How to Learn Probability Without Learning

Data → ? → P

Young-Han Kim, UCSD

Munich Workshop on Causal Inference and Information Theory
May 23, 2016
Laplace’s rule of succession (1812):

$$P(\bigcirc \mid n \times \bigcirc) = \frac{n + 1}{n + 2}$$
How to learn probability

\[ X^n = X_1, X_2, \ldots, X_n \]

010010100...11
GACGAACTGGC...GG
ONCE UPON...END
How to learn probability

\[ X^n = X_1, X_2, \ldots, X_n \]

010010100...11
GACGAACTGGC...GG
ONCE UPON...END

\[ P(X_{n+1} = x | X^n) \]
How to learn probability

\[ X^n = X_1, X_2, \ldots, X_n \] \quad \rightarrow \quad ? \quad \rightarrow \quad P(X_{n+1} = x \mid X^n) \]

010010100...11
GACGAACTGGC...GG
ONCE UPON...END

Forward estimation

Is there \( f_n \) such that

\[ |f_n(X^n_1) - P(X_{n+1} = x \mid X^n_1)| \rightarrow 0 \quad \text{a.s.} \]

for every stationary ergodic \( \{X_n\} \)?
How to learn probability

\[ X^n = X_1, X_2, \ldots, X_n \rightarrow ? \rightarrow P(X_{n+1} = x | X^n) \]

010010100...11
GACGAACTGGC...GG
ONCE UPON...END

**Forward estimation**
Is there \( f_n \) such that
\[
|f_n(X^n_1) - P(X_{n+1} = x | X^n_1)| \rightarrow 0 \text{ a.s.}
\]
for every stationary ergodic \( \{X_n\} \)?

**Backward estimation (Cover 1975)**
Is there \( g_n \) such that
\[
|g_n(X^{-1}_{-n}) - P(X_0 = x | X^{-1}_{-n})| \rightarrow 0 \text{ a.s.}
\]
for every stationary ergodic \( \{X_n\} \)?
How to learn probability

\[ X^n = X_1, X_2, \ldots, X_n \rightarrow ? \rightarrow P(X_{n+1} = x | X^n) \]

010010100...11
GACGAACTGGC...GG
ONCE UPON...END

**Forward estimation**

Is there \( f_n \) such that

\[ |f_n(X^n_1) - P(X_{n+1} = x | X^n_1)| \rightarrow 0 \text{ a.s.} \]

for every stationary ergodic \( \{X_n\} \)?

**Answer:** No! (Bailey 1976)

Getting older doesn’t make you wiser

**Backward estimation (Cover 1975)**

Is there \( g_n \) such that

\[ |g_n(X^{-1}_n) - P(X_0 = x | X^{-1}_n)| \rightarrow 0 \text{ a.s.} \]

for every stationary ergodic \( \{X_n\} \)?
How to learn probability

\[ X^n = X_1, X_2, \ldots, X_n \rightarrow P(X_{n+1} = x \mid X^n) \]

Forward estimation
Is there \( f_n \) such that

\[ |f_n(X_1^n) - P(X_{n+1} = x \mid X_1^n)| \rightarrow 0 \quad \text{a.s.} \]

for every stationary ergodic \( \{X_n\} \)?

Answer: No! (Bailey 1976)

Getting older doesn’t make you wiser

Backward estimation (Cover 1975)
Is there \( g_n \) such that

\[ |g_n(X_{-1}^{-n}) - P(X_0 = x \mid X_{-1}^{-n})| \rightarrow 0 \quad \text{a.s.} \]

for every stationary ergodic \( \{X_n\} \)?

Answer: Yes! (Ornstein 1978)

Learning more history does
How to learn probability

Forward estimation
Is there $f_n$ such that

$$|f_n(X_1^n) - P(X_{n+1} = x | X_1^n)| \to 0 \text{ a.s.}$$

for every stationary ergodic $\{X_n\}$?

Answer: No! (Bailey 1976)

Getting older doesn’t make you wiser

Backward estimation (Cover 1975)
Is there $g_n$ such that

$$|g_n(X_{-1}^{-n}) - P(X_0 = x | X_{-1}^{-1})| \to 0 \text{ a.s.}$$

for every stationary ergodic $\{X_n\}$?

Answer: Yes! (Ornstein 1978)

Learning more history does
How to learn probability without learning

\[ X \sim p(x) \quad \Rightarrow \quad \hat{p}(\cdot|X, p) \rightarrow \hat{p}(X) \]
How to learn probability without learning

There is no black box

$q \approx p$ for all $p \in \mathcal{P}$
How to learn probability without learning

There is no black box

$q \approx p \text{ for all } p \in \mathcal{P}$

- Universal
  - $\mathcal{P}$: parametric, IID, Markov, VMM, HMM, FSM, stationary ergodic, …
How to learn probability without learning

- Universal
  - $\mathcal{P}$: parametric, IID, Markov, VMM, HMM, FSM, stationary ergodic, ...

- Quick and dirty
  - Compression, prediction, filtering, denoising, portfolio, entropy estimation, classification
How to learn probability without learning

\[ X \sim p(x) \]

There is no black box

\[ q(X) \]

\[ q \approx p \text{ for all } p \in \mathcal{P} \]

• Universal
  - \( \mathcal{P} \): parametric, IID, Markov, VMM, HMM, FSM, stationary ergodic, ...

