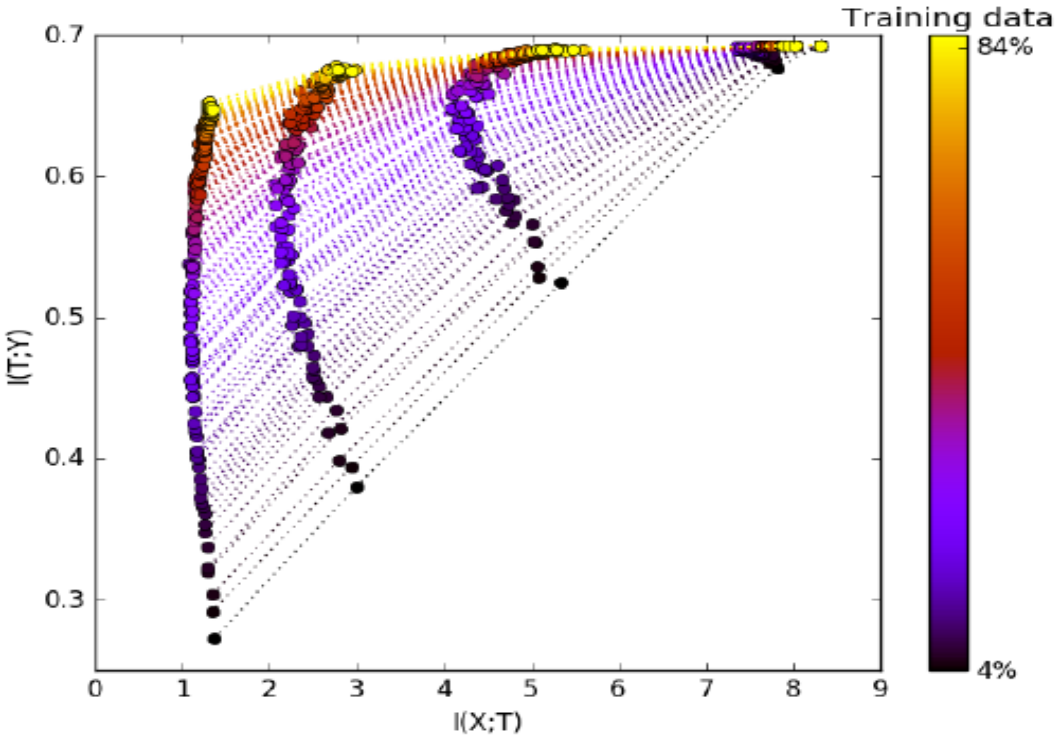
Part 3: Symmetries and bifurcations – what do the layers represent?

- **Critical points and training convergence times**
- **Symmetries and the IB bifurcation points**
  - **Topological transitions in the layers encoders**
- **Where do the layers converge to?**
  - **Water filling on irreducible representations**
  - **The IB bifurcation as codes of the relevant information**
  - **What can it tell us about the weights of DNNs?**

Where do the layers converge to?

- The layers are dominated by the IB encoder-decoder
- The Info-curve is dominated by its critical points
- Any local gradient optimization stalls near criticality
- Symmetries gather the IB critical points

- ➔ Layers are likely to be found near critical points



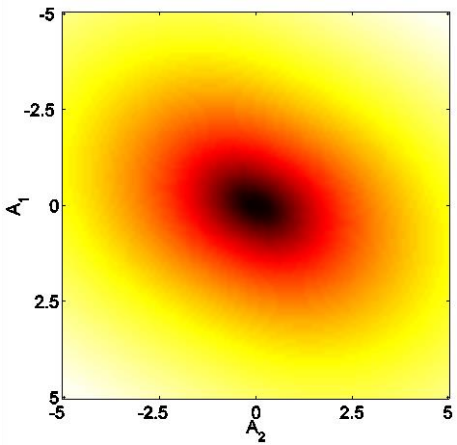
- Fitting larger training data require more information in the hidden layers.

