SSReflect による 可変長情報源符号化順定理の形式化

小尾　良介

joint work with

萩原 学（千葉大学）

Reynald Affeldt（AIST）

千葉大学 理学研究科

September 5, 2014

Table of Contents

形式化について

- **形式化の目的と利点**
  - **証明の正当性の検証**

    証明における曖昧さの排除，概念の形式的定義
  - **ライブラリの検証**

- Coq/SSReflect **による形式化**
  - **四色定理**（Gonthier，2008）
  - Feit-Thompson **の定理**（Gonthier et al.，2013）
  - Shannon **の定理**（Affeldt et al.，2014）

Shannon の定理

「A Mathematical Theory of Communication」(1948) における，
Shannon の定理一覧

	情報源符号化		通信路符号化
	固定長	可変長	
順定理	○	●	○
逆定理	○	●	○

○ · · · 形式化済
● · · · 本研究
● · · · 未形式化

情報源符号化について

- **情報源符号化とはデータ圧縮のこと．**
  代表的な圧縮形式として非可逆圧縮の mp3**，**jpeg **や，可逆圧縮の** zip**，**lzh**．**
- **情報源が分布になってる．**

  例

$$\mathcal{X} := \{\, \textbf{あ}, \textbf{い}, \textbf{う} \,\}$$
$$P(\textbf{あ}) = 0.7,\ P(\textbf{い}) = 0.2,\ P(\textbf{う}) = 0.1$$

あううあああいあいあいあああああいあああああ . . .

$$\textbf{あ} : \textbf{い} : \textbf{う} = 7 : 2 : 1$$

informal statement

Theorem （可変長情報源符号化順定理）

- X^n ： n 個の確率変数（それぞれが独立な同一分布に従う）
- $H(X)$ ： 確率変数 X のエントロピー

n が十分大きいとき，
次を満たす，単射 f が存在する．

$$\mathrm{E}\left[\frac{1}{n}l(X^n)\right] \leq H(X) + \varepsilon.$$

ただし，
- f ：X^n の事象からビット列への符号化写像 $(f : \mathcal{X}^n \to \mathbb{F}_2^*)$，
- $l(x)$：系列 x の像の長さ（$f(x)$ の長さ），とする．

像の長さを n で割った平均を，$H(X)$ ビットより少し大きい値まで圧縮できる．

$P(\text{あ}) = 0.7,\ P(\text{い}) = 0.2,\ P(\text{う}) = 0.1$ ならば，圧縮後の長さを n で割った平均を $H(P)(\approx 0.35)$ ビットくらいまで圧縮できる．

証明（符号化写像の定義）

Definition （長さ n の典型系列）

次の条件を満たす系列 $x \in \mathcal{X}^n$ を典型系列という．

$$2^{-n(H(P)+\varepsilon)} \leq P^n(x) \leq 2^{-n(H(P)-\varepsilon)}.$$

典型系列全体の集合を $\mathcal{TS}(n, P, \varepsilon)$ と記述する．

例 ： $n := 10, \quad \varepsilon := 0.82$ のとき，
ああいうあああああいあ $\Rightarrow$ 典型系列
うううういあううあううう $\Rightarrow$ 非典型系列

Lemma （$\mathcal{TS}(n, P, \varepsilon)$ の濃度の上界）

典型系列全体の集合の濃度 $|\mathcal{TS}(n, P, \varepsilon)|$ は次の上界をもつ．

$$|\mathcal{TS}(n, P, \varepsilon)| \leq 2^{n(H(P)+\varepsilon)}.$$

証明（符号化写像の定義）

- **符号化写像 $f : \mathcal{X}^n \to \mathbb{F}_2^*$ を定める．**
  定義域を二つに分ける．

$$\mathcal{X}^n = \underbrace{\mathcal{TS}(n,P,\varepsilon)}_{\text{高々}2^{n(H(P)+\varepsilon)}\text{ 個の系列}} \cup \underbrace{\mathcal{TS}^c(n,P,\varepsilon)}_{\text{高々}|\mathcal{X}^n|\text{ 個の系列}}$$

