Minimum Rates of Approximate Sufficient Statistics

Vincent Y. F. Tan (ECE and Mathematics, NUS)

Joint work with Prof. Masahito Hayashi (Nagoya University & NUS)

Group Meeting Seminar, March 2017

Outline

1. Sufficient Statistics, Motivation, and Main Contribution

Outline

Outline

Outline

Outline

Outline

Outline

Sufficient Statistics

- Random variable $X \in \mathcal{X}$ has distribution $P_{X|\theta}$ which depends on a parameter $\theta \in \Theta$.

Sufficient Statistics

- Random variable $X \in \mathcal{X}$ has distribution $P_{X|\theta}$ which depends on a parameter $\theta \in \Theta$.
- To estimate θ, we often don't need X but some function of X, say $Y = f(X) \in \mathcal{Y}$ is sufficient.

Sufficient Statistics

- Random variable $X \in \mathcal{X}$ has distribution $P_{X|\theta}$ which depends on a parameter $\theta \in \Theta$.
- To estimate θ, we often don't need X but some function of X, say $Y = f(X) \in \mathcal{Y}$ is sufficient.
- $Y = f(X)$ is called sufficient statistics for the family $\{P_{X|\theta}\}$.

Sufficient Statistics

- Random variable $X \in \mathcal{X}$ has distribution $P_{X|\theta}$ which depends on a parameter $\theta \in \Theta$.
- To estimate θ, we often don't need X but some function of X, say $Y = f(X) \in \mathcal{Y}$ is sufficient.
- $Y = f(X)$ is called sufficient statistics for the family $\{P_{X|\theta}\}$.
- In this case, $X -\!\!\circ\!\!- Y -\!\!\circ\!\!- \theta$ forms a Markov chain.

Sufficient Statistics

- Random variable $X \in \mathcal{X}$ has distribution $P_{X|\theta}$ which depends on a parameter $\theta \in \Theta$.
- To estimate θ, we often don't need X but some function of X, say $Y = f(X) \in \mathcal{Y}$ is sufficient.
- $Y = f(X)$ is called sufficient statistics for the family $\{P_{X|\theta}\}$.
- In this case, $X \multimap Y \multimap \theta$ forms a Markov chain. Equivalently,

$$P_{X|\theta}(x) = \sum_{y \in \mathcal{Y}} P_{X|Y}(x|y) P_{Y|\theta}(y) = \sum_{y \in \mathcal{Y}} P_{X,Y|\theta}(x,y) \quad \forall (x,\theta)$$

Sufficient Statistics

- Random variable $X \in \mathcal{X}$ has distribution $P_{X|\theta}$ which depends on a parameter $\theta \in \Theta$.
- To estimate θ, we often don't need X but some function of X, say $Y = f(X) \in \mathcal{Y}$ is sufficient.
- $Y = f(X)$ is called sufficient statistics for the family $\{P_{X|\theta}\}$.
- In this case, $X -\!\!\circ\!\!- Y -\!\!\circ\!\!- \theta$ forms a Markov chain. Equivalently,

$$P_{X|\theta}(x) = \sum_{y \in \mathcal{Y}} P_{X|Y}(x|y) P_{Y|\theta}(y) = \sum_{y \in \mathcal{Y}} P_{X,Y|\theta}(x,y) \quad \forall (x,\theta)$$

- In information theory language,

$$I(\theta; X) = I(\theta; f(X)) = I(\theta; Y).$$

Y provides as much information about θ as X does.

Examples

- $X^n = (X_1, \ldots, X_n) \in \{0, 1\}^n$ is i.i.d. Bernoulli with parameter $\theta = \Pr[X_i = 1]$. Then

$$X^n -\!\!\circ\!\!- \frac{1}{n}\sum_{i=1}^{n} X_i -\!\!\circ\!\!- \theta$$

forms a Markov chain so $Y = f(X^n) = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a sufficient statistic for the family $\{P_{X^n|\theta}\}$.

Examples

- $X^n = (X_1, \ldots, X_n) \in \{0,1\}^n$ is i.i.d. Bernoulli with parameter $\theta = \Pr[X_i = 1]$. Then

$$X^n \multimap \frac{1}{n}\sum_{i=1}^n X_i \multimap \theta$$

forms a Markov chain so $Y = f(X^n) = \frac{1}{n}\sum_{i=1}^n X_i$ is a sufficient statistic for the family $\{P_{X^n|\theta}\}$.

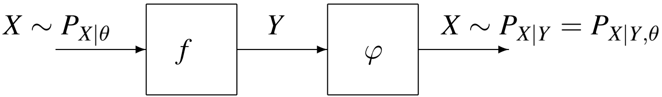
- Exponential family with natural parameter $\theta = (\theta_1, \ldots, \theta_d)$

$$P^n_{X|\theta}(x^n) = P^n_X(x^n)\exp\left[\langle Y^{(n)}(x^n), \theta\rangle - nA(\theta)\right].$$

Vector of sufficient statistics $Y^{(n)}(x^n) = (Y_1^{(n)}(x^n), \ldots, Y_d^{(n)}(x^n))$ with

$$Y_i^{(n)}(x^n) = \sum_{j=1}^n Y_i(x_j), \quad i = 1, \ldots, d.$$

Another Interpretation : Exact Reproduction of $P_{X|\theta}$

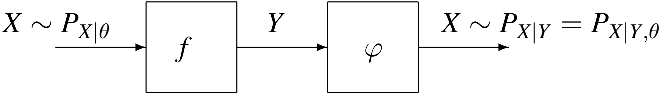


Another Interpretation : Exact Reproduction of $P_{X|\theta}$

$$X \sim P_{X|\theta} \longrightarrow \boxed{f} \xrightarrow{Y} \boxed{\varphi} \longrightarrow X \sim P_{X|Y} = P_{X|Y,\theta}$$

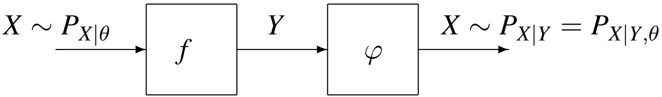
- If Y is a sufficient statistic relative to $\{P_{X|\theta}\}$, can find f and φ s.t. $P_{X|\theta}$ can be reproduced exactly using the code (f, φ).

Another Interpretation : Exact Reproduction of $P_{X|\theta}$



- If Y is a sufficient statistic relative to $\{P_{X|\theta}\}$, can find f and φ s.t. $P_{X|\theta}$ can be reproduced exactly using the code (f, φ).
- $P_{X|Y=y,\theta}$ does not depend on θ because $X \multimap Y \multimap \theta$. Can set decoder as

$$\varphi(y) = P_{X|Y=y} = P_{X|Y=y,\theta}$$

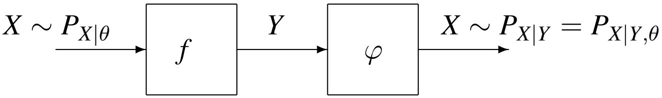


- If Y is a sufficient statistic relative to $\{P_{X|\theta}\}$, can find f and φ s.t. $P_{X|\theta}$ can be reproduced exactly using the code (f, φ).
- $P_{X|Y=y,\theta}$ does not depend on θ because $X \mathrel{-\!\!\circ\!\!-} Y \mathrel{-\!\!\circ\!\!-} \theta$. Can set decoder as

$$\varphi(y) = P_{X|Y=y} = P_{X|Y=y,\theta}$$

- Denote $(f \circ P_{X|\theta})(y) = \sum_{x \in \mathcal{X}} P_{X|\theta}(x) \Pr[f(x) = y]$.

Another Interpretation : Exact Reproduction of $P_{X|\theta}$



- If Y is a sufficient statistic relative to $\{P_{X|\theta}\}$, can find f and φ s.t. $P_{X|\theta}$ can be reproduced exactly using the code (f, φ).
- $P_{X|Y=y,\theta}$ does not depend on θ because $X -\!\circ\!- Y -\!\circ\!- \theta$. Can set decoder as

$$\varphi(y) = P_{X|Y=y} = P_{X|Y=y,\theta}$$

- Denote $(f \circ P_{X|\theta})(y) = \sum_{x \in \mathcal{X}} P_{X|\theta}(x) \Pr[f(x) = y]$. Hence,

$$\begin{aligned} \varphi \circ f \circ P_{X|\theta} &= \sum_{y \in \mathcal{Y}} (f \circ P_{X|\theta})(y) \varphi(y) \\ &= \sum_{y \in \mathcal{Y}} P_{X|\theta}\{x \in \mathcal{X} : f(x) = y\} P_{X|Y=y,\theta} = P_{X|\theta}. \end{aligned}$$

Memory Size

- How much memory to store the sufficient statistics?

Memory Size

- How much memory to store the sufficient statistics?
- Example 1: Binomial case. Since $\mathcal{X} = \{0, 1\}$, the sufficient statistic

$$\frac{1}{n}\sum_{j=1}^{n} X_j \in \left\{\frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n}{n}\right\}$$

can take on $n + 1 \sim n^1$ values.

Memory Size

- How much memory to store the sufficient statistics?
- Example 1: Binomial case. Since $\mathcal{X} = \{0, 1\}$, the sufficient statistic

$$\frac{1}{n}\sum_{j=1}^{n} X_j \in \left\{\frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\right\}$$

  can take on $n + 1 \sim n^1$ values.

- Example 2: k-nomial case, i.e., $\mathcal{X} = \{0, 1, \ldots, k-1\}$ and we have n samples. Size of sufficient statistics $Y^{(n)}(x^n)$ satisfies

$$\left|\{Y^{(n)}(x^n) : x^n \in \mathcal{X}^n\}\right| = \binom{n+k-1}{k-1} \sim \frac{n^{k-1}}{(k-1)!},$$

  so the size of the memory is $\asymp n^{k-1}$.

Memory Size

- How much memory to store the sufficient statistics?
- Example 1: Binomial case. Since $\mathcal{X} = \{0, 1\}$, the sufficient statistic

$$\frac{1}{n}\sum_{j=1}^{n} X_j \in \left\{\frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\right\}$$

  can take on $n + 1 \sim n^1$ values.

