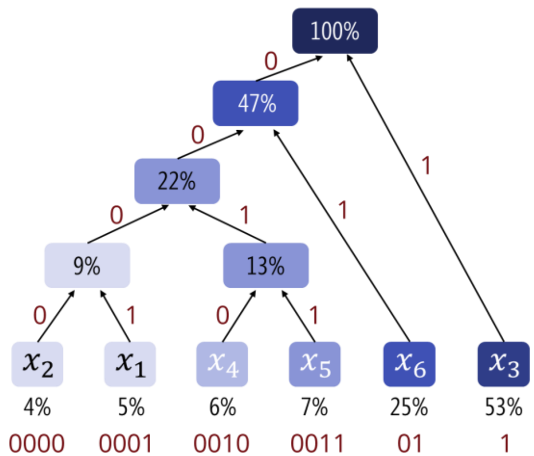
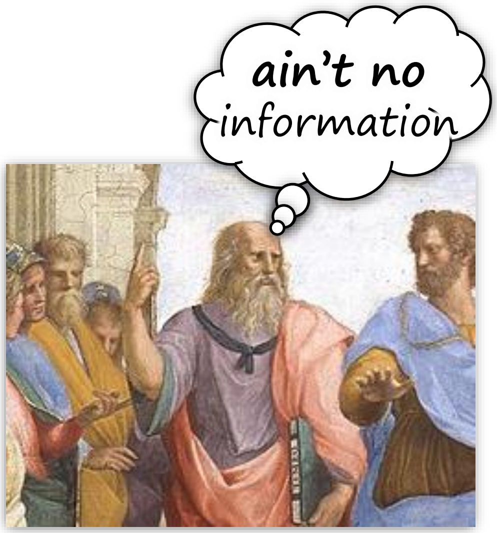
Modelling 2

STATISTICAL DATA MODELLING




0111001101001010


1110110111001111



Chapter 6

Information

Michael Wand · Institut für Informatik, JGU Mainz · michael.wand@uni-mainz.de

Video #06

Information Theory

- **Information & Entropy**
- **Algebra & Applications**

Information Theory

0111001101001010

1110110111001111

LITERATURE:

Massimiliano Tomassoli: *Information Theory for Machine Learning*
https://github.com/mtomassoli/papers/blob/master/inftheory.pdf, 2016.

David MacKay: *Information Theory, Inference, and Learning Algorithms*.
Cambridge University Press, 2003.

What is Information?

Defining Information

- Probability Theory
- Randomness = genuine new information

How much Information?

- Answer: “How random?”

Axioms of Information

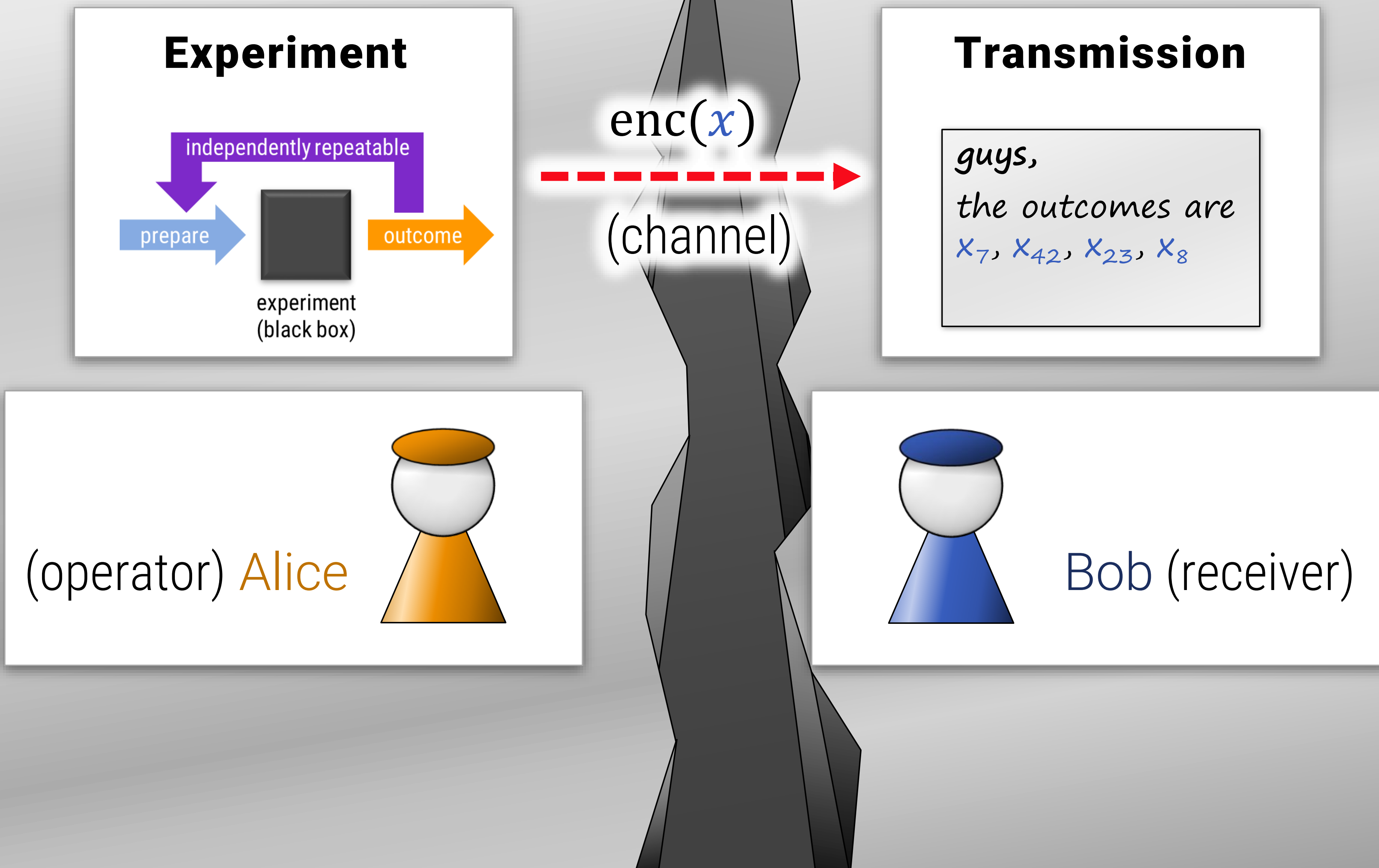
Random Information

- Random variable X
- Discrete probability distribution $p(x)$

Information

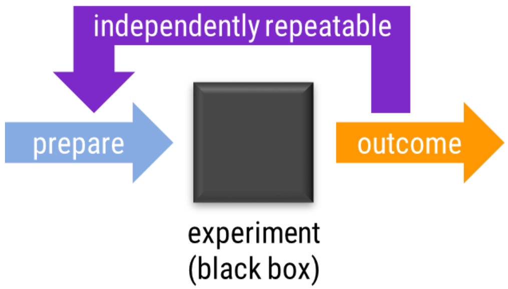
- $I(x)$ – Information contained in observation of x

“Frequentist” Model of Information



“Frequentist” Model of Information

Experiment



enc(x)

(channel)

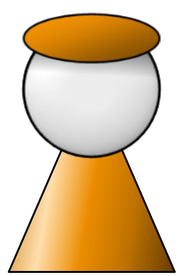
Transmission

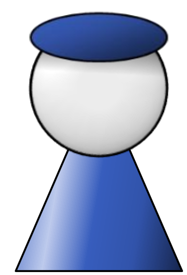
guys,

the outcomes are

x_7, x_{42}, x_{23}, x_8

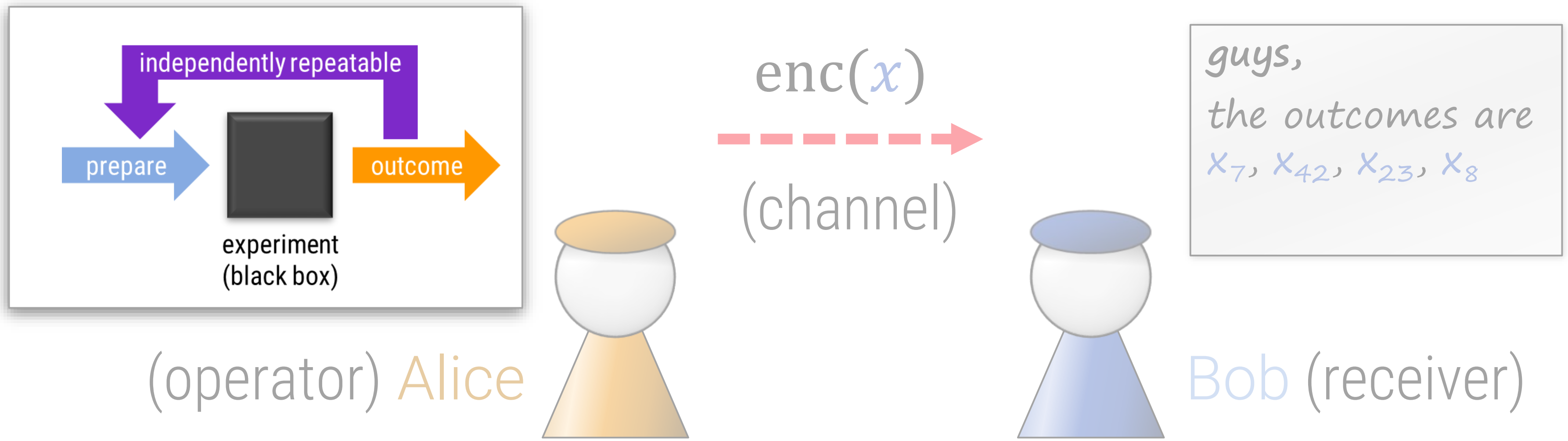
(operator) Alice





Bob (receiver)

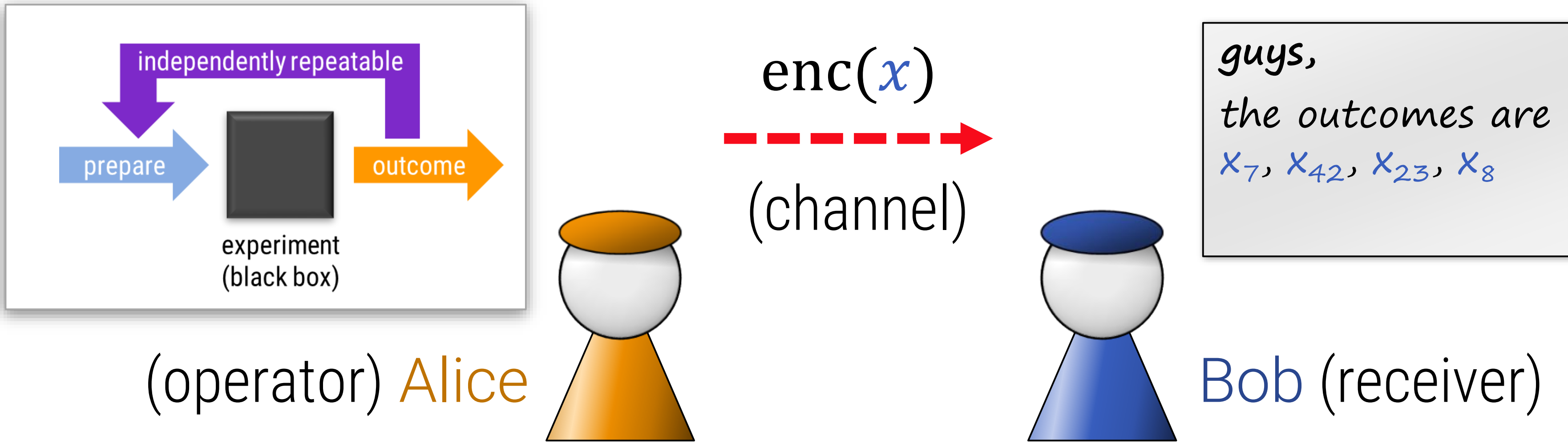
“Frequentist” Model of Information



How to understand information?