• Quick and dirty\[ \times \text{ CLEAN} \]
  - Compression, prediction, filtering, denoising, portfolio, entropy estimation, classification

• Avoids overfitting (built-in regularization)
Outline of the talk

• Brief overview of universal probability assignment
  #1. Definition, existence, and construction
  #2. Convergence control
  #3. Well-known applications
Outline of the talk

• Brief overview of universal probability assignment
  #1. Definition, existence, and construction
  #2. Convergence control
  #3. Well-known applications

• Directed information and its application to causality inference
Outline of the talk

• Brief overview of universal probability assignment
  #1. Definition, existence, and construction
  #2. Convergence control
  #3. Well-known applications

• Directed information and its application to causality inference

• Classification of DNA/RNA sequences using universal probability
Universal probability for stationary ergodic processes

$q$ is mean universal if

$$\frac{1}{n} D(p(x^n) \| q(x^n)) \to 0 \quad \forall p \in \mathcal{P}$$
Universal probability for stationary ergodic processes

\( q \) is mean universal if

\[
\frac{1}{n} D(p(x^n) \| q(x^n)) \to 0 \quad \forall p \in \mathcal{P}
\]

\( q \) is pointwise universal if

\[
\frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s.} \quad \forall p \in \mathcal{P}
\]
### Universal probability for stationary ergodic processes

$q$ is mean universal if
\[
\frac{1}{n} D(p(x^n) \| q(x^n)) \to 0 \quad \forall p \in \mathcal{P}
\]

$q$ is pointwise universal if
\[
\frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s.} \quad \forall p \in \mathcal{P}
\]

- Relative entropy (Kullback–Leibler divergence)

\[
D(p(x^n) \| q(x^n)) = \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = E_p \log \frac{p(X^n)}{q(X^n)}
\]
Universal probability for stationary ergodic processes

$q$ is mean universal if
\[
\frac{1}{n} D(p(x^n)\|q(x^n)) \to 0 \quad \forall p \in \mathcal{P}
\]

$q$ is pointwise universal if
\[
\frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s.} \quad \forall p \in \mathcal{P}
\]

- Relative entropy (Kullback–Leibler divergence)
  \[
  D(p(x^n)\|q(x^n)) = \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = \mathbb{E}_p \log \frac{p(X^n)}{q(X^n)}
  \]

- Simple construction using LZ78 incremental parsing (Ziv–Lempel 1978)
  \[x^n =\]
Universal probability for stationary ergodic processes

\( q \) is mean universal if
\[
\frac{1}{n} D(p(x^n) \| q(x^n)) \to 0 \quad \forall p \in \mathcal{P}
\]

\( q \) is pointwise universal if
\[
\frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s.} \quad \forall p \in \mathcal{P}
\]

- Relative entropy (Kullback–Leibler divergence)

\[
D(p(x^n) \| q(x^n)) = \sum_{x^n} \log \frac{p(x^n)}{q(x^n)} = E_p \log \frac{p(X^n)}{q(X^n)}
\]

- Simple construction using LZ78 incremental parsing (Ziv–Lempel 1978)

\[ x^n = 0 \]
Universal probability for stationary ergodic processes

- Relative entropy (Kullback–Leibler divergence)
  \[
  D(p(x^n) \| q(x^n)) = \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = \mathbb{E}_p \log \frac{p(X^n)}{q(X^n)}
  \]

- Simple construction using **LZ78 incremental parsing** (Ziv–Lempel 1978)
  \[x^n = 0 \ 00\]
Universal probability for stationary ergodic processes

$q$ is mean universal if

$$\frac{1}{n} D(p(x^n)\|q(x^n)) \to 0 \quad \forall p \in \mathcal{P}$$

$q$ is pointwise universal if

$$\frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s. } \forall p \in \mathcal{P}$$

- Relative entropy (Kullback–Leibler divergence)

$$D(p(x^n)\|q(x^n)) = \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = \mathbb{E}_p \log \frac{p(X^n)}{q(X^n)}$$

- Simple construction using **LZ78 incremental parsing** (Ziv–Lempel 1978)

$$x^n = 0 \ 00 \ 1$$
Universal probability for stationary ergodic processes

$q$ is mean universal if

$$\frac{1}{n} D(p(x^n)\|q(x^n)) \to 0 \quad \forall p \in \mathcal{P}$$

$q$ is pointwise universal if

$$\frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s. } \forall p \in \mathcal{P}$$

- Relative entropy (Kullback–Leibler divergence)

$$D(p(x^n)\|q(x^n)) = \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = \mathbb{E}_p \log \frac{p(X^n)}{q(X^n)}$$

- Simple construction using **LZ78 incremental parsing** (Ziv–Lempel 1978)

$$x^n = 0 0 0 1 1 0$$
Universal probability for stationary ergodic processes

$q$ is mean universal if
\[ \frac{1}{n} D(p(x^n)\|q(x^n)) \to 0 \quad \forall p \in \mathcal{P} \]
$q$ is pointwise universal if
\[ \frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s.} \quad \forall p \in \mathcal{P} \]

