Data Compression with Nearly Uniform Output

Rémi A. Chou and Matthieu R. Bloch
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332
GT-CNRS UMI 2958, 2 rue Marconi, 57070 Metz, France
e-mail: remi.chou@gatech.edu, matthieu.bloch@ece.gatech.edu

***Abstract*—For any lossless fixed-length compression scheme operating at the optimal coding rate, it is known that the encoder output is not uniform in variational distance, which yet might be desirable in some security schemes. In the case of independent and identically distributed (i.i.d.) sources, uniformity in divergence might be achieved if a uniformly distributed sequence, called seed, of length d_n negligible compared to the message length n, is shared between the encoder and the decoder. We show that the optimal scaling of d_n that jointly ensures an optimal coding rate and a uniform encoder output in divergence, is roughly on the order of $\sqrt{n}$. We also develop a near optimal achievability scheme using invertible extractors.**

I. Introduction

Communication with uniform messages might be desirable for security applications. For instance, in linear network coding [1], in random linear network coding for the α-order criterion [2], or in network coding for multi-resolution video streaming [3], the uniformity of the messages exchanged over the network is a sufficient condition to ensure security.

However, uniform messages are not easily obtained, and in particular, cannot be obtained by a regular compression scheme. By studying the joint code design for the intrinsic randomness problem [4] – which consists in extracting the highest rate of uniform random numbers from a source – and for lossless fixed-length source coding, Han's folklore theorem [5] shows that a source losselessly compressed at the optimal rate becomes uniform in normalized divergence but not necessarily in variational distance. In addition, for i.i.d. sources, Hayashi has shown a fundamental trade-off between error probability and uniformity of the encoder output with respect to (w.r.t.) the variational distance [6]. On this basis, at least three solutions can be adopted when dealing with security problems. One can

- try to derive sufficient conditions independent of the source distribution that ensure security, as done for instance in [1], [7], [8], [9] for network coding;
- study the robustness of the security criterion to non uniform messages, as in [10] for the wiretap channel or in [8] for network coding;
- try to obtain "more uniform" messages by modifying the operation of compression schemes.

In this paper, we do not compare these solutions, and we only focus on the analysis of compression schemes with better uniformity properties. Our objective is, for i.i.d. sources, to reach the optimal lossless source coding rate, while ensuring a uniform encoder output in divergence. To overcome the impossibility of joint code design at the optimal rate for source coding and the intrinsic randomness problem w.r.t. the variational distance [6], we assume that a uniformly distributed sequence, called "seed", is shared at the encoder and the decoder – the seed could for instance be set at the encoder and the decoder by secret key agreement [11], [12]. This setting is obviously only meaningful if the seed length is negligible compared to the message length, since a seed length on the order of the message length, times its entropy is enough to produce a uniform output by a one-time pad.

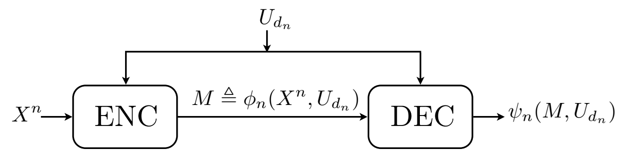

Fig. 1. Encoding/Decoding scheme

Our setting is related to the notion of invertible extractors [13] – which are functions that take as input an arbitrarily distributed sequence S and a uniform seed, to output a nearly uniformly distributed sequence, from which S can be reconstructed; the main difference is that we also require the compression of the source.

The main contributions of this work are

- the characterization of the optimal scaling of the seed length d_n to losslessly compress i.i.d. sources at the optimal rate, while ensuring a uniform encoder output in divergence;
- a near optimal scheme using invertible extractors that separates reliability and uniformity.

The remainder of the paper is organized as follows. In Section II, we formally introduce the problem and state our main result. Sections III and IV respectively deal with the achievability and the converse part of our main result. Section V gives a near optimal scheme that separates reliability and uniformity. Some proofs are omitted due to space limitation.

II. Problem Statement

Let $\mathcal{X}$, $\mathcal{Y}$ be finite alphabets. Let $n \in \mathbb{N}$ and assume X^n is obtained from a discrete memoryless source (DMS) $(\mathcal{X}, p_X)$. Let $d_n \in \mathbb{N}$ and let U_{d_n} be a uniform random variable over $\mathcal{U}_{d_n} \triangleq [\![1, 2^{d_n}]\!]$, independent of X^n, where $[\![p, q]\!]$ is the set of integers between $\lfloor p \rfloor$ and $\lceil q \rceil$, with $p, q \in \mathbb{R}$. In the following, we refer to U_{d_n} as the *seed* and d_n as its length. For $M_n \in \mathbb{N}$, we define $M'_n \triangleq M_n \times 2^{d_n}$ and $\mathcal{M}_n \triangleq [\![1, M_n]\!]$. As illustrated in Figure 1, we consider an encoder $\phi_n : \mathcal{X}^n \times \mathcal{U}_{d_n} \to \mathcal{M}_n$ and a decoder $\psi_n : \mathcal{M}_n \times \mathcal{U}_{d_n} \to \mathcal{X}^n$. We define the decoding error probability, and two different metrics concerning the encoder output uniformity:

$$\begin{aligned}\mathbf{P}_e &\triangleq \mathbb{P}[X^n \neq \psi_n(\phi_n(X^n, U_{d_n}), U_{d_n})],\\ \mathbf{U}_e^{(1)} &\triangleq \mathbb{V}[\phi_n(X^n, U_{d_n}), U_{M_n}],\\ \mathbf{U}_e^{(2)} &\triangleq \mathbb{V}[(\phi_n(X^n, U_{d_n}), U_{d_n}), U_{M_n'}],\end{aligned}$$

where $\mathbb{V}(\cdot,\cdot)$ is the variational distance, U_{M_n} has uniform distribution over $\mathcal{M}_n$. Observe that $\mathbf{U}_e^{(2)}$ is more restrictive than $\mathbf{U}_e^{(1)}$, since it requires the encoder output and the seed to be *jointly* uniformly distributed.

