Towards a Unified Theory of Learning and Information

Ibrahim Alabdulmohsin

November 2018

Key Takeaway

Q: When can we extrapolate from the **past** into the **future**?
A: Traditional approaches include statistical tools (e.g. VC theory, covering numbers, and multiple hypothesis testing) and algorithmic tools (e.g. stability).

However, there is a recent surge of interest in using **Information Theory** to answer this question.

- *Advantages*: robust guarantees, adaptive learning, general setting, simplified analysis, new tools, unifies results ...etc.

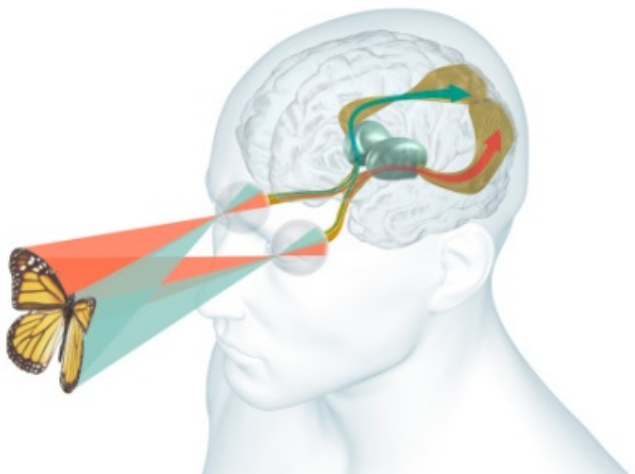
Background

A Bit of History

Ibn Al-Haitham (Alhazen) wrote his celebrated book *Optics* in around 1015 AD.

1. Introduced the scientific method, and emphasized the role of empirical evidence.
2. Established, among others: light travel in straight lines, eyes are not the source of light, refraction and reflection, ... etc.

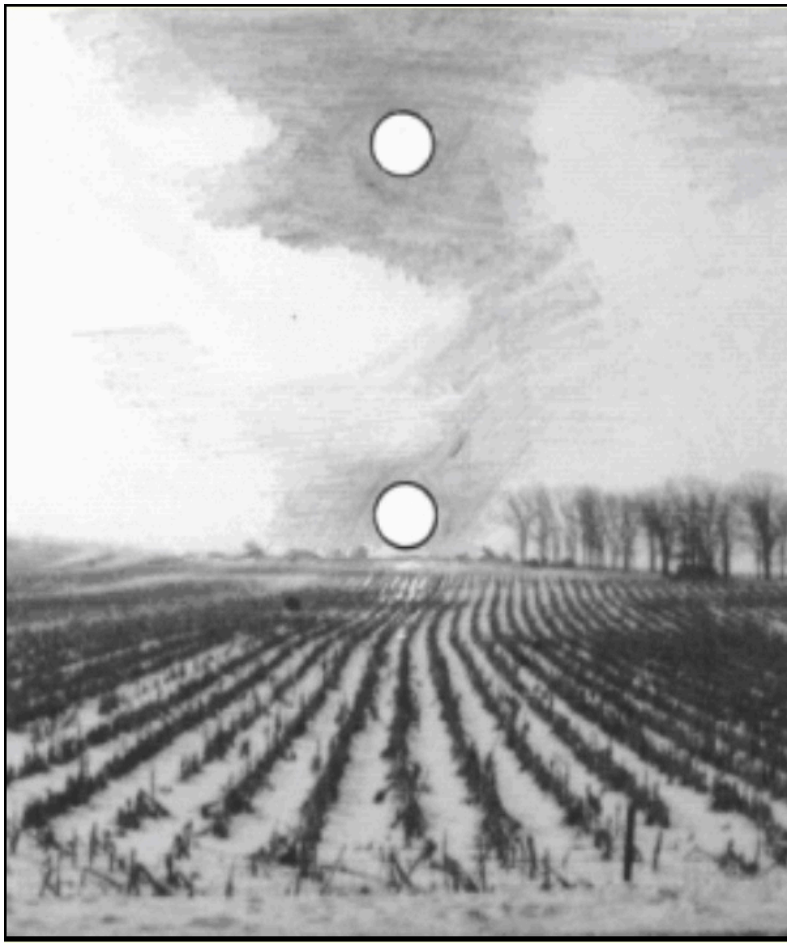
Ibn Al-Haitham is less known for a key significant contribution he made in his *Optics*.
He argued that "perception was inferential".