- Relative entropy (Kullback–Leibler divergence)

\[ D(p(x^n)\|q(x^n)) = \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = \mathbb{E}_p \log \frac{p(X^n)}{q(X^n)} \]

- Simple construction using **LZ78 incremental parsing** (Ziv–Lempel 1978)

\[ x^n = 0 \ 00 \ 1 \ 10 \ 101 \]
Universal probability for stationary ergodic processes

\( q \) is mean universal if

\[
\frac{1}{n} D(p(x^n) \| q(x^n)) \to 0 \quad \forall p \in \mathcal{P}
\]

\( q \) is pointwise universal if

\[
\frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s.} \quad \forall p \in \mathcal{P}
\]

- Relative entropy (Kullback–Leibler divergence)

\[
D(p(x^n) \| q(x^n)) = \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = \mathbb{E}_p \log \frac{p(X^n)}{q(X^n)}
\]

- Simple construction using **LZ78 incremental parsing** (Ziv–Lempel 1978)

\[
x^n = 0 \ 00 \ 1 \ 10 \ 101 \ 101 \ 1
\]
Universal probability for stationary ergodic processes

\( q \) is mean universal if
\[
\frac{1}{n} D(p(x^n)\|q(x^n)) \to 0 \quad \forall p \in \mathcal{P}
\]

\( q \) is pointwise universal if
\[
\frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s.} \quad \forall p \in \mathcal{P}
\]

- Relative entropy (Kullback–Leibler divergence)

\[
D(p(x^n)\|q(x^n)) = \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = E_p \log \frac{p(X^n)}{q(X^n)}
\]

- Simple construction using **LZ78 incremental parsing** (Ziv–Lempel 1978)

\[ x^n = 0 \ 00 \ 1 \ 10 \ 101 \ 1011 \ 000 \]
Universal probability for stationary ergodic processes

$q$ is mean universal if
\[ \frac{1}{n} D(p(x^n) \| q(x^n)) \to 0 \quad \forall p \in \mathcal{P} \]

$q$ is pointwise universal if
\[ \frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad \text{p-a.s.} \quad \forall p \in \mathcal{P} \]

- Relative entropy (Kullback–Leibler divergence)

\[
D(p(x^n) \| q(x^n)) = \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = \mathbb{E}_p \log \frac{p(X^n)}{q(X^n)}
\]

- Simple construction using **LZ78 incremental parsing** (Ziv–Lempel 1978)

\[
x^n = 0 \ 00 \ 1 \ 10 \ 101 \ 1011 \ 000 \quad \Rightarrow \quad q(x^n) = \frac{1}{(1 + c(x^n))!}
\]
Universal probability for stationary ergodic processes

$q$ is mean universal if
\[
\frac{1}{n} D(p(x^n)\|q(x^n)) \to 0 \quad \forall p \in \mathcal{P}
\]

$q$ is pointwise universal if
\[
\frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s.} \quad \forall p \in \mathcal{P}
\]

- Relative entropy (Kullback–Leibler divergence)

\[
D(p(x^n)\|q(x^n)) = \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = \mathbb{E}_p \log \frac{p(X^n)}{q(X^n)}
\]

- Simple construction using LZ78 incremental parsing (Ziv–Lempel 1978)

\[
x^n = 001101011000 \implies q(x^n) = \frac{1}{(1 + c(x^n))!}
\]

- Works “well” with deterministic (=real-life) sequences
Universal probability for stationary ergodic processes

\( q \) is mean universal if
\[
\frac{1}{n} D(p(x^n)\|q(x^n)) \to 0 \quad \forall p \in \mathcal{P}
\]

\( q \) is pointwise universal if
\[
\frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s.} \quad \forall p \in \mathcal{P}
\]

- Relative entropy (Kullback–Leibler divergence)

\[
D(p(x^n)\|q(x^n)) = \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = E_p \log \frac{p(X^n)}{q(X^n)}
\]

- Simple construction using LZ78 incremental parsing (Ziv–Lempel 1978)

\[ x^n = 0 \ 00 \ 1 \ 10 \ 101 \ 1011 \ 000 \Rightarrow q(x^n) = \frac{1}{(1 + c(x^n))!} \]

- Works “well” with deterministic (=real-life) sequences

- Faster convergence (minimax) for smaller classes \( \mathcal{P} \) (IID, Markov, CTM, …)
Minimax probability assignment
Minimax probability assignment

Minimax redundancy (Gallager 1974)

\[ R^* = \min_{q} \max_{p \in \mathcal{P}} D(p(x) \| q(x)) = \max_{F(p)} I(P; X) \]
Minimax probability assignment

Minimax redundancy (Gallager 1974)

\[ R^* = \min_{q} \max_{p \in \mathcal{P}} D(p(x) \parallel q(x)) = \max_{F(p)} I(P; X) \]

\[ q^*(x) = \int p(x) dF^*(p) \]
Minimax probability assignment

Minimax redundancy (Gallager 1974)

\[ R^* = \min_{q} \max_{p \in \mathcal{P}} D(p(x) \| q(x)) = \max_{F(p)} I(P; X) \]

\[ q^*(x) = \int p(x) dF^*(p) \]

- Mixture probability \( F(p) \) can upper and lower bound \( R^* \)
Minimax probability assignment