- Repeatable experiment
  - Outcomes $\Omega = \{x_1, \dots, x_n\}$
  - Elementary probabilities $p(x_1), \dots, p(x_n)$

“Frequentist” Model of Information

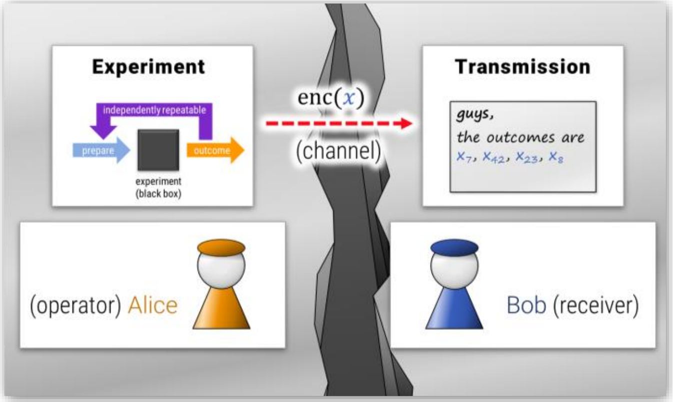


Communication

- Send outcomes over channel from Alice → Bob
  - Alice runs experiments
  - Both Alice and Bob know the experimental setting
  - Bob does not know the random outcome
- How much information is in outcome x_i?

Defining Information

Axioms of Information



Axioms

- $I(x) = f(p(x))$ for some f
  - Information should only depend on probability
- $p(x) < p(y) \Rightarrow f\big(p(x)\big) > f\big(p(y)\big)$
  - Rarer events should carry more information
  - f strictly decreasing
- $f(1) = 0$
  - Certain events carry no (new) information
- x, y independent $\Rightarrow I\big((\mathrm{x}, \mathrm{y})\big) = I(\mathrm{x}) + I(\mathrm{y})$
  - Information should add up
  - Independent experiments yield “totally new information”

Solution

Solution

$$f(p) = -\log p = \log \frac{1}{p}$$

this solution is unique (up to basis)

PS: we will usually use $\log_2$

Proving the properties

- $I(x) = \log \frac{1}{p(x)}$
- $p(x) < p(y) \Rightarrow \log \frac{1}{p(x)} > \log \frac{1}{p(y)}$
- $\log 1 = 0$
- x, y independent $\Rightarrow \log \frac{1}{p(x,y)} = \log \frac{1}{p(x)p(y)}$

$$= \log \frac{1}{p(x)} + \log \frac{1}{p(x)}$$

Summary so far…

Probability

- Independent events: Product of probabilities
- Number between 0 and 1

Information

- Information is additive
  - More info: larger value
  - No information = 0
- Information of event = negative logarithm of prob.
  - $I(x) = -\log p(x) = \log \frac{1}{p(x)}$
  - Usually: base 2 (measured in bits)

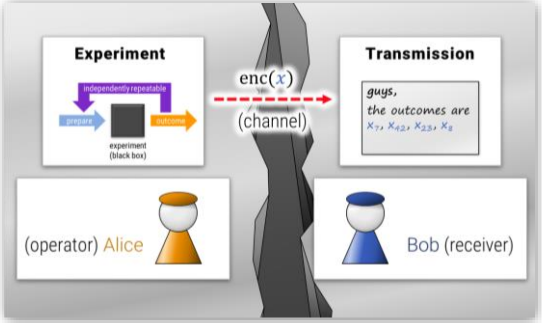
Neg-Log Likelihoods

quantify

Information Content

Next question: How much information is “in the whole distribution”?

Information in Outcomes



Alice observes an outcome x

- Alice needs to send Bob $I(x)$ bits

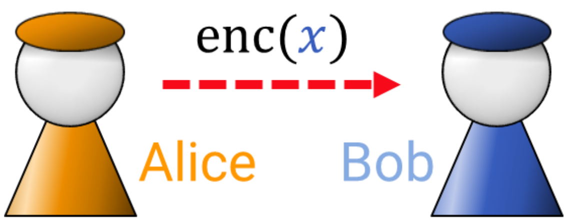
Alice observes an outcomes x, y

- Model: Two independent runs
- Alice needs to send Bob $I(x) + I(y)$ bits

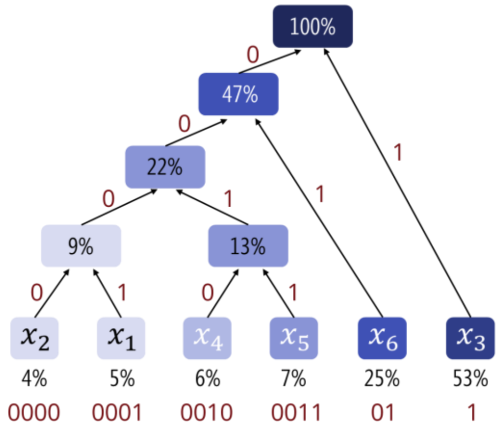
Alice keeps observing outcomes $x_1, x_2, x_3, \ldots$

- Model: independent repetitions
- Alice needs to send Bob $\mathbb{E}_{x\sim p}[I(x)]$ bits on average

Entropy



enc(x)
Alice
Bob



100%
0
47%
1
0
22%
1
0
9%
1
13%
0
1
0
1
x_2
x_1
x_4
x_5
x_6
x_3
4%
5%
6%
7%
25%
53%
0000
0001
0010
0011
01
1

Entropy

Definition: Entropy

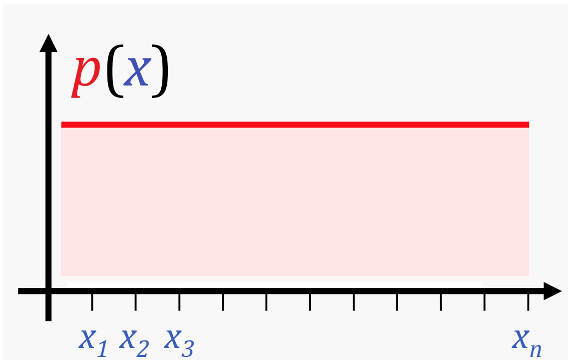
$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$ mean neg log prob

$$= \sum_{i=1}^{n} p(x_i) I(x_i)$$ mean information

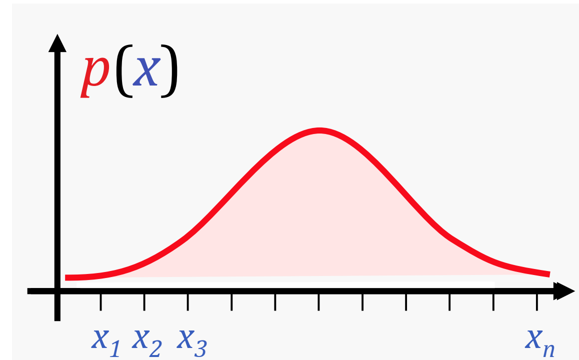
$$= \mathbb{E}_{x \sim p(x)}\big(I(x)\big)$$ expected information

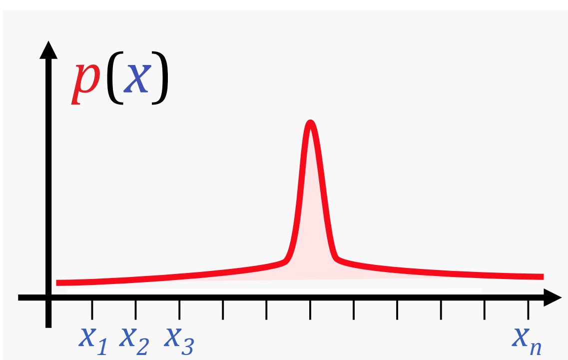
Measures: How "random" is p?

