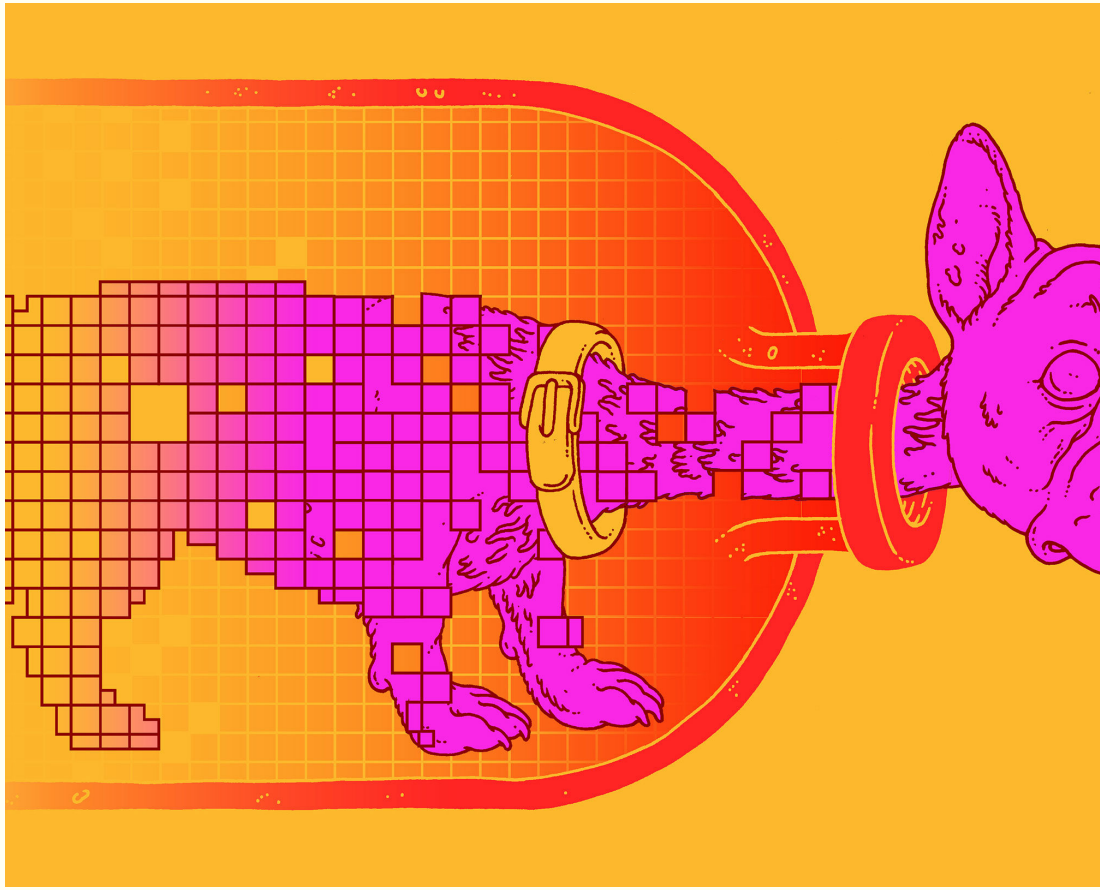
OPENING THE BLACK BOX OF DEEP NEURAL NETWORKS

Ravid Swhartz-Ziv and Naftali Tishby

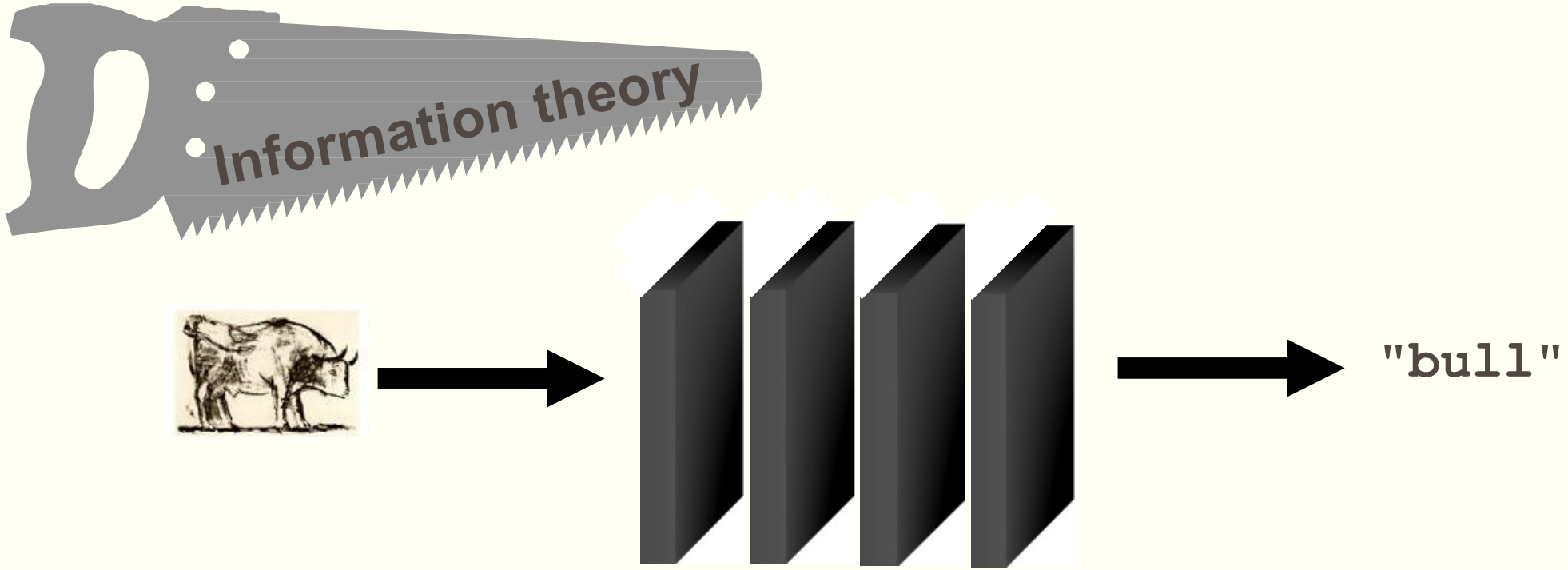
The Edmond & Lily Safra Center for Brain Sciences