- Example 2: k-nomial case, i.e., $\mathcal{X} = \{0, 1, \ldots, k - 1\}$ and we have n samples. Size of sufficient statistics $Y^{(n)}(x^n)$ satisfies

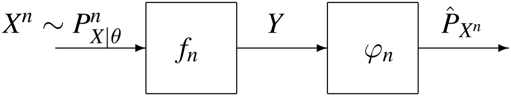
$$\left|\{Y^{(n)}(x^n) : x^n \in \mathcal{X}^n\}\right| = \binom{n+k-1}{k-1} \sim \frac{n^{k-1}}{(k-1)!},$$

  so the size of the memory is $\asymp n^{k-1}$.

- Example 3: $\theta \in \Theta = [0, 1]$ is the unknown mean of a Gaussian. Sufficient statistics can take uncountable number of values.

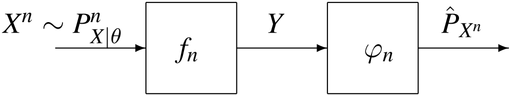
Our Contribution

- Reduce the exponent d in n^d by relaxing exact recovery condition on generating distribution $P^n_{X|\theta}$.



Our Contribution

- Reduce the exponent d in n^d by relaxing exact recovery condition on generating distribution $P^n_{X|\theta}$.



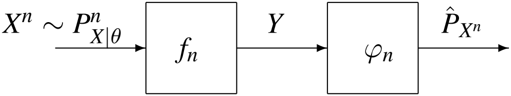
- Now instead of exact recovery $P^n_{X|\theta} = \varphi_n \circ f_n \circ P^n_{X|\theta}$ for every $n \in \mathbb{N}$, we only require that

$$\overline{\lim_{n\to\infty}} \int_\Theta F\left(P^n_{X|\theta}, \varphi_n \circ f_n \circ P^n_{X|\theta}\right) \mu(\mathrm{d}\theta) \leq \delta.$$

for some $\delta \geq 0$.

Our Contribution

- Reduce the exponent d in n^d by relaxing exact recovery condition on generating distribution $P^n_{X|\theta}$.



- Now instead of exact recovery $P^n_{X|\theta} = \varphi_n \circ f_n \circ P^n_{X|\theta}$ for every $n \in \mathbb{N}$, we only require that

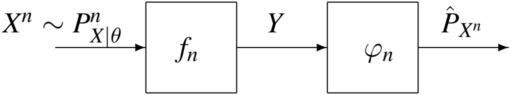
$$\varlimsup_{n\to\infty} \int_{\Theta} F\left(P^n_{X|\theta}, \varphi_n \circ f_n \circ P^n_{X|\theta}\right) \mu(\mathrm{d}\theta) \le \delta.$$

  for some $\delta \ge 0$.

- Most of the time, we can reduce the exponent to $d/2$ and this is optimal.

Outline

Definition of Code



Definition (Code)

A size-M_n code $\mathcal{C}_n = (f_n, \varphi_n)$ consists of

- A possibly stochastic encoder $f_n : \mathcal{X}^n \to \mathcal{Y}_n = \{1, \ldots, M_n\}$;
- A decoder $\varphi_n : \mathcal{Y}_n \to \mathcal{P}(\mathcal{X}^n)$ (set of distributions on $\mathcal{X}^n$)

Definition of Error

Definition (Average Error)

The average error is a code $\mathcal{C}_n = (f_n, \varphi_n)$ is defined as

$$\begin{aligned}\varepsilon(\mathcal{C}_n) &:= \int_\Theta F\left(\varphi_n \circ f_n \circ P^n_{X|\theta}, P^n_{X|\theta}\right)\ \mu(\mathrm{d}\theta) \\ &= \mathbb{E}_{\theta\sim\mu}\left[F\left(\varphi_n \circ f_n \circ P^n_{X|\theta}, P^n_{X|\theta}\right)\right]\end{aligned}$$

where $\mu(\cdot)$ is the prior distribution of θ.

Definition of Error

Definition (Average Error)

The average error is a code $\mathcal{C}_n = (f_n, \varphi_n)$ is defined as

$$\varepsilon(\mathcal{C}_n) := \int_\Theta F\left(\varphi_n \circ f_n \circ P^n_{X|\theta}, P^n_{X|\theta}\right)\ \mu(\mathrm{d}\theta)$$
$$= \mathbb{E}_{\theta\sim\mu}\left[F\left(\varphi_n \circ f_n \circ P^n_{X|\theta}, P^n_{X|\theta}\right)\right]$$

where $\mu(\cdot)$ is the prior distribution of θ.

Recall that

$$(f_n \circ P^n_{X|\theta})(y) = \sum_{x^n\in\mathcal{X}^n} P^n_{X|\theta}(x^n)\,\Pr[f(x^n) = y]$$

and

$$\varphi_n \circ f_n \circ P^n_{X|\theta} = \sum_{y\in\mathcal{Y}_n} (f_n \circ P^n_{X|\theta})(y)\varphi_n(y) \in \mathcal{P}(\mathcal{X}^n).$$

Measuring Errors Between Distributions

- Consider two commonly-used error criteria.

Measuring Errors Between Distributions

- Consider two commonly-used error criteria.
- Variational distance

$$F(P,Q) = \|P-Q\|_1 = 2 \sup_{A \subset \mathcal{X}} |P(A) - Q(A)| \in [0,2]$$

Measuring Errors Between Distributions

- Consider two commonly-used error criteria.
- Variational distance

$$F(P,Q) = \|P - Q\|_1 = 2 \sup_{A \subset \mathcal{X}} |P(A) - Q(A)| \in [0,2]$$

- Relative entropy (Kullback-Leibler distance)

$$D(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \in [0,\infty]$$

Measuring Errors Between Distributions

- Consider two commonly-used error criteria.
- Variational distance

$$F(P,Q) = \|P-Q\|_1 = 2 \sup_{A \subset \mathcal{X}} |P(A) - Q(A)| \in [0,2]$$

- Relative entropy (Kullback-Leibler distance)

$$D(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \in [0,\infty]$$

- Pinsker's inequality

$$\frac{\log e}{2} \|P-Q\|_1^2 \le D(P\|Q)$$

Minimum Compression Rate

- Given a code $\mathcal{C}_n$, denote its error under the variational distance and relative entropy as $\varepsilon^{(1)}(\mathcal{C}_n)$ and $\varepsilon^{(2)}(\mathcal{C}_n)$ resp.

Minimum Compression Rate

- Given a code $\mathcal{C}_n$, denote its error under the variational distance and relative entropy as $\varepsilon^{(1)}(\mathcal{C}_n)$ and $\varepsilon^{(2)}(\mathcal{C}_n)$ resp.
- Denote its size as $|\mathcal{C}_n|$.

Minimum Compression Rate

- Given a code $\mathcal{C}_n$, denote its error under the variational distance and relative entropy as $\varepsilon^{(1)}(\mathcal{C}_n)$ and $\varepsilon^{(2)}(\mathcal{C}_n)$ resp.
- Denote its size as $|\mathcal{C}_n|$.
- Find smallest exponent r in $|\mathcal{C}_n| \asymp n^r$ subject to a bounded error.

Minimum Compression Rate

- Given a code $\mathcal{C}_n$, denote its error under the variational distance and relative entropy as $\varepsilon^{(1)}(\mathcal{C}_n)$ and $\varepsilon^{(2)}(\mathcal{C}_n)$ resp.
- Denote its size as $|\mathcal{C}_n|$.
- Find smallest exponent r in $|\mathcal{C}_n| \asymp n^r$ subject to a bounded error.

Definition (Minimum Compression Rate)

Let $\delta \geq 0$. Define

$$\mathrm{R}^{(i)}(\delta) := \inf_{\{\mathcal{C}_n\}_{n\in\mathbb{N}}} \left\{ \varlimsup_{n\to\infty} \frac{\log|\mathcal{C}_n|}{\log n} \;:\; \varlimsup_{n\to\infty} \varepsilon^{(i)}(\mathcal{C}_n) \leq \delta \right\}, \quad i = 1, 2.$$

Minimum Compression Rate

- Given a code $\mathcal{C}_n$, denote its error under the variational distance and relative entropy as $\varepsilon^{(1)}(\mathcal{C}_n)$ and $\varepsilon^{(2)}(\mathcal{C}_n)$ resp.
- Denote its size as $|\mathcal{C}_n|$.
- Find smallest exponent r in $|\mathcal{C}_n| \asymp n^r$ subject to a bounded error.

Definition (Minimum Compression Rate)

Let $\delta \geq 0$. Define

$$\mathrm{R}^{(i)}(\delta) := \inf_{\{\mathcal{C}_n\}_{n\in\mathbb{N}}} \left\{ \varlimsup_{n\to\infty} \frac{\log |\mathcal{C}_n|}{\log n} \,:\, \varlimsup_{n\to\infty} \varepsilon^{(i)}(\mathcal{C}_n) \leq \delta \right\}, \quad i = 1, 2.$$

$$\lim_{n\to\infty} \frac{\log |\mathcal{C}_n|}{\log n} = r \quad \Longleftrightarrow \quad |\mathcal{C}_n| \asymp n^r$$

Minimum Compression Rate: Interpretation

- Suppose $\mathrm{R}^{(i)}(\delta) = r$. Then for every $\epsilon > 0$, there exists $\{\mathcal{C}_n\}_{n\in\mathbb{N}}$ whose asymptotic error under criterion $i = 1, 2$ is $\leq \delta$ and

$$|\mathcal{C}_n| \leq n^{r+\epsilon}.$$

Minimum Compression Rate: Interpretation

- Suppose $\mathrm{R}^{(i)}(\delta) = r$. Then for every $\epsilon > 0$, there exists $\{\mathcal{C}_n\}_{n\in\mathbb{N}}$ whose asymptotic error under criterion $i = 1, 2$ is $\leq \delta$ and

$$|\mathcal{C}_n| \leq n^{r+\epsilon}.$$

- Furthermore, no sequence of codes exists whose asymptotic error under criterion $i = 1, 2$ is $\leq \delta$ and whose size

$$|\mathcal{C}_n| \leq n^{r-\epsilon}.$$

Minimum Compression Rate: Interpretation

- Suppose $\mathrm{R}^{(i)}(\delta) = r$. Then for every $\epsilon > 0$, there exists $\{\mathcal{C}_n\}_{n\in\mathbb{N}}$ whose asymptotic error under criterion $i = 1, 2$ is $\leq \delta$ and

$$|\mathcal{C}_n| \leq n^{r+\epsilon}.$$

- Furthermore, no sequence of codes exists whose asymptotic error under criterion $i = 1, 2$ is $\leq \delta$ and whose size

$$|\mathcal{C}_n| \leq n^{r-\epsilon}.$$

- Note that $\delta \in [0, 2]$ for variational distance and $\delta \in [0, \infty]$ for relative entropy.

