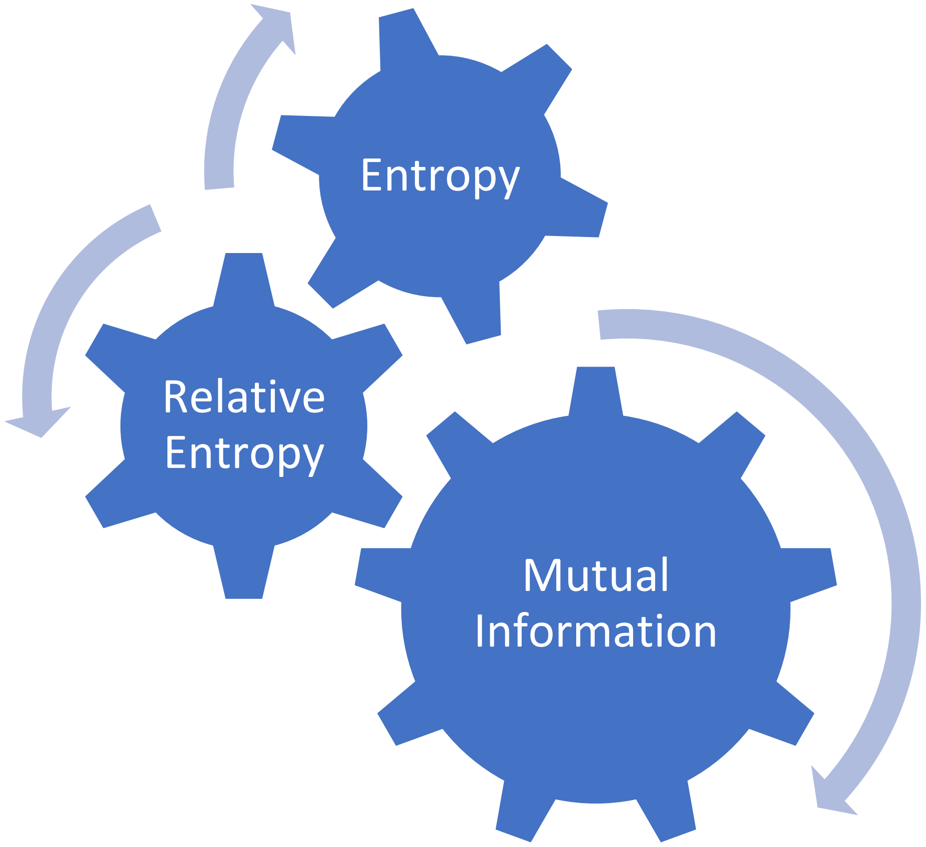
Brownbag on Mutual Information

Demystifying the "**Correlation of the** 21st **Century**." Terry Speed.

Matthias Hermann, HTWG-Konstanz, Institute for Optical Systems
30.11.2020

Overview

- Tishby‘s hypothesis
- Other applications
- Background on 5 slides
- Algorithms
- Flourish

