Journal of Mathematics and Statistics

Original Research Paper

Lossy Asymptotic Equipartition Property for Networked Data Structures

Kwabena Doku-Amponsah

University of Ghana, Ghana

Article history
Received: 18-09-2016
Revised: 11-01-2017
Accepted: 24-05-2017

Email: kdoku-amponsah@ug.edu.gh

Abstract: In this study, we prove a Generalized Information Theory for Networked Data Structures modelled as random graphs. The main tools in this study remain large deviation principles for properly defined empirical measures on random graphs. To motivate the paper, we apply our main result to a concrete example from the field of Biology.

Keywords: Asymptotic Equipartition Property, Rate-Distortion Theory, Process-Level Large Deviation Principle, Relative Entropy, Random Network, Metabolic Network

Introduction

Suppose we have a networked data structure $x = \{(x(u), x(v)): uv \in e\}$ generated by a memoryless source G with distribution $P(x)$ to be compressed with distortion no greater than $d \geq 0$, using a memoryless random codebook $\hat{G}$ with distribution $P(y)$. In this instance, the compression performance can be calculated by the generalized information theory or generalized Asymptotic Equipartition Property (AEP), which gives the probability of locating a d- close match between $x = \{(x(u), x(v)): uv \in e\}$ and any given networked data structure (codeword) $y = \{(y(u), y(v)): uv \in e\}$, as a number approximately equal $2^{-nR\left(P^{(x)}, P^{(y)}, d\right)}$. The rate function $R(P^{(x)}, P^{(y)}, d)$ may be presented as an infimum of relative entropies. The aim of this article is to extend the results presented in the recent paper Doku-Amponsah (2010) and the reference therein.

To be specific, in this article, we develop a Lossy AEP for networked structured data modelled as coloured random graphs. We present and prove process Large Deviation Principle (LDP) for the coloured random graph conditioned to have a given empirical colour measure and empirical pair measure, see Doku-Amponsah (2006), using similar coupling techniques as in the article by Boucheron *et al*. (2002). From this LDP and the techniques employed by Dembo and Kontoyiannis (2002) for the random field on Z^2, we obtain the proof of the Lossy AEP for the Networked Data Structures.

We apply our Lossy AEP to a concrete example from biology, Metabolic network. This is a graph of interactions forming a part of the energy generation and biosynthesis metabolism of the bacterium *E coli*. In this case, the colours represent substrates and products and edges represent interactions Newman (2002).

The article is organized as follows: Generalized AEP for Coloured Random Graph Model section contains the core result of the paper, Theorem 2.1. LDP for two-dimensional Coloured Random Graph Model section gives process level LDP's, Theorem 3.1 and 3.2, which form the bases of the proof of the core result of the article. The final part of the paper provides the proofs of all Process Level Large deviation principles (i.e., Proof of Theorem 2.1, 3.1 and 3.2) for the paper and hence the core result of the article.

Generalized AEP for Coloured Random Graph Process

Main Result

Consider two Coloured Random Graph processes $X = \{(X(u), X(v)): uv \in E\}$ and $Y = \{(Y(u), Y(v)): uv \in E\}$ which take values in $G = G(X)$ and $\hat{G} = \hat{G}(X)$, resp., the spaces of finite graphs on X. We equip $G(X)$, $\hat{G}(X)$ with their Borel σ fields $F(x)$ and $\hat{F}^{(x)}$. Let $P(x)$ and $P(y)$ denote the probability measures of the entire processes X and Y. By $P^{(x)}_{(\sigma,\pi)}$ and $P^{(y)}_{(\sigma,\pi)}$ we denote the coloured random graphs X and Y conditioned to have empirical colour distribution σ and empirical pair distribution π. See, example (Doku-Amponsah, 2006). We always assume that X and Y are independent of each other.

By X we denote a finite alphabet and we denote by $N(X)$ the space of counting measure on X equipped with the discrete topology. By $M(X)$ we denote the space of

probability vectors on X aimed with the weak topology and $M_*(X)$ denotes the space of finite measures on X aimed with the weak topology.

Throughout the remaining part of the paper we shall assume that X and Y are Coloured Random Graph processes, (Penman, 1998). For $n \geq 1$, let $P_n^{(y)}$ n denote the marginal distribution of X on $V = \{1, 2, 3, ..., n\}$ taken with respect to $P_{(\sigma,\pi)}^{(y)}$ and $Q_n^{(y)}$ denote the marginal distribution Y on $V = \{1, 2, 3, ..., n\}$ with respect to $P_{(\sigma,\pi)}^{(y)}$.