Remark We could also define the potentially stronger metrics

$$\begin{aligned}\mathbf{U}_e^{(1')} &\triangleq \mathbb{D}[\phi_n(X^n, U_{d_n}), U_{M_n}],\\ \mathbf{U}_e^{(2')} &\triangleq \mathbb{D}[(\phi_n(X^n, U_{d_n}), U_{d_n}), U_{M_n'}],\end{aligned}$$

where $\mathbb{D}(\cdot,\cdot)$ is the Kullback-Leibler divergence. However, for $i \in [\![1,2]\!]$, by [14, Lemma 2.7], $\mathbf{U}_e^{(i)}$ can be replaced by $\mathbf{U}_e^{(i')}$, if $\lim_{n\to\infty} n\mathbf{U}_e^{(i)} = 0$, which will be the case.

Definition 1. *A $(2^{nR}, n, 2^{d_n})$ code $\mathcal{C}_n$ for a DMS $(\mathcal{X}, p_X)$ consists of*

- *a message set $\mathcal{M}_n \triangleq [\![1, M_n]\!]$, with $M_n \triangleq 2^{nR}$,*
- *a seed set $\mathcal{U}_{d_n} \triangleq [\![1, 2^{d_n}]\!]$,*
- *an encoder ϕ_n and a decoder ψ_n.*

Definition 2 (Lossless compression with uniform encoder output)**.** *Let $i \in [\![1,2]\!]$. A rate $R \in \mathbb{R}_+$ is achievable for metric i, if there exists a sequence of $(2^{nR}, n, 2^{d_n})$ codes $\{\mathcal{C}_n\}_{n\in\mathbb{N}^*}$ for the source $(\mathcal{X}, p_X)$, such that*

$$\lim_{n\to\infty} \frac{1}{n}\log M_n \leqslant R, \quad \lim_{n\to\infty} \mathbf{P}_e = 0,$$
$$\lim_{n\to\infty} \mathbf{U}_e^{(i)} = 0, \quad \lim_{n\to\infty} \frac{d_n}{n} = 0.$$

Moreover, the infimum of achievable rates is called the capacity and is denoted by C_i.

In the following, we use the Landau notation to characterize the limiting behaviour of the seed scaling, with the convention that for any real functions f and g, $f = \Omega(g)$ means $f = o(g)$ is false. Our main result is presented in the following theorem.

Theorem 1. *Let $(\mathcal{X}, p_X)$ be a DMS. Then,*

(i) *$C_1 = H(X)$. Moreover, for a code length n, the optimal seed scaling of d_n verifies*

$$d_n \in \Omega(n^{1/2}) \cap O(n^{1/2+\epsilon}), \tag{1}$$

where $\epsilon > 0$ is arbitrary.

(ii) *$C_2 = 0$.*

Remark Theorem 1 extends to the case of lossless fixed-length source coding with side information, which can, for instance, find application in the problem described in [15]. Moreover, with a much more involved proof than the one of Proposition 1 relying on a technique developed in [16], the achievability part can be extended to lossy fixed-length source coding, i.e the rate distortion function is achievable with a uniform encoder ouput, if the seed scaling satisfies $d_n \in \omega(n^{1/2+\epsilon})$. We do not prove these extensions due to space constraint.

III. ACHIEVABILITY

Proposition 1 (Achievability)**.** *There exists a sequence of $(2^{nR}, n, 2^{d_n})$ codes $\{\mathcal{C}_n\}_{n\in\mathbb{N}^*}$ such that C_1 is achievable with a seed length d_n scaling as*

$$d_n = \Theta(n^{1/2+\epsilon}),$$

where $\epsilon > 0$ is arbitrary.

Proof: Let $\epsilon_1 > 0$, $\epsilon > 0$, $n \in \mathbb{N}$, $d_n \in \mathbb{N}$, $R > 0$. Define $M_n \triangleq 2^{nR}$ and $\mathcal{M}_n \triangleq [\![1, M_n]\!]$. Consider a random mapping $\Phi : \mathcal{X}^n \times \mathcal{U}_{d_n} \to \mathcal{M}_n$, and its associated decoder $\Psi : \mathcal{M}_n \times \mathcal{U}_{d_n} \to \mathcal{X}^n$. Given $(m, u_{d_n}) \in \mathcal{M}_n \times \mathcal{U}_{d_n}$, the decoder outputs $\hat{x}^n$ if it is the unique sequence such that $\hat{x}^n \in \mathcal{T}_{\epsilon_1}^n(X)$ and $\Phi(\hat{x}^n, u_{d_n}) = m$; otherwise it outputs an error. We let $M \triangleq \Phi(X^n, U_{d_n})$, and define $\mathbf{P}_e \triangleq \mathbb{P}(X^n \neq \Psi(\Phi(X^n, U_{d_n}), U_{d_n})$, $\mathbf{U}_e \triangleq \mathbb{V}(M, \mathcal{U}_{M_n})$.