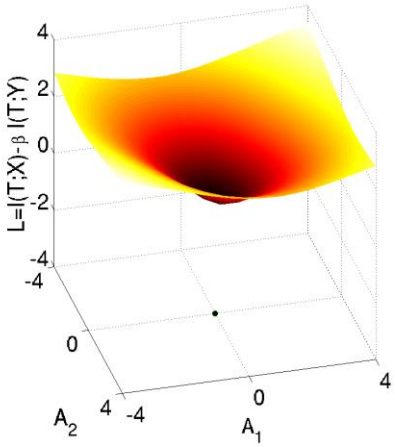
- It is the **mutual-information of the last hidden layer**, which determines generalization (unlike standard hypothesis class bounds)

The nature of the solution

$$A = \begin{cases} \frac{\beta(1\text{-}\lambda_1)\text{-}1}{\lambda_1} v_1 & \beta > (1\text{-}\lambda_1)^{-1} \\ 0 & \text{otherwise} \end{cases}$$

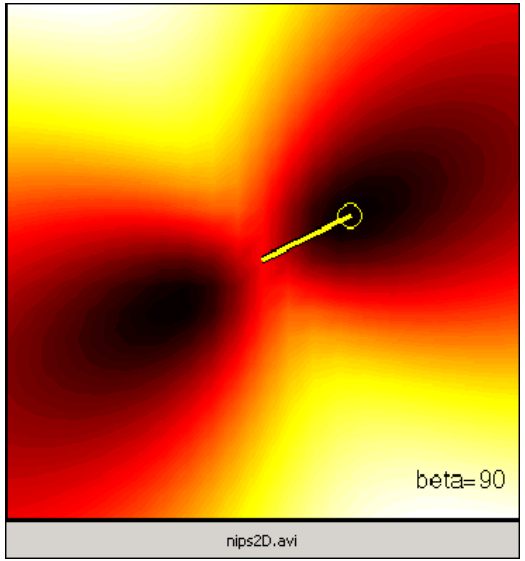
- As an illustration, look at the surface of *L* as a function of A for A = (1x2) vector. For example

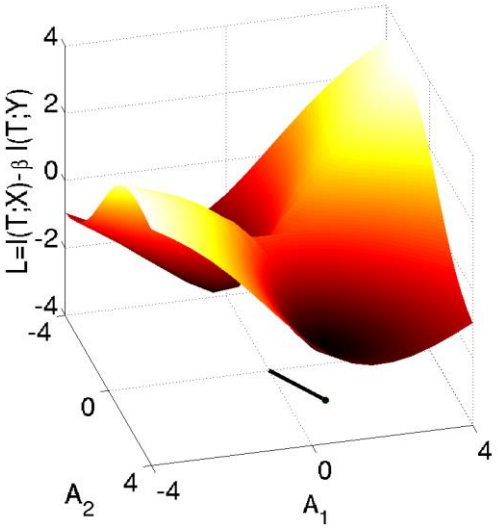




The nature of the solution

- As β increases the eigenvector solution emerges:





The IB critical (bifurcation) points

Defining the matrices:

$$C_{xx'}(\hat{x},b)=\sum_y \frac{p(y|x)}{p_b(y|\hat{x})}p_b(x'|\hat{x})p(y|x')$$

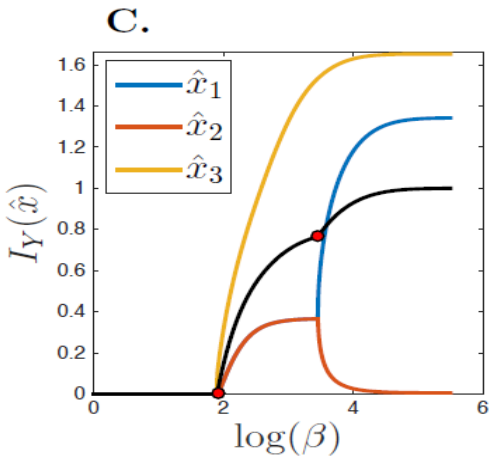
$$C_{yy'}(\hat{x},b)=\sum_x \frac{p(y|x)}{p_b(y|\hat{x})}p_b(x|\hat{x})p(y'|x)$$

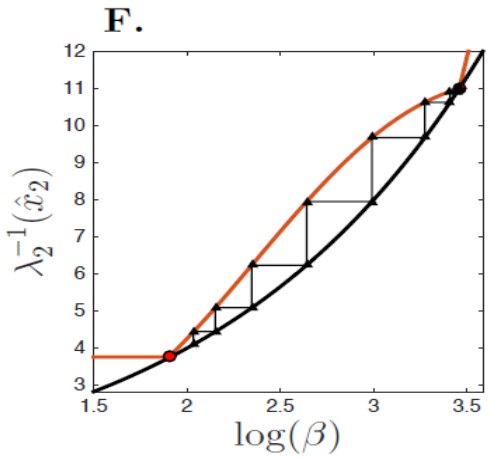
these equations can be combined into two (non-linear) eigenvalue problems:

$$\left[I - bC_{xx'}(\hat{x},b)\right]\frac{\partial \ln p_b(x'|\hat{x})}{\partial \hat{x}}=0$$