We take ρ: $X \times N(X) \times X \times N(X) \to [0,\infty)$ as an arbitrary non-negative function and we define a sequence of single-letter distortion measures $\rho^{(n)}$: $G \times \hat{G} \to [0, \infty)$, $n \geq 1$ by:

$$\rho^{(n)}(x,y) = \frac{1}{n}\sum_{v\in V}\rho\left(B_x(v), B_y(v)\right)$$

where, $B_x(v) = (x(v), L_x(v))$ and $B_y(v) = (y(v), L_y(v))$, Given $d \geq 0$ and $x \in G$, we denote the distortion-ball of radius d by:

$$B(x,d) = \left\{y \in \hat{G} : \rho^{(n)}(x,y) \leq d\right\}$$

For $(\sigma, \pi) \in M(X) \times M(X \times X)$, we write:

$$K_{(\sigma,\pi)}(a,l) = \sigma(a)\prod_{b\in X}\frac{e^{-\pi(a,b)/\sigma(a)}\left[\pi(a,b)/\sigma(a)\right]^{\ell(b)}}{\ell(b)!},\ for\ \ell \in N(X)$$

and define the rate function I_1:$M[(X \times N(X))^2] \to [0, \infty]$ by:

$$I_1(v) = \begin{cases} H\left(v \,\|\, K_{(\sigma,\pi)} \otimes K_{(\sigma,\pi)}\right), & if\ v\ is\ consistent\ and\ v_{1,1} = v_{1,2} = \sigma, \\ \infty & otherwise, \end{cases} \quad (2.1)$$

where:

$$K_{(\sigma,\pi)} \otimes K_{(\sigma,\pi)}\left(\left(a_x, a_y\right),\left(l_x, l_y\right)\right) = K_{(\sigma,\pi)}\left(a_x, l_x\right)K_{(\sigma,\pi)}\left(a_y, l_y\right)$$

By $x\ D\ p$ we mean x has distribution p. For $(\sigma, \pi) \in M(X) \times M(X \times X)$m we write:

$$d_{av}(\sigma,\pi) = \left\langle \log\left\langle e^{t\rho\left(B_X, B_Y\right)}, K_{(\sigma,\pi)}\right\rangle, K_{(\sigma,\pi)}\right\rangle$$

Assume:

$$d_{\min}^{(n)}(\sigma,\pi) = E_{P_n^{(x)}}\left[es\sin f_{Y\,D\,Q_n^{(y)}}\rho^{(n)}(X,Y)\right] \to d\min(\sigma,\pi)$$

For $n > 1$; we write:

$$R_n\left(P_n^x, Q_n^{(y)}, d\right) := \inf_{V_n}\left\{\frac{1}{n}H\left(V_n \,\|\, P_n^{(x)} \times Q_n^{(y)}\right) : V_n \in M\left(G \times \hat{G}\right)\right\}$$

and:

$$d_{\min}^{\infty}(\sigma,\pi) := \inf\left\{d \geq 0 : \sup_{n\geq 1} R_n\left(P_n^{(x)}, Q_n^{(y)}, d\right) < \infty\right\}$$

Theorem 2.1 (ii) below gives the Lossy AEP for networked data structures.

Theorem 2.1

Suppose X and Y are coloured random graphs. Assume ρ are bounded function. Then:

- with $P^{(x)}$- probability 1, conditional on the event $\{\Phi(L_{n,1}) = \Phi(L_{n,2}) = \sigma, \pi)\}$ the random variables $\{\rho^{(n)}(x, Y)\}$ satisfy an LDP with deterministic, convex rate-function:

$$I_\rho(z) := \inf_v\left\{I_1(v) : \langle\rho, v\rangle = z\right\}$$

- for all $d \in (d_{\min}(\sigma, \pi), d_{av}(\sigma, \pi)$, except possibly at $d_{\min}^{\infty}(\sigma,\pi)$:

$$\lim_{n\to\infty} -\frac{1}{n}\log Q_n^{(x)}\left(B(X,D)\right) = R\left(P_{(\sigma,\pi)}^{(x)}, P_{(\sigma,\pi)}^{(y)}, d\right) almost\ surely \quad (2.2)$$

where, $R(p, q, D) = \inf_v H(v \,\|\, p \times q)$.

Application (Doku-Amponsah, 2012)

Metabolic Network

Let us consider a metabolic network of the energy and biosynthesis metabolism of the bacterium *E coli* modelled as coloured random graph on n nodes partitioned into $n\sigma_n$(substrate) block of substrates and $n\sigma_n$(product) block of products and $n||\_\pi_n||$ number of interactions divided into $n\pi_n$(substrate, product), $n\pi_n$(substrate, product), $n\pi_n$(substrate, substrate)/2, $n\pi_n$(product, product)/2 different interactions, respectively. Assume σ_n converges σ and π_n converges π. If we take $\rho(s, r) = (s\text{-}r)^2$ then, by Theorem 2.1 we have the distortion-rate:

$$R(P,Q,D)=\begin{cases}0, & \text{if } D\geq 2\pi(subs,prod)+\pi(subs,subs)+\pi(prod,prod)+2\pi(subs,prod)\\ \infty & \text{otherwise}\end{cases} \tag{2.3}$$

where subs := substrate and prod := product.

