On The Information Bottleneck Theory of Deep Learning

Erfan Loweimi

Centre for Speech Technology Research (CSTR)

Opening the black box of Deep Neural Networks via Information

Ravid Schwartz-Ziv RAVID.ZIV@MAIL.HUJI.AC.IL
Edmond and Lilly Safra Center for Brain Sciences
The Hebrew University of Jerusalem
Jerusalem, 91904, Israel

Naftali Tishby* TISHBY@CS.HUJI.AC.IL
School of Engineering and Computer Science
and Edmond and Lilly Safra Center for Brain Sciences
The Hebrew University of Jerusalem
Jerusalem, 91904, Israel

Editor: ICRI-CI

Abstract

Despite their great success, there is still no comprehensive theoretical understanding of learning with Deep Neural Networks (DNNs) or their inner organization. Previous work [Tishby and Zaslavsky (2015)] proposed to analyze DNNs in the *Information Plane*; i.e., the plane of the Mutual Information values that each layer preserves on the input and output variables. They suggested that the goal of the network is to optimize the Information Bottleneck (IB) tradeoff between compression and prediction, successively, for each layer.

Deep Learning and the Information Bottleneck Principle

Naftali Tishby[1,2] Noga Zaslavsky[1]

***Abstract*—Deep Neural Networks (DNNs) are analyzed via the theoretical framework of the information bottleneck (IB) principle. We first show that any DNN can be quantified by the mutual information between the layers and the input and output variables. Using this representation we can calculate the optimal information theoretic limits of the DNN and obtain finite sample generalization bounds. The advantage of getting closer to the theoretical limit is quantifiable both by the generalization bound and by the network's simplicity. We argue that both the optimal architecture, number of layers and features/connections at each layer, are related to the bifurcation points of the information bottleneck tradeoff, namely, relevant compression of the input layer with respect to the output layer. The hierarchical representations at the layered network naturally correspond to the structural phase transitions along the information curve. We believe that this new insight can lead to new optimality bounds and deep learning algorithms.**

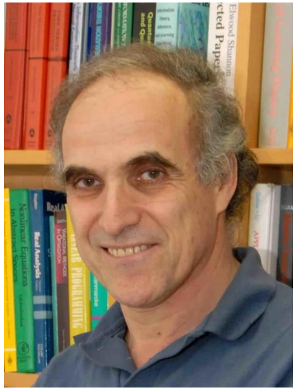
output. The information theoretic interpretation of minimal sufficient statistics [5] suggests a principled way of doing that: find a maximally compressed mapping of the input variable that preserves as much as possible the information on the output variable. This is precisely the goal of the Information Bottleneck (IB) method [6].

Several interesting issues arise when applying this principle to DNNs. First, the layered structure of the network generates a successive Markov chain of intermediate representations, which together form the (approximate) sufficient statistics. This is closely related to successive refinement of information in Rate Distortion Theory [7]. Each layer in the network can now be quantified by the amount of information it retains on the input variable, on the (desired) output vari-

ELSC
The Edmond & Lily Safra Center for Brain Sciences

ICRI-CI intel

THE HEBREW UNIVERSITY OF JERUSALEM



Professor Naftali Tishby

Noga Zaslavsky

Ravid Schwartz-Ziv

Opening the black box of Deep Neural Networks via Information

Ravid Schwartz-Ziv RAVID.ZIV@MAIL.HUJI.AC.IL
Edmond and Lilly Safra Center for Brain Sciences
The Hebrew University of Jerusalem
Jerusalem, 91904, Israel

Naftali Tishby* TISHBY@CS.HUJI.AC.IL
School of Engineering and Computer Science
and Edmond and Lilly Safra Center for Brain Sciences
The Hebrew University of Jerusalem
Jerusalem, 91904, Israel

Editor: ICRI-CI

Opening the Black Box of Deep Neural Networks via Information
https://arxiv.org › cs ▾
by R Shwartz-Ziv - 2017 - Cited by 183 - Related articles
2 Mar 2017 - This generalization through noise mechanism is unique to **Deep Neural Networks** and absent in one layer networks. (iv) The training time is ...

Rejected in both ICML and NIPS!

Deep Learning and the Information Bottleneck Principle

Naftali Tishby[1,2] Noga Zaslavsky[1]

***Abstract*— Deep Neural Networks (DNNs) are analyzed via the theoretical framework of the information bottleneck (IB) principle. We first show that any DNN can be quantified by the mutual information between the layers and the input and output variables. Using this representation we can calculate the optimal information theoretic limits of the DNN and** output. The information theoretic interpretation of minimal sufficient statistics [5] suggests a principled way of doing that: find a maximally compressed mapping of the input variable that preserves as much as possible the information on the output variable. This is precisely the goal of the

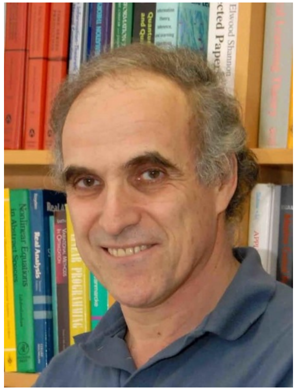
Deep Learning and the Information Bottleneck Principle
https://arxiv.org › cs ▾
by N Tishby - 2015 - Cited by 143 - Related articles
9 Mar 2015 - **Deep Learning and the Information Bottleneck Principle. Deep Neural Networks** (DNNs) are analyzed via the theoretical framework of the **information bottleneck** (IB) **principle**. We first show that any DNN can be quantified by the mutual **information** between the layers and the input and output variables.

19, Nov. , 2018

The Edmond & Lily Safra Center for Brain Sciences

ICRI-CI

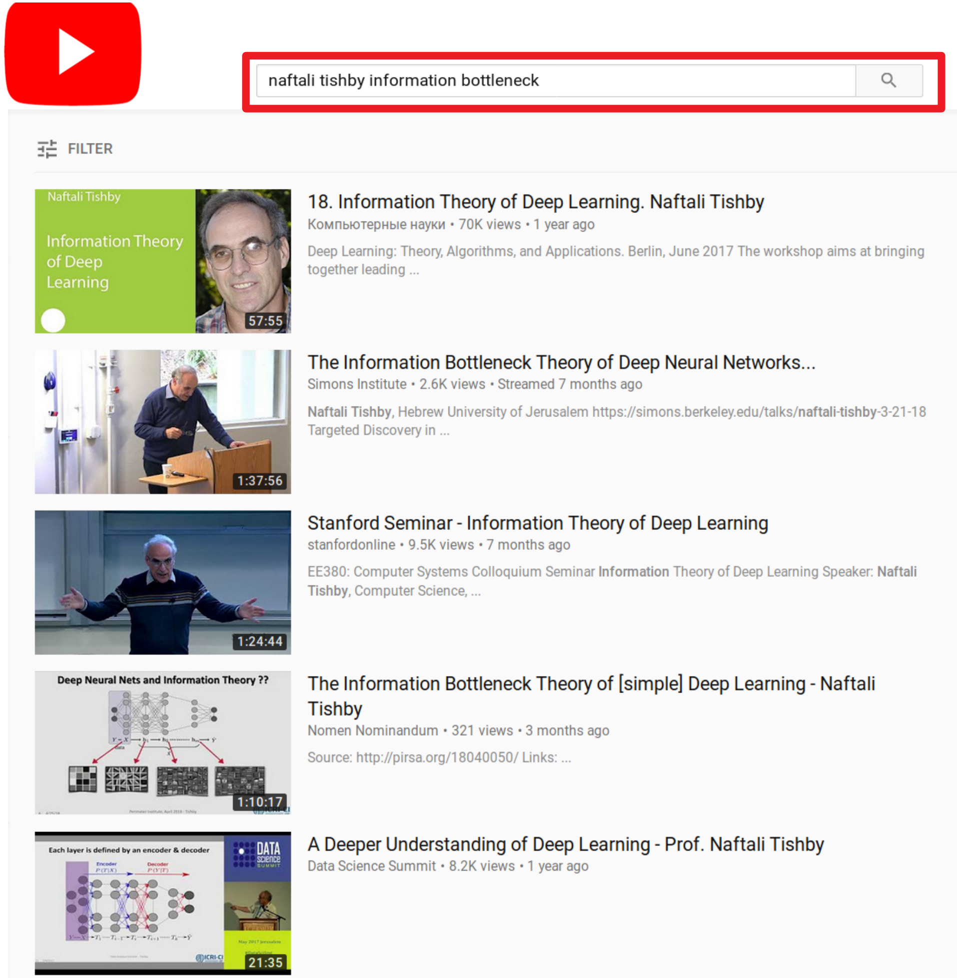
THE HEBREW UNIVERSITY OF JERUSALEM



Professor Naftali Tishby

Noga Zaslavsky

Ravid Schwartz-Ziv



naftali tishby information bottleneck

FILTER

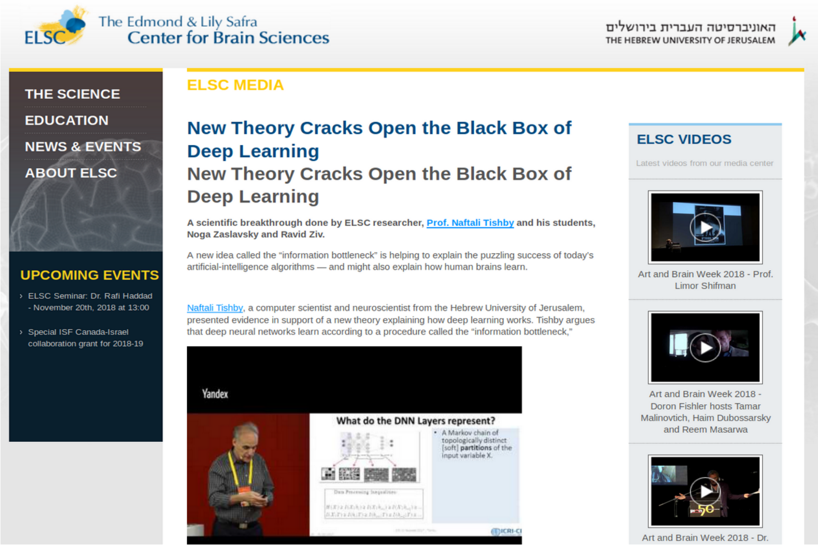
18. Information Theory of Deep Learning. Naftali Tishby
Компьютерные науки • 70K views • 1 year ago
Deep Learning: Theory, Algorithms, and Applications. Berlin, June 2017 The workshop aims at bringing together leading ...

The Information Bottleneck Theory of Deep Neural Networks...
Simons Institute • 2.6K views • Streamed 7 months ago
Naftali Tishby, Hebrew University of Jerusalem https://simons.berkeley.edu/talks/naftali-tishby-3-21-18 Targeted Discovery in ...

Stanford Seminar - Information Theory of Deep Learning
stanfordonline • 9.5K views • 7 months ago
EE380: Computer Systems Colloquium Seminar Information Theory of Deep Learning Speaker: Naftali Tishby, Computer Science, ...

The Information Bottleneck Theory of [simple] Deep Learning - Naftali Tishby
Nomen Nominandum • 321 views • 3 months ago
Source: http://pirsa.org/18040050/ Links: ...

A Deeper Understanding of Deep Learning - Prof. Naftali Tishby
Data Science Summit • 8.2K views • 1 year ago



ELSC The Edmond & Lily Safra Center for Brain Sciences

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

THE SCIENCE
EDUCATION
NEWS & EVENTS
ABOUT ELSC

UPCOMING EVENTS

› ELSC Seminar: Dr. Rafi Haddad - November 20th, 2018 at 13:00

› Special ISF Canada-Israel collaboration grant for 2018-19

ELSC MEDIA

New Theory Cracks Open the Black Box of Deep Learning

New Theory Cracks Open the Black Box of Deep Learning

A scientific breakthrough done by ELSC researcher, Prof. Naftali Tishby and his students, Noga Zaslavsky and Ravid Ziv.

A new idea called the "information bottleneck" is helping to explain the puzzling success of today's artificial-intelligence algorithms — and might also explain how human brains learn.

Naftali Tishby, a computer scientist and neuroscientist from the Hebrew University of Jerusalem, presented evidence in support of a new theory explaining how deep learning works. Tishby argues that deep neural networks learn according to a procedure called the "information bottleneck,"

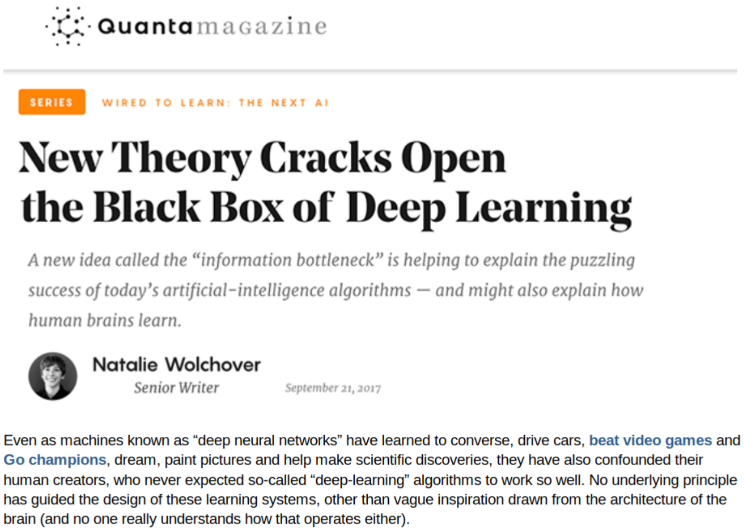
ELSC VIDEOS

Latest videos from our media center

Art and Brain Week 2018 - Prof. Limor Shifman

Art and Brain Week 2018 - Doron Fishler hosts Tamar Malinovitch, Haim Dubossarsky and Reem Masarwa

Art and Brain Week 2018 - Dr.



Quanta magazine

SERIES WIRED TO LEARN: THE NEXT AI

New Theory Cracks Open the Black Box of Deep Learning

A new idea called the "information bottleneck" is helping to explain the puzzling success of today's artificial-intelligence algorithms — and might also explain how human brains learn.

Natalie Wolchover
Senior Writer
September 21, 2017

Even as machines known as "deep neural networks" have learned to converse, drive cars, beat video games and Go champions, dream, paint pictures and help make scientific discoveries, they have also confounded their human creators, who never expected so-called "deep-learning" algorithms to work so well. No underlying principle has guided the design of these learning systems, other than vague inspiration drawn from the architecture of the brain (and no one really understands how that operates either).

Outlines

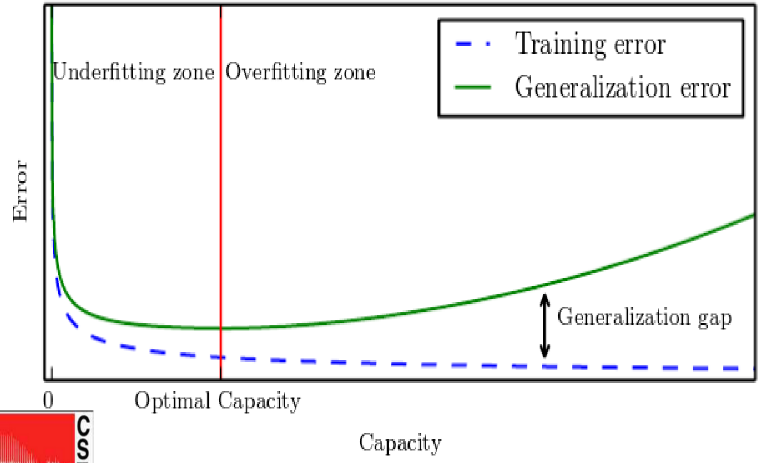
- Problem Statement
- Information Theory Review
- Information Bottleneck (IB)
- Opening the Black Box of DNNs via IB
- Other Views
- Conclusions

Outlines

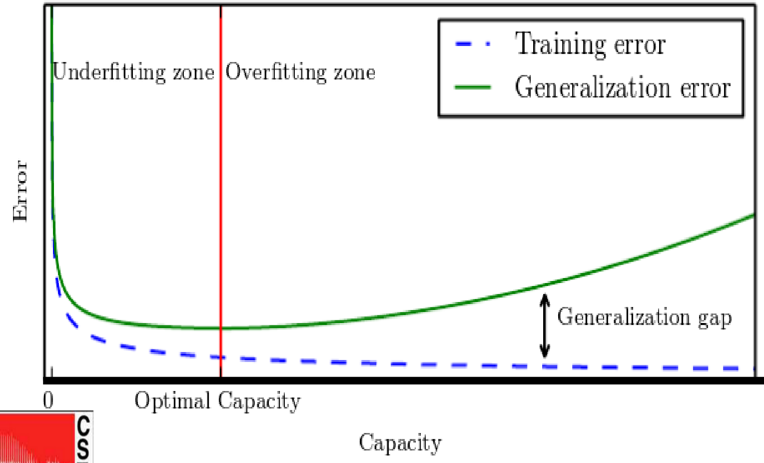
- Problem Statement
- Information Theory Review
- Information Bottleneck (IB)
- Opening the Black Box of DNNs via IB
- Criticisms
- Conclusions

- Why DNNs work/generalise well
- Interpretability/understanding

- Why DNNs work/generalise well
- Interpretability/understanding

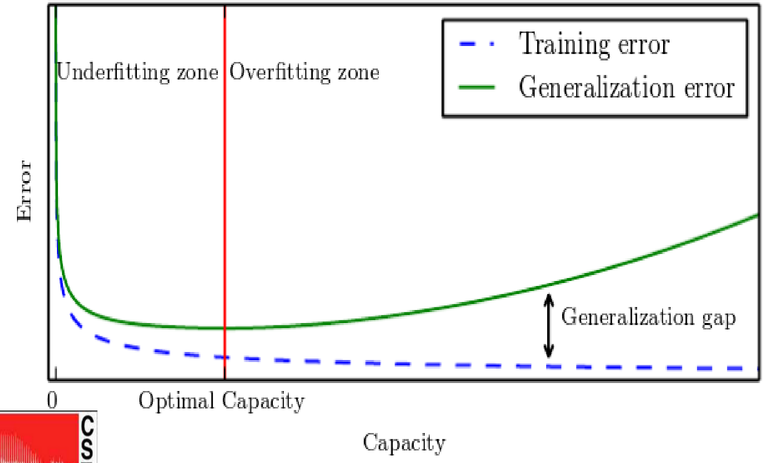


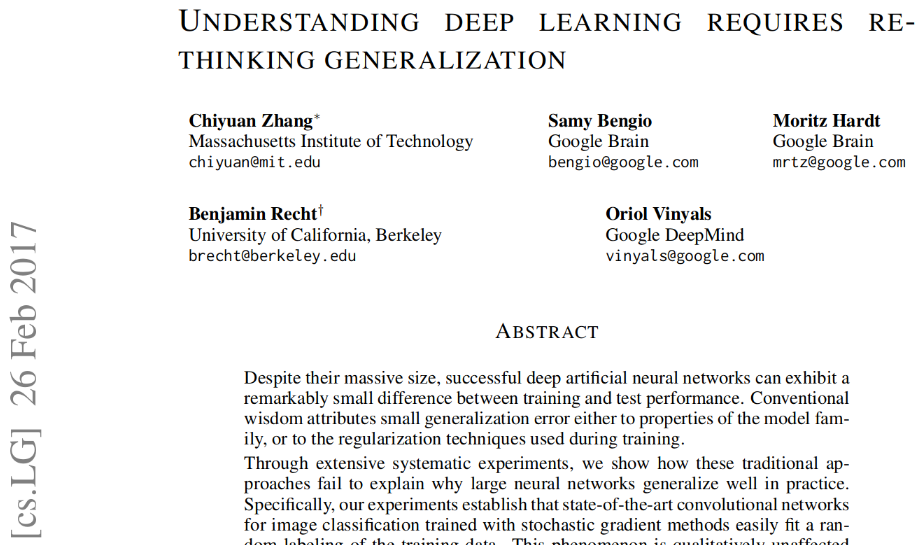
- Why DNNs work/generalise well
- Interpretability/understanding



DNNs

- Why DNNs work/generalise well
- Interpretability/understanding





[cs.LG] 26 Feb 2017

UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

Chiyuan Zhang*
Massachusetts Institute of Technology
chiyuan@mit.edu

Samy Bengio
Google Brain
bengio@google.com

Moritz Hardt
Google Brain
mrtz@google.com

Benjamin Recht†
University of California, Berkeley
brecht@berkeley.edu

Oriol Vinyals
Google DeepMind
vinyals@google.com

ABSTRACT

Despite their massive size, successful deep artificial neural networks can exhibit a remarkably small difference between training and test performance. Conventional wisdom attributes small generalization error either to properties of the model family, or to the regularization techniques used during training.

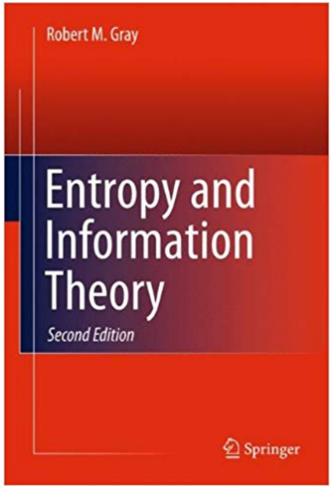
Through extensive systematic experiments, we show how these traditional approaches fail to explain why large neural networks generalize well in practice. Specifically, our experiments establish that state-of-the-art convolutional networks for image classification trained with stochastic gradient methods easily fit a ran-

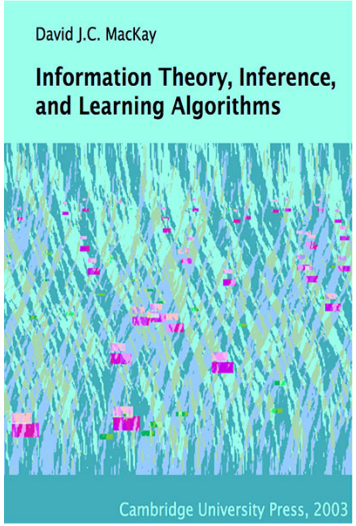
Outlines

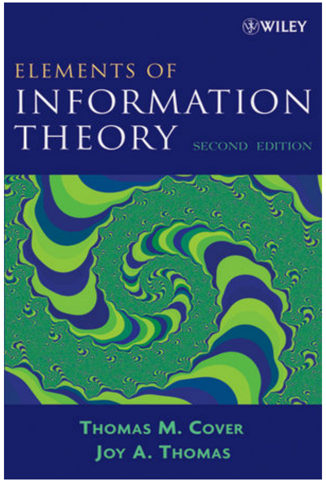
- Problem Statement
- Information Theory Review
- Information Bottleneck (IB)
- Opening the Black Box of DNNs via IB
- Criticisms
- Conclusions

Information Theory Review

- Information
- (Conditional, Joint) Entropy
- KL Divergence
- Mutual Information







Information Theory

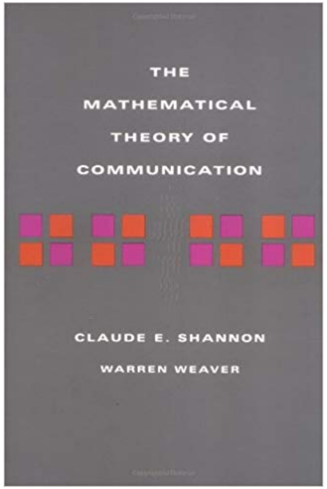
Reprinted with corrections from *The Bell System Technical Journal*,
Vol. 27, pp. 379–423, 623–656, July, October, 1948.

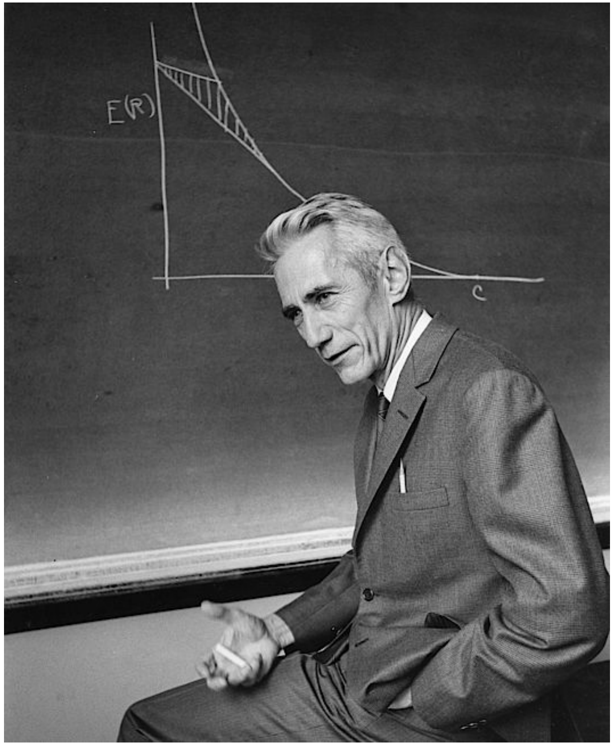
A Mathematical Theory of Communication

By C. E. SHANNON

Introduction

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.





Information Theory

Reprinted with corrections from *The Bell System Technical Journal*,
Vol. 27, pp. 379–423, 623–656, July, October, 1948.

