Non-asymptotic Equipartition Properties for Independent and Identically Distributed Sources

En-Hui Yang and Jin Meng
Department of Electrical and
Computer Engineering
University of Waterloo
Waterloo, Ontario N2L 3G1
Email: {ehyang, j4meng}@uwaterloo.ca

***Abstract*—Given an independent and identically distributed source $X = \{X_i\}_{i=1}^{\infty}$ with finite Shannon entropy or differential entropy (as the case may be) $H(X)$, the non-asymptotic equipartition property (NEP) with respect to $H(X)$ is established, which characterizes, for any finite block length n, how close $-\frac{1}{n}\ln p(X_1X_2\cdots X_n)$ is to $H(X)$ by determining the information spectrum of $X_1X_2\cdots X_n$, i.e., the distribution of $-\frac{1}{n}\ln p(X_1X_2\cdots X_n)$. Non-asymptotic equipartition properties (with respect to conditional entropy, mutual information, and relative entropy) in a similar nature can also be established [3]. These non-asymptotic equipartition properties are instrumental to the development of non-asymptotic coding (including both source and channel coding) results in information theory in the same way as the asymptotic equipartition property to all asymptotic coding theorems established so far in information theory. As an example, the NEP with respect to $H(X)$ is used to establish a non-asymptotic fixed rate source coding theorem, which reveals, for any finite block length n, a complete picture about the tradeoff between the minimum rate of fixed rate coding of $X_1\cdots X_n$ and error probability when the error probability is a constant, or goes to 0 with block length n at a sub-polynomial $n^{-\alpha}$, $0 < \alpha < 1$, polynomial $n^{-\alpha}$, $\alpha \geq 1$, or sub-exponential $e^{-n^{\alpha}}$, $0 < \alpha < 1$, speed. In particular, it is shown that for any finite block length n, the minimum rate (in nats per symbol) of fixed rate coding of $X_1X_2\cdots X_n$ with error probability $\Theta\left(\frac{n^{-\alpha}}{\sqrt{\ln n}}\right)$ is $H(X)+\sqrt{\sigma_H^2(X)(2\alpha)}\sqrt{\frac{\ln n}{n}}+O(\frac{\ln n}{n})$, where $\alpha > 0$ and $\sigma_H^2(X) = \mathbf{E}[-\ln p(X_1)]^2 - H^2(X)$ is the information variance of X. With the help of the NEP with respect to other information quantities, non-asymptotic channel coding theorems of similar nature will be established in a separate paper.**

***Index Terms*—Asymptotic equipartition property (AEP), conditional entropy, entropy, fixed rate coding, information spectrum, mutual information, non-asymptotic equipartition property (NEP).**

I. Introduction

Consider an independent and identically distributed (IID) source $X = \{X_i\}_{i=1}^{\infty}$ with source alphabet $\mathcal{X}$ and finite entropy $H(X)$, where $H(X)$ is the Shannon entropy of X_i if $\mathcal{X}$ is discrete, and the differential entropy of X_i if $\mathcal{X}$ is the real line and each X_i is a continuous random variable. Let $p(x)$ be the probability mass function (pmf) or probability density function (pdf) (as the case may be) of X_i. The asymptotic equipartition property (AEP) for X is the assertion that

$$-\frac{1}{n}\ln p(X_1X_2\cdots X_n) \to H(X) \tag{1.1}$$

either in probability or with probability one as n goes to ∞. It implies that for sufficiently large n, with high probability, the outcomes of $X_1X_2\cdots X_n$ are approximately equiprobable with their respective probability ranging from $e^{-n(H(X)+\epsilon)}$ to $e^{-n(H(X)-\epsilon)}$, where $\epsilon > 0$ is a small fixed number. Here and throughout the rest of the paper, $\ln$ stands for the logarithm with base e, and all information quantities are measured in nats.

The AEP is fundamental to information theory. It is not only instrumental to lossless source coding theorems, but also behind almost all asymptotic coding (including source, channel, and multi-user coding) theorems through the concepts of typical sets and typical sequences [1].

However, in the non-asymptotic regime where one wants to establish non-asymptotic coding results for finite block length n, the AEP in its current form can not be applied in general. In this paper, we aim to establish the non-asymptotic counterpart of the AEP, which is broadly referred to as the non-asymptotic equipartition property (NEP), so that the NEP can be applied to finite block length n. Specifically, with respect to $H(X)$, we first characterize, for any finite block length n, how close $-\frac{1}{n}\ln p(X_1X_2\cdots X_n)$ is to $H(X)$ by determining the information spectrum of $X_1X_2\cdots X_n$, i.e., the distribution of $-\frac{1}{n}\ln p(X_1X_2\cdots X_n)$; such a property is referred to as the NEP with respect to $H(X)$. NEP can be also established in a similar manner with respect to conditional entropy, mutual information, and relative entropy; for details, please refer to the full version of this paper [3].

In the same way as the AEP plays an important role in establishing the asymptotic coding (including source, channel, and multi-user coding) results in information theory, our established NEP is also instrumental to the development of non-asymptotic source and channel coding results. Using the NEP with respect to $H(X)$, we further establish a non-asymptotic fixed rate source coding theorem, which reveals, for any finite block length n, a complete picture about the tradeoff between the minimum rate of fixed rate coding of $X_1\cdots X_n$ and error probability when the error probability is a

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada under Grant RGPIN203035-11, and by the Canada Research Chairs Program.

constant, or goes to 0 with block length n at a sub-polynomial $n^{-\alpha}$, $0 < \alpha < 1$, polynomial $n^{-\alpha}$, $\alpha \geq 1$, or sub-exponential $e^{-n^{\alpha}}$, $0 < \alpha < 1$, speed. In particular, it is shown that for any finite block length n, the minimum rate (in nats per symbol) of fixed rate coding of $X_1 X_2 \cdots X_n$ with error probability $\Theta\left(\frac{n^{-\alpha}}{\sqrt{\ln n}}\right)$ is $H(X) + \sqrt{\sigma_H^2(X)(2\alpha)}\sqrt{\frac{\ln n}{n}} + O(\frac{\ln n}{n})$, where $\alpha > 0$ and $\sigma_H^2(X) = \mathbf{E}[-\ln p(X_1)]^2 - H^2(X)$ is the information variance of X. In a separate paper [4], non-asymptotic channel coding theorems of similar nature will be established with the help of the NEP with respect to other information quantities; in particular, it is shown [4] that for any binary input memoryless channel with uniform capacity achieving input X, random linear codes of block length n can reach within $\sqrt{\sigma_H^2(X|Y)(2\alpha)}\sqrt{\frac{\ln n}{n}} + O(\frac{\ln n}{n})$ of the channel capacity while maintaining word error probability $O(n^{-\alpha})$, where $\alpha > 0$ and $\sigma_H^2(X|Y) = \mathbf{E}[-\log p(X|Y)]^2 - H^2(X|Y)$ is the conditional information variance of X given Y with Y being the output of the channel in response to the input X.