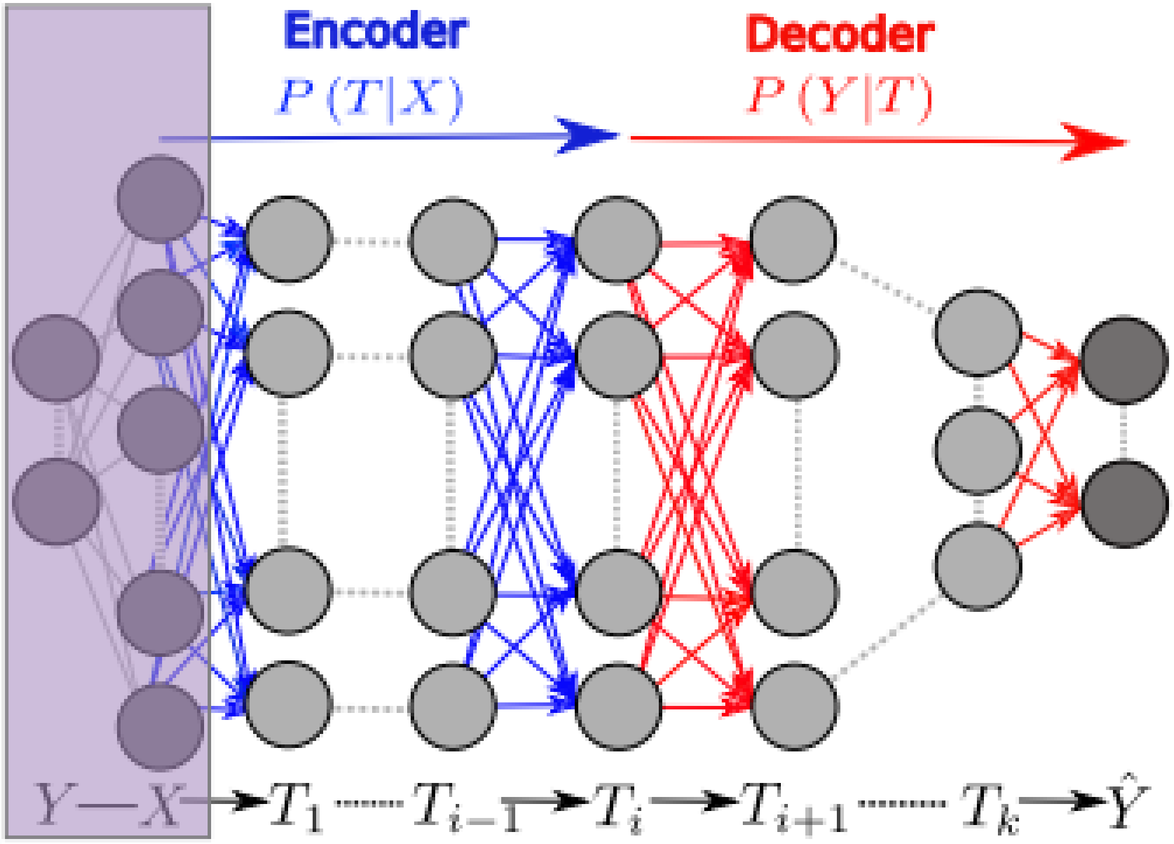


Tishby’s hypothesis

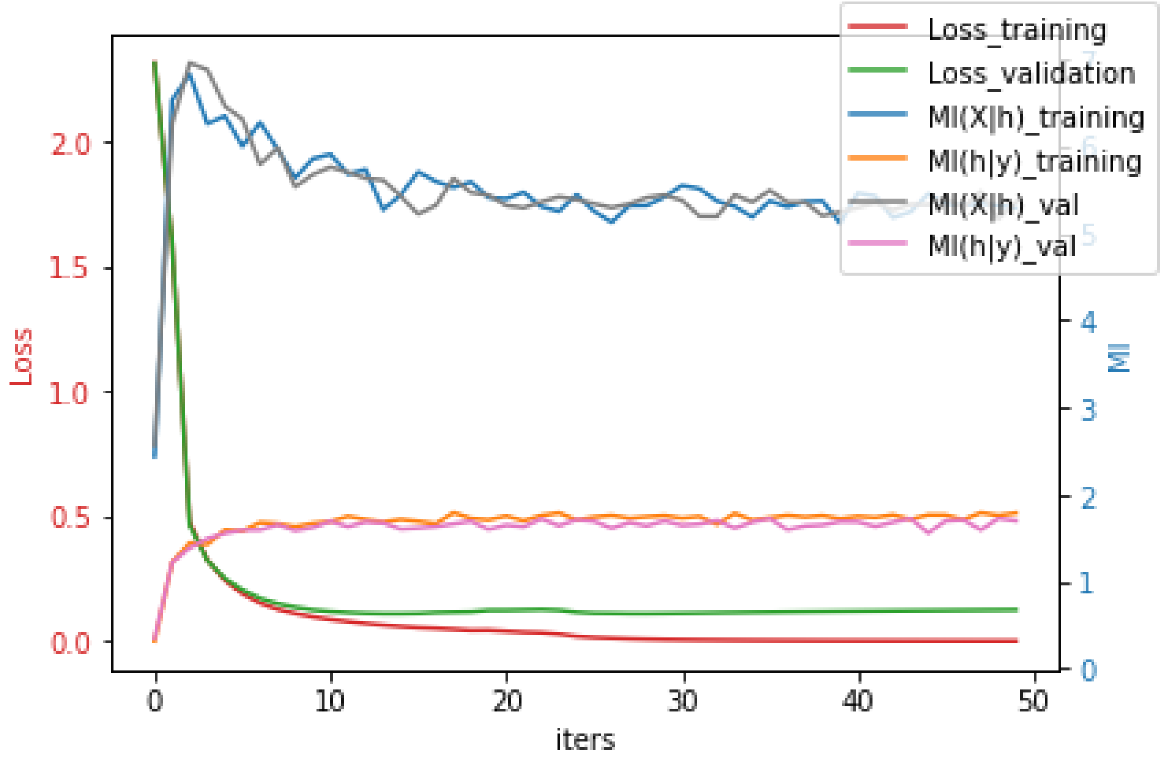
Naftali Tishby's information bottleneck for neural networks

Remember from compression:
Compression means minimizing MI(X,T) .