Hebrew University, Jerusalem, Israel



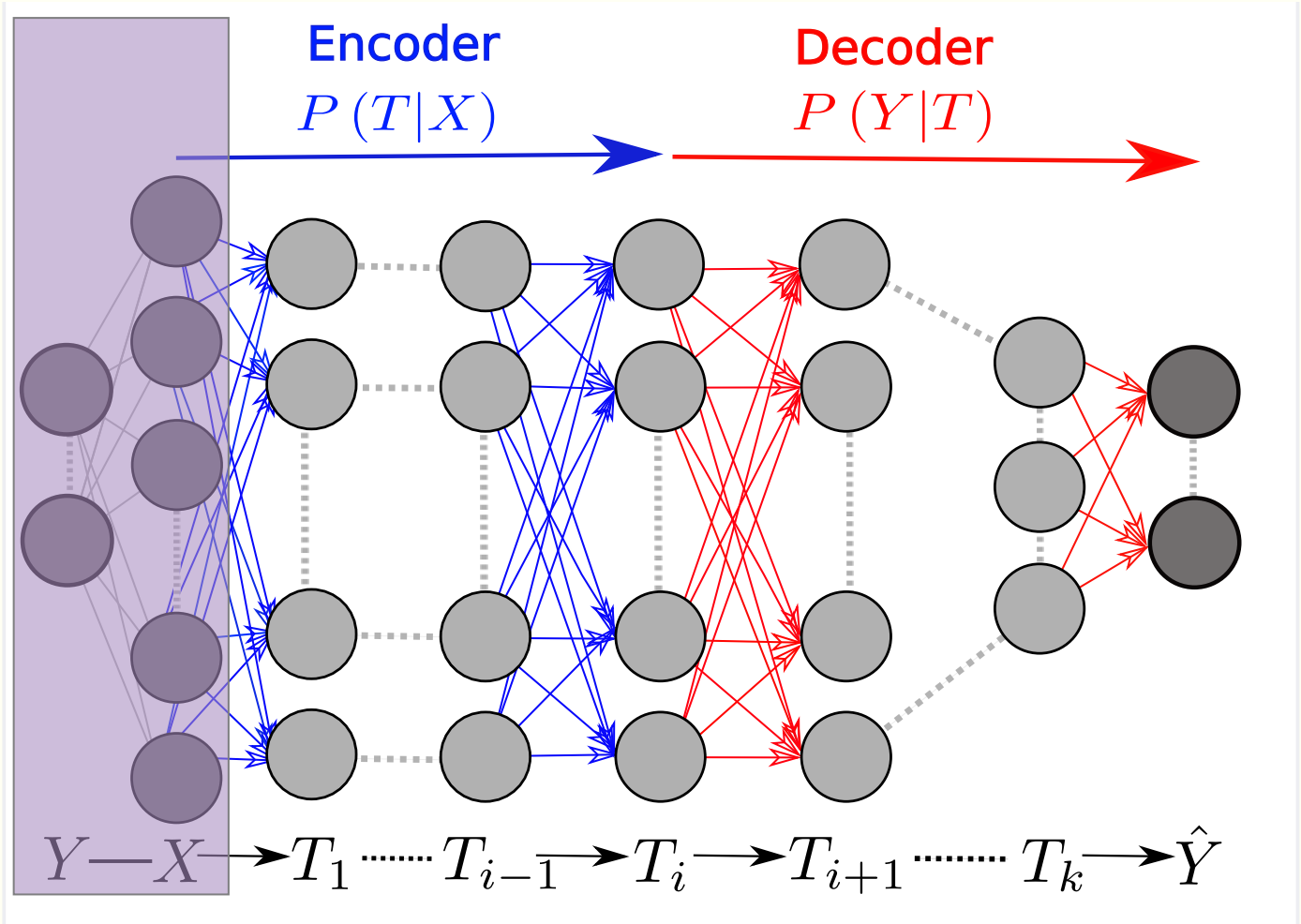
We need a theory for deep learning

Solution



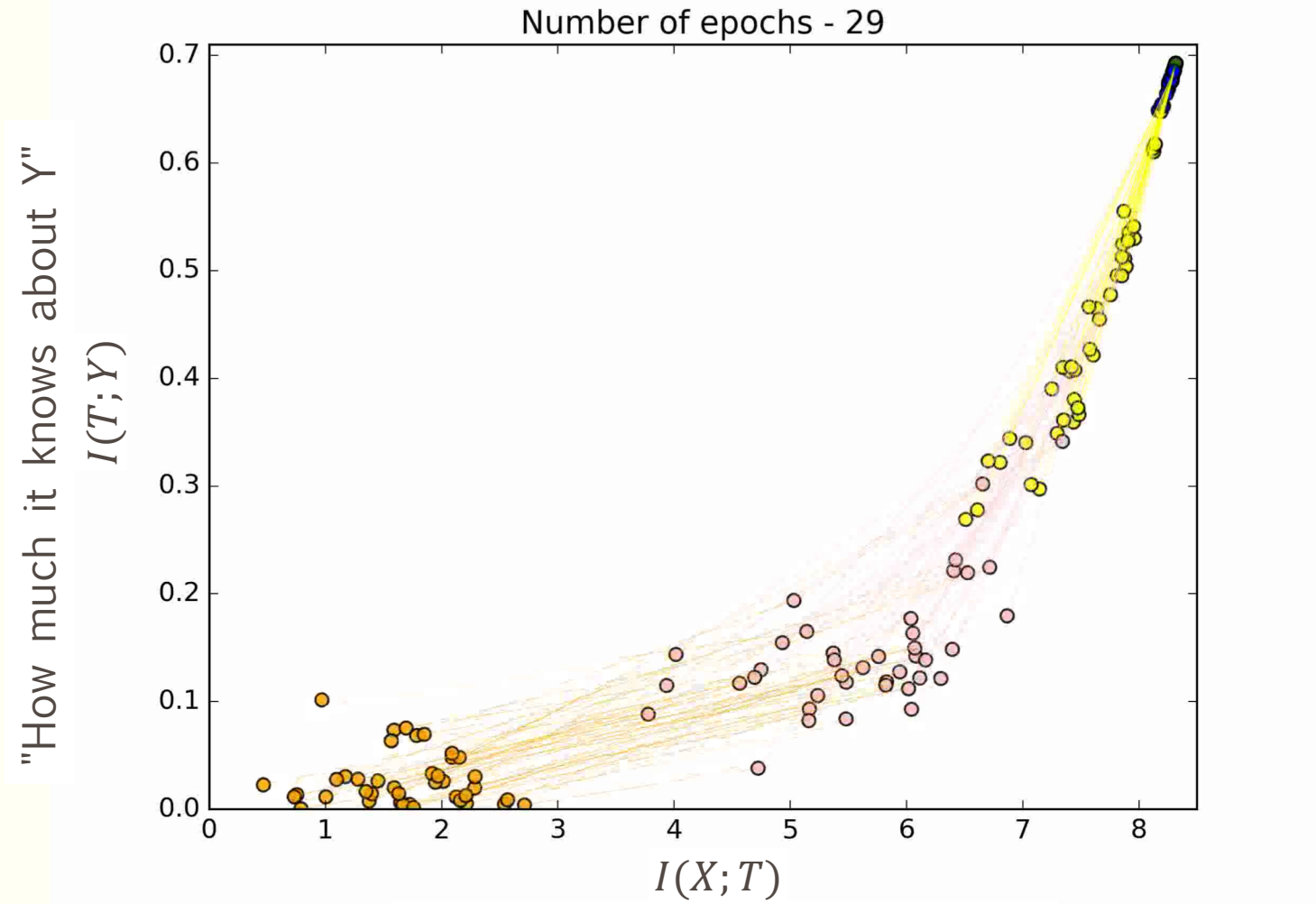
Encoder and decoder

- Encoder
  - Encodes the input
  - Complexity
- Decoder
  - Decodes predicted label
  - Generalization



