LOSSY ASYMMPTOTIC EQUIPARTITION PROPERTY FOR HIERARCHICAL DATA STRUCTURES

BY KWABENA DOKU-AMPONGSAH

University of Ghana

Abstract. This paper presents a rate-distortion theory for hierarchical networked data structures modelled as tree-indexed multitype process. To be specific, this paper gives a generalized Asymptotic Equipartition Property (AEP) for the Process. The general methodology of proof of the AEP are process level large deviation principles for suitably defined empirical measures for multitype Galton-Watson trees.

1. Introduction

Rate distortion theory (RDT) play crucial role in approximate pattern-matching in information theory. It provides the mathematical foundations for lossy data compression; it takes care of the problem of looking for the minimal number of bits per symbol, as measured by the rate $R$, that should be transmitted over a channel, so that the source (input signal) can be approximately decipher at the receiver (output signal) without exceeding a given distortion $D$. The RDT is mostly centered around a lossy version of the AEP, see example [CT91]. Several lossy versions of the AEP have been formulated for linear data sources including stationary ergodic random fields on the $Z^d$, the $d$– dimensional Lattice. See example [DK02] and the reference therein. This lossy AEP have been applied to strengthened versions of Shannon’s direct source coding and universal coding theorems, characterize the performance of ”mismatched” code books in lossy data compression, analyse the performance pattern-matching algorithms for lossy compression (including the Lempel-Ziv schemes), determine the first order asymptotics of waiting times (with distortion) between stationary process and characterize the best achievable rate of weighted codebooks as an optimal sphere-covering exponent. See [DK02]. In Doku-Amponsah [DA10] an AEP has been found for hierarchical structured data. Such naturally tree-like data exists and are usually encountered in communication studies, demographic studies, biological population studies and the field of physics. Example, the age structure of a given population is best modelled by genealogical trees. The lossy version of the AEP in [DA10] is yet to be developed.

In this paper we develop a Lossy AEP for hierarchical data structures modelled as multitype Galton-Watson trees. To be specific about this methodology, we use LDP for the empirical offspring measure of the critical, irreducible multitype Galon-Watson trees, see [DA06], to prove an LDP for two dimensional multitype multitype Galton-Watson trees. Using this LDP together with the techniques employed by Dembo and Kontoyiannis [DK02] for the random field on $Z^2$ we obtain the proof of the Lossy AEP for the hierarchical data structures.

Mathematics Subject Classification : 94A15, 94A24, 60F10, 05C80
Keywords: Asymptotic equipartition Property, rate-distortion theory, empirical measure, perron frobenious eigenvalue, perron frobenious eigenvector, weak irreducibility, relative entropy.

Address: Statistics Department, University of Ghana, Box LG 115, Legon,Ghana. E-mail: kdoku@ug.edu.gh.
The outline of the paper is given as follows. Generalized AEP for Multitype Galton-Watson Process section contain the main result of the paper, Theorem 2.1. LDP for two-dimensional multitype galton-watson process section gives process level LDP’s, Theorem 3.1 and 3.2 which form the bases of the proof the main result of the paper. Proof of Theorem 2.1, 3.1 and 3.2 section provides the proofs of all Process Level LDP’s for the paper and hence the main result of the paper.

2. Generalized AEP for Multitype Galton-Watson Process

2.1 Main Result

Consider two multitype Galton-Watson processes \( X = \{(X(v),C_X(v)) : v \in V\} \) and \( Y = \{(Y(v),C_Y(v)) : v \in V\} \) which take values in \( \mathcal{T} = \mathcal{T}(\mathcal{X}) \) and \( \hat{\mathcal{T}} = \hat{\mathcal{T}}(\mathcal{X}) \), resp., the spaces of finite trees on \( \mathcal{X} \). We equip \( \mathcal{T}(\mathcal{X}) \), \( \hat{\mathcal{T}}(\mathcal{X}) \) with their Borel \( \sigma \)-fields \( \mathcal{F} \) and \( \hat{\mathcal{F}} \). Let \( \mathbb{P}_x \) and \( \mathbb{P}_y \) denote the probability measures of the entire processes \( X \) and \( Y \). By \( \mathcal{X} \) we denote a finite alphabet and write \( \mathcal{X}_k^* = \bigcup_{n=0}^{k} \{n\} \times \mathcal{X}^n \), where \( k \in \mathbb{Z}^+ \). We always assume that \( X \) and \( Y \) are independent of each other.