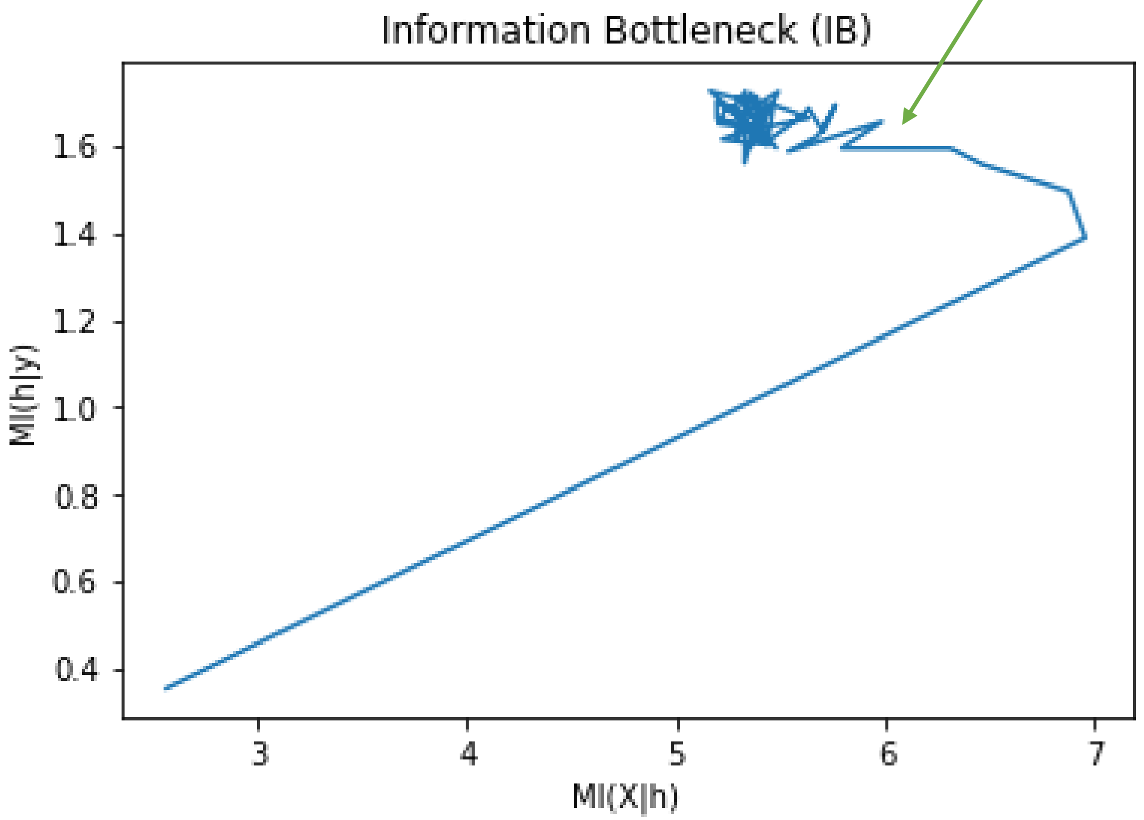
Tishby's hypothesis:
"A neural network learns by compressing the input optimally under the constraint given by the Loss($\widehat{Y}, Y$*)."*