The rest of the paper is organized as follows. Section II is devoted to the NEP with respect to $H(X)$. And in Section III, we apply the NEP with respect to $H(X)$ to investigate the performance of optimal fixed rate coding of $X_1 X_2 \cdots X_n$.

II. NEP With Respect to Entropy

Define

$$\lambda^*(X) \stackrel{\Delta}{=} \sup\left\{\lambda \geq 0 : \int p^{-\lambda+1}(x)dx < \infty\right\} \tag{2.1}$$

where $\int dx$ is understood throughout this paper to be the summation over the source alphabet of X if X is discrete. Suppose that

$$\lambda^*(X) > 0 \,. \tag{2.2}$$

Let

$$\sigma_H^2(X) \stackrel{\Delta}{=} \int p(x)[-\ln p(x)]^2 dx - H^2(X) \tag{2.3}$$

which will be referred to as the information variance of X. It is not hard to see that under the assumption (2.2),

$$\int \frac{p^{-\lambda+1}(x)}{\left[\int p^{-\lambda+1}(y)dy\right]} \left|-\ln p(x)\right|^k dx < \infty \tag{2.4}$$

and

$$\int p^{-\lambda+1}(x)dx < \infty$$

for any $\lambda \in [0, \lambda^*(X))$ and any positive integer k, and hence $\sigma_H^2(X)$ is finite in particular. Further assume that

$$\sigma_H^2(X) > 0 \,. \tag{2.5}$$

Then we have the following result, which will be referred to as the weak right NEP with respect to $H(X)$.

Theorem 1 (Weak Right NEP). *For any $\delta \geq 0$, let*

$$r_X(\delta) \stackrel{\Delta}{=} \sup_{\lambda \geq 0}\left[\lambda(H(X)+\delta) - \ln\int p^{-\lambda+1}(x)dx\right] .$$

Then the following hold:

(a) *For any positive integer n,*

$$\Pr\left\{-\frac{1}{n}\ln p(X^n) \geq H(X) + \delta\right\} \leq e^{-n r_X(\delta)} \tag{2.6}$$

where $X^n = X_1 X_2 \cdots X_n$.

(b) *Under the assumptions* (2.2) *and* (2.5)*, there exists a $\delta^* > 0$ such that for any $\delta \in (0, \delta^*]$ and any positive integer n,*

$$r_X(\delta) = \frac{1}{2\sigma_H^2(X)}\delta^2 + O(\delta^3) \tag{2.7}$$

and hence

$$\Pr\left\{-\frac{1}{n}\ln p(X^n) \geq H(X) + \delta\right\} \leq e^{-n(\frac{\delta^2}{2\sigma_H^2(X)} + O(\delta^3))} . \tag{2.8}$$

Proof of Theorem 1: The inequality (2.6) follows from the Chernoff bound. To see this is indeed the case, note that

$$\begin{aligned} & \Pr\left\{-\frac{1}{n}\ln p(X_1X_2\cdots X_n) \geq H(X)+\delta\right\} \\ = \; & \Pr\left\{-\ln p(X_1X_2\cdots X_n) \geq n(H(X)+\delta)\right\} \\ \leq \; & \inf_{\lambda\geq 0} \frac{\mathbf{E}[e^{-\lambda \ln p(X_1X_2\cdots X_n)}]}{e^{n\lambda(H(X)+\delta)}} \\ = \; & \inf_{\lambda\geq 0} e^{-n\left[\lambda(H(X)+\delta) - \ln \mathbf{E}[p^{-\lambda}(X_1)]\right]} \\ = \; & \inf_{\lambda\geq 0} e^{-n\left[\lambda(H(X)+\delta) - \ln\int p^{-\lambda+1}(x)dx\right]} \\ = \; & e^{-n r_X(\delta)} . \end{aligned} \tag{2.9}$$

To show (2.7) and (2.8), we first analyze the property of $r_X(\delta)$ as a function of δ over the region $\delta \geq 0$. It is easy to see that $r_X(\delta)$ is convex and non-decreasing. For any $\lambda \in [0, \lambda^*(X))$, define

$$\delta(\lambda) \stackrel{\Delta}{=} \int \frac{p^{-\lambda+1}(x)}{\left[\int p^{-\lambda+1}(y)dy\right]}\left[-\ln p(x)\right]dx - H(X) \tag{2.10}$$

which, in view of (2.4), is well defined. Using a similar argument as in [5, Properties 1 to 3], it is not hard to show that under the assumption (2.2), $\delta(\lambda)$ as a function of λ is continuously differentiable up to any order over $\lambda \in [0, \lambda^*(X))$. Taking the first order derivative of $\delta(\lambda)$ yields

$$\begin{aligned} \delta'(\lambda) \; = \; & \int \frac{p^{-\lambda+1}(x)}{\left[\int p^{-\lambda+1}(y)dy\right]}\left[-\ln p(x)\right]^2 dx \\ & - \left[\int \frac{p^{-\lambda+1}(x)}{\left[\int p^{-\lambda+1}(y)dy\right]}\left[-\ln p(x)\right]dx\right]^2 \\ > \; & 0 \end{aligned} \tag{2.11}$$

where the last inequality is due to (2.5). It is also easy to see that $\delta(0) = 0$ and $\delta'(0) = \sigma_H^2(X)$. Therefore, $\delta(\lambda)$ is strictly increasing over $\lambda \in [0, \lambda^*(X))$. On the other hand, it is not hard to verify that under the assumption (2.2), the function $\lambda(H(X)+\delta) - \ln\int p^{-\lambda+1}(x)dx$ as a function of

