Universal Information Processing

Young-Han Kim

Department of Electrical and Computer Engineering

University of California, San Diego

Pattern Recognition & Machine Learning Summer School
August 23, 2013

Information processing system

- Tasks: Compression, prediction, portfolio selection, filtering, estimation, denoising, decoding, classification, closeness testing, control, …

Information processing system

- Tasks: Compression, prediction, portfolio selection, filtering, estimation, denoising, decoding, classification, closeness testing, control, …
- Algorithms:
  - Causal vs. noncausal
  - Deterministic vs. randomized

Information processing system

- Tasks: Compression, prediction, portfolio selection, filtering, estimation, denoising, decoding, classification, closeness testing, control, …
- Algorithms:
  - Causal vs. noncausal
  - Deterministic vs. randomized
- Universality:
  - Probabilistic setting: "Optimal" performance when the source distribution is unknown
  - Deterministic setting: "Optimal" performance among a class of algorithms

Information processing system

- Tasks: Compression, prediction, portfolio selection, filtering, estimation, denoising, decoding, classification, closeness testing, control, …
- Algorithms:
  - Causal vs. noncausal
  - Deterministic vs. randomized
- Universality:
  - Probabilistic setting: "Optimal" performance when the source distribution is unknown
  - Deterministic setting: "Optimal" performance among a class of algorithms
- Questions:
  - Existence: Does there exist a universal algorithm for a given task?

Information processing system

- Tasks: Compression, prediction, portfolio selection, filtering, estimation, denoising, decoding, classification, closeness testing, control, …
- Algorithms:
  - Causal vs. noncausal
  - Deterministic vs. randomized
- Universality:
  - Probabilistic setting: "Optimal" performance when the source distribution is unknown
  - Deterministic setting: "Optimal" performance among a class of algorithms
- Questions:
  - Existence: Does there exist a universal algorithm for a given task?
  - Complexity: Is it practical to implement?

Information processing system

- Tasks: Compression, prediction, portfolio selection, filtering, estimation, denoising, decoding, classification, closeness testing, control, …

- Algorithms:
  - Causal vs. noncausal
  - Deterministic vs. randomized

- Universality:
  - Probabilistic setting: "Optimal" performance when the source distribution is unknown
  - Deterministic setting: "Optimal" performance among a class of algorithms

- Questions:
  - Existence: Does there exist a universal algorithm for a given task?
  - Complexity: Is it practical to implement?
  - Rate of convergence and nonasymptotic behavior: Does it perform "well" in real life?

Probabilistic setting

- Source
  - Parametric: $X \sim p_\theta(x)$, $\theta \in \mathcal{T}$

Probabilistic setting

- Source
  - Parametric: $X \sim p_\theta(x)$, $\theta \in \mathcal{T}$
  - Nonparametric: i.i.d., Markov, hidden Markov, stationary ergodic, …

Probabilistic setting

- Source
  - Parametric: $X \sim p_\theta(x)$, $\theta \in \mathcal{T}$
  - Nonparametric: i.i.d., Markov, hidden Markov, stationary ergodic, …
- Game: Nature chooses the distribution

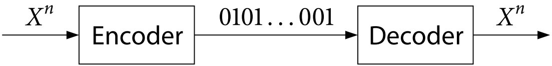
Probabilistic setting

- Source
  - Parametric: $X \sim p_\theta(x)$, $\theta \in \mathcal{T}$
  - Nonparametric: i.i.d., Markov, hidden Markov, stationary ergodic, …
- Game: Nature chooses the distribution
- Performance: Measured by an expected cost or reward

Probabilistic setting

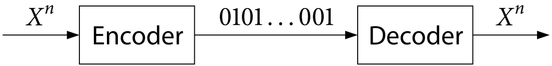
- Source
  - Parametric: $X \sim p_\theta(x)$, $\theta \in \mathcal{T}$
  - Nonparametric: i.i.d., Markov, hidden Markov, stationary ergodic, …
- Game: Nature chooses the distribution
- Performance: Measured by an expected cost or reward
- Goal: Achieve performance of optimal algorithm without prior knowledge of X

Example: Universal compression



- Let $X = \{X_i\}$ be a stationary ergodic source over a finite alphabet $\mathcal{X}$

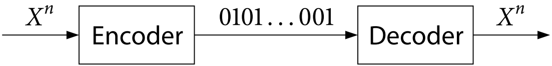
Example: Universal compression



- Let $X = \{X_i\}$ be a stationary ergodic source over a finite alphabet $\mathcal{X}$
- Minimum compression rate (average number of bits per source symbol):

$$\overline{H}(X) = \lim_{n\to\infty} \frac{1}{n} H(X^n)$$

Example: Universal compression

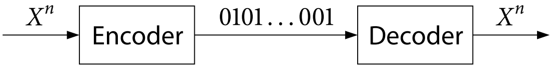


- Let $X = \{X_i\}$ be a stationary ergodic source over a finite alphabet $\mathcal{X}$
- Minimum compression rate (average number of bits per source symbol):

$$\overline{H}(X) = \lim_{n\to\infty} \frac{1}{n} H(X^n)$$

- Universal compression algorithms achieve $\overline{H}(X)$ for all stationary ergodic X

Example: Universal compression

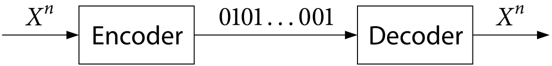


- Let $X = \{X_i\}$ be a stationary ergodic source over a finite alphabet $\mathcal{X}$
- Minimum compression rate (average number of bits per source symbol):

$$\overline{H}(X) = \lim_{n\to\infty} \frac{1}{n} H(X^n)$$

- Universal compression algorithms achieve $\overline{H}(X)$ for all stationary ergodic X
- Examples: Lempel–Ziv, Burrows–Wheeler transform, context-tree weighting

Example: Universal compression



- Let $X = \{X_i\}$ be a stationary ergodic source over a finite alphabet $\mathcal{X}$
- Minimum compression rate (average number of bits per source symbol):

$$\overline{H}(X) = \lim_{n\to\infty} \frac{1}{n} H(X^n)$$

- Universal compression algorithms achieve $\overline{H}(X)$ for all stationary ergodic X
- Examples: Lempel–Ziv, Burrows–Wheeler transform, context-tree weighting
- These algorithms can be implemented efficiently and perform well on real data
  - LZ78: `compress`, GIF, and TIFF
  - LZ77: `gzip`, PNG, and PDF
  - BWT: `bzip2`

Deterministic setting

- Source:
  - An arbitrary sequence
  - No statistical assumption (real world applications)

Deterministic setting

- Source:
  - An arbitrary sequence
  - No statistical assumption (real world applications)
- Game: Nature chooses the sequence (often maliciously)

Deterministic setting

- Source:
  - An arbitrary sequence
  - No statistical assumption (real world applications)
- Game: Nature chooses the sequence (often maliciously)
- Performance: Measured by a cost or reward for the specific sequence

Deterministic setting

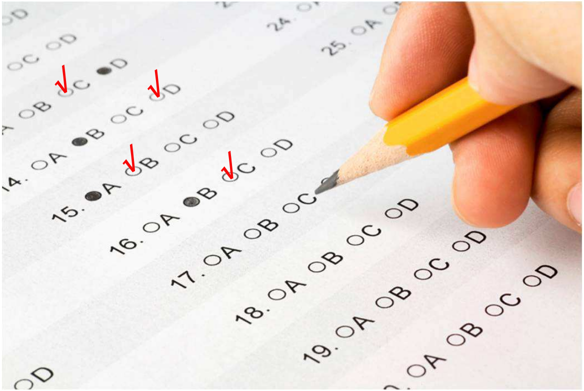
- Source:
  - An arbitrary sequence
  - No statistical assumption (real world applications)
- Game: Nature chooses the sequence (often maliciously)
- Performance: Measured by a cost or reward for the specific sequence
- Reference class of algorithms
  - Memoryless, sliding-window, finite-state, …
  - More generally, any class of "experts"

Deterministic setting

- Source:
  - An arbitrary sequence
  - No statistical assumption (real world applications)
- Game: Nature chooses the sequence (often maliciously)
- Performance: Measured by a cost or reward for the specific sequence
- Reference class of algorithms
  - Memoryless, sliding-window, finite-state, …
  - More generally, any class of "experts"
- Goal: Achieve performance of the optimal algorithm for each sequence

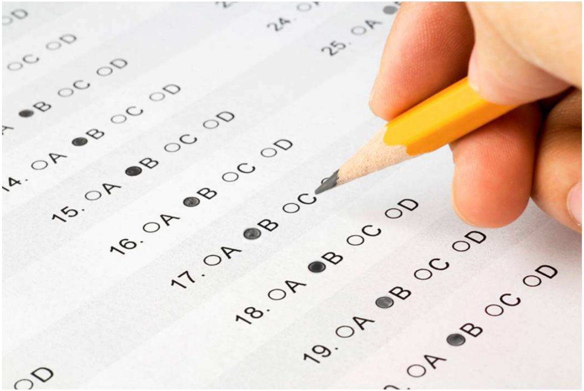
Example: Sequential prediction

- Multiple choice exam: Guess the answers one by one
- True answers (A, B, C, or D) are revealed right after



Example: Sequential prediction

- Multiple choice exam: Guess the answers one by one
- True answers (A, B, C, or D) are revealed right after
- Reference class of algorithms: Constant testers



Example: Sequential prediction

- Multiple choice exam: Guess the answers one by one
- True answers (A, B, C, or D) are revealed right after
- Reference class of algorithms: Constant testers
- Universal prediction algorithms achieve score of best tester for every exam