LDP for Two-Dimensional Coloured Random Graph Process

For any $n \in N$ we define:

$$M_n(X):=\{\sigma\in M(X):n\sigma(b)\in\mathbb{N}\ \text{for all}\ b\in X\},$$
$$\tilde{M}_n(X\times X):=\left\{\pi\in\tilde{M}_*(X\times X):\frac{n}{1+1\{b=a\}}\pi(b,a)\in\mathbb{N}\ \text{for all}\ b,a\in X\right\}$$

Throughout the proof, we shall assume that $\omega_n(a_x, a_y) > 0$, for all $a_x, a_y \in X$ and $\omega_{n,1}(a_x) = \sigma_n(a_x)$, $\omega_{n,2}(a_y) = \sigma_n(a_y)$. It is not too difficult to see that the law of the two-dimensional coloured random graph conditioned to have empirical colour distribution σ_n and empirical pair distribution π_n:

$$P(\sigma_n,\pi_n):=P\{\cdot\mid\Phi(L_{n,1})=(\omega_{n,1},\pi_n),\Phi(L_{n,2})=(\omega_{n,2},\pi_n)\}$$

Can be constructed in the following way:

- Colours are assigned to the vertices by sampling without replacement from the collection of n colours, which contains any colour $(a_x, a_y) \in X$ exactly $n\omega_n(a_x, a_y)$ times
- For each unordered pair $\{b, a\}$ of colours, we create exactly $m_n(b, a)$ edges by sampling without replacement from the pool of potential edges connecting vertices of colour b and a, where:

$$m_n(a,b):=\begin{cases}n\pi_n(a,b) & \text{if } a=a_x,b=b_x \text{ and } a_x\neq b_x\\ n\pi_n(a,b) & \text{if } a=a_y,b=b_y \text{ and } a_y\neq b_y\\ \frac{n}{2}\pi_n(a,b) & \text{if } a=a_x,b=b_x \text{ and } a_x=b_x\\ \frac{n}{2}\pi_n(a,b) & \text{if } a=a_y,b=b_y \text{ and } a_y=b_y\end{cases} \tag{3.1}$$

We define the process-level empirical measure L_n induced by X and Y on $G \times \hat{G}$ by:

$$L_n(\beta_x,\beta_y)=\frac{1}{n}\sum_{v\in V}\delta_{(B_X(v),B_Y(v))}(\beta_x,\beta_y),\ \text{for}\ (\beta_x,\beta_y)\in M\left[\left(X\times X_k^*\right)^2\right]$$

Note that we have:

$$\begin{aligned}&L_n\otimes\phi^{-1}\left(\left(x(v),y(v)\right),l_{x,y}(v)\right)\\&=\frac{1}{n}\sum_{v\in V}\delta_{(B_X(v),B_Y(v))}\left(\phi^{-1}\left(x(v),y(v)\right),l_{x,y}(v)\right)\\&=\frac{1}{n}\sum_{v\in V}\delta_{((X(v),Y(v)),L_{X,Y}(v))}\left(\left(x(v),y(v)\right),l_{x,y}(v)\right)\\&:=\tilde{L}_n\left(\left(x(v),y(v)\right),l_{x,y}(v)\right)\end{aligned}$$

where, $\phi(\beta_x,\beta_y) = ((x(v), y(v)), l_{x,y}(v)$. The next Theorem which is the LDP for L_n of the process X, Y is the main ingredient in the proof of the Lossy AEP.

Theorem 3.1

The sequence of empirical measures L_n obeys an LDP in the space of probability vectors on $(X \times N(X))^2$ equipped with the topology of weak convergence, with convex, good rate-function I_1.

The proof of Theorem 3.1 above is dependent on the LDP for $\tilde{L}_n$ given below:

Theorem 3.2

The sequence of empirical measures $\tilde{L}_n$ satisfies a large deviation principle in the space of probability measures on $X^2 \times N(X)^2$ equipped with the topology of weak convergence, with convex, good rate-function:

$$L_2(\omega)=\begin{cases}H\left(\omega\|K_{(\sigma,\pi)}\otimes K_{(\sigma,\pi)}\right), & \text{if } \omega \text{is consistent and } \omega_{1,1}=\omega_{1,2}=\sigma,\\ \infty & \text{otherwise}\end{cases} \tag{3.2}$$

where:

$$K_{(\sigma,\pi)}\otimes K_{(\sigma,\pi)}\left(\left(a_x,a_y\right),\left(l_x,l_y\right)\right)=K_{(\sigma,\pi)}\left(a_x,l_x\right)K_{(\sigma,\pi)}\left(a_y,l_y\right)$$