単射 f を次のように定義：

$$x \in \mathcal{TS}(n,P,\varepsilon) \Longrightarrow f(x) = 1 :: \Big[\text{長さ } L_0 = \lceil n(H(P)+\varepsilon)\rceil \text{ のビット列}\Big]$$

$$x \notin \mathcal{TS}(n,P,\varepsilon) \Longrightarrow f(x) = 0 :: \Big[\text{長さ } L_1 = \lceil \log_2(|\mathcal{X}^n|)\rceil \text{ のビット列}\Big]$$

証明

Theorem (可変長情報源符号化順定理)

n **が十分大きいとき，** $\cdots$ $\exists n_0, n_0 < n$ **に対して,**
次を満たす，単射（よって可逆）が存在する．

$$\mathrm{E}\left[\frac{1}{n} l(X^n)\right] \leq H(X) + \varepsilon.$$

$\varepsilon' := \frac{\varepsilon}{3+3\log_2 |\mathcal{X}|}$, $n_0 := \max\left[\left\lceil \frac{2}{1+\log_2 |\mathcal{X}|} \right\rceil, \left\lceil \frac{8}{\varepsilon} \right\rceil, \left\lceil \frac{\sigma^2}{\varepsilon'^3} \right\rceil\right]$ **とする．**

$$\begin{aligned}
\mathrm{E}[l(X^n)] &= \sum_{x \in \mathcal{X}^n} P^n(x) l(x) \\
&= \sum_{x \in \mathcal{TS}(n,P,\varepsilon')} P^n(x) l(x) + \sum_{x \in \mathcal{TS}(n,P,\varepsilon')^c} P^n(x) l(x) \\
&= \sum_{x \in \mathcal{TS}(n,P,\varepsilon')} P^n(x)(L_0 + 1) + \sum_{x \in \mathcal{TS}(n,P,\varepsilon')^c} P^n(x)(L_1 + 1)
\end{aligned}$$

Lemma（長い典型系列集合の生起確率の下界）

$\forall n, \frac{\sigma^2}{\varepsilon'^3} \leq n$ **のとき，典型系列集合の生起確率には次の下界をもつ．**

$$\sum_{x \in \mathcal{TS}(n,P,\varepsilon')} P^n(x) \geq 1 - \varepsilon'.$$

ただし， $\sigma^2 = \sum_{x \in \mathcal{X}^n} P(x)(\log P(x))^2 - (H(P))^2$ **とする．**

$$n_0 := \max\left[\left\lceil \frac{2}{1+\log_2 |\mathcal{X}|} \right\rceil, \left\lceil \frac{8}{\varepsilon} \right\rceil, \left\lceil \frac{\sigma^2}{\varepsilon'^3} \right\rceil\right]$$

$$\begin{aligned}
&= (L_0+1) \sum_{x \in \mathcal{TS}(n,P,\varepsilon')} P^n(x) + (L_1+1) \sum_{x \in \mathcal{TS}(n,P,\varepsilon')^c} P^n(x) \\
&\leq (L_0+1) + \varepsilon'(L_1+1) \\
&\leq n\left(H(P) + \frac{1}{3}\varepsilon + \frac{2}{3n(1+\log_2 |\mathcal{X}|)}\varepsilon + \frac{2}{n}\right) \\
\mathrm{E}[l(X^n)] &\leq n(H(P)+\varepsilon)
\end{aligned}$$

□

formal statement

- **符号化関数の型（$\mathcal{X}^n \to \mathbb{F}_2^*$）**

```
Definition var_enc X n := n.-tuple X → seq bool.
Variable f : var_enc X n.
```

- **平均符号長（$\mathrm{E}[l(X^n)]$）の定義**

```
Definition exp_len_cw f P :=
    ℰ (mkRvar (P^n) (fun x ⇒(size (f x)))).
```