Example: Sequential prediction

- Multiple choice exam: Guess the answers one by one
- True answers (A, B, C, or D) are revealed right after
- Reference class of algorithms: Constant testers
- Universal prediction algorithms achieve score of best tester for every exam
- Examples: Cover's binary prediction, Feder–Merhav–Gutman algorithm

This tutorial

- Universal information processing: Get something out of nothing!
  - Intersection of statistics, information theory, and learning theory
  - Many different terminologies, problems, approaches, and flavors

This tutorial

- Universal information processing: Get something out of nothing!
  - Intersection of statistics, information theory, and learning theory
  - Many different terminologies, problems, approaches, and flavors

- Goal: Provide a gentle overview of information-theoretic approaches

Outline

- Review of information measures
- Lossless compression and probability assignment (probabilistic / deterministic)
- Portfolio selection (deterministic)
- Sequential prediction (probabilistic)

Entropy

- Entropy of a discrete random variable $X \sim p(x)$ (pmf), $X \in \mathcal{X}$:

$$H(X) = -\sum p(x) \log p(x) = -\mathsf{E}_X \left(\log p(X) \right)$$

  - Nonnegative and concave function of $p(x)$
  - $H(X) \leq \log |\mathcal{X}|$ (by Jensen's inequality)

Entropy

- Entropy of a discrete random variable $X \sim p(x)$ (pmf), $X \in \mathcal{X}$:

$$H(X) = -\sum p(x) \log p(x) = -\mathsf{E}_X \left(\log p(X)\right)$$

  - Nonnegative and concave function of $p(x)$
  - $H(X) \le \log |\mathcal{X}|$ (by Jensen's inequality)

- Binary entropy function: For $p \in [0, 1]$,

$$H(p) = -p \log p - (1-p) \log(1-p)$$

Entropy

- Entropy of a discrete random variable $X \sim p(x)$ (pmf), $X \in \mathcal{X}$:

$$H(X) = -\sum p(x) \log p(x) = -\mathsf{E}_X \left(\log p(X)\right)$$

  - Nonnegative and concave function of $p(x)$
  - $H(X) \le \log |\mathcal{X}|$ (by Jensen's inequality)

- Binary entropy function: For $p \in [0, 1]$,

$$H(p) = -p \log p - (1-p) \log(1-p)$$

- Conditional entropy (equivocation): If $X \sim F(x)$ and $Y \,|\, \{X = x\} \sim p(y|x)$,

$$H(Y|X) = \int H(Y \,|\, X = x)\, dF(x) = -\mathsf{E}_{X,Y}\left(\log p(Y|X)\right)$$

  - $H(Y|X) \le H(Y)$ (with equality if X and Y are independent)

Entropy

- Joint entropy: If $(X, Y) \sim p(x, y)$,

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

Entropy

- Joint entropy: If $(X, Y) \sim p(x, y)$,

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

- Chain rule:

$$H(X^n) = \sum_{i=1}^{n} H(X_i|X^{i-1})$$

  - $H(X^n) \le \sum_{i=1}^{n} H(X_i)$ (with equality if $X_1, X_2, \ldots, X_n$ are independent)

Entropy

- Joint entropy: If $(X, Y) \sim p(x, y)$,

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

- Chain rule:

$$H(X^n) = \sum_{i=1}^{n} H(X_i|X^{i-1})$$

  - $H(X^n) \leq \sum_{i=1}^{n} H(X_i)$ (with equality if $X_1, X_2, \ldots, X_n$ are independent)

- Entropy rate: For a stationary random process $X = \{X_i\}$,

$$\overline{H}(X) = \lim_{n\to\infty} \frac{1}{n} H(X^n) = \lim_{n\to\infty} H(X_n|X^{n-1})$$

Relative entropy

- Relative entropy (Kullback-Leibler divergence) of a pair of pmfs $p(x)$ and $q(x)$:

$$\begin{aligned} D(p\|q) &= D(p(x)\|q(x)) \\ &= \sum p(x) \log \frac{p(x)}{q(x)} \\ &= \mathsf{E}\left[\log \frac{p(X)}{q(X)}\right] \end{aligned}$$

  - Nonnegativity: $D(p\|q) \geq 0$ and $D(p\|q) = 0$ iff $p \equiv q$
  - Convexity: $D(p\|q)$ is convex in (p, q), i.e., for any (p_1, q_1), (p_2, q_2), $\lambda \in [0, 1]$,

$$\lambda D(p_1\|q_1) + \bar{\lambda} D(p_2\|q_2) \geq D(\lambda p_1 + \bar{\lambda} p_2 \| \lambda q_1 + \bar{\lambda} q_2)$$

  - Chain rule: For any $p(x, y)$ and $q(x, y)$,

$$D(p(x,y)\|q(x,y)) = D(p(x)\|q(x)) + \sum_x p(x) D(p(y|x)\|q(y|x))$$

Mutual information

- Mutual information between $(X, Y) \sim p(x, y)$:

$$
\begin{aligned}
I(X; Y) &= D\big(p(x, y) \| p(x)p(y)\big) \\
&= \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\
&= \sum_{x} p(x) D(p(y|x) \| p(y)) \qquad \text{(chain rule)} \\
&= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X)
\end{aligned}
$$

  - Nonnegative function of $p(x, y) = p(x)p(y|x)$
  - Concave in $p(x)$ for fixed $p(y|x)$ and convex in $p(y|x)$ for fixed $p(x)$

Mutual information

- Mutual information between $(X, Y) \sim p(x, y)$:

$$\begin{aligned} I(X; Y) &= D\big(p(x,y) \| p(x)p(y)\big) \\ &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x} p(x) D(p(y|x) \| p(y)) && \text{(chain rule)} \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned}$$

  - Nonnegative function of $p(x, y) = p(x)p(y|x)$
  - Concave in $p(x)$ for fixed $p(y|x)$ and convex in $p(y|x)$ for fixed $p(x)$
  - Information capacity of a channel (conditional pmf) $p(y|x)$:

$$\begin{aligned} \max_{p(x)} I(X; Y) &= \max_{p(x)} \min_{q(y)} \sum_{x} p(x) D(p(y|x) \| q(y)) && \text{(nonnegativity)} \\ &= \min_{q(y)} \max_{p(x)} \sum_{x} p(x) D(p(y|x) \| q(y)) && \text{(minimax theorem)} \end{aligned}$$

Outline

- Review of information measures

- Lossless compression and probability assignment (probabilistic / deterministic)

- Portfolio selection (deterministic)

- Sequential prediction (probabilistic)

Lossless source codes

- Let $\mathcal{X}$ be a finite alphabet

Lossless source codes

- Let $\mathcal{X}$ be a finite alphabet
- Binary source code $C : \mathcal{X}^* \to \{0,1\}^*$

Lossless source codes

- Let $\mathcal{X}$ be a finite alphabet
- Binary source code $C : \mathcal{X}^* \to \{0,1\}^*$
- Prefix (instantaneous) code $C : \mathcal{X} \to \{0,1\}^*$
  - No codeword is a prefix of another

Lossless source codes

- Let $\mathcal{X}$ be a finite alphabet
- Binary source code $C : \mathcal{X}^* \to \{0,1\}^*$
- Prefix (instantaneous) code $C : \mathcal{X} \to \{0,1\}^*$
  - No codeword is a prefix of another
- Example:
  - $\mathcal{X} = \{a, b, c\}$, $C(a) = 0$, $C(b) = 10$, $C(c) = 111$
  - $C(acab) = 0111010$

Lossless source codes

- Let $\mathcal{X}$ be a finite alphabet
- Binary source code $C : \mathcal{X}^* \to \{0,1\}^*$
- Prefix (instantaneous) code $C : \mathcal{X} \to \{0,1\}^*$
  - No codeword is a prefix of another
- Example:
  - $\mathcal{X} = \{a, b, c\}$, $C(a) = 0$, $C(b) = 10$, $C(c) = 111$
  - $C(acab) = 0111010$
- Kraft inequality: $l(x) = |C(x)|$ is the length function of a prefix code iff

$$\sum 2^{-l(x)} \le 1$$

Lossless source codes

- Let $\mathcal{X}$ be a finite alphabet
- Binary source code $C : \mathcal{X}^* \to \{0,1\}^*$
- Prefix (instantaneous) code $C : \mathcal{X} \to \{0,1\}^*$
  - No codeword is a prefix of another
- Example:
  - $\mathcal{X} = \{a, b, c\}$, $C(a) = 0$, $C(b) = 10$, $C(c) = 111$
  - $C(acab) = 0111010$
- Kraft inequality: $l(x) = |C(x)|$ is the length function of a prefix code iff

$$\sum 2^{-l(x)} \leq 1$$

- Example: $1/2 + 1/4 + 1/8 < 1$

Average codeword length

- Performance of a prefix code is measured by its average codeword length

$$L = \mathsf{E}(l(X)) = \sum p(x)l(x)$$

Average codeword length

- Performance of a prefix code is measured by its average codeword length

$$L = \mathsf{E}(l(X)) = \sum p(x)l(x)$$

- Relationship between $l(x)$ and $p(x)$
  - $l(x)$ should be small for large $p(x)$ and large for small $p(x)$
  - The Kraft inequality suggests

$$\begin{aligned} p(x) &\Leftrightarrow 2^{-l(x)}, \\ \log(1/p(x)) &\Leftrightarrow l(x) \end{aligned}$$