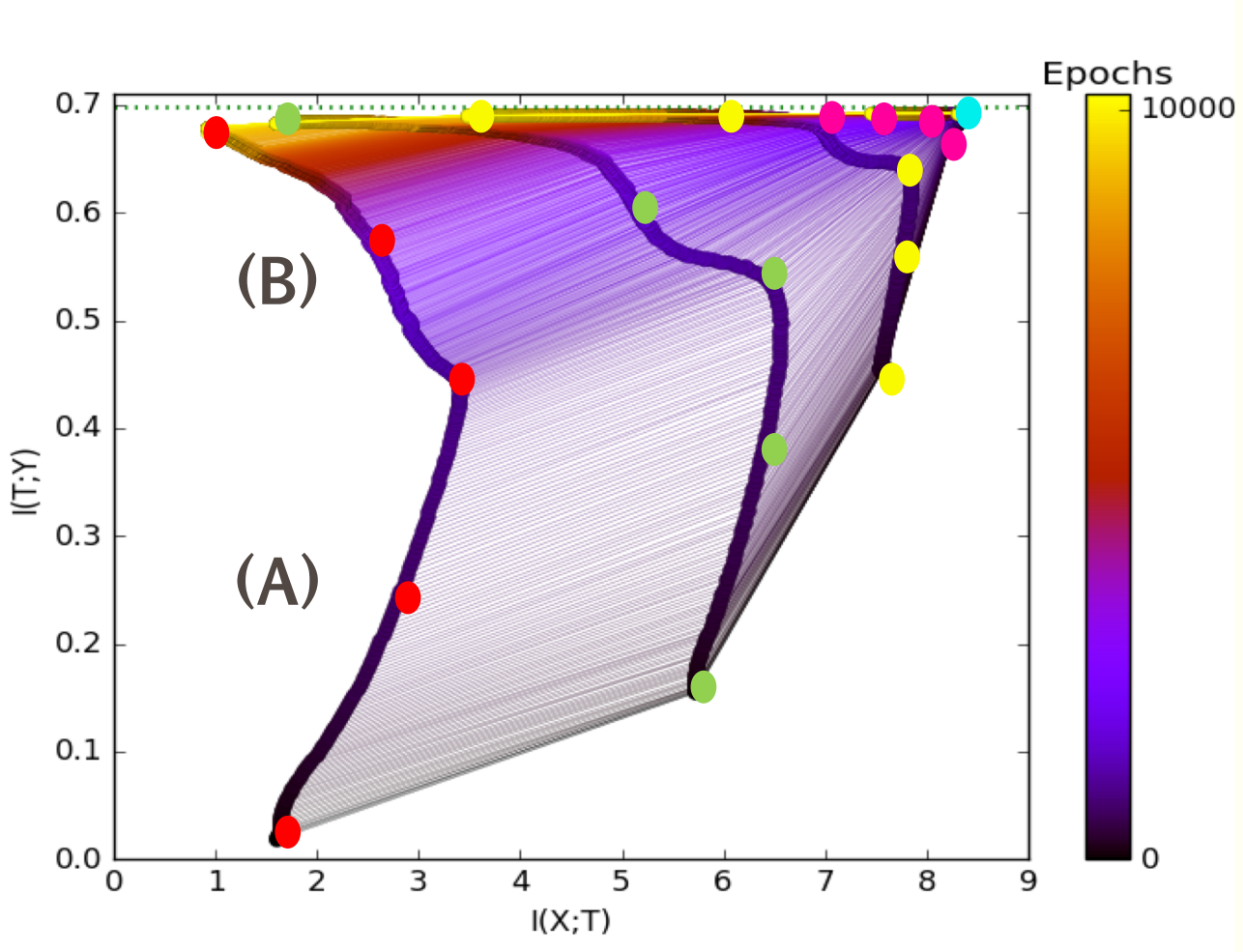
The Information During Learning

The Information During Learning

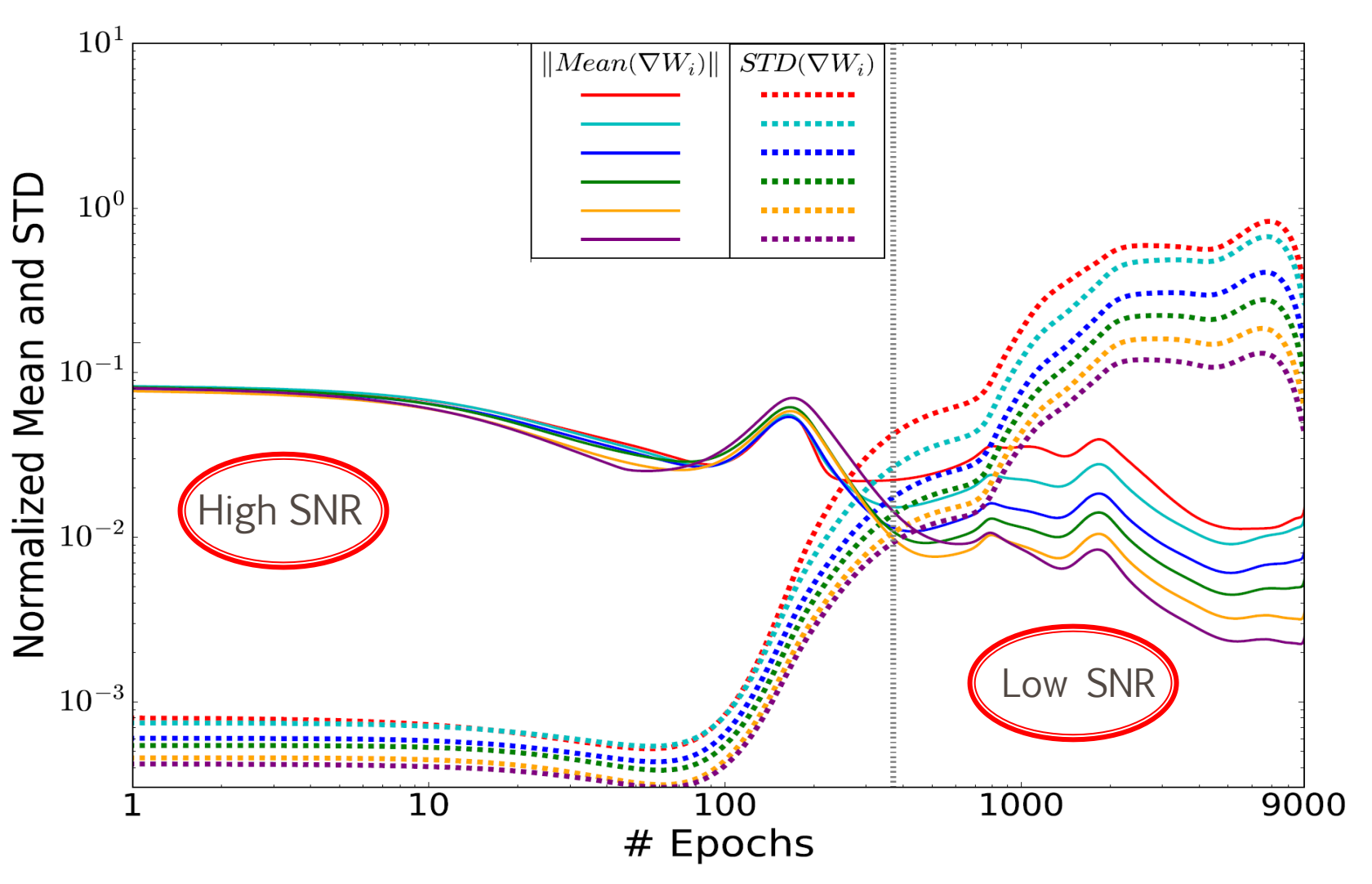


The information during learning

- Fitting (A)
- Compression (B)