λ is continuously differentiable over $\lambda \in [0, \lambda^*(X))$ with its derivative equal to

$$\delta - \delta(\lambda) . \tag{2.12}$$

To continue, we distinguish between two cases: (1) $\lambda^*(X) = \infty$, and (2) $\lambda^*(X) < \infty$. In case (1), since $\delta(\lambda)$ is strictly increasing over $\lambda \in [0, \infty)$, it follows that for any $\delta = \delta(\lambda)$ for some $\lambda \in [0, \lambda^*(X))$, the supremum in the definition of $r_X(\delta)$ is actually achieved at that particular λ, i.e.,

$$r_X(\delta(\lambda)) = \lambda(H(X) + \delta(\lambda)) - \ln \int p^{-\lambda+1}(x)dx . \tag{2.13}$$

In case (2), we have that for any $\delta = \delta(\lambda)$ for some $\lambda \in [0, \lambda^*(X))$,

$$\beta(H(X) + \delta(\lambda)) - \ln \int p^{-\beta+1}(x)dx < \lambda(H(X) + \delta(\lambda)) - \ln \int p^{-\lambda+1}(x)dx \tag{2.14}$$

for any $\beta \in [0, \lambda^*(X))$ with $\beta \neq \lambda$. In view of the definition of $\lambda^*(X)$, (2.14) remains valid for any $\beta > \lambda^*(X)$ since then the left side of (2.14) is $-\infty$. What remains to check is when $\beta = \lambda^*(X)$. If

$$\int p^{-\lambda^*(X)+1}(x)dx = \infty$$

it is easy to see that (2.14) holds as well when $\beta = \lambda^*(X)$. Suppose now

$$\int p^{-\lambda^*(X)+1}(x)dx < \infty .$$

In this case, it follows from the dominated convergence theorem that

$$\lim_{\beta \uparrow \lambda^*(X)} \int p^{-\beta+1}(x)dx = \int p^{-\lambda^*(X)+1}(x)dx$$

and hence by letting β go to $\lambda^*(X)$ from the left, we see that (2.14) holds as well when $\beta = \lambda^*(X)$. Putting all cases together, we always have that for any $\delta = \delta(\lambda)$ for some $\lambda \in [0, \lambda^*(X))$,

$$r_X(\delta(\lambda)) = \lambda(H(X) + \delta(\lambda)) - \ln \int p^{-\lambda+1}(x)dx . \tag{2.15}$$

Let

$$\Delta^*(X) \triangleq \lim_{\lambda \uparrow \lambda^*(X)} \delta(\lambda) .$$

Since both $\delta(\lambda)$ and $\ln \int p^{-\lambda+1}(x)dx$ are continuously differentiable with respect to $\lambda \in [0, \lambda^*(X))$ up to any order, it follows from (2.15) that $r_X(\delta)$ is also continuously differentiable with respect to $\delta \in [0, \Delta^*(X))$ up to any order. Taking the first and second order derivatives of $r_X(\delta)$ with respect to δ, we have

$$\begin{aligned} r'_X(\delta) &= \frac{dr_X(\delta)}{d\delta} \\ &= \frac{dr_X(\delta(\lambda))}{d\lambda} \frac{d\lambda}{d\delta} \\ &= \frac{dr_X(\delta(\lambda))}{d\lambda} \frac{1}{\delta'(\lambda)} \\ &= \frac{1}{\delta'(\lambda)} \Bigg[H(X) + \delta(\lambda) + \lambda\delta'(\lambda) \\ &\quad - \int \frac{p^{-\lambda+1}(x)}{[\int p^{-\lambda+1}(y)dy]} [-\ln p(x)]\, dx \Bigg] \\ &= \lambda \end{aligned} \tag{2.16}$$

and

$$\begin{aligned} r''_X(\delta) &= \frac{d\lambda}{d\delta} \\ &= \frac{1}{\delta'(\lambda)} \end{aligned} \tag{2.17}$$

where $\delta = \delta(\lambda)$. Therefore, $r_X(\delta)$ is convex, strictly increasing, and continuously differentiable up to any order over $\delta \in [0, \Delta^*(X))$. Note that from (2.16) and (2.17), we have $r'_X(0) = 0$ and $r''_X(0) = 1/\sigma^2_H(X)$. Expanding $r_X(\delta)$ at $\delta = 0$ by the Taylor expansion, we then have that there exists a $\delta^* > 0$ such that

$$r_X(\delta) = \frac{1}{2\sigma^2_H(X)}\delta^2 + O(\delta^3) \tag{2.18}$$

for $\delta \in (0, \delta^*]$. The inequality (2.8) now follows immediately from (2.6) and (2.18). This completes the proof of Theorem 1. ■

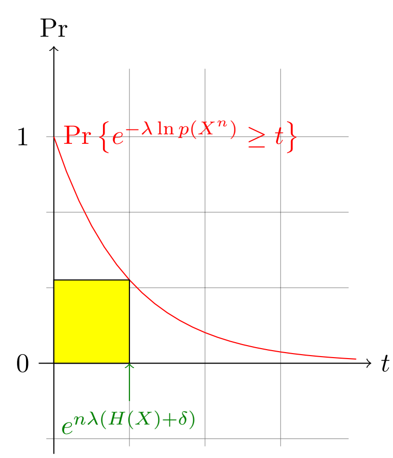


Fig. 1. Graphical interpretation of the weak right NEP

Graphical interpretation of the weak right NEP: Figure 1 provides a graphical interpretation of the weak right NEP, where

$$\mathbf{E}[e^{-\lambda \ln p(X^n)}] = \underbrace{\int \Pr\left\{e^{-\lambda \ln p(X^n)} \geq t\right\} dt}_{\text{area underneath the red curve}}$$

$$\geq \underbrace{e^{n\lambda(H(X)+\delta)} \Pr\left\{-\ln p(X^n) \geq n(H(X)+\delta)\right\}}_{\text{area of the yellow rectangle}}$$

which immediately gives us (2.6) by further optimizing λ.