The information bottleneck with MNIST

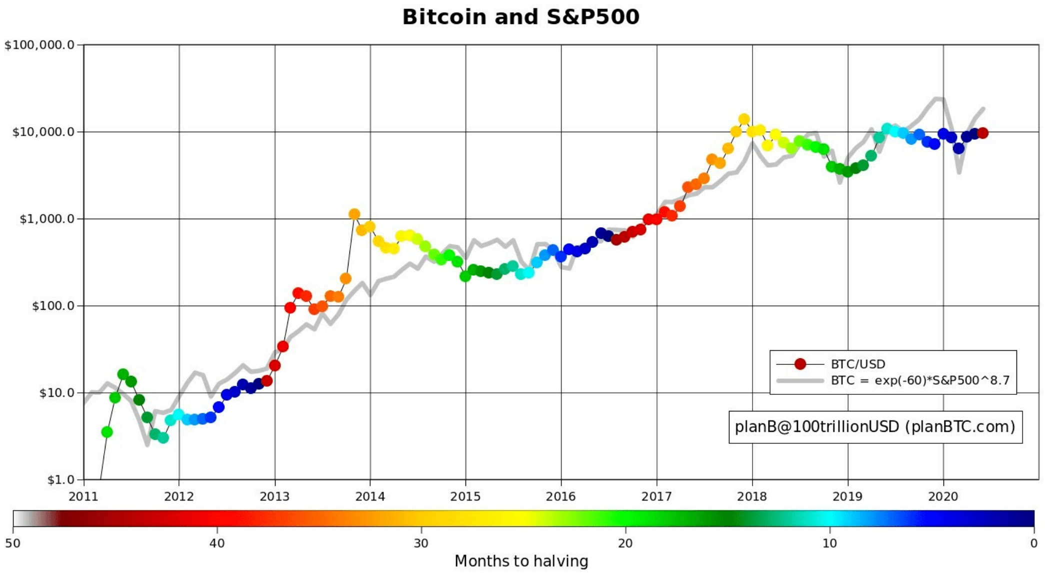


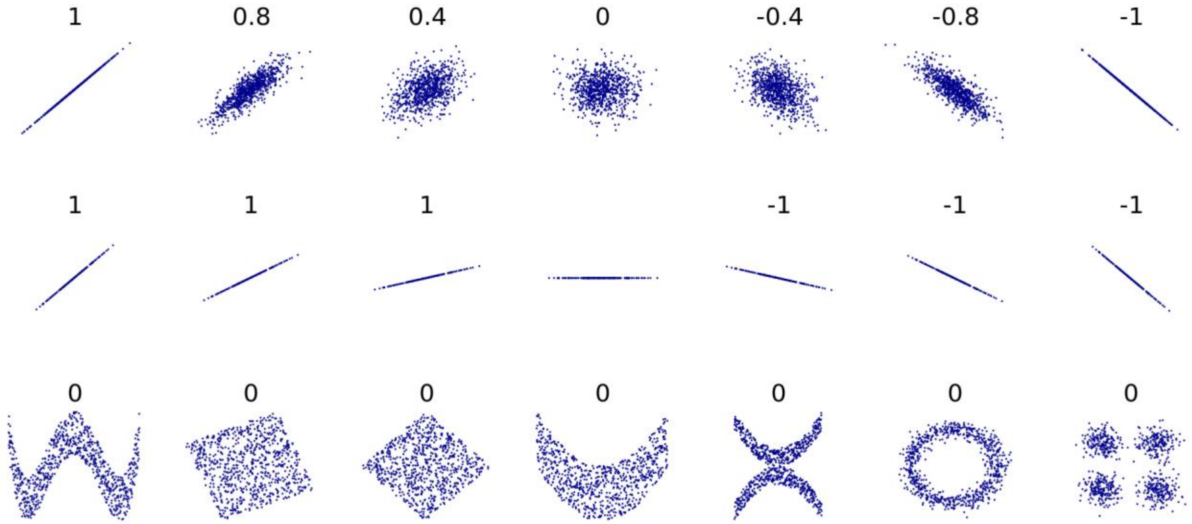
Compression appears!



Applications

Measuring correlation





Bitcoin X and S&P500 Y is highly correlated.

Pearson correlation coefficient measures linear relationships between X and Y. But fails with slope and non-linear relationships.

Mutual Information measures non-linear correlation.

Measuring compression rate

- Compression is a tradeoff of quality and needed bits.
- We usually want a code that is maximum independent of the input.
- Mutual information measures non-linear relationships (shared information).
- Mutual information between input and compressed code must be as small as possible.

Original JPEG JPEG2000

The **Input** should **not** be **predictable** given the code.

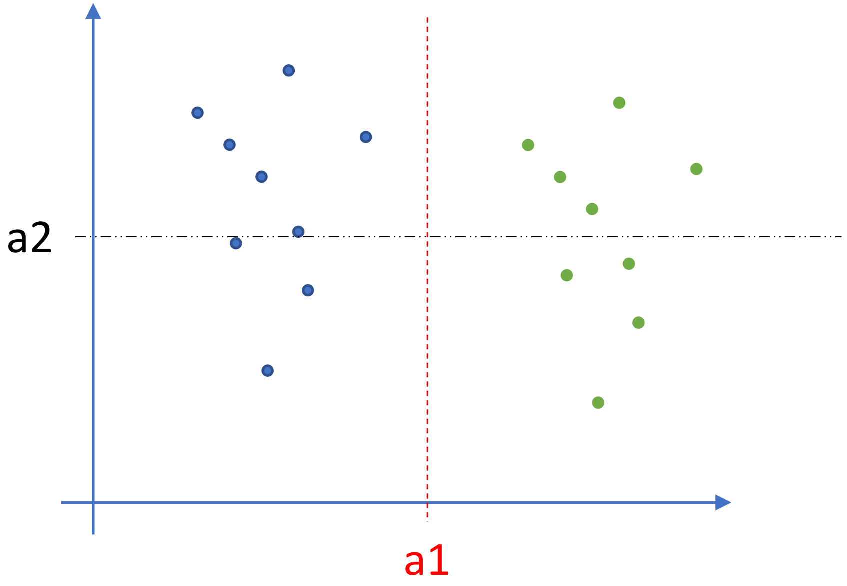
Measuring information gain

- The expected value of information gain is the mutual information

$$E_{a \sim A}[IG(X, a)] = I(X, A)$$

- Information gain

$$IG(X, a) = H[X] - H[X|a]$$



<u>**When we want to maximize IG:**</u>

Should we choose a2 or a1?

Criterion in decision trees

- Decision trees typically optimize mutual information criterion.
- Measuring mutual information of low dimensional categorial variables is easy.

Measuring $I(X,Y)$ of continuous high dimensional variables is difficult.