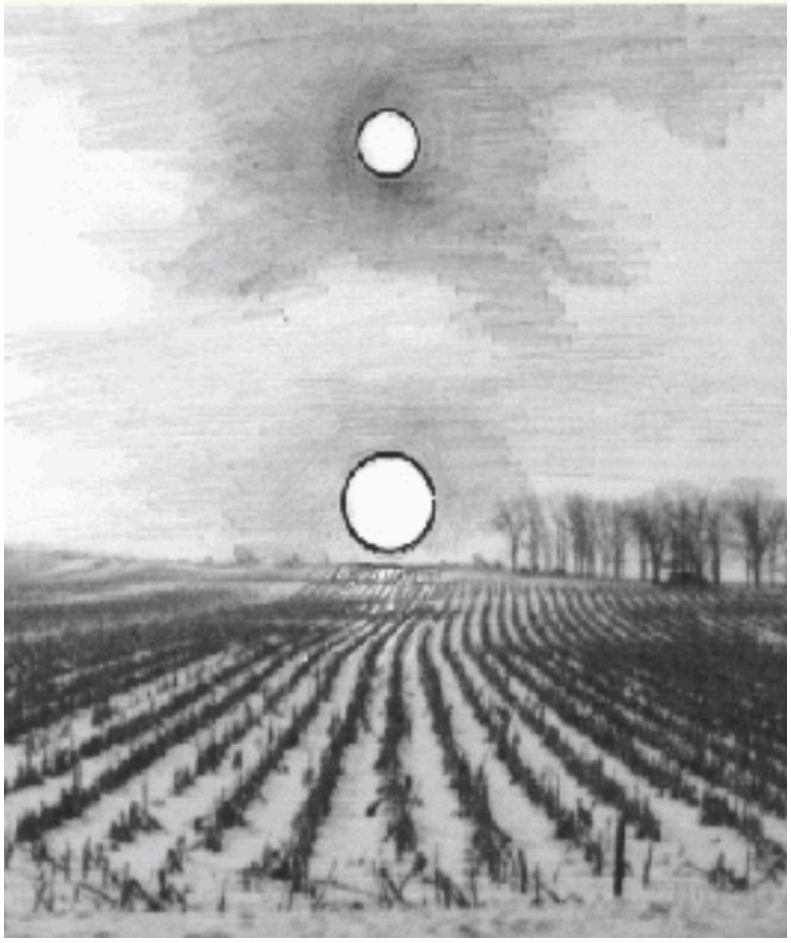


Ibn Al-Haitham is less known for a key significant contribution he made in his *Optics*.
He argued that "**perception was inferential**".

What we **observe** in "nature" may not be a phenomenon of nature.
It could be a phenomenon of **the human mind**.