Minimum Compression Rate: Properties

- Because $\delta \mapsto \mathrm{R}^{(i)}(\delta)$ is monotone

$$\mathrm{R}^{(i)}(\delta') \le \mathrm{R}^{(i)}(\delta), \quad \forall\, 0 \le \delta \le \delta'.$$

Minimum Compression Rate: Properties

- Because $\delta \mapsto \mathrm{R}^{(i)}(\delta)$ is monotone

$$\mathrm{R}^{(i)}(\delta') \le \mathrm{R}^{(i)}(\delta), \quad \forall\, 0 \le \delta \le \delta'.$$

- Due to Pinsker's inequality $\frac{\log e}{2}\|P - Q\|_1^2 \le D(P\|Q)$,

$$\mathrm{R}^{(1)}(0) \le \mathrm{R}^{(2)}(0).$$

Minimum Compression Rate: Properties

- Because $\delta \mapsto \mathrm{R}^{(i)}(\delta)$ is monotone

$$\mathrm{R}^{(i)}(\delta') \le \mathrm{R}^{(i)}(\delta), \quad \forall\, 0 \le \delta \le \delta'.$$

- Due to Pinsker's inequality $\frac{\log e}{2}\|P - Q\|_1^2 \le D(P\|Q)$,

$$\mathrm{R}^{(1)}(0) \le \mathrm{R}^{(2)}(0).$$

- Our goal is to characterize $\mathrm{R}^{(i)}(\delta)$ for all values of δ for statistical models $\{P_{X|\theta}\}$ under reasonable assumptions.

Minimum Compression Rate: Properties

- Because $\delta \mapsto \mathrm{R}^{(i)}(\delta)$ is monotone

$$\mathrm{R}^{(i)}(\delta') \le \mathrm{R}^{(i)}(\delta), \quad \forall\, 0 \le \delta \le \delta'.$$

- Due to Pinsker's inequality $\frac{\log e}{2}\|P - Q\|_1^2 \le D(P\|Q)$,

$$\mathrm{R}^{(1)}(0) \le \mathrm{R}^{(2)}(0).$$

- Our goal is to characterize $\mathrm{R}^{(i)}(\delta)$ for all values of δ for statistical models $\{P_{X|\theta}\}$ under reasonable assumptions.

- Typically for $\Theta \subset \mathbb{R}^d$,

$$\mathrm{R}^{(i)}(\delta) = \frac{d}{2}.$$

Outline

Assumptions

(i) Parameter space $\Theta \subset \mathbb{R}^d$ is bounded and has positive Lebesgue measure (in $\mathbb{R}^d$).

Assumptions

(i) Parameter space $\Theta \subset \mathbb{R}^d$ is bounded and has positive Lebesgue measure (in $\mathbb{R}^d$).

(ii) Local approximation of relative entropy: As $\theta' \to \theta$,

$$D\big(P_{X|\theta} \,\big\|\, P_{X|\theta'}\big) = \frac{1}{2}(\theta - \theta')^T J(\theta - \theta') + o(\|\theta - \theta'\|^2)$$

where J is the Fisher information matrix.

Assumptions

(i) Parameter space $\Theta \subset \mathbb{R}^d$ is bounded and has positive Lebesgue measure (in $\mathbb{R}^d$).

(ii) Local approximation of relative entropy: As $\theta' \to \theta$,

$$D\big(P_{X|\theta} \,\|\, P_{X|\theta'}\big) = \frac{1}{2}(\theta - \theta')^T J(\theta - \theta') + o(\|\theta - \theta'\|^2)$$

where J is the Fisher information matrix.

(iii) Asymptotic efficiency: Exists a sequence of estimators $\hat{\theta}_n(X^n)$ s.t.

$$\mathbb{E}_{\theta \sim \mu}\left[D\big(P_{X|\hat{\theta}_n(X^n)} \,\|\, P_{X|\theta}\big)\right] = \frac{d}{2n} + o\left(\frac{1}{n}\right).$$

Assumptions

(i) Parameter space $\Theta \subset \mathbb{R}^d$ is bounded and has positive Lebesgue measure (in $\mathbb{R}^d$).

(ii) Local approximation of relative entropy: As $\theta' \to \theta$,

$$D\big(P_{X|\theta} \,\big\|\, P_{X|\theta'}\big) = \frac{1}{2}(\theta - \theta')^T J(\theta - \theta') + o(\|\theta - \theta'\|^2)$$

where J is the Fisher information matrix.

(iii) Asymptotic efficiency: Exists a sequence of estimators $\hat{\theta}_n(X^n)$ s.t.

$$\mathbb{E}_{\theta\sim\mu}\left[D\big(P_{X|\hat{\theta}_n(X^n)} \,\big\|\, P_{X|\theta}\big)\right] = \frac{d}{2n} + o\left(\frac{1}{n}\right).$$

(iv) Local asymptotic normality of MLE

Assumptions

(i) Parameter space $\Theta \subset \mathbb{R}^d$ is bounded and has positive Lebesgue measure (in $\mathbb{R}^d$).

(ii) Local approximation of relative entropy: As $\theta' \to \theta$,

$$D\big(P_{X|\theta} \,\|\, P_{X|\theta'}\big) = \frac{1}{2}(\theta - \theta')^T J(\theta - \theta') + o(\|\theta - \theta'\|^2)$$

where J is the Fisher information matrix.

(iii) Asymptotic efficiency: Exists a sequence of estimators $\hat{\theta}_n(X^n)$ s.t.

$$\mathbb{E}_{\theta \sim \mu}\left[D\big(P_{X|\hat{\theta}_n(X^n)} \,\|\, P_{X|\theta}\big)\right] = \frac{d}{2n} + o\left(\frac{1}{n}\right).$$

(iv) Local asymptotic normality of MLE

(v) Local asymptotic sufficiency of MLE

Main Result

Theorem (Hayashi-T. (2016))

1. *Assume (i), (ii), (iv), and (v), under the variational distance criterion*

$$\mathrm{R}^{(1)}(\delta) = \frac{d}{2} \qquad \forall\, \delta \in [0,2).$$

Main Result

Theorem (Hayashi-T. (2016))

1. *Assume (i), (ii), (iv), and (v), under the variational distance criterion*
$$\mathrm{R}^{(1)}(\delta) = \frac{d}{2} \qquad \forall\, \delta \in [0, 2).$$
2. *Assume (i), (ii), and (iii), under the relative entropy criterion*
$$\mathrm{R}^{(2)}(\delta) = \frac{d}{2} \qquad \forall\, \delta \in \left[\frac{d}{2}, \infty\right).$$

Main Result

Theorem (Hayashi-T. (2016))

1. *Assume (i), (ii), (iv), and (v), under the variational distance criterion*
$$\mathrm{R}^{(1)}(\delta) = \frac{d}{2} \qquad \forall\, \delta \in [0,2).$$

2. *Assume (i), (ii), and (iii), under the relative entropy criterion*
$$\mathrm{R}^{(2)}(\delta) = \frac{d}{2} \qquad \forall\, \delta \in \left[\frac{d}{2}, \infty\right).$$

3. *If in addition* $\{P_{X|\theta}\}_{\theta\in\Theta}$ *is an exponential family,*
$$\mathrm{R}^{(2)}(\delta) = \frac{d}{2} \qquad \forall\, \delta \in [0,\infty).$$

Main Result : Remarks

- We construct codes $\mathcal{C}_n$ that achieve zero asymptotic error and have memory size $|\mathcal{C}_n| \asymp n^{d/2}$.

Main Result : Remarks

- We construct codes $\mathcal{C}_n$ that achieve zero asymptotic error and have memory size $|\mathcal{C}_n| \asymp n^{d/2}$.
- Compare to exact sufficient statistics in which $|\mathcal{C}_n| \asymp n^d$.

Main Result : Remarks

- We construct codes $\mathcal{C}_n$ that achieve zero asymptotic error and have memory size $|\mathcal{C}_n| \asymp n^{d/2}$.
- Compare to exact sufficient statistics in which $|\mathcal{C}_n| \asymp n^d$.
- But (this is more cool!), we show that even if the error is non-vanishing, i.e.,

$$\varlimsup_{n\to\infty} \varepsilon^{(1)}(\mathcal{C}_n) \le \delta, \quad \text{for any } \delta \in [0,2),$$
$$\varlimsup_{n\to\infty} \varepsilon^{(2)}(\mathcal{C}_n) \le \delta, \quad \text{for any } \delta \in [0,\infty),$$

the memory requirement $d/2$ is asymptotically the same.

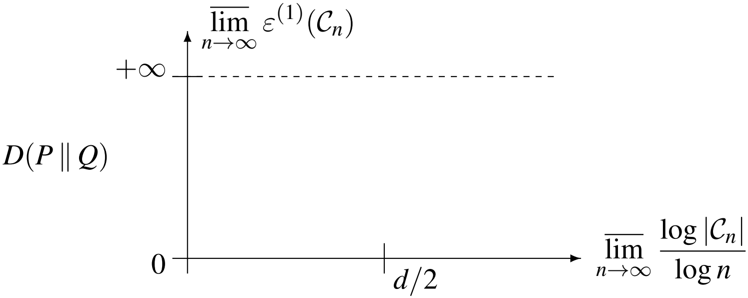
Main Result : Remarks

- We construct codes $\mathcal{C}_n$ that achieve zero asymptotic error and have memory size $|\mathcal{C}_n| \asymp n^{d/2}$.
- Compare to exact sufficient statistics in which $|\mathcal{C}_n| \asymp n^d$.
- But (this is more cool!), we show that even if the error is non-vanishing, i.e.,

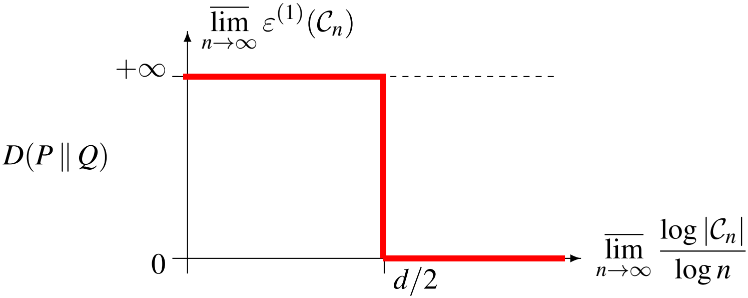
$$\varlimsup_{n\to\infty} \varepsilon^{(1)}(\mathcal{C}_n) \le \delta, \quad \text{for any } \delta \in [0,2),$$
$$\varlimsup_{n\to\infty} \varepsilon^{(2)}(\mathcal{C}_n) \le \delta, \quad \text{for any } \delta \in [0,\infty),$$

  the memory requirement $d/2$ is asymptotically the same.
- This is known in information theory as a strong converse.