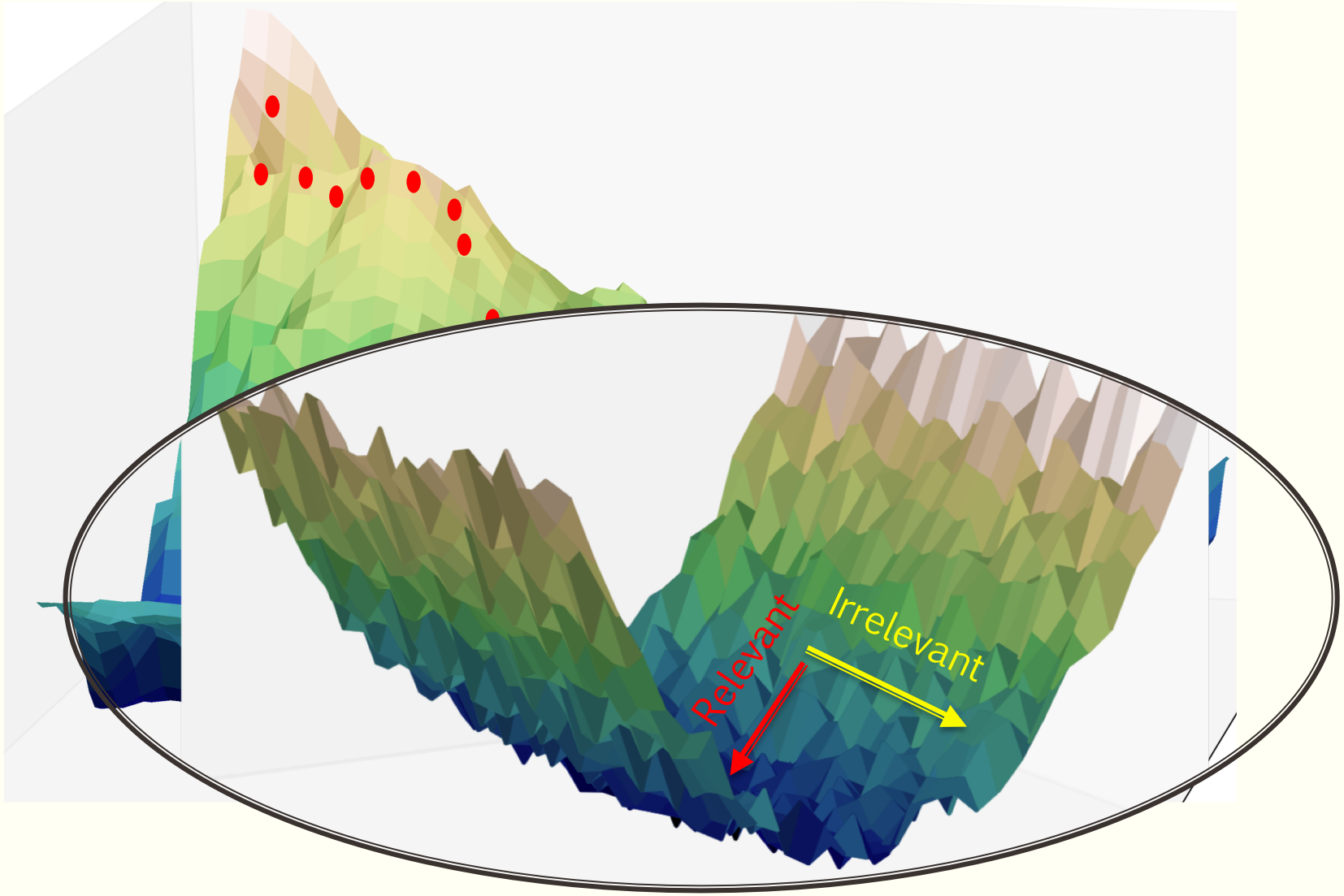


Layers' Gradients



Stochastic Relaxation and Compression

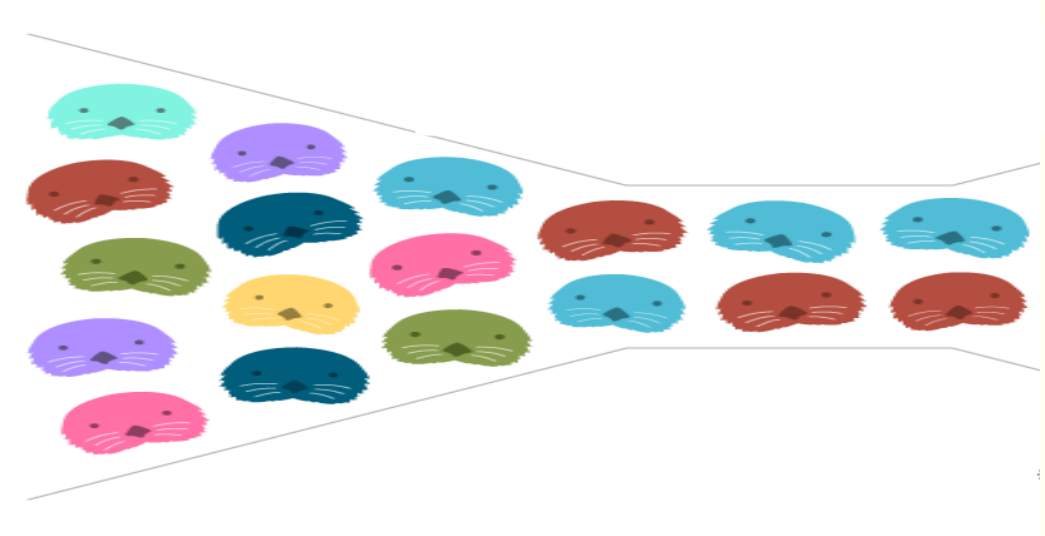
Stochastic relaxation and compression -



The Information Bottleneck method

(Tishby, Pereira, Bialek, 1999)

- Compress the input
- Captures relevant information

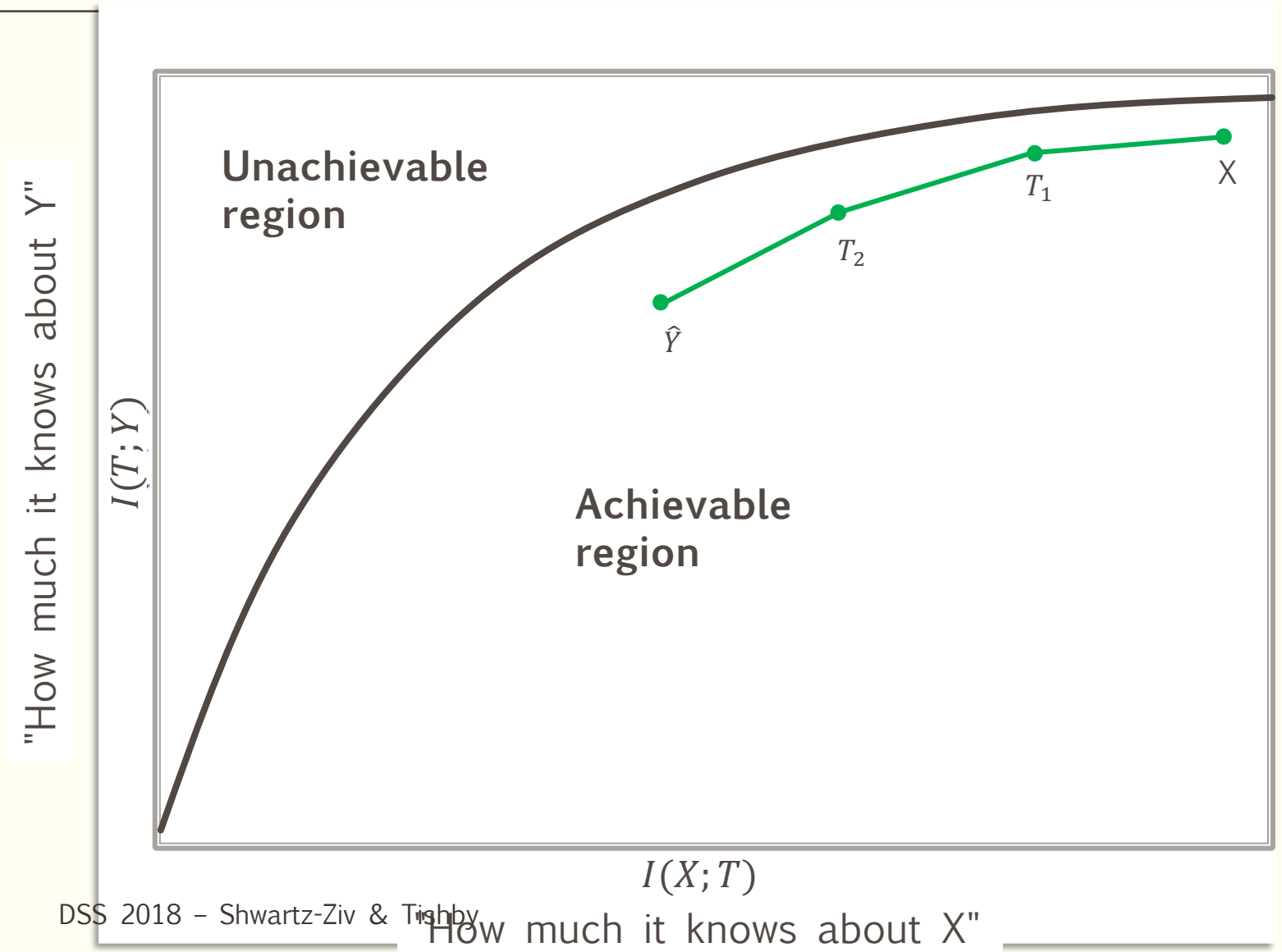


$$\underset{p(t|\mathrm{x})}{\operatorname{argmin}}\, I(T;X) - \beta I(T;Y), \qquad \beta > 0$$