Measuring uncertainty

Bayesian Active Learning for Classification and Preference Learning

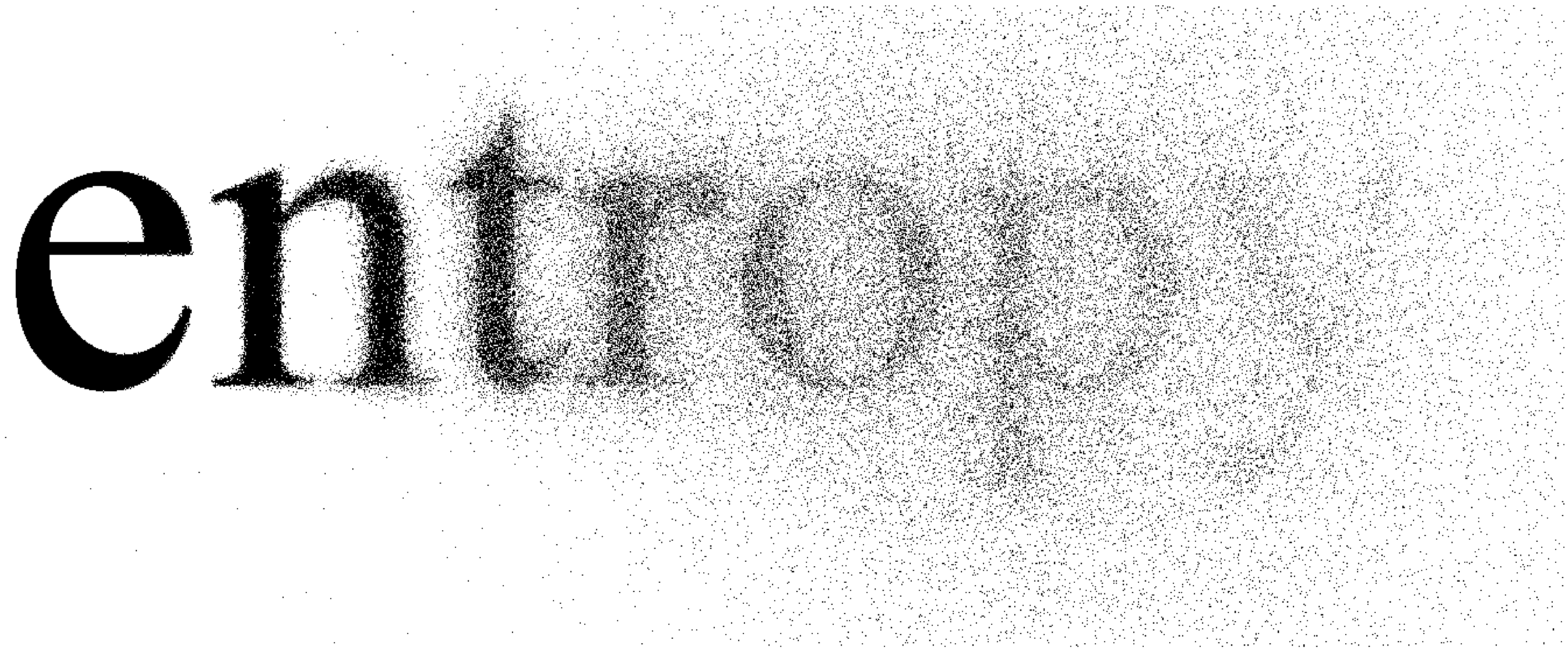
Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, Máté Lengyel
Computational and Biological Learning Laboratory
University of Cambridge

December 30, 2011

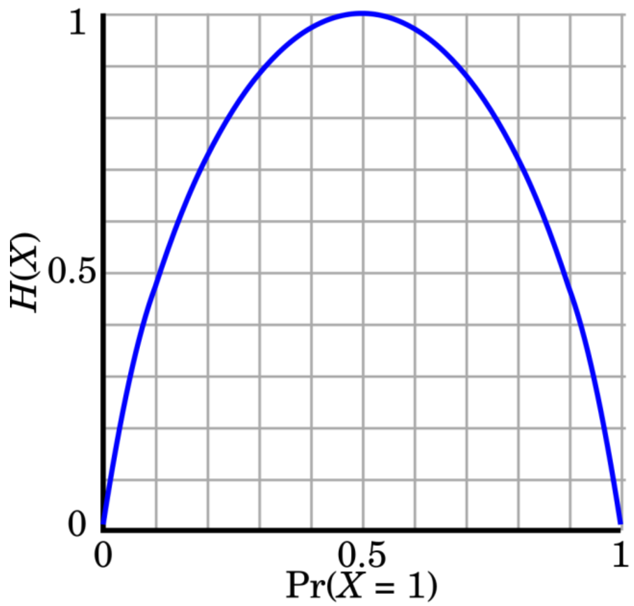
Bayesian **A**ctive **L**earning by **D**isagreement (BALD).

6 slides to go..