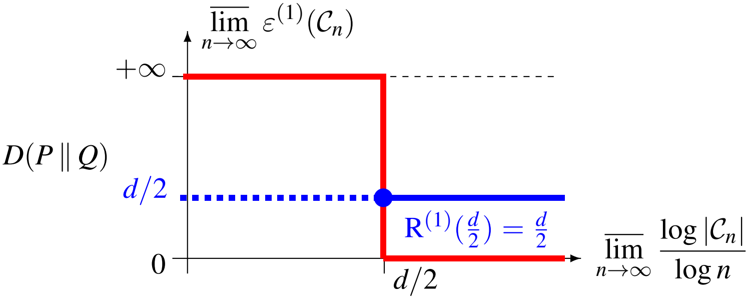
Main Result : Strong Converse



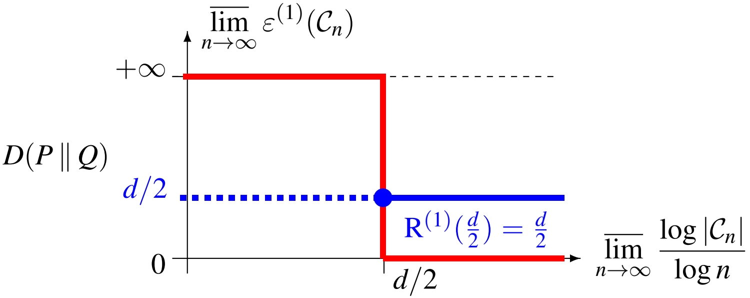
Main Result : Strong Converse

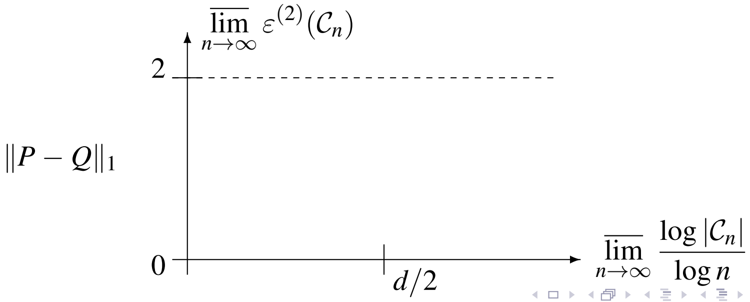


Main Result : Strong Converse

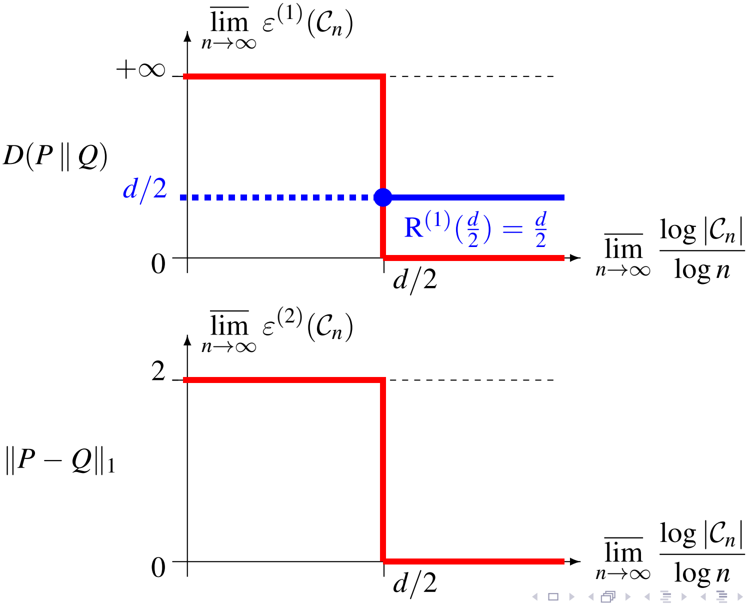


Main Result : Strong Converse





Main Result : Strong Converse



Outline

Weak Achievability for Relative Entropy: $\mathrm{R}^{(2)}(\frac{d}{2}) \leq \frac{d}{2}$

Weak Achievability for Relative Entropy: $\mathrm{R}^{(2)}(\frac{d}{2}) \leq \frac{d}{2}$

Universal Coding, Information, Prediction, and Estimation

JORMA RISSANEN

Abstract—A connection between universal codes and the problems of prediction and statistical estimation is established. A known lower bound for the mean length of universal codes is sharpened and generalized, and optimum universal codes constructed. The bound is defined to give the information in strings relative to the considered class of processes. The earlier derived minimum description length criterion for estimation of parameters, including their number, is given a fundamental information theoretic justification by showing that its estimators achieve the information in the strings. It is also shown that one cannot do prediction in Gaussian autoregressive moving average (ARMA) processes below a bound, which is determined by the information in the data.

Manuscript received July 13, 1983; revised January 16, 1984. This work was presented in part at the IEEE International Symposium on Information Theory, St. Jovite, Canada, September 26–30, 1983.

This work was done while the author was Visiting Professor at the Department of System Science, University of California, Los Angeles, while on leave from the IBM Research Laboratory, San Jose, CA 95193.

I. Introduction

THERE are three main problems in signal processing: prediction, data compression, and estimation. In the first, we are given a string of observed data points x_t, $t = 1, \cdots, n$, one after another, and the objective is to predict for each t the next outcome x_{t+1} from what we have seen so far. In the data compression problem we are given a similar sequence of observations, each truncated to some finite precision, and the objective is to redescribe the data with a suitably designed code as efficiently as possible, i.e., with a short code length.

J. Rissanen

Weak Achievability for Relative Entropy: $\mathrm{R}^{(2)}(\frac{d}{2}) \le \frac{d}{2}$

Universal Coding, Information, Prediction, and Estimation

JORMA RISSANEN

Abstract—A connection between universal codes and the problems of prediction and statistical estimation is established. A known lower bound for the mean length of universal codes is sharpened and generalized, and optimum universal codes constructed. The bound is defined to give the information in strings relative to the considered class of processes. The earlier derived minimum description length criterion for estimation of parameters, including their number, is given a fundamental information theoretic justification by showing that its estimators achieve the information in the strings. It is also shown that one cannot do prediction in Gaussian autoregressive moving average (ARMA) processes below a bound, which is determined by the information in the data.

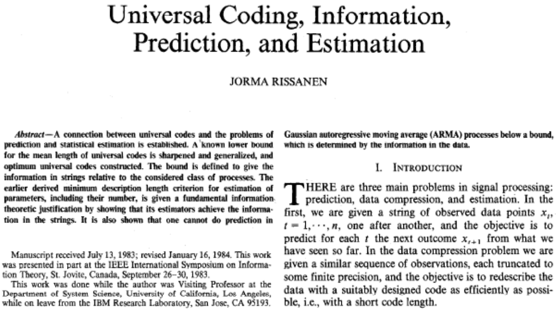
Manuscript received July 13, 1983; revised January 16, 1984. This work was presented in part at the IEEE International Symposium on Information Theory, St. Jovite, Canada, September 26–30, 1983.
This work was done while the author was Visiting Professor at the Department of System Science, University of California, Los Angeles, while on leave from the IBM Research Laboratory, San Jose, CA 95193.

I. Introduction

THERE are three main problems in signal processing: prediction, data compression, and estimation. In the first, we are given a string of observed data points x_t, $t = 1, \cdots, n$, one after another, and the objective is to predict for each t the next outcome x_{t+1} from what we have seen so far. In the data compression problem we are given a similar sequence of observations, each truncated to some finite precision, and the objective is to redescribe the data with a suitably designed code as efficiently as possible, i.e., with a short code length.

J. Rissanen

- Inventor of the minimum description length (MDL) principle for model selection (among many other things).



Universal Coding, Information, Prediction, and Estimation

JORMA RISSANEN

Abstract—A connection between universal codes and the problems of prediction and statistical estimation is established. A known lower bound for the mean length of universal codes is sharpened and generalized, and optimum universal codes constructed. The bound is defined to give the information in strings relative to the considered class of processes. The earlier derived minimum description length criterion for estimation of parameters, including their number, is given a fundamental information theoretic justification by showing that its estimators achieve the information in the strings. It is also shown that one cannot do prediction in Gaussian autoregressive moving average (ARMA) processes below a bound, which is determined by the information in the data.

Manuscript received July 13, 1983; revised January 16, 1984. This work was presented in part at the IEEE International Symposium on Information Theory, St. Jovite, Canada, September 26–30, 1983.
This work was done while the author was Visiting Professor at the Department of System Science, University of California, Los Angeles, while on leave from the IBM Research Laboratory, San Jose, CA 95193.

I. INTRODUCTION

THERE are three main problems in signal processing: prediction, data compression, and estimation. In the first, we are given a string of observed data points x_t, $t = 1, \cdots, n$, one after another, and the objective is to predict for each t the next outcome x_{t+1} from what we have seen so far. In the data compression problem we are given a similar sequence of observations, each truncated to some finite precision, and the objective is to redescribe the data with a suitably designed code as efficiently as possible, i.e., with a short code length.

J. Rissanen

- Inventor of the minimum description length (MDL) principle for model selection (among many other things).

- Quantize the MLE similarly to Rissanen.

Weak Achievability for Relative Entropy: $\mathrm{R}^{(2)}(\frac{d}{2}) \leq \frac{d}{2}$

- Compute MLE $\hat{\theta}_n$ from data X^n.

Weak Achievability for Relative Entropy: $\mathrm{R}^{(2)}(\frac{d}{2}) \leq \frac{d}{2}$

- Compute MLE $\hat{\theta}_n$ from data X^n.
- Encoder: Apply discretization to $\hat{\theta}_n$ with span $t/\sqrt{n}$ and store this discretized parameter $\hat{\theta}_n' \in \Theta_{n,t}$ in the memory $\Theta_{n,t}$.