Minimax redundancy (Gallager 1974)

\[ R^* = \min_{q} \max_{p \in \mathcal{P}} D(p(x) \| q(x)) = \max_{F(p)} I(P; X) \]

\[ q^*(x) = \int p(x) dF^*(p) \]

- Mixture probability \( F(p) \) can upper and lower bound \( R^* \)
- For the deterministic setting,

\[ R^* = \min_{q} \max_{p \in \mathcal{P}} \max_{x} \log \frac{p(x)}{q(x)} \]

\[ = \log \sum_{x} \max_{p \in \mathcal{P}} p(x) \]

\[ q^*(x) \propto \max_{p \in \mathcal{P}} p(x) \quad \text{(normalized ML)} \]
Bernoulli sources

- Let $X_1, X_2, \ldots$ be i.i.d. $\sim \text{Bern}(\theta)$, $\theta \in [0, 1]$ unknown
Bernoulli sources

- Let $X_1, X_2, \ldots$ be i.i.d. $\sim \text{Bern}(\theta)$, $\theta \in [0, 1]$ unknown
- **Uniform mixture (Laplace 1812):** $R \sim \log n$ (universal for Bernoulli sources!)

\[
q_L(x^n) = \int_0^1 \theta^k(1 - \theta)^{n-k} d\theta = \frac{1}{\binom{n}{k}(n + 1)}
\]

and

\[
q_L(1 \mid x^n) = \frac{k + 1}{n + 2}
\]
Bernoulli sources

- Let $X_1, X_2, \ldots$ be i.i.d. $\sim \text{Bern}(\theta)$, $\theta \in [0, 1]$ unknown

- **Uniform mixture (Laplace 1812):** $R \sim \log n$ (universal for Bernoulli sources!)

  \begin{align*}
  q_L(x^n) &= \int_0^1 \theta^k (1 - \theta)^{n-k} d\theta = \frac{1}{\binom{n}{k}(n+1)} \\
  q_L(1|x^n) &= \frac{k + 1}{n + 2}
  \end{align*}

- **Jeffreys mixture (Krichevsky–Trofimov 1981):** $R \sim \frac{1}{2} \log n \sim R^*$

  \begin{align*}
  q_{KT}(x^n) &= \int_0^1 \theta^k (1 - \theta)^{n-k} \frac{1}{\sqrt{\theta(1 - \theta)}} d\theta \\
  q_{KT}(1|x^n) &= \frac{k + 1/2}{n + 1}
  \end{align*}
Bernoulli sources

- Let \( X_1, X_2, \ldots \) be i.i.d. \( \sim \text{Bern}(\theta) \), \( \theta \in [0, 1] \) unknown

- **Uniform mixture (Laplace 1812):** \( R \sim \log n \) (universal for Bernoulli sources!)

\[
q_L(x^n) = \int_0^1 \theta^k (1 - \theta)^{n-k} d\theta = \frac{1}{\binom{n}{k}(n+1)} \quad \text{and} \quad q_L(1 \mid x^n) = \frac{k + 1}{n + 2}
\]

- **Jeffreys mixture (Krichevsky–Trofimov 1981):** \( R \sim \frac{1}{2} \log n \sim R^* \)

\[
q_{KT}(x^n) = \int_0^1 \theta^k (1 - \theta)^{n-k} \frac{1}{\sqrt{\theta(1 - \theta)}} d\theta \quad \text{and} \quad q_{KT}(1 \mid x^n) = \frac{k + 1/2}{n + 1}
\]

- For \( m \)-ary sources, \( R^* \sim \frac{(m - 1)}{2} \log n \) (both stochastic and deterministic)
Bernoulli sources

- Let $X_1, X_2, \ldots$ be i.i.d. $\sim \text{Bern}(\theta)$, $\theta \in [0, 1]$ unknown

- **Uniform mixture (Laplace 1812):** $R \sim \log n$ (universal for Bernoulli sources!)

\[
q_L(x^n) = \int_0^1 \theta^k (1 - \theta)^{n-k} d\theta = \frac{1}{\binom{n}{k}(n+1)} \quad \text{and} \quad q_L(1|x^n) = \frac{k+1}{n+2}
\]

- **Jeffreys mixture (Krichevsky–Trofimov 1981):** $R \sim \frac{1}{2} \log n \sim R^*$

\[
q_{KT}(x^n) = \int_0^1 \theta^k (1 - \theta)^{n-k} \frac{1}{\sqrt{\theta(1 - \theta)}} d\theta \quad \text{and} \quad q_{KT}(1|x^n) = \frac{k + 1/2}{n + 1}
\]

- For $m$-ary sources, $R^* \sim \frac{(m-1)}{2} \log n$ (both stochastic and deterministic)

- Can be generalized to Markov and tree (CTM) sources (Willems et al. 1995)
Applications

- **Compression**: Compress to the entropy rate using *arithmetic coding*
Applications

- **Compression**: Compress to the entropy rate using arithmetic coding

- **Prediction**: Take an action $a(X^n)$ for $X_{n+1}$
  - Bayes response: $a^*(p) = \arg\min_a E_p[l(X, a)]$
  - Merhav–Feder (1998): Choose action $a^*(q(x_{n+1}|X_1^n))$
Applications