The Moon Illusion





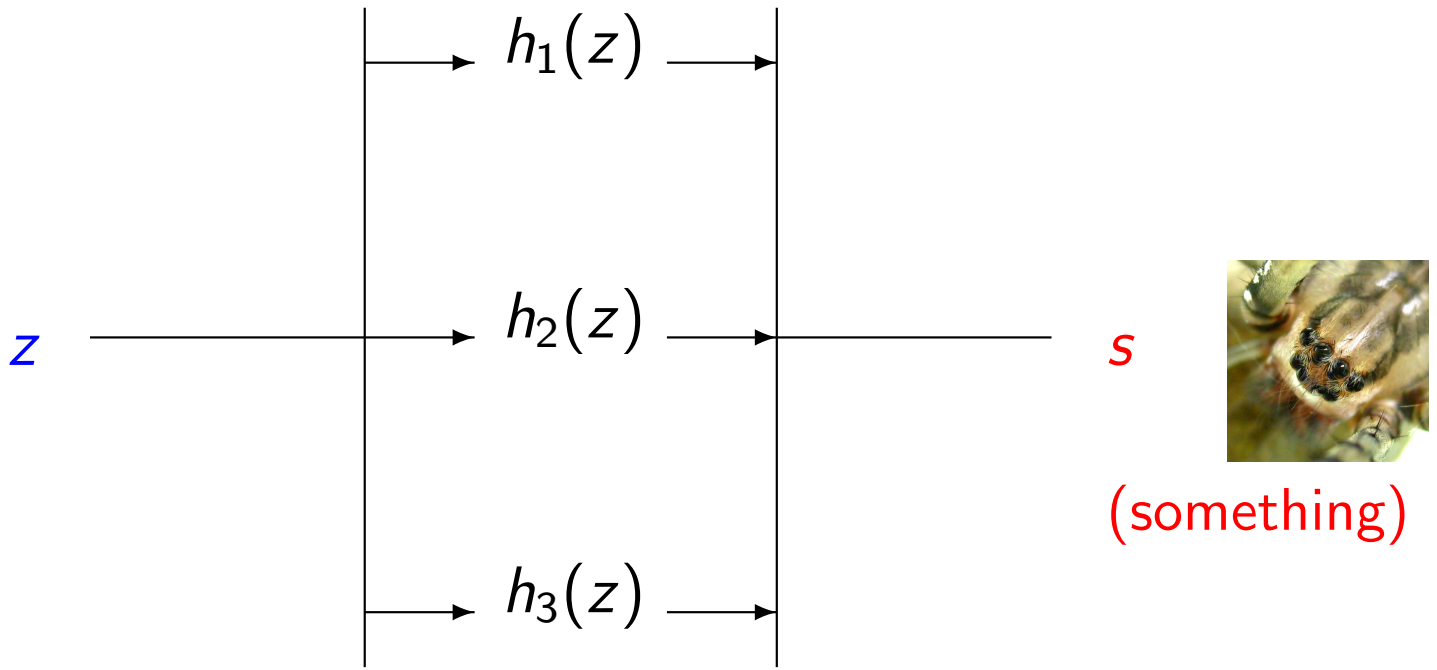
How Does it Work?

A possible explanation (*Angular Size-Contrast Theory*):

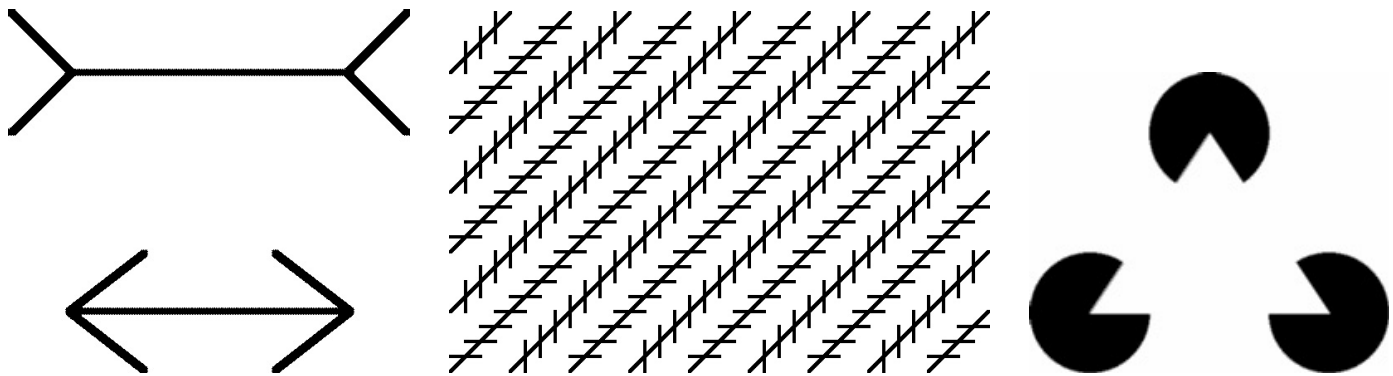
1. We can perceive the distance to the horizon (triangulation).
2. Objects at that distance are expected to be barely visible.
3. The moon is not barely visible.
4. Therefore, the moon must be really big. (Interpreation)

How does it work?

observation abstraction **perception**



Today, Ibn Al-Haitham's model is taken for granted. Many other optical illusions have been discovered and popularized.