Weak Achievability for Relative Entropy: $\mathrm{R}^{(2)}(\frac{d}{2}) \le \frac{d}{2}$

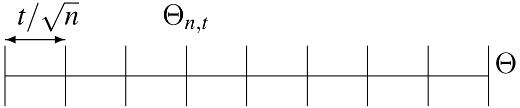
- Compute MLE $\hat{\theta}_n$ from data X^n.
- Encoder: Apply discretization to $\hat{\theta}_n$ with span $t/\sqrt{n}$ and store this discretized parameter $\hat{\theta}'_n \in \Theta_{n,t}$ in the memory $\Theta_{n,t}$.

Θ

Weak Achievability for Relative Entropy: $\mathrm{R}^{(2)}(\frac{d}{2}) \leq \frac{d}{2}$

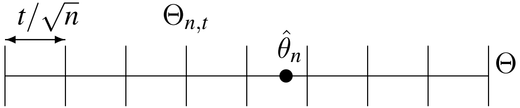
- Compute MLE $\hat{\theta}_n$ from data X^n.

- Encoder: Apply discretization to $\hat{\theta}_n$ with span $t/\sqrt{n}$ and store this discretized parameter $\hat{\theta}_n' \in \Theta_{n,t}$ in the memory $\Theta_{n,t}$.



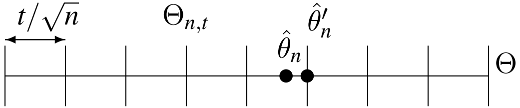
Weak Achievability for Relative Entropy: $\mathrm{R}^{(2)}(\frac{d}{2}) \leq \frac{d}{2}$

- Compute MLE $\hat{\theta}_n$ from data X^n.
- Encoder: Apply discretization to $\hat{\theta}_n$ with span $t/\sqrt{n}$ and store this discretized parameter $\hat{\theta}'_n \in \Theta_{n,t}$ in the memory $\Theta_{n,t}$.



Weak Achievability for Relative Entropy: $\mathrm{R}^{(2)}(\frac{d}{2}) \leq \frac{d}{2}$

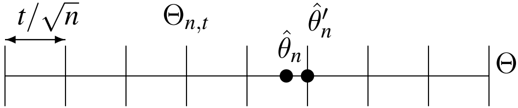
- Compute MLE $\hat{\theta}_n$ from data X^n.
- Encoder: Apply discretization to $\hat{\theta}_n$ with span $t/\sqrt{n}$ and store this discretized parameter $\hat{\theta}'_n \in \Theta_{n,t}$ in the memory $\Theta_{n,t}$.



Weak Achievability for Relative Entropy: $\mathrm{R}^{(2)}(\frac{d}{2}) \le \frac{d}{2}$

- Compute MLE $\hat{\theta}_n$ from data X^n.

- Encoder: Apply discretization to $\hat{\theta}_n$ with span $t/\sqrt{n}$ and store this discretized parameter $\hat{\theta}_n' \in \Theta_{n,t}$ in the memory $\Theta_{n,t}$.

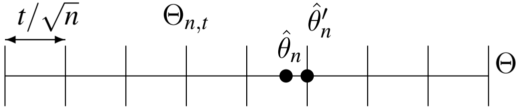


- Memory is $\Theta_{n,t} = \Theta \cap \frac{t}{\sqrt{n}}\mathbb{Z}^d$ and $|\Theta_{n,t}| \asymp n^{d/2}$.

Weak Achievability for Relative Entropy: $\mathrm{R}^{(2)}(\frac{d}{2}) \leq \frac{d}{2}$

- Compute MLE $\hat{\theta}_n$ from data X^n.

- Encoder: Apply discretization to $\hat{\theta}_n$ with span $t/\sqrt{n}$ and store this discretized parameter $\hat{\theta}'_n \in \Theta_{n,t}$ in the memory $\Theta_{n,t}$.



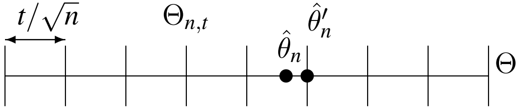
- Memory is $\Theta_{n,t} = \Theta \cap \frac{t}{\sqrt{n}}\mathbb{Z}^d$ and $|\Theta_{n,t}| \asymp n^{d/2}$.

- Decoder is the deterministic map from $\hat{\theta}'_n$ to distribution $P^n_{X|\hat{\theta}'_n}$.

Weak Achievability for Relative Entropy: $\mathrm{R}^{(2)}(\frac{d}{2}) \leq \frac{d}{2}$

- Compute MLE $\hat{\theta}_n$ from data X^n.

- Encoder: Apply discretization to $\hat{\theta}_n$ with span $t/\sqrt{n}$ and store this discretized parameter $\hat{\theta}'_n \in \Theta_{n,t}$ in the memory $\Theta_{n,t}$.



- Memory is $\Theta_{n,t} = \Theta \cap \frac{t}{\sqrt{n}}\mathbb{Z}^d$ and $|\Theta_{n,t}| \asymp n^{d/2}$.

- Decoder is the deterministic map from $\hat{\theta}'_n$ to distribution $P^n_{X|\hat{\theta}'_n}$.

- Can show that

$$\overline{\lim_{n\to\infty}}\, \varepsilon^{(2)}(\mathcal{C}_n) \leq \frac{d}{2}$$

by eventually taking $t \downarrow 0$. But error is non-vanishing. :(

Exponential Families yield Stronger Result: $\mathrm{R}^{(2)}(0) \leq \frac{d}{2}$

Exponential Families yield Stronger Result: $\mathrm{R}^{(2)}(0) \leq \frac{d}{2}$

- Assume that $\{P_{X|\theta}\}$ is an exponential family

$$P_{X|\theta}(x) = P_X(x) \exp\left[\langle \theta, Y(x) \rangle - A(\theta)\right].$$

Exponential Families yield Stronger Result: $\mathrm{R}^{(2)}(0) \leq \frac{d}{2}$

- Assume that $\{P_{X|\theta}\}$ is an exponential family

$$P_{X|\theta}(x) = P_X(x) \exp\left[\langle \theta, Y(x) \rangle - A(\theta)\right].$$

- Moment parametrization:

$$\eta(\theta) = \nabla_\theta A(\theta) = \mathbb{E}_\theta[Y(X)].$$

Set of feasible moment parameters $\mathcal{H} := \{\eta(\theta) : \theta \in \Theta\}$.

Exponential Families yield Stronger Result: $\mathrm{R}^{(2)}(0) \leq \frac{d}{2}$

- Assume that $\{P_{X|\theta}\}$ is an exponential family

$$P_{X|\theta}(x) = P_X(x) \exp\left[\langle \theta, Y(x) \rangle - A(\theta)\right].$$

- Moment parametrization:

$$\eta(\theta) = \nabla_\theta A(\theta) = \mathbb{E}_\theta[Y(X)].$$

Set of feasible moment parameters $\mathcal{H} := \{\eta(\theta) : \theta \in \Theta\}$.

- $\hat{\eta}_n = \frac{1}{n} \sum_{j=1}^n Y(X_j)$ is a sufficient statistic for $X^n \sim P_{X|\theta}^n$.

Exponential Families yield Stronger Result: $\mathrm{R}^{(2)}(0) \leq \frac{d}{2}$

- Assume that $\{P_{X|\theta}\}$ is an exponential family

$$P_{X|\theta}(x) = P_X(x) \exp\left[\langle \theta, Y(x)\rangle - A(\theta)\right].$$

- Moment parametrization:

$$\eta(\theta) = \nabla_\theta A(\theta) = \mathbb{E}_\theta[Y(X)].$$

  Set of feasible moment parameters $\mathcal{H} := \{\eta(\theta) : \theta \in \Theta\}$.

- $\hat{\eta}_n = \frac{1}{n}\sum_{j=1}^n Y(X_j)$ is a sufficient statistic for $X^n \sim P^n_{X|\theta}$.

- Encoder: Apply discretization to $\hat{\eta}$ with span $t/\sqrt{n}$, i.e.,

$$\hat{\eta}'_n = \beta_t(\hat{\eta}_n) = \underset{\eta' \in \mathcal{H}_{n,t}}{\arg\min} \|\eta' - \hat{\eta}_n\|_2, \quad \text{where} \quad \mathcal{H}_{n,t} = \mathcal{H} \cap \frac{t}{\sqrt{n}}\mathbb{Z}^d$$

Exponential Families yield Stronger Result: $\mathrm{R}^{(2)}(0) \leq \frac{d}{2}$

- Assume that $\{P_{X|\theta}\}$ is an exponential family

$$P_{X|\theta}(x) = P_X(x) \exp\left[\langle \theta, Y(x) \rangle - A(\theta)\right].$$

- Moment parametrization:

$$\eta(\theta) = \nabla_\theta A(\theta) = \mathbb{E}_\theta[Y(X)].$$

Set of feasible moment parameters $\mathcal{H} := \{\eta(\theta) : \theta \in \Theta\}$.

- $\hat{\eta}_n = \frac{1}{n}\sum_{j=1}^n Y(X_j)$ is a sufficient statistic for $X^n \sim P_{X|\theta}^n$.
- Encoder: Apply discretization to $\hat{\eta}$ with span $t/\sqrt{n}$, i.e.,

$$\hat{\eta}_n' = \beta_t(\hat{\eta}_n) = \arg\min_{\eta' \in \mathcal{H}_{n,t}} \|\eta' - \hat{\eta}_n\|_2, \quad \text{where} \quad \mathcal{H}_{n,t} = \mathcal{H} \cap \frac{t}{\sqrt{n}}\mathbb{Z}^d$$

$\mathcal{H}$

Exponential Families yield Stronger Result: $\mathrm{R}^{(2)}(0) \leq \frac{d}{2}$

- Assume that $\{P_{X|\theta}\}$ is an exponential family

$$P_{X|\theta}(x) = P_X(x) \exp\left[\langle \theta, Y(x) \rangle - A(\theta)\right].$$

- Moment parametrization:

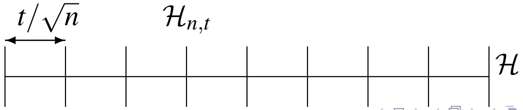
$$\eta(\theta) = \nabla_\theta A(\theta) = \mathbb{E}_\theta[Y(X)].$$

  Set of feasible moment parameters $\mathcal{H} := \{\eta(\theta) : \theta \in \Theta\}$.