- We first determine a condition over R to ensure $\mathbb{E}_\Phi\left[\mathbf{U}_e^{(1)}\right] \leqslant \epsilon$. Remark that

$$\forall m \in \mathcal{M}_n, p_M(m) = \sum_{x^n}\sum_u p(x^n, u)\mathbb{1}\{\Phi(x^n, u) = m\},$$

hence, on average $\forall m \in \mathcal{M}_n$, $\mathbb{E}_\Phi[p_M(m)] = 2^{-nR}$, which allows us to write

$$\begin{aligned}\mathbb{E}_\Phi\left[\mathbf{U}_e^{(1)}\right] &= \mathbb{E}_\Phi\left[\sum_m |p_M(m) - \mathbb{E}_\Phi[p_M(m)]|\right]\\ &\leqslant \sum_{i=1}^2 \mathbb{E}_\Phi\left[\sum_m \left|p_M^{(i)}(m) - \mathbb{E}_\Phi\left[p_M^{(i)}(m)\right]\right|\right],\end{aligned} \tag{2}$$

where $\forall m \in \mathcal{M}_n$, $\forall i \in [\![1,2]\!]$,

$$p_M^{(i)}(m) = \sum_{x^n\in\mathcal{A}_i}\sum_u p(x^n, u)\mathbb{1}\{\Phi(x^n, u) = m\},$$

with $\mathcal{A}_1 \triangleq \mathcal{T}_{\epsilon_1}^n(X)$ and $\mathcal{A}_2 \triangleq \mathcal{A}_1^c$. After some manipulations we bound the second term in (2) as follows

$$\mathbb{E}_\Phi\left[\sum_m \left|p_M^{(2)}(m) - \mathbb{E}_\Phi\left[p_M^{(2)}(m)\right]\right|\right] \leqslant 4|\mathcal{X}|e^{-n\epsilon_1^2\mu_X}, \tag{3}$$

with $\mu_X = \min_{x\in\text{supp}(P_X)} P_X(x)$. Then, we bound the first term in (2) by Jensen's inequality

$$\begin{aligned}&\mathbb{E}_\Phi\left[\sum_m \left|p_M^{(1)}(m) - \mathbb{E}_\Phi\left[p_M^{(1)}(m)\right]\right|\right]\\ &\qquad\leqslant \sum_m \sqrt{\text{Var}_\Phi\left(p_M^{(1)}(m)\right)}.\end{aligned} \tag{4}$$

Moreover, after some manipulations, we obtain

$$\text{Var}_\Phi\left(p_M^{(1)}(m)\right) \leqslant \exp_2\left[-n(1-3\epsilon_1)H(X)\right]2^{-d_n}2^{-nR}. \tag{5}$$

Thus, by combining (4) and (5), we obtain

$$\begin{aligned}
&\mathbb{E}_\Phi\left[\sum_m \left|p_M^{(1)}(m) - \mathbb{E}_\Phi\left[p_M^{(1)}(m)\right]\right|\right] \\
&\leqslant \sum_m \sqrt{\exp_2\left[-n(1-3\epsilon_1)H(X)\right]2^{-d_n}2^{-nR}} \qquad (6)\\
&= \sqrt{M_n}\exp_2\left[-\frac{n}{2}\left((1-3\epsilon_1)H(X)+\frac{d_n}{n}\right)\right]\\
&\leqslant \exp_2\left[\frac{n}{2}\left(R-(1-3\epsilon_1)H(X)-\frac{d_n}{n}\right)\right]. \qquad (7)
\end{aligned}$$

Hence, if $R < H(X) + \frac{d_n}{n} - 3\epsilon_1 H(X)$, then asymptotically $\mathbb{E}_\Phi\left[\mathbf{U}_e^{(1)}\right] \leqslant \epsilon$ by (3) and (7).

- We now derive a condition over R to ensure $\mathbb{E}_\Phi[\mathbf{P}_e] \leqslant \epsilon$. We define $\mathcal{E}_0 \triangleq \{X^n \notin \mathcal{T}_{\epsilon_1}^n(X)\}$, and $\mathcal{E}_1 \triangleq \{\exists \hat{x}^n \neq X^n, \Phi(\hat{x}^n, U) = \Phi(X^n, U)$ and $\hat{x}^n \in \mathcal{T}_{\epsilon_1}^n(X)\}$ so that by the union bound, $\mathbb{E}_\Phi[\mathbf{P}_e] \leqslant \mathbb{P}[\mathcal{E}_0] + \mathbb{P}[\mathcal{E}_1]$. We have

$$\mathbb{P}[\mathcal{E}_0] \leqslant 2|\mathcal{X}|e^{-n\epsilon_1^2\mu_X}, \qquad (8)$$

and defining $\mathbf{P}(x^n, \hat{x}^n, u) \triangleq \mathbb{P}[\exists \hat{x}^n \neq x^n, \Phi(\hat{x}^n, u) = \Phi(x^n, u)$ and $\hat{x}^n \in \mathcal{T}_{\epsilon_1}^n(X)]$, we have