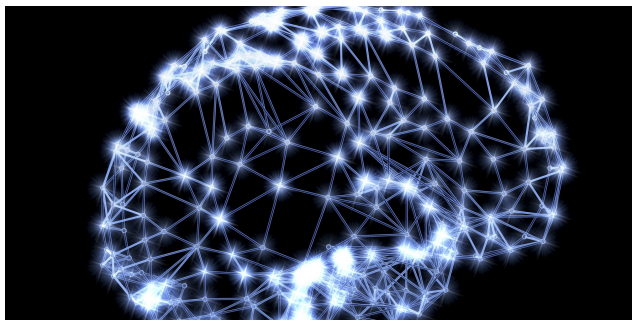


Why?

Q: Why does the brain create many layers of abstractions?

A: To extract meaningful patterns that will continue to hold in the future. This is called **generalization**.

- It does not help to memorize the past because experience will never be repeated exactly.



Generalization is a key goal in any machine learning algorithm.

1. computer vision
2. intrusion detection
3. spam filtering
4. demand forecasting
5. ...

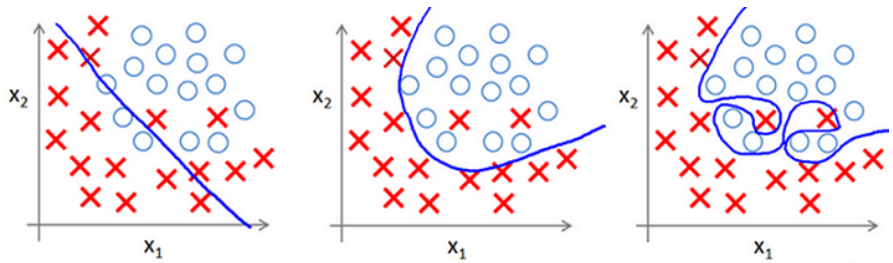
Conclusions based on experience may not always generalize:

- E.g. Fear of "Friday the 13^{th}"

What are the necessary and sufficient conditions for generalization?

Answering this questions will have applications for:

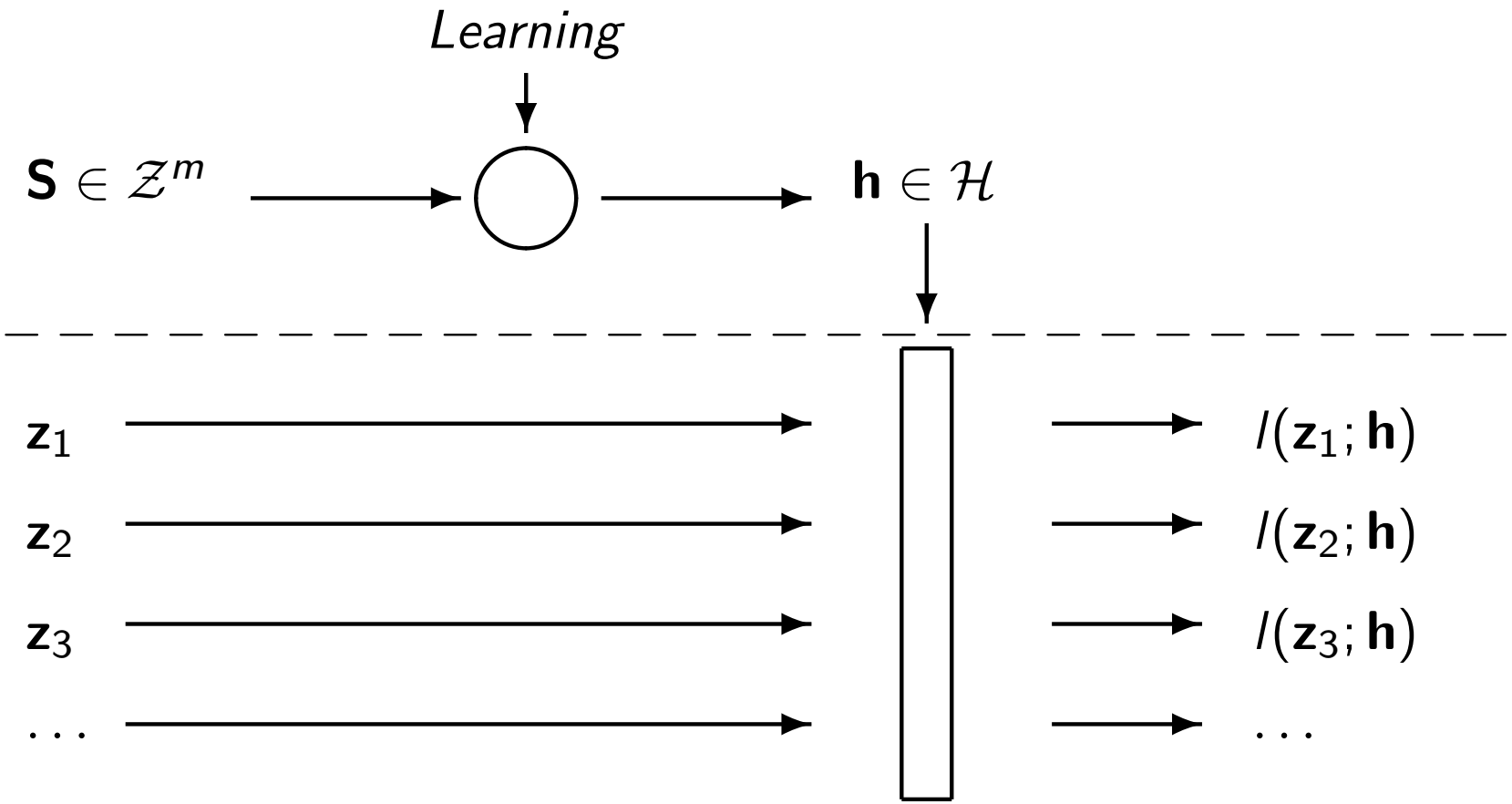
- *Quantifying Uncertainty*
- *Performing Model Selection*
- *Developing Better Machine Learning Algorithms*
- *Gaining Qualitative Insight*
- . . .



