On the Empirical Cumulant Generating Function of Code Lengths for Individual Sequences

Neri Merhav

The Andrew & Erna Viterbi Faculty of Electrical Engineering

Technion—Israel Institute of Technology

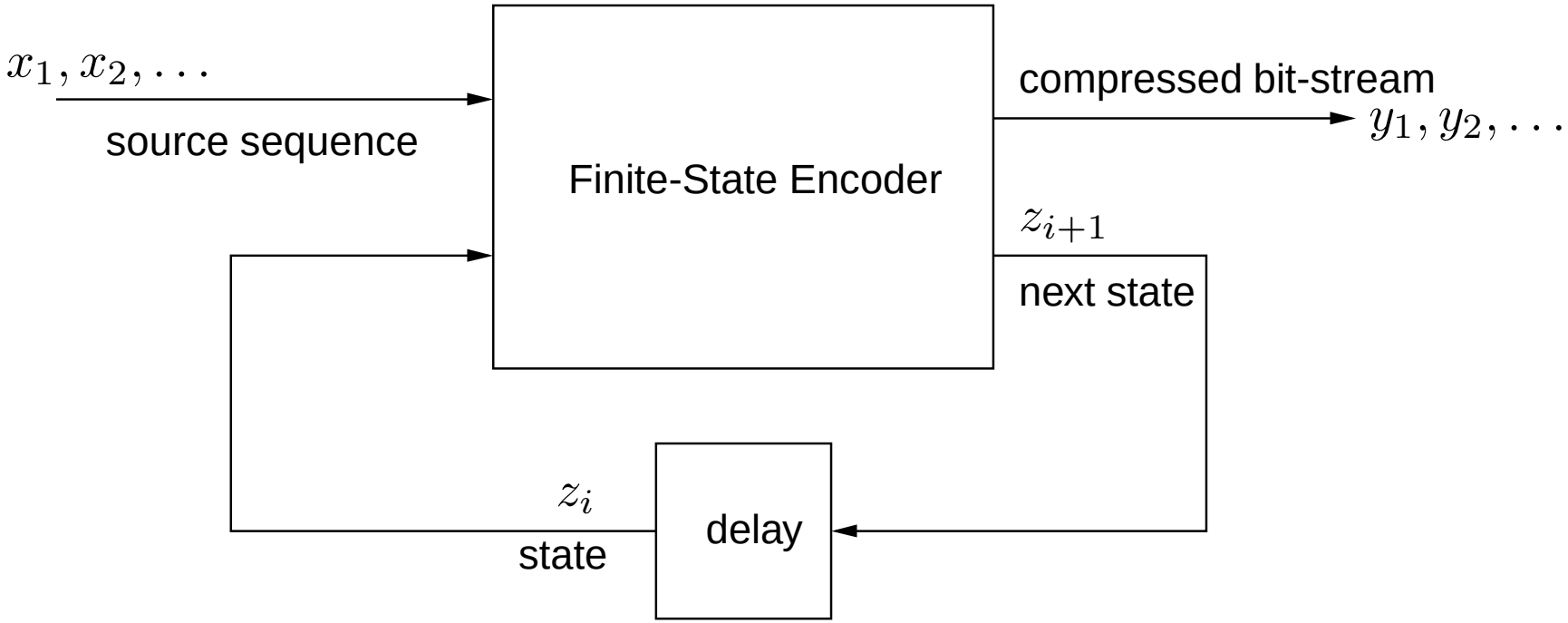
Haifa 3200004, Israel

ISIT 2017, Aachen, Germany, June 2017.

Motivation and Background

- Ziv & Lempel (1978): individual–sequence approach to data compression.
- Finite–State (FS) compressibility $\Leftrightarrow$ entropy rate.
- Campbell (1965): $\log \mathbf{E}\{\exp[\lambda \cdot \text{code-length}]\} \geq \lambda \cdot$ Rényi entropy.
- Motivation: risk–sensitivity, robustness, tail behavior (large deviations)...
- What would be the individual–sequence counterpart of the Rényi entropy?
- Wanted: tight lower bound on the empirical CGF of the code–length ...

Finite–State (FS), Information Lossless (IL) Encoder



$$
\begin{aligned}
y_i &= f(z_i, x_i,) \\
z_{i+1} &= g(z_i, x_i) \\
L(y^n) &= \sum_{t=1}^{n} l(y_t)
\end{aligned}
$$

Information losslessness = (z_1, y^n, z_n) uniquely determine x^n.