$$\begin{aligned}
\mathbb{P}[\mathcal{E}_1] &= \sum_{x^n}\sum_u p(x^n,u)\mathbf{P}(x^n,\hat{x}^n,u)\\
&\leqslant \sum_{x^n}\sum_u p(x^n,u)\sum_{\substack{\hat{x}^n\in\mathcal{T}_{\epsilon_1}^n(X)\\ \hat{x}^n\neq x^n}}\mathbb{P}[\Phi(\hat{x}^n,u)=\Phi(x^n,u)]\\
&= \sum_{x^n}\sum_u p(x^n,u)\sum_{\substack{\hat{x}^n\in\mathcal{T}_{\epsilon_1}^n(X)\\ \hat{x}^n\neq x^n}}2^{-nR}\\
&\leqslant \sum_{x^n}\sum_u p(x^n,u)|\mathcal{T}_{\epsilon_1}^n(X)|2^{-nR}\\
&\leqslant \sum_{x^n}\sum_u p(x^n,u)\exp_2\left[nH(X)(1+\epsilon_1)\right]2^{-nR}\\
&\leqslant \exp_2\left[n(H(X)(1+\epsilon_1)-R)\right]. \qquad (9)
\end{aligned}$$

Hence, if $R > H(X) + \epsilon_1 H(X)$, then asymptotically $\mathbb{E}_\Phi(\mathbf{P}_e) \leqslant \epsilon$ by (8) and (9).

All in all, if R is such that

$$H(X) + \epsilon_1 H(X) < R < H(X) + \frac{d_n}{n} - 3\epsilon_1 H(X),$$

then asymptotically by the selection lemma, $\mathbb{E}_\Phi[\mathbf{U}_e^{(1)}] \leqslant \epsilon$ and $\mathbb{E}_\Phi[\mathbf{P}_e] \leqslant \epsilon$. Thus, we choose d_n such that

$$4n\epsilon_1 H(X) < d_n \leqslant 4n\epsilon_1 H(X) + 1,$$

to obtain

$$H(X) + \epsilon_1 H(X) < H(X) + \frac{d_n}{n} - 3\epsilon_1 H(X).$$

We can also choose $\epsilon_1 = n^{-1/2+\epsilon_b}$, with any $\epsilon_b > 0$,[1] so that for any $\epsilon_a > \epsilon_b$

$$4n^{\epsilon_b-\epsilon_a}H(X) < \frac{d_n}{n^{1/2+\epsilon_a}} \leqslant 4n^{\epsilon_b-\epsilon_a}H(X) + n^{-1/2-\epsilon_a},$$

[1]Note that we cannot make ϵ_1 decrease faster because of Equations (3) and (8).

which means $d_n = o(n^{1/2+\epsilon_a})$. Finally, by means of the selection lemma applied to $\mathbf{P}_e$ and $\mathbf{U}_e$, there exists a realization of Φ such that $\mathbf{U}_e^{(1)} \leqslant \epsilon$ and $\mathbf{P}_e \leqslant \epsilon$. ■

Remark In Proposition 1, the same results holds if U_{d_n}, i.e. the seed, is not truly uniform but satisfies instead

$$\mathbb{V}(U_{d_n}, \mathcal{U}_{d_n}) \leqslant \exp\left[n^{1/2+\epsilon_0} - n^{1/2+\epsilon}\right],$$

where ϵ is such that $d_n = \Theta(n^{1/2+\epsilon})$ and $\epsilon_0 \in]0, \epsilon[$ is arbitrary.

IV. Converse

It can be shown without difficulty that any achievable rate R must satisfy $R \geqslant H(X)$ for the metric $\mathbf{U}_e^{(1)}$, hence it remains to show an upper bound for the optimal scaling of d_n. It is done by means of a second order asymptotics study, with which we also show that $C_2 = 0$.

In this section, we consider an arbitrary source $\mathbf{X} \triangleq \{X^n\}_{n=1}^\infty$, where X^n is a random variable taking values in $\mathcal{X}^n$ subject to P_{X^n}. Specifically, we generalize some results of [6] to our setup, and show that if $d_n = o(\sqrt{n})$, with n the code length, then the trade-off between error probability and uniformity of [6] cannot be improved.

For the fixed-length source coding problem, for $\epsilon > 0$, for $\mathbf{d} \triangleq \{d_n\}_n \in \mathbb{R}_+^{\mathbb{N}}$ and for a code $\mathcal{C}_n \triangleq (\phi_n, \psi_n, \mathcal{M}_n)$, we define the following first order asymptotics

$$a_0 \triangleq R(\mathbf{d},\epsilon|\mathbf{X}) \triangleq \inf_{\{\mathcal{C}_n\}}\left\{\overline{\lim}\left[\frac{1}{n}\log M_n\right]: \overline{\lim}\ \mathbf{P}_e < \epsilon\right\},$$

$$a_0^+ \triangleq R_+(\mathbf{d},\epsilon|\mathbf{X}) \triangleq \inf_{\{\mathcal{C}_n\}}\left\{\underline{\lim}\left[\frac{1}{n}\log M_n\right]: \overline{\lim}\ \mathbf{P}_e < \epsilon\right\},$$

as well as the following second order asymptotics

$$R(\mathbf{d},\epsilon,a_0\,|\mathbf{X}) \triangleq \inf_{\{\mathcal{C}_n\}}\left\{\overline{\lim}\left[\frac{1}{\sqrt{n}}\log\frac{M_n}{e^{na_0}}\right]: \overline{\lim}\ \mathbf{P}_e < \epsilon\right\},$$

$$R_+(\mathbf{d},\epsilon,a_0^+|\mathbf{X}) \triangleq \inf_{\{\mathcal{C}_n\}}\left\{\underline{\lim}\left[\frac{1}{\sqrt{n}}\log\frac{M_n}{e^{na_0^+}}\right]: \overline{\lim}\ \mathbf{P}_e < \epsilon\right\}.$$