- **Compression**: Compress to the entropy rate using **arithmetic coding**

- **Prediction**: Take an action $a(X^n)$ for $X_{n+1}$
  - Bayes response: $a^*(p) = \arg\min_a E_p[l(X,a)]$
  - Merhav–Feder (1998): Choose action $a^*(q(x_{n+1}|X^n))$

- **Portfolio selection**: Choose asset allocation $b_x(Y^n)$ for stock $x$
  - Fund of funds: Multi-period asset allocation using $q(x^n)$
  - Cover (1991): Minimax performance against constant-rebalanced portfolios
Applications

- **Compression**: Compress to the entropy rate using arithmetic coding

- **Prediction**: Take an action $a(X^n)$ for $X_{n+1}$
  - **Bayes response**: $a^*(p) = \arg \min_a E_p[l(X, a)]$
  - **Merhav–Feder (1998)**: Choose action $a^*(q(x_{n+1}|X^n))$

- **Portfolio selection**: Choose asset allocation $b_x(Y^n)$ for stock $x$
  - **Fund of funds**: Multi-period asset allocation using $q(x^n)$
  - **Cover (1991)**: Minimax performance against constant-rebalanced portfolios

- **Entropy estimation**: Estimate the entropy rate of $\{X_n\}$
  - **Shannon–McMillan–Breiman theorem**: $\frac{1}{n} \log \frac{1}{p(X^n)} \to \bar{H}(X)$
  - **Plug-in strategy**: Use $q$ in place of $p$
Outline of the talk

- Brief overview of universal probability assignment
- Directed information and its application to causality inference
- Classification of DNA/RNA sequences using universal probability
Correlation and causation between time series

![Graph showing time series data with HSI and DJIA indices]
Correlation and causation between time series

Are they “correlated”?
Correlation and causation between time series

Are they “correlated”?  
Yes if $I(X; Y) \gg 0$
Correlation and causation between time series

Are they “correlated”?

Yes if $I(X; Y) \gg 0$

Which “leads” the other?
Correlation and causation between time series

Are they “correlated”?

Yes if $I(X; Y) \gg 0$

Which “leads” the other?

$X$ if $I(X \rightarrow Y) \gg I(Y \rightarrow X)$

$Y$ if $I(Y \rightarrow X) \gg I(X \rightarrow Y)$
Directed information

\[ I(X \rightarrow Y) = H(Y) - H(Y \parallel X) = \sum H(Y_i | Y^{i-1}) - H(Y_i | Y^{i-1}, X^i) \]
Directed information

$$I(X \rightarrow Y)$$

$$H(Y) - H(Y\|X)$$

$$= \sum H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1}, X^i)$$

- **Causal information** from $X$ to $Y$
  
  Marko (1966, 1973)
  Massey (1990)

- Cybernetics, feedback comm., ...
  (Kramer 1998, Permuter 2008)
Directed information

\[ I(X \rightarrow Y) \]

\[ H(Y) - H(Y \| X) \]

\[ = \sum H(Y_i | Y^{i-1}) - H(Y_i | Y^{i-1}, X^i) \]

G(X → Y)

\[ \sum \log \frac{\text{MSE}(Y_i | Y_{i-1}^{i-1})}{\text{MSE}(Y_i | Y_{i-p}^{i-1}, X_{i-p}^i)} \]

- **Causal information** from X to Y

Marko (1966, 1973)
Massey (1990)

- Cybernetics, feedback comm., …
(Kramer 1998, Permuter 2008)
Directed information

\[ I(X \rightarrow Y) = H(Y) - H(Y \| X) = \sum H(Y_i | Y^i-1) - H(Y_i | Y^i-1, X^i) \]

- **Causal information** from \( X \) to \( Y \)
  - Marko (1966, 1973)
  - Massey (1990)
  - Cybernetics, feedback comm., …
    (Kramer 1998, Permuter 2008)

\[ G(X \rightarrow Y) = \sum \log \frac{\text{MSE}(Y_i | Y_{i-1}^i)}{\text{MSE}(Y_i | Y_{i-p}^{i-1}, X_{i-p}^i)} \]

- **Causal influence** of \( X \) on \( Y \)
  - Granger (1969)
  - Geweke (1982)
  - Econometrics, neuroscience, …
    (Sims 1972, Quinn et al. 2011)
Directed information

\[ I(X \rightarrow Y) \]
\[ = H(Y) - H(Y\|X) = \sum H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1}, X^i) \]

- **Causal information** from \( X \) to \( Y \)
  - Marko (1966, 1973)
  - Massey (1990)

- **Cybernetics, feedback comm., …**
  - (Kramer 1998, Permuter 2008)

- **Other interpretations** (Permuter–Kim–Weissman 2011, Kamath–Kim 2014)

\[ G(X \rightarrow Y) \]
\[ = \sum \log \frac{\text{MSE}(Y_i|Y_{i-1}^{i-1})}{\text{MSE}(Y_i|Y_{i-p}^{i-1}, X_{i-p}^i)} \]

- **Causal influence** of \( X \) on \( Y \)
  - Granger (1969)
  - Geweke (1982)