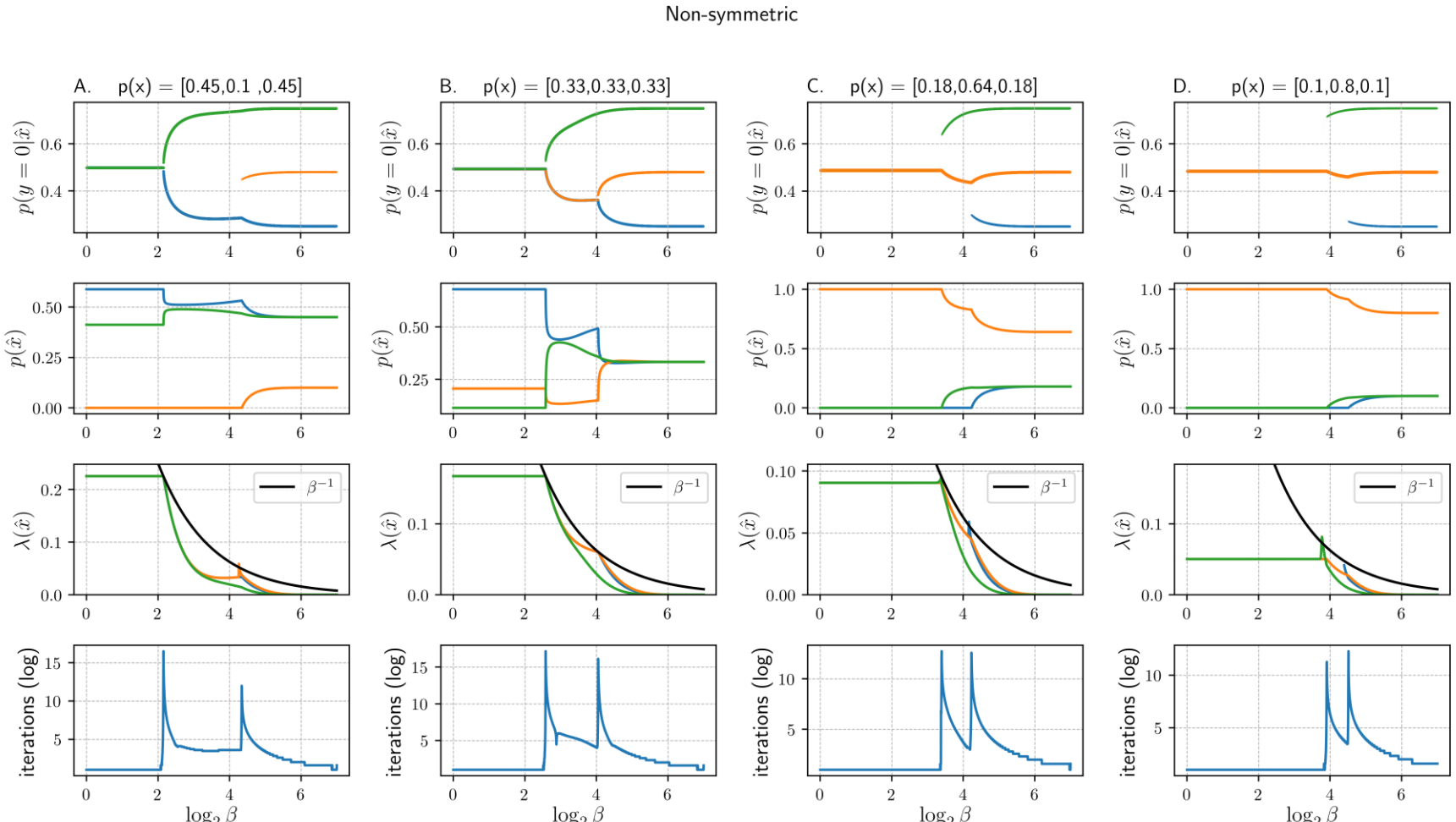
$$\left[I - bC_{yy'}(\hat{x},b)\right]\frac{\partial \ln p_b(y'|\hat{x})}{\partial \hat{x}}=0$$

These eigenvalue problems have non-trivial solutions (eigenvectors) only at the critical bifurcation points (second order phase transitions).