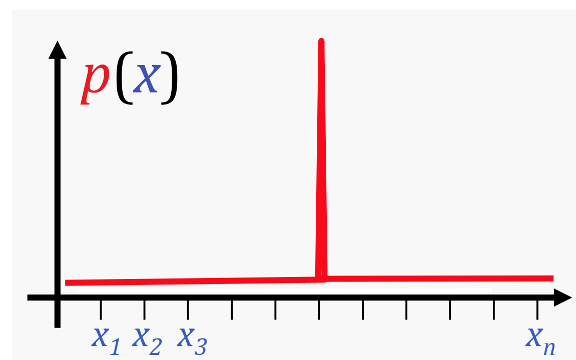
Examples



$$H(p) = -\sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n} = \log n$$







$$H(p) = 1 \cdot \log 1 = 0$$

Finite Outcome Spaces

Definition of $H(X)$ requires finite $\Omega(X)$

- Generalization to continuous variables non-trivial
- Just replacing $\sum$ by $\int$ leads to significant problems
  - Is done as „Differential Entropy“:
    $\sum_i p(x_i) \log p(x_i) \quad \rightarrow \quad \int_x p(x_i) \log p(x_i)$
- Problems include
  - Negative values (density > 1)
  - Not transformation invariant

Finite Outcome Spaces

“Proper” limit fixes some problems

- But entropy becomes infinite for continuous variable
- “Limiting density of discrete point”

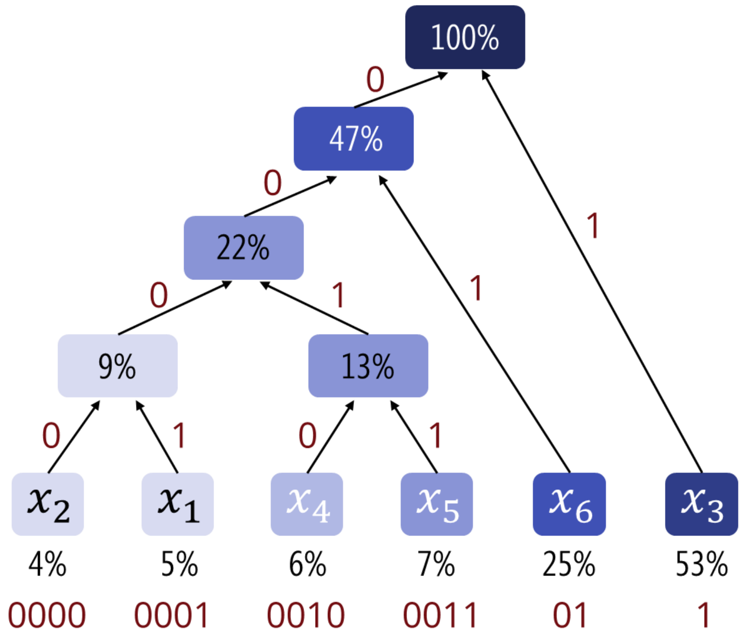
Coding length for continuous functions

- One can specify resolution limits / uncertainty
- Often no obvious resolution
  - Careful trade-off needed

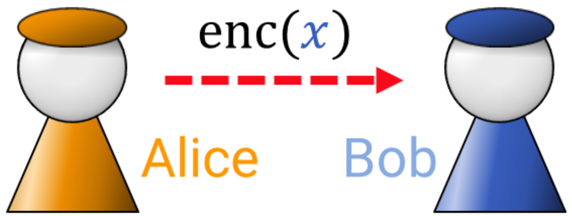
We stick to the discrete version

- Or just ignore the issue, when it comes up

Coding Theory



Coding Theory



Entropy

- Minimum number of bits required to transmit information about event x
  - We draw events i.i.d.
  - We send each outcome separately
    - After being asked for the answer
    - (Certain outcomes: no answer required)
- **Coding theorem**
  - $m(x)$ = message about x optimally encoded in bits
  - $H(X) \leq \mathbb{E}_{x \sim p(x)}\left(\text{length}\left(m(x)\right)\right) < H(X) + 1$

Random variable X distributed according to p(x)

Huffman Codes

Constructing a code

$\text{enc}(x)$

- Huffman algorithm
- Optimal for single events send in bits
  - Multiple symbols: Overhead up to one bit each
  - Optimality reached with “arithmetic coding”

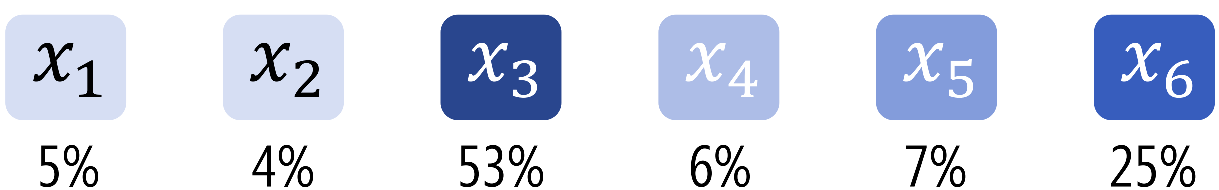
Huffman Codes

x_1	x_2	x_3	x_4	x_5	x_6
5%	4%	53%	6%	7%	25%

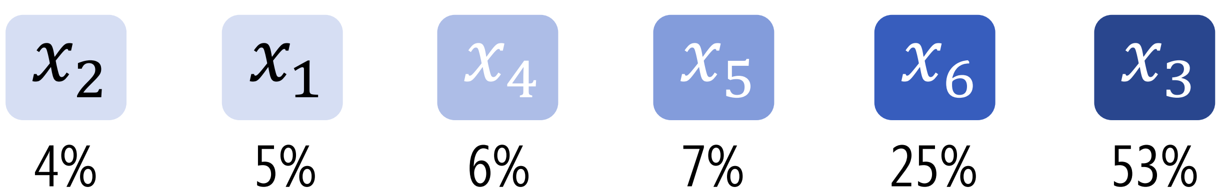
Algorithm

- Build a tree
  - Start with outcomes as leave nodes
- Iteratively:
  - Combine two lowest-probability nodes to new inner node
  - Until we have a root node
- Using a priority queue, if you care about run time