Average codeword length of the optimal code

Theorem

If $C^*(x)$ is a prefix code that minimizes $L = \mathsf{E}\, l(X) = \sum p(x) l(x)$, then

$$H(X) \le L^* < H(X) + 1$$

Average codeword length of the optimal code

Theorem

If $C^*(x)$ is a prefix code that minimizes $L = \mathsf{E}\, l(X) = \sum p(x) l(x)$, then

$$H(X) \le L^* < H(X) + 1$$

- Entropy is the fundamental limit for prefix codes (up to 1 bit)

Average codeword length of the optimal code

Theorem

If $C^*(x)$ is a prefix code that minimizes $L = \mathsf{E}\, l(X) = \sum p(x) l(x)$, then

$$H(X) \le L^* < H(X) + 1$$

- Entropy is the fundamental limit for prefix codes (up to 1 bit)
- If $X = \{X_i\}$ is stationary ergodic, then $L^*(X^n)/n \to \overline{H}(X)$ (entropy rate)

Average codeword length of the optimal code

Theorem

If $C^*(x)$ is a prefix code that minimizes $L = \mathsf{E}\, l(X) = \sum p(x) l(x)$, then

$$H(X) \le L^* < H(X) + 1$$

- Entropy is the fundamental limit for prefix codes (up to 1 bit)
- If $X = \{X_i\}$ is stationary ergodic, then $L^*(X^n)/n \to \overline{H}(X)$ (entropy rate)
- Lower bound: Kraft inequality and nonnegativity of relative entropy

Average codeword length of the optimal code

Theorem

If $C^*(x)$ is a prefix code that minimizes $L = \mathsf{E}\, l(X) = \sum p(x) l(x)$, then

$$H(X) \le L^* < H(X) + 1$$

- Entropy is the fundamental limit for prefix codes (up to 1 bit)
- If $X = \{X_i\}$ is stationary ergodic, then $L^*(X^n)/n \to \overline{H}(X)$ (entropy rate)
- Lower bound: Kraft inequality and nonnegativity of relative entropy
- Upper bound: Consider $l(x) = \lceil \log(1/p(x)) \rceil < \log(1/p(x)) + 1$

Average codeword length of the optimal code

Theorem

If $C^*(x)$ is a prefix code that minimizes $L = \mathsf{E}\, l(X) = \sum p(x)l(x)$, then

$$H(X) \le L^* < H(X) + 1$$

- Entropy is the fundamental limit for prefix codes (up to 1 bit)
- If $X = \{X_i\}$ is stationary ergodic, then $L^*(X^n)/n \to \overline{H}(X)$ (entropy rate)
- Lower bound: Kraft inequality and nonnegativity of relative entropy
- Upper bound: Consider $l(x) = \lceil \log(1/p(x)) \rceil < \log(1/p(x)) + 1$
- Equivalence between compression and probability assignment:

$$2^{-l^*(x)} \simeq p(x) \quad \text{or equivalently} \quad l^*(x) \simeq \log(1/p(x))$$

Average codeword length of the optimal code

Theorem

If $C^*(x)$ is a prefix code that minimizes $L = \mathsf{E}\, l(X) = \sum p(x) l(x)$, then

$$H(X) \le L^* < H(X) + 1$$

- Entropy is the fundamental limit for prefix codes (up to 1 bit)
- If $X = \{X_i\}$ is stationary ergodic, then $L^*(X^n)/n \to \overline{H}(X)$ (entropy rate)
- Lower bound: Kraft inequality and nonnegativity of relative entropy
- Upper bound: Consider $l(x) = \lceil \log(1/p(x)) \rceil < \log(1/p(x)) + 1$
- Equivalence between compression and probability assignment:

$$2^{-l^*(x)} \simeq p(x) \quad \text{or equivalently} \quad l^*(x) \simeq \log(1/p(x))$$

- Arithmetic coding: Efficient translation of probabilities $p(x_i | x^{i-1})$ to code phrases

Source coding with mismatch

- Suppose $X \sim p(x)$

Source coding with mismatch

- Suppose $X \sim p(x)$
- But we design a code assuming $X \sim q(x)$, i.e., $|C(x)| = \log(1/q(x))$

Source coding with mismatch

- Suppose $X \sim p(x)$
- But we design a code assuming $X \sim q(x)$, i.e., $|C(x)| = \log(1/q(x))$
- What is the regret (redundancy) of this mismatch?

Source coding with mismatch

- Suppose $X \sim p(x)$
- But we design a code assuming $X \sim q(x)$, i.e., $|C(x)| = \log(1/q(x))$
- What is the regret (redundancy) of this mismatch?

- Let

$$L = \sum p(x) \log \frac{1}{q(x)},$$
$$L^* = \sum p(x) \log \frac{1}{p(x)} = H(X)$$

Then

$$R := L - L^* = \sum p(x) \log \frac{p(x)}{q(x)} = D(p\|q) \geq 0$$

Minimax redundancy

- Now suppose $X \sim p_\theta$ for some unknown $\theta \in \mathcal{T}$

Minimax redundancy

- Now suppose $X \sim p_\theta$ for some unknown $\theta \in \mathcal{T}$
- How should we design our code (i.e., choose $q(x)$)?

Minimax redundancy

- Now suppose $X \sim p_\theta$ for some unknown $\theta \in \mathcal{T}$
- How should we design our code (i.e., choose $q(x)$)?

Minimax redundancy theorem

$$R^* = \min_{q(x)} \max_{\theta \in \mathcal{T}} D(p_\theta \| q)$$
$$= \max_{F(\theta)} I(\Theta; X)$$

Moreover,

$$q^*(x) = \int p_\theta(x) dF^*(\theta)$$

Minimax redundancy

- Now suppose $X \sim p_\theta$ for some unknown $\theta \in \mathcal{T}$
- How should we design our code (i.e., choose $q(x)$)?

Minimax redundancy theorem

$$R^* = \min_{q(x)} \max_{\theta \in \mathcal{T}} D(p_\theta \| q)$$
$$= \max_{F(\theta)} I(\Theta; X)$$

Moreover,

$$q^*(x) = \int p_\theta(x) dF^*(\theta)$$

- Lagrange duality (KKT condition)

Minimax redundancy

- Now suppose $X \sim p_\theta$ for some unknown $\theta \in \mathcal{T}$
- How should we design our code (i.e., choose $q(x)$)?

Minimax redundancy theorem

$$\begin{aligned} R^* &= \min_{q(x)} \max_{\theta \in \mathcal{T}} D(p_\theta \| q) \\ &= \max_{F(\theta)} I(\Theta; X) \end{aligned}$$

Moreover,

$$q^*(x) = \int p_\theta(x) dF^*(\theta)$$

- Lagrange duality (KKT condition)
- Each $q(x)$ leads to an upper bound on R^*, while each $F(\theta)$ leads to a lower bound

Minimax redundancy of Bernoulli sources

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

Minimax redundancy of Bernoulli sources

- Let $X_1, X_2, \ldots$ be i.i.d. $\sim \mathrm{Bern}(\theta)$, $\theta \in [0,1]$
- If θ is known, then $L^*(X^n) = nH(\theta)$

Minimax redundancy of Bernoulli sources

- Let $X_1, X_2, \ldots$ be i.i.d. $\sim \mathrm{Bern}(\theta)$, $\theta \in [0,1]$
- If θ is known, then $L^*(X^n) = nH(\theta)$
- What is the best performance with no prior knowledge of θ?

Minimax redundancy of Bernoulli sources

- Let $X_1, X_2, \ldots$ be i.i.d. $\sim \mathrm{Bern}(\theta)$, $\theta \in [0,1]$
- If θ is known, then $L^*(X^n) = nH(\theta)$
- What is the best performance with no prior knowledge of θ?

Theorem

$$R^*(X^n) = \frac{1}{2}\log n + o(1)$$

Minimax redundancy of Bernoulli sources

- Let $X_1, X_2, \ldots$ be i.i.d. $\sim \mathrm{Bern}(\theta)$, $\theta \in [0,1]$
- If θ is known, then $L^*(X^n) = nH(\theta)$
- What is the best performance with no prior knowledge of θ?

Theorem

$$R^*(X^n) = \frac{1}{2}\log n + o(1)$$

- We focus on the upper bound on R^*

Laplace mixture

- We first try $\Theta \sim \text{Unif}\,[0, 1]$

Laplace mixture

- We first try $\Theta \sim \mathrm{Unif}\,[0,1]$
- Since

$$p_\theta(x^n) = \theta^{k(x^n)}(1-\theta)^{n-k(x^n)},$$

where $k = k(x^n)$ is the number of 1s in x^n, the mixture probability is

$$q_{\mathrm{L}}(x^n) = \int_0^1 \theta^k (1-\theta)^{n-k} d\theta = \frac{1}{\binom{n}{k}(n+1)}$$

Laplace mixture

- We first try $\Theta \sim \mathrm{Unif}[0,1]$
- Since
$$p_\theta(x^n) = \theta^{k(x^n)}(1-\theta)^{n-k(x^n)},$$
where $k = k(x^n)$ is the number of 1s in x^n, the mixture probability is
$$q_{\mathrm{L}}(x^n) = \int_0^1 \theta^k (1-\theta)^{n-k} d\theta = \frac{1}{\binom{n}{k}(n+1)}$$
- This mixture has a simple horizon-free sequential probability assignment, namely,
$$q_{\mathrm{L}}(1|x^n) = \frac{k+1}{n+2}$$
(What is the probability that the sun will rise tomorrow morning?)