Defining the Empirical CGF

Compression ratio = $\frac{1}{n}\sum_{t=1}^{n} l(y_t)$ = empirical expectation of code–length.

First attempt to define empirical CGF:

$$\frac{1}{\lambda}\log\left[\frac{1}{n}\sum_{t=1}^{n}2^{\lambda l(y_t)}\right].$$

Difficulty: For many codes, $\ell(y_t)=0$ for most t.

Possible solutions:

- Simply ignore terms with $\ell(y_t)=0$.
- Define in the block level: fixed–to–variable CGF:

$$\frac{1}{\lambda\ell}\log\left[\frac{\ell}{n}\sum_{t=0}^{n/\ell}\exp_2\{\lambda L(y_{i\ell+1}^{i\ell+\ell})\}\right].$$

Main Result for Fixed–to–Variable Length CGF's

Theorem: For every IL encoder with s states,

$$\frac{1}{\lambda\ell}\log_2\left[\frac{\ell}{n}\sum_{t=0}^{n/\ell-1}2^{\lambda L(y_{t\ell+1}^{t\ell+\ell})}\right]\geq\hat{H}_\lambda^\ell(x^n)-\frac{\gamma(s,\ell)}{\ell},$$

where

$$\hat{H}_\lambda^\ell(x^n)=\frac{1+\lambda}{\lambda\ell}\log_2\left(\sum_{a^\ell\in\mathcal{X}^\ell}[\hat{P}(a^\ell)]^{1/(1+\lambda)}\right),$$

$\hat{P}(a^\ell)$ being the empirical probability of a^ℓ in x^n along its n/ℓ non–overlapping ℓ–blocks, and

$$\gamma(s,\ell)=2\log s+\log\left[1+\log\left(\frac{s^2+\alpha^\ell}{s^2}\right)\right].$$

The proof is based on the generalized Kraft inequality [ZL78].

"Achievability" – by an optimal Campbell code w.r.t. $\{\hat{P}(\cdot)\}$.

Another (Conceptually Simple) Lower Bound

Theorem: For every IL encoder with s states,

$$\frac{1}{\lambda\ell}\log\left[\frac{\ell}{n}\sum_{t=0}^{n\ell-1}\exp_2\{\lambda L(y_{t\ell+1}^{t\ell+\ell})\}\right]$$

$$\geq \frac{1}{\lambda\ell}\log\left[\frac{\ell}{n}\sum_{t=0}^{n/\ell-1}\exp_2\left\{\lambda(c_t+s^2)\log[(c_t+s^2)/4s^2]\right\}\right]$$

where $c_t =$ maximum number of phrases at block no. t.

This is a simple application of the lower bound of [ZL78] on $L(y_{t\ell+1}^{t\ell+\ell})$.

Variable–to–Variable Length CGFs – Discussion

- Problem with V–F CGF: large ℓ – large fluctuations.
- Extending the scope to V–V setting: more flexibility to reduce fluctuations.
- Sequence–dependent segmentation instead of fixed–length blocks.
- Dictionary of different phrases with $\hat{P}(\text{phrase}) \sim$ Unif.
- Same probabilities – same code lengths.
- Simple strategy: parse x^n to c distinct phrases, each appearing just once.

V–V Length CGFs – Discussion (Cont'd)

In particular, let x^n be parsed as

$$x_1^{n_1}, x_{n_1+1}^{n_2}, \ldots, x_{n_{c-1}+1}^{n}$$

and define

$$\rho_E^{\lambda}(x^n) = \frac{c}{n\lambda} \log \left[\frac{1}{c} \sum_{i=1}^{c} 2^{\lambda L(y_{n_{i-1}+1}^{n_i})} \right], \quad n_0 \equiv 0, \quad n_c \equiv n$$

Observation: Even if the decoder knew the dictionary in advance and there was no FS structure, $L(y_{n_{i-1}+1}^{n_i}) \sim \log c$, and so, one would expect

$$\rho_E^{\lambda}(x^n) \geq \frac{c}{n\lambda} \log \left[\frac{1}{c} \sum_{i=1}^{c} 2^{\lambda \log c} \right] = \frac{c \log c}{n}$$

in agreement with the ordinary compressibility.

Main Result for V–V CGFs

Theorem: For any IL encoder with no more than s states, and given a source sequence x^n with c distinct phrases,

$$\sum_{i=1}^{c} \exp_2\{\lambda L(y_{n_{i-1}+1}^{n_i})\} \geq \frac{s^2\left[\exp_2\left\{(\lambda+1)\log\left(\frac{c+s^2}{2s^2}\right)\right\}-1\right]}{2^{\lambda+1}-1}.$$

The proof is based on the assumed IL property, like in the converse theorem of [ZL78].