Having analyzed the function $r_X(\delta)$, we are now ready for a stronger version of the right NEP. For any $\lambda \in [0, \lambda^*(X))$, define

$$f_\lambda(x) \triangleq \frac{p^{-\lambda}(x)}{\int p^{-\lambda+1}(y)dy} \tag{2.19}$$

$$\sigma_H^2(X,\lambda) \triangleq \int f_\lambda(x)p(x) \left|-\ln p(x) - (H(X)+\delta(\lambda))\right|^2 dx \tag{2.20}$$

$$M_H(X,\lambda) \triangleq \int f_\lambda(x)p(x) \left|-\ln p(x) - (H(X)+\delta(\lambda))\right|^3 dx \tag{2.21}$$

and

$$f_\lambda(x^n) \triangleq \prod_{i=1}^n f_\lambda(x_i) \tag{2.22}$$

where $\delta(\lambda)$ is defined in (2.10). Write $M_H(X,0)$ as $M_H(X)$. It is easy to see that $\sigma_H^2(X,0) = \sigma_H^2(X)$, $\sigma_H^2(X,\lambda) = \delta'(\lambda)$, and

$$M_H(X) = \int p(x) \left|-\ln p(x) - H(X)\right|^3 dx\ . \tag{2.23}$$

Then we have the following stronger result.

Theorem 2 (Strong Right NEP). *Under the assumptions* (2.2) *and* (2.5)*, the following hold:*

(a) *For any $\delta \in (0, \Delta^*(X))$ and any positive integer n*

$$\begin{aligned}\Pr&\left\{-\frac{1}{n}\ln p(X^n) \geq H(X)+\delta\right\} \\ &\leq \frac{1}{1-e^{-\lambda}}\left[\frac{1}{\sqrt{2\pi}\sigma_H(X,\lambda)} + \frac{2CM_H(X,\lambda)}{\sigma_H^3(X,\lambda)}\right] \\ &\quad \times e^{-nr_X(\delta)-\frac{1}{2}\ln n}\end{aligned} \tag{2.24}$$

and

$$\begin{aligned}\Pr&\left\{-\frac{1}{n}\ln p(X^n) \geq H(X)+\delta\right\} \\ &\geq e^{-\lambda d}\left[\frac{de^{-\frac{d^2}{2n\sigma_H^2(X,\lambda)}}}{\sqrt{2\pi}\sigma_H(X,\lambda)} - \frac{2CM_H(X,\lambda)}{\sigma_H^3(X,\lambda)}\right] \\ &\quad \times e^{-nr_X(\delta)-\frac{1}{2}\ln n}\end{aligned} \tag{2.25}$$

for any $d > 0$, where $\lambda = r'_X(\delta) > 0$, and $C < 1$ is the universal constant in the central limit theorem of Berry and Esseen.

(b) *For any $\delta \leq c\sqrt{\frac{\ln n}{n}}$, where $c < \sigma_H(X)$ is a constant,*

$$\begin{aligned}&Q\left(\frac{\delta\sqrt{n}}{\sigma_H(X)}\right) - \frac{CM_H(X)}{\sqrt{n}\sigma_H^3(X)} \\ &\leq \Pr\left\{-\frac{1}{n}\ln p(X^n) \geq H(X)+\delta\right\} \\ &\leq Q\left(\frac{\delta\sqrt{n}}{\sigma_H(X)}\right) + \frac{CM_H(X)}{\sqrt{n}\sigma_H^3(X)}\end{aligned} \tag{2.26}$$

where $Q(t) = \frac{1}{\sqrt{2\pi}}\int_t^\infty e^{-u^2/2}du$.

Proof of Theorem 2: From (2.15), it follows that with $\lambda = r'_X(\delta)$

$$r_X(\delta) = \lambda(H(X)+\delta) - \ln\int p^{-\lambda+1}(x)dx\ . \tag{2.27}$$

Define for any integer $k \geq 0$

$$B_k \triangleq \left\{x^n : \frac{k}{n} \leq -\frac{1}{n}\ln p(x^n) - (H(X)+\delta) < \frac{k+1}{n}\right\}.$$

Then it is not hard to verify that

$$\begin{aligned}&\Pr\left\{-\frac{1}{n}\ln p(X^n) \geq H(X)+\delta\right\} \\ &= \int_{-\frac{1}{n}\ln p(x^n) \geq H(X)+\delta} p(x^n)dx^n \\ &= \sum_{k=0}^\infty \int_{x^n \in B_k} p(x^n)dx^n \\ &= \sum_{k=0}^\infty \int_{x^n \in B_k} f_\lambda^{-1}(x^n)f_\lambda(x^n)p(x^n)dx^n \\ &= \sum_{k=0}^\infty \int_{x^n \in B_k} f_\lambda(x^n)p(x^n) \\ &\quad \times \exp\left\{-n\left[-\frac{1}{n}\lambda\ln p(x^n) - \ln\int p^{-\lambda+1}(y)dy\right]\right\}dx^n \\ &\leq \sum_{k=0}^\infty \int_{x^n \in B_k} f_\lambda(x^n)p(x^n)dx^n \\ &\quad \times \exp\left\{-n\left[\lambda(H(X)+\delta+\frac{k}{n}) - \ln\int p^{-\lambda+1}(y)dy\right]\right\} \\ &= e^{-nr_X(\delta)}\sum_{k=0}^\infty e^{-\lambda k}\int_{x^n \in B_k} f_\lambda(x^n)p(x^n)dx^n\end{aligned} \tag{2.28}$$

where the last equality is due to (2.27). At this point, we invoke the following central limit theorem of Berry and Esseen [2, Theorem 1.2].

Lemma 1. *Let $V_1, V_2, \cdots$ be independent real random variables with zero means and finite third moments, and set*

$$\sigma_n^2 = \sum_{i=1}^n \mathbf{E}V_i^2.$$

Then there exists a universal constant $C < 1$ such that for any $n \geq 1$,

$$\sup_{-\infty<t<+\infty} |\Pr\{\sum_{i=1}^n V_i \leq \sigma_n t\} - \Phi(t)| \leq C\sigma_n^{-3} \sum_{i=1}^n \mathbf{E}|V_i|^3,$$

where $\Phi(t) = (2\pi)^{-1/2} \int_{-\infty}^t e^{-u^2/2} du$.