For any bin $v \in \{1,..., n\}$, we denote its colours by $\left(\tilde{X}(v),\tilde{Y}(v)\right)$ and for $h = x, y$, the number of balls of colour $b_h \in X$ it contains it contains is denoted by $l^v(b_h)$. Now we define an empirical process-level occupancy distribution of this constellation by:

$$\tilde{L}_n^+\left(a_x,a_y,\ell_{x,y}\right)=\frac{1}{n}\sum_{v\in V}\delta_{\left(\tilde{X}(v),\tilde{Y}(v),\tilde{L}_{X,Y}(v)\right)}\left(\left(a_x,a_y\right),\ell_{x,y}\right),$$
$$for\left(a_x,a_y,\ell_{x,y}\right)\in X^2\times N^2(X)$$

where, $\tilde{L}_{X,Y}(v) = (l^v(b_x), l^v(b_y), (b_x, b_y) \in X \times X)$ is the colour distribution in bin v. In the next theorem we prove

exponential equivalence of the law of the empirical process-level distribution $\tilde{L}_n$ under $P_{(\sigma_n,\varpi_n)}$, the law of the coloured random graph conditioned to have colour law σ_n and edge distribution π_n and the law of the empirical process-level occupancy distribution $\tilde{L}_n^+$ in the random allocation model $\tilde{P}_{(\sigma_n,\pi_n)}$ (We refer to (Dembo and Zeitouni, 1998), Definition 4.2.10) for the definition of definition of exponential equivalence).

Lemma 3.3

The law of $\tilde{L}_n^+$ under $\tilde{P}_{(\sigma_n,\pi_n)}$ and the law of $\tilde{L}_n$ under $P_{(\sigma_n,\pi_n)}$ are exponentially equiv-alent.

We define the metric d of total variation by:

$$d(v,\tilde{v})=\frac{1}{2}\sum_{\left((a_x,a_y),(l_x,l_y)\right)\in X^2\times N^2(X)}|v\left((a_x,a_y),(l_x,l_y)\right)$$
$$-\tilde{v}\left((a_x,a_y),(l_x,l_y)\right)|,\ for\, v,\tilde{v}\in M\left(X^2\times N^2(X)\right)$$

As this metric generates the weak topology, the proof of Lemma 3.3 is the equivalent to proving that for every $\varepsilon > 0$:

$$\lim_{n\to\infty}\frac{1}{n}\log P\left\{d\left(\tilde{L}_n^+,\tilde{L}_n\right)\geq\varepsilon\right\}=-\infty \tag{3.3}$$

where, P indicates a properly defined coupling measure between the random allocation model and the coloured random graph.

We begin, by denoting by $V(a)$ the set of vertices (bins) which have colour $a \in X$ and observe that:

$$\#V(a)=n\sigma_n(a)$$

For $h = x, y$ and every $a_h, b_h \in X$, begin: At each step $k = 1,..., m_n(a_h, b_h)$, we pick at random two vertices $V_1^k \in V(a_h)$ and $V_2^k \in V(b_h)$. Place one ball of colour b_h in bin V_1^k and one ball of colour a_h in V_2^k and link V_1^k to V_2^k by an edge unless $V_1^k = V_2^k$ or the two vertices are already connected. When one of these two things occur, then we simply choose an edge at random from the set of all possible edges connecting colours ah and bh, which are not yet an edge in the graph. This gives a graph with:

$$\Phi\left(\tilde{L}_{n,1}\right)=\Phi\left(\tilde{L}_{n,2}\right)=\left(\omega_n,\pi_n\right)$$

and:

$$d\left(\tilde{L}_n^+,\tilde{L}_n\right)\leq\frac{2}{n}\left(\sum_{a,b\in X}B^n\left(a_x,b_x\right)+\sum_{a,b\in X}B^n\left(a_y,b_y\right)\right) \tag{3.4}$$

where, $B^n(a, b)$ is the total number of steps $k \in \{1,..., m_n(a, b)\}$ where there is a difference between the vertices V_1^k, V_2^k drawn and the vertices which formed the kth edge linkimg a and b in the random graph constructed.