*Assuming Markovian

DNNs and Information Bottleneck

- Trade-off between compression and prediction
- Optimality for each layer



SGD in continuous time

$$dw_{1\ldots K}(t) = -\nabla L(w_{1\ldots K})dt + \sqrt{\beta^{-1}D(w_{1\ldots K})}dB\ (t)$$

- $L(w_k\ )$- loss function
- $D_k(x)$ - covariance gradients
- $\mathrm{B}(t)$ - Brownian motion, β_k - noise level

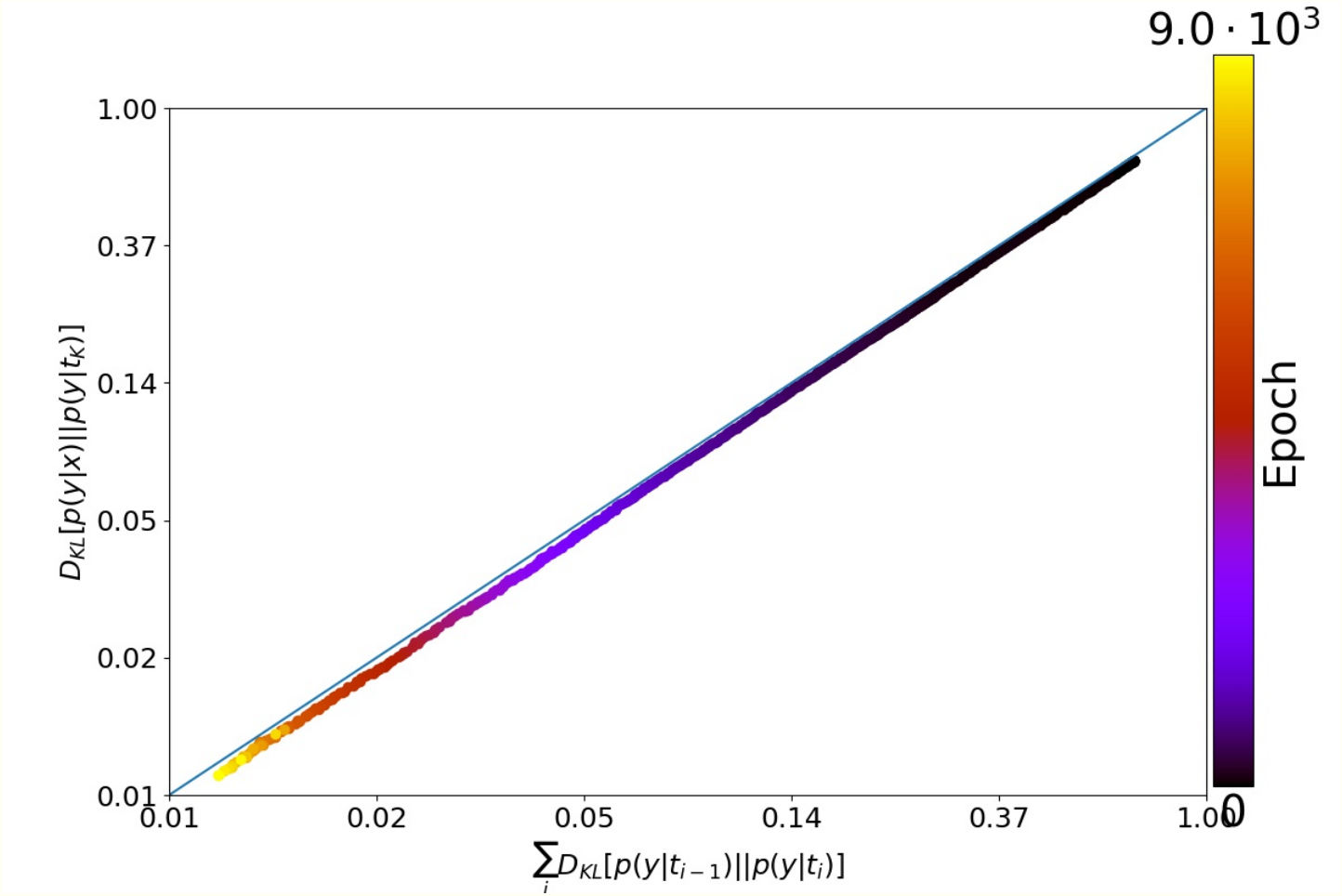
(Li et al, 2016)

Successive Refinable IB and DNNs

- Decompose the error by a set of projections
- Each one depends only on the previous one
- Pythagorean theorem

$$L = D_{KL}[p(y|x)||p(y|t_k)] \geq \sum_{i=1}^{k} D_{KL}[p(y|t_{i-1})||p(y|t_i)\,] = \tilde{L}$$

Successively Refinable Approximation of DNNs



SGD converged to the IB bound

Decompose the SGD over the layers –

$$dw_k(t) = -\nabla L_k(w_k)dt + \sqrt{\beta_k^{-1} D_k(w_k)}dB\ (t)$$

$L(w_k)$- loss function
$D_k(w_k)$ - covariance gradients
$\mathrm{B}(t)$ - Brownian motion, β_k - noise level