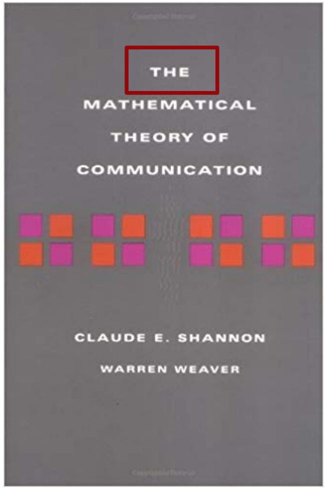
A Mathematical Theory of Communication

By C. E. SHANNON

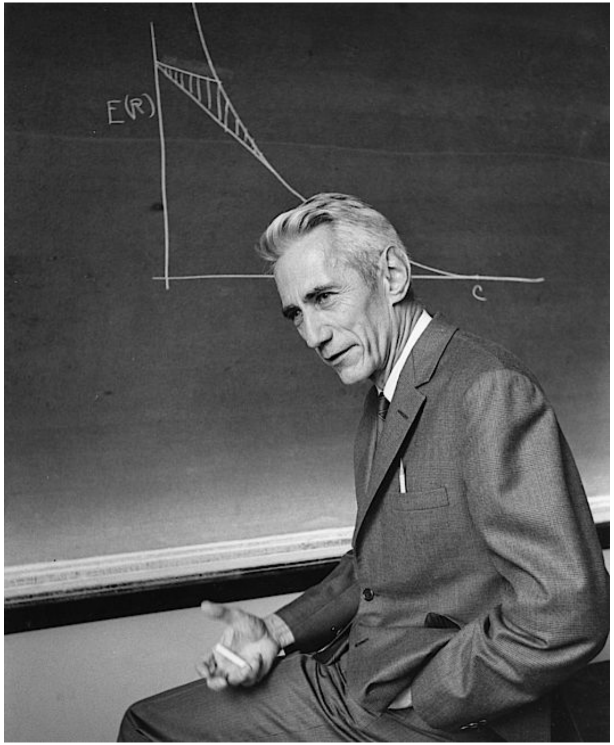
INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

1948



1949



Defining Information – Qualitatively

- Fundamental properties of information (I)

Defining Information – Qualitatively

- Fundamental properties of information (I)
  - I(p) is monotonically decreasing in probability (p)

Defining Information – Qualitatively

- Fundamental properties of information (I)
  - I(p) is monotonically decreasing in probability (p)

$$
\begin{aligned}
\text{Information} &\equiv \text{Surprise} \\
\text{Informaiton} &\equiv \text{Uncertainty}
\end{aligned}
$$

Defining Information – Qualitatively

- Fundamental properties of information (I)
  - I(p) is monotonically decreasing in probability (p)
  - $I(p) \geq 0$
  - $I(p=1) = 0$

Defining Information – Qualitatively

- Fundamental properties of information (I)
  - I(p) is monotonically decreasing in probability (p)
  - $I(p) \geq 0$
  - $I(p=1) = 0$
  - $I(x_1,x_2) = I(x_1) + I(x_2) \leftrightarrow x_1 \perp\!\!\!\perp x_2$

Independent

Defining Information – Quantitatively

- Information ≡ Average surprise/uncertainty

$$H(X) = \mathbb{E}[\,\log \frac{1}{P(X)}\,] = -\sum_{x \in \mathcal{X}} P(x_i)\,\,\log_b P(x_i)$$

Defining Information – Quantitatively

- Information ≡ Average surprise/uncertainty

Entropy

Self information

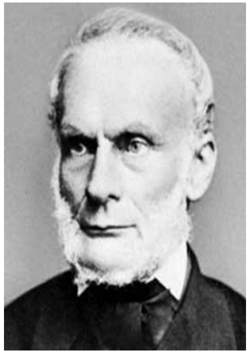
$$H(X) = \mathbb{E}\left[\ \boxed{\log \frac{1}{P(X)}}\ \right] = -\sum_{x \in \mathcal{X}} P(x_i)\ \log_b P(x_i)$$

Set of char

b = 2 → bit
b = e → nat
b = 10 → Hartley

History of Entropy

R. Clausius



$$dS = \frac{dQ}{T}$$

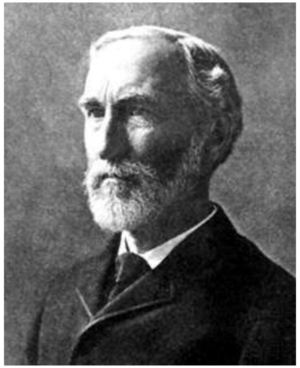
1865

L. Boltzmann

$$S = k_B \ \log W$$

1870

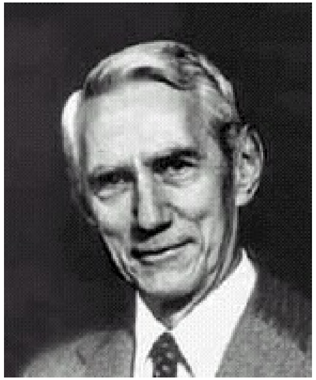
JW Gibbs



$$S = -k_B \sum_i p_i \ \log p_i$$

1876

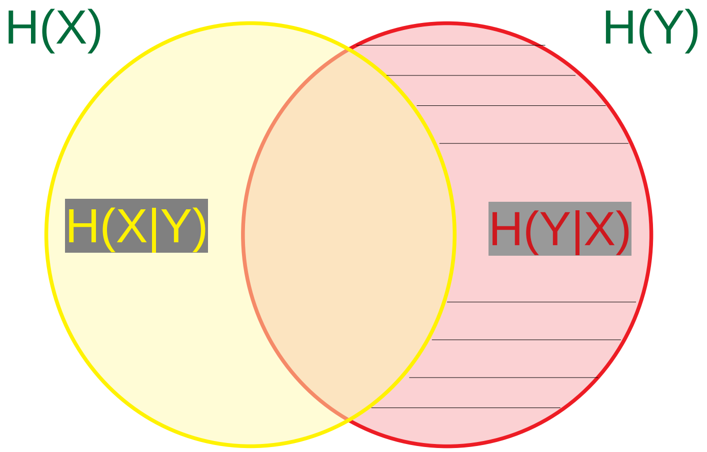
C. Shannon



$$H(X) = -\sum_i p_i \ \log_2 p_i$$

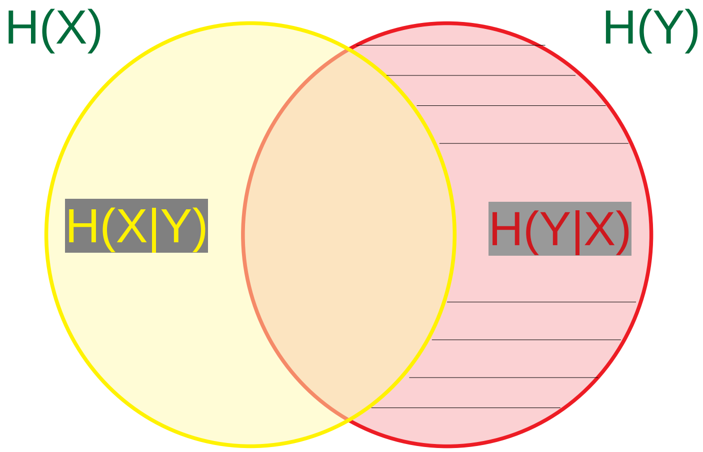
1948

Conditional Entropy



Conditional Entropy

- $H(Y|X) \rightarrow$ Remaining uncertainty in Y, when X is known

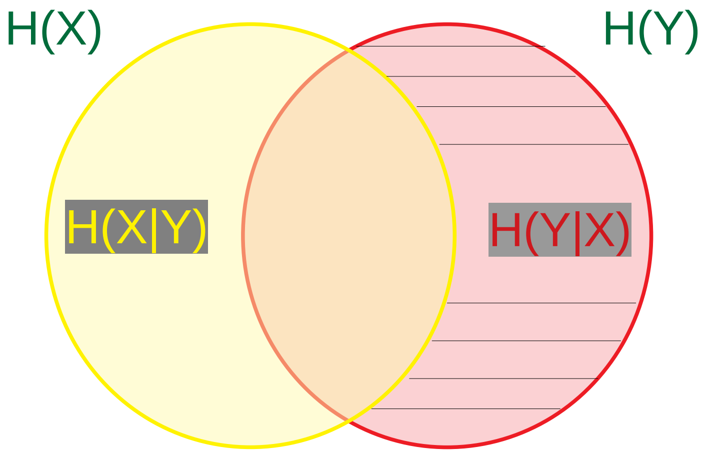


Conditional Entropy

- H(Y|X) → Remaining uncertainty in Y, when X is known

$$H(Y|X = x_i) = \mathbb{E}[\mathbb{I}(Y)|X = x_i]$$
$$= -\sum_{y \in \mathcal{Y}} P(y|x_i) \log P(y|x_i)$$

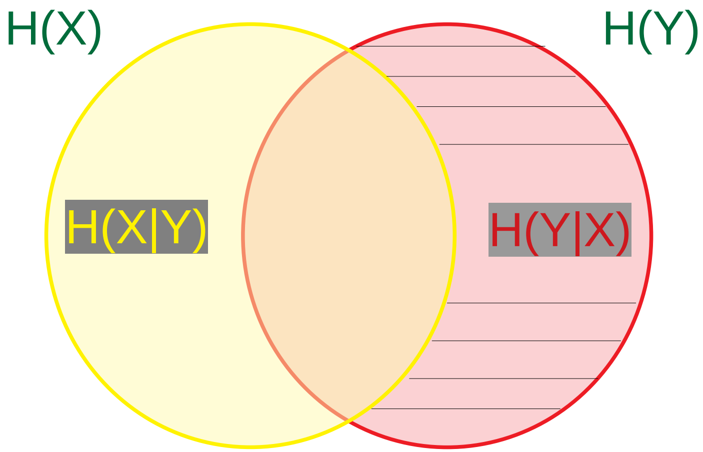
$$H(Y|X) = -\sum_{x_i \in \mathcal{X}} p(x_i) H(Y|X = x_i)$$



Conditional Entropy

- H(Y|X) → Remaining uncertainty in Y, when X is known

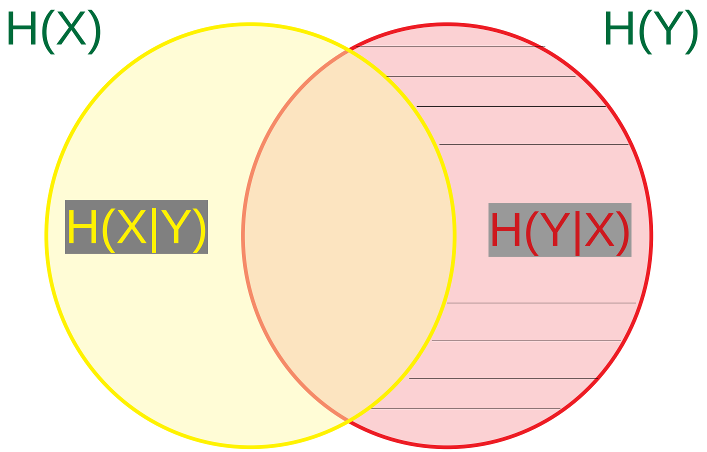
$$\boxed{H(Y) \geq H(Y|X) \geq 0}$$



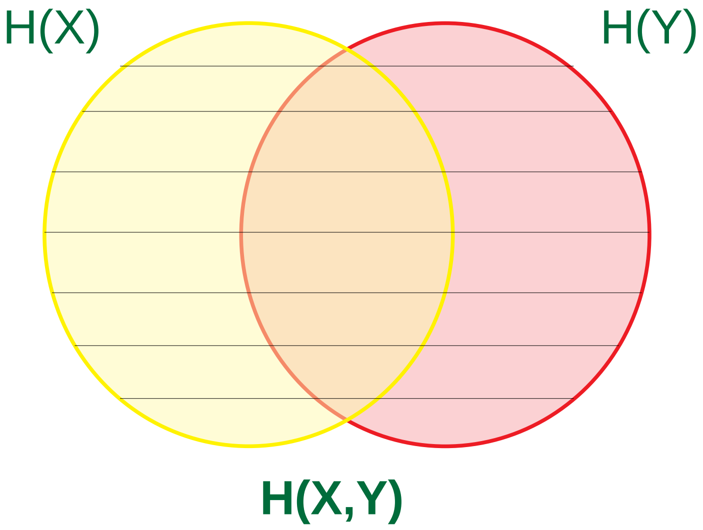
Conditional Entropy

- H(Y|X) $\rightarrow$ Remaining uncertainty in Y, when X is known

$$\boxed{H(Y) \underset{\text{Independent}}{\geq} H(Y|X) \underset{\text{Deterministic}}{\geq} 0}$$

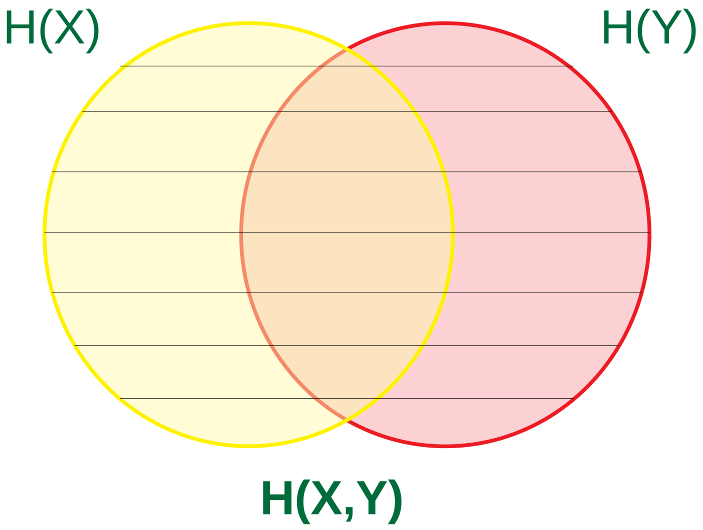


Joint Entropy



Joint Entropy

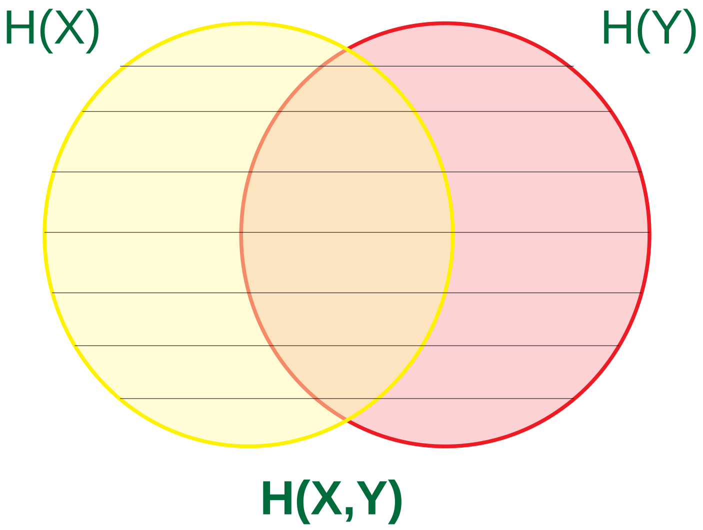
- H(X,Y) → Uncertainty associated with a set of Variables



Joint Entropy

- H(X,Y) → Uncertainty associated with a set of Variables

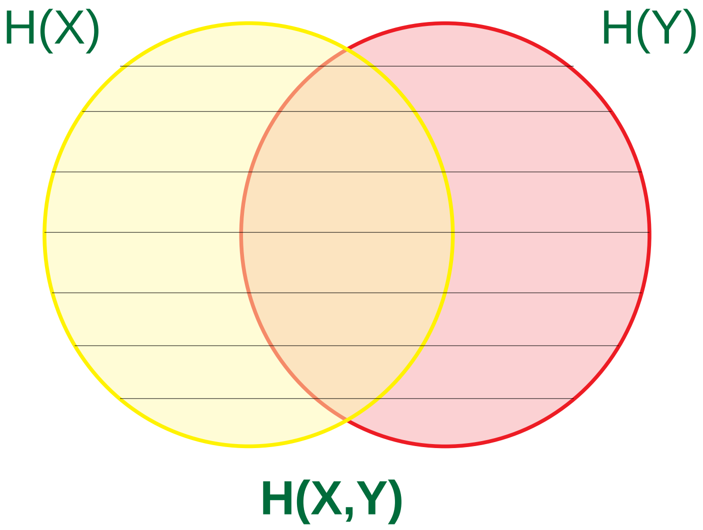
$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x,y)\log P(x,y)$$



Joint Entropy

- H(X,Y) → Uncertainty associated with a set of Variables

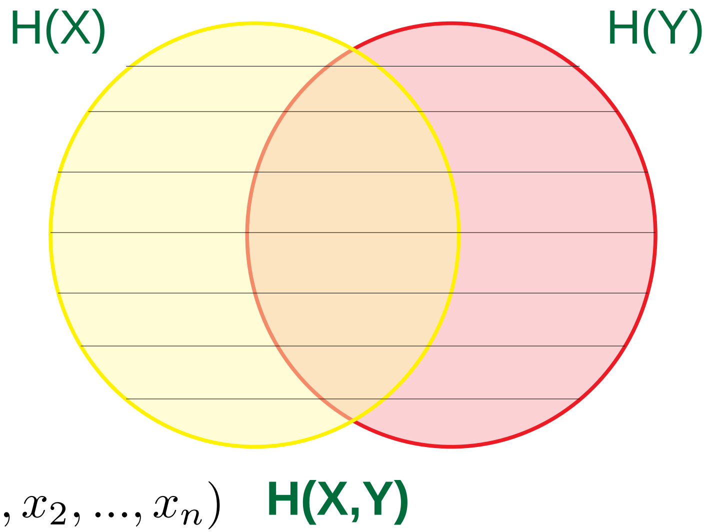
$$
\begin{aligned}
H(X,Y) &= H(Y|X) + H(X) \\
&= H(X|Y) + H(Y)
\end{aligned}
$$



Joint Entropy

- H(X,Y) → Uncertainty associated with a set of Variables

$$H(X,Y) = H(Y|X) + H(X) \\ = H(X|Y) + H(Y)$$



$$H(X_1, X_2, ..., X_n) = -\sum_{x_1}\sum_{x_2}...\sum_{x_n} P(x_1, x_2, ..., x_n) \log P(x_1, x_2, ..., x_n)$$

Kullback–Leibler Divergence (KLD)

$$D_{KL}(P||Q) = -\sum_i P(i)\ \log \frac{Q(i)}{P(i)} = \boxed{H(P,Q)} - H(P)$$

Cross Entropy

Kullback–Leibler Divergence (KLD)

$$D_{KL}(P||Q) = -\sum_i P(i)\ \log \frac{Q(i)}{P(i)} = H(P,Q) - H(P)$$

Interpretation:

– Distance measure (P: ref; Q: est)

– Information gain (P: posterior; Q: prior)

Kullback–Leibler Divergence (KLD)

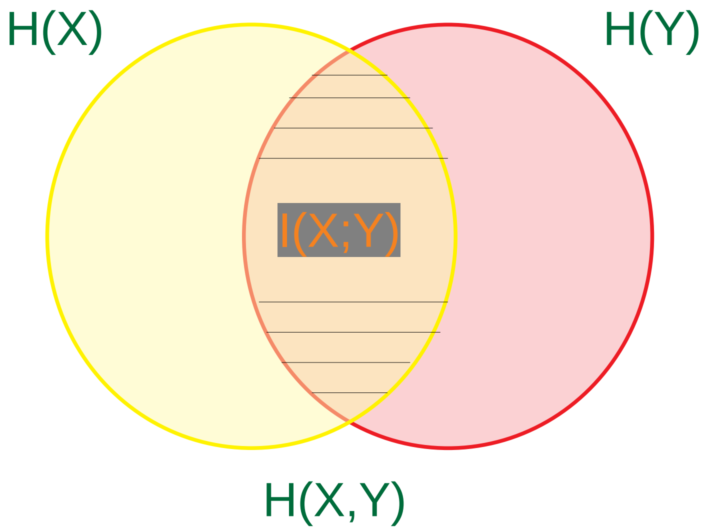
$$D_{KL}(P||Q) = -\sum_i P(i)\ \log \frac{Q(i)}{P(i)} = H(P,Q) - H(P)$$

Properties

$D_{KL}(P||Q) \geq 0$ Gibbs Inequality

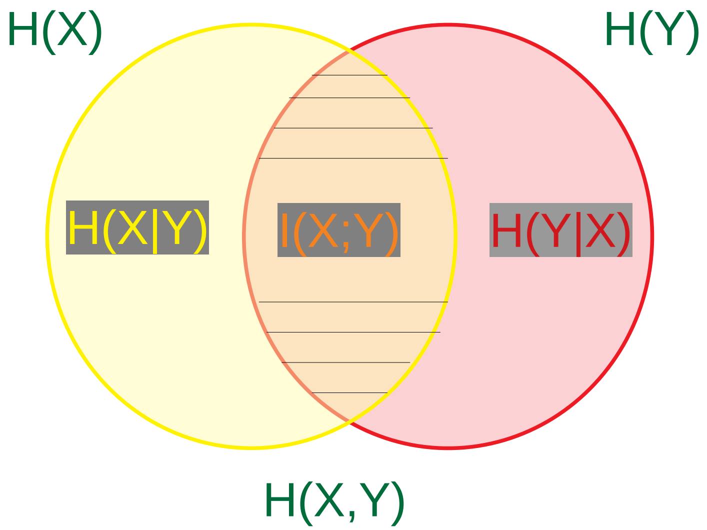
$D_{KL}(P||Q) \neq D_{KL}(Q||P)$ Asymmetric

Mutual Information (MI)



Mutual Information

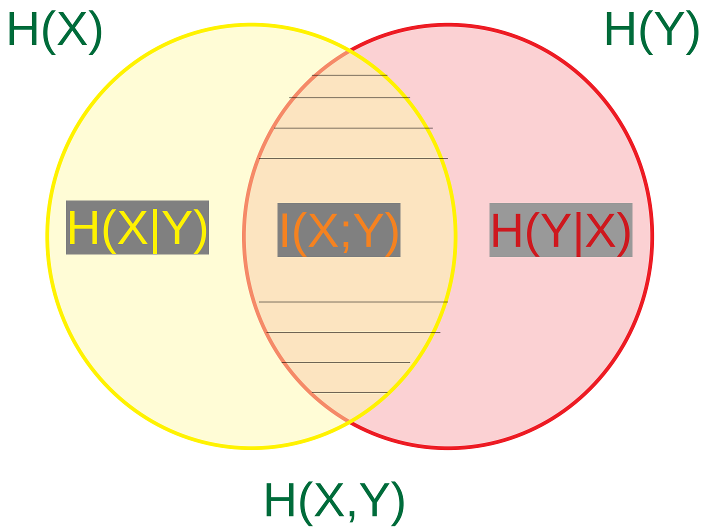
- Information X gives about Y



Mutual Information

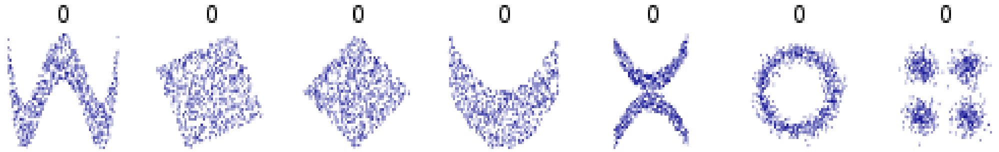
- Information X gives about Y, or *vice verse*

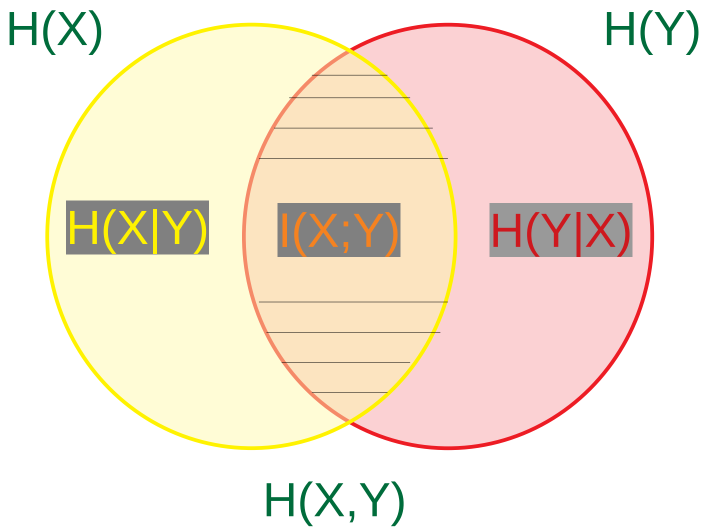
$$I(X;Y) = I(Y;X)$$



Mutual Information

- Information X gives about Y, or *vice verse*
- More general form of correlation

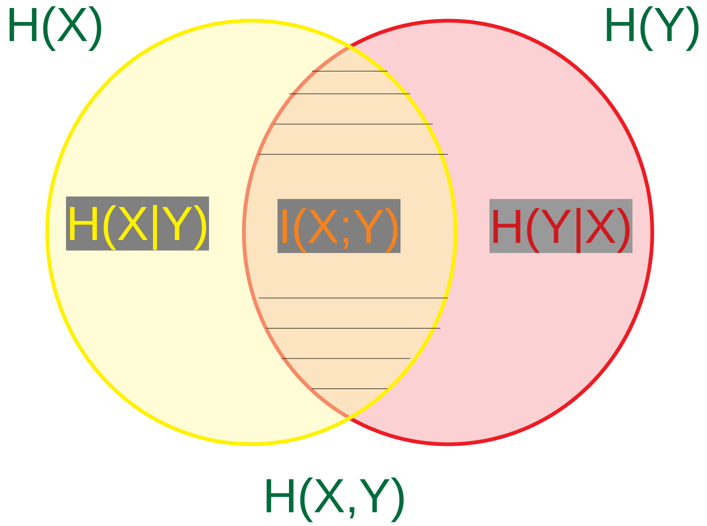




Mutual Information

- Information X gives about Y, or vice verse

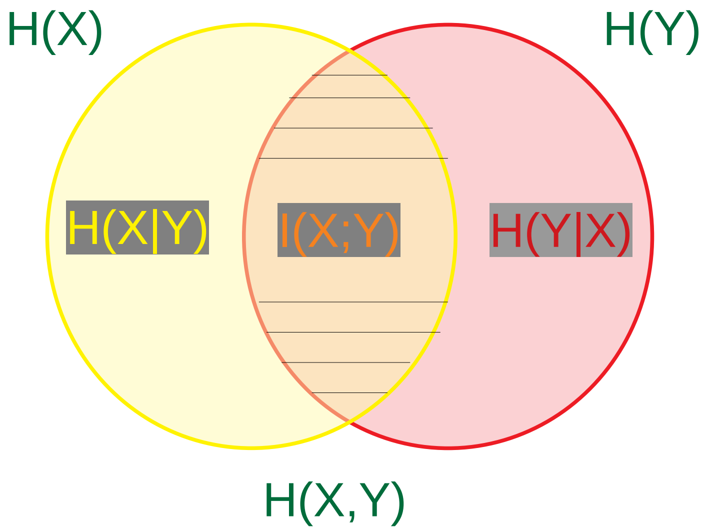
$$I(X;Y) = D_{KL}(P(X,Y)||P(X)P(Y))$$



Mutual Information

- Information X gives about Y, or vice verse

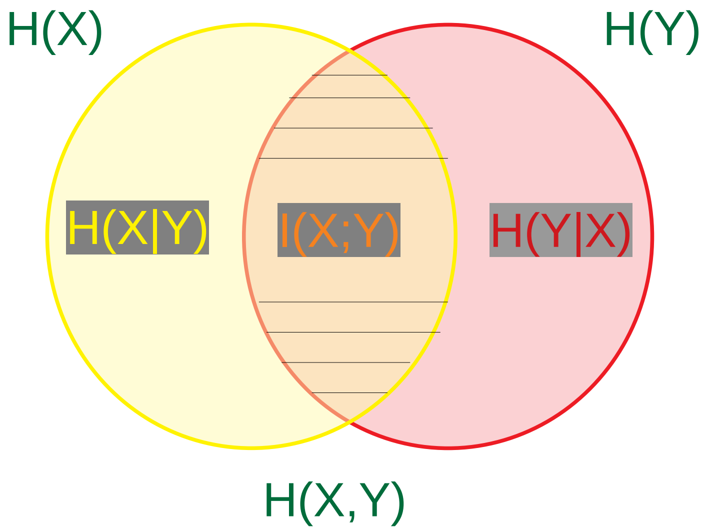
$$
\begin{aligned}
I(X;Y) &= D_{KL}(P(X,Y)||P(X)P(Y)) \\
&= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X)
\end{aligned}
$$



Mutual Information

- Information X gives about Y, or vice verse

$$
\begin{aligned}
I(X;Y) &= D_{KL}(P(X,Y)||P(X)P(Y)) \\
&= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X) \\
&= H(X,Y) - H(X|Y) - H(Y|X)
\end{aligned}
$$



Mutual Information – Properties

- Data Processing Inequality (DPI)
  - For Markov Chain: $X \rightarrow T1 \rightarrow T2 \rightarrow T3$

$$I(X;T1) \geq I(X;T2) \geq I(X;T3)$$

Mutual Information – Properties

- Data Processing Inequality (DPI)
  - For Markov Chain: $X \rightarrow T1 \rightarrow T2 \rightarrow T3$

$$I(X;T1) \geq I(X;T2) \geq I(X;T3)$$

- Transformation Invariance
  - For invertible functions *f* and *g*

$$I(X;Y) = I(\, f(X) \,;\, g(Y) \,)$$

Mutual Information – Estimation

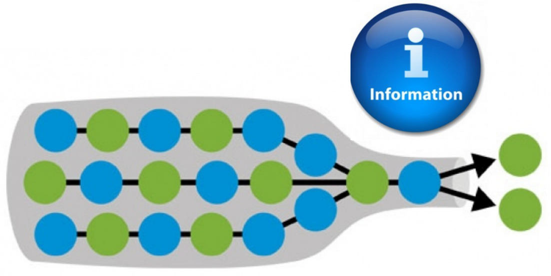
- Estimation (tricky, specially in high-dim space)
  - Ensemble Dependence Graph (Noshad 2018)
  - Kernel-based (Kolchinsky 2017)
  - K-NN-based (Kraskov 2004)
  - ...