- $\hat{\eta}_n = \frac{1}{n}\sum_{j=1}^n Y(X_j)$ is a sufficient statistic for $X^n \sim P^n_{X|\theta}$.

- Encoder: Apply discretization to $\hat{\eta}$ with span $t/\sqrt{n}$, i.e.,

$$\hat{\eta}'_n = \beta_t(\hat{\eta}_n) = \operatorname*{arg\,min}_{\eta' \in \mathcal{H}_{n,t}} \|\eta' - \hat{\eta}_n\|_2, \quad \text{where} \quad \mathcal{H}_{n,t} = \mathcal{H} \cap \frac{t}{\sqrt{n}}\mathbb{Z}^d$$



Exponential Families yield Stronger Result: $\mathrm{R}^{(2)}(0) \leq \frac{d}{2}$

- Assume that $\{P_{X|\theta}\}$ is an exponential family

$$P_{X|\theta}(x) = P_X(x) \exp\left[\langle \theta, Y(x) \rangle - A(\theta)\right].$$

- Moment parametrization:

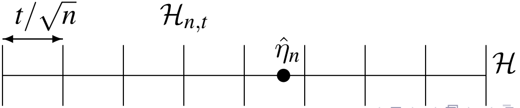
$$\eta(\theta) = \nabla_\theta A(\theta) = \mathbb{E}_\theta[Y(X)].$$

  Set of feasible moment parameters $\mathcal{H} := \{\eta(\theta) : \theta \in \Theta\}$.

- $\hat{\eta}_n = \frac{1}{n}\sum_{j=1}^n Y(X_j)$ is a sufficient statistic for $X^n \sim P_{X|\theta}^n$.

- Encoder: Apply discretization to $\hat{\eta}$ with span $t/\sqrt{n}$, i.e.,

$$\hat{\eta}_n' = \beta_t(\hat{\eta}_n) = \operatorname*{arg\,min}_{\eta' \in \mathcal{H}_{n,t}} \|\eta' - \hat{\eta}_n\|_2, \quad \text{where} \quad \mathcal{H}_{n,t} = \mathcal{H} \cap \frac{t}{\sqrt{n}}\mathbb{Z}^d$$



Exponential Families yield Stronger Result: $\mathrm{R}^{(2)}(0) \leq \frac{d}{2}$

- Assume that $\{P_{X|\theta}\}$ is an exponential family

$$P_{X|\theta}(x) = P_X(x) \exp\left[\langle \theta, Y(x) \rangle - A(\theta)\right].$$

- Moment parametrization:

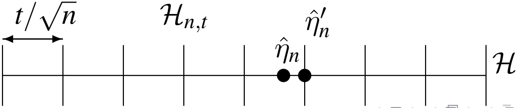
$$\eta(\theta) = \nabla_\theta A(\theta) = \mathbb{E}_\theta[Y(X)].$$

  Set of feasible moment parameters $\mathcal{H} := \{\eta(\theta) : \theta \in \Theta\}$.

- $\hat{\eta}_n = \frac{1}{n}\sum_{j=1}^n Y(X_j)$ is a sufficient statistic for $X^n \sim P_{X|\theta}^n$.

- Encoder: Apply discretization to $\hat{\eta}$ with span $t/\sqrt{n}$, i.e.,

$$\hat{\eta}_n' = \beta_t(\hat{\eta}_n) = \operatorname*{arg\,min}_{\eta' \in \mathcal{H}_{n,t}} \|\eta' - \hat{\eta}_n\|_2, \quad \text{where} \quad \mathcal{H}_{n,t} = \mathcal{H} \cap \frac{t}{\sqrt{n}}\mathbb{Z}^d$$



Exponential Families yield Stronger Result: $\mathrm{R}^{(2)}(0) \leq \frac{d}{2}$

- Decoder: Uniform mixture of conditional distributions whose moment parameter is discretized to $\hat{\eta}_n'$:

$$\varphi(\hat{\eta}_n') = \frac{1}{|\beta_t^{-1}(\hat{\eta}_n')|} \sum_{\eta \in \beta_t^{-1}(\hat{\eta}_n')} P_{X^n|Y=n\eta}$$

where

$$\beta_t^{-1}(\hat{\eta}_n') := \left\{ \hat{\eta}_n : \beta_t(\hat{\eta}_n) = \hat{\eta}_n' \right\}.$$

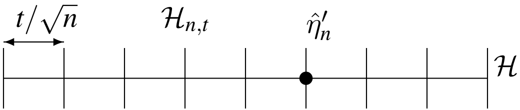
Exponential Families yield Stronger Result: $\mathrm{R}^{(2)}(0) \le \frac{d}{2}$

- Decoder: Uniform mixture of conditional distributions whose moment parameter is discretized to $\hat{\eta}_n'$:

$$\varphi(\hat{\eta}_n') = \frac{1}{|\beta_t^{-1}(\hat{\eta}_n')|} \sum_{\eta \in \beta_t^{-1}(\hat{\eta}_n')} P_{X^n|Y=n\eta}$$

where

$$\beta_t^{-1}(\hat{\eta}_n') := \left\{ \hat{\eta}_n : \beta_t(\hat{\eta}_n) = \hat{\eta}_n' \right\}.$$



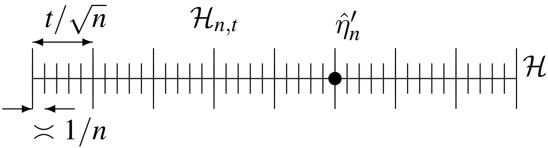
Exponential Families yield Stronger Result: $\mathrm{R}^{(2)}(0) \le \frac{d}{2}$

- Decoder: Uniform mixture of conditional distributions whose moment parameter is discretized to $\hat{\eta}'_n$:

$$\varphi(\hat{\eta}'_n) = \frac{1}{|\beta_t^{-1}(\hat{\eta}'_n)|} \sum_{\eta \in \beta_t^{-1}(\hat{\eta}'_n)} P_{X^n|Y=n\eta}$$

where

$$\beta_t^{-1}(\hat{\eta}'_n) := \left\{ \hat{\eta}_n : \beta_t(\hat{\eta}_n) = \hat{\eta}'_n \right\}.$$

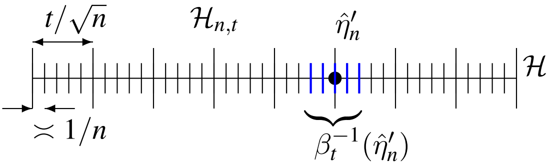


- Decoder: Uniform mixture of conditional distributions whose moment parameter is discretized to $\hat{\eta}_n'$:

$$\varphi(\hat{\eta}_n') = \frac{1}{|\beta_t^{-1}(\hat{\eta}_n')|} \sum_{\eta \in \beta_t^{-1}(\hat{\eta}_n')} P_{X^n|Y=n\eta}$$

where

$$\beta_t^{-1}(\hat{\eta}_n') := \left\{ \hat{\eta}_n : \beta_t(\hat{\eta}_n) = \hat{\eta}_n' \right\}.$$



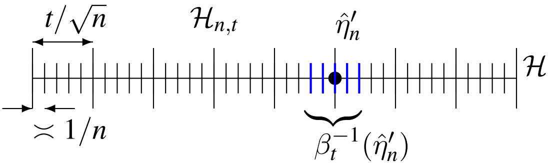
Exponential Families yield Stronger Result: $\mathrm{R}^{(2)}(0) \le \frac{d}{2}$

- Decoder: Uniform mixture of conditional distributions whose moment parameter is discretized to $\hat{\eta}'_n$:

$$\varphi(\hat{\eta}'_n) = \frac{1}{|\beta_t^{-1}(\hat{\eta}'_n)|} \sum_{\eta \in \beta_t^{-1}(\hat{\eta}'_n)} P_{X^n|Y=n\eta}$$

where

$$\beta_t^{-1}(\hat{\eta}'_n) := \left\{ \hat{\eta}_n : \beta_t(\hat{\eta}_n) = \hat{\eta}'_n \right\}.$$



- Asymptotic error under relative entropy is zero and $|\mathcal{H}_{n,t}| \asymp n^{d/2}$.

Achievability for Variational Distance: $\mathrm{R}^{(1)}(0) \leq \frac{d}{2}$

Achievability for Variational Distance: $\mathrm{R}^{(1)}(0) \leq \frac{d}{2}$

- Need to assume local asymptotic normality and local asymptotic sufficiency of MLE.

Achievability for Variational Distance: $\mathrm{R}^{(1)}(0) \leq \frac{d}{2}$

- Need to assume local asymptotic normality and local asymptotic sufficiency of MLE.

- Discretize MLE with span $t/\sqrt{n}$.

Achievability for Variational Distance: $\mathrm{R}^{(1)}(0) \leq \frac{d}{2}$

- Need to assume local asymptotic normality and local asymptotic sufficiency of MLE.
- Discretize MLE with span $t/\sqrt{n}$.
- Variational distance is a norm $\Rightarrow$ triangle inequality

Achievability for Variational Distance: $\mathrm{R}^{(1)}(0) \le \frac{d}{2}$

- Need to assume local asymptotic normality and local asymptotic sufficiency of MLE.
- Discretize MLE with span $t/\sqrt{n}$.
- Variational distance is a norm $\Rightarrow$ triangle inequality
- Uniform mixture idea.

Outline

Weak Converse Variational Distance: $\mathrm{R}^{(1)}(0) \geq \frac{d}{2}$

Weak Converse Variational Distance: $\mathrm{R}^{(1)}(0) \geq \frac{d}{2}$

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 36, NO. 3, MAY 1990 453

Information-Theoretic Asymptotics of Bayes Methods

BERTRAND S. CLARKE AND ANDREW R. BARRON, MEMBER, IEEE

***Abstract*—In the absence of knowledge of the true density function, Bayesian models take the joint density function for a sequence of n random variables to be an average of densities with respect to a prior. We examine the relative entropy distance D_n between the true density and the Bayesian density and show that the asymptotic distance is $(d/2)(\log n)+c$, where d is the dimension of the parameter vector. Therefore, the relative entropy rate D_n/n converges to zero at rate $(\log n)/n$. The constant c, which we explicitly identify, depends only on the prior density function and the Fisher information matrix evaluated at the true parameter value. Consequences are given for density estimation, universal data compression, composite hypothesis testing, and stock-market portfolio selection.**

I. Introduction

THE RELATIVE entropy is a mathematical expres-

we identify. We note that if the mixture excludes a neighborhood of the true density, then the behavior of the relative entropy is, asymptotically, of the order of the sample size; in addition, if the prior is discrete and assigns positive mass at θ_o, the relative entropy then asymptotically tends to a constant.

