信息论导论

第5讲 渐近均分性、典型集，Markov链、数据处理不等式、Fano不等式

[信息论教材中页码范围] 渐近均分性、典型集：p57~p63，Markov链 p71~74，数据处理不等式、Fano不等式：p34~p41

信息学部-信息科学与技术学院 吴绍华

hitwush@hit.edu.cn

随机变量的收敛

- （严格）收敛
  $X_n \xrightarrow[n\to\infty]{} Y \Rightarrow \forall \varepsilon > 0, \exists m$ 使得 $\forall n > m, |X_n - Y| < \varepsilon$
- 依概率收敛（弱于严格收敛）
  $X_n \xrightarrow[n\to\infty]{\text{prob}} Y \Rightarrow \forall \varepsilon > 0, \Pr(|X_n - Y| > \varepsilon) \xrightarrow[n\to\infty]{} 0$

例

- $X_n = \pm 2^{-n}, p = [1/2; 1/2]$
  $X_n \xrightarrow[n\to\infty]{} 0$ (选择 $m = 1 - \log \varepsilon$)
- $X_n \in \{0, 1\}$, $p = [1 - n^{-1}; n^{-1}]$
  $\forall \varepsilon > 0, p(|X_n - 0| > \varepsilon) = n^{-1} \xrightarrow[n\to\infty]{} 0$, 因此 $X_n \xrightarrow[n\to\infty]{\text{prob}} 0$

弱大数定律（WLLN）

定理——弱大数定律（Weak Law of Large Numbers, WLLN）

给定 i.i.d. $\{X_i\}$, $\mathrm{E}\,X_i = \mu$, $\mathrm{Var}\,X_i = \sigma^2 < \infty$, 则样本均值 $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ 依概率收敛至 μ，即,

$$S_n \xrightarrow[n\to\infty]{\text{prob}} \mu \Leftrightarrow \forall \varepsilon > 0, \Pr\left(|S_n - \mu| > \varepsilon\right) \xrightarrow[n\to\infty]{} 0$$

切比雪夫不等式（Chebyshevs Inequality）

$$\begin{aligned}\mathrm{Var}\,Y &= \mathrm{E}(Y-\mu)^2 = \sum_{y\in\mathscr{Y}} p(y)(Y-\mu)^2 \\ &\geq \sum_{y:|y-\mu|>\varepsilon} p(y)(Y-\mu)^2 \geq \sum_{y:|y-\mu|>\varepsilon} p(y)\varepsilon^2 \\ &= \varepsilon^2 \Pr\left(|Y-\mu| > \varepsilon\right)\end{aligned}$$

弱大数定律的证明

WLLN 的证明.

$$\operatorname{Var} S_n = \frac{\sigma^2}{n} \geq \varepsilon^2 \Pr(|S_n - \mu| > \varepsilon)$$

$$\Rightarrow \Pr(|S_n - \mu| > \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$$

$$\Rightarrow \Pr(|S_n - \mu| > \varepsilon) \xrightarrow[n\to\infty]{} 0$$

$$\text{因此}, S_n \xrightarrow[n\to\infty]{\text{prob}} \mu$$

WLLN有很多重要推论，其中之一便是**渐近均分性**
（**A**symptotic **E**quipartition **P**roperty, **AEP**）

渐近均分性（AEP）

定理 3.1.1 AEP

如果 $\{X_i\}$ 为 i.i.d. $\sim p(x)$, 则

$$-\frac{1}{n}\log p(X_{1:n}) \xrightarrow{\text{prob}} H(X)$$

证明.