Learning and Information

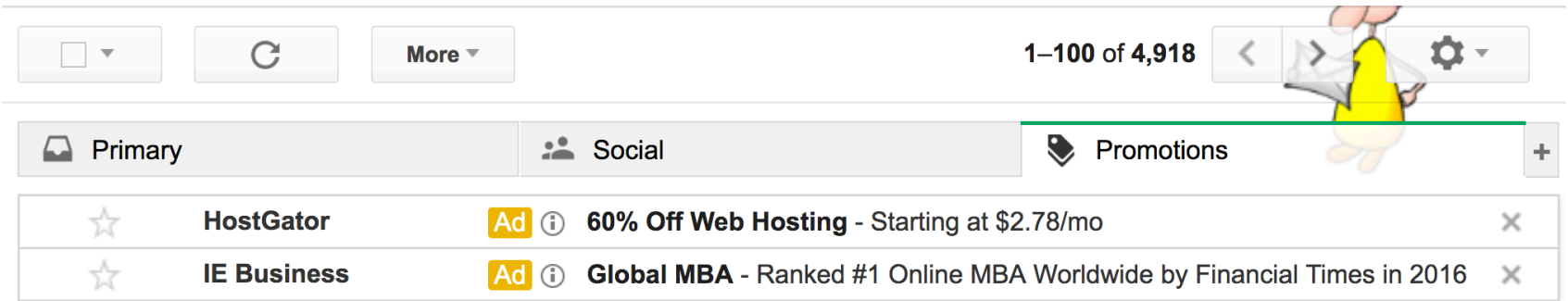
Mathematical Formalism

$$\textit{Generalization} = \textit{Future} - \textit{Past}$$



Example 1

Consider a classification task

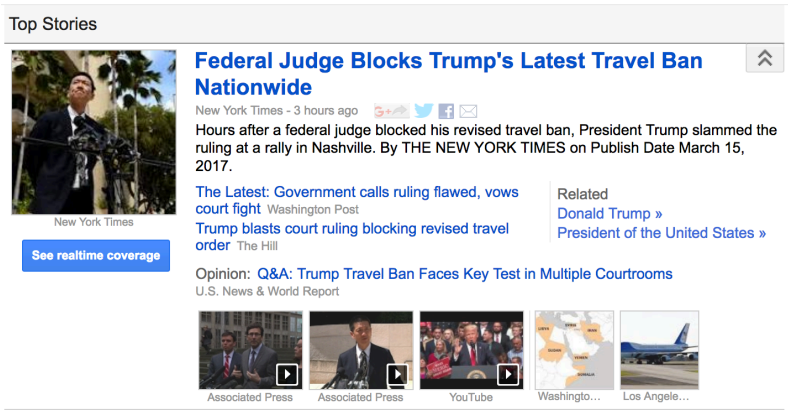


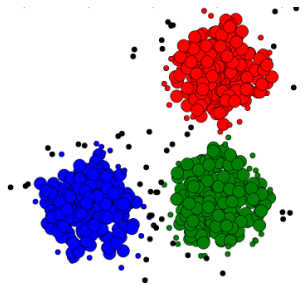
We want to **minimize** the misclassification error rate.

- Early systems used expert advice to create rules manually. Did not work well! In fact, experts disagreed.
- Optimization based on data solved the problem successfully.

Example 2

Consider a clustering task (e.g. location assignment, market segmentation, news aggregation, topic discovery, community detection)

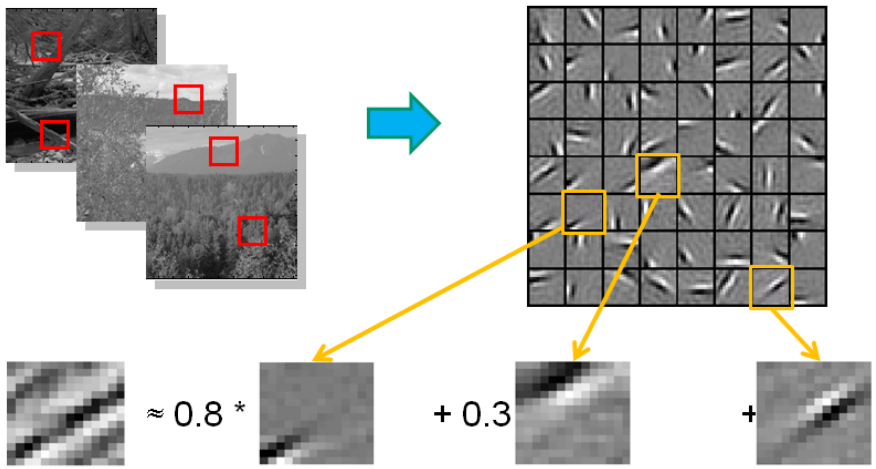




We want to **minimize** the intra-cluster distances and **maximize** the inter-cluster distances.

Example 3

Consider a sparse coding task:



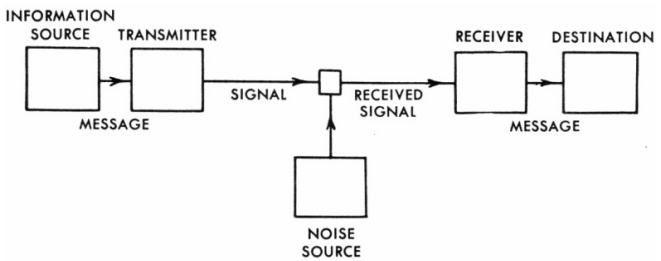
(image credit: Andrew Ng)

- Denoising and Compression

We want to **minimize** the description length and **minimize** the information loss.

Information-Theoretic Approach

A learning algorithm as a **channel** from $\mathcal{Z}^m$ to $\mathcal{H}$. From this, we define a "learning capacity" $C(\mathcal{L})$ for machine learning algorithms.