Outlines

- Problem Statement
- Information Theory Review
- Information Bottleneck (IB)
- Opening the Black Box of DNNs via IB
- Criticisms
- Conclusions

Information Bottleneck

- Motivation
- Definition



The Information Bottleneck Method

Naftali Tishby
The Hebrew University
Jerusalem 91904, Israel
tishby@cs.huji.ac.il

Fernando C. Pereira
AT&T Labs – Research
Florham Park, NJ 07932
pereira@research.att.com

William Bialek
NEC Research Institute
Princeton, NJ 08540
bialek@research.nj.nec.com

Abstract

We define the relevant information in a signal $x \in X$ as being the information that this signal provides about another signal $y \in Y$. Examples include the information that face images provide about the names of the people portrayed, or the information that speech sounds provide about the words spoken. Understanding the signal x requires more than just predicting y, it also requires specifying which features of X play a role in the prediction. We formalize the problem as that of finding a short code for X that preserves the maximum information about Y. That is, we squeeze the information that X provides about Y through a 'bottleneck' formed by a limited set of codewords $\tilde{X}$. This constrained optimization problem can be seen as a generalization of rate distortion theory in which the distortion measure $d(x, \tilde{x})$ emerges from the joint statistics of X and Y. The approach yields an exact set of self-consistent equations for the coding rules $X \to \tilde{X}$ and $\tilde{X} \to Y$. Solutions to these equations can be found by a convergent re-estimation method that generalizes the Blahut–Arimoto algorithm. Our variational principle provides a surprisingly rich framework for discussing a variety of problems in signal processing and learning, as will be described in detail elsewhere.

1999

Rate-Distortion Theory

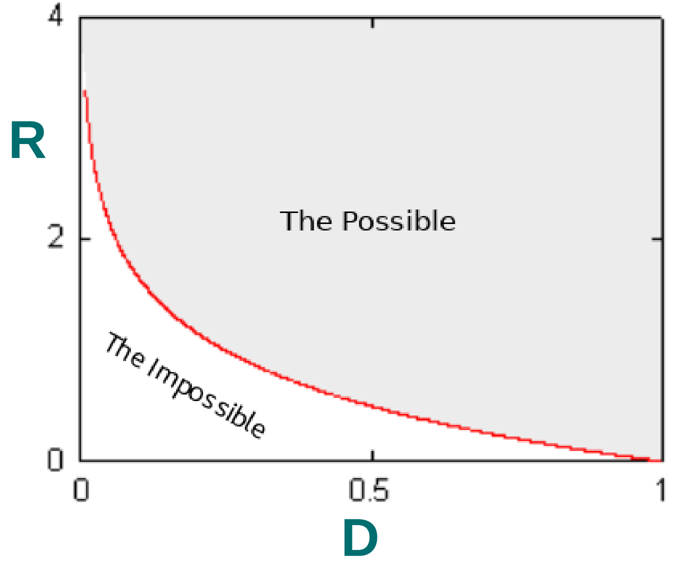
- GOAL:
  - Find Minimal **R**ate

    subject to: **D**istortion $\leq D_{Max}$

Rate-Distortion Theory

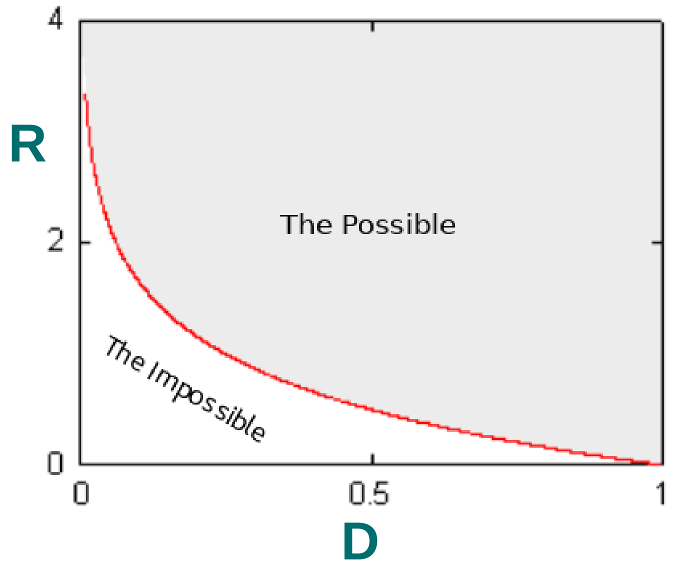
- GOAL:
  - Find Minimal **R**ate
    subject to: **D**istortion ≤ D_{Max}

Trade-off: Rate vs **Distortion**



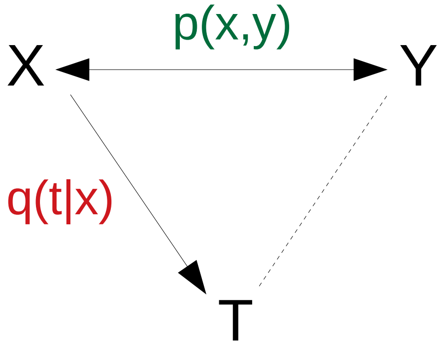
Rate-Distortion Theory

- Distortion Measure
  - Purely Mathematical
  - Relevant/Irrelevant info
    - Perceptual
    - Side-information (Wyner 1975)



Information Bottleneck (IB)

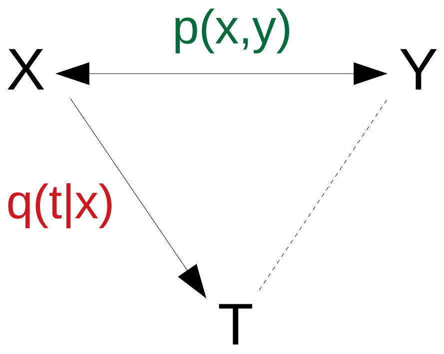
- Idea
  - Shannon Information + Learning



X: observation
Y: variable of interest
T: representation of X

Information Bottleneck (IB)

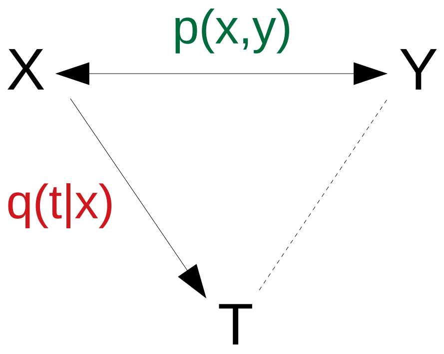
- How
  - Compress X into T subject to ...
  - Maintain relevant info about Y



X: observation
Y: variable of interest
T: representation of X

Information Bottleneck (IB) – Interpretation

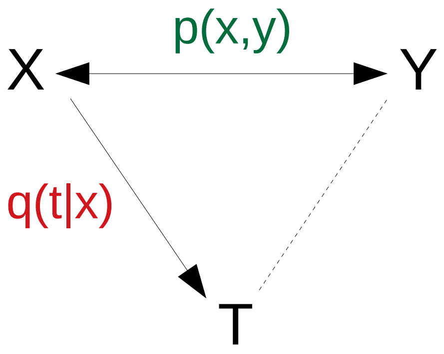
- Statistics
  - Generalised minimal sufficient statistics for Y
  - $Y \perp\!\!\!\perp X|T \iff I(X;Y) = I(T;Y)$



X: observation
Y: variable of interest
T: representation of X

Information Bottleneck (IB) – Interpretation

- Statistics
  - Generalised minimal sufficient statistics for Y
  - $Y \perp\!\!\!\perp X|T \iff I(X;Y) = I(T;Y)$
- Machine Learning
  - Kind of clustering or VQ

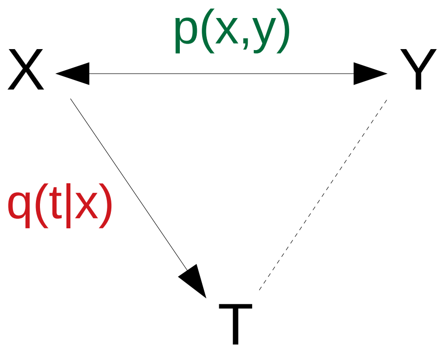


X: observation
Y: variable of interest
T: representation of X

IB – Objective Function

$$\min_{q(t|x)} \{I(T;X) \;-\; \beta I(T;Y)\}$$

Lagrange
multiplier



X: observation
Y: variable of interest
T: representation of X

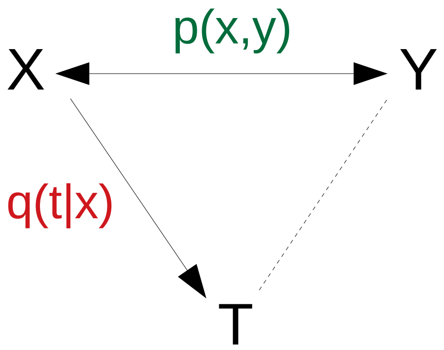
IB – Objective Function

Minimality/
Compression/
Complexity

Fidelity/
Sufficiency/
Accuracy

$$\min_{q(t|x)} \{ \boxed{I(T;X)} - \beta \boxed{I(T;Y)} \}$$

Lagrange
multiplier



X: observation
Y: variable of interest
T: representation of X

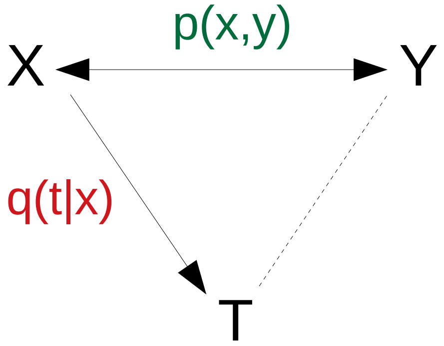
IB – Objective Function

Minimality/ Compression/ Complexity

Fidelity/ Sufficiency/ Accuracy

$$\min_{q(t|x)} \{\boxed{I(T;X)} - \beta \boxed{I(T;Y)}\}$$

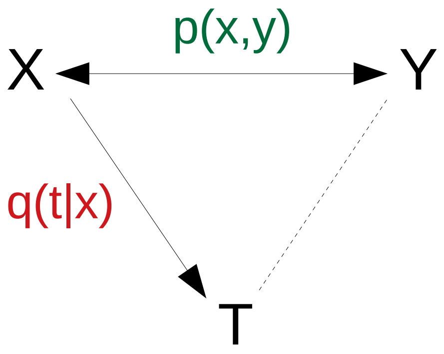
Trade-off parameter



X: observation
Y: variable of interest
T: representation of X

IB – Objective Function

$$\min_{q(t|x)} \{\overbrace{I(T;X)}^{\text{Minimality/ Compression/ Complexity}} - \beta \overbrace{I(T;Y)}^{\text{Fidelity/ Sufficiency/ Accuracy}}\}$$

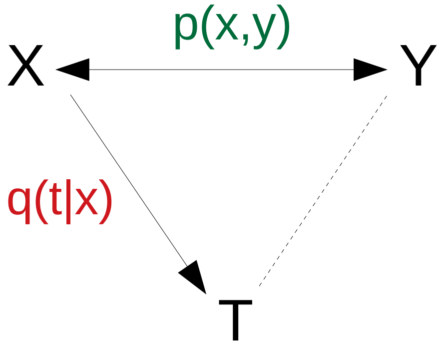


X: observation
Y: variable of interest
T: representation of X

IDEALLY
- I(T;X) ↔ as LOW as possible
- I(T;Y) ↔ as HIGH as possible

IB – Objective Function

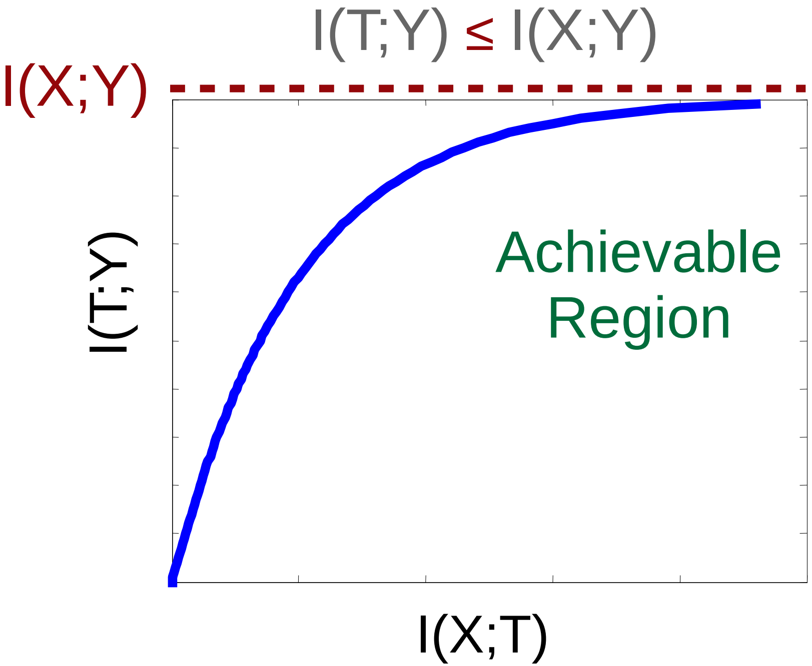
$$\min_{q(t|x)} \left\{ \overbrace{I(T;X)}^{\text{Minimality/ Compression/ Complexity}} - \beta \overbrace{I(T;Y)}^{\text{Fidelity/ Sufficiency/ Accuracy}} \right\}$$



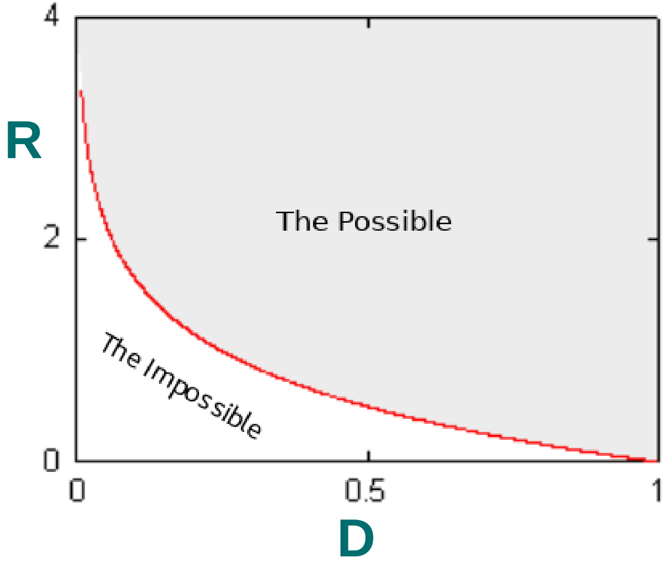
X: observation
Y: variable of interest
T: representation of X

$$X \xrightarrow{\text{Encoding}} T \xrightarrow{\text{Decoding}} Y$$

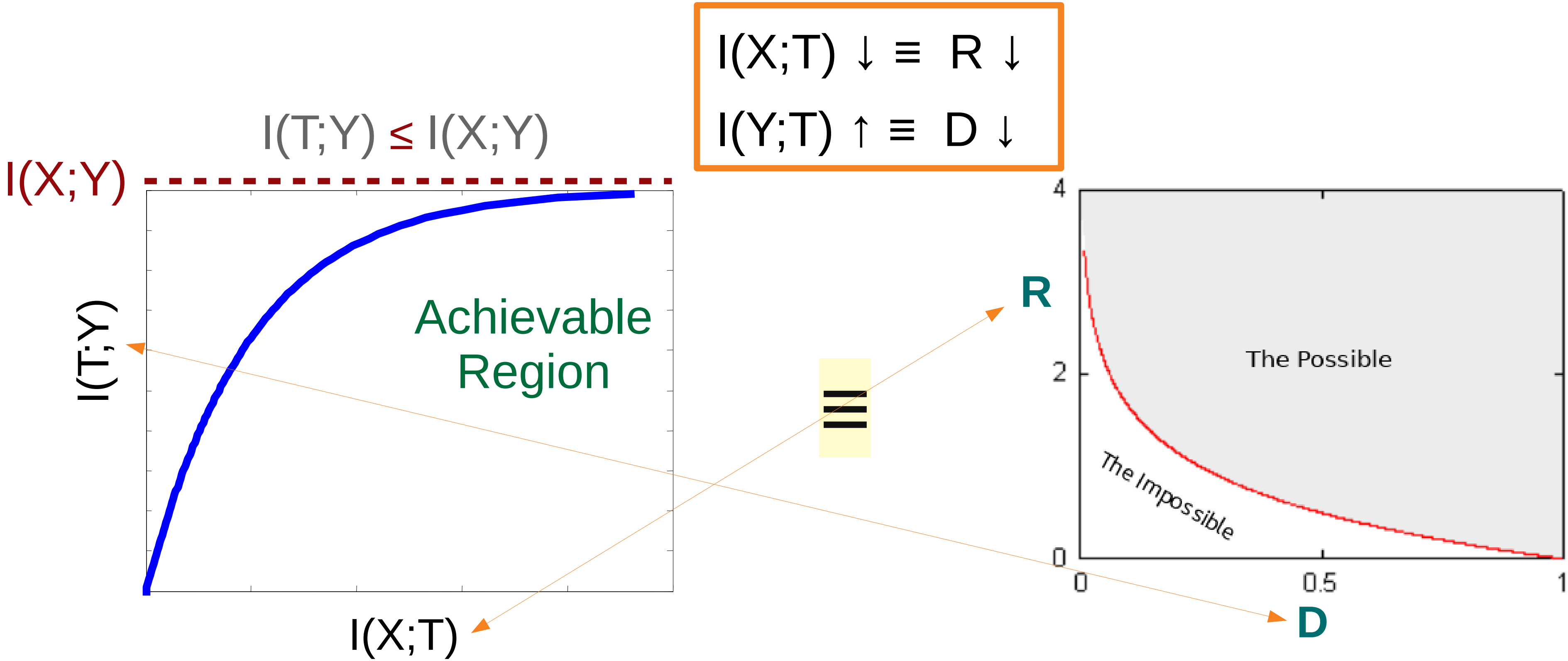
Accuracy-Compression Tradeoff



≡



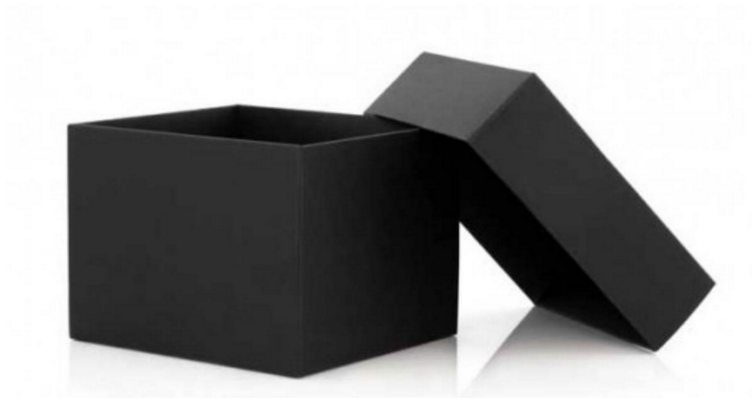
Accuracy-Compression Tradeoff

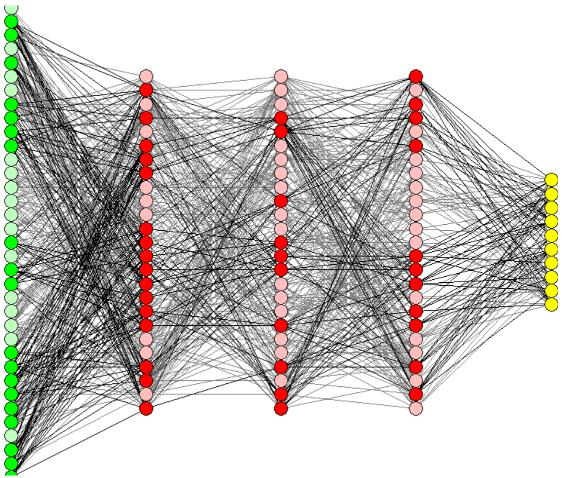


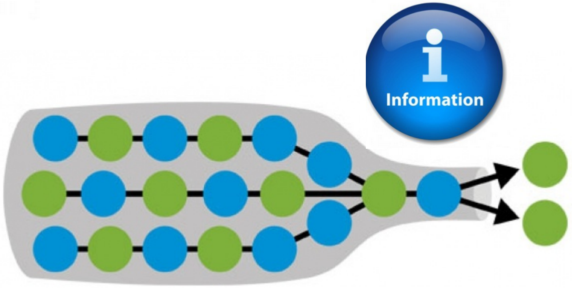
Outlines

- Problem Statement
- Information Theory Review
- Information Bottleneck (IB)
- Opening the Black Box of DNNs via IB
- Criticisms
- Conclusions

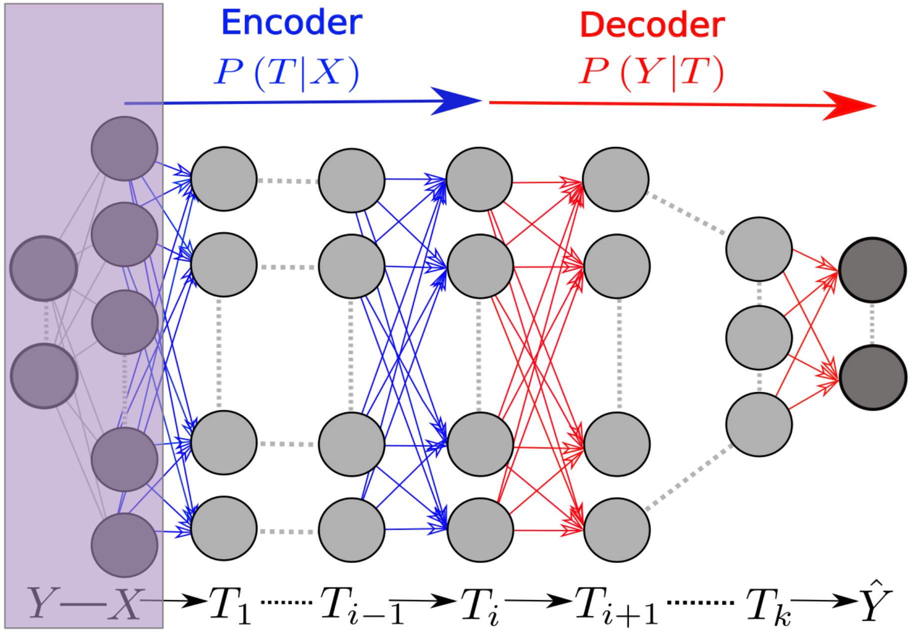
Opening the Black Box of DNNs via Information Bottleneck







NN as a Markov Chain

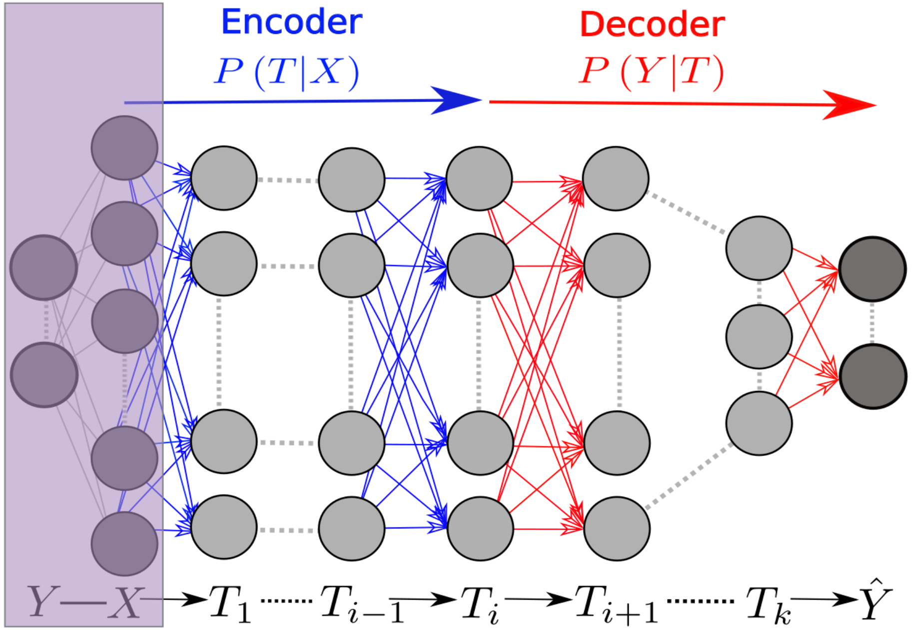


$$\text{Markov Chain}: Y \leftrightarrow X \rightarrow T \rightarrow \hat{Y}$$

$$\text{Data}: \{(x_i, y_i)\}_{i=1}^{n} \sim p(x, y)$$

NN as a Markov Chain

- Each layer characterised by
  - Encoder
  - Decoder

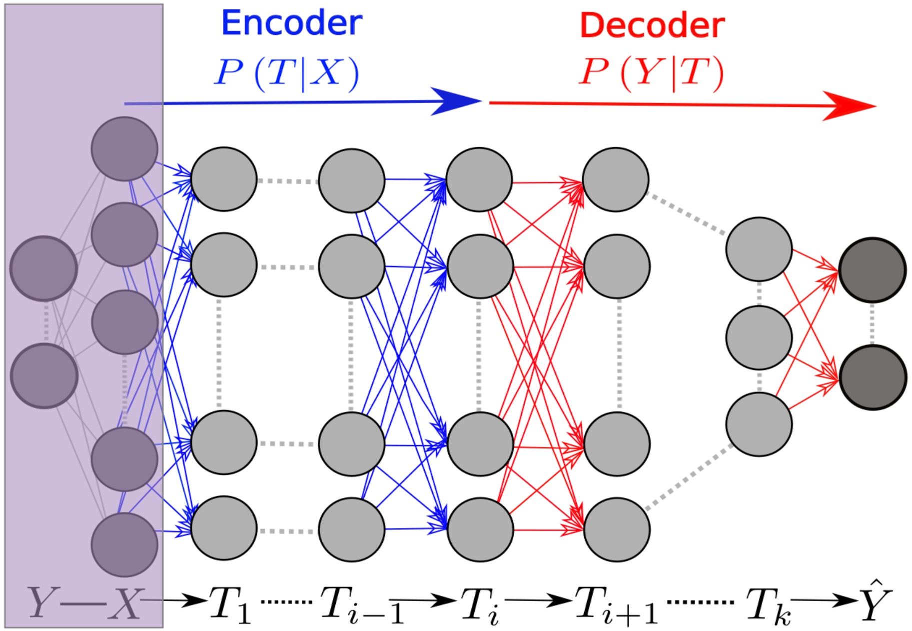


$$\text{Markov Chain} : Y \leftrightarrow X \to T \to \hat{Y}$$

$$\text{Data} : \{(x_i, y_i)\}_{i=1}^{n} \sim p(x, y)$$

NN as a Markov Chain

- Each layer characterised by
  - Encoder
  - Decoder
- Across layers, complexity shifts from De to En

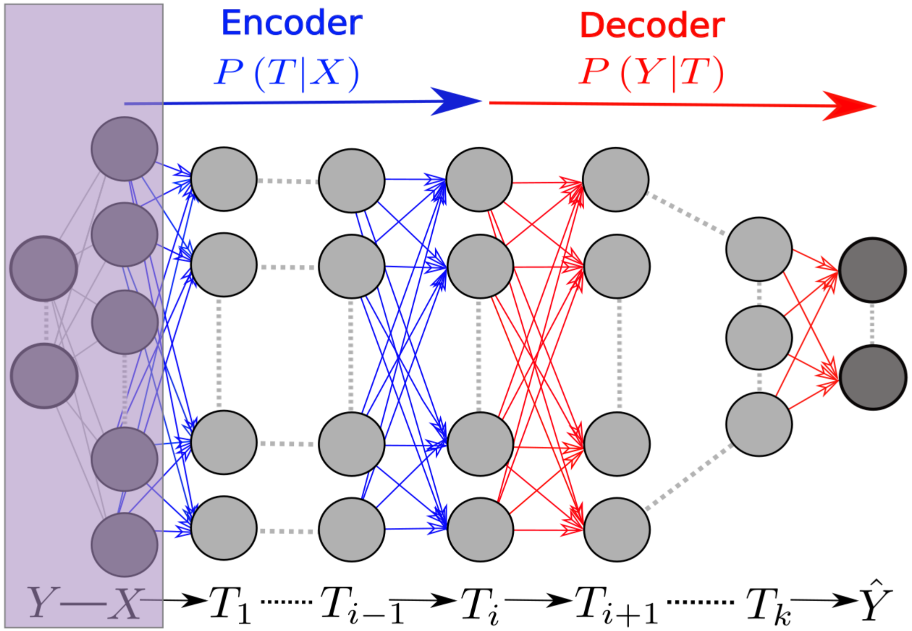


$$\text{Markov Chain} : Y \leftrightarrow X \rightarrow T \rightarrow \hat{Y}$$

$$\text{Data} : \{(x_i, y_i)\}_{i=1}^{n} \sim p(x, y)$$

NN as a Markov Chain

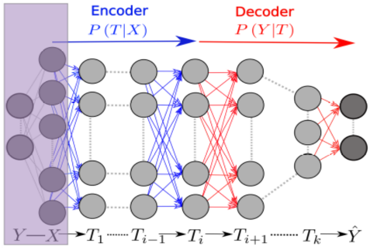
- Each layer characterised by
  - Encoder
  - Decoder
- Across layers, complexity shifts from De to En
- **GOAL**:
  - Successive Refinement of relevant information