$$-\frac{1}{n}\log p(X_{1:n}) = -\frac{1}{n}\log\prod_{i=1}^{n} p(X_i)$$

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

由WLLN ← $\xrightarrow{\text{prob}} -\mathrm{E}\log p(X) = H(X)$

□

典型序列与典型集

定理 3.1.1 AEP

如果 $\{X_i\}$ 为 i.i.d. $\sim p(x)$, 则

$$-\frac{1}{n}\log p(X_{1:n}) \xrightarrow{\text{prob}} H(X)$$

定义

关于 $p(x)$ 的典型集 $T_\varepsilon^{(n)}$ 指的是具有以下性质的序列 $x_{1:n} \in \mathscr{X}^n$ 的集合：

$$2^{-n(H(X)+\varepsilon)} \leq p(x_{1:n}) \leq 2^{-n(H(X)-\varepsilon)}$$

典型集的性质

定理 3.1.2 典型集的性质

1. $x_{1:n} \in T_\varepsilon^{(n)} \Rightarrow H(X) - \varepsilon \leq -\frac{1}{n}\log p(x_{1:n}) \leq H(X) + \varepsilon$
2. $\Pr\left\{T_\varepsilon^{(n)}\right\} > 1 - \varepsilon$, 当 $n > N_\varepsilon$
3. $\left|T_\varepsilon^{(n)}\right| \leq 2^{n(H(X)+\varepsilon)}$
4. $\left|T_\varepsilon^{(n)}\right| \geq (1-\varepsilon)2^{n(H(X)-\varepsilon)}$, 当 $n > N_\varepsilon$

性质 (1) 与 (2) 的证明

- 由 $T_\varepsilon^{(n)}$ 的定义易得性质 (1)
- $\Pr\{T_\varepsilon^{(n)}\} = p(x_{1:n} \in T_\varepsilon^{(n)})$. 由定理 3.1.1，事件 $X_{1:n} \in T_\varepsilon^{(n)}$ 的概率随着 $n \to \infty$ 趋近于 1
  从而 $\forall \varepsilon > 0, \exists N_\varepsilon$ 使得 $\forall n > N_\varepsilon$,
  $p(x_{1:n} \in T_\varepsilon^{(n)}) = p(|-\frac{1}{n}\log p(x_{1:n}) - H(X)| < \varepsilon) > 1 - \varepsilon$

典型集的性质（续）

性质 (3) 的证明

- 性质 (3)

$$
\begin{aligned}
1 &= \sum_{x_{1:n}\in\mathcal{X}^n} p(x_{1:n}) \geq \sum_{x_{1:n}\in T_\varepsilon^{(n)}} p(x_{1:n}) \\
&\geq \sum_{x_{1:n}\in T_\varepsilon^{(n)}} 2^{-n(H(X)+\varepsilon)} = 2^{-n(H(X)+\varepsilon)} |T_\varepsilon^{(n)}|
\end{aligned}
$$

典型集的性质（续）

性质 (4) 的证明

- 性质 (4)

$$1-\varepsilon \overset{\text{for } n>N_\varepsilon}{<} p(x_{1:n}\in T_\varepsilon^{(n)}) = \sum_{x_{1:n}\in T_\varepsilon^{(n)}} p(x_{1:n})$$
$$\leq \sum_{x_{1:n}\in T_\varepsilon^{(n)}} 2^{-n(H(X)-\varepsilon)} = 2^{-n(H(X)-\varepsilon)}|T_\varepsilon^{(n)}|$$