Laplace mixture

- Corresponding redundancy:

$$R = \max_{\theta} D(p_{\theta} \| q)$$

Laplace mixture

- Corresponding redundancy:

$$R = \max_{\theta} D(p_\theta \| q)$$

$$= \max_{\theta} \sum_{x^n} p_\theta(x^n) \log \frac{p_\theta(x^n)}{q_{\mathrm{L}}(x^n)}$$

Laplace mixture

- Corresponding redundancy:

$$R = \max_{\theta} D(p_\theta \| q)$$

$$= \max_{\theta} \sum_{x^n} p_\theta(x^n) \log \frac{p_\theta(x^n)}{q_{\mathrm{L}}(x^n)}$$

$$= \max_{\theta} \sum_{x^n} \theta^k (1-\theta)^{n-k} \log \frac{\theta^k (1-\theta)^{n-k}}{q_{\mathrm{L}}(x^n)}$$

Laplace mixture

- Corresponding redundancy:

$$
\begin{aligned}
R &= \max_{\theta} D(p_\theta \| q) \\
&= \max_{\theta} \sum_{x^n} p_\theta(x^n) \log \frac{p_\theta(x^n)}{q_{\mathrm{L}}(x^n)} \\
&= \max_{\theta} \sum_{x^n} \theta^k (1-\theta)^{n-k} \log \frac{\theta^k (1-\theta)^{n-k}}{q_{\mathrm{L}}(x^n)} \\
&= \max_{\theta} \sum_{x^n} \theta^k (1-\theta)^{n-k} \log \left(\theta^k (1-\theta)^{n-k} \binom{n}{k} (n+1) \right)
\end{aligned}
$$

Laplace mixture

- Corresponding redundancy:

$$
\begin{aligned}
R &= \max_\theta D(p_\theta \| q) \\
&= \max_\theta \sum_{x^n} p_\theta(x^n) \log \frac{p_\theta(x^n)}{q_{\text{L}}(x^n)} \\
&= \max_\theta \sum_{x^n} \theta^k (1-\theta)^{n-k} \log \frac{\theta^k (1-\theta)^{n-k}}{q_{\text{L}}(x^n)} \\
&= \max_\theta \sum_{x^n} \theta^k (1-\theta)^{n-k} \log \left(\theta^k (1-\theta)^{n-k} \binom{n}{k} (n+1) \right) \\
&= \max_\theta \left[\sum_{k=0}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k} \log \left(\theta^k (1-\theta)^{n-k} \binom{n}{k} \right) \right] + \log(n+1)
\end{aligned}
$$

Laplace mixture

- Corresponding redundancy:

$$
\begin{aligned}
R &= \max_{\theta} D(p_\theta \| q) \\
&= \max_{\theta} \sum_{x^n} p_\theta(x^n) \log \frac{p_\theta(x^n)}{q_{\mathrm{L}}(x^n)} \\
&= \max_{\theta} \sum_{x^n} \theta^k (1-\theta)^{n-k} \log \frac{\theta^k (1-\theta)^{n-k}}{q_{\mathrm{L}}(x^n)} \\
&= \max_{\theta} \sum_{x^n} \theta^k (1-\theta)^{n-k} \log \left(\theta^k (1-\theta)^{n-k} \binom{n}{k} (n+1) \right) \\
&= \max_{\theta} \left[\sum_{k=0}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k} \log \left(\theta^k (1-\theta)^{n-k} \binom{n}{k} \right) \right] + \log(n+1) \\
&= \max_{\theta} \left[-H(\mathrm{Binom}(n,\theta)) \right] + \log(n+1)
\end{aligned}
$$

Laplace mixture

- Corresponding redundancy:

$$
\begin{aligned}
R &= \max_{\theta} D(p_\theta \| q) \\
&= \max_{\theta} \sum_{x^n} p_\theta(x^n) \log \frac{p_\theta(x^n)}{q_{\mathrm{L}}(x^n)} \\
&= \max_{\theta} \sum_{x^n} \theta^k (1-\theta)^{n-k} \log \frac{\theta^k (1-\theta)^{n-k}}{q_{\mathrm{L}}(x^n)} \\
&= \max_{\theta} \sum_{x^n} \theta^k (1-\theta)^{n-k} \log \left(\theta^k (1-\theta)^{n-k} \binom{n}{k} (n+1) \right) \\
&= \max_{\theta} \left[\sum_{k=0}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k} \log \left(\theta^k (1-\theta)^{n-k} \binom{n}{k} \right) \right] + \log(n+1) \\
&= \max_{\theta} [-H(\mathrm{Binom}(n,\theta))] + \log(n+1) \\
&\le \log(n+1)
\end{aligned}
$$

Jeffreys–Krichevsky–Trofimov mixture

- We now try $\Theta \sim$ Beta(1/2, 1/2), i.e., $f(\theta) = 1/(\pi\sqrt{\theta(1-\theta)})$

Jeffreys–Krichevsky–Trofimov mixture

- We now try $\Theta \sim \text{Beta}(1/2, 1/2)$, i.e., $f(\theta) = 1/(\pi\sqrt{\theta(1-\theta)})$
- Then, by Stirling's approximation,

$$q_{\text{JKT}}(x^n) = \int_0^1 \theta^k (1-\theta)^{n-k} f(\theta)\, d\theta \geq \frac{1}{\sqrt{2n}} \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k}$$

Jeffreys–Krichevsky–Trofimov mixture

- We now try $\Theta \sim \text{Beta}(1/2, 1/2)$, i.e., $f(\theta) = 1/(\pi\sqrt{\theta(1-\theta)})$
- Then, by Stirling's approximation,

$$q_{\text{JKT}}(x^n) = \int_0^1 \theta^k (1-\theta)^{n-k} f(\theta) d\theta \geq \frac{1}{\sqrt{2n}} \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k}$$

- Corresponding redundancy:

$$R \leq \log(\sqrt{2n}) = \frac{1}{2} \log 2n$$

(Essentially optimal!)

Jeffreys–Krichevsky–Trofimov mixture

- We now try $\Theta \sim \mathrm{Beta}(1/2, 1/2)$, i.e., $f(\theta) = 1/(\pi\sqrt{\theta(1-\theta)})$
- Then, by Stirling's approximation,

$$q_{\mathrm{JKT}}(x^n) = \int_0^1 \theta^k (1-\theta)^{n-k} f(\theta) d\theta \geq \frac{1}{\sqrt{2n}} \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k}$$

- Corresponding redundancy:

$$R \leq \log(\sqrt{2n}) = \frac{1}{2} \log 2n$$

(Essentially optimal!)

- Horizon-free sequential probability assignment:

$$q_{\mathrm{JKT}}(1|x^n) = \frac{k+1/2}{n+1}$$

Extensions

- Nonbinary alphabet: For m-ary memoryless source,

$$R^* \simeq \frac{m-1}{2} \log n$$

achieved by $\Theta \sim \text{Dirichlet}(1/2, 1/2, \ldots, 1/2)$

Extensions

- Nonbinary alphabet: For m-ary memoryless source,

$$R^* \simeq \frac{m-1}{2} \log n$$

  achieved by $\Theta \sim \text{Dirichlet}(1/2, 1/2, \ldots, 1/2)$

- Memory:
  - There exists $q(x^n)$ such that

$$\lim_{n\to\infty} \frac{1}{n} D(p(x^n) \| q(x^n)) = 0$$

    for every stationary ergodic $p(x^n)$!

Extensions

- Nonbinary alphabet: For m-ary memoryless source,

$$R^* \simeq \frac{m-1}{2} \log n$$

achieved by $\Theta \sim \mathrm{Dirichlet}(1/2, 1/2, \ldots, 1/2)$

- Memory:
  - There exists $q(x^n)$ such that

$$\lim_{n\to\infty} \frac{1}{n} D(p(x^n) \| q(x^n)) = 0$$

for every stationary ergodic $p(x^n)$!

  - Arbitrarily slow convergence

Universal compression of individual sequences

- Let $\{p_\theta : \theta \in \mathcal{T}\}$ be a collection of codes (probability assignments)

Universal compression of individual sequences

- Let $\{p_\theta \colon \theta \in \mathcal{T}\}$ be a collection of codes (probability assignments)
- Let $q(x^n)$ be our code

Universal compression of individual sequences

- Let $\{p_\theta : \theta \in \mathcal{T}\}$ be a collection of codes (probability assignments)
- Let $q(x^n)$ be our code
- Regret for a given sequence x^n:

$$R(q, x^n) = \log \frac{1}{q(x^n)} - \min_\theta \log \frac{1}{p_\theta(x^n)}$$

Universal compression of individual sequences

- Let $\{p_\theta : \theta \in \mathcal{T}\}$ be a collection of codes (probability assignments)
- Let $q(x^n)$ be our code
- Regret for a given sequence x^n:

$$R(q, x^n) = \log \frac{1}{q(x^n)} - \min_\theta \log \frac{1}{p_\theta(x^n)}$$

- Minimax redundancy:

$$\begin{aligned} R^* &= \min_{q(x^n)} \max_{x^n} R(q, x^n) \\ &= \min_{q(x^n)} \max_{\theta} \max_{x^n} \log \frac{p_\theta(x^n)}{q(x^n)} \end{aligned}$$

Universal compression of individual sequences

- Let $\{p_\theta : \theta \in \mathcal{T}\}$ be a collection of codes (probability assignments)
- Let $q(x^n)$ be our code
- Regret for a given sequence x^n:

$$R(q, x^n) = \log \frac{1}{q(x^n)} - \min_\theta \log \frac{1}{p_\theta(x^n)}$$

- Minimax redundancy:

$$\begin{aligned} R^* &= \min_{q(x^n)} \max_{x^n} R(q, x^n) \\ &= \min_{q(x^n)} \max_{\theta} \max_{x^n} \log \frac{p_\theta(x^n)}{q(x^n)} \end{aligned}$$