For the intrinsic randomness problem, for $\epsilon > 0$, for $\mathbf{d} \in \mathbb{R}_+^{\mathbb{N}}$, for $i \in [\![1, 2]\!]$ and for a code $\mathcal{C}'_n \triangleq (\phi_n, \mathcal{M}_n)$, we define the following first order asymptotics

$$a_i \triangleq S^{(i)}(\mathbf{d},\epsilon|\mathbf{X}) \triangleq \sup_{\{\mathcal{C}'_n\}}\left\{\underline{\lim}\left[\frac{1}{n}\log M_n\right]: \overline{\lim}\ \mathbf{U}_e^{(i)} < \epsilon\right\},$$

$$a_i^- \triangleq S_-^{(i)}(\mathbf{d},\epsilon|\mathbf{X}) \triangleq \sup_{\{\mathcal{C}'_n\}}\left\{\overline{\lim}\left[\frac{1}{n}\log M_n\right]: \overline{\lim}\ \mathbf{U}_e^{(i)} < \epsilon\right\},$$

as well as the following second order asymptotics

$$S^{(i)}(\mathbf{d},\epsilon,a_i|\mathbf{X}) \triangleq \sup_{\{\mathcal{C}'_n\}}\left\{\underline{\lim}\left[\frac{1}{\sqrt{n}}\log\frac{M_n}{e^{na_i}}\right]: \overline{\lim}\ \mathbf{U}_e^{(i)} < \epsilon\right\},$$

$$S_-^{(i)}(\mathbf{d},\epsilon,a_i^-|\mathbf{X}) \triangleq \sup_{\{\mathcal{C}'_n\}}\left\{\overline{\lim}\left[\frac{1}{\sqrt{n}}\log\frac{M_n}{e^{na_i^-}}\right]: \overline{\lim}\ \mathbf{U}_e^{(i)} < \epsilon\right\}.$$

We express the first order and the second order asymptotics, defined above, in the following lemmas – we omit the proof for brevity.

Lemma 1. *Let $\epsilon > 0$. Let $\mathbf{d} \in \mathbb{R}_+^{\mathbb{N}}$. The first order asymptotics have the following expression*

$$\begin{aligned} R(\mathbf{d},\epsilon|\mathbf{X}) &= \overline{H}(\mathbf{0},1-\epsilon|\mathbf{X}),\\ R_+(\mathbf{d},\epsilon|\mathbf{X}) &= \underline{H}(\mathbf{0},1-\epsilon|\mathbf{X}),\\ S^{(1)}(\mathbf{d},\epsilon|\mathbf{X}) &= \underline{H}(\mathbf{d},\epsilon|\mathbf{X}),\\ S_-^{(1)}(\mathbf{d},\epsilon|\mathbf{X}) &= \overline{H}(\mathbf{d},\epsilon|\mathbf{X}),\\ S^{(2)}(\mathbf{d},\epsilon|\mathbf{X}) &= \underline{H}(\mathbf{0},\epsilon|\mathbf{X}),\\ S_-^{(2)}(\mathbf{d},\epsilon|\mathbf{X}) &= \overline{H}(\mathbf{0},\epsilon|\mathbf{X}), \end{aligned}$$

where,

$$\underline{H}(\mathbf{d},\epsilon|\mathbf{X}) \triangleq \inf_x \left\{ x : \overline{\lim}\ \mathbb{P}\left[\frac{1}{n}\log\frac{1}{P_{X^n}(X^n)} < x - \frac{d_n}{n}\right] \geqslant \epsilon \right\},$$
$$\overline{H}(\mathbf{d},\epsilon|\mathbf{X}) \triangleq \inf_x \left\{ x : \underline{\lim}\ \mathbb{P}\left[\frac{1}{n}\log\frac{1}{P_{X^n}(X^n)} < x - \frac{d_n}{n}\right] \geqslant \epsilon \right\}.$$

Lemma 2. *Let $\epsilon > 0$. Let $\mathbf{d} \in \mathbb{R}_+^{\mathbb{N}}$. The second order asymptotics have the following expression*

$$\begin{aligned} R(\mathbf{d},\epsilon,a_0|\mathbf{X}) &= \overline{H}(\mathbf{0},1-\epsilon,a_0|\mathbf{X}),\\ R_+(\mathbf{d},\epsilon,a_0^+|\mathbf{X}) &= \underline{H}(\mathbf{0},1-\epsilon,a_0^+|\mathbf{X}),\\ S^{(1)}(\mathbf{d},\epsilon,a_1|\mathbf{X}) &= \underline{H}(\mathbf{d},\epsilon,a_1|\mathbf{X}),\\ S_-^{(1)}(\mathbf{d},\epsilon,a_1^-|\mathbf{X}) &= \overline{H}(\mathbf{d},\epsilon,a_1^-|\mathbf{X}),\\ S^{(2)}(\mathbf{d},\epsilon,a_2|\mathbf{X}) &= \underline{H}(\mathbf{0},\epsilon,a_2|\mathbf{X}),\\ S_-^{(2)}(\mathbf{d},\epsilon,a_2^-|\mathbf{X}) &= \overline{H}(\mathbf{0},\epsilon,a_2^-|\mathbf{X}), \end{aligned}$$