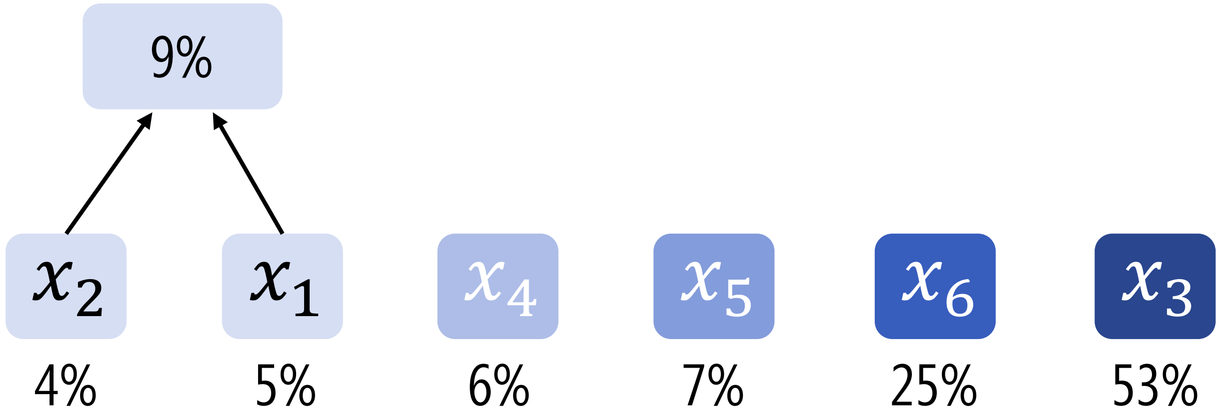
Huffman Codes



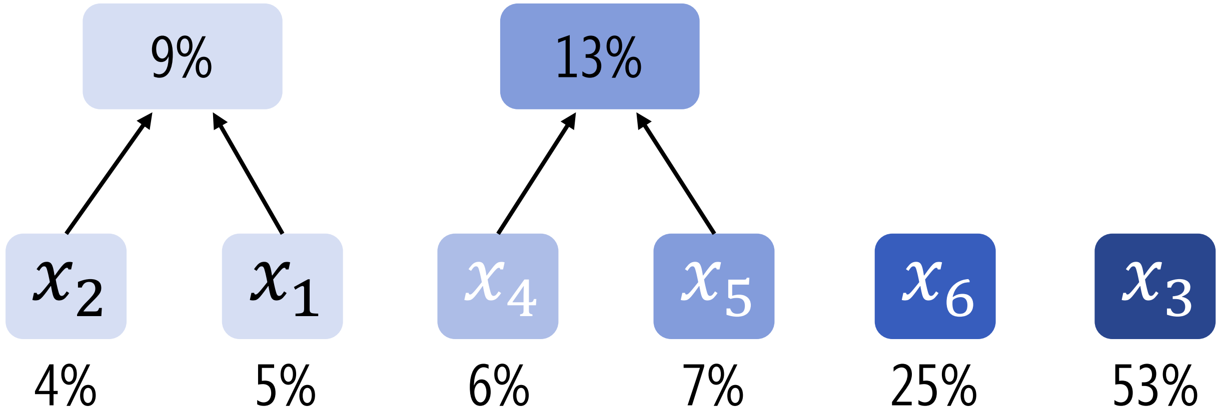
Huffman Codes



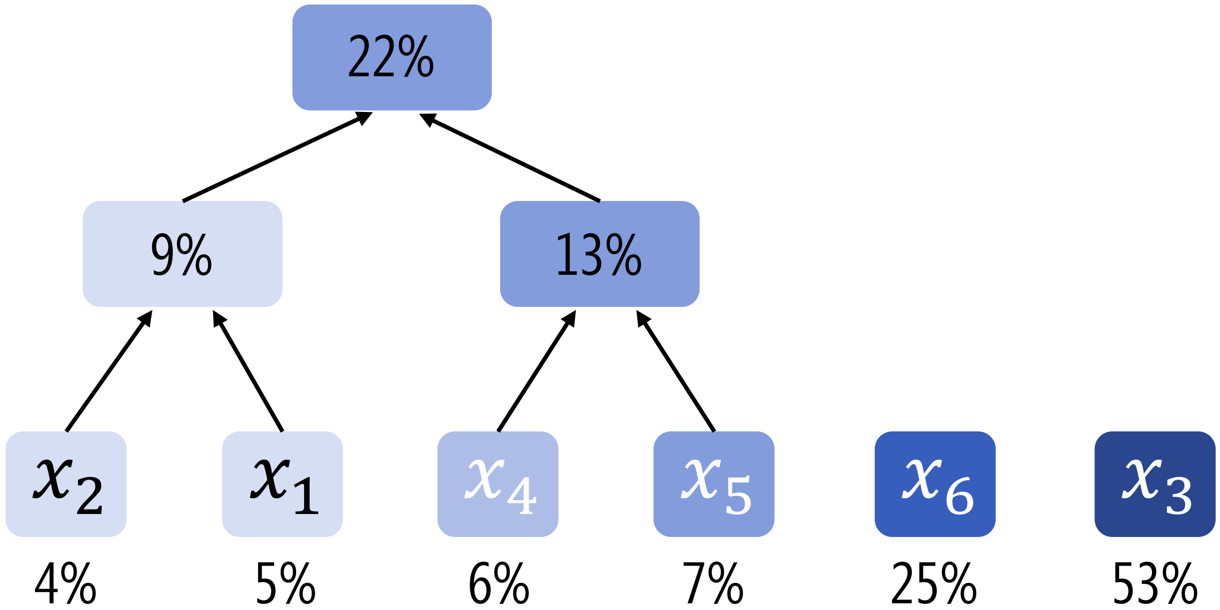
Huffman Codes



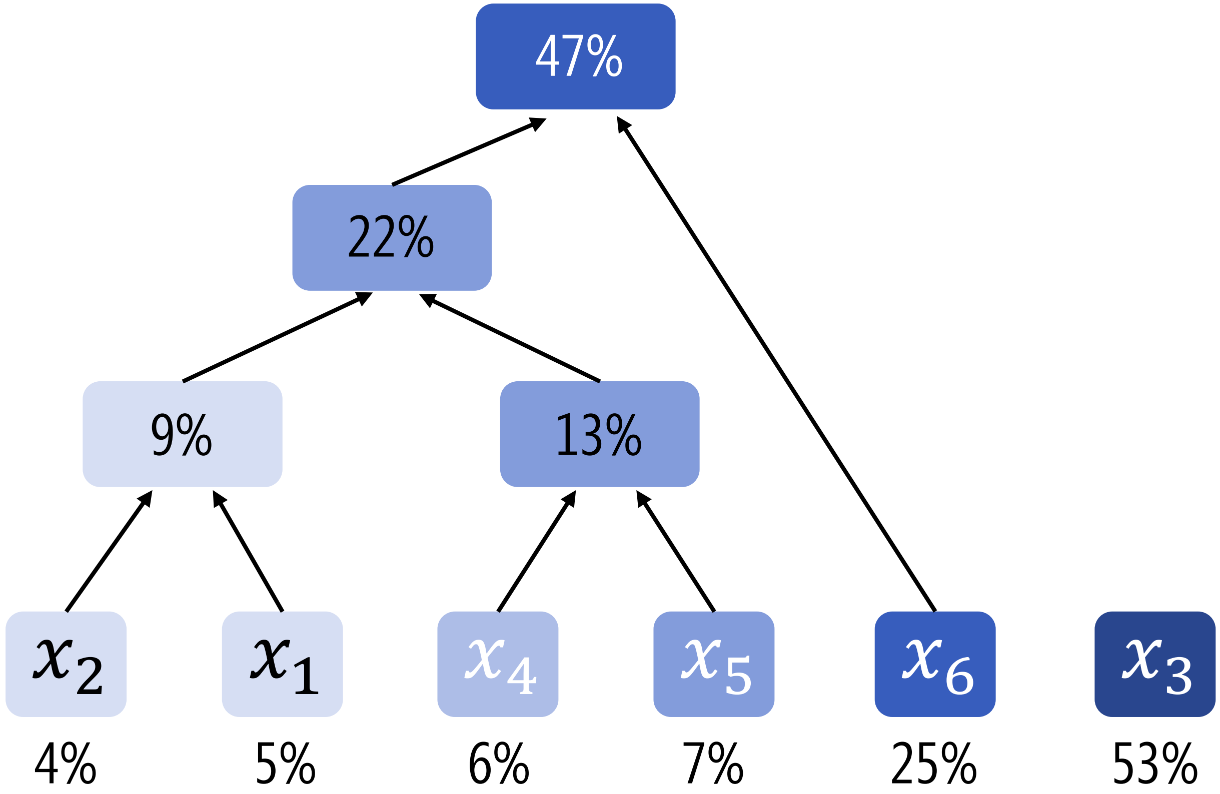
Huffman Codes



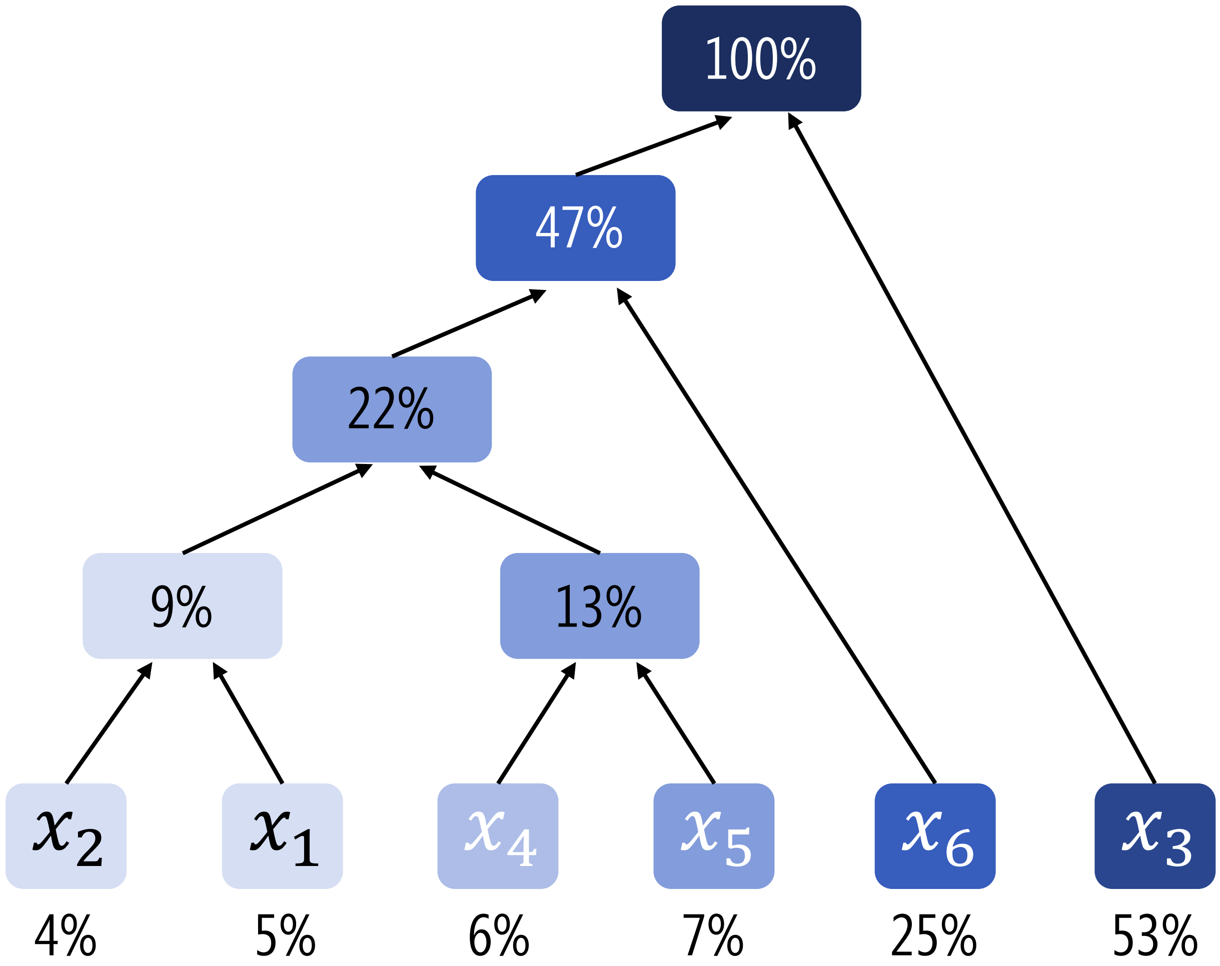
Huffman Codes



Huffman Codes

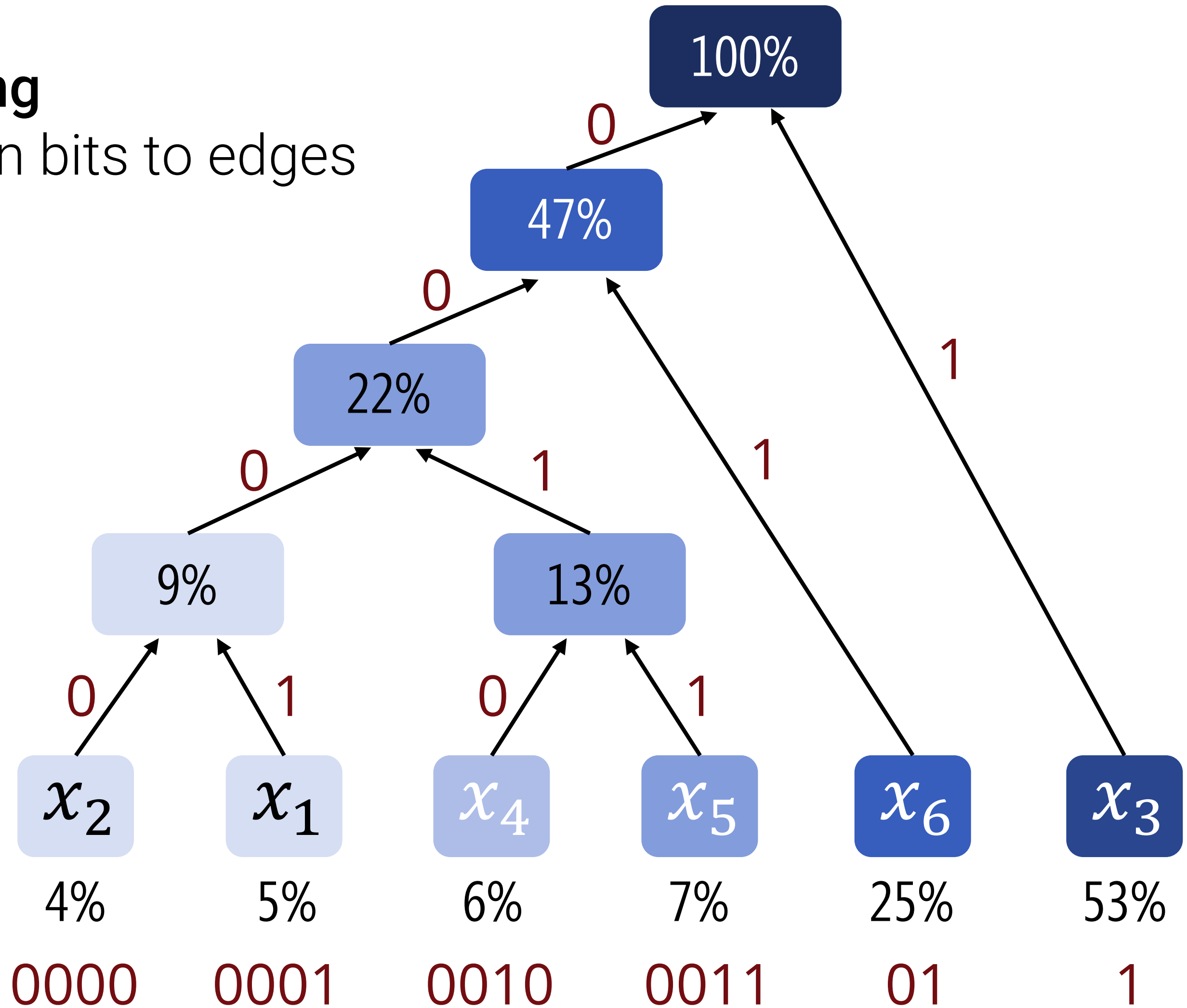


Huffman Codes



Huffman Codes

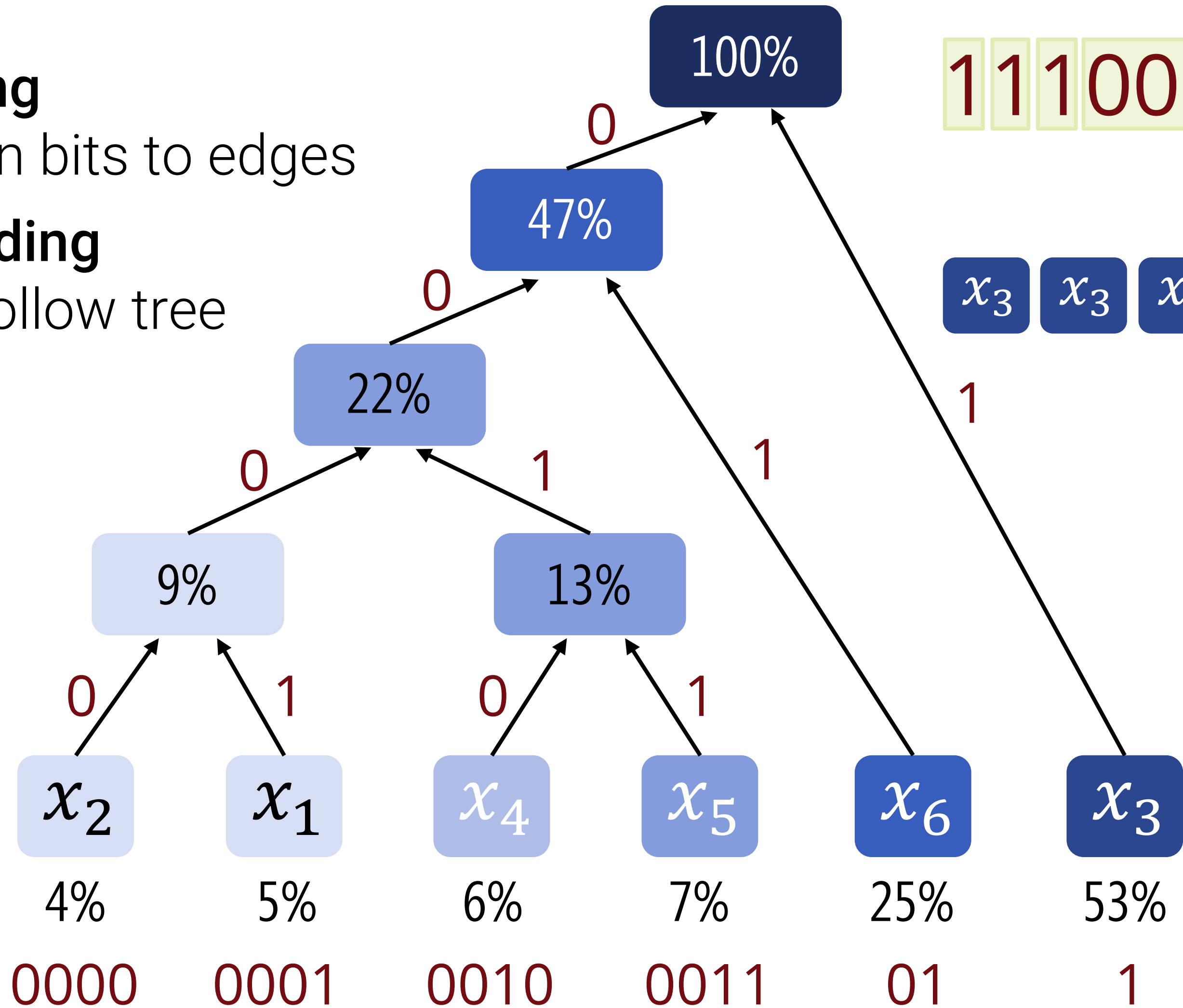
Coding
assign bits to edges

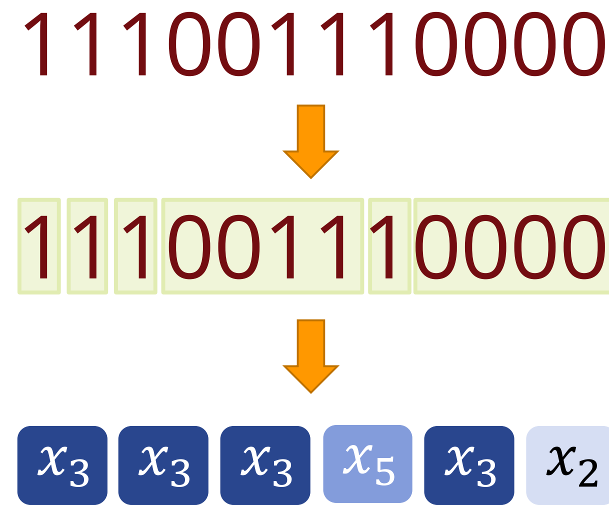


Huffman Codes

Coding
assign bits to edges

Decoding
just follow tree





Bit-Coding

Coding of Symbols

- Number of bits $\leq \log \frac{1}{p(x)} + 1$
- Information = code length (up to one bit)
- Entropy = expected code length (up to one bit)

Summary

Summary: Information & Entropy

Information is randomness

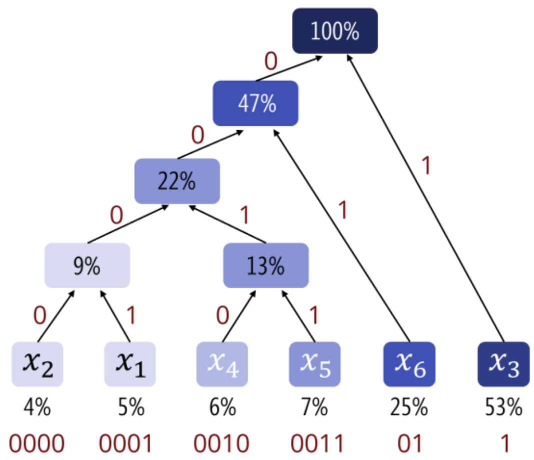
- "Frequentist" repeated coding scenario
- Analysis of coding length
  - Information $I(x) = -\log p(x)$
  - Entropy $H(p) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$
    $= \mathbb{E}_{x \sim p}[I(x)]$

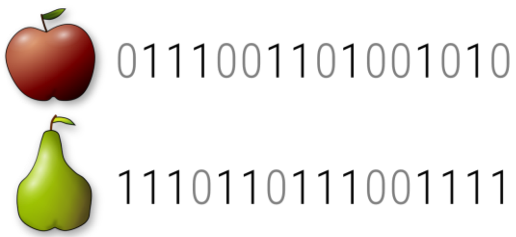
Not just pure theory

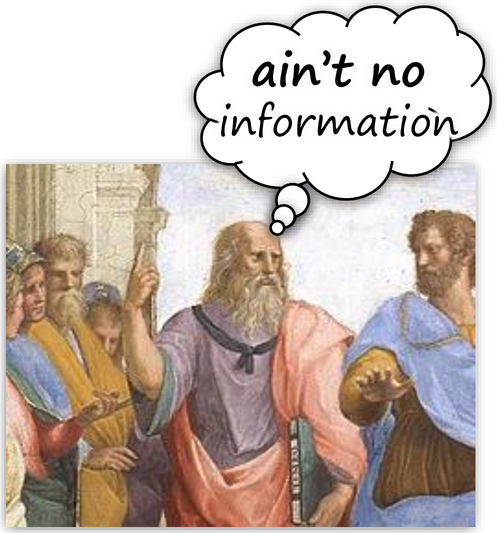
- Coding can be achieved (and is used) in practice

Modelling 2

STATISTICAL DATA MODELLING







Chapter 6

Information

Michael Wand · Institut für Informatik, JGU Mainz · michael.wand@uni-mainz.de

Video #06

Information Theory