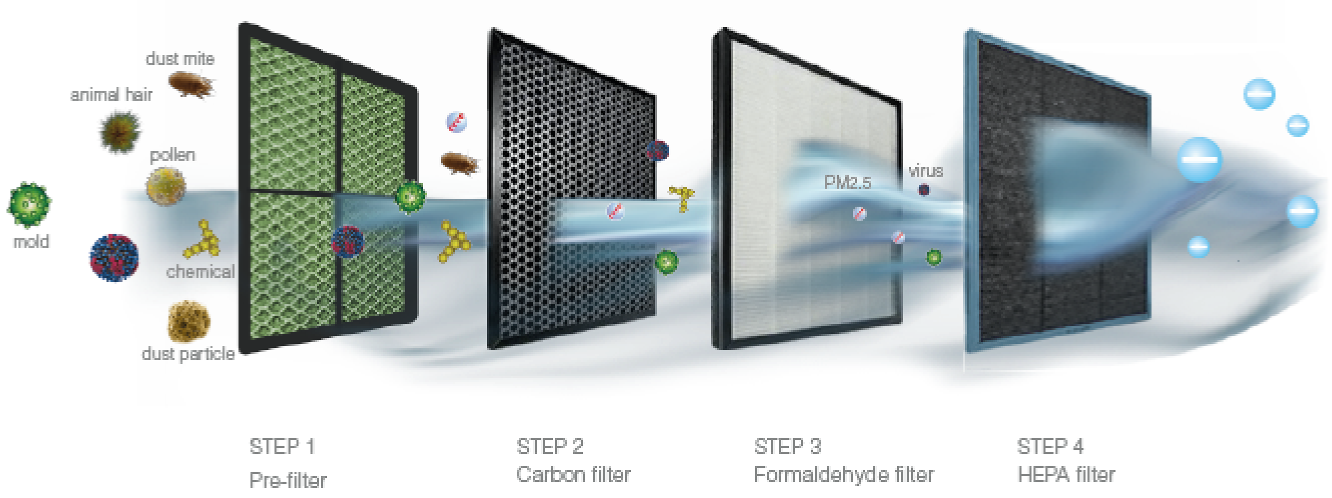


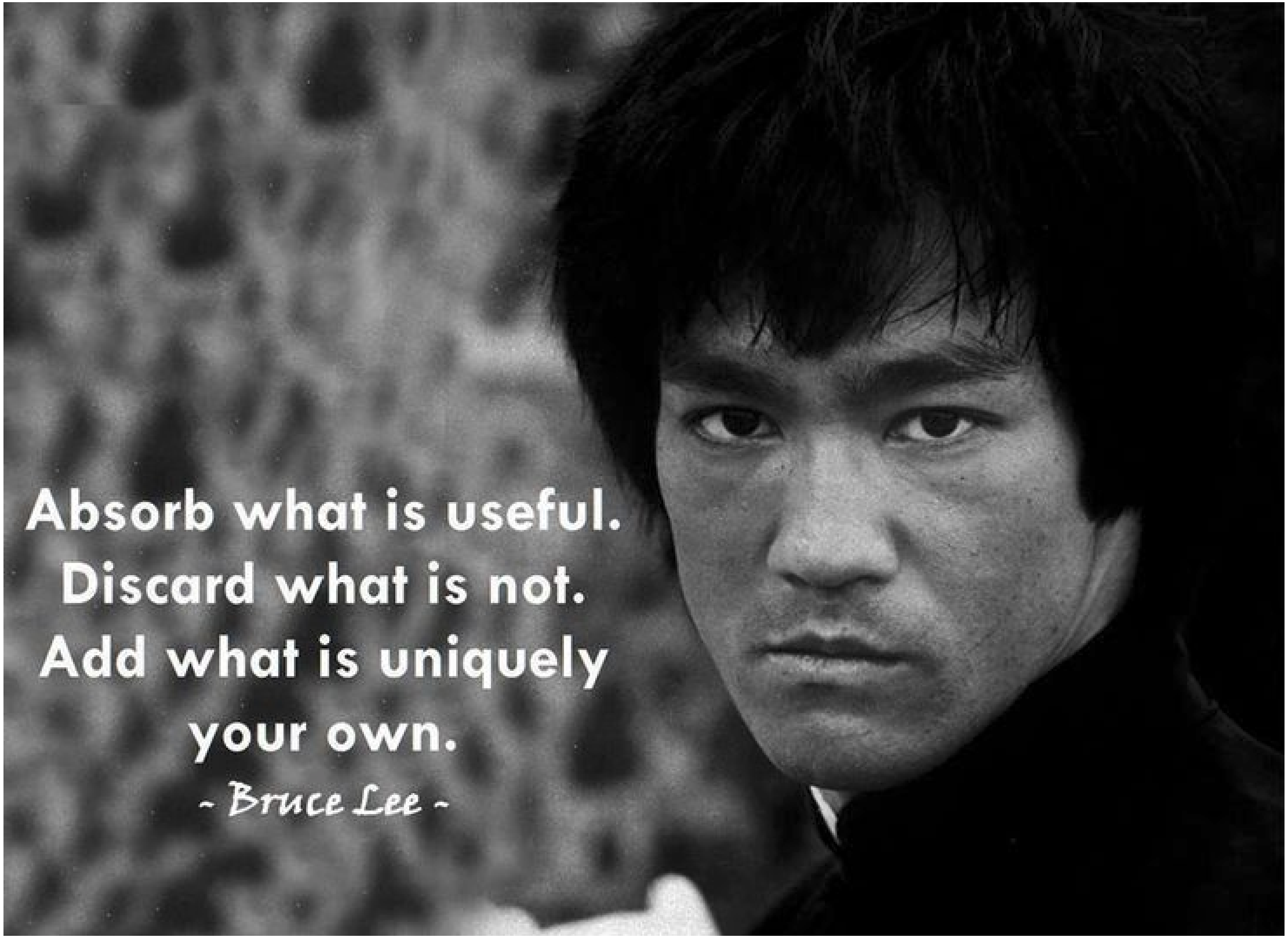
$$\text{Markov Chain} : Y \leftrightarrow X \rightarrow T \rightarrow \hat{Y}$$

$$\text{Data} : \{(x_i, y_i)\}_{i=1}^{n} \sim p(x, y)$$

Successive Refinement of Relevant Info

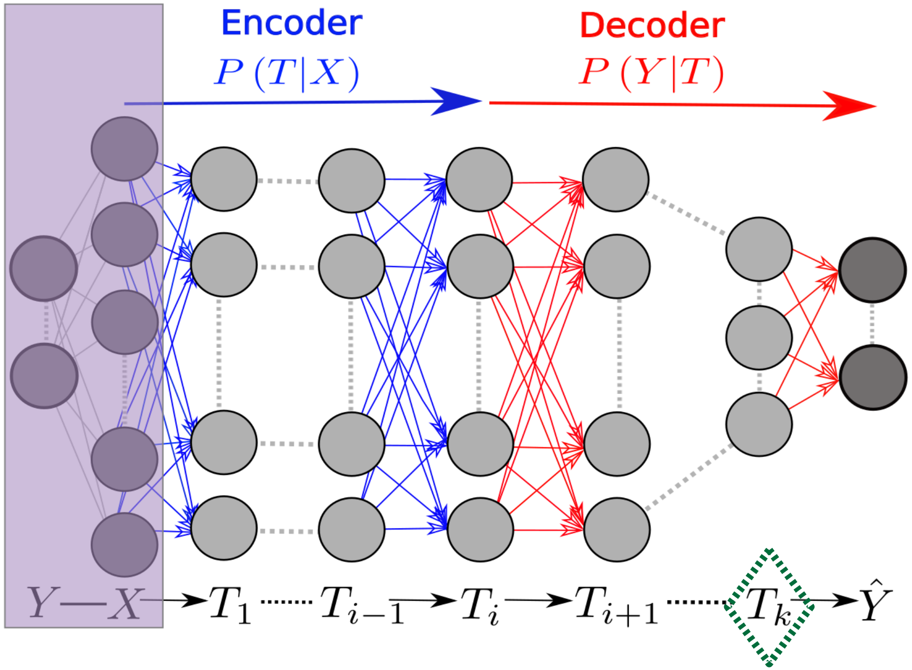





Absorb what is useful.
Discard what is not.
Add what is uniquely
your own.
~ Bruce Lee ~

NN as a Markov Chain

- Ideally
  - Compression: $I(X;T_k)$ ↓
  - Accuracy: $I(Y;T_k)$ ↑

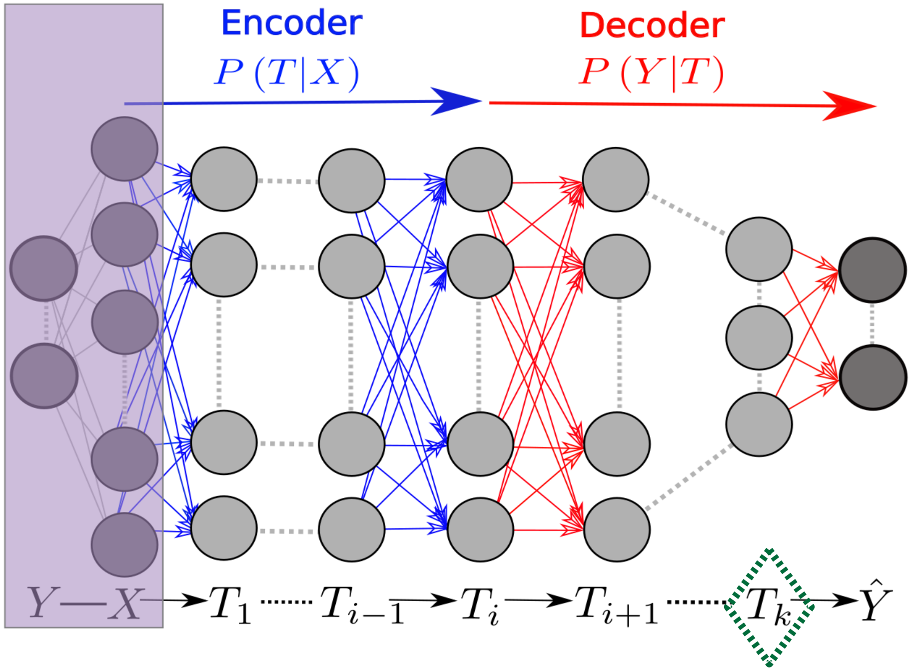


$$H(X) \geq I(X;T_i) \geq I(X;T_{i+1}) \geq \ldots$$

$$I(X;Y) \geq I(T_i;Y) \geq I(T_{i+1};Y) \geq \ldots$$

NN as a Markov Chain

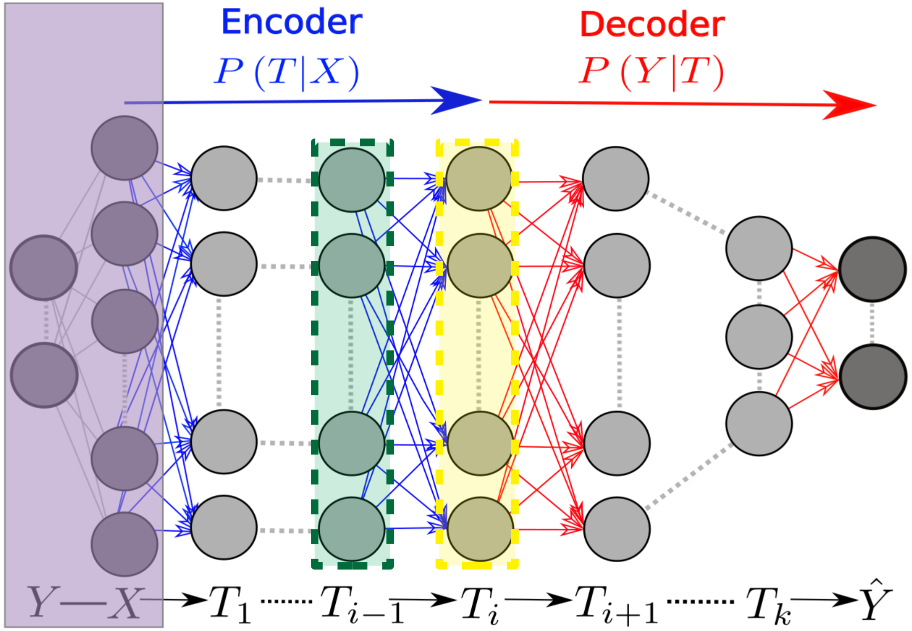
- Theorem (T_k: last hidden layer)
  - $I(X;T_k) \leftrightarrow$ sample complexity
  - $I(Y;T_k) \leftrightarrow$ generalisation error

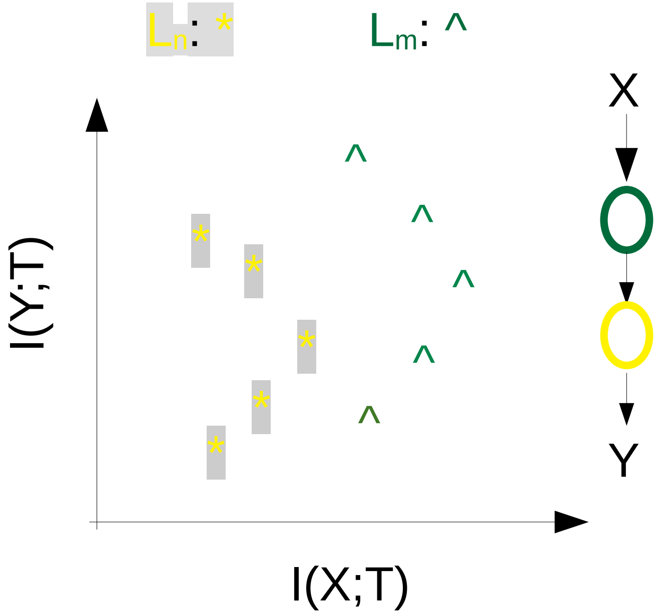


$$H(X) \geq I(X;T_i) \geq I(X;T_{i+1}) \geq \ldots$$

$$I(X;Y) \geq I(T_i;Y) \geq I(T_{i+1};Y) \geq \ldots$$

Information Plane



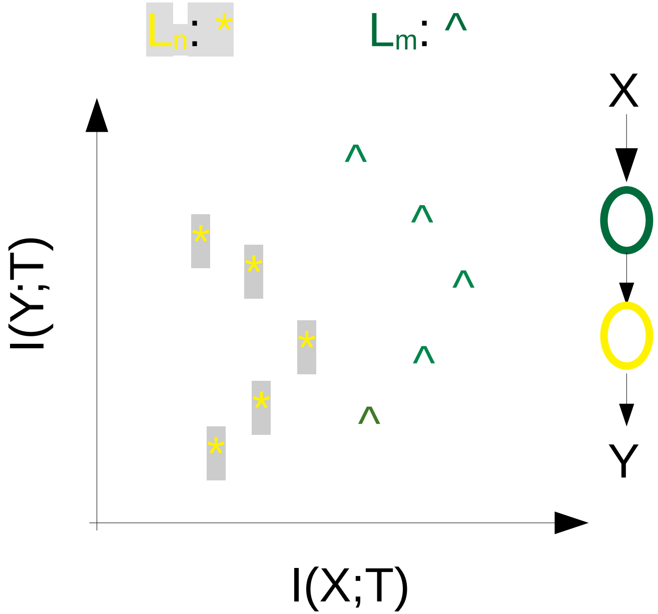


$I_X = I(X;T)$

$I_Y = I(Y;T)$

Information Plane

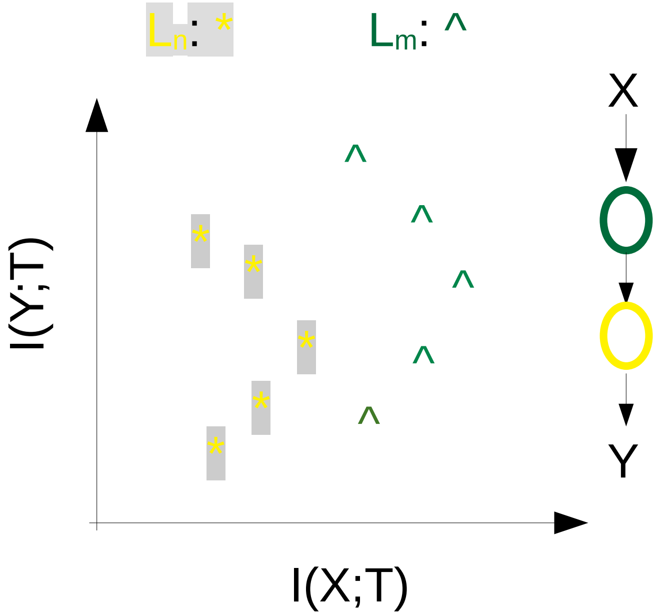
- GOAL
  - Study the dynamics of learning



$$I_X = I(X;T)$$
$$I_Y = I(Y;T)$$

Information Plane

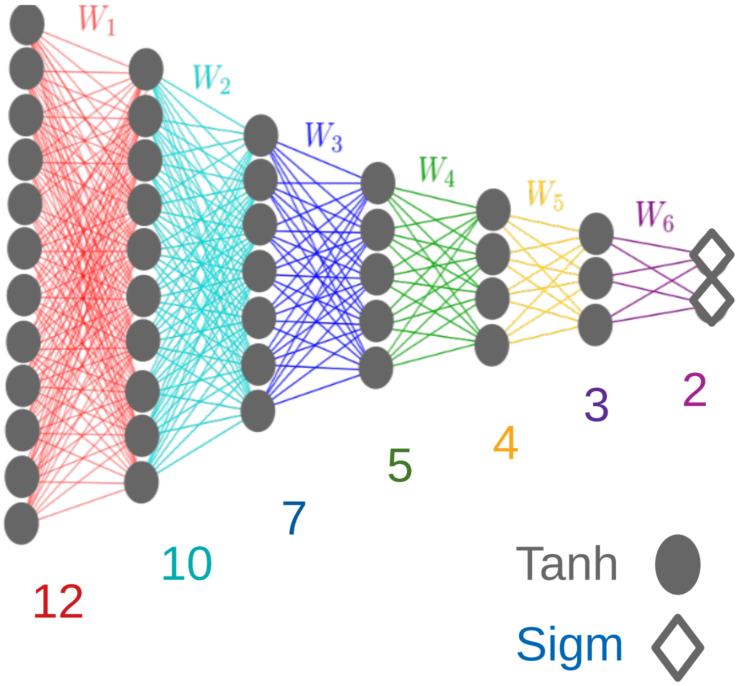
- GOAL
  - Study the dynamics of learning
- How
  - Estimate I_X & I_Y for all layers, all epochs



$$I_X = I(X;T)$$
$$I_Y = I(Y;T)$$

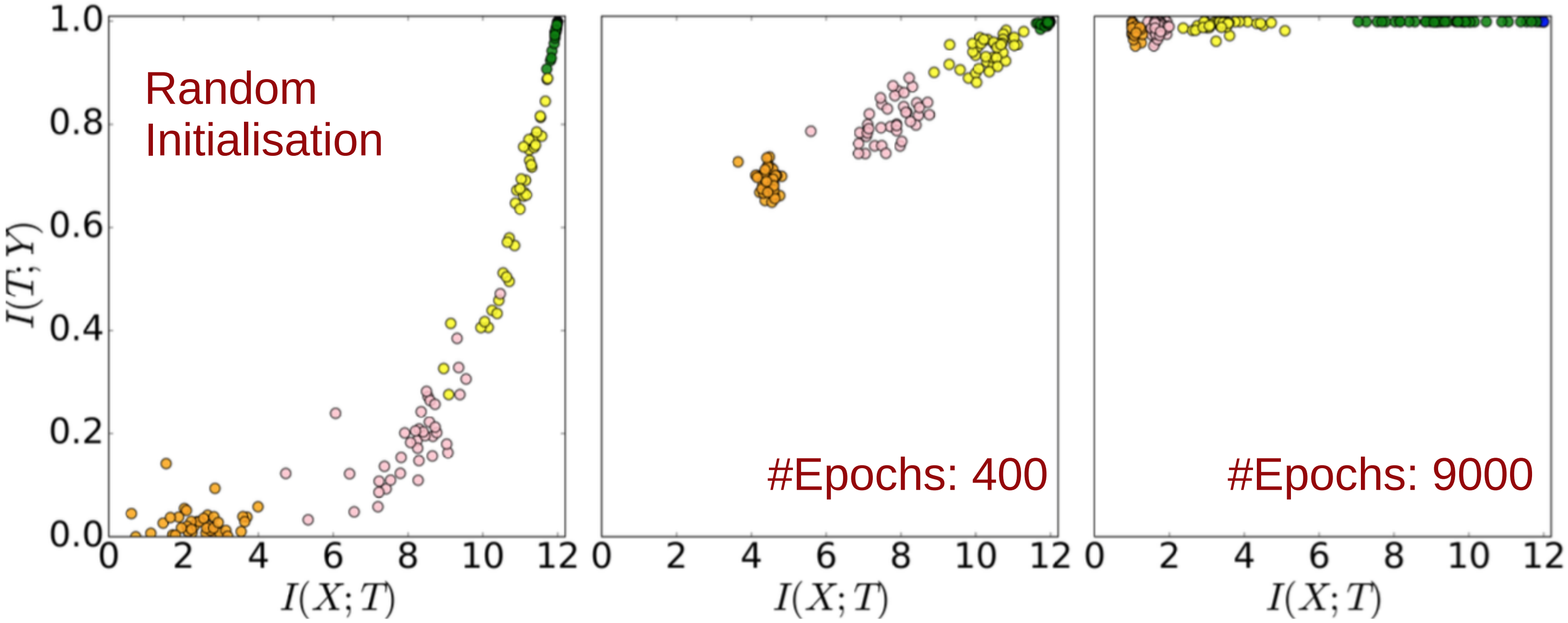
Network Architecture

- Fully-connected FFNN
- Training data:
  - 4096 samples
  - Batch size: 64
- 10k epochs
- 50 initialisations



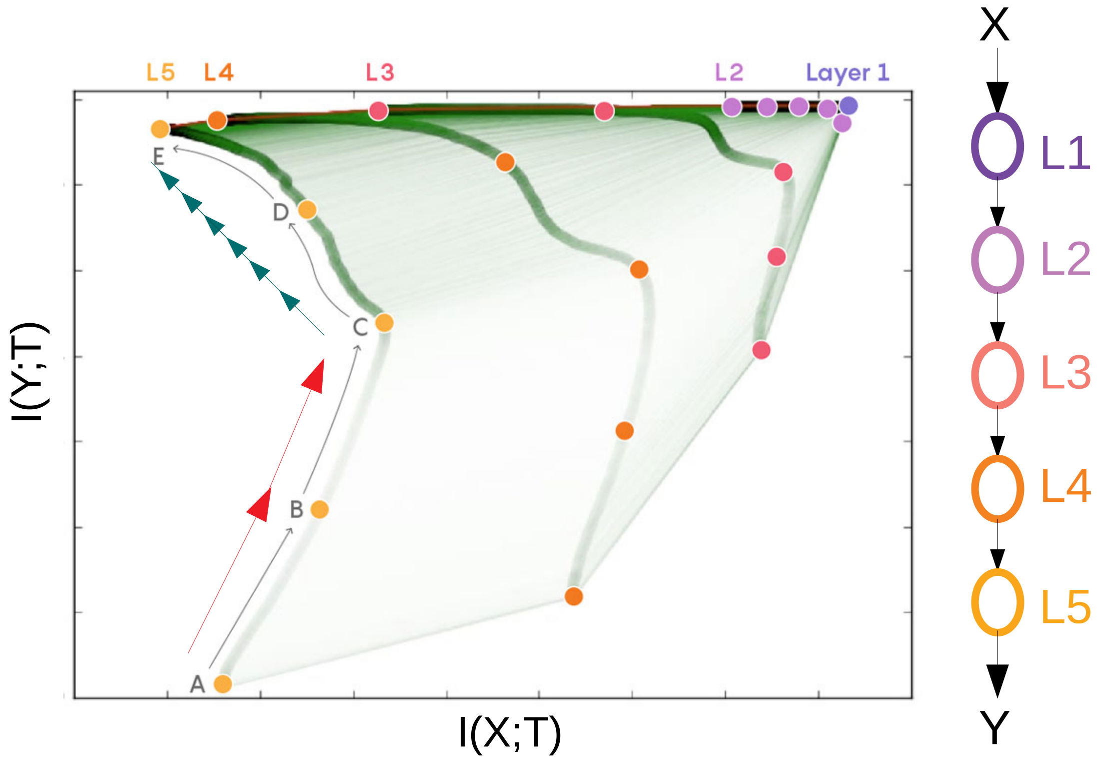
Mutual Information at Different Epochs

Animation



Each circle represents an initialisation (50)

Dynamics of Learning – Info Path



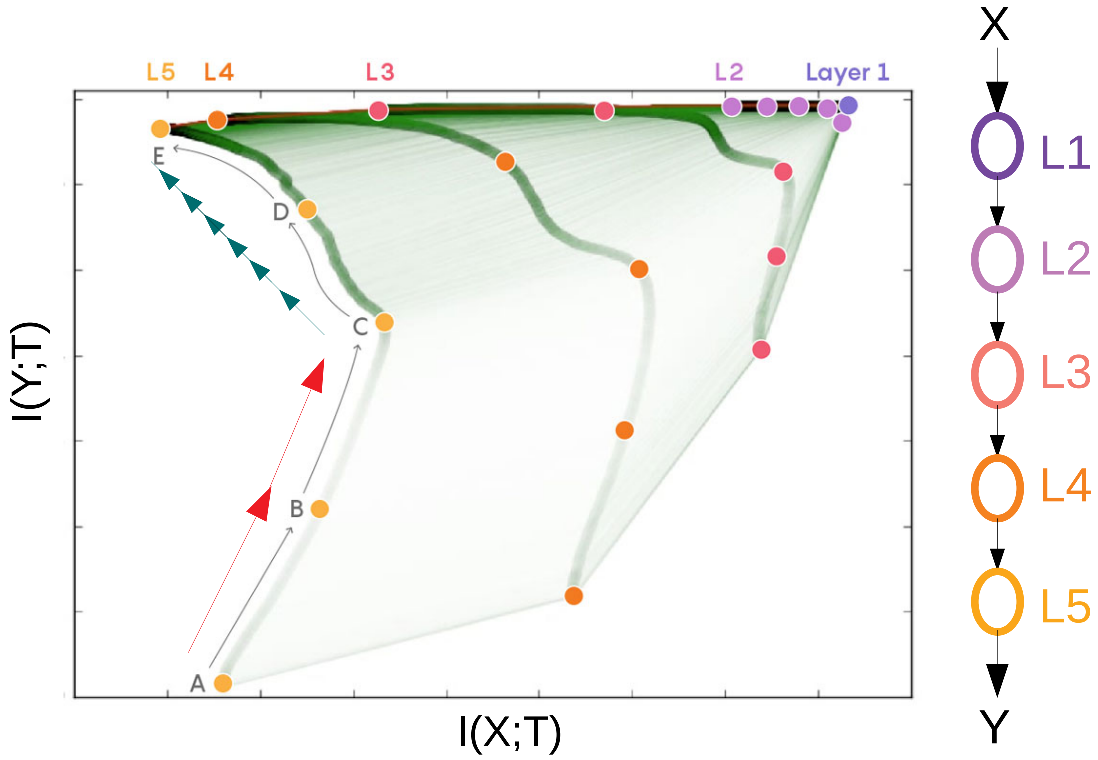
Similar path for all initialisations → Average