where,

$$\underline{H}(\mathbf{d},\epsilon,a|\mathbf{X}) \triangleq \inf_x \left\{ x : \overline{\lim}\ \mathbb{P}\left[\frac{1}{n}\log\frac{1}{P_{X^n}(X^n)} < a + \frac{x}{\sqrt{n}} - \frac{d_n}{n}\right] \geqslant \epsilon \right\},$$
$$\overline{H}(\mathbf{d},\epsilon,a|\mathbf{X}) \triangleq \inf_x \left\{ x : \underline{\lim}\ \mathbb{P}\left[\frac{1}{n}\log\frac{1}{P_{X^n}(X^n)} < a + \frac{x}{\sqrt{n}} - \frac{d_n}{n}\right] \geqslant \epsilon \right\}.$$

From the first order and the second order asymptotics derived in Lemma 1 and Lemma 2, we study the trade-off between $\mathbf{P}_e$ and $\mathbf{U}_e^{(i)}$, $i \in [\![1,2]\!]$ for i.i.d. sources following the same method as in [6]. We consider the intrinsic randomness problem for the code $\mathcal{C}'_n = (\phi_n, \mathcal{M}_n)$ and the fixed-length source coding for the code $\mathcal{C}_n = (\phi_n, \psi_n, \mathcal{M}_n)$. For $i \in [\![1,2]\!]$, we want to know whether there exists a sequence of triplet $\{(\phi_n, \psi_n, \mathcal{M}_n)\}_{n\in\mathbb{N}}$ such that $\overline{\lim}\ \mathbf{P}_e = \epsilon$ and $\overline{\lim}\ \mathbf{U}_e^{(i)} = \epsilon'$, where $\epsilon, \epsilon' \in]0,1[$ can be chosen arbitrarily small, while ensuring d_n negligible compared to n. We first simplify the first order asymptotics of Lemma 1, when $d_n = o(n)$.

Lemma 3. *Let $\mathbf{d} \in \mathbb{R}_+^{\mathbb{N}}$. Assume i.i.d. sources and assume $d_n = o(n)$. Then, $\overline{H}(\mathbf{0},\epsilon|\mathbf{X})$, $\underline{H}(\mathbf{0},\epsilon|\mathbf{X})$, $\underline{H}(\mathbf{d},\epsilon|\mathbf{X})$, $\overline{H}(\mathbf{d},\epsilon|\mathbf{X})$, $\underline{H}(\mathbf{0},\epsilon|\mathbf{X})$, $\overline{H}(\mathbf{0},\epsilon|\mathbf{X})$ are all equal to $H(X)$.*

Proposition 2 (Converse)**.** *Let $\mathbf{d} \in \mathbb{R}_+^{\mathbb{N}}$. Assume i.i.d. sources.*

(i) If $d_n = o(n)$, then

$$\overline{\lim}\ \mathbf{P}_e + \overline{\lim}\ \mathbf{U}_e^{(2)} \geqslant 1,$$

which implies $C_2 = 0$.

(ii) If $d_n = o(\sqrt{n})$, then

$$\overline{\lim}\ \mathbf{P}_e + \overline{\lim}\ \mathbf{U}_e^{(1)} \geqslant 1.$$

Proof: We prove the two statements in order.

(i) Note that, for i.i.d. sources, by Lemma 1 and Lemma 3, all the first asymptotics considered are equal, hence by definition of the second order asymptotics, the following must hold

$$S_-^{(i)}(\mathbf{d},\epsilon',a|\mathbf{X}) \geqslant \overline{\lim}\left[\frac{1}{\sqrt{n}}\log\frac{M_n}{e^{na}}\right] \geqslant R(\mathbf{d},\epsilon,a|\mathbf{X}), \tag{10}$$

$$S^{(i)}(\mathbf{d},\epsilon',a|\mathbf{X}) \geqslant \underline{\lim}\left[\frac{1}{\sqrt{n}}\log\frac{M_n}{e^{na}}\right] \geqslant R_+(\mathbf{d},\epsilon,a|\mathbf{X}). \tag{11}$$

Then, for $i = 2$, (10) and (11) together with Lemma 2 give

$$\begin{aligned} \overline{H}(\mathbf{0},\epsilon',a|\mathbf{X}) &\geqslant \overline{H}(\mathbf{0},1-\epsilon,a|\mathbf{X}),\\ \underline{H}(\mathbf{0},\epsilon',a|\mathbf{X}) &\geqslant \underline{H}(\mathbf{0},1-\epsilon,a|\mathbf{X}). \end{aligned}$$

Thus, for i.i.d. sources, since $\overline{H}(\mathbf{0},\epsilon,a|\mathbf{X})$ and $\underline{H}(\mathbf{0},\epsilon,a|\mathbf{X})$ are continuous and increasing w.r.t. ϵ, we find that

$$\overline{\lim}\ \mathbf{P}_e + \overline{\lim}\ \mathbf{U}_e^{(2)} \geqslant 1.$$

(ii) For $i = 1$, we assume $d_n = o(\sqrt{n})$. By Equations (10), (11), we have by Lemma 2

$$\overline{H}(\mathbf{d},\epsilon',a|\mathbf{X}) \geqslant \overline{H}(\mathbf{0},1-\epsilon,a|\mathbf{X}), \tag{12}$$
$$\underline{H}(\mathbf{d},\epsilon',a|\mathbf{X}) \geqslant \underline{H}(\mathbf{0},1-\epsilon,a|\mathbf{X}). \tag{13}$$