Given $a, b \in X$, the probability that $V_1^k = V_2^k$ or the two vertices are already connected is given by:

$$p_{[k]}\left(a_h,b_h\right):=\frac{1}{m_n\left(a_h,b_h\right)}1_{\{a_h,b_h\}}$$
$$+\left(1-\frac{1}{m_n\left(a_h,b_h\right)}1_{\{a_h,b_h\}}\right)\frac{(k-1)}{\left(m_n\left(a_h,b_h\right)\right)^2}$$

$B^n(a_h, b_h)$ is a sum of independent Bernoulli random variables $X_1^{(h)},...,X_{n\varpi_n(a_h,b_h)/2}^{(h)}$ with 'success' probabilities given by $p_{[1]}\left(a_h,b_h\right),...,p_{n\varpi_n(a_h,b_h)/2}\left(a_h,b_h\right)$. Note that $E[X_k] = p_{[k]}(a_h, b_h)$ and:

$$Var\left[X_k^{(h)}\right]=p_{[k]}\left(a_h,b_h\right)\left(1-p_{[k]}\left(a_h,b_h\right)\right)$$

Now, we have:

$$EB^n\left(a_h,b_h\right)=\sum_{k=1}^{n\left(a_h,b_h\right)}p_{[k]}\left(a_h,b_h\right)=1_{\{a_h,b_h\}}$$
$$+\left(1-1_{\{a_h=b_h\}}\frac{1}{m_n\left(a_h,b_h\right)}\right)\left(1-\frac{1}{m_n\left(a_h,b_h\right)}\right)\leq 1+1\{a_h=b_h\}$$

We write:

$$\sigma_n^2\left(a_h,b_h\right):=\frac{1}{m_n\left(a_h,b_h\right)}\sum_{k=1}^{m_n\left(a_h,b_h\right)}Var\left[X_k^{(h)}\right]$$

and observe that:

$$\lim_{n\to\infty}E\left(B_n\left(a_h,b_h\right)\right)=\lim_{n\to\infty}Var\left(B^n\left(a_h,b_h\right)\right)$$
$$=\lim_{n\to\infty}m_n\left(a_h,b_h\right)\sigma_n^2(a,b)=1_{\{a_h,b_h\}}+1$$

We Define $e(z) = (1 + z)\log(1 + z) - z$, for $z \geq 0$ and use Bennett's inequality, (Bennett, 1962), to arrive at, for sufficiently large n:

$$P\left\{\frac{1}{n}\sum_{h=x,y}B^n\left(a_h,b_h\right)\geq\frac{\sum_{h=x,y}1_{\{a_h=b_h\}}+1}{n}+\delta_1\right\}$$
$$\leq\exp\left[-\sum_{h=x,y}m_n\left(a_h,b_h\right)\sigma_n^2\left(a_h,b_h\right)e\left(\frac{n\delta_1}{\sum_{h=x,y}m_n a_h,b_h\sigma_n^2\left(a_h,b_h\right)}\right)\right]$$

for any $\delta_1 > 0$. Let $\varepsilon \geq 0$ and choose $\delta_1 = \frac{\varepsilon}{2m^2}$. Suppose that we have $B^n(a_h, b_h) \leq \delta_1$, for $h = x, y$. Then, by (3.4):

$$d\left(\tilde{L}, v_n\right) \leq 2\delta_1 m^2 = \varepsilon$$

Hence:

$$P\left\{d\left(\tilde{L}, \tilde{L}^+\right) > \varepsilon\right\} \leq \max_{h=x,y} \sum_{a_h, b_h \in X} P\left\{B^n\left(a_h, b_h\right) \geq n\delta_1\right\}$$

$$\leq m^2 \max_{h=x,y} \sup_{a_h, b_h \in X} P\left\{B^n\left(a_h, b_h\right) \geq 1_{\{a_h, b_h\}} + 1 + \left(n\delta_1\right)/2\right\}$$

$$\leq m^2 \max_{h=x,y} \sup_{a,b \in X} \exp\left[\begin{matrix} -m_n\left(a_h, b_h\right)\sigma_n^2\left(a_h, b_h\right) \\ e\left(\frac{n\delta_1}{m_n\left(a_h, b_h\right)\sigma_n^2\left(a_h, b_h\right)}\right)\end{matrix}\right]$$

Let $0 \leq \delta_2 \leq 1$. The, for very large n we obtain:

$$\frac{1}{n}\log P\left\{d\left(\tilde{L}, \tilde{L}^+\right) > \varepsilon\right\} \leq -\left(1-\delta_2\right)e\left(\frac{n\delta_1}{2\left(1+\delta_2\right)}\right)$$

$$= -\left(1_{(b=a)} + 1 - \delta_2\right)\left[\begin{matrix} \left(\frac{1}{n} + \frac{\delta_1}{2\left(1_{\{b=a\}} + 1 + \delta_2\right)}\right) \\ \log\left(1 + \frac{n\delta_1}{2\left(1_{\{a=b\}} + 1 + \delta_2\right)}\right) - \frac{\delta_1}{2\left(1_{\{a=b\}} + 1 + \delta_2\right)}\end{matrix}\right] \quad (3.5)$$

This ends the proof of the lemma.