Dynamics of Learning – Info Path

- Two distinct Phases

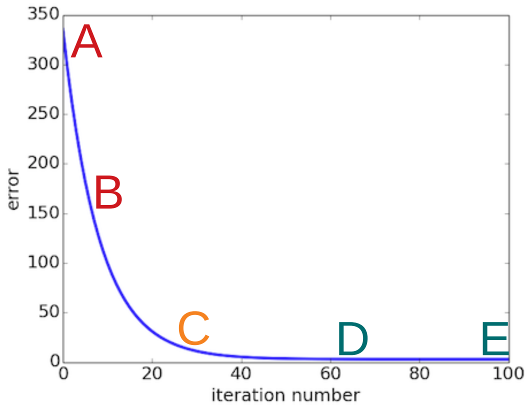
  (1) A → C

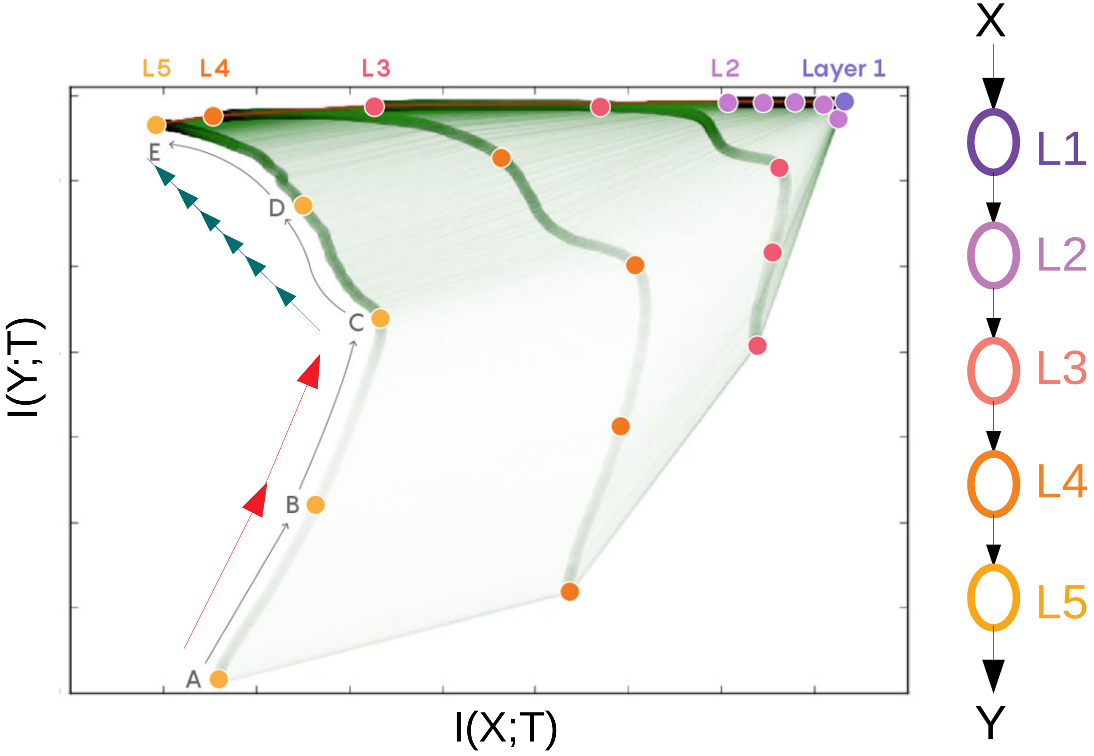
  (2) C → E



Dynamics of Learning – Info Path

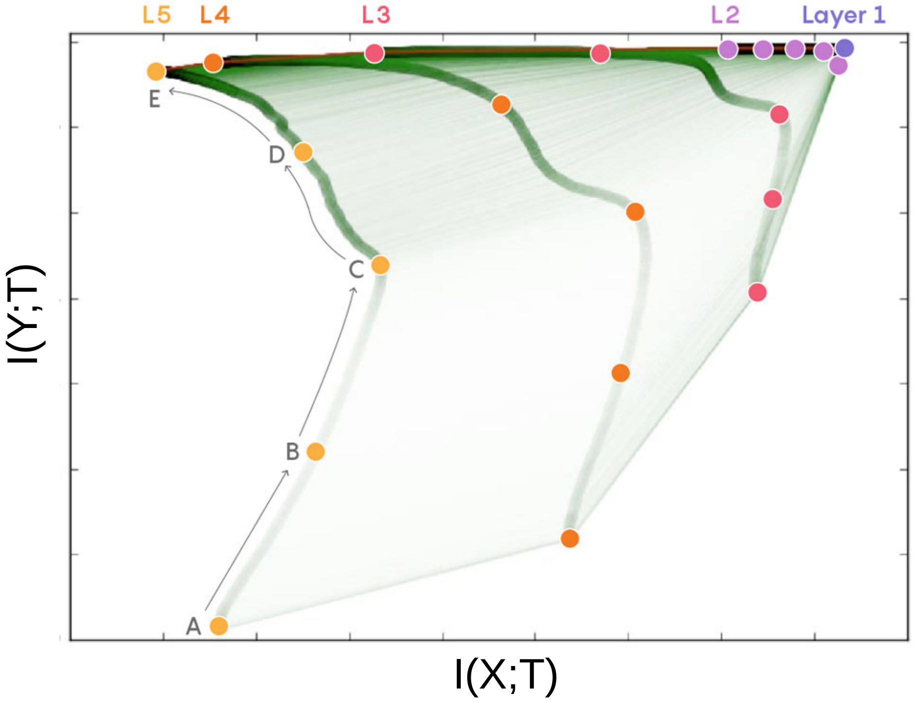
- Two distinct Phases
  - (1) A → C
  - (2) C → E

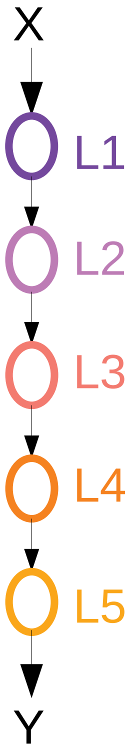




Information Path – First Phase

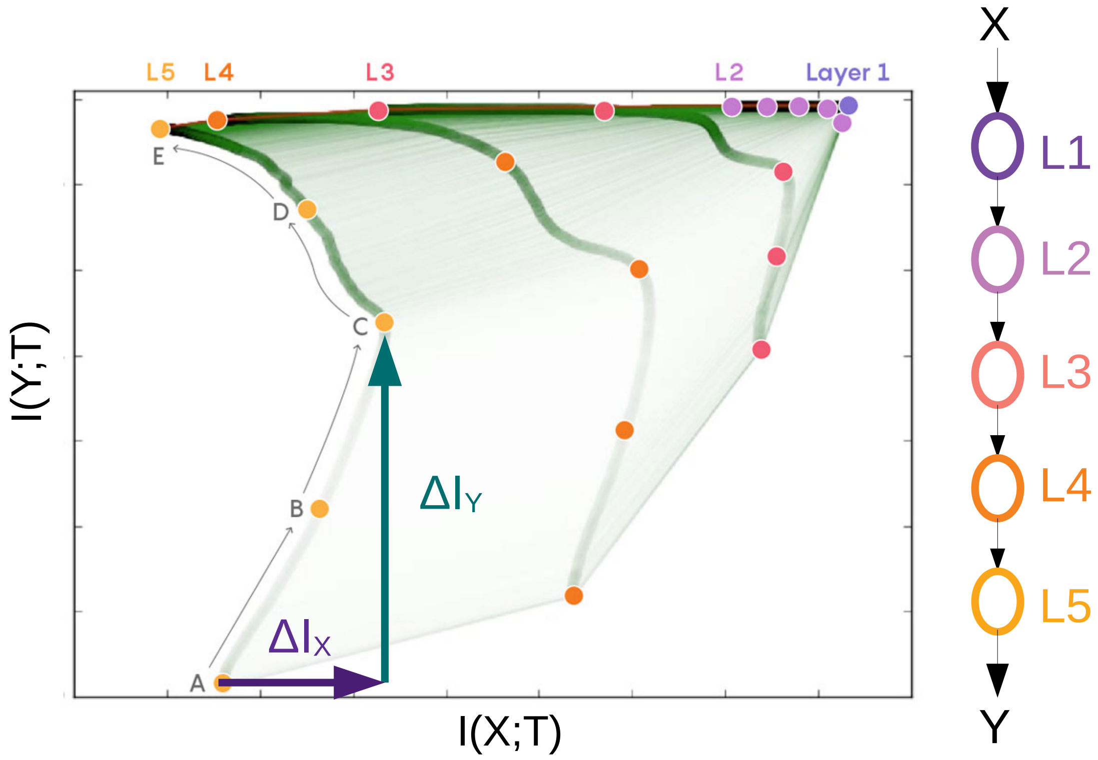
- A → C
  - Fitting the labels
  - Empirical Risk Min
  - Fast





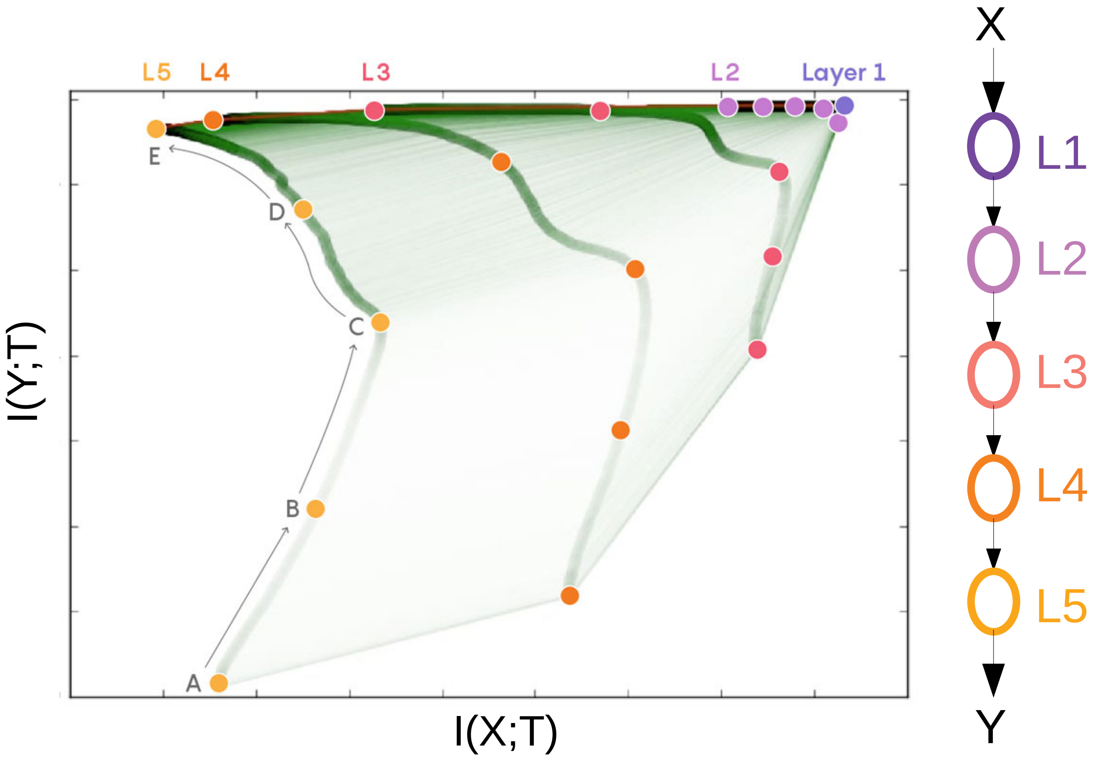
Information Path – First Phase

- A → C
  - Fitting the labels
  - Empirical Risk Min
  - Fast
  - $\Delta I_Y > 0$ and $\Delta I_X > 0$



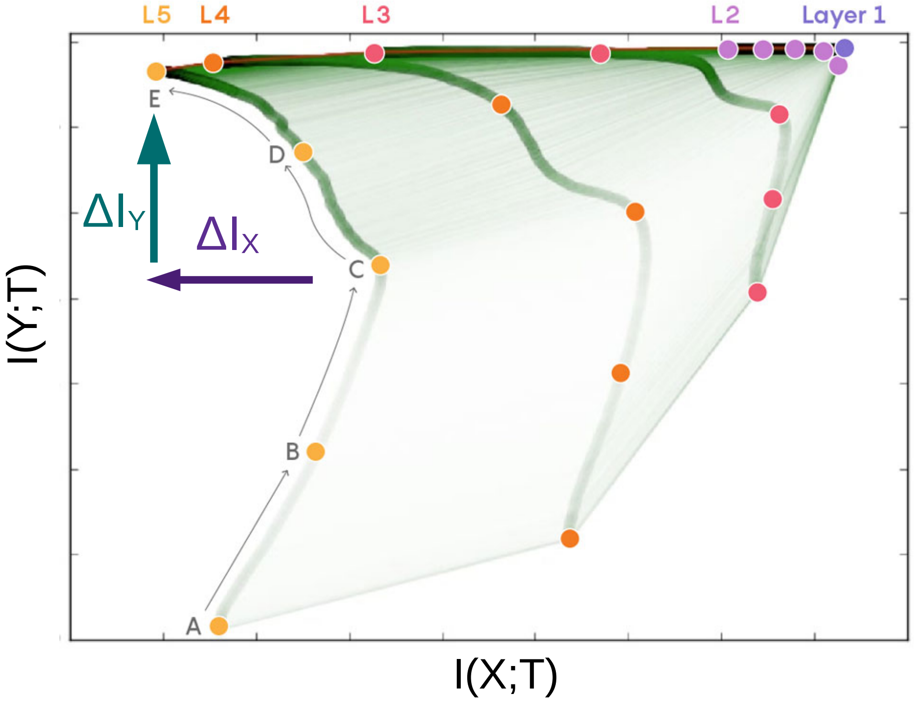
Information Path – Second Phase

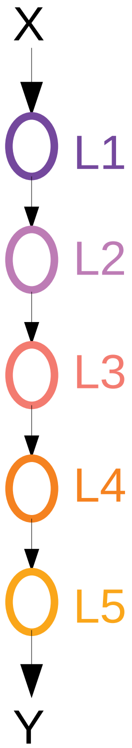
- C → E
  - Empirical risk ≈ constant
  - Slow



Information Path – Second Phase

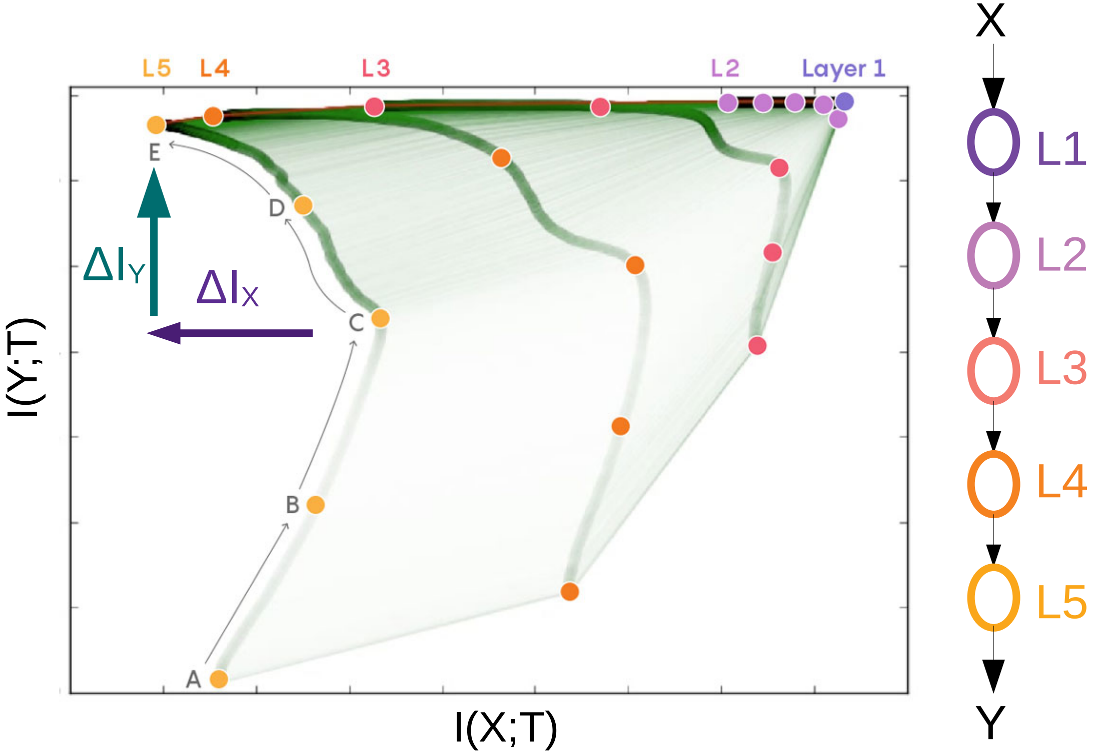
- C → E
  - Empirical risk ≈ constant
  - Slow
  - $\Delta I_X < 0$ (and $\Delta I_Y \geq 0$)





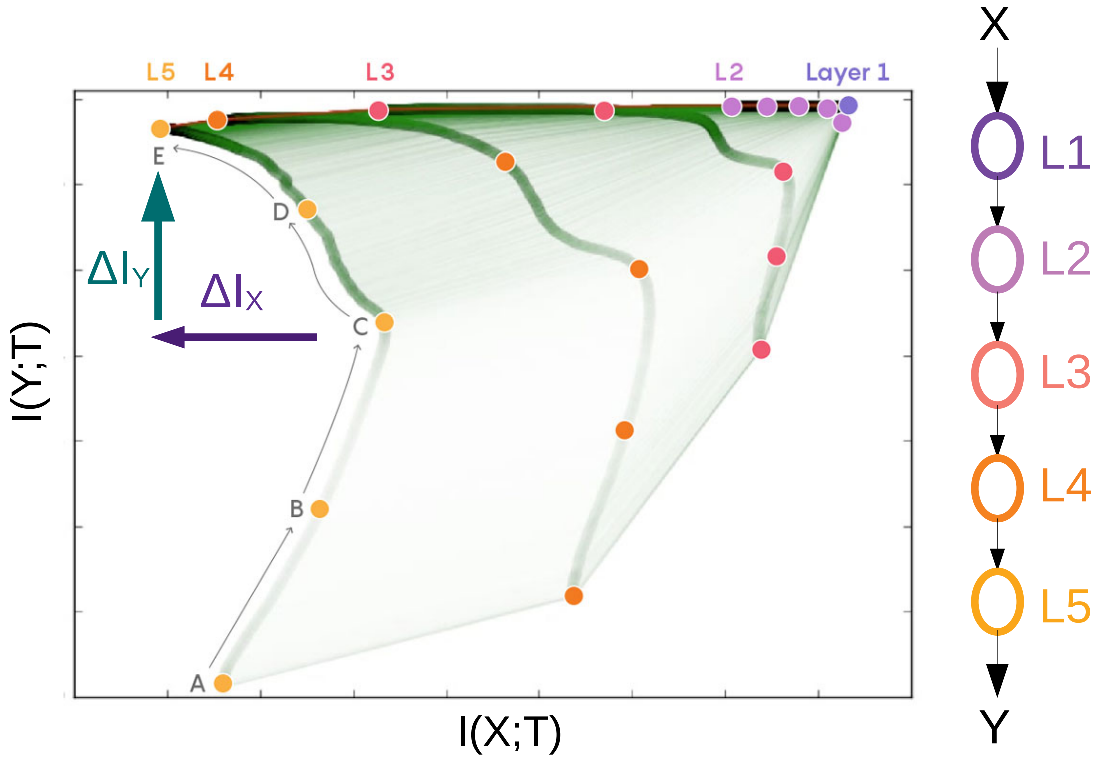
Information Path – Second Phase

- C → E
  - Empirical risk ≈ constant
  - Slow
  - $\Delta I_X < 0$ (and $\Delta I_Y \geq 0$)
    - Compression
    - Forget irrelevant info



Information Path – Second Phase

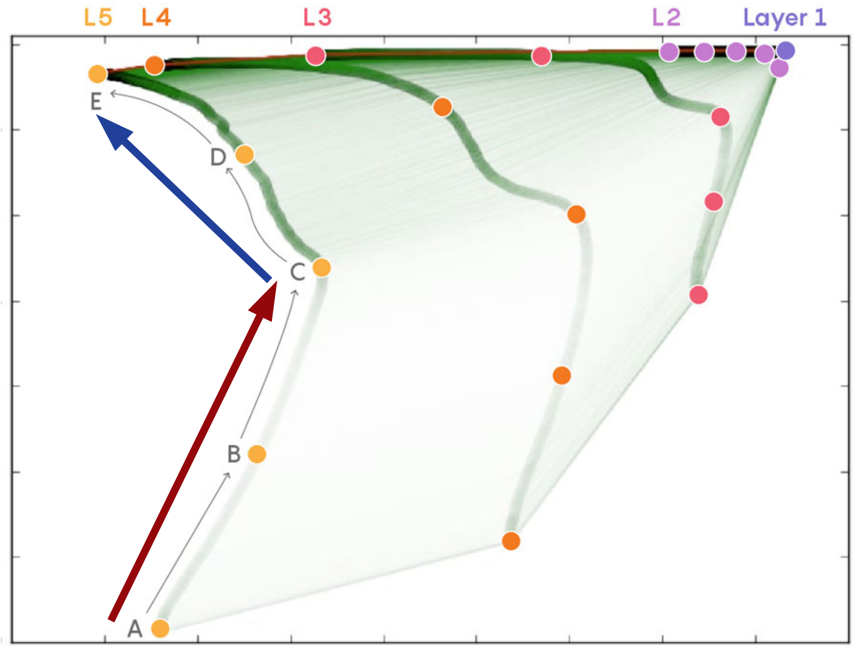
- C → E
  - Empirical risk ≈ constant
  - Slow
  - $\Delta I_X < 0$ (and $\Delta I_Y \geq 0$)
    - Compression
    - Forget irrelevant info
  - Responsible for generalisation