Remark that for any $\epsilon_0 > 0$, since $d_n = o(\sqrt{n})$, we have

$$\begin{aligned} &\overline{\lim}\ \mathbb{P}\left[\frac{1}{n}\log\frac{1}{P_{X^n}(X^n)} < a + \frac{b - d_n/\sqrt{n}}{\sqrt{n}}\right]\\ &\geqslant \overline{\lim}\ \mathbb{P}\left[\frac{1}{n}\log\frac{1}{P_{X^n}(X^n)} < a + \frac{b-\epsilon_0}{\sqrt{n}}\right], \end{aligned}$$

hence,

$$\begin{aligned} &\underline{H}(\mathbf{d},\epsilon',a|\mathbf{X})\\ &= \inf_b \left\{ b : \overline{\lim}\ \mathbb{P}\left[\frac{1}{n}\log\frac{1}{P_{X^n}(X^n)} < a + \frac{b - d_n/\sqrt{n}}{\sqrt{n}}\right] \geqslant \epsilon \right\}\\ &\leqslant \inf_b \left\{ b : \overline{\lim}\ \mathbb{P}\left[\frac{1}{n}\log\frac{1}{P_{X^n}(X^n)} < a + \frac{b-\epsilon_0}{\sqrt{n}}\right] \geqslant \epsilon \right\}\\ &= \epsilon_0 + \inf_b \left\{ b : \overline{\lim}\ \mathbb{P}\left[\frac{1}{n}\log\frac{1}{P_{X^n}(X^n)} < a + \frac{b}{\sqrt{n}}\right] \geqslant \epsilon \right\}\\ &= \epsilon_0 + \underline{H}(\mathbf{0},\epsilon',a|\mathbf{X}), \end{aligned}$$

and similarly

$$\overline{H}(\mathbf{d},\epsilon',a|\mathbf{X}) \leqslant \epsilon_0 + \overline{H}(\mathbf{0},\epsilon',a|\mathbf{X}).$$

Thus, by (12), (13), we have

$$\begin{aligned} \epsilon_0 + \overline{H}(\mathbf{0},\epsilon',a|\mathbf{X}) \geqslant \overline{H}(\mathbf{d},\epsilon',a|\mathbf{X}) \geqslant \overline{H}(\mathbf{0},1-\epsilon,a|\mathbf{X}),\\ \epsilon_0 + \underline{H}(\mathbf{0},\epsilon',a|\mathbf{X}) \geqslant \underline{H}(\mathbf{d},\epsilon',a|\mathbf{X}) \geqslant \underline{H}(\mathbf{0},1-\epsilon,a|\mathbf{X}), \end{aligned}$$

which means

$$\begin{aligned} \overline{H}(\mathbf{0},\epsilon',a|\mathbf{X}) &\geqslant \overline{H}(\mathbf{0},1-\epsilon,a|\mathbf{X}),\\ \underline{H}(\mathbf{0},\epsilon',a|\mathbf{X}) &\geqslant \underline{H}(\mathbf{0},1-\epsilon,a|\mathbf{X}), \end{aligned}$$

since ϵ_0 is arbitrary. Consequently, if $d_n = o(\sqrt{n})$, we conclude as in (i), to show that

$$\overline{\lim}\ \mathbf{P}_e + \overline{\lim}\ \mathbf{U}_e^{(1)} \geqslant 1.$$

■

Disappointedly, for $i = 2$, we find a result similar to the one in [6] when there is no additional randomness available at the encoder and the decoder, that is, we cannot compress a source at the optimal rate and ensure $\overline{\lim}\ \mathbf{U}_e^{(2)} = 0$.

On the other hand, for $i = 1$, Proposition 2 shows that $d_n = \Omega(n^{1/2})$ is a sufficient condition on the seed scaling, to losslessly compress a source at optimal rate and ensure $\overline{\lim}\ \mathbf{U}_e^{(1)} = 0$. The characterization of the optimal seed scaling, given by Propositions 1 and 2, can be seen as a way to quantify the gap between uniformity in normalized divergence and uniformity in variational distance in Han's folklore theorem [5] for the i.i.d. case.

V. A SCHEME WITH INVERTIBLE EXTRACTORS

In this section, we consider the situation in which $(\mathcal{X}, p_X)$ is a binary memoryless source. We propose a schemes that achieve C_1 with a near optimal seed rate. The scheme involves invertible extractors and separates reliability and uniformity.

Definition 3 ([17]). *Let $\epsilon > 0$. Let $m, d, l \in \mathbb{N}$ and let $t \in \mathbb{R}^+$. A polynomial time probabilistic function* Ext : $\{0,1\}^m \times \{0,1\}^d \mapsto \{0,1\}^l$ *is called a (m, d, l, t, ϵ)-extractor, if for all binary source X satisfying $\mathbb{H}_\infty(X) \geqslant t$, we have*

$$\mathbb{V}(\text{Ext}(X, U_d), U_l) \leqslant \epsilon,$$

where U_d is a sequence of d uniformly distributed bits, U_l has uniform distribution over $\{0,1\}^l$. Moreover, a (m, d, l, t, ϵ)-extractor is said invertible if the input can be reconstructed from the output and U_d.