Theorem （可変長情報源符号化順定理）

```
Variable X : finType.
Variable n : nat.
Variable ε : R.
Hypothesis ep_pos : 0 < ε.
Definition n0 := ... .

Theorem vscode : n0 < n →
  ∃ f : var_enc X n,
    injective f ∧
    exp_len_cw f P / n ≤ ℋ P + ε.
```

Theorem

$\exists n_0, n_0 \leq n$ **に対して，**
次を満たす，単射 f **が存在する．**

$$\mathrm{E}\left[\frac{1}{n} l(X^n)\right] < H(X) + \varepsilon.$$

符号化写像の形式化

単射 f を次のように定義：

$$x \in \mathcal{TS}(n,P,\varepsilon) \Rightarrow f(x) = 1 :: \Big[\textbf{長さ } L_0 = \lceil n(H(P)+\varepsilon)\rceil \textbf{ のビット列}\Big]$$

$$x \notin \mathcal{TS}(n,P,\varepsilon) \Rightarrow f(x) = 0 :: \Big[\textbf{長さ } L_1 = \lceil \log_2(|\mathcal{X}^n|)\rceil \textbf{ のビット列}\Big]$$

L_0, L_1 の定義

```
Definition L0 := ⌈ n * (ℋP + ε) ⌉.
Definition L1 := ⌈ log (♯ |[set : n.-tuple X]|) ⌉.
```

```
Definition ceil r : Z := - Int_part (- r).
Lemma ceil_upper:∀ r, ceil r < r + 1.
Lemma ceil_bottom:∀ r, r ≤ ceil r.
```

符号化写像の形式化

単射 f を次のように定義：

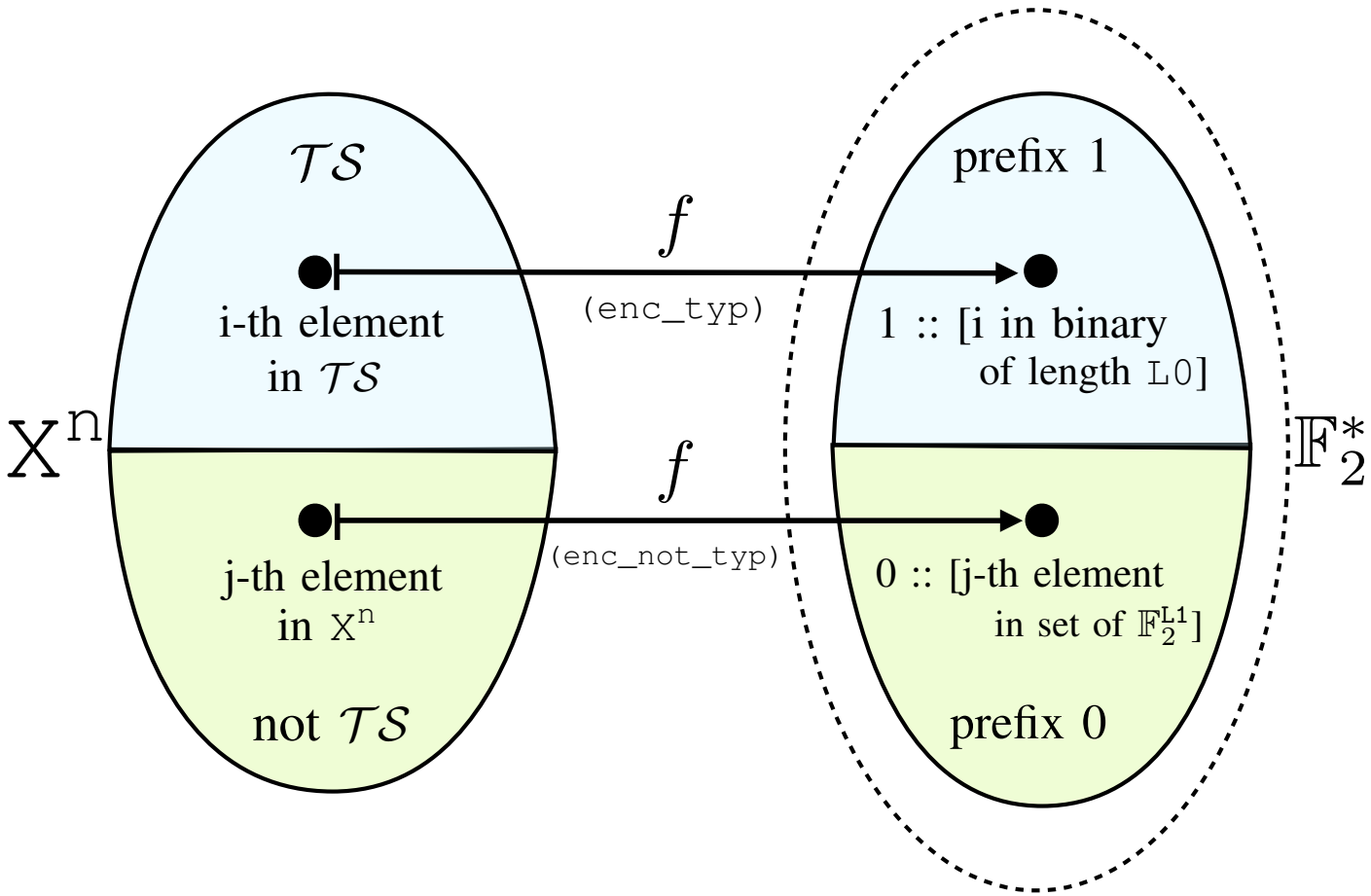
$$x \in \mathcal{TS}(n,P,\varepsilon) \Rightarrow f(x) = 1 :: \Big[\textbf{長さ } L_0 = \lceil n(H(P)+\varepsilon)\rceil \textbf{ のビット列}\Big]$$