Alternative Lower Bound

Theorem: For any IL encoder with no more than s states, and given a source sequence x^n with c distinct phrases,

$$\sum_{i=1}^{c} \exp_2\{\lambda L(y_{n_{i-1}+1}^{n_i})\} \geq \exp_2\{(\lambda+1)\log c - \lambda\gamma(s, \log_\alpha c)\}.$$

This one is based on the generalized Kraft inequality.

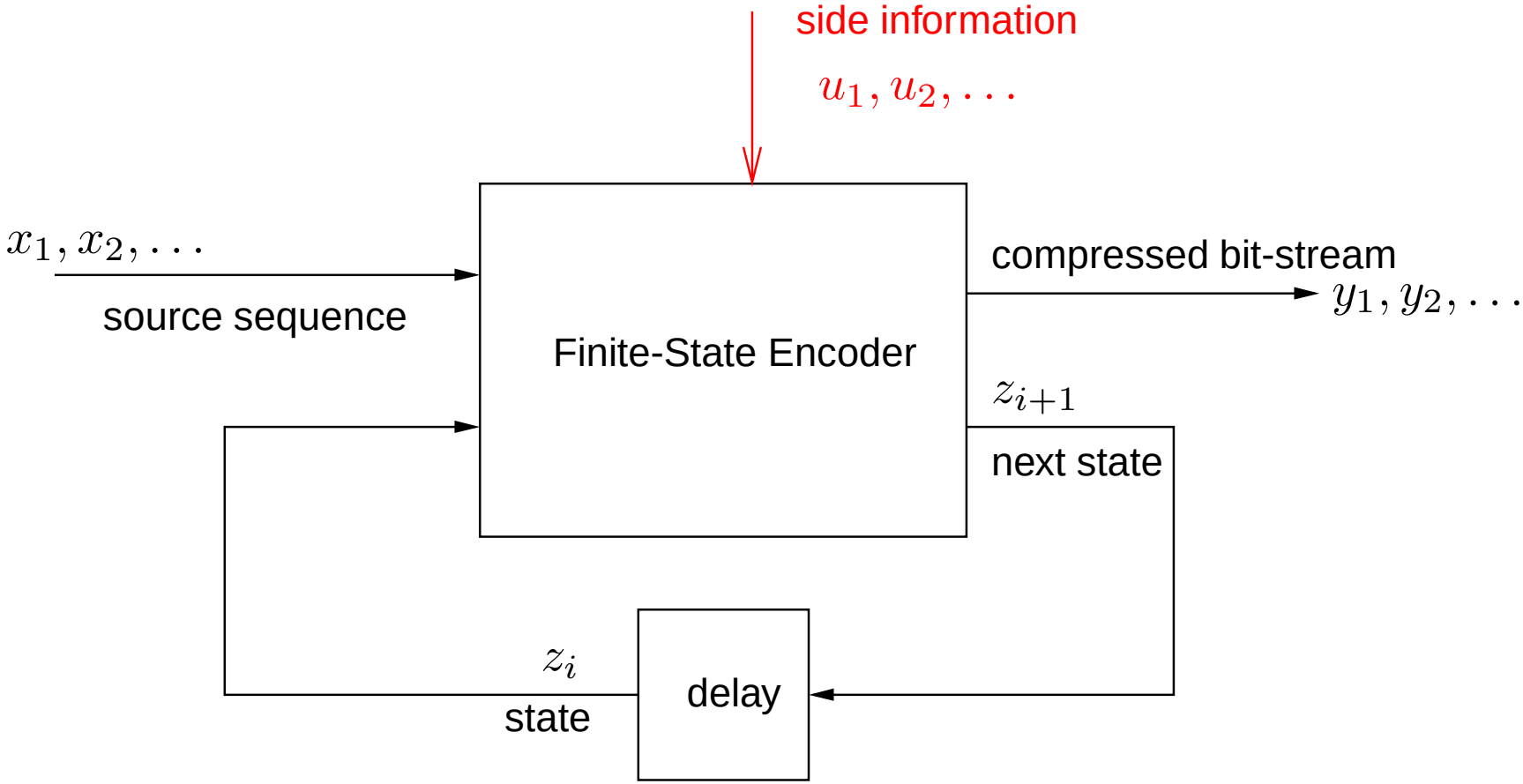
Compatible Achievability Result

Theorem: Let x^n be given and let c denote the number of phrases resulting from the incremental parsing procedure. Let $L_{\mathsf{LZ}}(x_{n_{i-1}+1}^{n_i})$ denote the total length associated with the compression of the i–th phrase according to the LZ78 algorithm. Then,

$$\sum_{i=1}^{c} \exp_2\{\lambda L_{\mathsf{LZ}}(x_{n_{i-1}+1}^{n_i})\} \leq (2\alpha)^{\lambda} 2^{(\lambda+1)\log c}.$$