Theorem 2 ([17],[13]). *Let $\epsilon > 0$. Let $m, d \in \mathbb{N}$ and $t \in \mathbb{R}^+$. There exists an invertible (m, d, m, t, ϵ)-extractor such that*

$$d = m - t + 2\log m + 2\log\frac{1}{\epsilon} + O(1). \tag{14}$$

Remark In Theorem 2, the extractor is explicitly constructed with a δ-biased set (see [18] for construction) used as a generator to construct an invertibly labelled (natural labelling is invertible) Cayley graph, see [13] for details. In (14), the term "$2\log m$" can be removed by using Ramanujan expander graphs, however, it increases the construction complexity [17].

Proposition 3. *For any $\epsilon > 0$, there exists $\mathbf{d} \in \mathbb{R}_+^{\mathbb{N}}$, satisfying $d_n = \Theta(n^{1/2+\epsilon})$, such that if U_{d_n} is shared by the emitter and the receiver, then there exists a sequence of $(2^{nR}, n, 2^{d_n})$ codes achieving C_1 in the sense of Definition 2, where the encoder ϕ_n and the decoder ψ_n are made of the composition of a typical sequence based compression scheme and an invertible extractor, as described in Figure 2.*

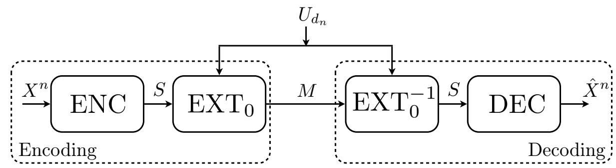

Fig. 2. Encoding/Decoding scheme

Note that, the scheme described in Proposition 3 suffers from a lack of practicality as far as the typical sequence based compression part is concerned. Whether it can be replaced by a practical compression scheme remains an open question.

Remark A more practical scheme achieving C_1, which jointly deals with reliability and uniformity, although operating with a non-optimal seed scaling $d_n = o(n)$, can be obtained with polar codes following the proof of [19, Theorem 1].

ACKNOWLEDGEMENTS

Work supported through CNRS PEPS project OPTO-ALEA.

REFERENCES

[1] N. Cai and T. Chan, "Theory of Secure Network Coding," *Proceedings of the IEEE*, vol. 99, no. 3, pp. 421–437, 2011.

[2] L. Lima, "Network Coding Security Algebraic Properties and Lightweight Solutions," Ph.D. dissertation, Faculdade de Ciências da Universidade do Porto, 2010.

[3] L. Lima, S. Gheorghiu, J. Barros, M. Médard, and A. Toledo, "Secure Network Coding for Multi-Resolution Wireless Video Streaming," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 3, pp. 377–388, 2010.

[4] T. S. Han, *Information-Spectrum Methods in Information Theory*. Springer, 2002, vol. 50.

[5] ——, "Folklore in Source Coding: Information-Spectrum Approach," *IEEE Trans. Inf. Theory*, vol. 51, no. 2, pp. 747–753, 2005.

[6] M. Hayashi, "Second-Order Asymptotics in Fixed-Length Source Coding and Intrinsic Randomness," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4619–4637, 2008.

[7] N. Cai and R. Yeung, "A Security Condition for Multi-Source Linear Network Coding," in *Proc. IEEE Int. Symp. Inf. Theory*, 2007, pp. 561–565.

[8] R. Matsumoto and M. Hayashi, "Universal Strongly Secure Network Coding with Dependent and Non-Uniform Messages," *arXiv preprint arXiv:1111.4174*, 2011.

[9] M. Hayashi and R. Matsumoto, "Secure Multiplex Coding with Dependent and Non-Uniform Multiple Messages," *arXiv preprint arXiv:1202.1332*, 2012.

[10] M. Bloch and J. Kliewer, "On Secure Communication with Constrained Randomization," in *Proc. IEEE Int. Symp. Inf. Theory*, 2012, pp. 1172–1176.

[11] U. Maurer, "Secret Key Agreement by Public Discussion from Common Information," *IEEE Trans. Inf. Theory*, vol. 39, pp. 733–742, 1993.

[12] R. Ahlswede and I. Csiszár, "Common Randomness in Information Theory and Cryptography Part I: Secret Sharing," *IEEE Trans. Inf. Theory*, vol. 39, pp. 1121–1132, 1993.

[13] Y. Dodis and A. Smith, "Entropic Security and the Encryption of High-Entropy Messages," in *Theory of Cryptography*, 2005.

[14] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge Univ Pr, 1981.

[15] S. Watanabe, R. Matsumoto, and T. Uyematsu, "Strongly Secure Privacy Amplification Cannot Be Obtained by Encoder of Slepian-Wolf Code," *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 93, no. 9, pp. 1650–1659, 2010.

[16] M. Yassaee, M. Aref, and A. Gohari, "Achievability Proof via Output Statistics of Random Binning," *Arxiv preprint arXiv:1203.0730*, 2012.

[17] Y. Dodis, "On Extractors, Error-Correction and Hiding All Partial Information," in *Proc. IEEE Inf. Theory Workshop*, 2005.

[18] N. Alon, O. Goldreich, J. Håstad, and R. Peralta, "Simple Constructions of Almost k-Wise Independent Random Variables," in *Proc. of the 31st lEEE Symp. on Foundations of Computer Science*, 1990.

[19] R. Chou, M. Bloch, and E. Abbe, "Polar Coding for Secret-Key Generation," *submitted to IEEE Inf. Theory Workshop*, 2013.