- Note: Minimax redundancy in the probabilistic setting

$$\begin{aligned} R^* &= \min_{q(x^n)} \max_{\theta} D(p_\theta \| q) \\ &= \min_{q(x^n)} \max_{\theta} \sum_{x^n} p_\theta(x^n) \log \frac{p_\theta(x^n)}{q(x^n)} \end{aligned}$$

Normalized maximum likelihood code

- Normalized maximum likelihood (NML) code:

$$q_{\mathrm{NML}}(x^n) \propto \max_\theta p_\theta(x^n)$$
$$= \frac{\max_\theta p_\theta(x^n)}{\sum_{y^n} \max_\theta p_\theta(y^n)}$$

Normalized maximum likelihood code

- Normalized maximum likelihood (NML) code:

$$q_{\mathrm{NML}}(x^n) \propto \max_\theta p_\theta(x^n) = \frac{\max_\theta p_\theta(x^n)}{\sum_{y^n} \max_\theta p_\theta(y^n)}$$

- Corresponding regret:

$$R(q_{\mathrm{NML}}, x^n) = \log \frac{1}{q_{\mathrm{NML}}(x^n)} - \min_\theta \log \frac{1}{p_\theta(x^n)} = \log \sum_{y^n} \max_\theta p_\theta(y^n)$$

Normalized maximum likelihood code

- Normalized maximum likelihood (NML) code:

$$q_{\text{NML}}(x^n) \propto \max_\theta p_\theta(x^n) = \frac{\max_\theta p_\theta(x^n)}{\sum_{y^n} \max_\theta p_\theta(y^n)}$$

- Corresponding regret:

$$R(q_{\text{NML}}, x^n) = \log \frac{1}{q_{\text{NML}}(x^n)} - \min_\theta \log \frac{1}{p_\theta(x^n)} = \log \sum_{y^n} \max_\theta p_\theta(y^n)$$

- Equalizer: Regret is independent of x^n!

Normalized maximum likelihood code

- Normalized maximum likelihood (NML) code:

$$q_{\mathrm{NML}}(x^n) \propto \max_\theta p_\theta(x^n) = \frac{\max_\theta p_\theta(x^n)}{\sum_{y^n} \max_\theta p_\theta(y^n)}$$

- Corresponding regret:

$$R(q_{\mathrm{NML}}, x^n) = \log \frac{1}{q_{\mathrm{NML}}(x^n)} - \min_\theta \log \frac{1}{p_\theta(x^n)} = \log \sum_{y^n} \max_\theta p_\theta(y^n)$$

- Equalizer: Regret is independent of x^n!

Theorem

$$R^* = R(q_{\mathrm{NML}}, x^n)$$

Normalized maximum likelihood code

- Normalized maximum likelihood (NML) code:

$$q_{\mathrm{NML}}(x^n) \propto \max_\theta p_\theta(x^n)$$
$$= \frac{\max_\theta p_\theta(x^n)}{\sum_{y^n} \max_\theta p_\theta(y^n)}$$

- Corresponding regret:

$$R(q_{\mathrm{NML}}, x^n) = \log \frac{1}{q_{\mathrm{NML}}(x^n)} - \min_\theta \log \frac{1}{p_\theta(x^n)} = \log \sum_{y^n} \max_\theta p_\theta(y^n)$$

- Equalizer: Regret is independent of x^n!

Theorem

$$R^* = R(q_{\mathrm{NML}}, x^n)$$

- q_{NML} is horizon-dependent and difficult to make sequential

Example: Memoryless codes

- Let $\mathcal{X} = \{0, 1\}$ and $p_\theta(x^n) = \theta^{k(x^n)}(1 - \theta)^{n-k(x^n)}$ (memoryless codes)

Example: Memoryless codes

- Let $\mathcal{X} = \{0, 1\}$ and $p_\theta(x^n) = \theta^{k(x^n)}(1-\theta)^{n-k(x^n)}$ (memoryless codes)
- Normalized maximum likelihood code:

$$q_{\mathrm{NML}}(x^n) = \frac{\max_\theta p_\theta(x^n)}{\sum_{y^n} \max_\theta p_\theta(y^n)}$$

$$= \frac{\left(\frac{k(x^n)}{n}\right)^{k(x^n)} \left(\frac{n-k(x^n)}{n}\right)^{n-k(x^n)}}{\sum_{y^n} \left(\frac{k(y^n)}{n}\right)^{k(y^n)} \left(\frac{n-k(y^n)}{n}\right)^{n-k(y^n)}}$$

Example: Memoryless codes

- Let $\mathcal{X} = \{0, 1\}$ and $p_\theta(x^n) = \theta^{k(x^n)}(1-\theta)^{n-k(x^n)}$ (memoryless codes)
- Normalized maximum likelihood code:

$$q_{\mathrm{NML}}(x^n) = \frac{\max_\theta p_\theta(x^n)}{\sum_{y^n} \max_\theta p_\theta(y^n)}$$

$$= \frac{\left(\frac{k(x^n)}{n}\right)^{k(x^n)} \left(\frac{n-k(x^n)}{n}\right)^{n-k(x^n)}}{\sum_{y^n} \left(\frac{k(y^n)}{n}\right)^{k(y^n)} \left(\frac{n-k(y^n)}{n}\right)^{n-k(y^n)}}$$

- Minimax regret:

$$R^* = \log \sum_{y^n} \left(\frac{k(y^n)}{n}\right)^{k(y^n)} \left(\frac{n-k(y^n)}{n}\right)^{n-k(y^n)}$$

$$= \log \sum_{k=0}^{n} \binom{n}{k} \left(\frac{k}{n}\right)^{k} \left(\frac{n-k}{n}\right)^{n-k}$$

$$\simeq \frac{1}{2} \log \frac{n\pi}{2}$$

Example: Memoryless codes

- As in the probabilistic setting, we now try the mixture approach

Example: Memoryless codes

- As in the probabilistic setting, we now try the mixture approach
- Recall

$$q_{\mathrm{JKT}}(x^n) \geq \frac{1}{\sqrt{2n}} \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k}$$

Example: Memoryless codes

- As in the probabilistic setting, we now try the mixture approach
- Recall

$$q_{\mathrm{JKT}}(x^n) \geq \frac{1}{\sqrt{2n}} \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k}$$

- Note the maximum likelihood $\theta = k/n$:

$$\max_{\theta} \theta^k (1-\theta)^{n-k} = \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k}$$

Example: Memoryless codes

- As in the probabilistic setting, we now try the mixture approach
- Recall

$$q_{\mathrm{JKT}}(x^n) \geq \frac{1}{\sqrt{2n}} \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k}$$

- Note the maximum likelihood $\theta = k/n$:

$$\max_\theta \theta^k (1-\theta)^{n-k} = \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k}$$

- Corresponding redundancy:

$$R \leq \frac{1}{2} \log 2n$$

(Again essentially optimal!)

Summary and extensions

- Normalized regret (for both probabilistic and deterministic settings):

$$\lim_{n\to\infty} \frac{R}{n} = \lim_{n\to\infty} \frac{\log n}{2n} = 0$$

Summary and extensions

- Normalized regret (for both probabilistic and deterministic settings):

$$\lim_{n\to\infty} \frac{R}{n} = \lim_{n\to\infty} \frac{\log n}{2n} = 0$$

- Duality between probabilistic and deterministic settings:

memoryless sources $\leftrightarrow$ constant codes

Summary and extensions

- Normalized regret (for both probabilistic and deterministic settings):

$$\lim_{n\to\infty} \frac{R}{n} = \lim_{n\to\infty} \frac{\log n}{2n} = 0$$

- Duality between probabilistic and deterministic settings:

memoryless sources $\leftrightarrow$ constant codes

- Jeffreys–Krichevsky–Trofimov mixture:
  - Essentially optimal redundancy (up to a constant)
  - Sequential (causal) probability assignment (arithmetic coding)

Summary and extensions

- Normalized regret (for both probabilistic and deterministic settings):

$$\lim_{n\to\infty} \frac{R}{n} = \lim_{n\to\infty} \frac{\log n}{2n} = 0$$

- Duality between probabilistic and deterministic settings:

memoryless sources $\leftrightarrow$ constant codes

- Jeffreys–Krichevsky–Trofimov mixture:
  - Essentially optimal redundancy (up to a constant)
  - Sequential (causal) probability assignment (arithmetic coding)

- Nonbinary alphabet: $R^* \simeq \frac{m-1}{2} \log n$

Summary and extensions

- Normalized regret (for both probabilistic and deterministic settings):

$$\lim_{n\to\infty} \frac{R}{n} = \lim_{n\to\infty} \frac{\log n}{2n} = 0$$

- Duality between probabilistic and deterministic settings:

memoryless sources $\leftrightarrow$ constant codes

- Jeffreys–Krichevsky–Trofimov mixture:
  - Essentially optimal redundancy (up to a constant)
  - Sequential (causal) probability assignment (arithmetic coding)

- Nonbinary alphabet: $R^* \simeq \frac{m-1}{2} \log n$

- There exists $q(x^n)$ such that

$$\lim_{n\to\infty} \min_{q(x^n)} \max_{x^n} \frac{1}{n} \log \frac{p(x^n)}{q(x^n)} = 0$$

for all finite-state probability assignments $p(x^n)$!