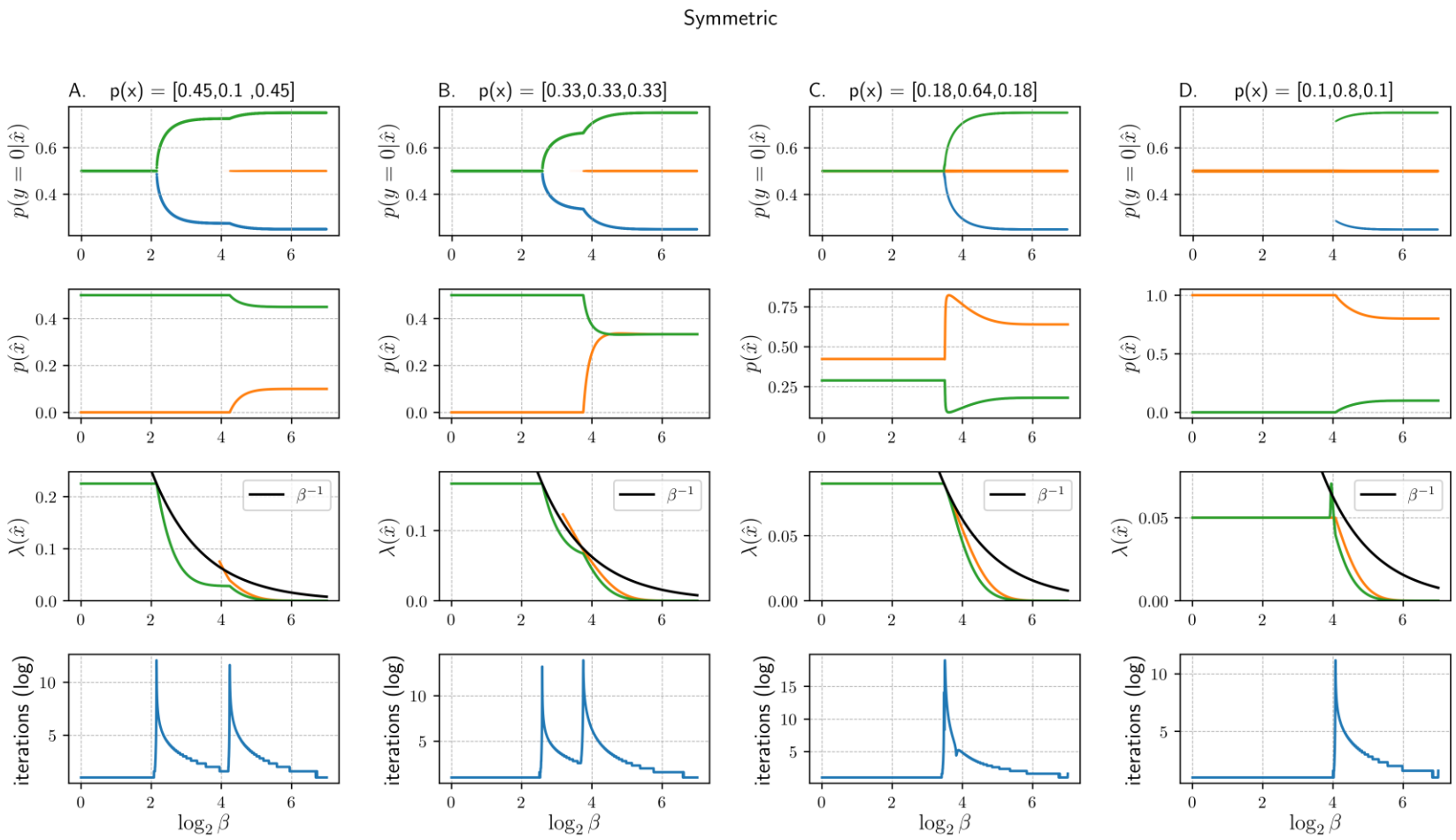




Convergence times near critical points



Convergence times near critical points



Critical slowing down and convergence times

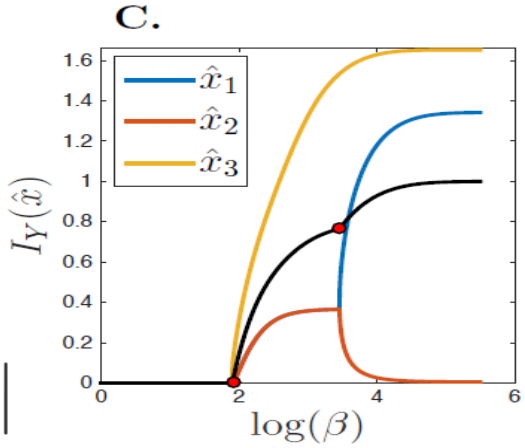
$$x_{t+1} = G(x_t) \qquad x^* = G(x^*)$$