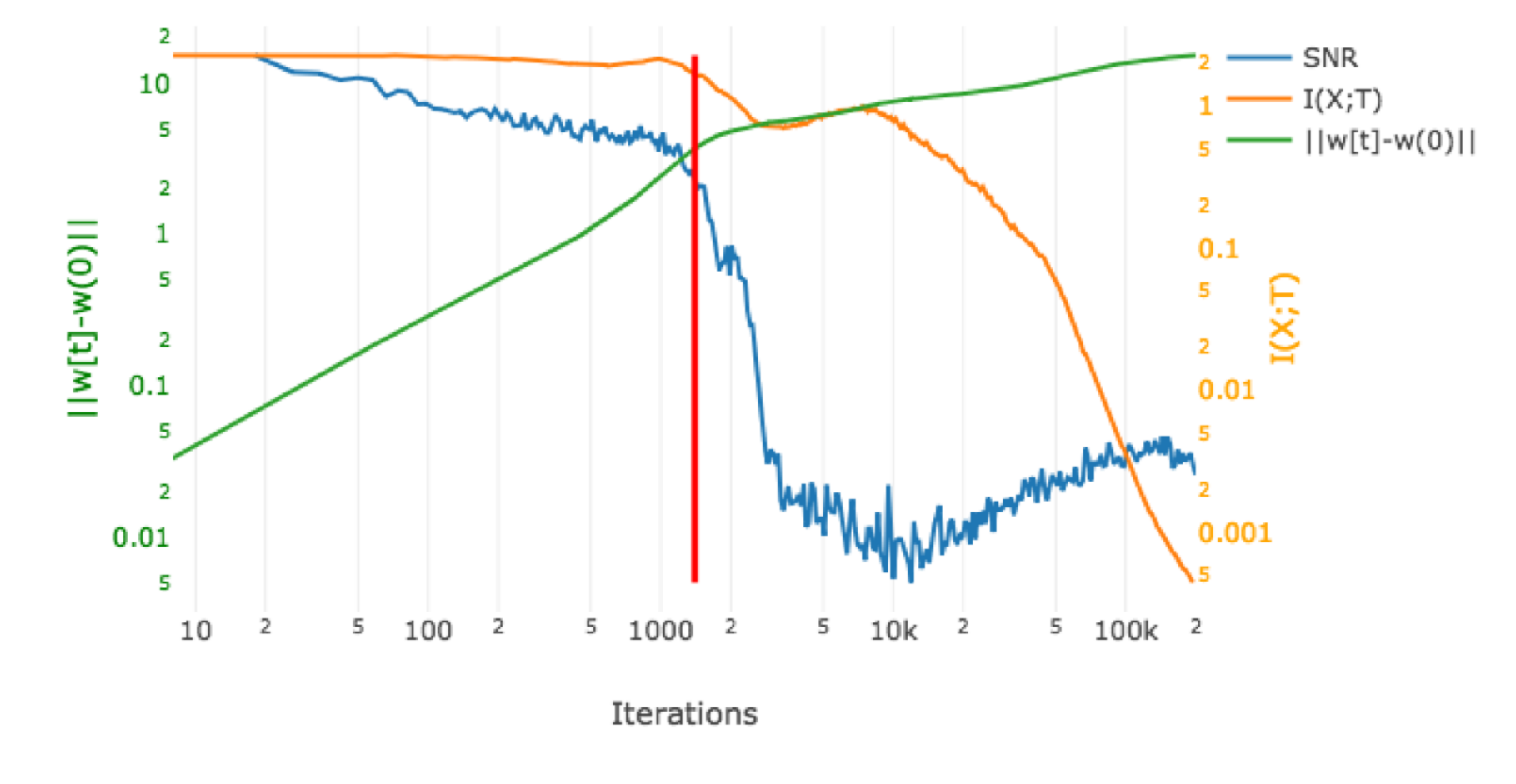
Random walk with constrain

- SGD Converged to Gibbs distribution for each layer
  - $p(W_k|X) \approx \exp(-\beta_k L(W_k|X))$
- The global minimizer of the free energy functional
- SGD brings $I(X;T_k)$ to minimum under constraint