Note that with $\lambda = r'_X(\delta)$, we have $\delta = \delta(\lambda)$ and hence

$$\int f_\lambda(x)p(x)[-\ln p(x)]dx = H(X) + \delta \ .$$

Consider now an IID random variables $Z_1, Z_2, \cdots, Z_n$ with pmf or pdf $f_\lambda(z)p(z)$ (as the case may be). Applying Lemma 1 to the IID sequence $\{-\ln p(Z_i) - (H(X)+\delta)\}_{i=1}^n$, we then have

$$\begin{aligned} &\int\limits_{x^n \in B_k} f_\lambda(x^n)p(x^n)dx^n \\ &\leq \frac{1}{\sqrt{2\pi}} \int_0^{\frac{1}{\sqrt{n}\sigma_H(X,\lambda)}} e^{-\frac{t^2}{2}} dt + 2C\frac{1}{\sqrt{n}} \frac{M_H(X,\lambda)}{\sigma_H^3(X,\lambda)} \\ &\leq \frac{1}{\sqrt{n}} \left(\frac{1}{\sqrt{2\pi}\sigma_H(X,\lambda)} + \frac{2CM_H(X,\lambda)}{\sigma_H^3(X,\lambda)} \right) \end{aligned} \tag{2.29}$$

for any $k \geq 0$. Combining (2.29) with (2.28) yields

$$\begin{aligned} &\Pr\left\{-\frac{1}{n}\ln p(X^n) \geq H(X)+\delta\right\} \\ &\leq \left(\frac{1}{\sqrt{2\pi}\sigma_H(X,\lambda)} + \frac{2CM_H(X,\lambda)}{\sigma_H^3(X,\lambda)} \right) \sum_{k=0}^\infty e^{-\lambda k} \\ &\quad \times e^{-nr_X(\delta) - \frac{1}{2}\ln n} \\ &= \frac{1}{1-e^{-\lambda}} \left(\frac{1}{\sqrt{2\pi}\sigma_H(X,\lambda)} + \frac{2CM_H(X,\lambda)}{\sigma_H^3(X,\lambda)} \right) \\ &\quad \times e^{-nr_X(\delta) - \frac{1}{2}\ln n} \ . \end{aligned} \tag{2.30}$$

This completes the proof of (2.24).

To prove (2.25), note that for any $d > 0$

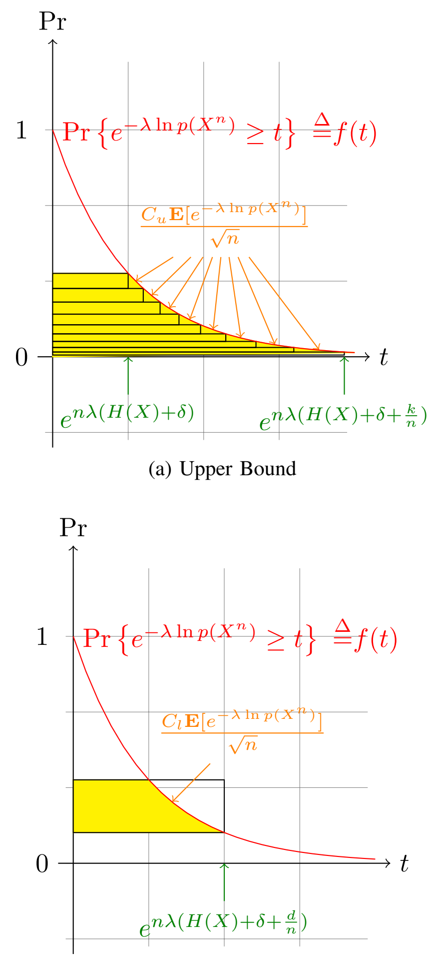
$$\begin{aligned} &\Pr\left\{-\frac{1}{n}\ln p(X^n) \geq H(X)+\delta\right\} \\ &\geq \int\limits_{0 \leq -\frac{1}{n}\ln p(x^n) - (H(X)+\delta) < \frac{d}{n}} p(x^n)dx^n \\ &= \int\limits_{0 \leq -\frac{1}{n}\ln p(x^n) - (H(X)+\delta) < \frac{d}{n}} f_\lambda^{-1}(x^n) f_\lambda(x^n)p(x^n)dx^n \\ &\geq e^{-nr_X(\delta) - \lambda d} \\ &\quad \times \int\limits_{0 \leq -\frac{1}{n}\ln p(x^n) - (H(X)+\delta) < \frac{d}{n}} f_\lambda(x^n)p(x^n)dx^n \ . \end{aligned} \tag{2.31}$$

Applying Lemma 1 to the IID sequence $\{-\ln p(Z_i) - (H(X)+\delta)\}_{i=1}^n$ again, we have

$$\begin{aligned} &\int\limits_{0 \leq -\frac{1}{n}\ln p(x^n) - (H(X)+\delta) < \frac{d}{n}} f_\lambda(x^n)p(x^n)dx^n \\ &\geq \frac{1}{\sqrt{2\pi}} \int_0^{\frac{d}{\sqrt{n}\sigma_H(X,\lambda)}} e^{-\frac{t^2}{2}} dt - 2C\frac{1}{\sqrt{n}} \frac{M_H(X,\lambda)}{\sigma_H^3(X,\lambda)} \\ &\geq \frac{1}{\sqrt{n}} \left(\frac{d}{\sqrt{2\pi}\sigma_H(X,\lambda)} e^{-\frac{d^2}{2n\sigma_H^2(X,\lambda)}} - \frac{2CM_H(X,\lambda)}{\sigma_H^3(X,\lambda)} \right) \end{aligned} \tag{2.32}$$

which, combined with (2.31), implies (2.25). This completes the proof Part (a) of Theorem 2.

Applying Lemma 1 directly to the IID sequence $\{-\ln p(X_i) - H(X)\}_{i=1}^n$, we get (2.26). This completes the proof of Theorem 2. ■



(a) Upper Bound

(b) Lower Bound

Fig. 2. Graphical interpretation of the strong right NEP

Graphical interpretation of the strong right NEP: Figure 2 provides a graphical interpretation of the upper and lower bounds in the strong right NEP. For the upper bound (2.24), as illustrated in Figure 2(a), we have

$$\begin{aligned}
&\Pr\left\{e^{-\lambda \ln p(X^n)} \geq e^{n\lambda(H(X)+\delta)}\right\} \\
=& \sum_{k=0}^{\infty}\Big(\Pr\left\{e^{-\lambda \ln p(X^n)} \geq e^{n\lambda(H(X)+\delta+\frac{k+1}{n})}\right\} \\
&- \Pr\left\{e^{-\lambda \ln p(X^n)} \geq e^{n\lambda(H(X)+\delta+\frac{k}{n})}\right\}\Big) \\
\triangleq& \sum_{k=1}^{\infty} h_k
\end{aligned}$$