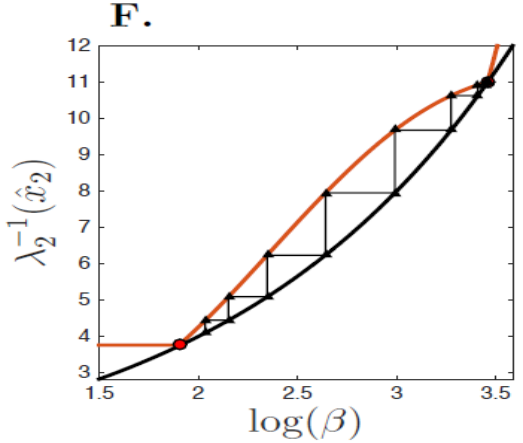
$$(x_{t+1} - x_t) \simeq \nabla G(x^*)(x_t - x_{t-1})$$

$$|x_{t+1} - x_t| \leq \gamma_{\max} |x_t - x_{t-1}| \qquad \gamma_{\max} = \max |\nabla G(x^*)|$$

$$\delta^2 L_{IB} \propto 1 - \nabla G(x^*) \qquad \gamma_{\max} \to 1$$

$$\tau \sim \frac{1}{1 - \gamma_{\max}}$$





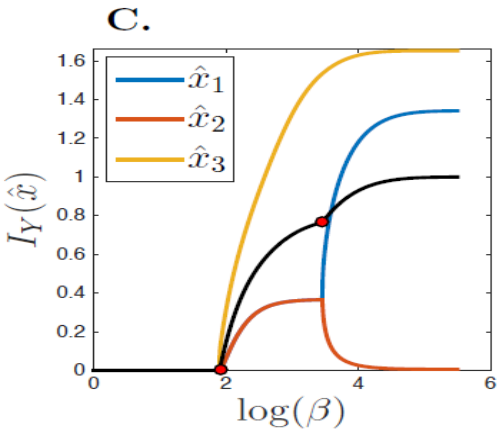
Critical slowing down and convergence times

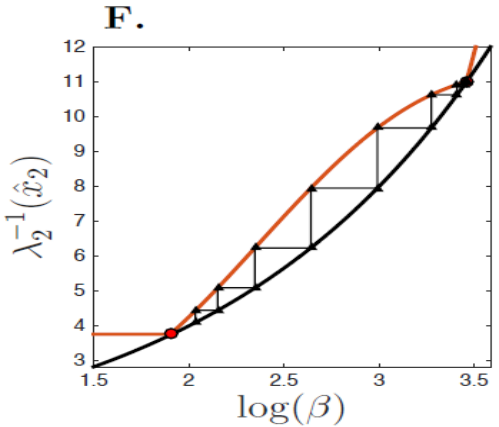
$$(x_{t+1}-x_t)\simeq\nabla G(x^*)(x_t-x_{t-1})$$

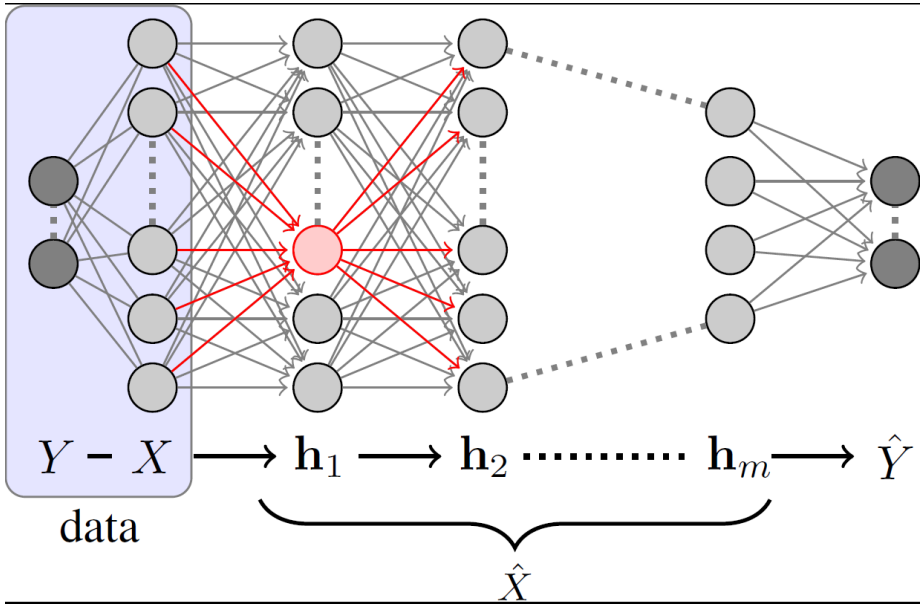
$$\ln p(x|\hat{x}_{t+1})-\ln p(x|\hat{x}_t)\simeq\frac{\partial \ln p(x|\hat{x}_t)}{\partial \hat{x}_t}\delta\hat{x}_t$$

$$\frac{\partial \ln p(x|\hat{x}_t)}{\partial \hat{x}_t}\delta\hat{x}_t=\beta\sum_{x\cdot}C_{xx}(\beta,\hat{x}_t)\frac{\partial \ln p(x|\hat{x}_{t-1})}{\partial \hat{x}_{t-1}}\delta\hat{x}_{t-1}$$