- **Econometrics, neuroscience, …**
  - (Sims 1972, Quinn et al. 2011)
Directed information

\[ I(X \rightarrow Y) \]

\[ H(Y) - H(Y \| X) = \sum H(Y_i | Y^{i-1}) - H(Y_i | Y^{i-1}, X^i) \]

- **Causal information** from \( X \) to \( Y \)
  
  Marko (1966, 1973)
  Massey (1990)

- Cybernetics, feedback comm., …
  (Kramer 1998, Permuter 2008)

- Other interpretations (Permuter–Kim–Weissman 2011, Kamath–Kim 2014)

- Can be generalized to continuous time (Weissman–Kim–Permuter 2013)

\[ G(X \rightarrow Y) \]

\[ \sum \log \frac{\text{MSE}(Y_i | Y_{i-1}^{i-p})}{\text{MSE}(Y_i | Y_{i-p}^{i-1}, X_{i-p}^i)} \]

- **Causal influence** of \( X \) on \( Y \)
  
  Granger (1969)
  Geweke (1982)

- Econometrics, neuroscience, …
  (Sims 1972, Quinn et al. 2011)
**Directed information**

\[ I(X \rightarrow Y) \]

\[
H(Y) - H(Y|X) = \sum H(Y_i|Y_i^{i-1}) - H(Y_i|Y_i^{i-1}, X^i)
\]

- **Causal information** from \( X \) to \( Y \)
  - Marko (1966, 1973)
  - Massey (1990)

- **Causal influence** of \( X \) on \( Y \)
  - Granger (1969)
  - Geweke (1982)

- **Cybernetics, feedback comm., …**
  - (Kramer 1998, Permuter 2008)

- **Econometrics, neuroscience, …**
  - (Sims 1972, Quinn et al. 2011)

- **Other interpretations** (Permuter–Kim–Weissman 2011, Kamath–Kim 2014)

- **Can be generalized to continuous time** (Weissman–Kim–Permuter 2013)

- **Conservation law:**
  \[ I(X; Y) = I(X \rightarrow Y) + I(Y \rightarrow X) \]

\[ G(X \rightarrow Y) \]

\[
\sum \log \frac{\text{MSE}(Y_i|Y_i^{i-1})}{\text{MSE}(Y_i|Y_i^{i-1}, X_i^{i-p})}
\]

12/22
**Directed information estimation (Jiao et al. 2013)**

### Algorithm 1

\[
\hat{I}_1(X \rightarrow Y) = \hat{H}_1(Y) - \hat{H}_1(Y \| X)
\]

\[
\frac{1}{n} \log \frac{1}{q(Y^n)}
\]

- Very good convergence (a.s. & \(L_1\))
- Erratic for small \(n\)
- Unbounded support

### Algorithm 2

\[
\hat{I}_2(X \rightarrow Y) = \hat{H}_2(Y) - \hat{H}_2(Y \| X)
\]

\[
\frac{1}{n} \sum_{i=1}^{n} H(q(y_i | Y^{i-1}))
\]

- Similar convergence rate
- Smooth and bounded support
- Can be negative

### Algorithms 3 & 4

\[
\hat{I}_3(X \rightarrow Y) = \frac{1}{n} \sum_{i=1}^{n} D(q(y_i | X^i, Y^{i-1}) \| q(y_i | Y^{i-1}))
\]

\[
\hat{I}_4(X \rightarrow Y) = \frac{1}{n} \sum_{i=1}^{n} D(q(x_i, y_i | X^i, Y^{i-1}) \| q(y_i | Y^{i-1})q(x_i | X^i, Y^i))
\]
HSI (X) versus DJIA (Y)

Algorithm 1

Algorithm 2

Algorithm 3

Algorithm 4
Outline of the talk

- Brief overview of universal probability assignment
- Directed information and its application to causality inference
- Classification of DNA/RNA sequences using universal probability
Classification of nucleic acid sequences

Query sequence
TTCTTTTGAGAGTTTGATCCTGGGCTC

Family 1
GACGAACGCTGGCGGCCTGCTTAACAC
CACATGCAAGTCAGCGGTAAGGGCT

Family 2
AGAGTTTGATCCCTGGGCTCAGGACGAAC
ATTGAACGCTGGCGGCATGCCTCATG

Family k
GATGAACGCTGACAGAACGCTTAACAC
GATGAACGCTGACAGAATGCTTACACATG
Classification of nucleic acid sequences

Query sequence

TTCTTTTGGAGAGTTTGATCCTGGGCT

Family 1

GACGAACGCTGGCGGCCTAGCTTAACAC
CACATGCAAGTCGACGCTGTAACGGGCT

Family 2

AGAGTTTGATCCTGGCTCAGGACGAAC
ATTGAACGCTGGCGGCATGCCTCATG

Family k

GATGAACGCTGACAGAAGCTTAACAC
GATGAACGCTGACAGAATGCTTACACATG
Classification of nucleic acid sequences

Query sequence

TTTTTTGGAGAGTTTGATCCTGGCTC

Family 1
GACGAACGCTGGCGGCCTGCTTAACAC
CACATGCAAGTCAGCCGCTAAGGGCT

Family 2
AGAGTTTGATCCTGGCTCAGGACGAAC
ATTGAACGCTGGCGGCATGCCTCATG

Family k
GATGAACGCTGACAGAACGCTTAACAC
GATGAACGCTGACAGAAATGCTTACACATG

- Alignment-based methods: BLAST, USEARCH, UBLAST, caBLAST, BLAT, ...