and

$$\begin{aligned}
&\underbrace{h_k e^{n\lambda(H(X)+\delta+\frac{k}{n})}}_{\text{area of the } k\text{th rectangle}} \\
\leq& \underbrace{\int_{e^{n\lambda(H(X)+\delta+\frac{k+1}{n})}}^{e^{n\lambda(H(X)+\delta+\frac{k}{n})}} t df(t)}_{\text{area of the } k\text{th slice under the curve}} \\
=& \int_{e^{n\lambda(H(X)+\delta+\frac{k+1}{n})}}^{e^{n\lambda(H(X)+\delta+\frac{k}{n})}} t\left(-\int_{x^n: e^{-\lambda \ln p(x^n)}=t} p(x^n)dx^n\right)dt \\
=& \int_{e^{n\lambda(H(X)+\delta+\frac{k}{n})}}^{e^{n\lambda(H(X)+\delta+\frac{k+1}{n})}}\left(\int_{x^n: p^{-\lambda}(x^n)=t} p^{-\lambda}(x^n)p(x^n)dx^n\right)dt \\
=& \int_{x^n\in B_k} p^{-\lambda}(x^n)p(x^n)dx^n \\
=& \mathbf{E}[e^{-\lambda \ln p(X^n)}]\int_{x^n\in B_k} f_\lambda(x^n)p(x^n)dx^n \\
\leq& \frac{C_u \mathbf{E}[e^{-\lambda \ln p(X^n)}]}{\sqrt{n}}
\end{aligned}$$

where $C_u = \left(\frac{1}{\sqrt{2\pi}\sigma_H(X,\lambda)} + \frac{2CM_H(X,\lambda)}{\sigma_H^3(X,\lambda)}\right)$. In a similar manner, for the lower bound (2.25), as illustrated in Figure 2(b), we have

$$\begin{aligned}
&\Pr\left\{e^{-\lambda \ln p(X^n)} \geq e^{n\lambda(H(X)+\delta)}\right\} \\
\geq& \Pr\left\{e^{-\lambda \ln p(X^n)} \geq e^{n\lambda(H(X)+\delta)}\right\} \\
&- \Pr\left\{e^{-\lambda \ln p(X^n)} \geq e^{n\lambda(H(X)+\delta+\frac{d}{n})}\right\} \\
\triangleq& h_d
\end{aligned}$$

and

$$\begin{aligned}
&\underbrace{h_d e^{n\lambda(H(X)+\delta+\frac{d}{n})}}_{\text{area of the rectangle}} \\
\geq& \underbrace{\int_{e^{n\lambda(H(X)+\delta+\frac{d}{n})}}^{e^{n\lambda(H(X)+\delta)}} t df(t)}_{\text{area of the slice marked in yellow}} \\
=& \int_{0\leq -\frac{1}{n}\ln p(x^n)-(H(X)+\delta)<\frac{d}{n}} p^{-\lambda}(x^n)p(x^n)dx^n \\
=& \mathbf{E}[e^{-\lambda \ln p(X^n)}] \\
&\times \int_{0\leq -\frac{1}{n}\ln p(x^n)-(H(X)+\delta)<\frac{d}{n}} f_\lambda(x^n)p(x^n)dx^n \\
\geq& \frac{C_l \mathbf{E}[e^{-\lambda \ln p(X^n)}]}{\sqrt{n}}
\end{aligned}$$

where $C_l = \left(\frac{d}{\sqrt{2\pi}\sigma_H(X,\lambda)} e^{-\frac{d^2}{2n\sigma_H^2(X,\lambda)}} - \frac{2CM_H(X,\lambda)}{\sigma_H^3(X,\lambda)}\right)$.

Remark 1. *In view of the proof of Theorem 2, it is easy to see that the upper bound in* (2.24) *can be replaced with the following improved version:*

$$\begin{aligned}
&\left[\sum_{k=0}^{\infty} \frac{e^{-\lambda k}}{\sqrt{2\pi}\sigma_H(X,\lambda)} e^{-\frac{k^2}{2n\sigma_H^2(X,\lambda)}} \right. \\
&\left. + \frac{1}{1-e^{-\lambda}} \frac{2CM_H(X,\lambda)}{\sigma_H^3(X,\lambda)}\right] \\
&\times e^{-nr_X(\delta)-\frac{1}{2}\ln n}
\end{aligned} \tag{2.33}$$

The probability that $-\frac{1}{n}\ln p(X^n)$ is away from $H(X)$ to the left can be bounded similarly. Define

$$\lambda_-^*(X) \triangleq \sup\left\{\lambda \geq 0 : \int p^{\lambda+1}(x)dx < \infty\right\} . \tag{2.34}$$

Suppose that

$$\lambda_-^*(X) > 0 . \tag{2.35}$$

Define for any $\delta \geq 0$

$$r_{X,-}(\delta) \triangleq \sup_{\lambda\geq 0}\left[\lambda(\delta - H(X)) - \ln\int p^{\lambda+1}(x)dx\right]$$

and for any $\lambda \in [0, \lambda_-^*(X))$

$$\delta_-(\lambda) \triangleq \int \frac{p^{\lambda+1}(x)}{\left[\int p^{\lambda+1}(y)dy\right]}[\ln p(x)]\, dx + H(X) .$$

Then under the assumption (2.5), $\delta_-(\lambda)$ is strictly increasing over $\lambda \in [0, \lambda_-^*(X))$ with $\delta_-(0) = 0$. Let

$$\Delta_-^*(X) = \lim_{\lambda\uparrow\lambda_-^*(X)} \delta_-(\lambda) .$$

Following the proof of Theorem 1, we have that $r_{X,-}(\delta)$ is strictly increasing, convex, and continuously differentiable up to any order over $\delta \in [0, \Delta_-^*(X))$, and furthermore

$$r_{X,-}(\delta) = \lambda(\delta - H(X)) - \ln\int p^{\lambda+1}(x)dx$$

with $\lambda = r'_{X,-}(\delta)$ satisfying

$$\delta_-(\lambda) = \delta .$$

Define

$$\sigma^2_{H,-}(X,\lambda) \triangleq \int \frac{p^{\lambda+1}(x)}{\left[\int p^{\lambda+1}(y)dy\right]} \left|-\ln p(x) - (H(X) - \delta_-(\lambda))\right|^2 dx$$

and

$$M_{H,-}(X,\lambda) \triangleq \int \frac{p^{\lambda+1}(x)}{\left[\int p^{\lambda+1}(y)dy\right]} \left|-\ln p(x) - (H(X) - \delta_-(\lambda))\right|^3 dx \,.$$

In parallel with Theorems 1 and 2, we have the following result, which is referred to as the left NEP with respect to $H(X)$ and can be proved similarly.