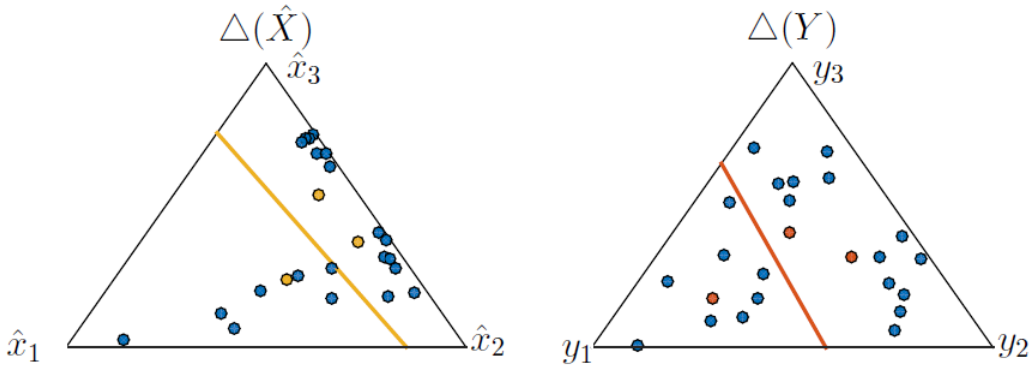
$$\eta=\frac{|\delta\hat{x}_{t+1}|}{|\delta\hat{x}_t|} \qquad \tau\sim\frac{1}{1-\eta\beta\lambda(C(\beta,\hat{x}))}$$







Each neuron corresponds to a hyperplane that separates two cluster-siblings at their split transition



- Conditioned on x
- Conditioned on $\hat{x}$
- Conditioned on y
- Decision between $\hat{x}_1$ and $\hat{x}_3$
- Decision between y_1 and y_3

Symmetry Groups & Representations

- Representations are homeomorphisms of the group to [unitary] Matrices.

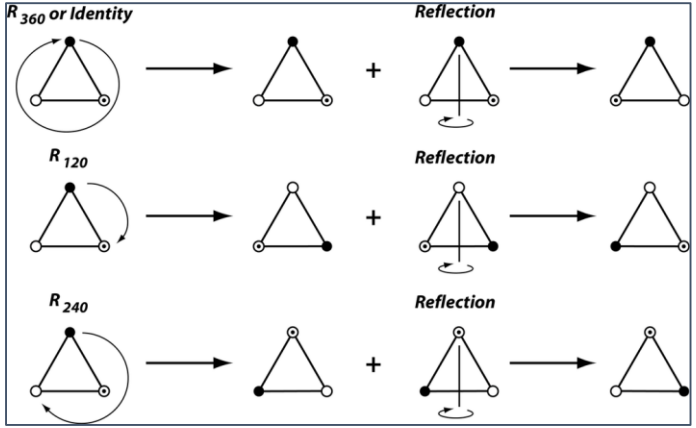
$$\mathrm{R}: G \to M(F), \quad \text{such that } \forall g_1, g_2 \in G, \quad R(g_1 \circ g_2) = R(g_1) \cdot \mathrm{R}(\mathrm{g}_2)$$

- Irreducible Representations – can't be block diagonalized
- Non-Abelian (non-commutative) groups have high dimensional irreducible representations.

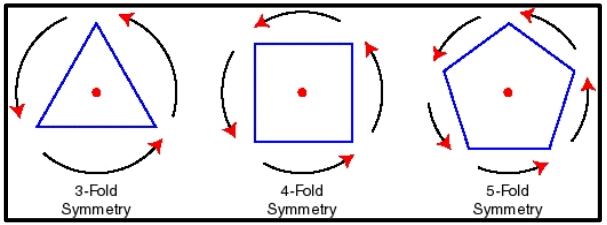
- d-dimensional Irreducible Unitary Representations induce d-dimensional sunspace of orthogonal functions that is invariant under the actions of the group.

- Such functions define a generalized Fourier series (with generalized FFT) over the group.

- A set of transformations, $g \in G$, that transform an object to itself. Forms a closed group, with identity & inverses.



Symmetries of the equilateral triangle



Different subgroups of the 2D Rotation
Group, O(2).