Background


ent

Entropy



$$\operatorname*{argmax}_{p} H[Bernoulli(p)] = 0.5 \;\Rightarrow H[Bernoulli(0.5)] = \; 1\; bit$$

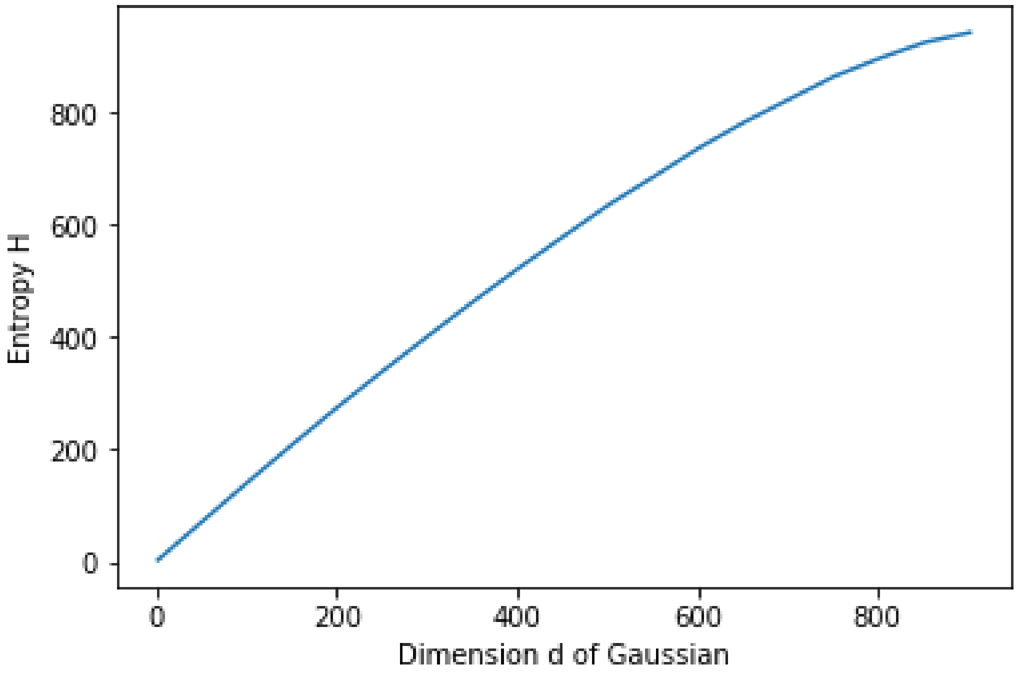
Entropy

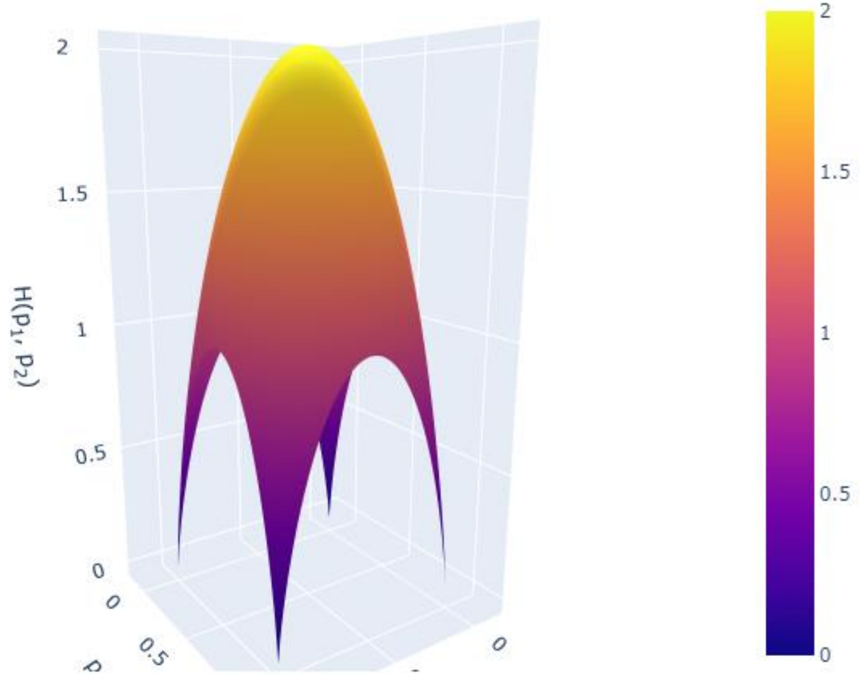
$$H[X] = H[p(x)] = E_{x \sim X}[-\log p(x)]$$

- Entropy of a variable X depends on p(x)
- Maximum entropy distribution for continuous variables on [-inf, +inf] with fixed variance is Gaussian
- Maximum entropy distribution for discrete variables is uniform
- Entropy is measured in bits or nats

The **better** the model fits, the **lower** the entropy under that model.

Entropy increases with dimension





Entropy of a Gaussian variable with increasing dimension (left) and a two-dimensional Bernoulli variable (right).