Theorem 3 (Left NEP). *For any positive integer n,*

$$\Pr\left\{-\frac{1}{n}\ln p(X^n) \le H(X) - \delta\right\} \le e^{-nr_{X,-}(\delta)} \,. \tag{2.36}$$

Furthermore, under the assumptions (2.35) *and* (2.5)*, the following also hold:*

(a) *There exists a $\delta^* > 0$ such that for any $\delta \in (0, \delta^*]$ and any positive integer n,*

$$r_{X,-}(\delta) = \frac{1}{2\sigma^2_H(X)}\delta^2 + O(\delta^3) \tag{2.37}$$

and hence

$$\Pr\left\{-\frac{1}{n}\ln p(X^n) \le H(X) - \delta\right\} \le e^{-n\left(\frac{\delta^2}{2\sigma^2_H(X)} + O(\delta^3)\right)} \,. \tag{2.38}$$

(b) *For any $\delta \in (0, \Delta^*_-(X))$ and any positive integer n*

$$\Pr\left\{-\frac{1}{n}\ln p(X^n) \le H(X) - \delta\right\} \le \frac{1}{1-e^{-\lambda}}\left[\frac{1}{\sqrt{2\pi}\sigma_{H,-}(X,\lambda)} + \frac{2CM_{H,-}(X,\lambda)}{\sigma^3_{H,-}(X,\lambda)}\right] \times e^{-nr_{X,-}(\delta) - \frac{1}{2}\ln n} \tag{2.39}$$

and

$$\Pr\left\{-\frac{1}{n}\ln p(X^n) \le H(X) - \delta\right\} \ge e^{-\lambda d}\left[\frac{de^{-\frac{d^2}{2n\sigma^2_{H,-}(X,\lambda)}}}{\sqrt{2\pi}\sigma_{H,-}(X,\lambda)} - \frac{2CM_{H,-}(X,\lambda)}{\sigma^3_{H,-}(X,\lambda)}\right] \times e^{-nr_{X,-}(\delta) - \frac{1}{2}\ln n} \tag{2.40}$$

for any $d > 0$, where $\lambda = r'_{X,-}(\delta) > 0$.

(c) *For any $\delta \le c\sqrt{\frac{\ln n}{n}}$, where $c < \sigma_H(X)$ is a constant,*

$$Q\left(\frac{\delta\sqrt{n}}{\sigma_H(X)}\right) - \frac{CM_H(X)}{\sqrt{n}\sigma^3_H(X)} \le \Pr\left\{-\frac{1}{n}\ln p(X^n) \le H(X) - \delta\right\} \le Q\left(\frac{\delta\sqrt{n}}{\sigma_H(X)}\right) + \frac{CM_H(X)}{\sqrt{n}\sigma^3_H(X)} \,. \tag{2.41}$$

Remark 2. *In parallel with* (2.33)*, the upper bound in* (2.39) *can be replaced with the following improved version:*

$$\left[\sum_{k=0}^{\infty}\frac{e^{-\lambda k}}{\sqrt{2\pi}\sigma_{H,-}(X,\lambda)}e^{-\frac{k^2}{2n\sigma^2_{H,-}(X,\lambda)}} + \frac{1}{1-e^{-\lambda}}\frac{2CM_{H,-}(X,\lambda)}{\sigma^3_{H,-}(X,\lambda)}\right] \times e^{-nr_{X,-}(\delta) - \frac{1}{2}\ln n} \tag{2.42}$$

III. NEP Application to Fixed Rate Source Coding

Assume that the source alphabet $\mathcal{X}$ is finite. In this section, we make use of the NEP with respect to $H(X)$ to establish a non-asymptotic fixed rate source coding theorem, which reveals, for any finite block length n, a complete picture about the tradeoff between the minimum rate of fixed rate coding of $X_1 \cdots X_n$ and error probability when the error probability is a constant, or goes to 0 with block length n at a sub-polynomial $n^{-\alpha}$, $0 < \alpha < 1$, polynomial $n^{-\alpha}$, $\alpha \ge 1$, or sub-exponential $e^{-n^{\alpha}}$, $0 < \alpha < 1$, speed. We begin with the definition of fixed rate source code.

Definition 1. *Given a source from alphabet $\mathcal{X}$, a fixed rate source code with coding length n is defined as a mapping $i : S_n \to \{1, 2, \ldots, |S_n|\}$, where S_n is a subset of $\mathcal{X}^n$. The performance of the code is measured by the rate $R_n = \frac{1}{n}\ln|S_n|$ (in nats) and error probability $\epsilon_n = \Pr\{X^n \notin S_n\}$.*

As can be seen from the definition, the design of a fixed rate source code is equivalent to picking a subset of $\mathcal{X}^n$. Given the source statistics $p(x)$, one can easily show that the optimal way to pick S_n is to order x^n in the non-increasing order of $p(x^n)$, and include those x^n with rank less than or equal to $|S_n|$. Then we have the following non-asymptotic fixed rate source coding theorem, the proof of which can be found in [3].

Theorem 4. *Let R_n denote the minimum rate (in nats) of fixed rate coding of $X_1X_2\cdots X_n$ subject to the error probability ϵ_n. Under the assumptions* (2.2) *and* (2.5)*, we have for any n,*

$$\delta \ge R_n - H(X) \ge \delta - r_X(\delta) - \frac{\ln n}{2n} - O(n^{-1}) \tag{3.1}$$

whenever

$$\left|\frac{\ln\epsilon_n}{n} + r_X(\delta) + \frac{\ln n}{2n} + \frac{\ln\lambda}{n}\right| \le O(n^{-1}) \tag{3.2}$$

for $\Omega\left(\frac{1}{\sqrt{n}}\right) = \delta \le \ln|\mathcal{X}| - H(X)$, *where* $\lambda = r'_X(\delta)$. *In particular, the following hold, depending on whether* ϵ_n *is a constant, or how fast* ϵ_n *goes to* 0.