- Information & Entropy
- **Algebra & Applications**

Entropy:

Additional Definitions & Theorems

Joint Entropy

Joint Entropy

$$H(X,Y) = -\sum_{i=1}^{n_x}\sum_{j=1}^{n_y} p(x_i,y_j)\log_2 p(x_i,y_j)$$

- Simply the entropy of the joint distribution $p(x,y)$

Theorem

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

$$\Leftrightarrow \; p(x,y) = p(x)p(y)$$

- Additive iff independent

Attention: Do not mix up with $H(p_1,p_2)$ for cross-entropy

Conditional Entropy

Conditional Entropy

$$H(X|Y) = -\sum_{i=1}^{n_x}\sum_{j=1}^{n_y} p(x_i|y_j)\log_2 p(x_i|y_j)$$

- Simply the entropy of the conditional distribution $p(x|y)$

Conditional Entropy

Marginal Entropy

$$H(X) = -\sum_{i=1}^{n_x} p(x_i) \log_2 p(x_i)$$

$$= -\sum_{i=1}^{n_x} \left(\sum_{j=1}^{n_y} p(x_i, y_j) \right) \left(\log_2 \sum_{j=1}^{n_y} p(x_i, y_j) \right)$$

- Simply the entropy of the marginal distribution $p(x)$

Conditional Entropy

Theorem: Chain Rule

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

Proof

- Very simple :-)

“Divergences”

Comparing Probability Distributions

Cross Entropy

Situation

- Two different distributions p_1, p_2
  (same probability space)

Definition: Cross Entropy (aka Relative Entropy)

$$H(p_1, p_2) = -\sum_{i=1}^{n} p_1(x_i) \log_2 p_2(x_i)$$

$$= \mathbb{E}_{x \sim p_1}\left[I_{p_2}(x)\right]$$

Idea

- Coding events $x \sim p_1$ with codes optimized for p_2

How to Read This...