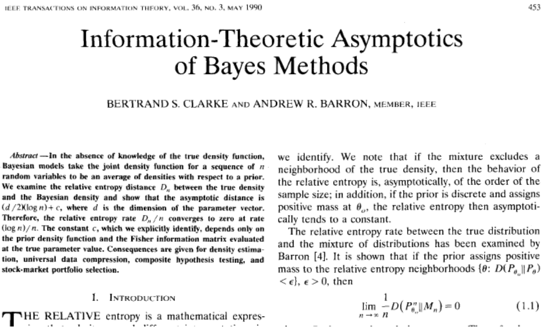
The relative entropy rate between the true distribution and the mixture of distributions has been examined by Barron [4]. It is shown that if the prior assigns positive mass to the relative entropy neighborhoods $\{\theta: D(P_{\theta_o}\|P_\theta) < \epsilon\}$, $\epsilon > 0$, then

$$\lim_{n\to\infty} \frac{1}{n} D\left(P^n_{\theta_o}\|M_n\right) = 0 \qquad (1.1)$$

B. Clarke

A. Barron

Weak Converse Variational Distance: $\mathrm{R}^{(1)}(0) \geq \frac{d}{2}$



IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 36, NO. 3, MAY 1990 453

Information-Theoretic Asymptotics of Bayes Methods

BERTRAND S. CLARKE AND ANDREW R. BARRON, MEMBER, IEEE

Abstract —In the absence of knowledge of the true density function, Bayesian models take the joint density function for a sequence of n random variables to be an average of densities with respect to a prior. We examine the relative entropy distance D_n between the true density and the Bayesian density and show that the asymptotic distance is $(d/2)(\log n) + c$, where d is the dimension of the parameter vector. Therefore, the relative entropy rate D_n/n converges to zero at rate $(\log n)/n$. The constant c, which we explicitly identify, depends only on the prior density function and the Fisher information matrix evaluated at the true parameter value. Consequences are given for density estimation, universal data compression, composite hypothesis testing, and stock-market portfolio selection.

I. INTRODUCTION

THE RELATIVE entropy is a mathematical expres-

we identify. We note that if the mixture excludes a neighborhood of the true density, then the behavior of the relative entropy is, asymptotically, of the order of the sample size; in addition, if the prior is discrete and assigns positive mass at θ_o, the relative entropy then asymptotically tends to a constant.

The relative entropy rate between the true distribution and the mixture of distributions has been examined by Barron [4]. It is shown that if the prior assigns positive mass to the relative entropy neighborhoods $\{\theta: D(P_{\theta_o} \| P_\theta) < \epsilon\}$, $\epsilon > 0$, then

$$\lim_{n \to \infty} \frac{1}{n} D(P_{\theta_o}^n \| M_n) = 0 \qquad (1.1)$$

B. Clarke A. Barron

$$D\left(P^n_{X|\theta} \,\middle\|\, \underbrace{\int_\Theta P^n_{X|\theta'}\,\mu(\mathrm{d}\theta')}_{\text{mixture}}\right) = ??$$

Weak Converse Variational Distance: $\mathrm{R}^{(1)}(0) \geq \frac{d}{2}$

- Can obtain a weak converse $\mathrm{R}^{(1)}(0) \geq \frac{d}{2}$ by using Clarke and Barron's asymptotic formula:

$$D\left(P^n_{X|\theta} \,\middle\|\, \int_\Theta P^n_{X|\theta'}\,\mu(\mathrm{d}\theta')\right) = \frac{d}{2}\log n + O(1).$$

Weak Converse Variational Distance: $\mathrm{R}^{(1)}(0) \geq \frac{d}{2}$

- Can obtain a weak converse $\mathrm{R}^{(1)}(0) \geq \frac{d}{2}$ by using Clarke and Barron's asymptotic formula:

$$D\left(P^n_{X|\theta} \,\middle\|\, \int_\Theta P^n_{X|\theta'}\, \mu(\mathrm{d}\theta')\right) = \frac{d}{2}\log n + O(1).$$

- Additionally use the fact that

$$\varepsilon^{(1)}(\mathcal{C}_n) \to 0$$

and the uniform continuity of mutual information, i.e.,

$$|I_P(A;B) - I_{P'}(A;B)| \leq 3\nu \log(|\mathcal{A}||\mathcal{B}| - 1) + 3H(\nu)$$

where

$$\nu = \frac{1}{2}\|P - P'\|_1.$$

Strong Converse Variational Distance : $\mathrm{R}^{(1)}(2^-) \geq \frac{d}{2}$

Strong Converse Variational Distance : $\mathrm{R}^{(1)}(2^-) \geq \frac{d}{2}$

- We want to show that for any sequence of codes $\{\mathcal{C}_n\}_{n\in\mathbb{N}}$ such that

$$\overline{\lim_{n\to\infty}}\, \varepsilon^{(1)}(\mathcal{C}_n) < 2$$

the memory size cannot be smaller than $n^{d(\frac{1}{2}-\gamma)}$ for any $\gamma > 0$.

Strong Converse Variational Distance : $\mathrm{R}^{(1)}(2^-) \geq \frac{d}{2}$

- We want to show that for any sequence of codes $\{\mathcal{C}_n\}_{n\in\mathbb{N}}$ such that

$$\varlimsup_{n\to\infty} \varepsilon^{(1)}(\mathcal{C}_n) < 2$$

the memory size cannot be smaller than $n^{d(\frac{1}{2}-\gamma)}$ for any $\gamma > 0$.

- Assume, to the contrary, that there exists a code $\mathcal{C}_n$ with error

$$\mathbb{E}_{\theta\sim\mu}\left[\left\|P^n_{X|\theta} - (\varphi\circ f)(\theta)\right\|_1\right] \leq 2-\alpha,$$

with memory size $M_n = O(n^{\frac{1}{2}-\gamma})$ for some $\gamma > 0$.

Strong Converse Variational Distance : $\mathrm{R}^{(1)}(2^-) \geq \frac{d}{2}$

- We want to show that for any sequence of codes $\{\mathcal{C}_n\}_{n\in\mathbb{N}}$ such that

$$\varlimsup_{n\to\infty} \varepsilon^{(1)}(\mathcal{C}_n) < 2$$

the memory size cannot be smaller than $n^{d(\frac{1}{2}-\gamma)}$ for any $\gamma > 0$.

- Assume, to the contrary, that there exists a code $\mathcal{C}_n$ with error

$$\mathbb{E}_{\theta\sim\mu}\left[\left\|P^n_{X|\theta} - (\varphi\circ f)(\theta)\right\|_1\right] \leq 2-\alpha,$$

with memory size $M_n = O(n^{\frac{1}{2}-\gamma})$ for some $\gamma > 0$.

- Define $\mathcal{S} = \left\{\theta\in\Theta : \|P^n_{X|\theta} - (\varphi\circ f)(\theta)\|_1 \leq 2-\frac{\alpha}{2}\right\}$. Markov inequality says

$$\mu(\mathcal{S}) \geq \frac{\alpha}{4-\alpha} > 0.$$

Strong Converse Variational Distance : $\mathrm{R}^{(1)}(2^-) \geq \frac{d}{2}$

- Assume $\lambda \ll \mu$. Then $\lambda(\mathcal{S}) > 0$.

Strong Converse Variational Distance : $\mathrm{R}^{(1)}(2^-) \geq \frac{d}{2}$

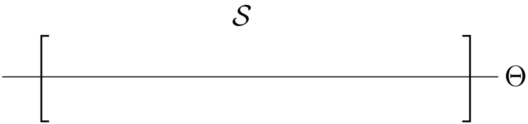
- Assume $\lambda \ll \mu$. Then $\lambda(\mathcal{S}) > 0$.
- Can choose $\frac{5}{\alpha}M_n$ points $\{\theta_i : i = 1, \ldots, \frac{5}{\alpha}M_n\} \subset \mathcal{S}$ such that

$$\|P^n_{X|\theta_i} - (\varphi \circ f)(\theta_i)\|_1 \leq 2 - \frac{\alpha}{2}, \quad |\theta_i - \theta_j| > \lambda(\mathcal{S}) \left(\frac{5}{\alpha}M_n\right)^{-1}$$

Θ

- Assume $\lambda \ll \mu$. Then $\lambda(\mathcal{S}) > 0$.
- Can choose $\frac{5}{\alpha}M_n$ points $\{\theta_i : i = 1, \ldots, \frac{5}{\alpha}M_n\} \subset \mathcal{S}$ such that

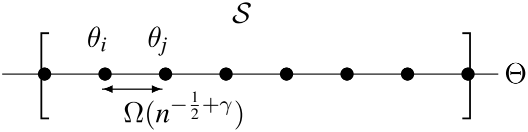
$$\|P^n_{X|\theta_i} - (\varphi \circ f)(\theta_i)\|_1 \leq 2 - \frac{\alpha}{2}, \quad |\theta_i - \theta_j| > \lambda(\mathcal{S}) \left(\frac{5}{\alpha}M_n\right)^{-1}$$



Strong Converse Variational Distance : $\mathrm{R}^{(1)}(2^-) \geq \frac{d}{2}$

- Assume $\lambda \ll \mu$. Then $\lambda(\mathcal{S}) > 0$.
- Can choose $\frac{5}{\alpha}M_n$ points $\{\theta_i : i = 1, \ldots, \frac{5}{\alpha}M_n\} \subset \mathcal{S}$ such that

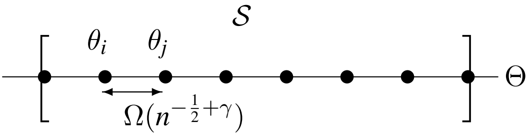
$$\|P^n_{X|\theta_i} - (\varphi \circ f)(\theta_i)\|_1 \leq 2 - \frac{\alpha}{2}, \quad |\theta_i - \theta_j| > \lambda(\mathcal{S}) \left(\frac{5}{\alpha}M_n\right)^{-1}$$



Strong Converse Variational Distance : $\mathrm{R}^{(1)}(2^-) \geq \frac{d}{2}$

- Assume $\lambda \ll \mu$. Then $\lambda(\mathcal{S}) > 0$.
- Can choose $\frac{5}{\alpha}M_n$ points $\{\theta_i : i = 1, \ldots, \frac{5}{\alpha}M_n\} \subset \mathcal{S}$ such that

$$\|P^n_{X|\theta_i} - (\varphi \circ f)(\theta_i)\|_1 \leq 2 - \frac{\alpha}{2}, \quad |\theta_i - \theta_j| > \lambda(\mathcal{S}) \left(\frac{5}{\alpha}M_n\right)^{-1}$$



- Because separation is $\Omega(n^{-\frac{1}{2}+\gamma})$, there exists disjoint $\mathcal{D}_i \subset \mathcal{X}^n$, $i = 1, \ldots, \frac{5}{\alpha}M_n$ such that

$$P^n_{X|\theta_i}(\mathcal{D}_i) \geq 1 - \epsilon.$$

Strong Converse Variational Distance : $\mathrm{R}^{(1)}(2^{-}) \geq \frac{d}{2}$

- Note that $\frac{1}{2}\|P - Q\|_1 = \sup_A |P(A) - Q(A)|$.