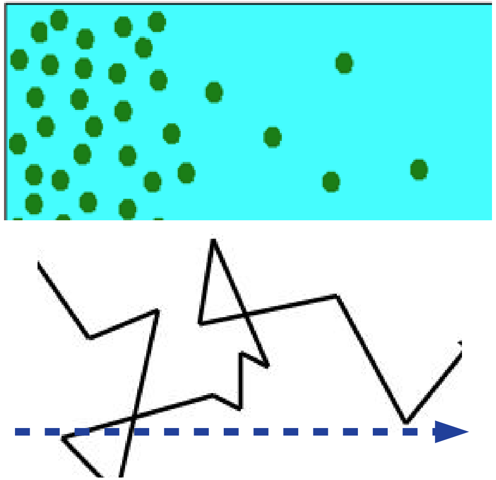
Drift vs Diffusion

Diffusion: C → E

Drift: A → C

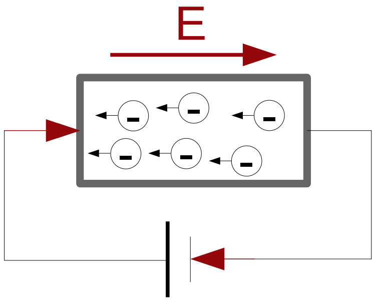


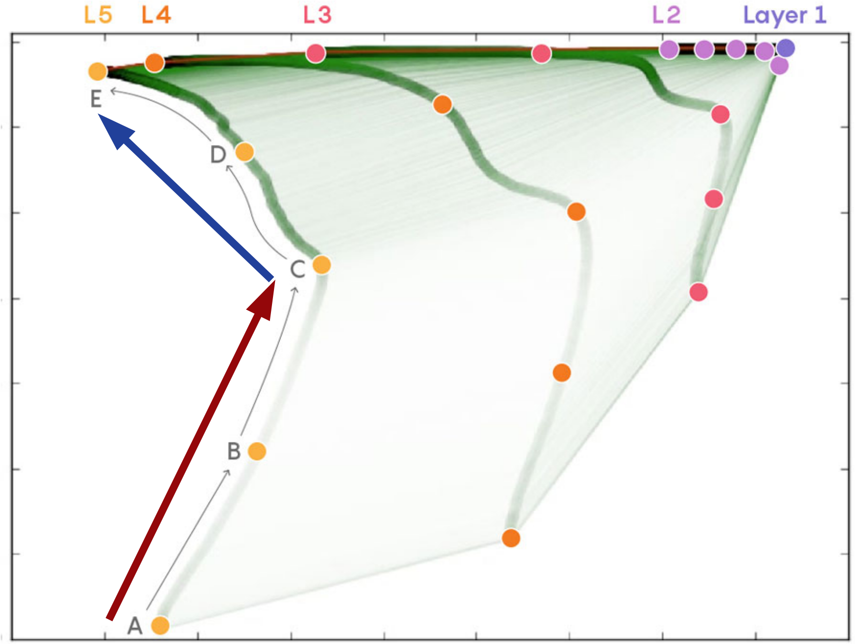
Drift vs Diffusion



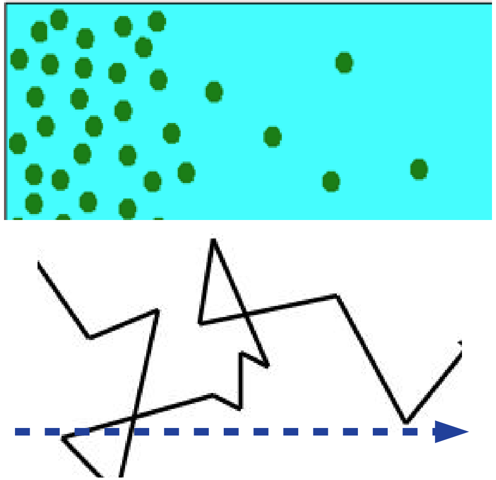
Diffusion

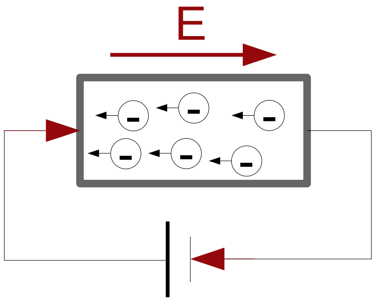
Drift

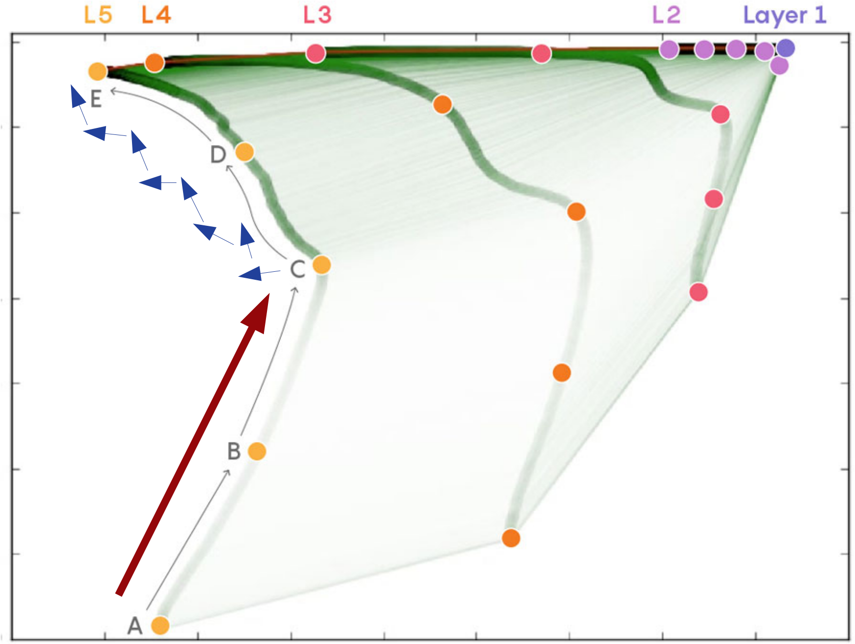




Drift vs Diffusion

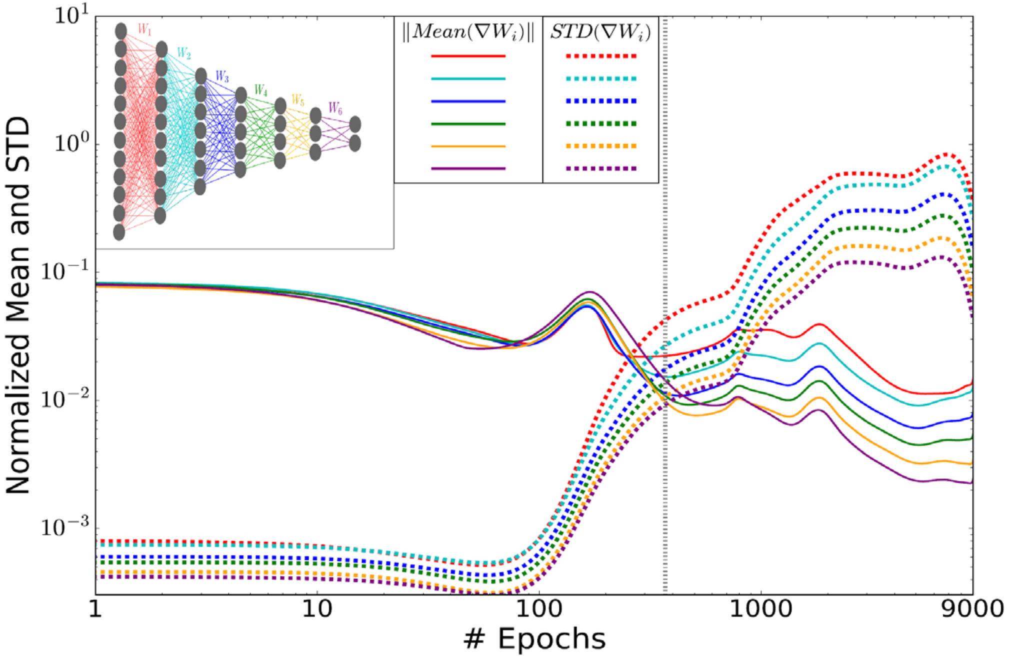






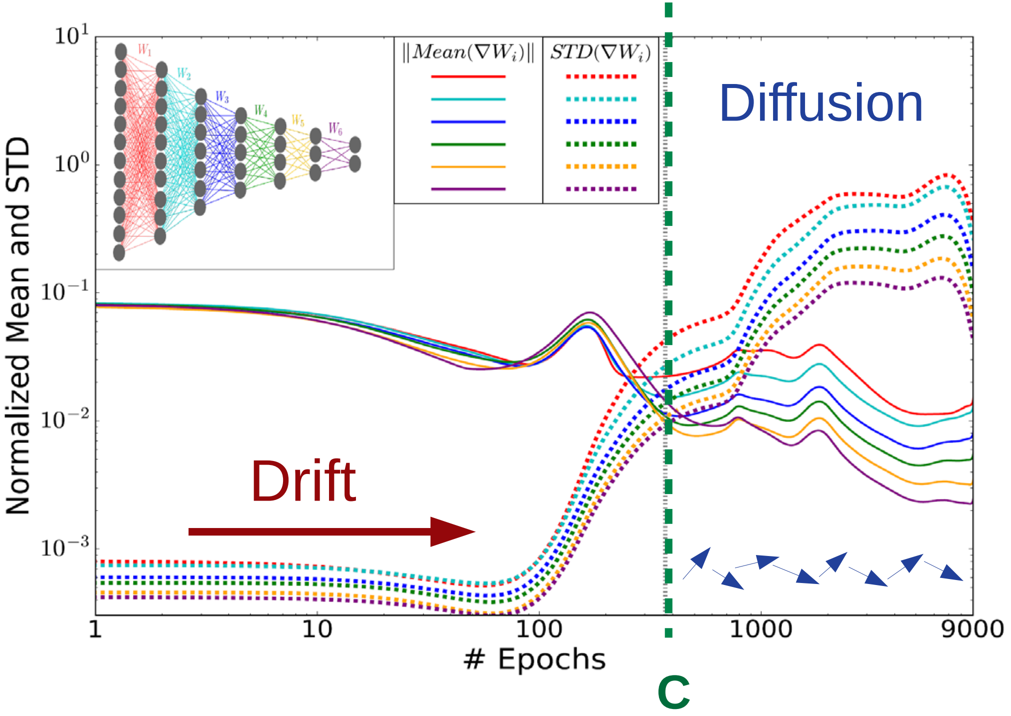
SNR of Gradient

$$\mathrm{SNR} \triangleq \frac{Mean(\|\nabla W_i\|)}{STD(\nabla W_i)}$$



SNR of Gradient

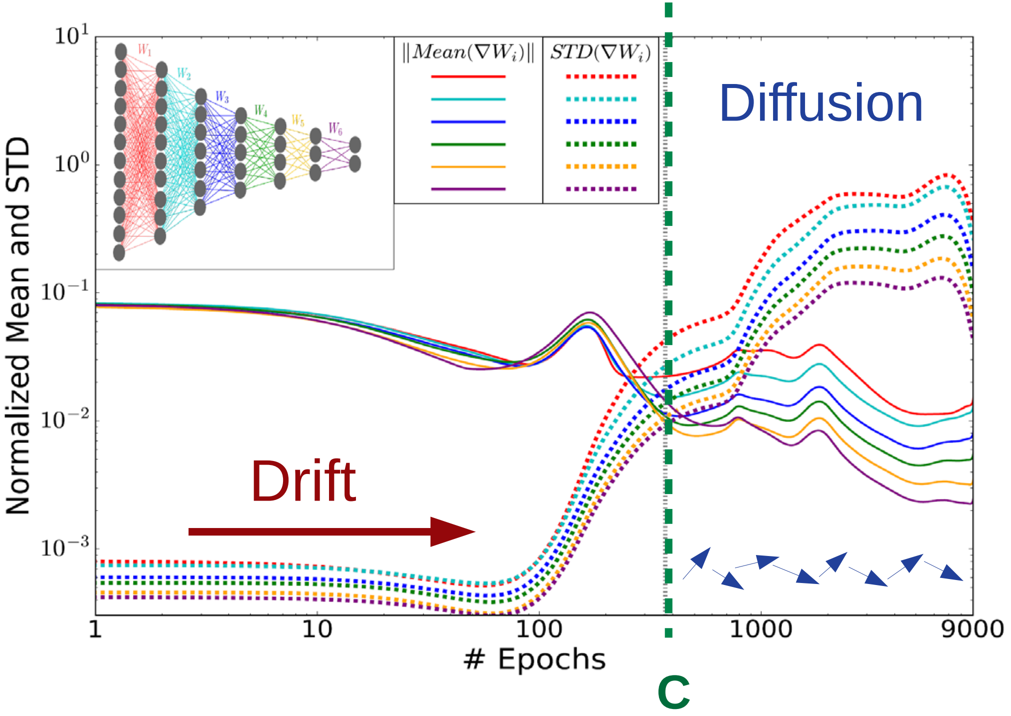
$$\text{SNR} \triangleq \frac{Mean(\|\nabla W_i\|)}{STD(\nabla W_i)}$$