Often: Searching for "codes"

first argument	second argument	output
data distribution	coding distribution	coding length

$$H(p_1, p_2) = -\sum_{i=1}^{n} p_1(x_i) \log_2 p_2(x_i)$$

Properties

- Non-symmetric!
- $\forall p_2: H(p_1, p_2) \geq H(p_1, p_1) = H(p_1)$
  - Reverse ($H(p_2, p_1)$ vs $H(p_1)$) is not true!
- **In optimization problems:** Usually vary p_2

Kullback-Leibler Divergence

Kullback-Leibler Divergence

$$KL(p_1 \parallel p_2) = \sum_{i=1}^{n} p_1(x_i) \log_2 \frac{p_1(x_i)}{p_2(x_i)}$$

$$= H(p_1, p_2) - H(p_1, p_1)$$

$$= H(p_1, p_2) - H(p_1)$$

Idea

- Measure coding efficiency p_1 using p_2-codes
  - Price to pay for coding in p_2 rather than p_1
- Compare with optimum for p_1
  - Measures how far distribution p_2 is from p_1

Kullback-Leibler Divergence

Kullback-Leibler Divergence

first argument
data distribution

second argument
coding distribution

output
increase in coding length

$$KL(p_1 \parallel p_2) = H(p_1, p_2) - H(p_1)$$

Idea

- Compare two distributions
  - Loss in coding efficiency [in bits]
  - Extra message length (Alice → Bob)
  - Just cross-entropy minus baseline $H(p_1, p_1)$
- Again, not symmetric

KL and JS Divergences

Kullback-Leibler Divergence

- Distance ≥ 0
- Zero distance means same distribution
- Not symmetric:

  $KL(p_1 \parallel p_2)$ different from $KL(p_2 \parallel p_1)$
- "Almost a metric"

Jensen–Shannon Divergence

- Symmetrized version
- $JSD(p_1 \parallel p_2) \coloneqq \frac{1}{2} KL(p_1 \parallel p_2) + \frac{1}{2} KL(p_2 \parallel p_1)$

What kind of metric is this?

KL-Divergence

$$KL(p_1 \parallel p_2) = \sum_{i=1}^{n} p_1(x_i) \log_2 \frac{p_1(x_i)}{p_2(x_i)}$$

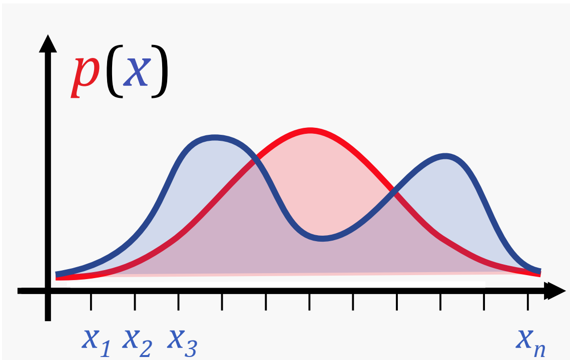
$$= \sum_{i=1}^{n} \underbrace{p_1(x_i)}_{\text{weighted by probability of occurrence in } p_1} \overbrace{[\log_2 p_1(x_i) - \log_2 p_2(x_i)]}^{\text{difference in Information for the same outcomes } x_i}$$

difference in Information for the same outcomes x_i

weighted by probability of occurrence in p_1

What kind of metric is this?

KL-Divergence



difference in Information
for the same outcomes x_i

$$KL(p_1 \parallel p_2) = \sum_{i=1}^{n} p_1(x_i)[\log_2 p_1(x_i) - \log_2 p_2(x_i)]$$

weighted by probability
of occurrence in p_1

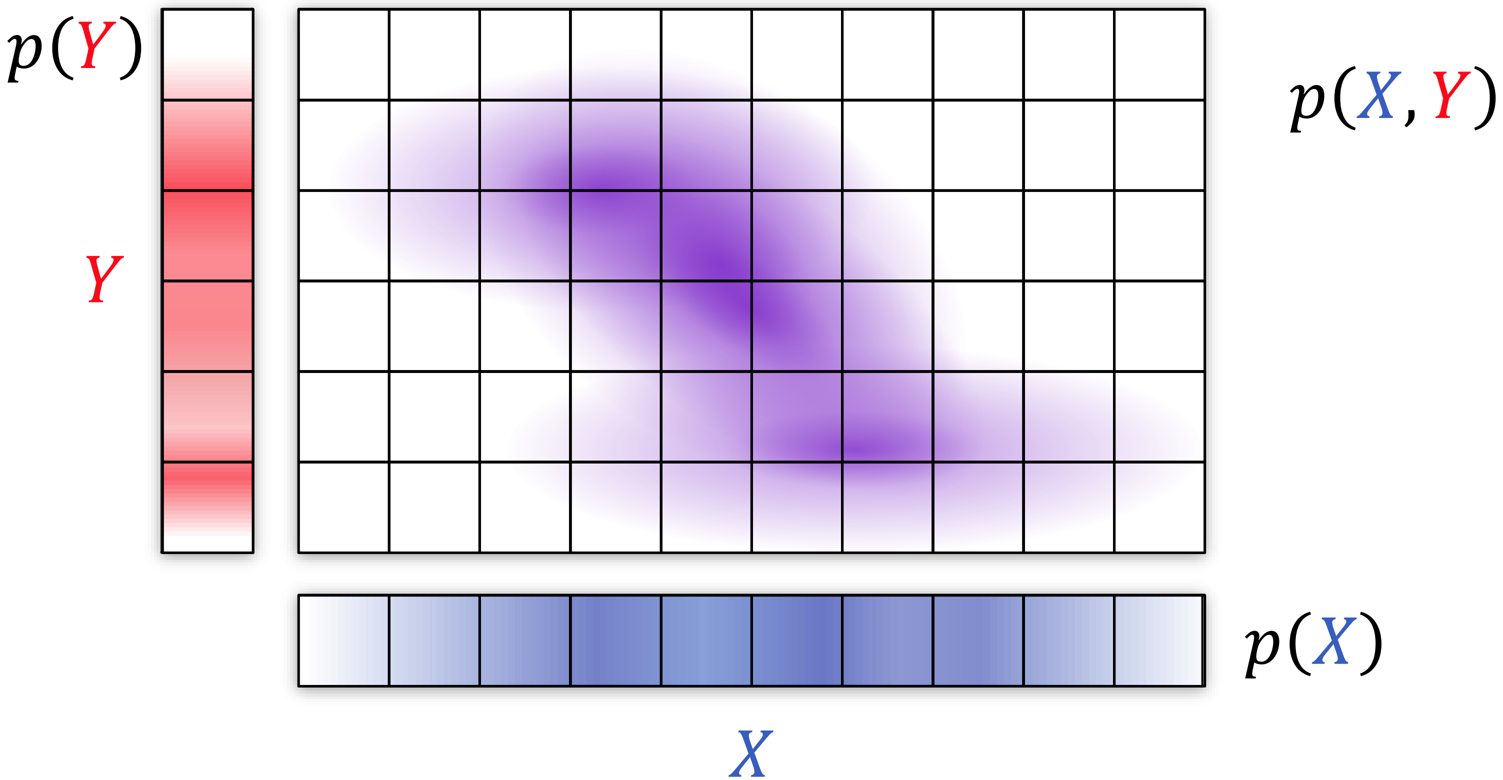
Mutual Information

Mutual Information

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

- Entropy of the marginal distributions minus that of the joint distribution

Mutual Information



Marginal & Joint Histograms

- Consider $H(X), H(Y), H(X,Y)$

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

Mutual Information

Alternative Formulas

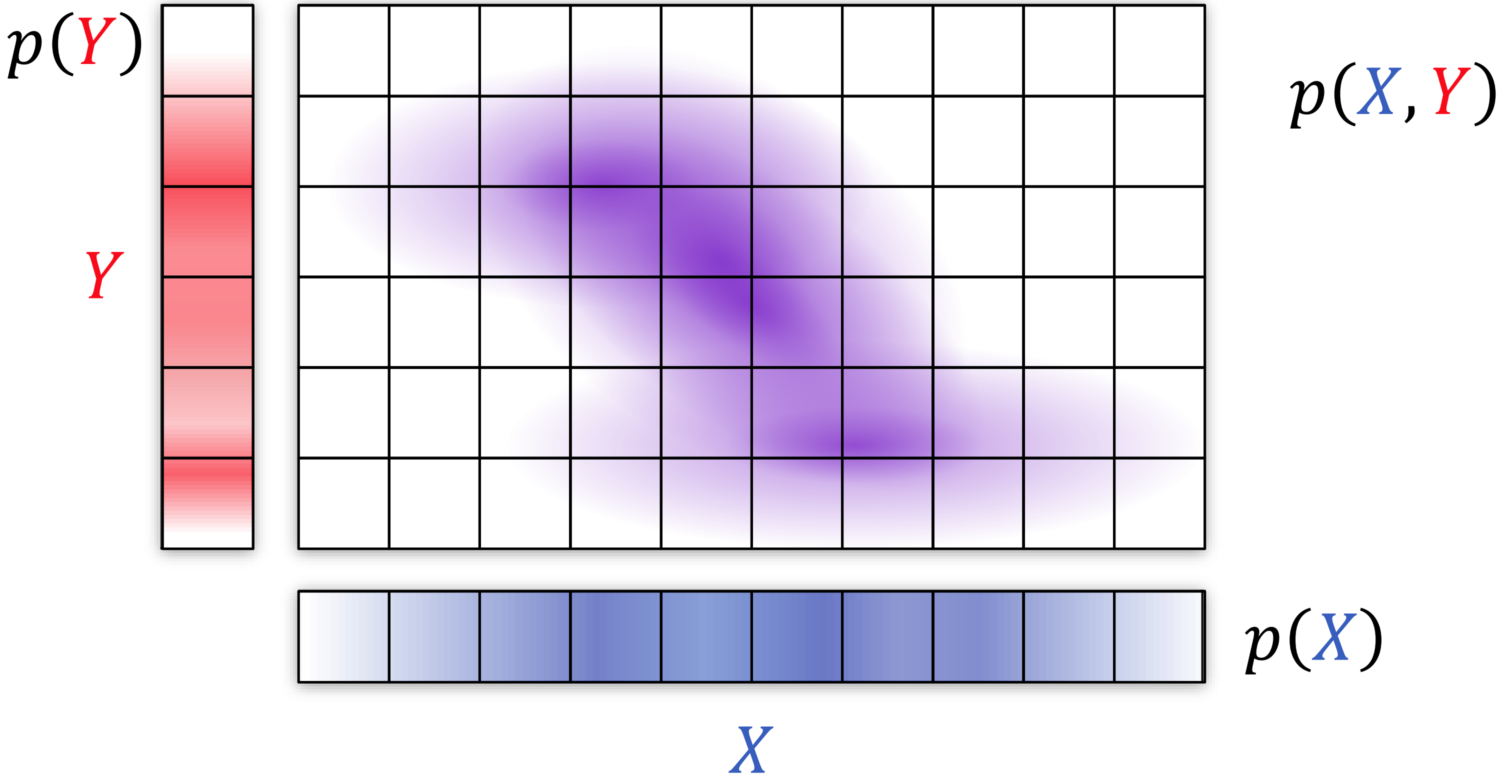
$$
\begin{aligned}
I(X;Y) &= H(X) + H(Y) - H(X,Y) \\
&= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X) \\
&= -\sum_{i=1}^{n_x}\sum_{j=1}^{n_y} p(x_i, y_j)\log_2\left(\frac{p(x_i, y_j)}{p(x_i)p(y_j)}\right) \\
&= KL\left(p(x_i, y_j) \parallel p(x_i)p(y_j)\right)
\end{aligned}
$$