(a) *Let* δ *be a constant with respect to* n. *Then*

$$\begin{aligned} & r_X^{(inv)}\left(-\frac{\ln \epsilon_n}{n} - \frac{\ln n}{2n}\right) + O(n^{-1}) \\ \ge\ & R_n - H(X) \\ \ge\ & r_X^{(inv)}\left(-\frac{\ln \epsilon_n}{n} - \frac{\ln n}{2n}\right) + \frac{\ln \epsilon_n}{n} - O(n^{-1}) \end{aligned} \tag{3.3}$$

whenever ϵ_n *decreases exponentially with respect to* n, *where* $r_X^{(inv)}(\cdot)$ *is the inverse function of* $r_X(\cdot)$.

(b) *Let* $\delta = \sigma_H(X)\sqrt{\frac{2\alpha \ln n}{n}}$ *for some* $\alpha > 0$. *Then*

$$\begin{aligned} & \sigma_H(X)\sqrt{\frac{2\alpha \ln n}{n}} \\ \ge\ & R_n - H(X) \\ \ge\ & \sigma_H(X)\sqrt{\frac{2\alpha \ln n}{n}} - \left(\frac{1}{2} + \alpha\right)\frac{\ln n}{n} - O(n^{-1}) \end{aligned} \tag{3.4}$$

whenever

$$\epsilon_n = \Theta\left(\frac{n^{-\alpha}}{\sqrt{\ln n}}\right). \tag{3.5}$$

(c) *Let* $\delta = \frac{c}{\sqrt{n}}$ *for a constant* c. *Then*

$$\frac{c}{\sqrt{n}} \ge R_n - H(X) \ge \frac{c}{\sqrt{n}} - \frac{\ln n}{2n} - O(n^{-1}) \tag{3.6}$$

whenever

$$\left|\epsilon_n - Q\left(\frac{c}{\sigma_H(X)}\right)\right| \le \frac{C M_H(X)}{\sqrt{n}\sigma_H^3(X)} \tag{3.7}$$

where $Q(t) = \frac{1}{\sqrt{2\pi}}\int_t^\infty e^{-u^2/2}du$, *and* $C < 1$ *is the universal constant in the central limit theorem of Berry and Esseen.*

Remark 3. To show Theorem 4 provides a non-trivial bound, we claim that

$$\delta > r_X(\delta)$$

for $0 < \delta < \ln|\mathcal{X}| - H(X)$. Indeed, recall the definition of $\delta(\lambda)$ and

$$0 \le r_X(\delta(1)) = H(X) + \delta(1) - \ln|\mathcal{X}|$$

which implies that $\delta(1) \ge \ln|\mathcal{X}| - H(X)$ or $r'_X(\delta) < 1$ for $0 < \delta < \ln|\mathcal{X}| - H(X)$. The claim then follows immediately from the fact that $r_X(0) = 0$.

Remark 4. In Part (c) of Theorem 4, we can see that if $c < 0$ is selected, then R_n could be less than $H(X)$ while ϵ_n approaches a constant $Q\left(\frac{c}{\sigma_H(X)}\right)$! This means that if the error probability is allowed to be a little larger than 0.5, the rate of source code can be even less than the entropy rate. For an IID binary source with $p = \Pr\{X_1 = 1\} = 0.12$, Figure 3 shows the tradeoff between the error probability and block length when the code rate is 0.21% below the entropy rate, where in Figure 3, both the entropy rate and code rate are expressed in terms of bits. As can be seen from Figure 3, at the block length 1000, the error probability is around 0.65, and the code rate is 0.21% below the entropy rate. Similar phenomenon can be seen for channel coding shown in [4].

Remark 5. Related to Part (c) of Theorem 4 is the second order source coding analysis in [6] with a fixed error probability $0 < \epsilon < 1$. Both results are concerned with the scenario where the rate is around the entropy rate in the order of $\frac{1}{\sqrt{n}}$ and the error probability is a constant. However, the work in [6] is asymptotic, while Part (c) of Thereom 4 part (c) is non-asymptotic and valid for any block length n. Moreover, Thereom 4 also shows the tradeoff between the rate and error probability when the error probability approaches 0 with block length n at an exponential (part (a)), a polynomial (part (b)), or other (e.g. sub-exponential, sub-polynomial) speed, which can be derived directly from (3.1) and (3.2).

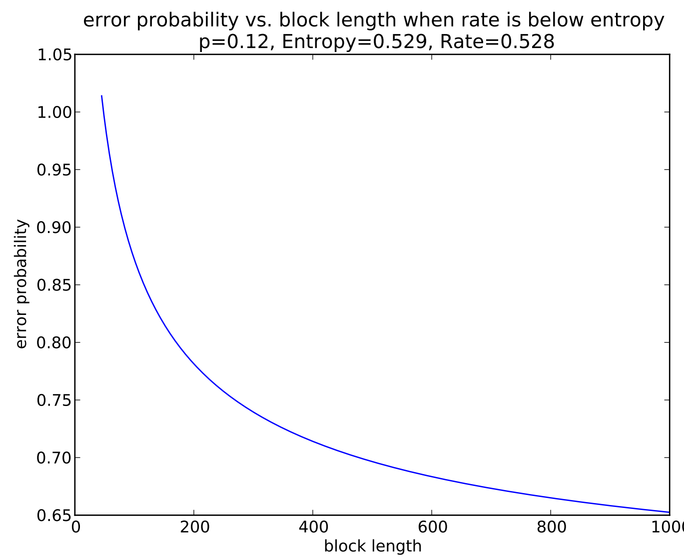


Fig. 3. Tradeoff between the error probability and block length when the rate is below the entropy rate with $p = 0.12$

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory (second edition)*. Hoboken, NJ: Wiley, 2006.

[2] P. Hall, *Rates of Convergence in the Central Limit Theorem.* Pitman Books Limited, Boston, 1982.

[3] E. -H. Yang and J. Meng, "Non-asymptotic Equipartition Properties for Independent and Identically Distributed Sources," *submitted to IEEE Trans. Inform. Theory* for publications, 2011.

[4] E. -H. Yang and J. Meng, "Jar decoding: Basic Concepts and Non-asymptotic Capacity Achieving Coding Theorems for Channels with Discrete Inputs," *submitted to IEEE Trans. Inform. Theory* for publications, 2011.

[5] E.-H. Yang and Zhen Zhang, "On the redundancy of lossy source coding with abstract alphabets," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1092–1110, May 1999.

[6] M. Hayashi, "Second order asymptotics in fixed-length source coding and intrinsic randomness", *IEEE Trans. Inform. Theory*, vol. 54, pp. 4619 - 4637, Oct. 2008.