典型集示例

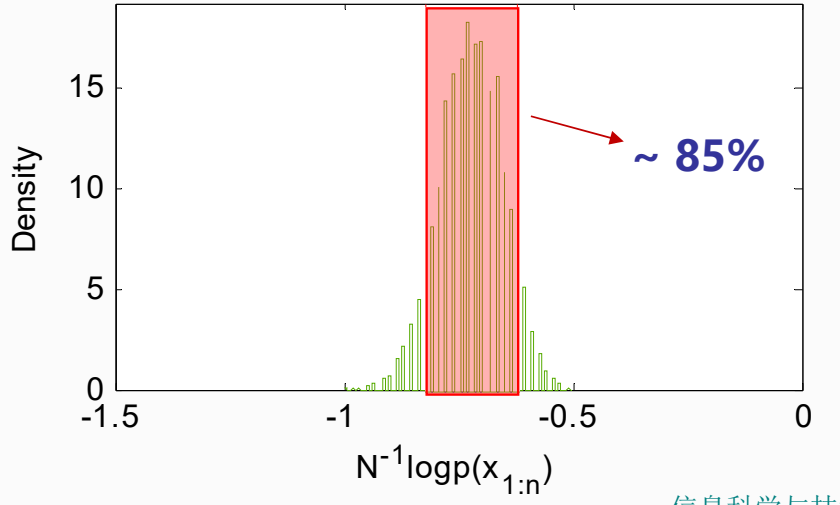
例

随机变量 X_i 服从伯努利分布，且 $p(X_i = 1) = p$, 则

$$p(X_{1:n}) = p^{\sum X_i}(1-p)^{n-\sum X_i}$$

对于 $p = 0.2$, $H(X) = 0.72$bits

下图红色区域显示了 $T_{0.1}^{(128)}$



并非“→1”，为什么?

AEP的简单解读

- 对于任何 ε 和 $n > N_\varepsilon$
  ——“几乎所有事情都令人同等的意外”
  ——”Almost all events are almost equally surprising”

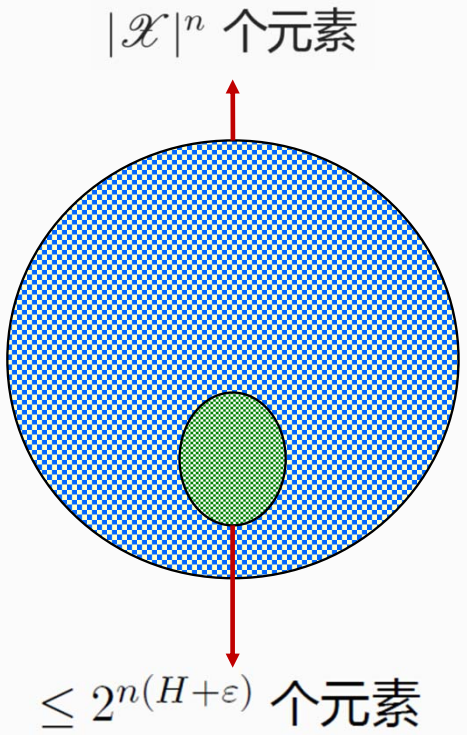
定理 3.1.2 典型集的性质

1. $x_{1:n} \in T_\varepsilon^{(n)} \Rightarrow H(X) - \varepsilon \leq -\frac{1}{n}\log p(x_{1:n}) \leq H(X) + \varepsilon$
2. $\Pr\left\{T_\varepsilon^{(n)}\right\} > 1 - \varepsilon$, 当 $n > N_\varepsilon$
3. $\left|T_\varepsilon^{(n)}\right| \leq 2^{n(H(X)+\varepsilon)}$
4. $\left|T_\varepsilon^{(n)}\right| \geq (1-\varepsilon)2^{n(H(X)-\varepsilon)}$, 当 $n > N_\varepsilon$

典型集的应用举例——信源编码

- 编码方案
  - $x_{1:n} \in T_\varepsilon^{(n)}$: '0' + 至多 $1 + n(H+\varepsilon)$ bits
  - $x_{1:n} \notin T_\varepsilon^{(n)}$: '1' + 至多 $1 + n\log|\mathscr{X}|$ bits
  - 平均码长:

$$\begin{aligned} L &\le p(x_{1:n} \in T_\varepsilon^{(n)})[2 + n(H+\varepsilon)] \\ &\quad + p(x_{1:n} \notin T_\varepsilon^{(n)})[2 + n\log|\mathscr{X}|] \\ &\le n(H+\varepsilon) + 2 + \varepsilon(2 + n\log|\mathscr{X}|) \\ &= n(H + \varepsilon + \varepsilon\log|\mathscr{X}| + \frac{2+2\varepsilon}{n}) \\ &= n(H+\varepsilon') \end{aligned}$$



使用典型集证明Shannon第一定理

对于任意 $\varepsilon > 0$, 通过选择足够大的序列长度 n:

- 使用前述基于典型集的无损信源编码，平均每个符号所需码字长度仅为 $H + \varepsilon$ bits，即

$$\frac{L}{n} \leq H + \varepsilon \underset{n\to\infty}{\to} H$$

- 这实际上是对信源编码定理（Shannon 第一定理）的另一种证明。
- 需注意：前述基于典型集的编码方案仅理论上可行——由于编码、解码的复杂度都以 n 的指数递增，是无法实用的。

阶段小结1

- AEP: $-\frac{1}{n}\log p(X_{1:n}) \xrightarrow{\text{prob}} H(X)$
- 典型集:
  - 个体概率: $H(X)-\varepsilon \le -\frac{1}{n}\log p(x_{1:n}) \le H(X)+\varepsilon$
  - 总体概率: $\Pr\{T_\varepsilon^{(n)}\} = p(x_{1:n}\in T_\varepsilon^{(n)}) > 1-\varepsilon$, 对于 $n > N_\varepsilon$
  - 集合大小: $(1-\varepsilon)2^{n(H(X)-\varepsilon)} \overset{\text{for } n>N_\varepsilon}{\le} |T_\varepsilon^{(n)}| \le 2^{n(H(X)+\varepsilon)}$
- AEP 的简单解读与应用:
  - "Almost all events are almost equally surprising"
  - 使用基于典型集的无损信源编码，平均每符号码长仅需 $H+\varepsilon$ bits ——Shannon 第一定理的另一种证明方式

马尔可夫（Markov）链

- 考虑三个随机变量：X, Y, Z，有

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

若满足：

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

则它们形成了一条Markov 链$X \to Y \to Z$

- Markov 链 $X \to Y \to Z$ 意味着
  - X 只能通过 Y 来影响 Z
  - 若已知 Y，则观测到 X 并不能获得关于 Z 的额外信息，即

$$I(X;Z|Y) = 0 \Leftrightarrow H(Z|Y) = H(Z|X,Y)$$

  - 同理（互信息的对称性）：若已知 Y，则观测到 Z 并不能获得关于 X 的额外信息

两个简单结论

马尔可夫性意味着条件独立

若 $X \to Y \to Z$，那么

$$p(x,z|y) = \frac{p(x,y,z)}{p(y)} = \frac{p(x,y)p(z|y)}{p(y)} = p(x|y)p(z|y)$$

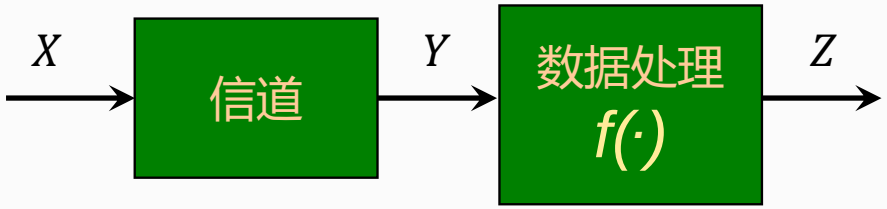
因此，在给定 Y 的条件下，随机变量 X 和 Z 是条件独立的。

马尔可夫性是对称的

$X \to Y \to Z \Leftrightarrow Z \to Y \to X$ ，因为

$$p(x|y) = p(x|y)\frac{p(z|y)p(y)}{p(y,z)} = \frac{p(x,z|y)p(y)}{p(y,z)} = \frac{p(x,y,z)}{p(y,z)} = p(x|y,z)$$

重要的Markov链——数据传输与处理



传统处理

压缩感知、
随机共振、
……

- $Z = f(Y)$，其中 $f(\cdot)$ 是一个确定性的或是随机性的函数。如此构成了一个与信息流通系统相对应的 Markov 链

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

- 通过对 Y 进行处理，能否获取到更多关于 X 的信息？

数据处理不等式

定理 2.8.1

若 $X \to Y \to Z$，那么 $I(X;Y) \geq I(X;Z)$

证明.