$$x \notin \mathcal{TS}(n,P,\varepsilon) \Rightarrow f(x) = 0 :: \Big[\textbf{長さ } L_1 = \lceil \log_2(|\mathcal{X}^n|)\rceil \textbf{ のビット列}\Big]$$

f を次のように定義：

```
Definition f : var_enc X n := fun x ⇒
  if x ∈ 𝒯𝒮 P n ε then
    true :: enc_typ x
  else
    false :: enc_not_typ x.
```

```
Definition enc_typ x :=
 let i := index x (enum (𝒯𝒮 P n ε))
 in Tuple (size_nat2bin i L0).
```




```
Definition enc_not_typ x := enum_val
    (widen_ord le_n_L1_tuple (enum_rank x)).
```

今回の形式化について

Theorem （可変長情報源符号化順定理）

```
Variable X : finType.
Variable n : nat.
Variable ε : R.
Hypothesis ep_pos : 0 < ε.
Definition n0 := ... .

Theorem vscode : n0 < n →
  ∃ f : var_enc X n,
    injective f ∧
    exp_len_cw f P / n ≤ ℋ P + ε.
```

f に関連した証明

- 補題：17 個，約 230 行
  - L_0, L_1 に関する不等式
  - 単射性
  - 逆写像の構成
  - 一意復号可能性

n_0 に関連した証明

- 補題：10 個，約 200 行

行数のかかる証明の例：

$$\frac{2}{3n(1+\log_2|\mathcal{X}|)}\varepsilon \leq \frac{1}{3}\varepsilon$$

まとめ

- 本研究にて形式化した定理，定義：
  - 可変長情報源符号化順定理
  - 天井関数の定義
  - 天井関数の上界，下界
- 今後の研究題目：
  - 別証明での，可変長情報源符号化順定理の形式化
  - 可変長情報源符号化逆定理の形式化