Mutual Information

As a measure of dependency

- Most general *gradual* measure of dependency
- $I(X;Y) = 0 \;\Leftrightarrow\; X, Y$ are independent
- $I(X;Y) \to H(X) + H(Y)$: "maximally" dependent
  - Joint histogram becomes very sparse
  - $H(X,Y)$ very small
    - Zero not possible for discrete Ω if $H(X), H(Y) > 0$
    - Limit for $\#\Omega(X), \#\Omega(Y) \to \infty$
- Alternative measures such as correlation miss cases
  - Example: Linear correlation iff PCA spectrum flat
  - Does (e.g.) not detect quadratic dependencies

Computation



Actual Histograms

- Compute $H(X), H(Y), H(X, Y)$
- Costly: $O(|\Omega_X| \times |\Omega_Y|)$ (e.g., exponential in dimension)

Computation

Parametric Distributions

- Closed-Form Expressions for Gaussians etc.
- $H(\mathcal{N}_{\mu,\Sigma}) = \frac{1}{2}\ln\left((2\pi e)^d \det(\Sigma)\right)$ (differential entropy)

Computation

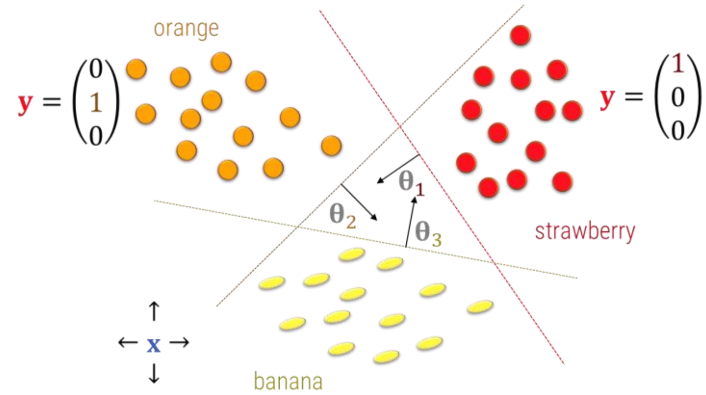
Approximations

- Sample-based Entropy
  - Measure only on input/training data of a DA/ML application
- Nearest-neighbors-methods
- Lower-bounds by “variational Bayes”
  - Build neural network f: predicting Y from X (or vice versa)
  - Least-squares fit $\|Y - f(X)\|^2$
  - Entropy of Gaussian error (covariance of errors)
    - Gives an upper bound of $H(X,Y)$
    - Upper bound of entropy of the joint Histogram
    - Has negative contribution, i.e.: lower bound for $I(X;Y)$

Application

Softmax Regression

Multi-Label Case



Task

- n Data points, indexed by $i = 1 \dots n$
  - Data $\mathbf{x}_i \in \mathbb{R}^d$ with…
  - …label vectors $\mathbf{y}_i \in \{0,1\}^K$
    - "One hot vectors"
- Learn class-specific parameters $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K \in \mathbb{R}^d$

Notation

- $y(\mathbf{x}) \in \{1, \dots, K\}$ denotes class index of input $\mathbf{x}$

Multi-Label Case

Unnormalized classifier

$$\mathbf{u}_{\theta}(\mathbf{x}) = \begin{pmatrix} - & \theta_1 & - \\ & \vdots & \\ - & \theta_k & - \end{pmatrix} \mathbf{x}$$

Class probabilities via softmax $\sigma: \mathbb{R}^K \to \mathbb{R}^K$

$$\sigma_m(\mathbf{y}_i) := \frac{e^{y_m}}{\sum_{j=1}^{K} e^{z_j}},$$

$$f_{\theta}(\mathbf{x}) := \begin{pmatrix} P(y(\mathbf{x}) = 1) \\ \vdots \\ P(y(\mathbf{x}) = K) \end{pmatrix} = \begin{pmatrix} \sigma_1(\mathbf{u}_{\theta}(\mathbf{x})) \\ \vdots \\ \sigma_K(\mathbf{u}_{\theta}(\mathbf{x})) \end{pmatrix}$$

Softmax Regression

MLE Training via

$$\theta = \underset{\theta \in \mathbb{R}^{K \times d}}{\arg\max} \prod_{i=1}^{n} f_{\theta}(\mathbf{x})_{y(\mathbf{x})}$$

$$= \underset{\theta \in \mathbb{R}^{K \times d}}{\arg\min} \sum_{i=1}^{n} -\log\left(f_{\theta}(\mathbf{x})_{y(\mathbf{x})}\right)$$

$$= \underset{\theta \in \mathbb{R}^{K \times d}}{\arg\min} \sum_{i=1}^{n} \left[\underbrace{\log\left(\sum_{m=1}^{K} e^{\theta_m^T \cdot \mathbf{x}}\right)}_{\text{normalization}} - \underbrace{\theta_{y(\mathbf{x})}^T \cdot \mathbf{x}}_{\substack{\text{(neg)}-\text{log}-\text{likelihood} \\ \text{of correct class}}} \right]$$

Cross Entropy Loss

Alternative formulation

- One-hot vectors $\mathbf{y}_i$ are “ground truth” distribution
  - Over classes $1 \dots K$
- Training: Make output distribution $f(\mathbf{x})$ similar to $\mathbf{y}_i$
  - Use KL-divergence to compare

$$KL\left(\mathbf{y}_i \parallel f_\theta(\mathbf{x}_i)\right) = \sum_{i=1}^{n} \mathbf{y}_i \log_2 \frac{\mathbf{y}_i}{f_\theta(\mathbf{x}_i)}$$

  - We will see: Minimization same for cross-entropy

$$H(\mathbf{y}_i, p_2) = -\sum_{i=1}^{n} \mathbf{y}_i \log_2 f_\theta(\mathbf{x}_i)$$

  - Which is per-class maximum-likelihood

KL as Cross Entropy as MLE

$$\arg\min_{\boldsymbol{\theta}} KL\left(\mathbf{y}_i \parallel f_{\boldsymbol{\theta}}(\mathbf{x}_i)\right)$$

← KL-Divergence

KL as Cross Entropy as MLE

$$\arg\min_{\theta} KL(\mathbf{y}_i \parallel f_\theta(\mathbf{x}_i))$$ ← KL-Divergence

$$= \arg\min_{\theta} \sum_{k=1}^{n_l} [\mathbf{y}_i]_k \log_2 \frac{[\mathbf{y}_i]_k}{[f_\theta(\mathbf{x}_i)]_k}$$

$$= \arg\min_{\theta} \left(H(\mathbf{y}_i, f_\theta(\mathbf{x}_i)) - H(\mathbf{y}_i) \right)$$

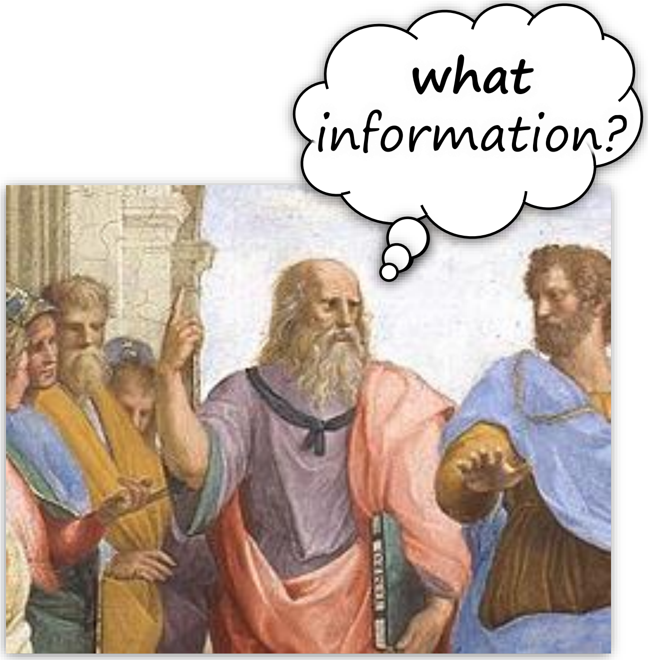
$$= \arg\min_{\theta} \left(H(\mathbf{y}_i, f_\theta(\mathbf{x}_i)) \right)$$ ← X-Entropy

$$= \arg\min_{\theta} \sum_{k=1}^{n_l} [\mathbf{y}_i]_k \log_2 [f_\theta(\mathbf{x}_i)]_k$$

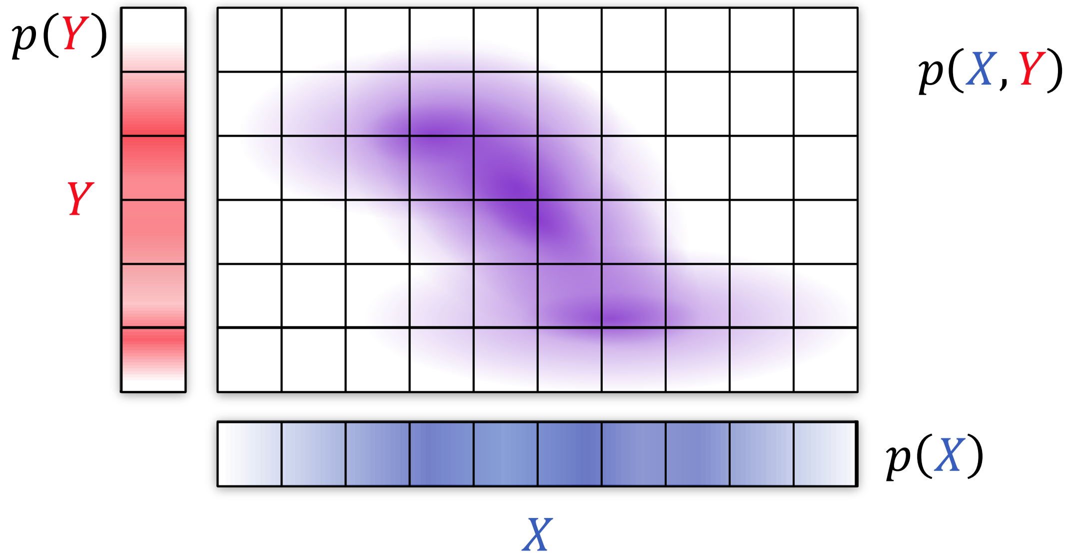
$$= \arg\min_{\theta} \log_2 [f_\theta(\mathbf{x}_i)]_{y(\mathbf{x}_i)}$$ ← MLE