Outline

- Review of information measures
- Lossless compression and probability assignment (probabilistic / deterministic)
- Portfolio selection (deterministic)
- Sequential prediction (probabilistic)

Investment in a stock market

- Consider a stock market with m stocks

Investment in a stock market

- Consider a stock market with m stocks
- Let $\mathbf{x}^n$ denote an arbitrary sequence of price relative vectors

Investment in a stock market

- Consider a stock market with m stocks
- Let $\mathbf{x}^n$ denote an arbitrary sequence of price relative vectors
- Let $\mathbf{a}_i(\mathbf{x}^{i-1})$, $i = 1, 2, \ldots$, be a sequence of portfolios (investment strategies)

Investment in a stock market

- Consider a stock market with m stocks
- Let $\mathbf{x}^n$ denote an arbitrary sequence of price relative vectors
- Let $\mathbf{a}_i(\mathbf{x}^{i-1})$, $i = 1, 2, \ldots$, be a sequence of portfolios (investment strategies)
- Wealth at time n:

$$S_n(\mathbf{a}, \mathbf{x}^n) = \prod_{i=1}^{n} \mathbf{a}_i^T(\mathbf{x}^{i-1})\mathbf{x}_i = \prod_{i=1}^{n} \sum_{j=1}^{m} a_{ij}(\mathbf{x}^{i-1}) x_{ij}$$

Investment in a stock market

- Consider a stock market with m stocks
- Let $\mathbf{x}^n$ denote an arbitrary sequence of price relative vectors
- Let $\mathbf{a}_i(\mathbf{x}^{i-1})$, $i = 1, 2, \ldots$, be a sequence of portfolios (investment strategies)
- Wealth at time n:

$$S_n(\mathbf{a}, \mathbf{x}^n) = \prod_{i=1}^{n} \mathbf{a}_i^T(\mathbf{x}^{i-1})\mathbf{x}_i = \prod_{i=1}^{n} \sum_{j=1}^{m} a_{ij}(\mathbf{x}^{i-1}) x_{ij}$$

- Minimax regret:

$$R^* = \min_{\mathbf{b}} \max_{\mathbf{x}^n} \max_{\mathbf{a}} \log S_n(\mathbf{a}, \mathbf{x}^n) - \log S_n(\mathbf{b}, \mathbf{x}^n) = \min_{\mathbf{b}} \max_{\mathbf{a}} \max_{\mathbf{x}^n} \log \frac{S_n(\mathbf{a}, \mathbf{x}^n)}{S_n(\mathbf{b}, \mathbf{x}^n)}$$

Constant rebalanced portfolios

- Reference class of portfolios: $\mathbf{a}_i(\mathbf{x}^{i-1}) \equiv \mathbf{a}$, $i = 1, 2, \ldots$

Constant rebalanced portfolios

- Reference class of portfolios: $\mathbf{a}_i(\mathbf{x}^{i-1}) \equiv \mathbf{a}$, $i = 1, 2, \ldots$

- Minimax regret:

$$
\begin{aligned}
R^* &= \min_{\mathbf{b}} \max_{\mathbf{a}} \max_{\mathbf{x}^n} \log \frac{S_n(\mathbf{a}, \mathbf{x}^n)}{S_n(\mathbf{b}, \mathbf{x}^n)} \\
&= \log \left(\sum_{n_1+n_2+\cdots+n_m=n} \binom{n}{n_1, n_2, \ldots, n_m} \prod_{i=1}^{m} \left(\frac{n_i}{n}\right)^{n_i} \right) \\
&\simeq \frac{m-1}{2} \log \frac{n}{2\pi} + \log \frac{\Gamma(1/2)^m}{\Gamma(m/2)}
\end{aligned}
$$

(the minimax regret for binary sequence compression!)

Constant rebalanced portfolios

- Reference class of portfolios: $\mathbf{a}_i(\mathbf{x}^{i-1}) \equiv \mathbf{a}$, $i = 1, 2, \ldots$
- Minimax regret:

$$\begin{aligned} R^* &= \min_{\mathbf{b}} \max_{\mathbf{a}} \max_{\mathbf{x}^n} \log \frac{S_n(\mathbf{a}, \mathbf{x}^n)}{S_n(\mathbf{b}, \mathbf{x}^n)} \\ &= \log \left(\sum_{n_1+n_2+\cdots+n_m=n} \binom{n}{n_1, n_2, \ldots, n_m} \prod_{i=1}^{m} \left(\frac{n_i}{n}\right)^{n_i} \right) \\ &\simeq \frac{m-1}{2} \log \frac{n}{2\pi} + \log \frac{\Gamma(1/2)^m}{\Gamma(m/2)} \end{aligned}$$

(the minimax regret for binary sequence compression!)

- For $m = 2$,

$$R^* = \log \sum_{j=0}^{n} \binom{n}{j} \left(\frac{j}{n}\right)^{j} \left(\frac{n-j}{n}\right)^{n-j} \simeq \frac{1}{2} \log \frac{\pi n}{2}$$

Probability assignment and portfolio

- Each probability assignment $p(y^n)$ on $\{1, 2, \dots, m\}^n$ induces a portfolio

$$a_{iy_i}(\mathbf{x}^{i-1}) = \frac{\sum_{y^{i-1}} p(y^i)\mathbf{x}(y^{i-1})}{\sum_{y^{i-1}} p(y^{i-1})\mathbf{x}(y^{i-1})}$$

where $\mathbf{x}(y^n) = \prod_{i=1}^n x_{iy_i}$

Probability assignment and portfolio

- Each probability assignment $p(y^n)$ on $\{1, 2, \ldots, m\}^n$ induces a portfolio

$$a_{iy_i}(\mathbf{x}^{i-1}) = \frac{\sum_{y^{i-1}} p(y^i)\mathbf{x}(y^{i-1})}{\sum_{y^{i-1}} p(y^{i-1})\mathbf{x}(y^{i-1})}$$

where $\mathbf{x}(y^n) = \prod_{i=1}^n x_{iy_i}$

- By telescoping,

$$S_n(\mathbf{a}, \mathbf{x}^n) = \sum_{y^n} p(y^n)\mathbf{x}(y^n)$$

Probability assignment and portfolio

- Each probability assignment $p(y^n)$ on $\{1, 2, \ldots, m\}^n$ induces a portfolio

$$a_{iy_i}(\mathbf{x}^{i-1}) = \frac{\sum_{y^{i-1}} p(y^i)\mathbf{x}(y^{i-1})}{\sum_{y^{i-1}} p(y^{i-1})\mathbf{x}(y^{i-1})}$$

  where $\mathbf{x}(y^n) = \prod_{i=1}^n x_{iy_i}$

- By telescoping,

$$S_n(\mathbf{a}, \mathbf{x}^n) = \sum_{y^n} p(y^n)\mathbf{x}(y^n)$$

- Fund of funds: Each $p(y^n)$ invests all money in stocks y_i on day i

Probability assignment and portfolio

- Each probability assignment $p(y^n)$ on $\{1, 2, \dots, m\}^n$ induces a portfolio

$$a_{iy_i}(\mathbf{x}^{i-1}) = \frac{\sum_{y^{i-1}} p(y^i)\mathbf{x}(y^{i-1})}{\sum_{y^{i-1}} p(y^{i-1})\mathbf{x}(y^{i-1})}$$

  where $\mathbf{x}(y^n) = \prod_{i=1}^n x_{iy_i}$

- By telescoping,

$$S_n(\mathbf{a}, \mathbf{x}^n) = \sum_{y^n} p(y^n)\mathbf{x}(y^n)$$

- Fund of funds: Each $p(y^n)$ invests all money in stocks y_i on day i

- Conversely, each portfolio $\mathbf{a}$ induces a probability assignment $p(y^n)$

$$\mathbf{a} \Leftrightarrow p,$$
$$\mathbf{b} \Leftrightarrow q$$

Probability assignment and portfolio

- Each probability assignment $p(y^n)$ on $\{1, 2, \dots, m\}^n$ induces a portfolio

$$a_{iy_i}(\mathbf{x}^{i-1}) = \frac{\sum_{y^{i-1}} p(y^i)\mathbf{x}(y^{i-1})}{\sum_{y^{i-1}} p(y^{i-1})\mathbf{x}(y^{i-1})}$$

  where $\mathbf{x}(y^n) = \prod_{i=1}^n x_{iy_i}$

- By telescoping,

$$S_n(\mathbf{a}, \mathbf{x}^n) = \sum_{y^n} p(y^n)\mathbf{x}(y^n)$$

- Fund of funds: Each $p(y^n)$ invests all money in stocks y_i on day i

- Conversely, each portfolio $\mathbf{a}$ induces a probability assignment $p(y^n)$

$$\mathbf{a} \Leftrightarrow p,$$
$$\mathbf{b} \Leftrightarrow q$$

- Thus, the question boils down to choosing the right $q(y^n)$

Minimax portfolio

- Let $q_{\text{NML}}(y^n)$ be the normalized maximum likelihood pmf

Minimax portfolio

- Let $q_{\mathrm{NML}}(y^n)$ be the normalized maximum likelihood pmf
- For $m = 2$,

$$q_{\mathrm{NML}}(x^n) = \frac{\left(\frac{k(x^n)}{n}\right)^{k(x^n)} \left(\frac{n-k(x^n)}{n}\right)^{n-k(x^n)}}{\sum_{y^n} \left(\frac{k(y^n)}{n}\right)^{k(y^n)} \left(\frac{n-k(y^n)}{n}\right)^{n-k(y^n)}}$$