Proof of Theorem 3.2, 3.1 and 2.1

Proof of Theorem 3.2

We write $\vartheta_2^{(n)} := \vartheta_2^{(n)}\left(\varpi_n, v_n\right)$, $\vartheta_1^{(n)} := \vartheta_1^{(n)}\left(\varpi_n, v_n\right)$ and state the following Lemmma. Denote by $\Sigma^{(n)}(\sigma_n, \pi_n)$ the space of all empirical neighbourhood measures with empirical colour distribution σ_n and empirical pair distributions π_n.

*Lemma 4.1 (*Doku-Amponsah, 2014*)*

For any process level empirical measure v_n with $v_{n,1}$, $v_{n,2} \in \Sigma^{(n)}(\sigma_n, \pi_n)$, we have:

$$e^{-n\left(H\left(vn,1\|K_{\left(\sigma_n,\pi_n\right)}\right)+H\left(v_{n,2}\|K_{\left(\sigma_n,\pi_n\right)}\right)\right)+\vartheta_1^{(n)}} \leq \tilde{P}_{\left(\sigma_n,\pi_n\right)}\left(\tilde{L}_n^+ = v_n\right)$$
$$\leq |\Sigma^{(n)}_{\left(\sigma_n,\pi_n\right)}|^{-2}\, e^{-n\left(H\left(vn,1\|K_{\left(\sigma_n,\pi_n\right)}\right)+H\left(v_{n,2}\|K_{\left(\sigma_n,\pi_n\right)}\right)\right)+\vartheta_2^{(n)}} \quad (4.1)$$

where:

$$K_{\left(\sigma_n,\pi_n\right)}\left(a_h, l_h\right) = \sigma_n\left(a_h\right)K_{\pi_n}\left\{l_h \mid a_h\right\}$$

and:

$$K_{\pi_n}\left\{l_h \mid a_h\right\} = \prod_{b_h \in X} \frac{e^{-\pi_n\left(a_h,b_h\right)/\sigma_n\left(a_h\right)}\left[\pi_n\left(a_h,b_h\right)/\sigma_n\left(a_h\right)\right]^{\ell\left(b_h\right)}}{\ell\left(b_h\right)!},$$

for $\ell_h \in N(X)$ *and* $h = x, y$.

$$\lim_{n\to\infty}\vartheta_2^{(n)} = \lim_{n\to\infty}\vartheta_1^{(n)} = 0$$

Proof

Note, by construction, for any process level empirical measure, v_n with $v_{n,1}$, $v_{n,2} \in \Sigma^{(n)}(\sigma_n, \pi_n)$, we have:

$$\tilde{P}_{\left(\sigma_n,\pi_n\right)}\left(\tilde{L}_n^+ = v_n\right) = \tilde{P}\left\{\tilde{L}_n^+ = v_n \mid \Phi\left(\tilde{L}_{n,1}^+\right) = \Phi\left(\tilde{L}_{n,2}^+\right) = \left(\sigma_n, \pi_n\right)\right\} \quad (4.2)$$

$$= \prod_{h=x,y}\prod_{a_h \in X}\left(\begin{matrix} n\sigma_n\left(a_h\right) \\ nv_{n,u(h)}\left(a_h, \ell_h\right), \ell_h \in N(X)\end{matrix}\right) \prod_{a_h, b_h \in X}\left(\begin{matrix} n\pi_n\left(a_h, b_h\right) \\ \ell_{a_h}^{(j)}\left(b_h\right), j = 1, \ldots, n\omega_n\left(a_h\right)\end{matrix}\right)\left(\frac{1}{n\sigma_n\left(a_h\right)}\right)^{n\pi_n\left(a_h,b_h\right)} \quad (4.3)$$

while $\tilde{P}_{\left(\sigma_n,\pi_n\right)}\left(\tilde{L}_n^+\right) = 0$ when $\Phi\left(\tilde{L}_{n,1}^+\right) \neq (\sigma_n, \pi_n)$ or $\Phi\left(\tilde{L}_{n,2}^+\right) \neq (\sigma_n, \pi_n)$ by convention. Therefore, by similar combinatoric computations as in the proof of (Doku-Amponsah, 2014), Lemma 0.6) and the Sterling's formula see, (Feller, 1968) we have 4.1.

The proof of Theorem 3.2 is derived from Lemma 4.1 and similar arguments as (Doku-Amponsah, 2014, Page 13).