Thoughts About

The Nature of Information



Properties of Mutual Information



Bijection invariant

- Discrete $\Omega(X) = \{1, \dots, n_X\}, \Omega(Y) = \{1, \dots, n_Y\}$
- For bijective $\pi_X : \Omega(X) \to \Omega(X),\ \pi_Y : \Omega(Y) \to \Omega(Y)$

$$I(X;Y) = I\big(\pi_X(X); \pi_Y(Y)\big)$$

- Invertible functions do not change information

Bijection Invariance

Applies to other measures

- Entropy

$$H(X) = H\big(\pi(X)\big)$$

For any bijection $\pi: X \to X$

- Proof

$$H(X) = \sum_{i=1}^{n} p(x_i) \log p(x_i)$$

$$= \sum_{i=1}^{n} p\big(x_{\pi(i)}\big) \log p\big(x_{\pi(i)}\big) \quad \text{(identifying } x_i \text{ with } i\text{)}$$

$$= H\big(\pi(X)\big)$$

Bijection Invariance

Information theoretic measures

- Entropy
- Mutual Information

are invariant under

- Bijective mappings,
- i.e.: application of “information preserving functions”
- Applies to divergences only if both p_1, p_2 are transformed the same way
  - Cross-Entropy, KL-Divergence, J-S-Divergence

Data Processing

Deterministic Information Processing

- Arbitrary function

$$f : \Omega(X) \to \Omega(Y)$$

- We can only lose information

$$H(X) \geq H\big(f(X)\big)$$

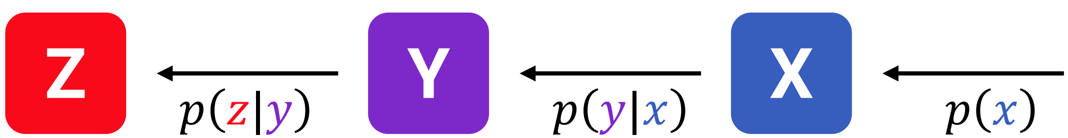
Data Processing

(Probabilistic) **Data Processing Inequality**

- Random variables with densities

$$X, Y, Z \text{ with } p(x, y, z)$$

- Chain-like dependency structure



$$p(x, y, z) = p(z|y) \cdot p(y|x) \cdot p(x)$$

btw, this is called a “Markov chain”

- Data processing inequality

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

Information

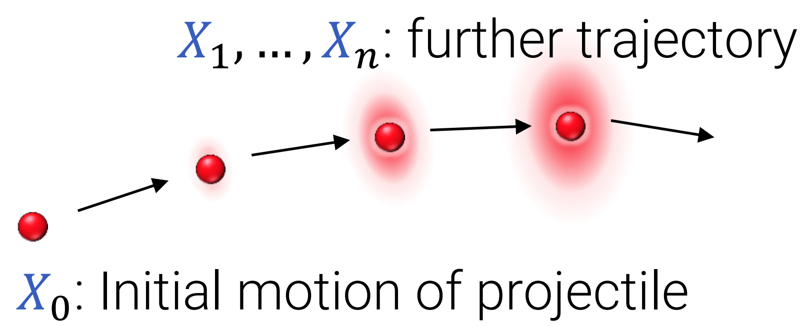
Information

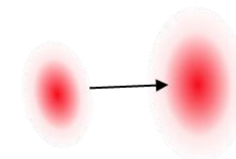
- Originate from random process X

Processing / Calculating

- Deterministic processes can only reduce information
- Probabilistic processes can add information, but cannot add information on original X
- Bijections (invertible maps) do not change anything

Probabilistic Evolution of Information



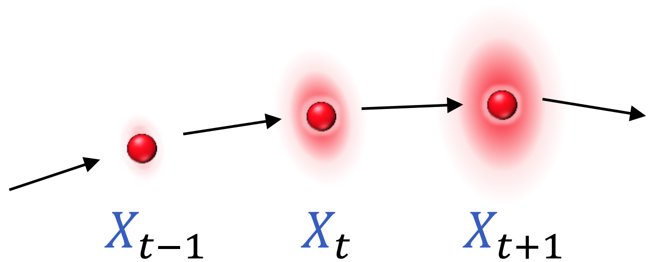


$p(X_{t+1}|X_t)$
models physics
+ randomness
(e.g. air turbulence)

Example

- Trajectory of a projectile
  - Imprecision due to limited knowledge (wind)
- If motion was deterministic
  - No information loss: $\forall t \geq 0 \colon H(X_t) = H(X_0)$
    - Physics is reversible (≡ bijective)
    - But we have incomplete knowledge

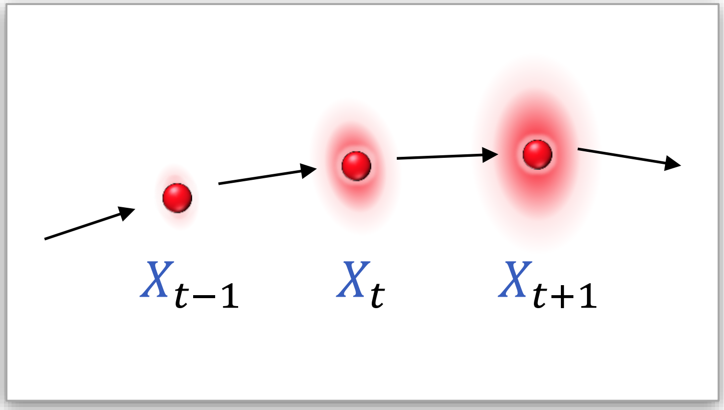
Probabilistic Evolution of Information

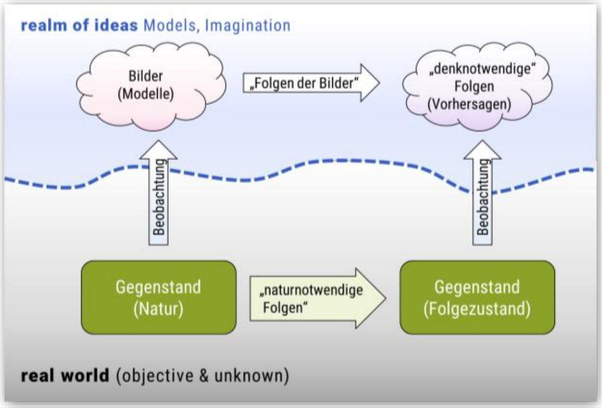


With random perturbations

- Old information gradually replaced by new randomness
  - **Loss:** X_t cannot be fully reconstructed from X_{t+1}
  - **Gain:** X_t not fully predictable from X_{t-1} (new random info.)
- Information is probabilistic
  - Available knowledge reduces entropy of $P(X_t)$

In one sentence





Information in machine learning

Being able to predict (e.g., the future)

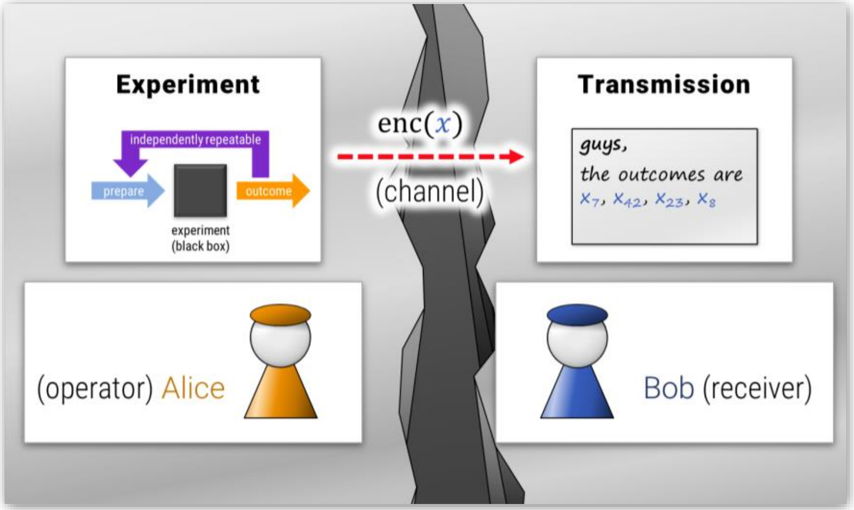
means

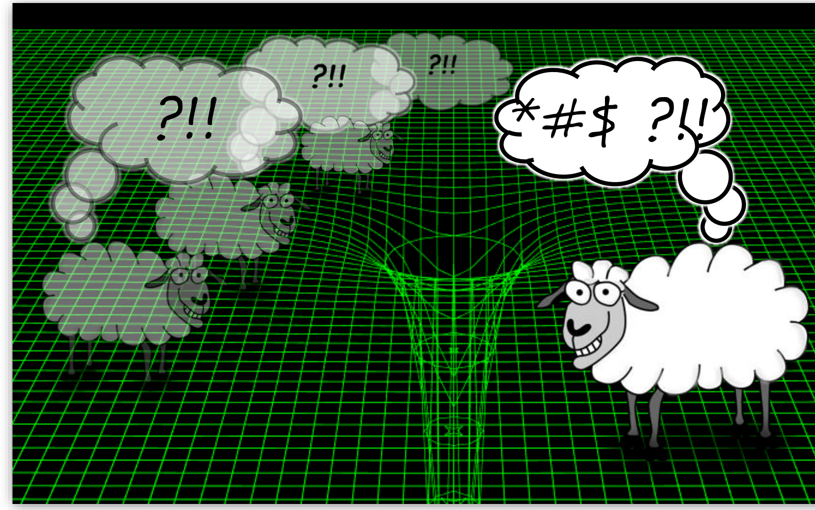
reducing the uncertainty/entropy

(of the probability distribution of the outcome)

So – What *is* Information?

What is Information?





Frequentist Information **Bayesian Information?**

Bayesian Probabilities → Information

- One time-events
- Model uncertainty & subjective knowledge
- Information = "I learned something new"

(Note: Personal view/interpretation)

Summary

Divergences: Comparing Distributions

Divergences

- **Cross entropy** (a.k.a. relative entropy)
- KL divergence & JS divergence
- Mutual Information

Computation

- Analytical solution
- Numerics: very expensive
  - Linear/quadratic in $|\Omega|$ usually means exponential in input
  - There are many dirty tricks / approximations

Divergences: Comparing Distributions

What do they do?

- Measure differences in distributions wrt. information
- Pure “information”
  - Every bit of random noise counts

X-Entropy, KL/JS-Divergence

- Compare information of corresponding outcomes

Mutual Information

- MI is fully bijection invariant (XE/KL/JS are not!)

Use with care

- Pure information is not always what you want!