Minimax portfolio

- Let $q_{\mathrm{NML}}(y^n)$ be the normalized maximum likelihood pmf
- For $m = 2$,

$$q_{\mathrm{NML}}(x^n) = \frac{\left(\frac{k(x^n)}{n}\right)^{k(x^n)} \left(\frac{n-k(x^n)}{n}\right)^{n-k(x^n)}}{\sum_{y^n} \left(\frac{k(y^n)}{n}\right)^{k(y^n)} \left(\frac{n-k(y^n)}{n}\right)^{n-k(y^n)}}$$

- Then

$$\begin{aligned}
R(\mathbf{b}_{\mathrm{NML}}, \mathbf{x}^n) &= \max_{\mathbf{x}^n} \max_{\mathbf{a}} \log \frac{S_n(\mathbf{a}, \mathbf{x}^n)}{S_n(\mathbf{b}, \mathbf{x}^n)} \\
&= \max_{\mathbf{x}^n} \max_{p} \log \frac{\sum_{y^n} p(y^n)\mathbf{x}(y^n)}{\sum_{y^n} q_{\mathrm{NML}}(y^n)\mathbf{x}(y^n)} \\
&\le \max_{p} \max_{y^n} \log \frac{p(y^n)}{q_{\mathrm{NML}}(y^n)} \\
&\simeq \frac{1}{2} \log \frac{n\pi}{2}
\end{aligned}$$

Horizon-free portfolios

- Laplace mixture:

$$R(\mathbf{b}_{\mathrm{L}}, \mathbf{x}^n) = \log \binom{m+n-1}{m-1}$$

Horizon-free portfolios

- Laplace mixture:

$$R(\mathbf{b}_{\mathrm{L}}, \mathbf{x}^n) = \log \binom{m+n-1}{m-1}$$

- Jeffreys–Krichevsky–Trofimov mixture:

$$R(\mathbf{b}_{\mathrm{JKT}}, \mathbf{x}^n) = \frac{m-1}{2} \log 2n + \log \frac{\Gamma(1/2)^m}{\Gamma(m/2)} + o(1)$$

(Essentially optimal!)

Horizon-free portfolios

- Laplace mixture:

$$R(\mathbf{b}_{\mathrm{L}}, \mathbf{x}^n) = \log \binom{m+n-1}{m-1}$$

- Jeffreys–Krichevsky–Trofimov mixture:

$$R(\mathbf{b}_{\mathrm{JKT}}, \mathbf{x}^n) = \frac{m-1}{2} \log 2n + \log \frac{\Gamma(1/2)^m}{\Gamma(m/2)} + o(1)$$

  (Essentially optimal!)

- For any mixture $q(y^n) = \int \prod_{i=1}^n p(y_i)\, dF(p)$,

$$\begin{aligned} b_{iy_i} &= \frac{\sum_{y^{i-1}} q(y^i)\mathbf{x}(y^{i-1})}{\sum_{y^{i-1}} q(y^{i-1})\mathbf{x}(y^{i-1})} \\ &= \frac{\int p(y_i) S_{i-1}(p, x^{i-1})\, dF(p)}{\int S_{i-1}(p, x^{i-1})\, dF(p)} \end{aligned}$$

Horizon-free portfolios

- Laplace mixture:

$$R(\mathbf{b}_{\mathrm{L}}, \mathbf{x}^n) = \log \binom{m+n-1}{m-1}$$

- Jeffreys–Krichevsky–Trofimov mixture:

$$R(\mathbf{b}_{\mathrm{JKT}}, \mathbf{x}^n) = \frac{m-1}{2} \log 2n + \log \frac{\Gamma(1/2)^m}{\Gamma(m/2)} + o(1)$$

(Essentially optimal!)

- For any mixture $q(y^n) = \int \prod_{i=1}^n p(y_i)\, dF(p)$,

$$b_{iy_i} = \frac{\sum_{y^{i-1}} q(y^i)\mathbf{x}(y^{j-1})}{\sum_{y^{i-1}} q(y^{i-1})\mathbf{x}(y^{i-1})}$$

$$= \frac{\int p(y_i) S_{i-1}(p, x^{i-1})\, dF(p)}{\int S_{i-1}(p, x^{i-1})\, dF(p)}$$

- Fund of funds:

$$S_n(\mathbf{b}, \mathbf{x}^n) = \int S_n(p, \mathbf{x}^n)\, dF(p)$$

Outline

- Review of information measures
- Lossless compression and probability assignment (probabilistic / deterministic)
- Portfolio selection (deterministic)
- Sequential prediction (probabilistic)

Sequential prediction

- Observation: $x^n = (x_1, x_2, \ldots, x_n), x_i \in \mathcal{X}$

Sequential prediction

- Observation: $x^n = (x_1, x_2, \ldots, x_n)$, $x_i \in \mathcal{X}$
- Action: $a^n = (a_1, a_2, \ldots, a_n)$, $a_i \in \mathcal{A}$

Sequential prediction

- Observation: $x^n = (x_1, x_2, \ldots, x_n)$, $x_i \in \mathcal{X}$
- Action: $a^n = (a_1, a_2, \ldots, a_n)$, $a_i \in \mathcal{A}$
- Loss function: $l : \mathcal{X} \times \mathcal{A} \to [0, \infty)$

Sequential prediction

- Observation: $x^n = (x_1, x_2, \ldots, x_n)$, $x_i \in \mathcal{X}$
- Action: $a^n = (a_1, a_2, \ldots, a_n)$, $a_i \in \mathcal{A}$
- Loss function: $l : \mathcal{X} \times \mathcal{A} \to [0, \infty)$
- Cumulative loss: $l(x^n, a^n) = \sum_{i=1}^n l(x_i, a_i)$

Sequential prediction

- Observation: $x^n = (x_1, x_2, \ldots, x_n)$, $x_i \in \mathcal{X}$
- Action: $a^n = (a_1, a_2, \ldots, a_n)$, $a_i \in \mathcal{A}$
- Loss function: $l : \mathcal{X} \times \mathcal{A} \to [0, \infty)$
- Cumulative loss: $l(x^n, a^n) = \sum_{i=1}^n l(x_i, a_i)$
- Deterministic strategy: $a_i(x^{i-1}, a^{i-1}) = a_i(x^{i-1})$, $i \in [1 : n]$

Sequential prediction

- Observation: $x^n = (x_1, x_2, \dots, x_n)$, $x_i \in \mathcal{X}$
- Action: $a^n = (a_1, a_2, \dots, a_n)$, $a_i \in \mathcal{A}$
- Loss function: $l : \mathcal{X} \times \mathcal{A} \to [0, \infty)$
- Cumulative loss: $l(x^n, a^n) = \sum_{i=1}^n l(x_i, a_i)$
- Deterministic strategy: $a_i(x^{i-1}, a^{i-1}) = a_i(x^{i-1})$, $i \in [1:n]$
- Randomized strategy: $F(a_i | x^{i-1}, a^{i-1})$, $i \in [1:n]$

Sequential prediction

- Observation: $x^n = (x_1, x_2, \ldots, x_n)$, $x_i \in \mathcal{X}$
- Action: $a^n = (a_1, a_2, \ldots, a_n)$, $a_i \in \mathcal{A}$
- Loss function: $l : \mathcal{X} \times \mathcal{A} \to [0, \infty)$
- Cumulative loss: $l(x^n, a^n) = \sum_{i=1}^n l(x_i, a_i)$
- Deterministic strategy: $a_i(x^{i-1}, a^{i-1}) = a_i(x^{i-1})$, $i \in [1:n]$
- Randomized strategy: $F(a_i | x^{i-1}, a^{i-1})$, $i \in [1:n]$
- Examples:
  - Probability assignment (log loss): $a \in \mathscr{P}(\mathcal{X})$ and $l(x, a) = -\log a(x)$

Sequential prediction

- Observation: $x^n = (x_1, x_2, \ldots, x_n)$, $x_i \in \mathcal{X}$
- Action: $a^n = (a_1, a_2, \ldots, a_n)$, $a_i \in \mathcal{A}$
- Loss function: $l : \mathcal{X} \times \mathcal{A} \to [0, \infty)$
- Cumulative loss: $l(x^n, a^n) = \sum_{i=1}^n l(x_i, a_i)$
- Deterministic strategy: $a_i(x^{i-1}, a^{i-1}) = a_i(x^{i-1})$, $i \in [1:n]$
- Randomized strategy: $F(a_i | x^{i-1}, a^{i-1})$, $i \in [1:n]$
- Examples:
  - Probability assignment (log loss): $a \in \mathscr{P}(\mathcal{X})$ and $l(x, a) = -\log a(x)$
  - Portfolio selection: $\mathbf{a} \in \mathscr{P}(\mathcal{X})$ and $l(\mathbf{x}, \mathbf{a}) = -\log \mathbf{a}^T \mathbf{x}$

Bayes envelope

- Let $X \sim p(x)$

Bayes envelope

- Let $X \sim p(x)$
- Bayes envelope:

$$U(X) = U(p) = \inf_a \mathsf{E}[l(X, a)]$$

  - Concave in $p(x)$ (since it is the infimum over linear functions of $p(x)$)

Bayes envelope

- Let $X \sim p(x)$
- Bayes envelope:

$$U(X) = U(p) = \inf_a \mathsf{E}[l(X, a)]$$

  - Concave in $p(x)$ (since it is the infimum over linear functions of $p(x)$)
- Bayes response $a^*(X) = a^*(p)$
  - Any a that attains $U(X)$
  - Assume without loss of generality deterministic

Bayes envelope

- Let $X \sim p(x)$

- Bayes envelope:

$$U(X) = U(p) = \inf_a \mathsf{E}[l(X, a)]$$

  - Concave in $p(x)$ (since it is the infimum over linear functions of $p(x)$)