Irreducible matrix representations

$$A' = x^{-1}Ax = x^{-1}\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n1} & \cdots & a_{nn} \end{pmatrix} x = \begin{pmatrix} \begin{matrix} a_{11}' & a_{12}' & \cdots & a_{1m}' \\ a_{21}' & a_{22}' & \cdots & a_{2m}' \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1}' & a_{m1}' & \cdots & a_{mm}' \end{matrix} & 0 & \cdots & 0 \\ 0 & \begin{matrix} a_{11}'' & a_{12}'' & \cdots & a_{1l}'' \\ a_{21}'' & a_{22}'' & \cdots & a_{2l}'' \\ \vdots & \vdots & \vdots & \vdots \\ a_{l1}'' & a_{l1}'' & \cdots & a_{ll}'' \end{matrix} & & \vdots \\ \vdots & & & 0 \\ 0 & 0 & \cdots & \begin{matrix} a_{11}''' & a_{12}''' & \cdots & a_{1k}''' \\ a_{21}'' & a_{22}'' & \cdots & a_{2k}''' \\ \vdots & \vdots & \vdots & \vdots \\ a_{k1}''' & a_{k1}''' & \cdots & a_{kk}''' \end{matrix} \end{pmatrix} = \begin{pmatrix} A' & 0 & 0 \\ 0 & A'' & 0 \\ 0 & 0 & A'' \end{pmatrix}$$

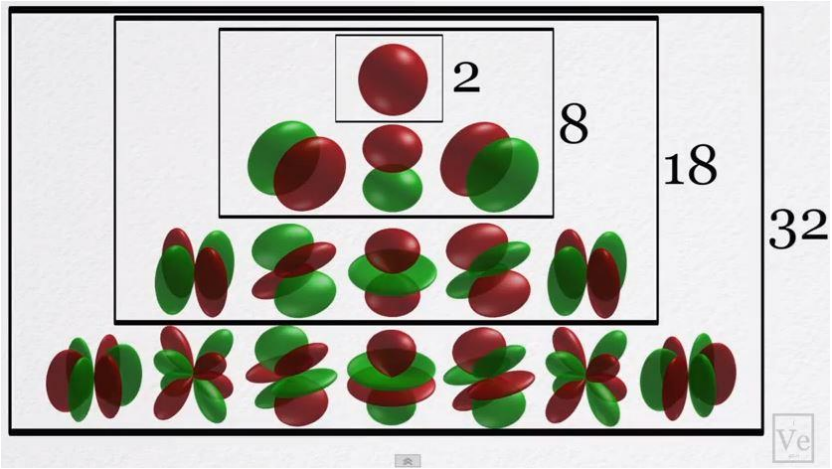
Example: O(3) & Spherical Harmonics

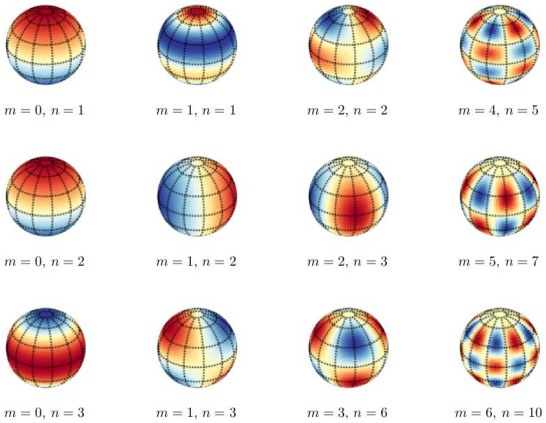
- The group of rotations in 3D, *O(3)*, is non-abelian and has Irreducible Representations in all odd dimensions:

$$d = 2l + 1, \quad l = 0, 1, 2, \ldots$$

- The orthogonal invariant functions are the Spherical Harmonics,

$$Y_{l,m}(\theta, \phi), -l \le m \le l$$





Any pattern on the sphere has a unique expansion in the $Y_{l,m}(\theta, \phi)$

Information Bottleneck for Symmetric problems
("ExpIB")

Def: *The learning problem (X,Y) is G symmetric if:*

$$P(Y \mid g.X) = P(Y \mid X) \quad \forall g \in G$$

Thm 1: *For G-Symmetric (X,Y) the* ***sufficient statistics*** *are projections of X on Irreducible Representations of G,* $T_l(X)$.