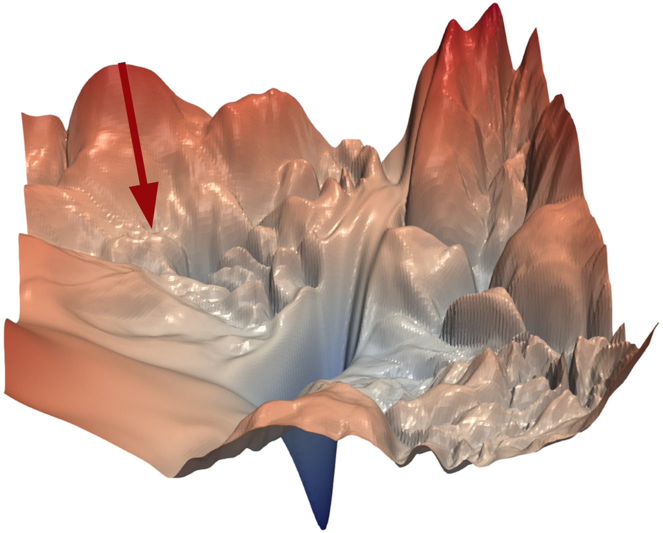
SNR of Gradient

$$\text{SNR} \triangleq \frac{Mean(\|\nabla W_i\|)}{STD(\nabla W_i)}$$

- Drift: High SNR $\rightarrow$ Fast
- Diffusion: Low SNR $\rightarrow$ Slow

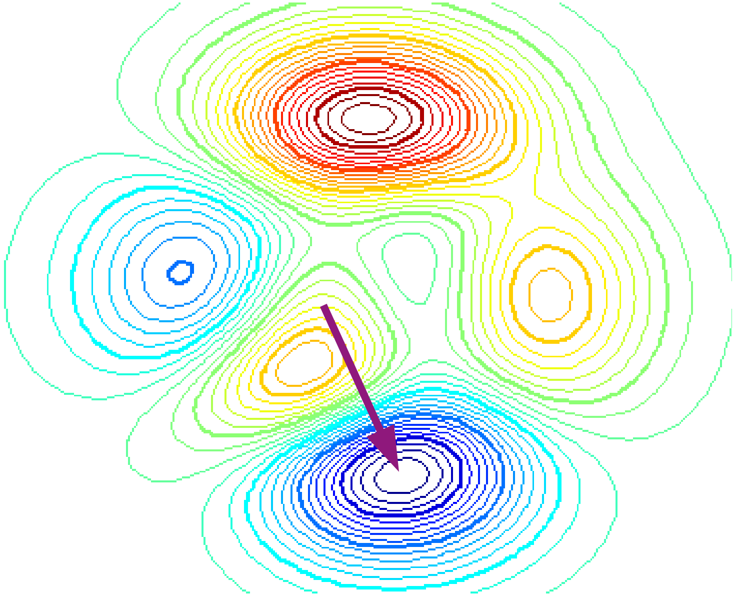


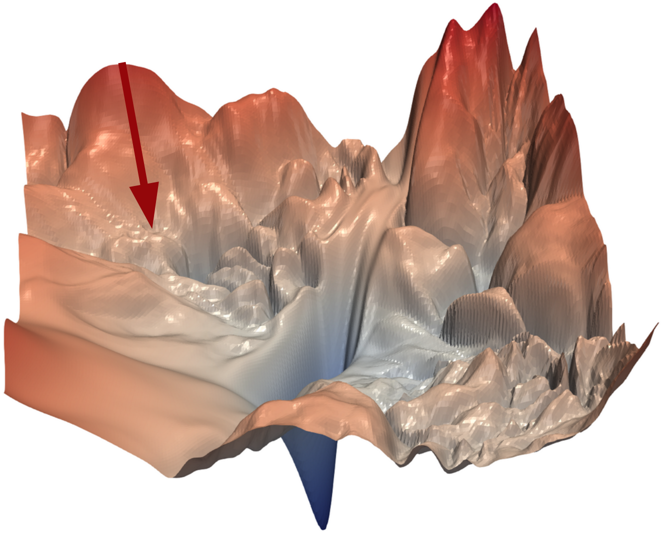
Diffusion Improves Generalisation



Drift (A → C) → High SNR

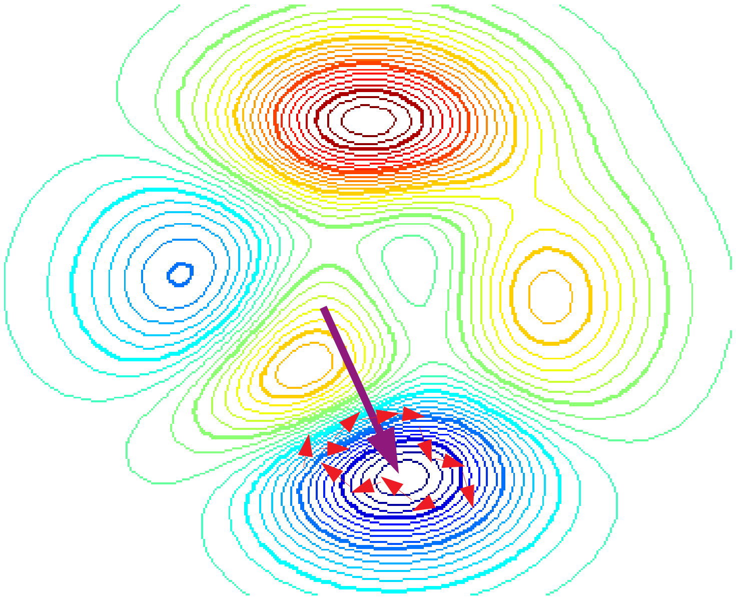
Diffusion Improves Generalisation



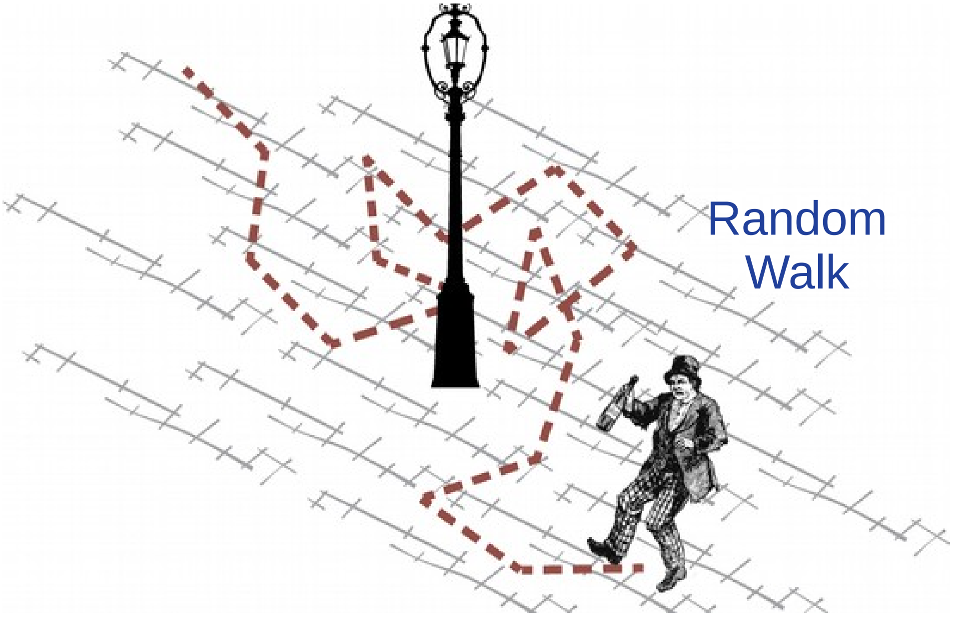


Drift (A→C) → High SNR

Diffusion Improves Generalisation

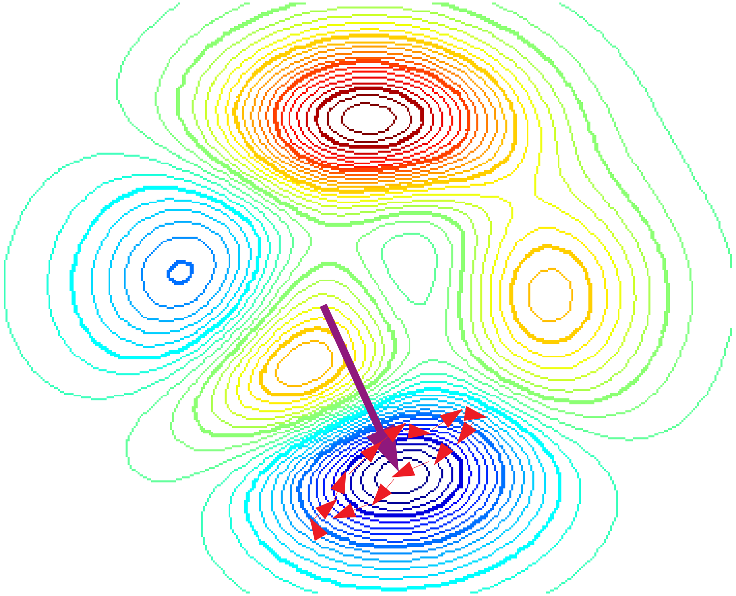


Drift (A→C) → High SNR

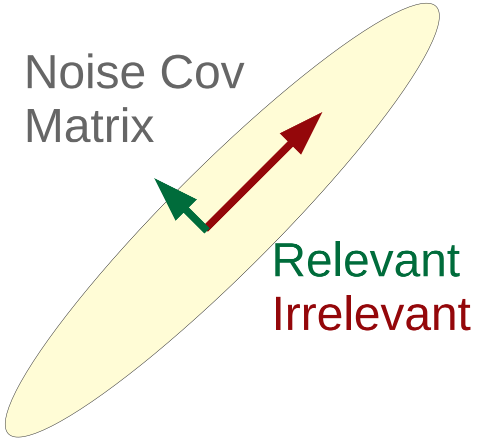


Diffusion (C→E) → Low SNR
Large stochasticity

Diffusion Improves Generalisation



Drift (A→C) → High SNR

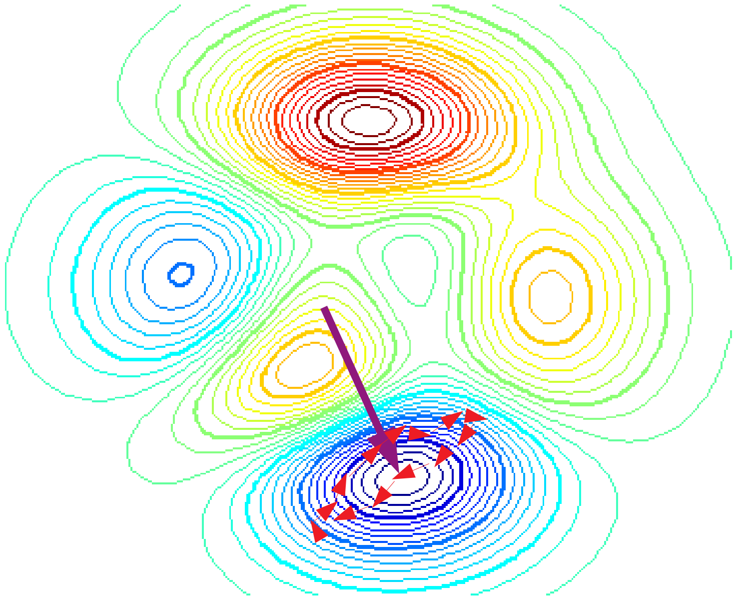


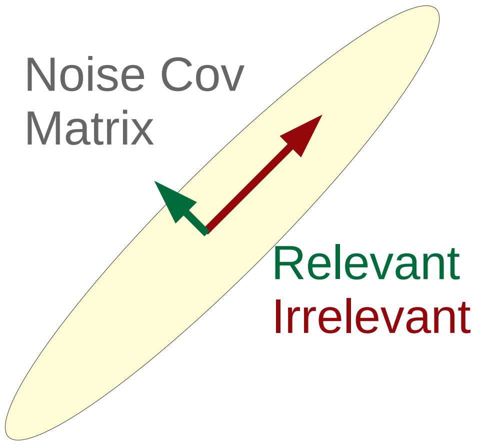
Diffusion → Large stochasticity

→ Add noise to irrelevant features

→ Forget irrelevant details

Diffusion Improves Generalisation





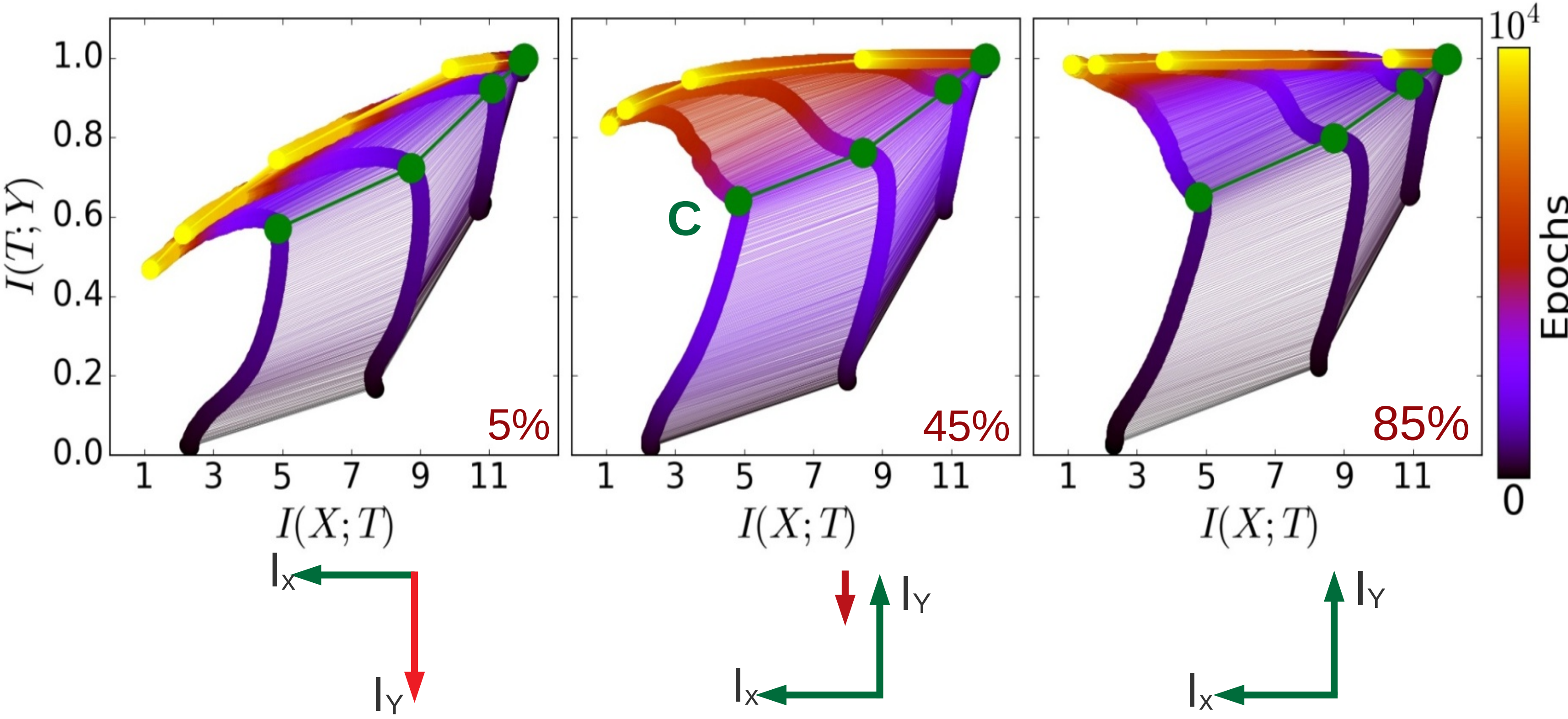
Drift (A → C) → High SNR

Diffusion → Large stochasticity

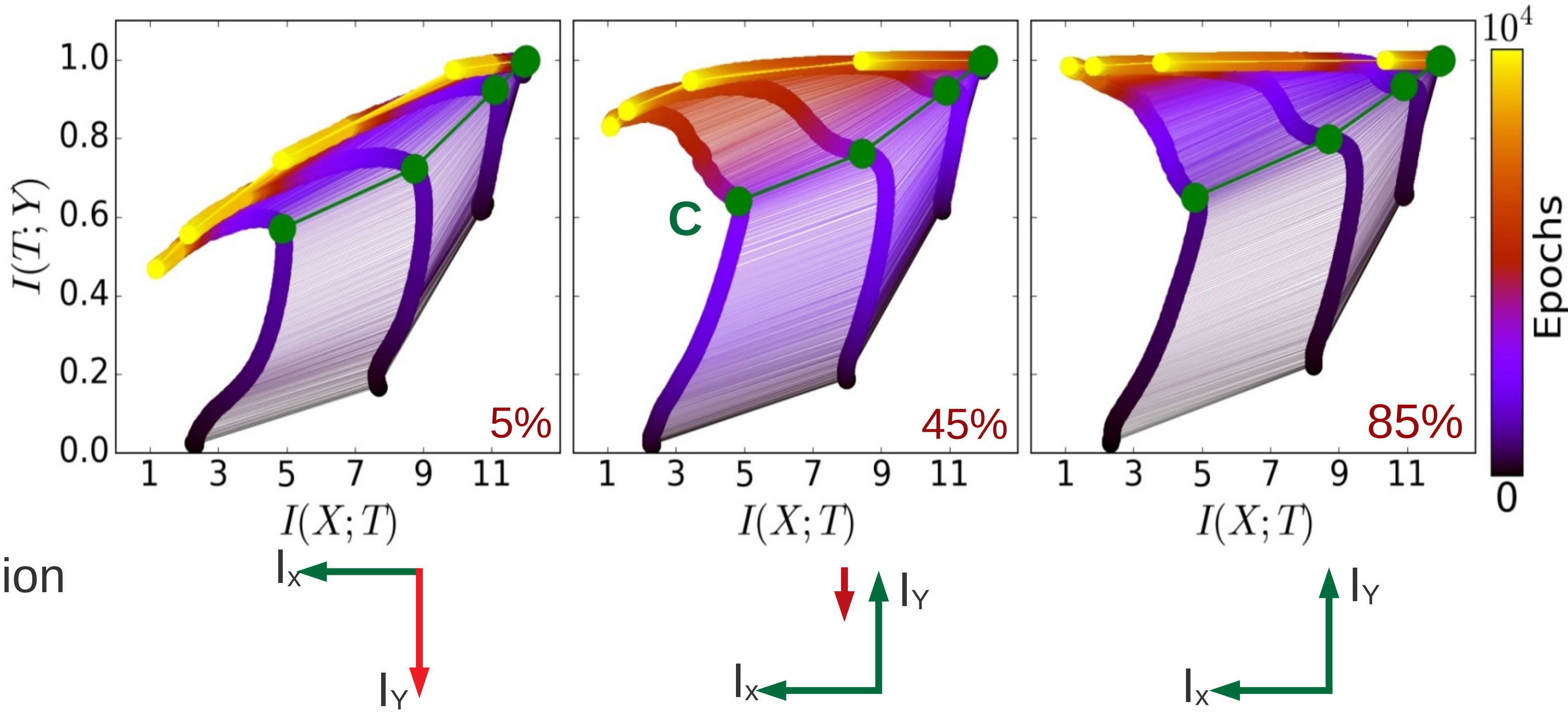
→ Add noise to irrelevant features

→ Forget irrelevant details

Effect of Amount of Data



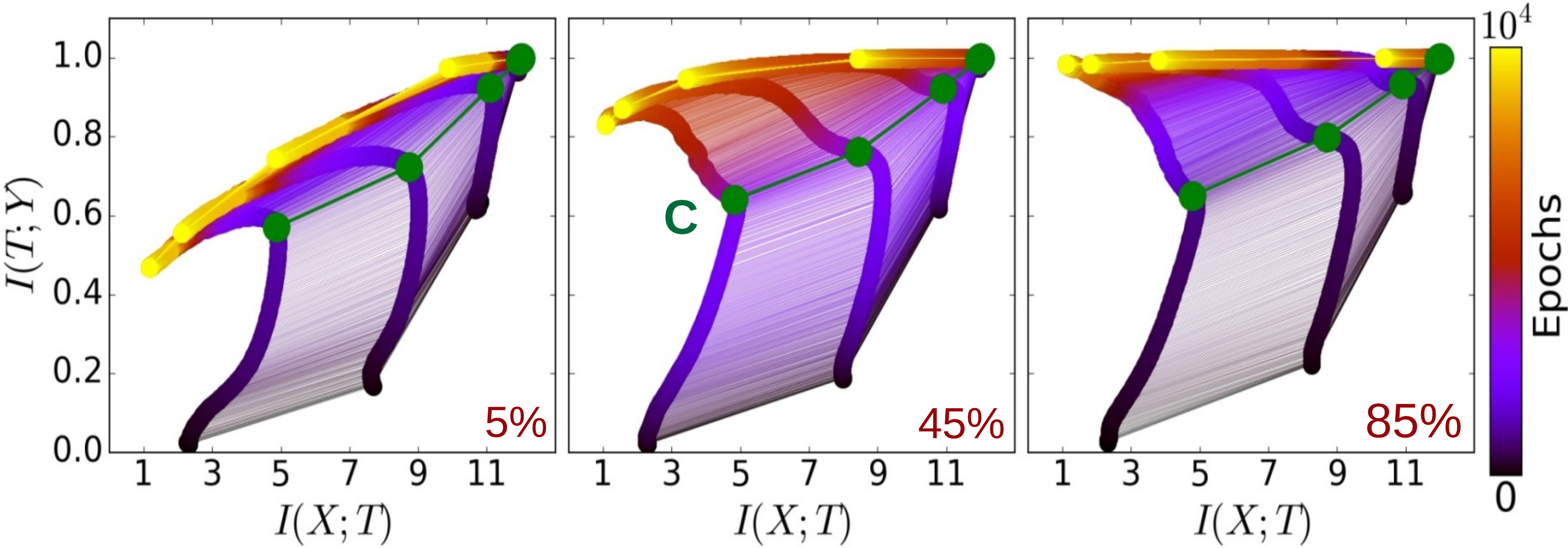
Effect of Amount of Data



Over-training

↔ $I_Y \downarrow$ in Diffusion

Effect of Amount of Data

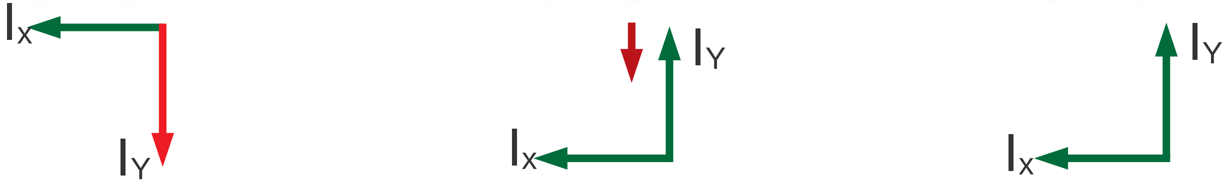


Over-training

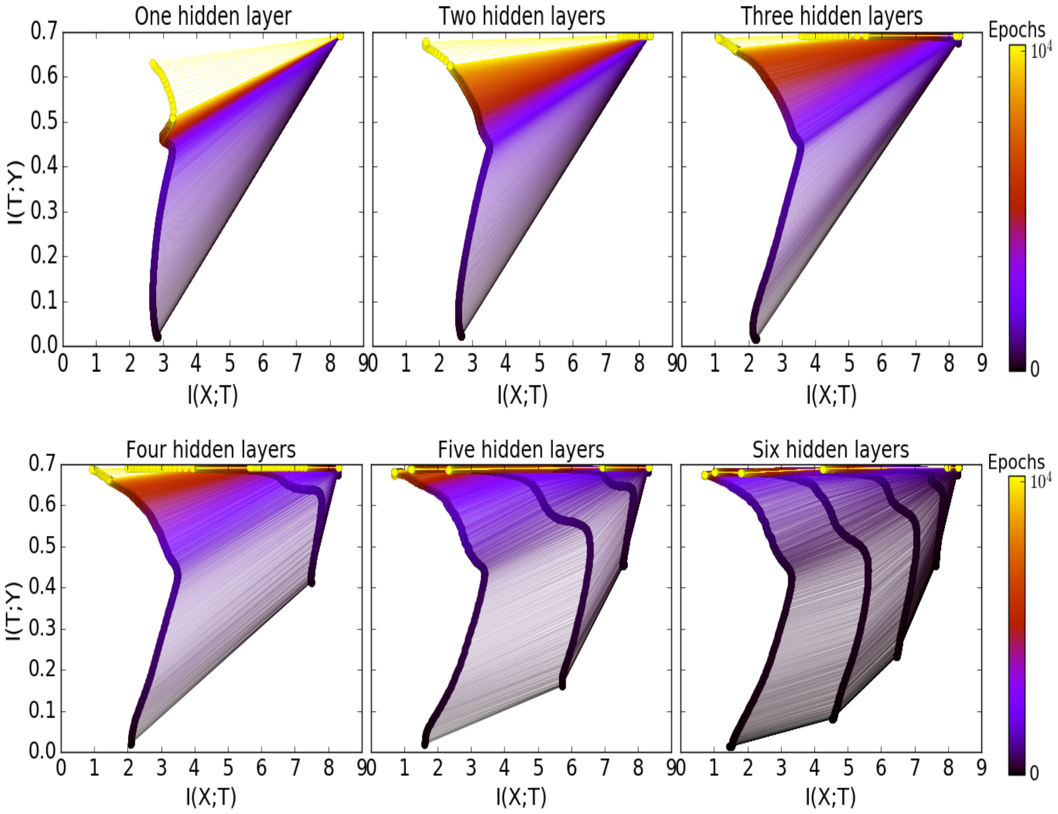
↔ I_Y ↓ in Diffusion

Ideal early stop ...

Just before I_Y ↓

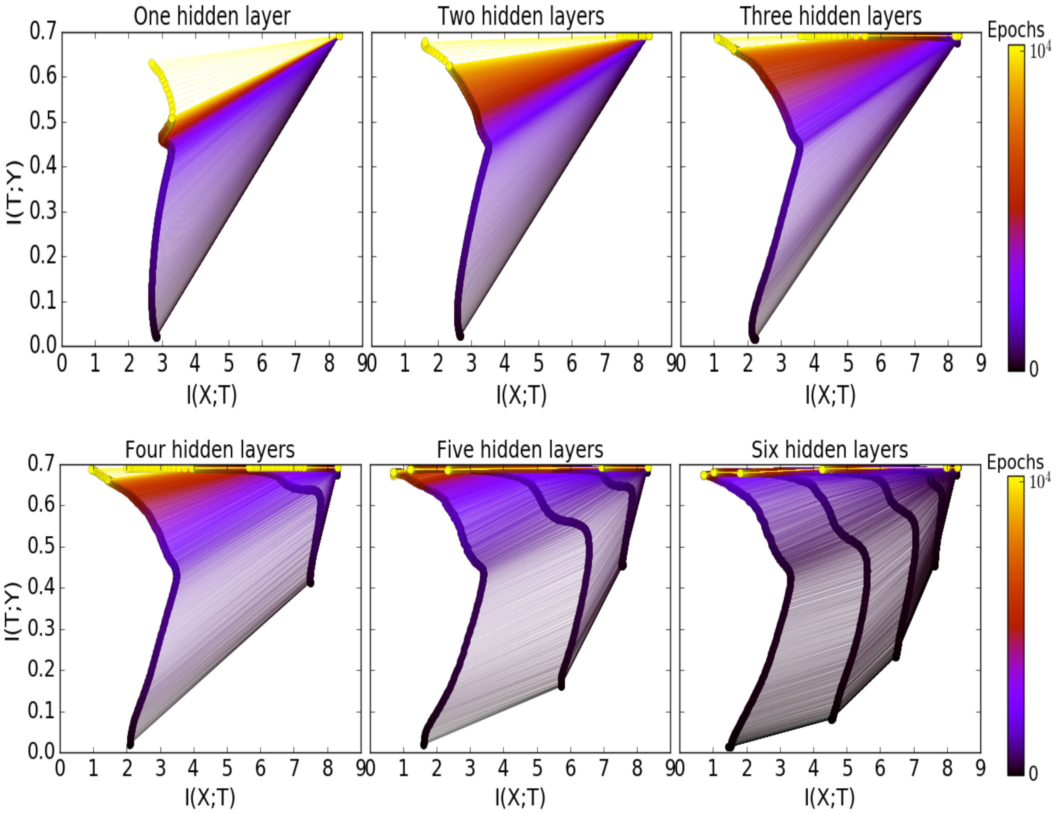


Advantage of More Layers



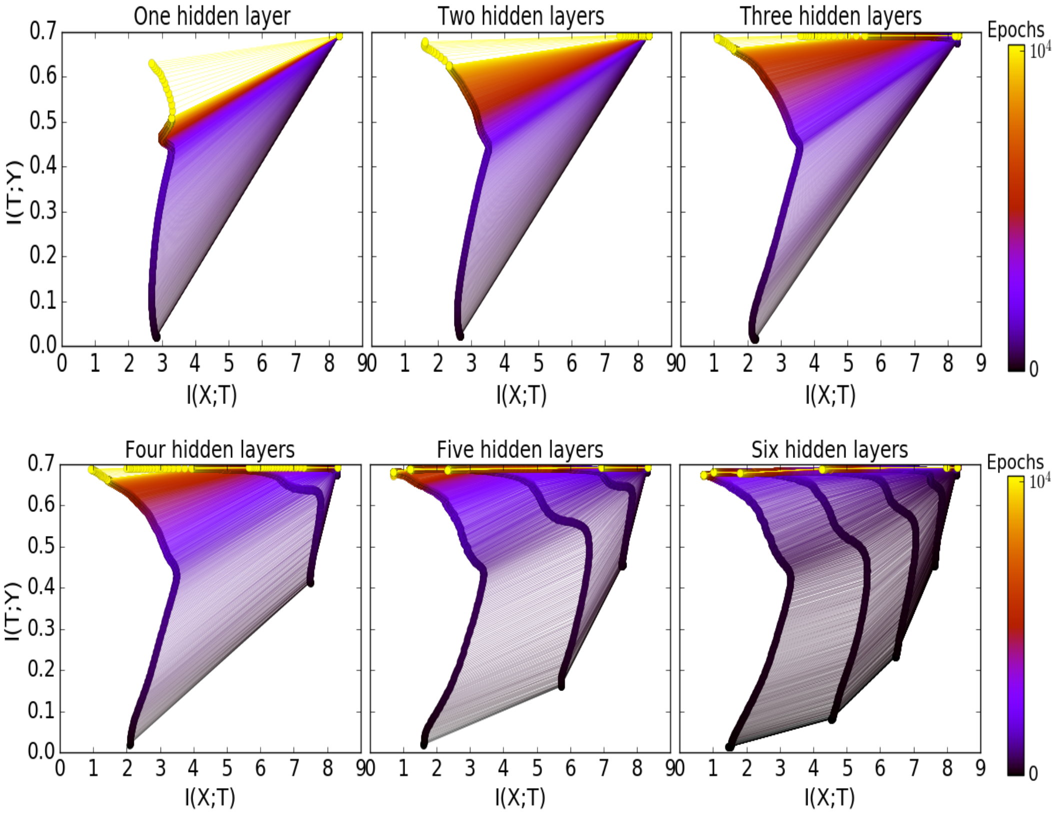
Advantage of More Layers

- Faster diffusion/convergence to good generalisation



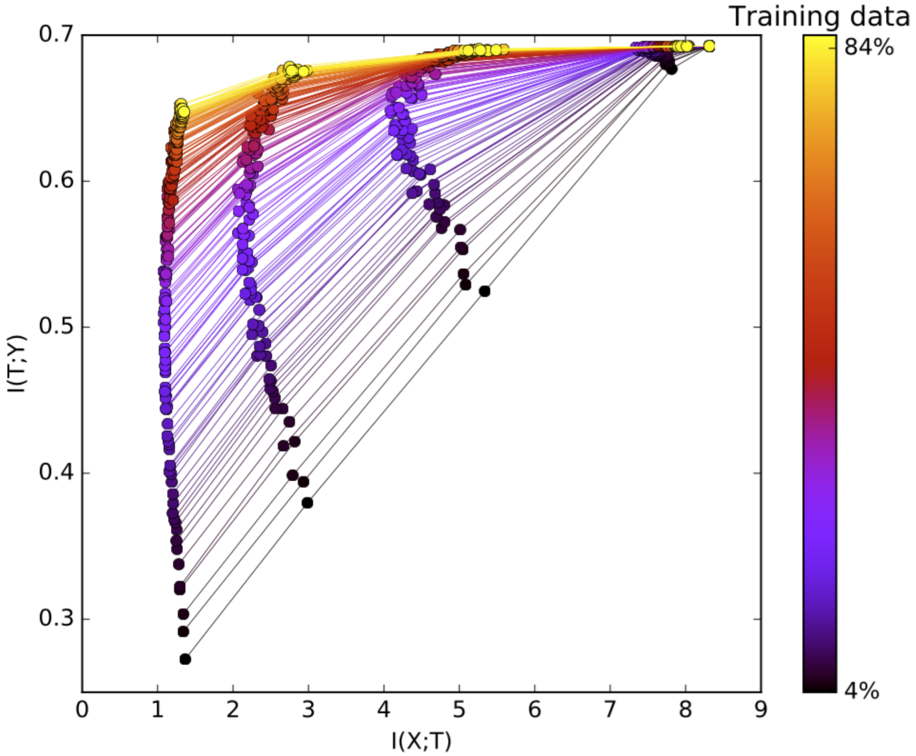
Advantage of More Layers

- Faster diffusion/convergence to good generalisation
- Convergence time scales as a negative power of the number of effective layers



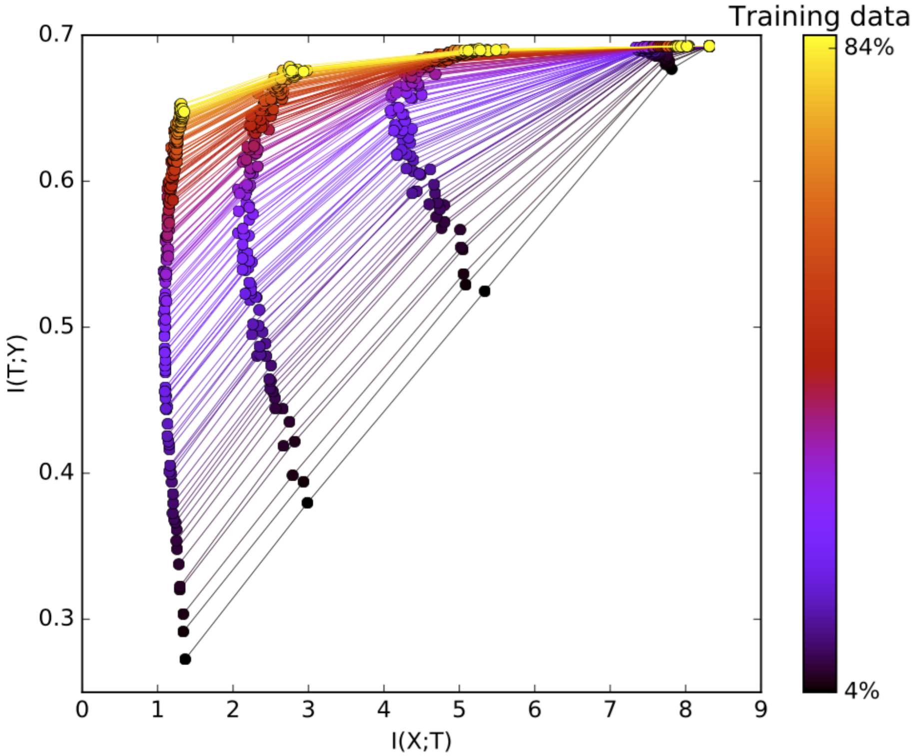
Effect of Training Data

- Better generalisation
  - Higher I_Y

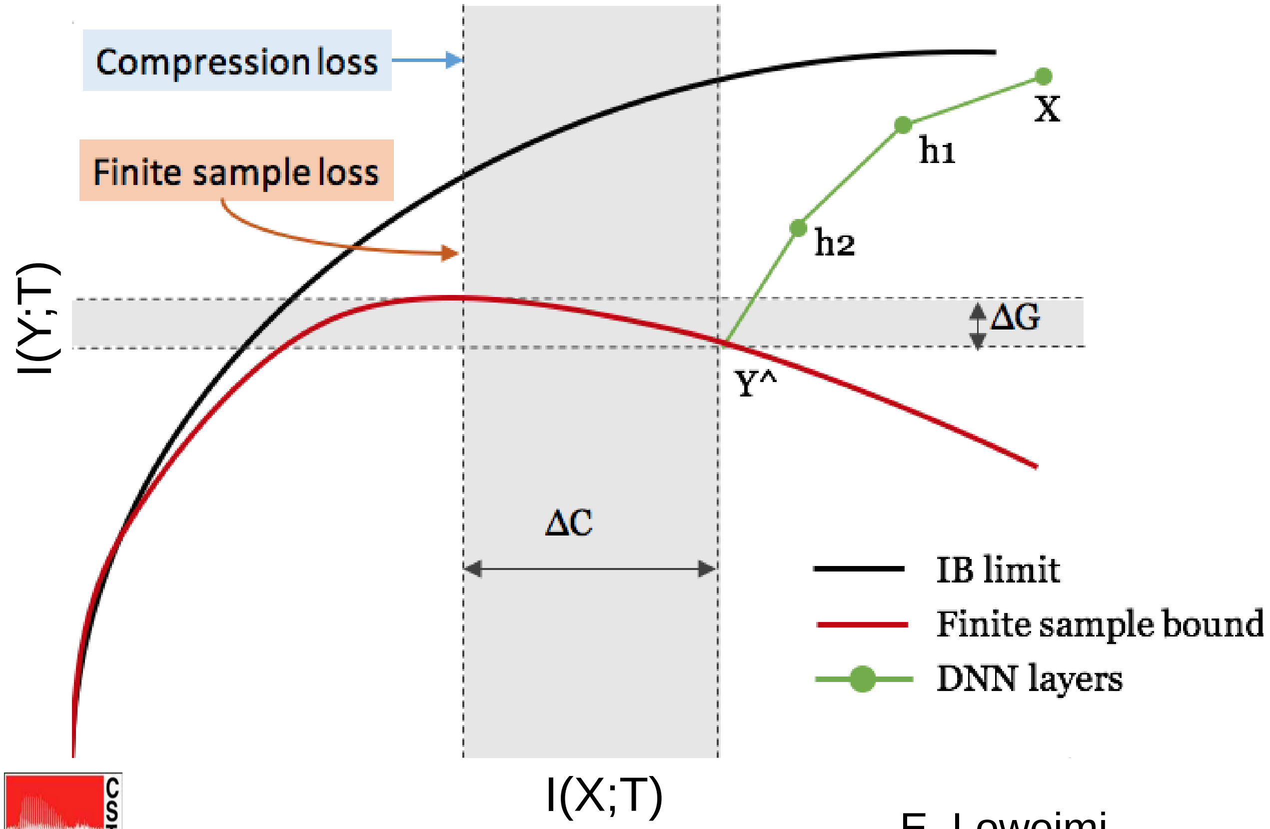


Effect of Training Data

- Better generalisation
  - Higher I_Y
- Limited effect on I_X (compression)
  - Remains almost constant

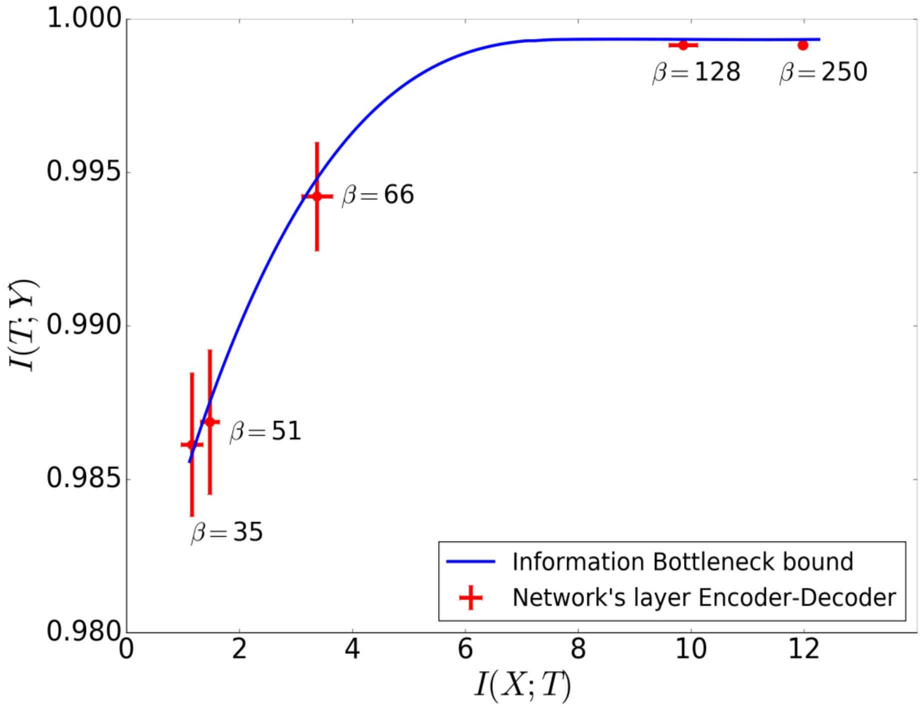


Theoretical IB Bound



Encoder-decoder structure for each layer satisfies the IB bounds

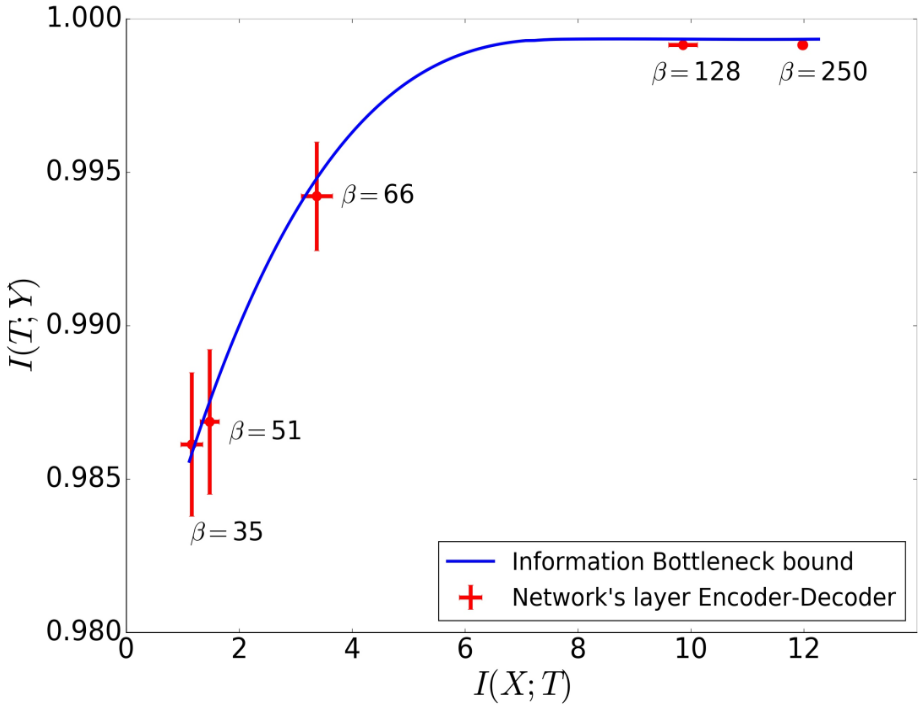
Theoretical IB Bound



Beta for different layers

$$\min_{q(t|x)} \{I(T;X) - \beta I(T;Y)\}$$

Theoretical IB Bound

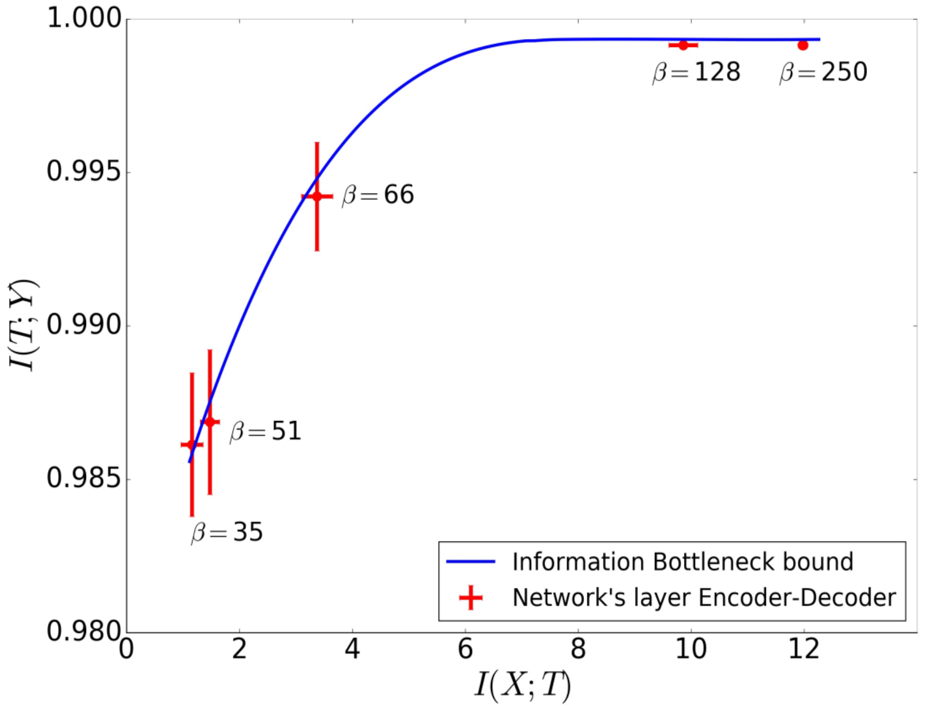


Beta for different layers

$$\min_{q(t|x)} \{I(T;X) - \beta I(T;Y)\}$$

β is inversely proportional with the slope of tangent line

Theoretical IB Bound



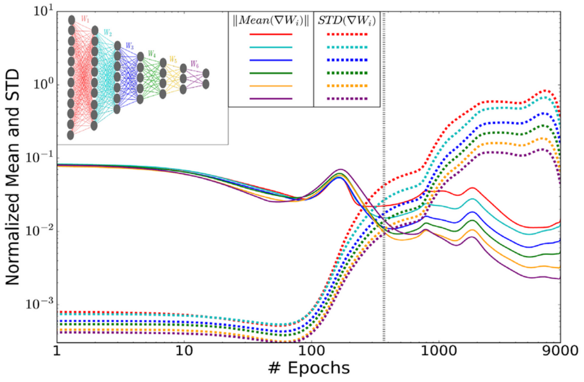
Beta for different layers

$$\min_{q(t|x)} \{I(T;X) - \beta I(T;Y)\}$$

β decreases by moving towards higher layers

How well the results generalise to other architectures and data?

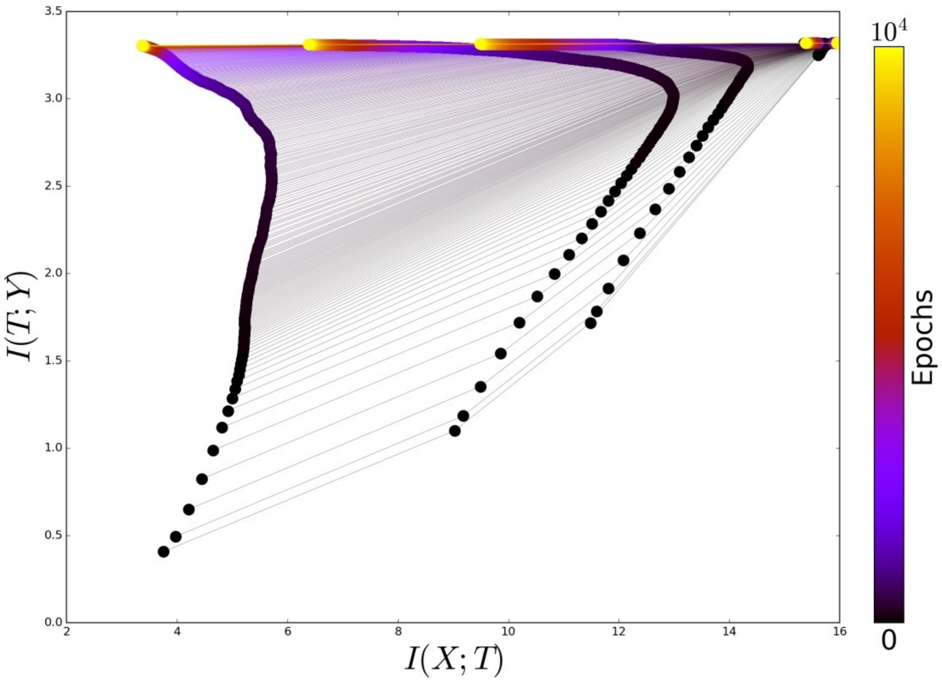
How well the results generalise to other architectures and data?