Strong Converse Variational Distance : $\mathrm{R}^{(1)}(2^-) \geq \frac{d}{2}$

- Note that $\frac{1}{2}\|P - Q\|_1 = \sup_A |P(A) - Q(A)|$.
- Take $P = (\varphi \circ f(\theta_i))$ and $Q = P^n_{X|\theta_i}$.

Strong Converse Variational Distance : $\mathrm{R}^{(1)}(2^-) \geq \frac{d}{2}$

- Note that $\frac{1}{2}\|P - Q\|_1 = \sup_A |P(A) - Q(A)|$.
- Take $P = (\varphi \circ f(\theta_i))$ and $Q = P^n_{X|\theta_i}$.
- This implies

$$1 - \frac{\alpha}{4} \geq (\varphi \circ f(\theta_i))(\mathcal{D}_i^c) - P^n_{X|\theta_i}(\mathcal{D}_i^c)$$

Strong Converse Variational Distance : $\mathrm{R}^{(1)}(2^-) \geq \frac{d}{2}$

- Note that $\frac{1}{2}\|P - Q\|_1 = \sup_A |P(A) - Q(A)|$.
- Take $P = (\varphi \circ f(\theta_i))$ and $Q = P^n_{X|\theta_i}$.
- This implies

$$1 - \frac{\alpha}{4} \geq (\varphi \circ f(\theta_i))(\mathcal{D}_i^c) - P^n_{X|\theta_i}(\mathcal{D}_i^c) \geq (\varphi \circ f(\theta_i))(\mathcal{D}_i^c) - \epsilon$$

Strong Converse Variational Distance : $\mathrm{R}^{(1)}(2^-) \geq \frac{d}{2}$

- Note that $\frac{1}{2}\|P - Q\|_1 = \sup_A |P(A) - Q(A)|$.
- Take $P = (\varphi \circ f(\theta_i))$ and $Q = P^n_{X|\theta_i}$.
- This implies

$$1 - \frac{\alpha}{4} \geq (\varphi \circ f(\theta_i))(\mathcal{D}_i^c) - P^n_{X|\theta_i}(\mathcal{D}_i^c) \geq (\varphi \circ f(\theta_i))(\mathcal{D}_i^c) - \epsilon$$

- We have

$$(\varphi \circ f(\theta_i))(\mathcal{D}_i) \geq \frac{\alpha}{4} - \epsilon, \quad \forall\, i = 1, \ldots, \frac{5}{\alpha} M_n.$$

Strong Converse Variational Distance : $\mathrm{R}^{(1)}(2^-) \geq \frac{d}{2}$

$$M_n \geq \sum_{j=1}^{M_n} (\varphi(j)) \left(\bigcup_{i=1}^{\frac{5}{\alpha} M_n} \mathcal{D}_i \right)$$

[$\varphi(j)$ is a prob. meas.]

Strong Converse Variational Distance : $\mathrm{R}^{(1)}(2^-) \geq \frac{d}{2}$

$$M_n \geq \sum_{j=1}^{M_n} (\varphi(j)) \left(\bigcup_{i=1}^{\frac{5}{\alpha} M_n} \mathcal{D}_i \right) \qquad [\varphi(j) \text{ is a prob. meas.}]$$

$$= \sum_{i=1}^{\frac{5}{\alpha} M_n} \left(\sum_{j=1}^{M_n} (\varphi(j))(\mathcal{D}_i) \right) \qquad [\mathcal{D}_i \text{ are disjoint}]$$

Strong Converse Variational Distance : $\mathrm{R}^{(1)}(2^-) \geq \frac{d}{2}$

$$M_n \geq \sum_{j=1}^{M_n} (\varphi(j)) \left(\bigcup_{i=1}^{\frac{5}{\alpha}M_n} \mathcal{D}_i \right)$$

$[\varphi(j)$ is a prob. meas.$]$

$$= \sum_{i=1}^{\frac{5}{\alpha}M_n} \left(\sum_{j=1}^{M_n} (\varphi(j))(\mathcal{D}_i) \right)$$

$[\mathcal{D}_i$ are disjoint$]$

$$\geq \sum_{i=1}^{\frac{5}{\alpha}M_n} (\varphi \circ f(\theta_i))(\mathcal{D}_i)$$

$[\varphi \circ f$ is a cvx. comb. of $\varphi(j)]$

Strong Converse Variational Distance : $\mathrm{R}^{(1)}(2^-) \geq \frac{d}{2}$

$$M_n \geq \sum_{j=1}^{M_n} (\varphi(j)) \left(\bigcup_{i=1}^{\frac{5}{\alpha}M_n} \mathcal{D}_i \right)$$ [$\varphi(j)$ is a prob. meas.]

$$= \sum_{i=1}^{\frac{5}{\alpha}M_n} \left(\sum_{j=1}^{M_n} (\varphi(j))(\mathcal{D}_i) \right)$$ [$\mathcal{D}_i$ are disjoint]

$$\geq \sum_{i=1}^{\frac{5}{\alpha}M_n} (\varphi \circ f(\theta_i))(\mathcal{D}_i)$$ [$\varphi \circ f$ is a cvx. comb. of $\varphi(j)$]

$$\geq \sum_{i=1}^{\frac{5}{\alpha}M_n} \left(\frac{\alpha}{4} - \epsilon \right) = \frac{5}{\alpha} M_n \left(\frac{\alpha}{4} - \epsilon \right)$$

Strong Converse Variational Distance : $\mathrm{R}^{(1)}(2^-) \geq \frac{d}{2}$

$$M_n \geq \sum_{j=1}^{M_n} (\varphi(j)) \left(\bigcup_{i=1}^{\frac{5}{\alpha}M_n} \mathcal{D}_i \right) \qquad [\varphi(j) \text{ is a prob. meas.}]$$

$$= \sum_{i=1}^{\frac{5}{\alpha}M_n} \left(\sum_{j=1}^{M_n} (\varphi(j))(\mathcal{D}_i) \right) \qquad [\mathcal{D}_i \text{ are disjoint}]$$

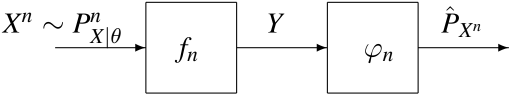
$$\geq \sum_{i=1}^{\frac{5}{\alpha}M_n} (\varphi \circ f(\theta_i))(\mathcal{D}_i) \qquad [\varphi \circ f \text{ is a cvx. comb. of } \varphi(j)]$$

$$\geq \sum_{i=1}^{\frac{5}{\alpha}M_n} \left(\frac{\alpha}{4} - \epsilon \right) = \frac{5}{\alpha} M_n \left(\frac{\alpha}{4} - \epsilon \right)$$

Contradiction if $0 < \epsilon < \frac{\alpha}{20}$.

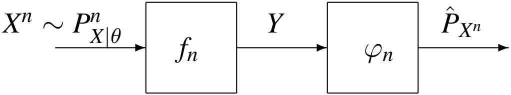
Outline

Concluding Remarks



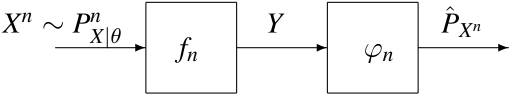
- Approximate sufficient statistics and minimum size of memory Y.

Concluding Remarks



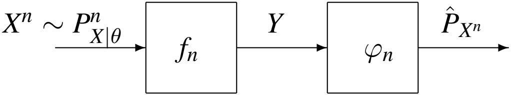
- Approximate sufficient statistics and minimum size of memory Y.
- The optimal rate $\frac{d}{2}$ (exponent in $n^{d/2}$) is reduced from d (cf. exact sufficient statistics) for multinomial distributions.

Concluding Remarks



- Approximate sufficient statistics and minimum size of memory Y.
- The optimal rate $\frac{d}{2}$ (exponent in $n^{d/2}$) is reduced from d (cf. exact sufficient statistics) for multinomial distributions.
- Weak results (weak converse and weak achievability) follow from the results by Rissanen and Clarke-Barron.

Concluding Remarks



- Approximate sufficient statistics and minimum size of memory Y.
- The optimal rate $\frac{d}{2}$ (exponent in $n^{d/2}$) is reduced from d (cf. exact sufficient statistics) for multinomial distributions.
- Weak results (weak converse and weak achievability) follow from the results by Rissanen and Clarke-Barron.
- Achievability and strong converse parts do not follow from them. We invented new methods.

Concluding Remarks

$$X^n \sim P^n_{X|\theta} \longrightarrow \boxed{f_n} \xrightarrow{\;Y\;} \boxed{\varphi_n} \xrightarrow{\;\hat{P}_{X^n}\;}$$

- Approximate sufficient statistics and minimum size of memory Y.
- The optimal rate $\frac{d}{2}$ (exponent in $n^{d/2}$) is reduced from d (cf. exact sufficient statistics) for multinomial distributions.
- Weak results (weak converse and weak achievability) follow from the results by Rissanen and Clarke-Barron.
- Achievability and strong converse parts do not follow from them. We invented new methods.
- arXiv 1612.02542 and submitted to the IEEE Trans. on Inform. Th.