Proof of Theorem 3.1

Let $\Gamma \in M[(X \times N(X))^2]$ and write $\Gamma_\phi = \{\omega \otimes \phi^{-1}: \omega \in \Gamma\}$. Note that if A is closed (open) then Γ_ϕ is a closed (an open) since ϕ is linear. Now suppose F is a closed subset of $M[(X \times N(X))^2]$ then by Theorem 3.2 we have:

$$-\inf_{\omega\in F} I_2\left(\omega \otimes \phi^{-1}\right) = -\inf_{v\in F_\phi} I_2\left(v\right) \leq \liminf_{n\to\infty}\frac{1}{n}\log P\left\{\tilde{L}_n \in F_\phi\right\}$$
$$\leq \lim_{n\to\infty}\frac{1}{n}\log P\left\{L_n \in F\right\} \leq \liminf_{n\to\infty}\frac{1}{n}\log P\left\{\tilde{L}_n \in F_\phi\right\}$$
$$\leq -\lim_{v\in F_\phi} I_2\left(v\right) = -\inf_{\omega\in F} I_2\left(\omega \otimes \phi^{-1}\right)$$

The form of the rate function in Theorem 3.1 is obtained if we solve the optimization problem:

$$\inf\left\{I_2(v):\omega\otimes\phi^{-1}=v\right\}=I_1(\omega)$$

Proof of Theorem 2.1

We write $M:=M[(X\times N(X))^2]$ and define the set C^ε by:

$$C^\varepsilon(\sigma,\pi)=\left\{\begin{matrix}v\in M:\sup\limits_{\beta_x,\beta_y\in X\times N(X)}|v(\beta_x,\beta_y)\\ -K_{(\sigma,\pi)}(\beta_x,\beta_y)|\geq\varepsilon\end{matrix}\right\}$$

Lemma 4.2

Suppose the sequence of measures (σ_n, π_n) converges to the pair of measures (σ, π): For any $\varepsilon > 0$ we have $\lim_{n\to\infty}P_{(\sigma_n,\pi_n)}(C^\varepsilon)=0$.

Proof

Observe that C^ε defined above is a closed subset of M and so by Theorem 3.1 we have that:

$$\limsup_{n\to\infty}\frac{1}{n}\log P(\sigma_n,\pi_n)(C^\varepsilon)\leq-\inf_{v\in C^\varepsilon}I_1(v)\qquad(4.4)$$

We use proof by contradiction to show that the right hand side of (4.4) is negative. Suppose that there exists sequence v_n in C^ε such that $I_1(v_n)\downarrow 0$. Then, there is a limit point $v\in F_1$ with $I(v) = 0$. Note I is a good rate function and its level sets are compact and the mapping $v\mapsto I(v)$ lower semi-continuity. Now $I_1(v) = 0$ implies $v(\beta_x, \beta_y) = K_{(\sigma, \pi)}\otimes K_{(\sigma, \pi)}(\beta_x, \beta_y)$, for all $\beta_x, \beta_y\in X\times N(X)$ which contradicts $v\in C^\varepsilon$.

Notice $\rho^{(n)}(X, Y) = \langle\rho,L_n\rangle$ and if Γ is an open (a closed) subset of M then:

$$\Gamma_\rho:=\left\{v:\langle\rho,v\rangle\in\Gamma\right\}$$

is also an open (a closed) set since ρ is bounded function:

$$\begin{aligned}&-\inf_{z\in In(\Gamma)}I_\rho(z)=-\inf_{v\in\ln(\Gamma_\rho)}I_1(v)\\&\leq\liminf_{n\to\infty}\frac{1}{n}\log P\left\{\begin{matrix}\rho^{(n)}(X,Y)\in\Gamma X\\ =x,\Phi(L_{n-1})=\Phi(L_{n,2})=(\sigma_n,\pi_n)\end{matrix}\right\}\\&\leq\lim_{n\to\infty}\frac{1}{n}\log P\left\{\begin{matrix}\rho^{(n)}(X,Y)\in\Gamma\mid X\\ =x,\Phi(L_{n,1})=\Phi(L_{n,2})=(\sigma_n,\pi_n)\end{matrix}\right\}\\&\leq\limsup_{n\to\infty}\frac{1}{n}\log P\left\{\begin{matrix}\rho^{(n)}(X,Y)\in\Gamma\mid X\\ =x,\Phi(L_{n,1})=\Phi(L_{n,2})=(\sigma_n,\pi_n)\end{matrix}\right\}\\&\leq-\inf_{v\in cl(\Gamma_\rho)}I_1(v)=-\inf_{z\in cl(\Gamma)}I_\rho(z)\end{aligned}$$

Observe that ρ are bounded, therefore by Varadhan's Lemma and convex duality, we have:

$$R(P^x,P^y,d)=\sup_{t\in\mathbb{R}}\left[td-\Lambda_\infty(t)\right]=\Lambda_\infty^*(d)$$

where:

$$\Lambda_\infty^*(t):=\lim_{n\to\infty}\frac{1}{n}\log\int e^{nt\langle\rho,L_n\rangle}dQ_n(y)$$