- Model/feature-based methods: nhmmer, ICM, RDP, …
Classification of nucleic acid sequences

Query sequence

\[ X \]

Family 1
\[ Y_1 \sim P_1 \]

Family 2
\[ Y_2 \sim P_2 \]

\[ \vdots \]

Family \( k \)
\[ Y_k \sim P_k \]
Classification of nucleic acid sequences

Query sequence $X$

Family 1
$Y_1 \sim P_1$

Family 2
$Y_2 \sim P_2$

...$

Family k
$Y_k \sim P_k$

Were $P_1, \ldots, P_k$ known ...

$$j^* = \arg\max_j P_j(X)$$
Method

- Context tree models
Method

- Context tree models
Method

- Context tree models

\[ p_A, p_C, p_G, p_T \]
Method

- Context tree models

```
AA CA GA TA
AT CT GT TT
```

\[ p_A, p_C, p_G, p_T \]

\[ \cdots \]

\[ \cdots \]

\[ \text{For each family } j \text{ and its sequence } Y_j \]

\[ \text{find the best context tree model } \]

\[ M_j^* = \arg \max_M Q_M(Y_j) \]

- \( Q_M \): Universal prob. for model \( M \)

\[ Q_{M_j^*} \approx P_j \]

- Simple recursive maximization
Method

- Context tree models

For each family $j$ and its sequence $Y_j$ find the best context tree model

$$M_j^* = \arg \max_M Q_M(Y_j)$$

- $Q_M$: Universal prob. for model $M$

$$Q_{M_j^*}\approx P_j$$

- Simple recursive maximization

Modeling

Classification

Given a query sequence $X$ find the best family

$$j^* = \arg \max_j Q_{M_j^*}(X|Y_j)$$

- Close approximation of ML

$$Q_{M_j^*}(X|Y_j) \approx P_j(X|Y_j) \approx P_j(X)$$

- Simple Bayesian update (Dirichlet)
### Performance highlights

- Nine RNA datasets of different types (including large pyrosequencing databases)

<table>
<thead>
<tr>
<th>Classification category of the dataset</th>
<th>ID</th>
<th>Dataset name (version)</th>
<th>AIFD ‡</th>
<th># families*</th>
<th># total † sequences</th>
<th>Sequence length</th>
<th>Ground truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional non-coding RNA</td>
<td>RF</td>
<td>Rfam (11.0)</td>
<td>0.33</td>
<td>1,320</td>
<td>170,881</td>
<td>20–1,875</td>
<td>Accession</td>
</tr>
<tr>
<td>RRNA database (16S, 18S, 23S/28S)</td>
<td>RD</td>
<td>RDP (10.0)</td>
<td>0.08</td>
<td>134</td>
<td>3,838</td>
<td>320–1,833</td>
<td>Taxonomy</td>
</tr>
<tr>
<td>GG</td>
<td></td>
<td>Greengenes (13.5)</td>
<td>0.12</td>
<td>464</td>
<td>23,142</td>
<td>1,254–2,146</td>
<td>Genus level</td>
</tr>
<tr>
<td>SS</td>
<td></td>
<td>SILVA-SSU (119.1)</td>
<td>0.15</td>
<td>313</td>
<td>17,625</td>
<td>902–3,749</td>
<td></td>
</tr>
<tr>
<td>SL</td>
<td></td>
<td>SILVA-LSU (119)</td>
<td>0.21</td>
<td>107</td>
<td>4,593</td>
<td>1,900–4,954</td>
<td></td>
</tr>
<tr>
<td>Pyrosequencing data (16S rRNA)</td>
<td>AR</td>
<td>Artificial</td>
<td>0.18</td>
<td>60</td>
<td>44,407</td>
<td>40–294</td>
<td>Reference</td>
</tr>
<tr>
<td></td>
<td>DV</td>
<td>Divergent</td>
<td>0.14</td>
<td>23</td>
<td>55,466</td>
<td>38–521</td>
<td></td>
</tr>
<tr>
<td>Coding/non-coding RNA</td>
<td>CN</td>
<td>RefSeq,Rfam</td>
<td>0.60</td>
<td>2</td>
<td>103,136</td>
<td>22–9,993</td>
<td>Specified</td>
</tr>
<tr>
<td></td>
<td>HS</td>
<td>Ensembl (human)</td>
<td>0.67</td>
<td>2</td>
<td>112,180</td>
<td>20–15,945</td>
<td></td>
</tr>
</tbody>
</table>

* the number of families with more than 10 sequences
† the total number of sequences after the preprocessing
‡ average intra-family distance (the normalized pairwise distance between the sequences within a family)
Performance highlights

- Nine RNA datasets of different types (including large pyrosequencing databases)
- Comparison to 9 existing methods (BLAST, RDP, USEARCH, HMMER, ICM, …)
Performance highlights

- Nine RNA datasets of different types (including large pyrosequeencing databases)

- Comparison to 9 existing methods (BLAST, RDP, USEARCH, HMMER, ICM, …)