$$F(p) = -I(Y;T_k) + \beta_k^{-1} I(X;T_k)$$

-> SGD converged to the optimal IB bound

Random walk

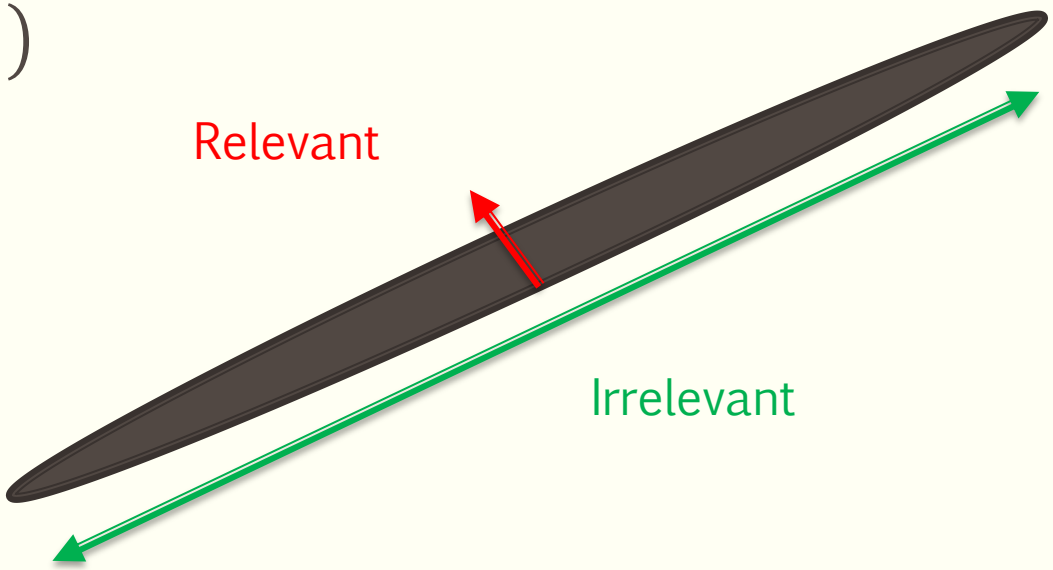


Irrelevant information compression

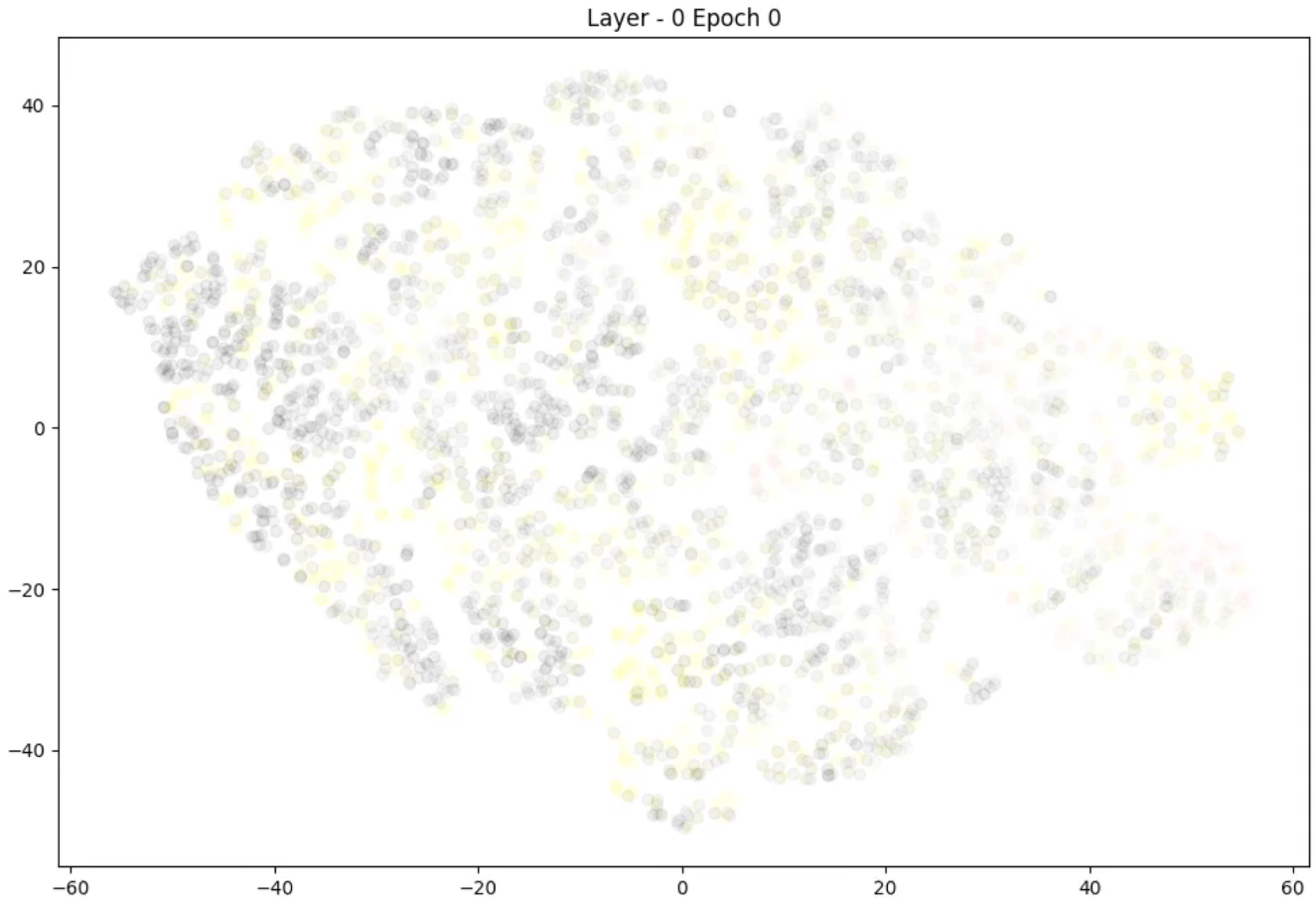
$$T_k = \sigma(W_k^* T_{k-1} + \beta_k \, \xi_k)$$

$$\xi_k \sim N\left(0, H_k^{-1} | T_{k-1} \; |\right)$$

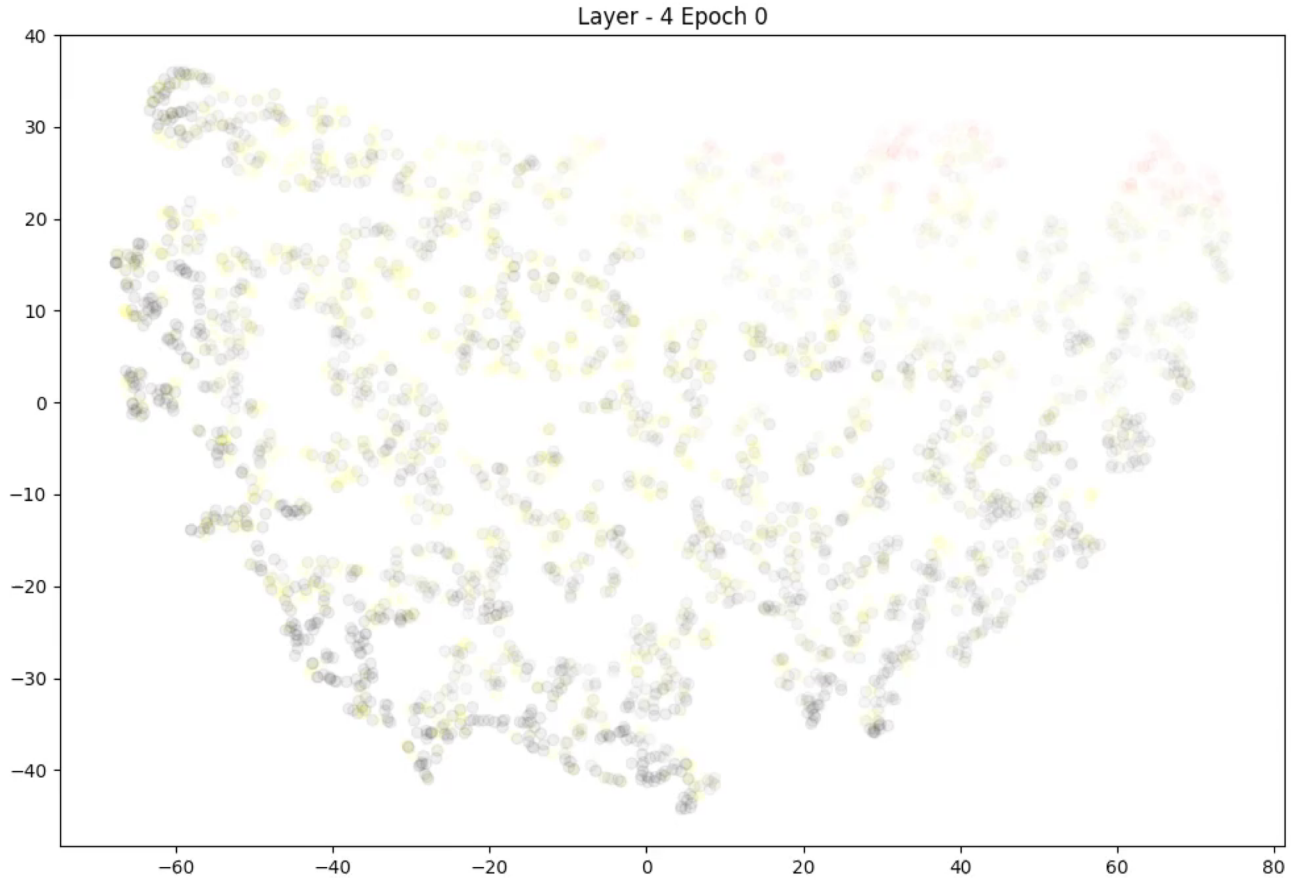
- Gaussian channel capacity
- Compression irrelevant directions



Irrelevant information compression – TSNE



Irrelevant information compression – TSNE



Summary

- Fitting and compression of information
- Stochastic relaxation and representation compression
  - SGD converged to the optimal IB bound
  - Compression of the irrelevant information

Questions?