- Bayes response $a^*(X) = a^*(p)$
  - Any a that attains $U(X)$
  - Assume without loss of generality deterministic

- Example
  - For $a \in \mathscr{P}(\mathcal{X})$ and $l(x, a) = -\log a(x)$, $U(X) = H(X)$, which is attained by $a^* = p$

Bayes envelope

- Let $X \sim p(x)$
- Bayes envelope:

$$U(X) = U(p) = \inf_a \mathsf{E}[l(X, a)]$$

  - Concave in $p(x)$ (since it is the infimum over linear functions of $p(x)$)
- Bayes response $a^*(X) = a^*(p)$
  - Any a that attains $U(X)$
  - Assume without loss of generality deterministic
- Example
  - For $a \in \mathscr{P}(\mathcal{X})$ and $l(x, a) = -\log a(x)$, $U(X) = H(X)$, which is attained by $a^* = p$
- Every admissible strategy is a Bayes response to some $q(x)$

Bayes envelope

- Let $X \sim p(x)$
- Bayes envelope:

$$U(X) = U(p) = \inf_a \mathsf{E}[l(X, a)]$$

  - Concave in $p(x)$ (since it is the infimum over linear functions of $p(x)$)
- Bayes response $a^*(X) = a^*(p)$
  - Any a that attains $U(X)$
  - Assume without loss of generality deterministic
- Example
  - For $a \in \mathscr{P}(\mathcal{X})$ and $l(x, a) = -\log a(x)$, $U(X) = H(X)$, which is attained by $a^* = p$
- Every admissible strategy is a Bayes response to some $q(x)$
- Conversely, for every $q(x^n)$, $a_i(x^{i-1}) = a^*(q(x_i|x^{i-1}))$ is a valid prediction strategy

Generalized divergence

- Suppose $X \sim p(x)$

Generalized divergence

- Suppose $X \sim p(x)$
- Let $a^*(q)$ be the Bayes response to $X \sim q(x)$

Generalized divergence

- Suppose $X \sim p(x)$
- Let $a^*(q)$ be the Bayes response to $X \sim q(x)$
- Regret:

$$\mathsf{E}_p[l(X, a^*(q)) - l(X, a^*(p))] = \mathsf{E}_p[l(X, a^*(q))] - U(p) =: \Delta(p\|q)$$

Generalized divergence

- Suppose $X \sim p(x)$
- Let $a^*(q)$ be the Bayes response to $X \sim q(x)$
- Regret:

$$\mathsf{E}_p[l(X, a^*(q)) - l(X, a^*(p))] = \mathsf{E}_p[l(X, a^*(q))] - U(p) =: \Delta(p\|q)$$

- $\Delta(p\|q) \geq 0$ is referred to as the generalized divergence

Generalized divergence

- Suppose $X \sim p(x)$
- Let $a^*(q)$ be the Bayes response to $X \sim q(x)$
- Regret:

$$\mathsf{E}_p[l(X, a^*(q)) - l(X, a^*(p))] = \mathsf{E}_p[l(X, a^*(q))] - U(p) =: \Delta(p\|q)$$

- $\Delta(p\|q) \geq 0$ is referred to as the generalized divergence

Lemma

$$\Delta(p\|q) \leq l_{\max} \sqrt{2(\ln 2) D(p\|q)}$$

where $l_{\max} = \max_{a,x} l(x, a)$

Generalized divergence

- Suppose $X \sim p(x)$
- Let $a^*(q)$ be the Bayes response to $X \sim q(x)$
- Regret:

$$\mathsf{E}_p[l(X, a^*(q)) - l(X, a^*(p))] = \mathsf{E}_p[l(X, a^*(q))] - U(p) =: \Delta(p\|q)$$

- $\Delta(p\|q) \geq 0$ is referred to as the generalized divergence

Lemma

$$\Delta(p\|q) \leq l_{\max} \sqrt{2(\ln 2) D(p\|q)}$$

where $l_{\max} = \max_{a,x} l(x, a)$

- Proof idea: Pinsker's inequality $\sum |p(x) - q(x)| \leq \sqrt{2(\ln 2) D(p||q)}$

Consistency of the plug-in strategy

Theorem

If $\lim_{n\to\infty}(1/n)D(p(x^n)\|q(x^n)) = 1$ for every $p(x^n) \in \mathscr{P}$, then

$$\lim_{n\to\infty} \frac{1}{n}\Delta(p(x^n)\|q(x^n)) = 0$$

for every $p(x^n) \in \mathscr{P}$

Consistency of the plug-in strategy

Theorem

If $\lim_{n\to\infty}(1/n)D(p(x^n)\|q(x^n)) = 1$ for every $p(x^n) \in \mathscr{P}$, then

$$\lim_{n\to\infty} \frac{1}{n}\Delta(p(x^n)\|q(x^n)) = 0$$

for every $p(x^n) \in \mathscr{P}$

- Example (multiple choice exam):

Consistency of the plug-in strategy

Theorem

If $\lim_{n\to\infty}(1/n)D(p(x^n)\|q(x^n)) = 1$ for every $p(x^n) \in \mathscr{P}$, then

$$\lim_{n\to\infty} \frac{1}{n}\Delta(p(x^n)\|q(x^n)) = 0$$

for every $p(x^n) \in \mathscr{P}$

- Example (multiple choice exam):
  - For $\Theta \sim \mathrm{Dirichlet}(1/2, 1/2, 1/2, 1/2)$, $D(p(x^n)\|q(x^n)) = O((\log n)/n)$

Consistency of the plug-in strategy

Theorem

If $\lim_{n\to\infty}(1/n)D(p(x^n)\|q(x^n)) = 1$ for every $p(x^n) \in \mathscr{P}$, then

$$\lim_{n\to\infty} \frac{1}{n}\Delta(p(x^n)\|q(x^n)) = 0$$

for every $p(x^n) \in \mathscr{P}$

- Example (multiple choice exam):
  - For $\Theta \sim \text{Dirichlet}(1/2, 1/2, 1/2, 1/2)$, $D(p(x^n)\|q(x^n)) = O((\log n)/n)$
  - $R = O(\sqrt{(\log n)/n}) \to 0$ is achievable

Consistency of the plug-in strategy

Theorem

If $\lim_{n\to\infty}(1/n)D(p(x^n)\|q(x^n)) = 1$ for every $p(x^n) \in \mathscr{P}$, then

$$\lim_{n\to\infty} \frac{1}{n}\Delta(p(x^n)\|q(x^n)) = 0$$

for every $p(x^n) \in \mathscr{P}$

- Example (multiple choice exam):
  - For $\Theta \sim \text{Dirichlet}(1/2, 1/2, 1/2, 1/2)$, $D(p(x^n)\|q(x^n)) = O((\log n)/n)$
  - $R = O(\sqrt{(\log n)/n}) \to 0$ is achievable
  - In other words, the plug-in strategy can track optimal performance

Consistency of the plug-in strategy

Theorem

If $\lim_{n\to\infty}(1/n)D(p(x^n)\|q(x^n)) = 1$ for every $p(x^n) \in \mathscr{P}$, then

$$\lim_{n\to\infty} \frac{1}{n}\Delta(p(x^n)\|q(x^n)) = 0$$

for every $p(x^n) \in \mathscr{P}$

- Example (multiple choice exam):
  - For $\Theta \sim \text{Dirichlet}(1/2, 1/2, 1/2, 1/2)$, $D(p(x^n)\|q(x^n)) = O((\log n)/n)$
  - $R = O(\sqrt{(\log n)/n}) \to 0$ is achievable
  - In other words, the plug-in strategy can track optimal performance
  - Can we do better?

Concluding remarks

- Compression = probability assignment

Concluding remarks

- Compression = probability assignment
- Investment $\simeq$ probability assignment

Concluding remarks

- Compression = probability assignment
- Investment $\simeq$ probability assignment
- "Good" probability assignment $\Rightarrow$ "good" prediction algorithm

Concluding remarks

- Compression = probability assignment
- Investment $\simeq$ probability assignment
- "Good" probability assignment $\Rightarrow$ "good" prediction algorithm
- JKT/Dirichlet mixture: Best asymptotic probability assignment

Concluding remarks

- Compression = probability assignment
- Investment ≃ probability assignment
- "Good" probability assignment ⇒ "good" prediction algorithm
- JKT/Dirichlet mixture: Best asymptotic probability assignment

- We can actually do better than the plug-in strategy

Concluding remarks

- Compression = probability assignment
- Investment ≃ probability assignment
- "Good" probability assignment ⇒ "good" prediction algorithm
- JKT/Dirichlet mixture: Best asymptotic probability assignment
- We can actually do better than the plug-in strategy
- Regrets for probabilistic and deterministic settings can be different

Concluding remarks

- Compression = probability assignment
- Investment ≃ probability assignment
- "Good" probability assignment ⇒ "good" prediction algorithm
- JKT/Dirichlet mixture: Best asymptotic probability assignment

- We can actually do better than the plug-in strategy
- Regrets for probabilistic and deterministic settings can be different

- Many more tasks, approaches, and algorithms

To learn more

- Cover and Thomas (2006), Elements of Information Theory, 2nd ed, Wiley
- Lugosi and Cesa–Bianchi (1977), Prediction, Learning, and Games, Cambridge
- Merhav (1998), "Universal prediction," IEEE Trans. Inf. Theory
- Willems (2013), "Lossless source coding algorithms," IEEE Int. Symp. Inf. Theory