The proof is by a simple performance analysis of the LZ78 algorithm.

FS, IL Encoder with Side Information



$$y_i = f(z_i, x_i, u_i)$$

$$z_{i+1} = g(z_i, x_i, u_i)$$

$$L(y^n) = \sum_{t=1}^{n} l(y_t)$$

Information losslessness = (z_1, y^n, u^n, z_n) uniquely determine x^n.

F–V CGFs

Theorem: For every IL encoder with s states and SI,

$$\frac{\ell}{n}\sum_{t=0}^{n/\ell-1}\exp_2\{\lambda L(y_{t\ell+1}^{t\ell+\ell})\} \geq 2^{-\lambda\gamma(s,\ell)}\sum_{u^\ell}\left\{\sum_{x^\ell}[\hat{P}(x^\ell,u^\ell)]^{1/(1+\lambda)}\right\}^{1+\lambda}.$$

Proof: very similar to the case without SI.

Conditional LZ Parsing [Ziv85]

Given $(\boldsymbol{x}, \boldsymbol{u}) = [(x_1, u_1), \ldots, (x_n, u_n)]$, apply LZ parsing to this sequence pair.

- $c(\boldsymbol{x}, \boldsymbol{u}) =$ number of phrases.
- $c(\boldsymbol{u}) =$ number of distinct phrases of $\boldsymbol{u}$.
- $\boldsymbol{u}(l) =$ the lth distinct $\boldsymbol{u}$–phrase, $l = 1, 2, ..., c(\boldsymbol{u})$.
- $c_l(\boldsymbol{x}|\boldsymbol{u}) =$ number of $\boldsymbol{u}(l)$ in parsing of $\boldsymbol{u}$.

$$\text{conditional compressibility} = \frac{1}{n} \sum_{l=1}^{c(\boldsymbol{u})} c_l(\boldsymbol{x}|\boldsymbol{u}) \log c_l(\boldsymbol{x}|\boldsymbol{u}).$$

For example, $n = 6$ and

$$\begin{aligned} \boldsymbol{x} &= 0 \mid 1 \mid 0\,0 \mid 0\,1| \\ \boldsymbol{u} &= 0 \mid 1 \mid 0\,1 \mid 0\,1| \end{aligned}$$

then

$$c(\boldsymbol{x}, \boldsymbol{u}) = 4, \;\; c(\boldsymbol{u}) = 3, \;\; \boldsymbol{u}(1) = 0, \; \boldsymbol{u}(2) = 1, \; \boldsymbol{u}(3) = 01,$$

$$c_1(\boldsymbol{x}|\boldsymbol{u}) = c_2(\boldsymbol{x}|\boldsymbol{u}) = 1, \;\; c_3(\boldsymbol{x}|\boldsymbol{u}) = 2.$$

A Lower Bound to the V–V CGF

For every IL encoder with s states and SI,

$$\sum_{i=1}^{c} \exp_2\{\lambda L(y_{n_{i-1}+1}^{n_i})\} \geq \frac{s^2}{2^{\lambda+1}-1} \cdot \sum_{k=1}^{c(u^n)} \left(\exp_2 \left\{ (\lambda+1) \log \left[\frac{c_k(x^n|u^n)+s^2}{2s^2} \right] \right\} - 1 \right)$$

On the other hand, the conditional LZ algorithm achieves

$$(2\alpha)^{\lambda} \sum_{k=1}^{c(u^n)} \exp_2\{(\lambda+1) \log c_k(x^n|u^n)\}.$$

Here, in contrast to the case without SI, there is a difference between the best achievable CGF and the compressibility,

$$\frac{1}{n} \sum_{l=1}^{c(\boldsymbol{u})} c_l(\boldsymbol{x}|\boldsymbol{y}) \log c_l(\boldsymbol{x}|\boldsymbol{u}).$$