- Channel capacity is not restricted to "communication" channels *per se*. More generally, it is about **random variables**.
- Given two random variables $\mathbf{x}$ and $\mathbf{y}$, the conditional distribution $p(\mathbf{y} \mid \mathbf{x})$ defines a channel.

It can be shown that $C(\mathcal{L})$ captures all of the contributions to the risk of over-fitting.

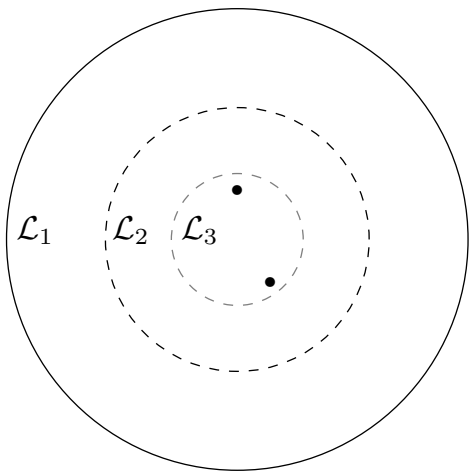
- **Finite Domain**:
  If $|\mathcal{Z}| < \infty$, then $C(\mathcal{L}) \lesssim \sqrt{\frac{|\mathcal{Z}|-1}{2\pi m}}$.
- **Finite Hypothesis Space**:
  If $|\mathcal{H}| < \infty$, then $C(\mathcal{L}) \leq \sqrt{\frac{\log|\mathcal{H}|}{2m}}$.
- **Differential Privacy**:
  If $\mathcal{L}$ is (ϵ, δ) differentially private, then $C(\mathcal{L}) \leq (e^{\epsilon} - 1 + \delta)/2$.
- **Stochastic Convex Optimization**: $C(\mathcal{L}) \leq \sqrt{d/(2m)}$.

Qualitative Insight

- **Stability**: A learning algorithm has a small learning capacity if and only if it is algorithmically stable.
- **Information**: A learning algorithm has a small learning capacity if and only if its hypothesis does not reveal a lot of information about the training sample.
- **Composition**: Learning "more" about the training sample, e.g. $(\mathbf{h}_1, \mathbf{h}_2)$, increases the risk for overfitting, compared to $\mathbf{h}_1$.
- **Universality**: A problem is learnable *if and only if* it is learnable via an algorithm with a vanishing learning capacity.

- **Data Processing Inequality**: If $\mathbf{s} \to \mathbf{h}_1 \to \mathbf{h}_2$, then the risk of overfitting in $\mathbf{h}_2$ is smaller.
  - E.g. decision tree pruning and sparsification.
  - Induces a partial order on learning algorithms.

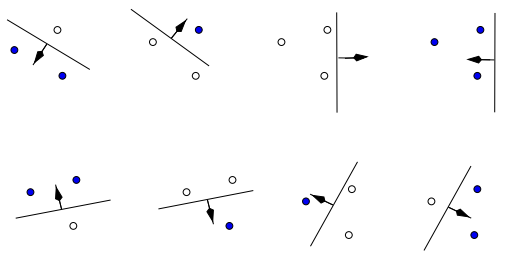
$$\mathcal{L}_2 \subseteq \mathcal{L}_1 \qquad \Leftrightarrow \qquad \mathbf{s} \to \mathbf{h}_1 \to \mathbf{h}_2$$



Equivalence Result

In Learning Theory: The **VC dimension**.

- Agnostic PAC learnability $\equiv$ finite VC dimension.



In Information Theory: The **Shannon channel capacity**.

- Highest achievable information rate with arbitrarily small error probability.

Note: An equivalence relation exists between the two quantities if the channel capacity is measured in the total variation distance.

Applications

Large Scale Optimization

$$\textit{Accuracy} \quad = \quad \textit{Empirical Risk} \quad + \quad \textit{Generalization Risk}$$

- Stochastic convex optimization can be parallelized trivially (e.g. SVM, logistic regression, least squares, ...etc).