Properties of KL divergence/relative entropy

- The KL-divergence represents the number of extra bits needed when using Q instead of P.

$$KL[P|\,Q] = \mathrm{CE}[\mathrm{P}, Q] - H[P]$$

- Non-negative
- Dimensionally consistent i.e. invariant under parameter transformations.
- Classical loss for optimizing classifier $f_\theta(x)$:

$$\hat{y} = \log softmax(\frac{e^{f_\theta(x_i)}}{\sum_j e^{f_\theta(x_j)}})$$

$$CE[p(y), p(\hat{y})] = \sum_{y_j} y \log \hat{y}_j$$

$$KL[p(y)|p(\hat{y})] \propto CE[p(y), p(\hat{y})]$$

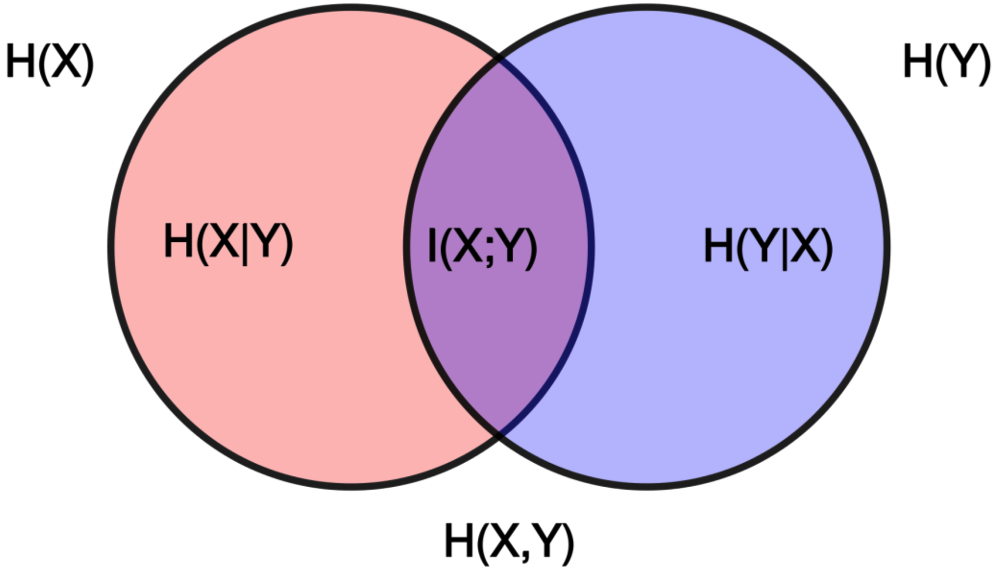
Mutual Information

Entropies and Conditional Entropies.

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

KL divergence between joint and marginal.

$$I(X,Y) = KL[p(x,y)|p(x)p(y)]$$



Venn-diagramm for Mutual Information I(X,Y).

Algorithms

Measuring uncertainty with BALD

I(X,Y) = H[X] - H[X|Y]
= H[Y] - H[Y|X]

BALD Idea: Propose example x that greedily maximizes the decrease in posterior entropy:

$$I(\theta, Y|x) = H[\theta] - H(\theta|Y, x)$$

$$\arg\max_x H[\theta] - H[\theta|Y] = H[\theta] - E_{y\sim p(y|x)}\big[H[\theta|y, x]\big]$$

$$\arg\max_x H[p(y|x)] - E_{\theta\sim q(\theta)}\big[H[p(y|x, \theta)]\big]$$

Estimating entropy H[Y|X] is easy as we do know that p(y|x) is multinomial.

Levels of concern

- Maximizing mutual information $\mathrm{I}(X|f_\theta(x))$
- Estimating entropy H[X] = E$[-\log p_\theta(x)]$
- Minimizing mutual information $\mathrm{I}(\mathrm{X}, Y)$
- Estimating joint entropy H[A,…,Z]
- Estimating I(X,Y)
- Maximizing KL-divergence $KL(p_\theta(y)|p_\theta(x))$
- Minimizing KL-divergence $KL(p(y)|p(f_\theta(x)))$

- Autoencoder
- Density estimation
- Decorrelation
- Absolute measuring
- Absolute measuring
- Contrastive learning
- Regression and classification

Estimation Dual KL-Divergence

- Donsker-Varadhan representation of KL (e.g. contrastive learning):