- **Accuracy** of 95.2% (next to 96.5% of BLAST)

<table>
<thead>
<tr>
<th>method/data</th>
<th>RF</th>
<th>RD</th>
<th>GG</th>
<th>SS</th>
<th>SL</th>
<th>AR</th>
<th>DV</th>
<th>CN</th>
<th>HS</th>
<th>average</th>
<th>geomean</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLAST</td>
<td>95.8%</td>
<td>98.5%</td>
<td>98.4%</td>
<td>96.3%</td>
<td>97.4%</td>
<td>96.5%</td>
<td>98.8%</td>
<td>92.9%</td>
<td>94.2%</td>
<td>96.5%</td>
<td>96.5%</td>
</tr>
<tr>
<td>NASCUP</td>
<td>95.8%</td>
<td><strong>99.0%</strong></td>
<td>98.2%</td>
<td><strong>96.8%</strong></td>
<td>96.5%</td>
<td><strong>97.6%</strong></td>
<td>99.0%</td>
<td>89.3%</td>
<td>84.5%</td>
<td><strong>95.2%</strong></td>
<td><strong>95.1%</strong></td>
</tr>
<tr>
<td>USEARCH</td>
<td><strong>96.5%</strong></td>
<td>98.6%</td>
<td><strong>98.6%</strong></td>
<td>96.7%</td>
<td>97.4%</td>
<td>89.8%</td>
<td>98.8%</td>
<td>26.7%</td>
<td>84.3%</td>
<td>87.5%</td>
<td>82.5%</td>
</tr>
<tr>
<td>UBLAST</td>
<td>79.9%</td>
<td>98.5%</td>
<td>97.9%</td>
<td>95.9%</td>
<td>97.1%</td>
<td>96.4%</td>
<td>98.7%</td>
<td>23.1%</td>
<td>88.2%</td>
<td>86.2%</td>
<td>80.3%</td>
</tr>
<tr>
<td>RDP</td>
<td>52.6%</td>
<td>99.0%</td>
<td>98.3%</td>
<td>96.5%</td>
<td>96.9%</td>
<td>97.1%</td>
<td><strong>99.1%</strong></td>
<td>60.6%</td>
<td>70.7%</td>
<td>85.6%</td>
<td>83.5%</td>
</tr>
<tr>
<td>BLAT</td>
<td>79.1%</td>
<td>97.2%</td>
<td>92.1%</td>
<td>92.3%</td>
<td>95.2%</td>
<td>94.7%</td>
<td>98.9%</td>
<td>18.7%</td>
<td>87.8%</td>
<td>84.0%</td>
<td>77.1%</td>
</tr>
<tr>
<td>ICM</td>
<td>93.6%</td>
<td>77.5%</td>
<td>76.7%</td>
<td>39.5%</td>
<td>93.2%</td>
<td>95.0%</td>
<td>98.9%</td>
<td>92.8%</td>
<td>87.3%</td>
<td>83.8%</td>
<td>81.4%</td>
</tr>
<tr>
<td>gzip</td>
<td>62.7%</td>
<td>96.3%</td>
<td>90.3%</td>
<td>80.1%</td>
<td>77.6%</td>
<td>80.9%</td>
<td>96.3%</td>
<td>59.5%</td>
<td>69.1%</td>
<td>79.2%</td>
<td>78.2%</td>
</tr>
<tr>
<td>caBLAST</td>
<td>39.4%</td>
<td>97.1%</td>
<td>86.9%</td>
<td>90.5%</td>
<td>93.9%</td>
<td>95.5%</td>
<td>97.0%</td>
<td>18.8%</td>
<td>82.8%</td>
<td>78.0%</td>
<td>70.1%</td>
</tr>
<tr>
<td>HMMER</td>
<td>96.1%</td>
<td>98.4%</td>
<td>80.1%</td>
<td>14.9%</td>
<td>80.3%</td>
<td>41.9%</td>
<td>64.0%</td>
<td>#</td>
<td>#</td>
<td>68.0%</td>
<td>58.8%</td>
</tr>
</tbody>
</table>
Performance highlights

- Nine RNA datasets of different types (including large pyrosequencing databases)
- Comparison to 9 existing methods (BLAST, RDP, USEARCH, HMMER, ICM, …)
- Accuracy of 95.2% (next to 96.5% of BLAST)
- Scalability
Concluding remarks

Maslow’s axiom (1966)

If all you have is a hammer, everything looks like a nail.
Concluding remarks

Maslow’s axiom (1966)

If all you have is a hammer, everything looks like a nail.

- Our hammer: Universal probability $q$
Concluding remarks

Maslow’s axiom (1966)

If all you have is a hammer, everything looks like a nail.

- Our hammer: Universal probability $q$
- Versatile and often on par with custom tools
Concluding remarks

Maslow’s axiom (1966)
If all you have is a hammer, everything looks like a nail.

- Our hammer: Universal probability $q$
- Versatile and often on par with custom tools
- Many classical results, but still more to explore
Concluding remarks

Maslow’s axiom (1966)

If all you have is a hammer, everything looks like a nail.

- Our hammer: Universal probability $q$
- Versatile and often on par with custom tools
- Many classical results, but still more to explore

Towards information-theoretic data science
References


References (cont.)