Algorithm	Test Error	Training Time
ERM ($m = 50,000$)	$14.8 \pm 0.3\%$	2.93s
ERM ($m = 1,000$)	$15.3 \pm 0.5\%$	0.03s
Parallelized ($K = 50, m = 1,000$)	$14.8 \pm 0.1\%$	0.03s

Table: ℓ_2 regularized logistic regression on the MiniBooNE Particle Identification problem.

Model Selection

Uniform Information Criterion: Model selection according to $UIC = \mathbb{E}_{\mathbf{z}\sim S}[l(\hat{\mathbf{h}}, \mathbf{z})] + \sqrt{\frac{d}{2m}}$, where d is the number of learned parameters.

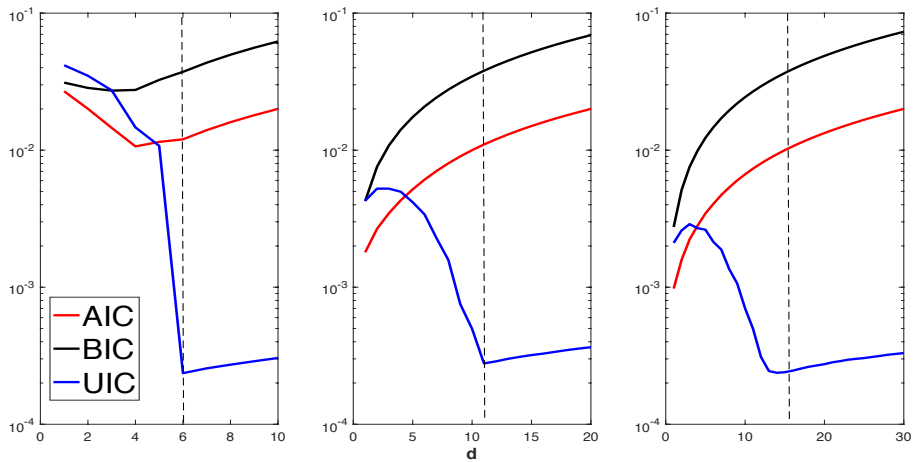


Figure: Least-squares polynomial regression where $\mathcal{X} = [-1, +1]$, $\mathbf{y} = f(\mathbf{x}) + \epsilon$, f is a polynomial of degree $d^\star$ (marked by dashed line), and ϵ is white Gaussian noise.

Future Research Directions

Adaptive Learning

A typical machine learning project involves multiple rounds of analysis (e.g. feature selection, normalization, cross validation, trying multiple algorithms, ...etc). The chain rule provides a simple recipe for analyzing the generalization risk.

- Given the analysis conducted (e.g. in the form of a code or a flowchart), can we design a procedure for estimating the generalization risk?

Information Criterion

Information criteria (such as BIC), are sometimes used in the unsupervised learning setting for model selection (e.g. in k-means clustering).

- Given that uniform generalization is developed in the *general* setting of learning, should the learning capacity serve as a model selection criterion in the unsupervised setting? Why or why not?

References

Learning Capacity

- Alabdulmohsin, "Algorithmic Stability and Uniform Generalization." NIPS, 2015.
- Alabdulmohsin, "An information-theoretic route from generalization in expectation to generalization in probability." AISTATS, 2017.
- Alabdulmohsin, "Information-Theoretic Guarantees for ERM with Applications to Model Selection and Large-Scale Optimization." ICML, 2018.

The Shannon Mutual Information and Adaptive Learning

- D Russo and J Zou. "Controlling bias in adaptive data analysis using information theory." AISTATS, 2017.

Generalization via Differential Privacy

- C Dwork, V Feldman, M Hardt, T Pitassi, O Reingold, and A Roth. "Preserving statistical validity in adaptive data analysis." STOC 2015.

Generalization based on leave-one-out mutual information

- M Raginsky, A Rakhlin, M Tsao, Y Wu, and A Xu. "Information-theoretic analysis of stability and bias of learning algorithms." ITW, 2016.
- A Xu, M Raginsky. "Information-theoretic analysis of generalization capability of learning algorithms." NIPS 2017.