$$Dual\ KL[P|Q] > \sup_{T} E_P[T] - \log(E_Q[e^T])$$

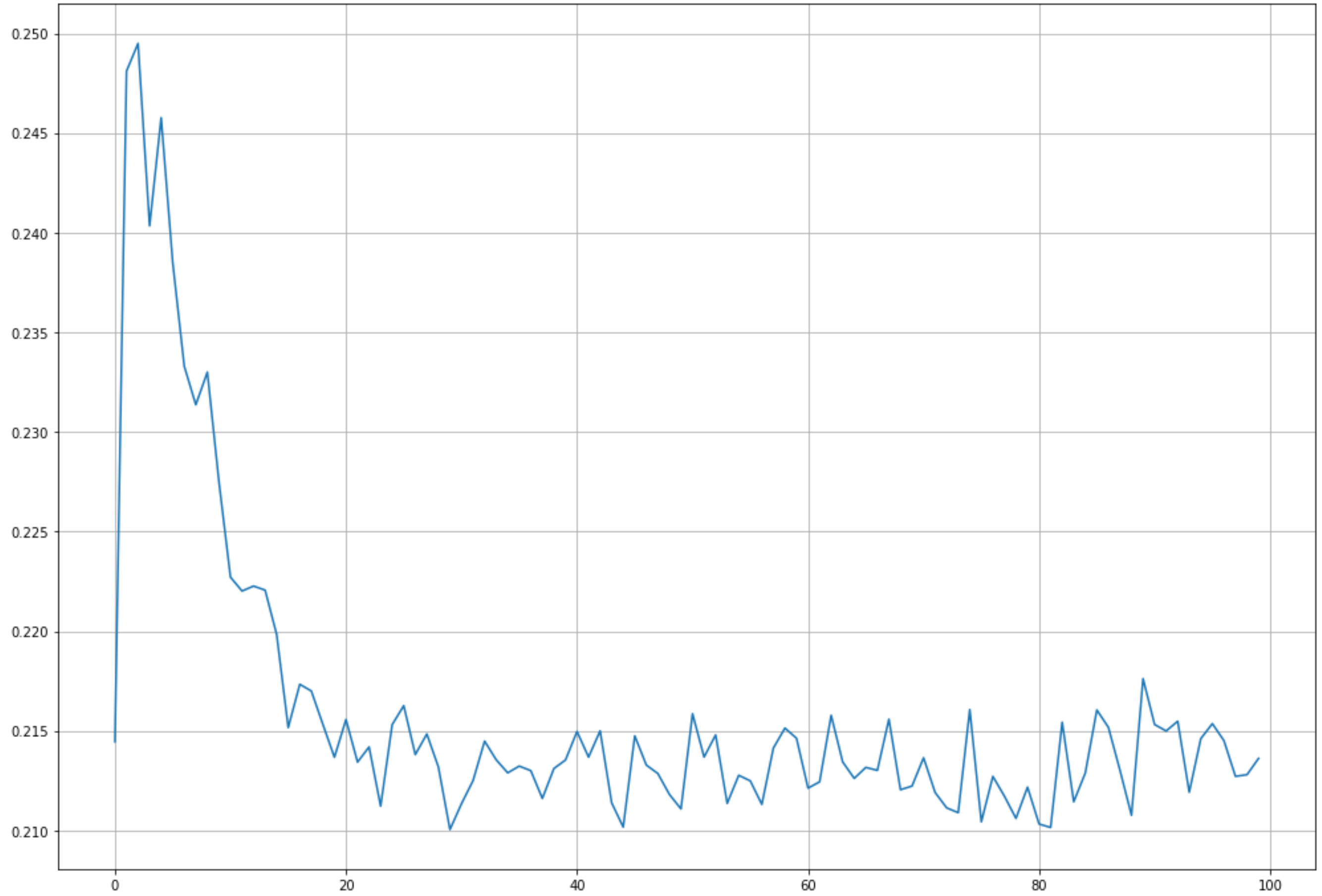
- Maximizing (lower bound) mutual information (neural estimation):

$$Dual\ KL[X|Y] \geq \sup_{f_\theta} E_{x\sim X}[f_\theta(x)] - \log(E_{y\sim Y}\left[e^{f_\theta(y)}\right])$$

$$I_{neural}(X,Y) \geq \sup E_{x^+\sim(X,Y^*)}[f_\theta] - \log(E_{x^-\sim(X,Y)}\left[e^{f_\theta}\right])$$

- Auxilary dataset $\mathrm{D}^- = (X, Y^*)$ by sampling y^* without replacement.

Negentropy over epochs



References

- Tishby, Naftali; Pereira, Fernando C.; Bialek, William (September 1999). The Information Bottleneck Method (PDF). The 37th annual Allerton Conference on Communication, Control, and Computing. pp. 368–377.
- Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., & Hjelm, D. (2018, July). Mutual information neural estimation. In International Conference on Machine Learning (pp. 531-540)
- Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein gan. arXiv preprint arXiv:1701.07875.
- https://en.wikipedia.org/wiki/Correlation_and_dependence
- http://www.ncbi.nlm.nih.gov/pubmed/8947780
- https://de.wikipedia.org/wiki/JPEG_2000
- https://en.wikipedia.org/wiki/Information_gain_in_decision_trees
- Houlsby, N., Huszár, F., Ghahramani, Z., & Lengyel, M. (2011). Bayesian active learning for classification and preference learning. arXiv preprint arXiv:1112.5745.
- https://homes.cs.washington.edu/~ewein/blog/2020/07/30/joint-entropy

Thanks.

Matthias Hermann

Hochschule Konstanz
Institute for Optical Systems

matthias.hermann@htwg-konstanz.de

www.ios.htwg-konstanz.de