Thm 2: *For G-Symmetric (X,Y) all IB optimal representations are G-symmetric:*

$$P(Y \mid g.\hat{X}) = P(Y \mid \hat{X}) \quad \forall g \in G$$

Thm 3: *For any G-Symmetric* $(Y, \hat{X}(X))$

$$\log P(Y \mid \hat{X}) = \sum_l \eta_l(Y) T_l(X)$$

All topologically distinct IB phases are different projections on irreducible representations of G.

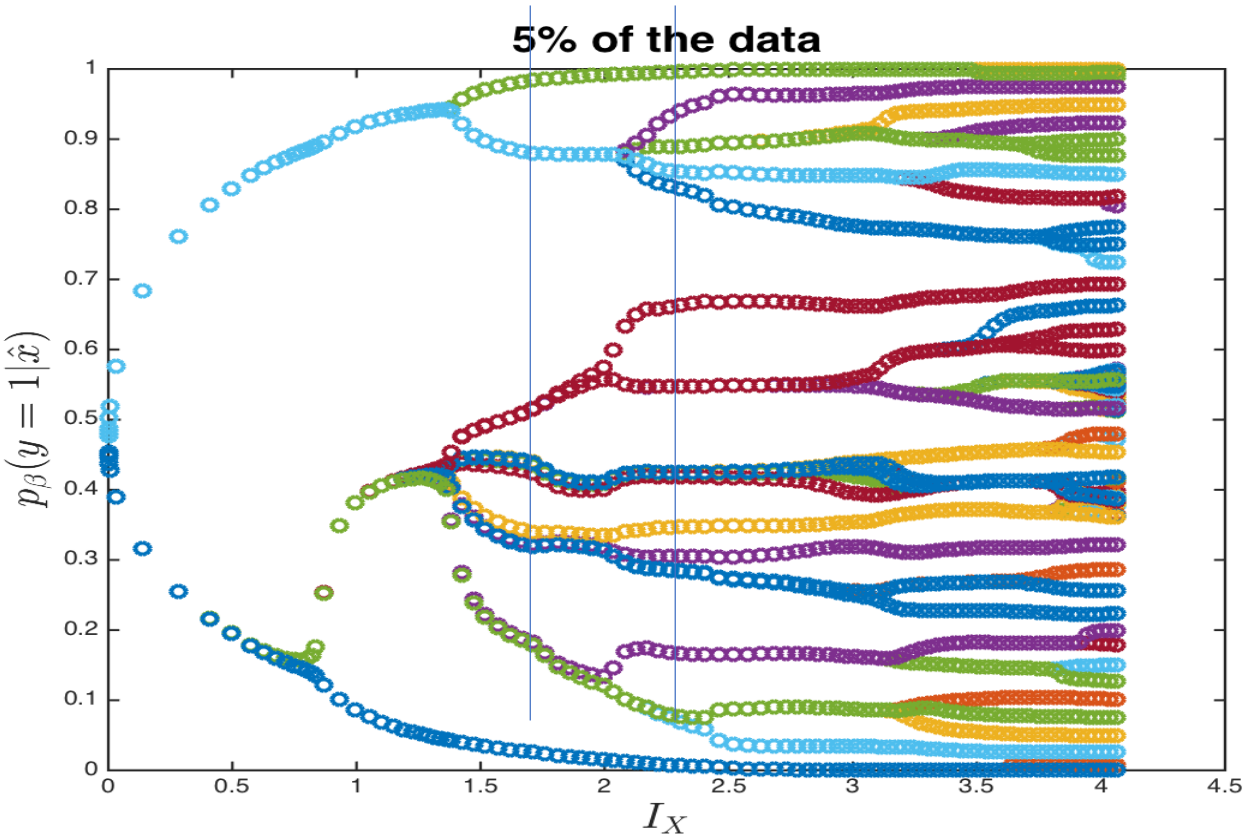
Information Bottleneck for Symmetric problems
("Group CCA")

- $T_l(X)$ *act like* ***generalized power-spectrum*** *at some frequency, which invariant under linear translations. Here these are the set of functions that are invariant under G.*

- *The IB phase transitions have multiplicity with the dimensionality of the irreducible representation, and the weights are vectors that span the irreducible representation –* ***generalized convolution****. Similar to the Hamiltonian symmetry in physics.*

- *Lower dimensional irreducible representations are strict subgroups of G, and the corresponding statistics* $T_l(X)$ *can be estimated from much smaller samples, hence the advantage for learning.*

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



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

Summary

- **Only 2 numbers per-layer matter: The Information Plane**
  - Most of the learning time goes to input compression –
  - Forgetting irrelevant details
  - Much more than curve-fitting
  - The layers converge to special (critical?) points on the IB optimal bound
- **The advantage of the many layers is mostly computational**
  - Relaxation times are super-linear (exponential?) in the compression 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 …

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.