利用互信息的链式法则，将互信息以两种不同方式展开：

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

然而，$I(X;Z|Y) = 0$

从而，$I(X;Y) = I(X;Z) + I(X;Y|Z)$

因此，$I(X;Y) \geq I(X;Z)$ 且 $I(X;Y) \geq I(X;Y|Z)$ □

处理Y并不能增加关于X的新信息

知道Z仅会减少Y中所包含的关于X的信息

相关推论

推论（第 35 页）

若 $X \to Y \to Z$，那么 $I(X;Y) \geq I(X;Y|Z)$

推论（第 35 页）

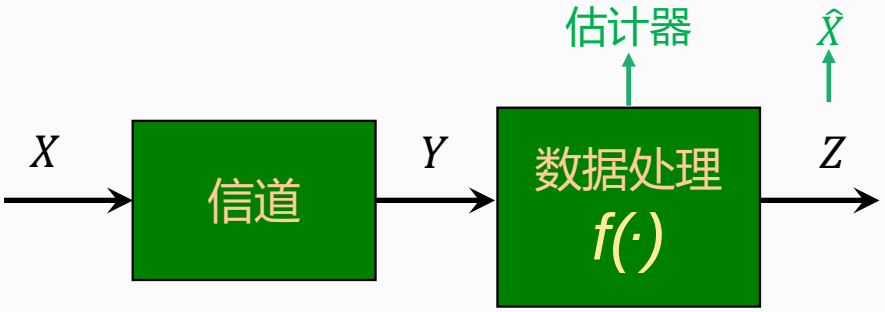
若 $Z = f(Y)$，那么 $I(X;Y) \geq I(X;f(Y))$

长马尔可夫链

若 $X_1 \to X_2 \to X_3 \to X_4 \to X_5 \to X_6$，那么随着两个变量距离变近，其互信息将会增加，例如

$$I(X_3;X_4) \geq I(X_2;X_4) \geq I(X_1;X_5) \geq I(X_1;X_6)$$

为什么还要进行数据处理?



- 既然处理数据并不能增加额外的信息，那为什么还要对数据进行处理呢?
- 处理的目的通常是旨在通过 Y 估计出 X
- 并且：是有可能做到不损失信息的（当数据处理不等式等号成立时）
- 充分统计量：Z 包含了 Y 中所有关于 X 的信息，即 $I(X;Y) = I(X;Z)$

充分统计量

设 $T(Y)$ 是针对 Y 的处理结果（如：基于 Y 的某个统计量），则

$$X \to Y \to T(Y) \quad \Rightarrow \quad I(X;T(Y)) \le I(X;Y)$$

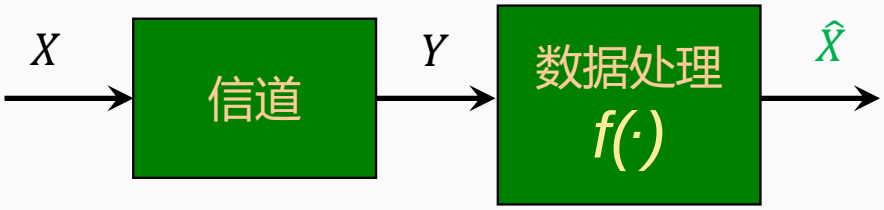
定义

给定 $T(Y)$ 的条件下，如果 Y 与 X 条件独立，即 $X \to T(Y) \to Y$ 也构成马尔可夫链，则称 $T(Y)$ 是关于 X 的充分统计量。

- 一些结论：

$$\begin{aligned} X \to Y \to T(Y) \to X &\Leftrightarrow I(X;T(Y)) = I(X;Y) \\ &\Leftrightarrow X \to T(Y) \to Y \to X \\ &\Leftrightarrow p(Y|T(Y),X) = p(Y|T(Y)) \end{aligned}$$

Fano不等式



定理 2.11.1 (Fano 不等式)

设 $\hat{X}$ 为通过 Y 所得到的对 X 的估计，则估计误差概率 $P_e = \Pr(\hat{X} \neq X)$ 满足如下不等式：

$$H(X|Y) \leq H(P_e) + P_e \log |\mathscr{X}|$$

$$\Rightarrow P_e \geq \frac{H(X|Y) - H(P_e)}{\log |\mathscr{X}|} \geq \frac{H(X|Y) - 1}{\log |\mathscr{X}|}$$

这个形式更弱，但更易于使用

Fano不等式的证明

证明.