Exists for P almost everywhere x. Using bounded convergence, we can show that:

$$\Lambda_\infty(t):=\lim_{n\to\infty}\Lambda_n(t)=\lim_{n\to\infty}\frac{1}{n}\int\left[\log\int e^{nt\langle\rho,L_n\rangle}dQ_n^{(y)}(y)\right]dP_n^{(x)}(x)$$

Using Lemma 4.4, by boundedness of ρ we have that:

$$\begin{aligned}&\frac{1}{n}\Lambda_n(nt)=\frac{1}{n}\sum_{j=1}^n\log E_{Q_n^{(y)}}\left(e^{t\rho(B_x(j),B_y(j))}\right)\\&\to\left\langle\log\left\langle e^{t\rho(B_X,B_Y)},K_{(\sigma,\pi)}\right\rangle,K_{(\sigma,\pi)}\right\rangle=d_{av}(\sigma,\pi)\end{aligned}$$

Also let:

$$D_{\min}^{(n)}(\sigma,\pi):=\lim_{t\downarrow-\infty}\frac{\Lambda_n(t)}{t}$$

so that $\Lambda_n^*(d)=\infty$ for $d<d<d_{\min}^{(n)}(\sigma,\pi)$, while $\Lambda_n^*(D)<\infty$ for $d_{\min}^{(n)}(\sigma,\pi)$. Observe that for $n < 1$ we have $D_{\min}^{(n)}(\sigma,\pi)=E_{P_n}\left[essinf_{YDQ_n}\rho(n)(X,Y)\right]$, which converges to $d_{\min}(\sigma, \pi)$. Using similar arguments as (Dembo and Kontoyiannis, 2002, Proposition 2) we obtain:

$$R_n\left(P_n^{(x)},Q_n^{(y)},d\right)=\sup_{t\in\mathbb{R}}\left(td-\Lambda_n(t)\right):=\Lambda_n^*(d)$$

Now we observe from (Dembo and Kontoyiannis, 2002, Page 41) that the converge of $\Lambda_n^*(\cdot)\to\Lambda_\infty(\cdot)$ is uniform on compact subsets of R. Moreover, Λ_n is convex, continuous function converging informally to Λ_∞ and hence we can invoke (Shannon, 1948, Theorem 5) to obtain:

$$\Lambda_n^*(d)=\lim_{\delta\to0}\limsup_{n\to\infty}\inf_{|\hat{d}-d|<\delta}\Lambda_n^*(\hat{d})$$

Using similar arguments as (Dembo and Kontoyiannis, 2002, Page 41) in the lines after equation (64), we have (2.3) which completes the proof.

Conclusion

In this article we have found a rate distortion theorem for networked data structures. As an application the abstract theorem was applied to biosynthesis metabolism of the bacterium *E coli*. This theorem could serve as the basis for providing efficient coding/compressing algorithm and/or approximate pattern matching algorithms for networked data structures modeled as coloured random graphs.

Acknowledgement

This extension has been discussed in the author's PhD Thesis at University of Bath.

Conflict of Interest

The author declares that he has no conict of interest.

Reference

Bennett, G., 1962. Probability inequalities for the sum of independent random variables. J. Am. Stat. Assoc., 57: 33-45. DOI: 10.2307/2282438

Boucheron, S., F. Gamboa and C. Leonard, 2002. Bins and balls: Large deviations of the empirical occupancy process. Ann. Applied Probab., 12: 607-636. DOI: 10.1214/aoap/1026915618

Dembo, A. and I. Kontoyiannis, 2002. Source coding, large deviations and approximate pattern matching. IEEE Trans. Inform. Theory, 48: 1590-1615. DOI: 10.1109/TIT.2002.1003841

Dembo, A. and O. Zeitouni, 1998. Large Deviations Techniques and Applications. 1st Edn., Springer, New York, ISBN-10: 0387984062, pp: 396.

Doku-Amponsah, K., 2006. Large deviations and basic information theory for hierarchical and networked data structures. PhD Thesis, Bath.

Doku-Amponsah, K., 2010. Large deviation results for critical multitype Galton-Watson trees.

Doku-Amponsah, K., 2012. Asymptotic equipartition properties for simple hierarchical and networked structures. ESAIM: PS, 16: 114-138. DOI: 10.1051/ps/2010016

Doku-Amponsah, K., 2014. Exponential approximation, method of types for empirical neighbourhood distributions of random graphs by random allocations. Int. J. Stat. Probability, 3: 110-120. DOI: 10.5539/ijsp.v3n2p110

Feller, W., 1968. An Introduction to Probability Theory and its Applications. 3rd Edn., Wiley, New York, ISBN-10: 0471257087, pp: 528.

Newman, M.E., 2002. Random graphs as models of networks.

Penman, D.B., 1998. Random graphs with correlation structure. PhD Thesis, Sheeld.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J., 27: 623-656. DOI: 10.1002/j.1538-7305.1948.tb00917.x