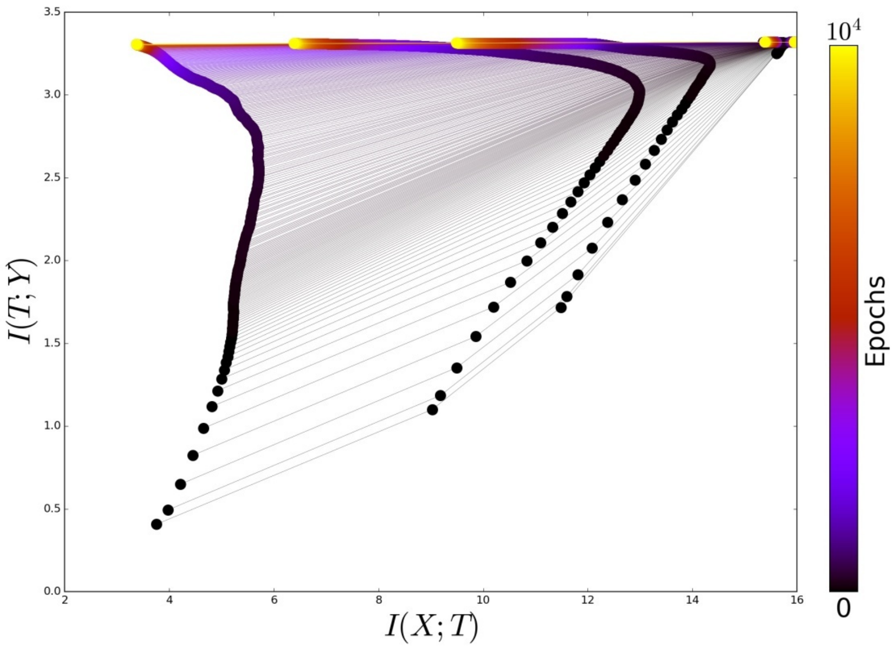
When your whole neural network can be drawn as the legend to the plot analyzing it (Figure 4), you know you are in trouble!

Zeeshan Zia, a discussion in Quora

MNIST (or CIFAR-10 ?) CNN + ReLU



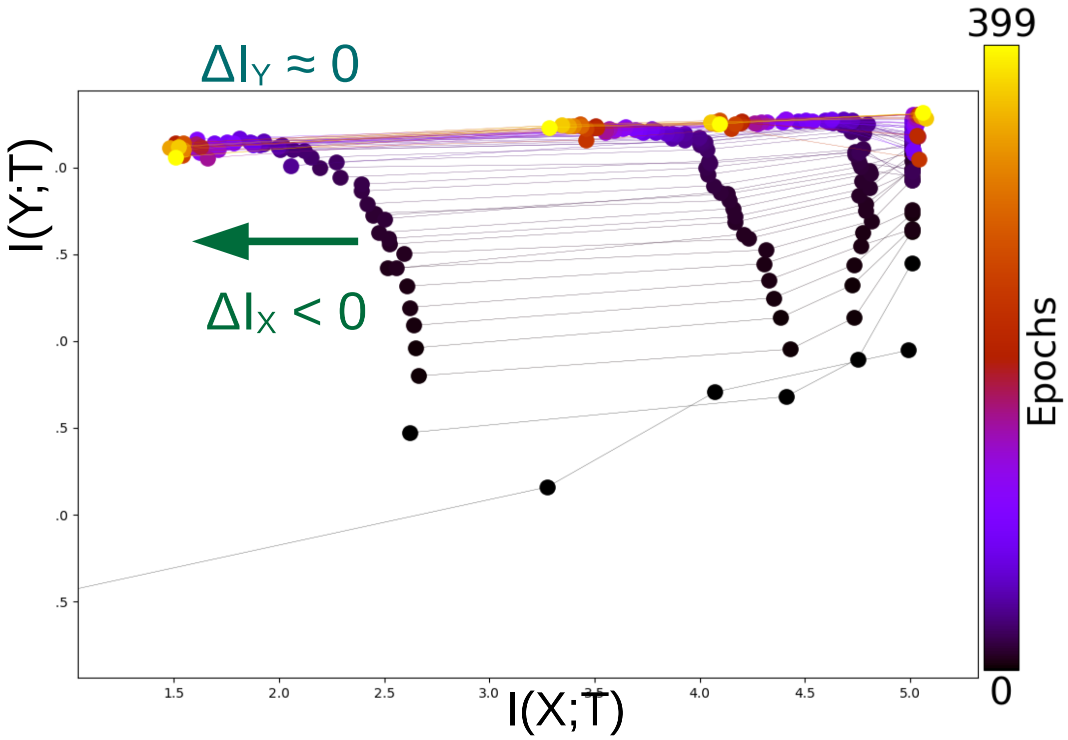
MNIST + CNN



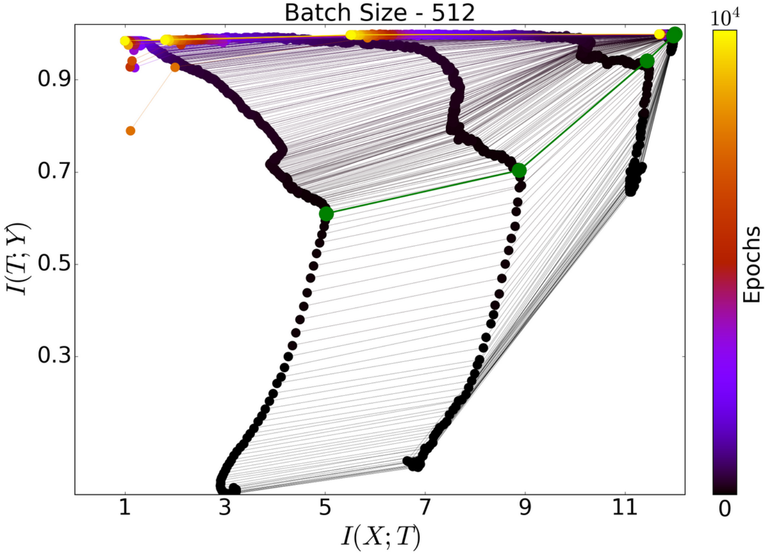
CIFAR-10 + CNN + ReLU

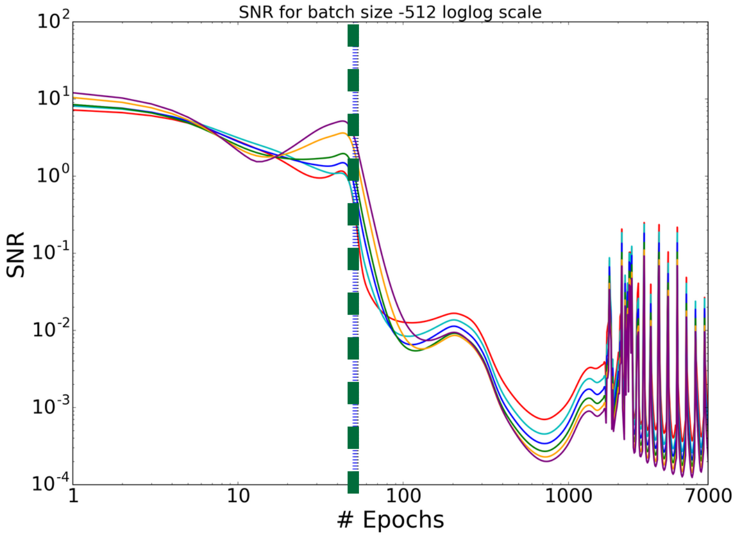
CIFAR-10

Argument holds if in diffusion
$\Delta I_X < 0$ and $\Delta I_Y \geq 0$,
which is the case here

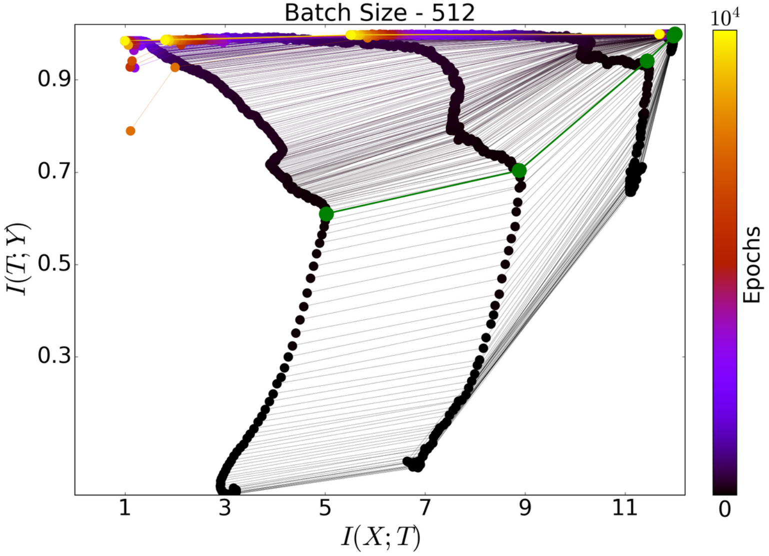


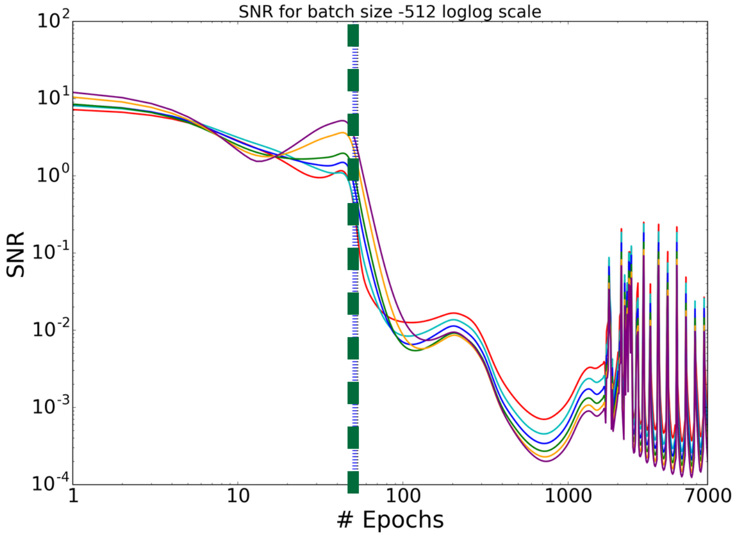
Role of the Batch Size





Role of the Batch Size



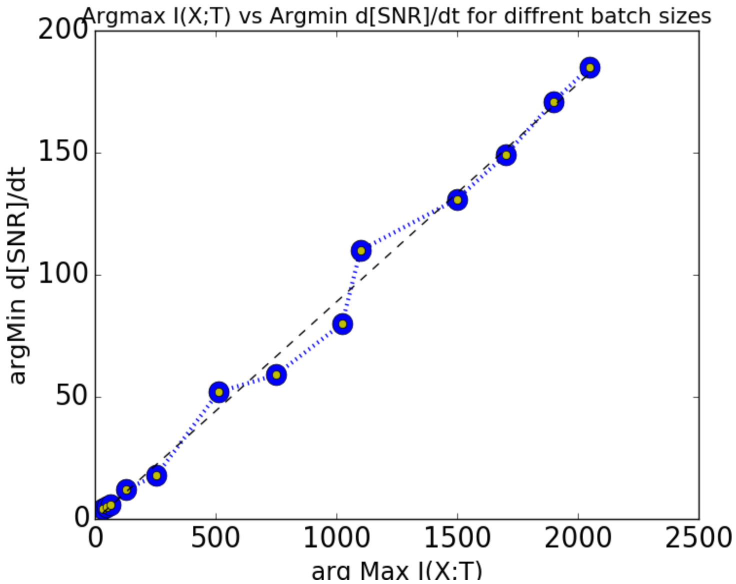


Drift to *diffusion* transition:

$$\operatorname{argmin} \frac{d}{dt} SNR \approx \operatorname{argmax} I(X;T)$$

E. Loweimi

Role of the Batch Size

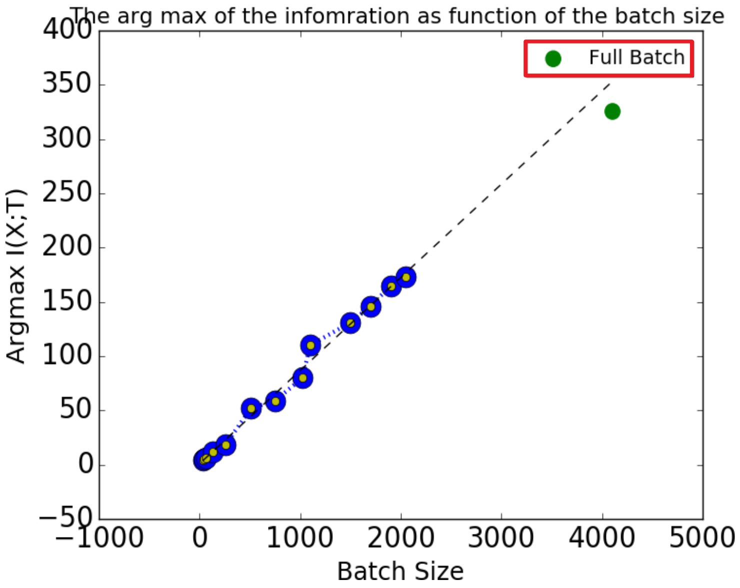


Larger batch size →
delays transition to diffusion
→ more iterations required
(less randomness in diffusion)

– x-axis: batch size (incorrect xlabel/title)

– y-axis: argmin d/dt SNR

Role of the Batch Size



Larger batch size →
delays transition to diffusion
→ more iterations required
(less randomness in diffusion)

– x-axis: batch size

– y-axis: argmax I(X;T)

Outlines

- Problem Statement
- Information Theory Review
- Information Bottleneck (IB)
- Opening the Black Box of DNNs via IB
- Criticisms
- Conclusions

OpenReview.net Search ICLR 2018 Conference Login

← Go to ICLR 2018 Conference homepage

On the Information Bottleneck Theory of Deep Learning

Andrew Michael Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Daniel Tracey, David Daniel Cox

15 Feb 2018 (modified: 24 Feb 2018) ICLR 2018 Conference Blind Submission Readers: Everyone Show Bibtex Show Revisions

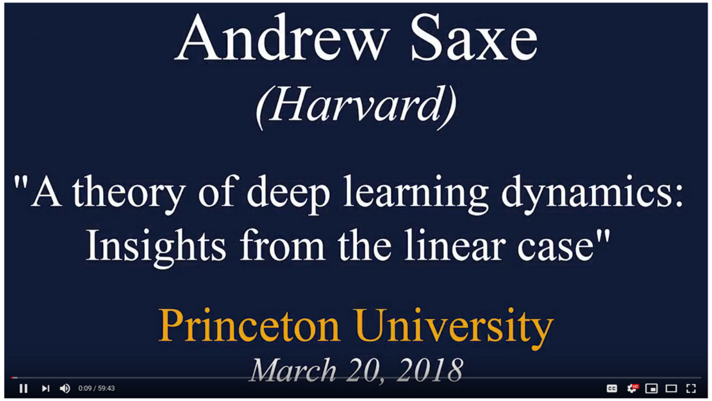
Abstract: The practical successes of deep neural networks have not been matched by theoretical progress that satisfyingly explains their behavior. In this work, we study the information bottleneck (IB) theory of deep learning, which makes three specific claims: first, that deep networks undergo two distinct phases consisting of an initial fitting phase and a subsequent compression phase; second, that the compression phase is causally related to the excellent generalization performance of deep networks; and third, that the compression phase occurs due to the diffusion-like behavior of stochastic gradient descent. Here we show that none of these claims hold true in the general case. Through a combination of analytical results and simulation, we demonstrate that the information plane trajectory is predominantly a function of the neural nonlinearity employed: double-sided saturating nonlinearities like tanh yield a compression phase as neural activations enter the saturation regime, but linear activation functions and single-sided saturating nonlinearities like the widely used ReLU in fact do not. Moreover, we find that there is no evident causal connection between compression and generalization: networks that do not compress are still capable of generalization, and vice versa. Next, we show that the compression phase, when it exists, does not arise from stochasticity in training by demonstrating that we can replicate the IB findings using full batch gradient descent rather than stochastic gradient descent. Finally, we show that when an input domain consists of a subset of task-relevant and task-irrelevant information, hidden representations do compress the task-irrelevant information, although the overall information about the input may monotonically increase with training time, and that this compression happens concurrently with the fitting process rather than during a subsequent compression period.

TL;DR: We show that several claims of the information bottleneck theory of deep learning are not true in the general case.

Keywords: information bottleneck, deep learning, deep linear networks

21 Replies

YouTube andre saxe linear neural network



Andrew Saxe: A theory of deep learning dynamics: Insights from the linear case

OpenReview.net Search ICLR 2018 Conference Login

← Go to ICLR 2018 Conference homepage

i-RevNet: Deep Invertible Networks

Jörn-Henrik Jacobsen, Arnold W.M. Smeulders, Edouard Oyallon

15 Feb 2018 (modified: 23 Feb 2018) ICLR 2018 Conference Blind Submission Readers: Everyone Show Bibtex Show Revisions

Abstract: It is widely believed that the success of deep convolutional networks is based on progressively discarding uninformative variability about the input with respect to the problem at hand. This is supported empirically by the difficulty of recovering images from their hidden representations, in most commonly used network architectures. In this paper we show via a one-to-one mapping that this loss of information is not a necessary condition to learn representations that generalize well on complicated problems, such as ImageNet. Via a cascade of homeomorphic layers, we build the i-RevNet, a network that can be fully inverted up to the final projection onto the classes, i.e. no information is discarded. Building an invertible architecture is difficult, for one, because the local inversion is ill-conditioned, we overcome this by providing an explicit inverse.
An analysis of i-RevNet's learned representations suggests an alternative explanation for the success of deep networks by a progressive contraction and linear separation with depth. To shed light on the nature of the model learned by the i-RevNet we reconstruct linear interpolations between natural image representations.

20 Replies

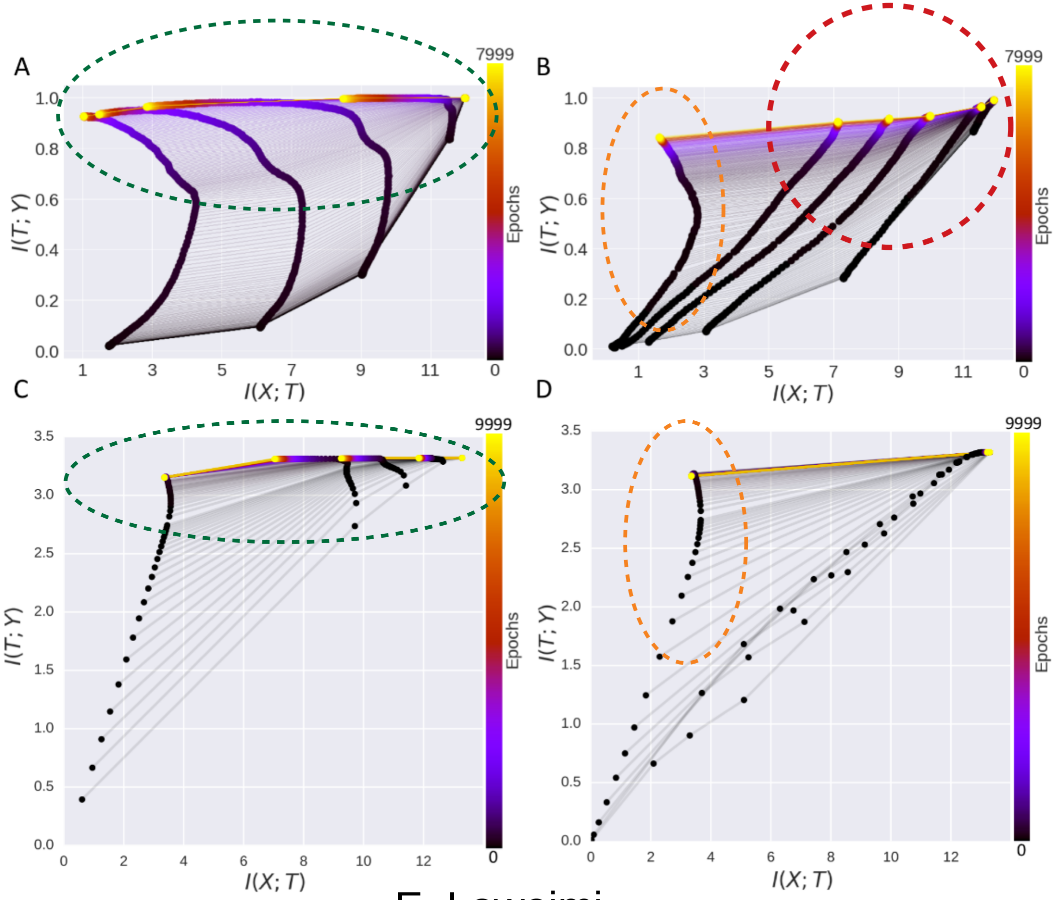
Rebuttal

- Compression is an artefact of the double saturation of Tanh, it does not happen for ReLU

Two-phase process is not generic!

Tanh,
small data,
MI: Binnig

Tanh,
MNIST,
MI: kernel-based



ReLU,
small data,
MI: Binnig

ReLU,
MNIST,
MI: kernel-based

E. Loweimi

Rebuttal

- Compression is an artefact of the double saturation of Tanh, it does not happen for ReLU

Rebuttal

- Compression is an artefact of the double saturation of Tanh, it does not happen for ReLU
- Slow diffusion $\leftrightarrow$ due to small gradient at saturation

Rebuttal

- Compression is an artefact of the double saturation of Tanh, it does not happen for ReLU
- Slow diffusion $\leftrightarrow$ due to small gradient at saturation
- No causal relationship between compression (stochasticity of SGD) and generalisation

Rebuttal

- Compression is an artefact of the double saturation of Tanh, it does not happen for ReLU
- Slow diffusion $\leftrightarrow$ due to small gradient at saturation
- No causal relationship between compression (stochasticity of SGD) and generalisation
- i-RevNet: loss of information is not a necessary condition to learn representations that generalise well

Conclusions

- Novel approach: Using Information Theory to study DNNs
- Mutual information between input/hidden/output layers is investigated to understand/visualise the DNNs and their learning dynamics
- Why DNNs generalise well?
  - Stochasticity of SGD in diffusion → forgetting irrelevant info by adding noise
- Trend is claimed to be general but ...
  - ReLU → no compression
  - iRev-Net → no forgetting

and still generalise well!

That’s It!

- Thanks for Your Attention
- Q & A

Appendices

(A1) Sufficient Statistics

(A2) Information Bottleneck – Solution

(A3) Some Useful Resources

(A4) Information Theory and Statistical Mechanics

(A1) Sufficient Statistics

- Fisher's Factorisation Theorem (*g* and *h* ≥ 0)

$$p_\theta(x) = h(x)\ g_\theta(T(x))$$

- T is sufficient statistics for Y if

$$\text{Markov Chain} : Y \rightarrow X \rightarrow T$$

$$I(T;Y) = I(X;Y) \longleftrightarrow Y \perp\!\!\!\perp X | T$$

- T is minimum sufficient statistics for Y if

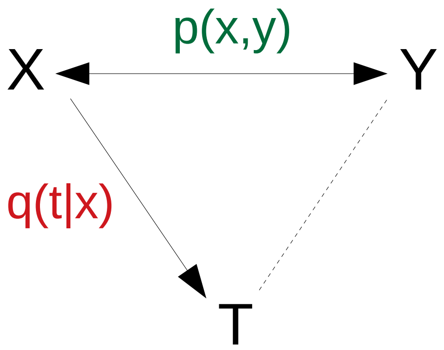
$$T(X) = \text{argmin}\ I(T;X)$$

$$\text{s.t.}\ \ I(T;X) = I(T;Y)$$

(A2) Info Bottleneck – Solution

$$\min_{q(t|x),q(t),q(y|t)} \{I(T;X) - \beta I(T;Y)\}$$

$$\begin{cases} q(t|x) & = \frac{q(t)}{Z} \exp(-\beta D_{KL}[p(y|x)||q(y|t)]) \\ q(t) & = \sum_x p(x)q(t|x) \\ q(y|t) & = \frac{1}{q(t)} \sum_x p(y|x)q(t|x)p(x) \end{cases}$$



X: observation
Y: variable of interest
T: representation of X

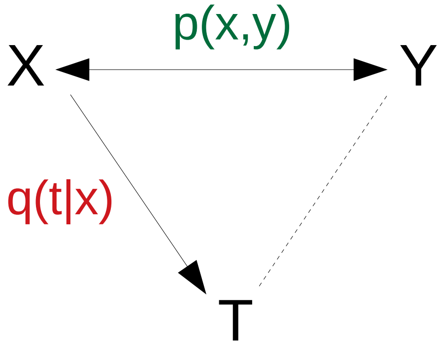
Named *self-consistent* equations

Solution: Blahut-Arimoto (iterative)

(A2) Info Bottleneck – Solution

$$\min_{q(t|x),q(t),q(y|t)} \{I(T;X) - \beta I(T;Y)\}$$

$$\begin{array}{l} \text{Key} \\[1.2em] \text{Sum} \\[1.2em] \text{Bayes} \end{array} \begin{cases} \boxed{q(t|x) = \frac{q(t)}{Z}\exp(-\beta D_{KL}[p(y|x)||q(y|t)])} \\ q(t) = \sum_x p(x)q(t|x) \\ q(y|t) = \frac{1}{q(t)}\sum_x p(y|x)q(t|x)p(x) \end{cases}$$



X: observation
Y: variable of interest
T: representation of X

Named *self-consistent* equations

Solution: Blahut-Arimoto (iterative)

(A3) Useful Resources

- Tishby's Talk in Simon Institute

  https://www.youtube.com/watch?v=EQTtBRM0sIs

- Information Bottleneck Workshop
  http://www.cs.huji.ac.il/~tishby/NIPS-Workshop/

- The optimization process in the Information Plane

  https://www.youtube.com/watch?v=P1A1yNsxMjc

- Ravid Schwarz-Ziv, Data Science Summit 2018

  https://www.youtube.com/watch?v=gOn8Po_NPe4

(A4) Info Theory & Statistical Mechanics

PHYSICAL REVIEW VOLUME 106, NUMBER 4 MAY 15, 1957

Information Theory and Statistical Mechanics

E. T. Jaynes
Department of Physics, Stanford University, Stanford, California
(Received September 4, 1956; revised manuscript received March 4, 1957)

Information theory provides a constructive criterion for setting up probability distributions on the basis of partial knowledge, and leads to a type of statistical inference which is called the maximum-entropy estimate. It is the least biased estimate possible on the given information; i.e., it is maximally noncommittal with regard to missing information. If one considers statistical mechanics as a form of statistical inference rather than as a physical theory, it is fou
rules, starting with the detern
are an immediate consequence
In the resulting "subjective sta
are thus justified independent
in particular independently of

or not the results agree with experiment, they still represent the best estimates that could have been made on the basis of the information available.

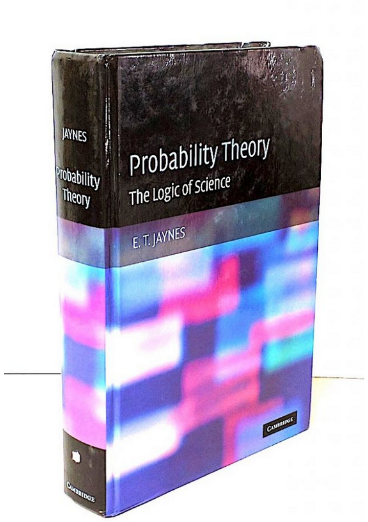
It is concluded that statistical mechanics need not be regarded as a physical theory dependent for its validity on the truth of additional assumptions not contained in the laws of mechanics (such as ergodicity, metric transitivity, equal *a priori* probabilities,

IEEE Transactions On Systems Science and Cybernetics, vol. sec-4, no. 3, 1968, pp. 227-241

Prior Probabilities

Edwin T. Jaynes
Department of Physics, Washington University, St. Louis, Missouri

In decision theory, mathematical analysis shows that once the sampling distributions, loss function, and sample are specified, the only remaining basis for a choice among different admissible decisions lies in the prior probabilities. Therefore, the logical foundations of decision theory cannot be put in fully satisfactory form until the old problem of arbitrariness (sometimes called "subjectiveness") in assigning prior probabilities is resolved.



Edwin Thompson Jaynes
(1922-1998)