- 定义一个误差随机变量 $E = (\hat{X} \neq X)?1 : 0$
- 利用熵的链式法则，将 $H(E, X|Y)$ 以两种不同方式展开

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

- 由于 $H(E|X, Y) = 0$ 且 $H(E|Y) \leq H(E) = H(P_e)$，有

$$\begin{aligned} H(X|Y) + 0 &\leq H(P_e) + H(X|E, Y) \\ &= H(P_e) + H(X|E = 0, Y)p(E = 0) + H(X|E = 1, Y)p(E = 1) \\ &\leq H(P_e) + 0 \times (1 - P_e) + \log |\mathscr{X}| P_e \end{aligned}$$

□

Fano不等式暗含的结果

$$P_e \geq \frac{H(X|Y) - H(P_e)}{\log|\mathscr{X}|} \geq \frac{H(X|Y) - 1}{\log|\mathscr{X}|}$$

- 零误差概率（即 $P_e = 0$）$\Rightarrow H(X|Y) = 0$
- 若 $H(X|Y)$ 较小，则误差概率可能也较低
- 若 $H(X|Y)$ 较大，则误差概率必然也高
- 可略微强化为

$$P_e \geq \frac{H(X|Y) - H(P_e)}{\log(|\mathscr{X}| - 1)} \geq \frac{H(X|Y) - 1}{\log(|\mathscr{X}| - 1)}$$

- Fano 不等式可用于任意需要对误差概率进行定量描述或是需要表明误差不可避免的情形。例如：Shannon 第二定理逆定理的证明。

Fano不等式的例子

例

- $\mathscr{X} = [1; 2; 3; 4; 5]$, $\mathbf{p}_X = [0.35; 0.35; 0.1; 0.1; 0.1]$
  $\mathscr{Y} = [1; 2]$，若 $X \leq 2$ 则 $Y = X$ 有 $6/7$ 的概率发生，而若 $X > 2$ 则 $Y = 1$ 或 $Y = 2$ 按均等概率发生。
- 易知估计 X 的最好策略是令 $\hat{X} = Y$，其对应的实际误差概率为 $P_e = 0.4$
- $H(X|Y)$ 与 Fano 界：

$$H(X|Y) = 1.771$$

$$P_e \geq \frac{H(X|Y) - 1}{\log(|\mathscr{X}| - 1)} = \frac{1.771 - 1}{\log(4)} = 0.3855$$

思考：Fano不等式取“=”对应什么情形?

阶段小结2

- 马尔可夫链： $X \to Y \to Z \Leftrightarrow p(z|x,y) = p(z|y) \Leftrightarrow I(X;Z|Y) = 0$
- 数据处理不等式：若 $X \to Y \to Z$，那么
  - $I(X;Y) \geq I(X;Z)$
  - $I(X;Y) \geq I(X;Y|Z)$
- 充分统计量：在 $X \to Y \to T(Y)$ 中，若 $I(X;T(Y)) = I(X;Y)$ 则称 $T(Y)$ 是 X 的一个充分统计量
- Fano 不等式：若 $X \to Y \to \hat{X}$

$$P_e \geq \frac{H(X|Y) - H(P_e)}{\log(|\mathscr{X}| - 1)} \geq \frac{H(X|Y) - 1}{\log(|\mathscr{X}| - 1)} \geq \frac{H(X|Y) - 1}{\log|\mathscr{X}|}$$

由于与P_e无关，因此尽管更弱但更易于使用

结束