Throughout the rest of the article we will assume that \( X \) and \( Y \) are irreducible, critical multitype Galton-Watson processes. See example [DMS03]. For \( n \geq 1 \), let \( P_n \) denote the marginal distribution of \( X \) given \( |V(T)| = n \) taking with respect to \( \mathbb{P}_x \) and \( Q_n \) denote the marginal distribution \( Y \) given \( |V(T)| = n \) with respect to \( \mathbb{P}_y \). Let \( \rho : \mathcal{X} \times \mathcal{X}^* \times \mathcal{X} \times \mathcal{X}^* \to [0,\infty) \) be an arbitrary non-negative function and define a sequence of single-letter distortion measures \( \rho^{(n)} : \mathcal{T} \times \hat{\mathcal{T}} \to [0,\infty), n \geq 1 \) by

\[
\rho^{(n)}(x,y) = \frac{1}{n} \sum_{v \in V} \rho\left(\mathcal{A}_x(v), \mathcal{A}_y(v)\right),
\]

where \( \mathcal{A}_x(v) = (x(v),c_x(v)) \) and \( \mathcal{A}_y(v) = (y(v),c_y(v)) \). Given \( d \geq 0 \) and \( x \in \mathcal{T} \), we denote the distortion-ball of radius \( d \) by

\[
B(x,d) = \left\{ y \in \hat{\mathcal{T}} : \rho^{(n)}(x,y) \leq d \right\}.
\]

Theorem 2.1 below is the generalized Shannon-McMillan-Breiman Theorem or Lossy Asymptotic Equipartition Property for the hierarchical data structures. Define the matrix \( A : \mathcal{X}^2 \times \mathcal{X}^2 \to \mathbb{R}_+ \cup \{0\} \) by

\[
A[(a,\hat{a}),(b,\hat{b})] = \sum_{(c,\hat{c}) \in \mathcal{X}^2} m(a,c)m(\hat{a},\hat{c})K_x(c|b)K_y(\hat{c}|\hat{b}).
\]

By \( x \triangleright p \) we mean \( x \) has distribution \( p \). For \( \pi \) the eigen vector corresponding to the largest eigen value 1 of the matrix \( A \), we write

\[
d_{av} = \langle \log(e^{\rho(\mathcal{A}_x,\mathcal{A}_y)}), \pi_1 \otimes K_x \rangle,
\]

and assume \( d_{mn}^{(n)} = E_{P_n}[\text{essinf}_{Y \triangleright Q_n} \rho^{(n)}(X,Y)] \) converges to \( D_{mn} \). For \( n > 1 \), we write

\[
R_n(P_n,Q_n,d) := \inf_{V_n} \left\{ \frac{1}{n} H(V_n \parallel P_n \times Q_n) : V_n \in \mathcal{M}(\mathcal{T} \times \hat{\mathcal{T}}) \right\}
\]

and write

\[
d_{mn}^\infty := \inf \left\{ d \geq 0 : \sup_{n \geq 1} R_n(P_n,Q_n,d) < \infty \right\}.
\]
We call \( \nu \in \mathcal{M}[(X \times X^*)^2] \) with marginals \( \nu_1 \) and \( \nu_2 \) respectively, shift-invariant if
\[
\nu_{11}(a) = \sum_{(b,c) \in X \times X^*} m(a,c) \nu_1(b,c), \ a \in X
\]
and
\[
\nu_{21}(a) = \sum_{(b,c) \in X \times X^*} m(a,c) \nu_2(b,c), \ a \in X,
\]
where \( m(a,c) \) is the multiplicity of the symbol \( a \) in \( c \). See [DMS03]. We define the rate function \( I_1 : \mathcal{M}[(X \times X^*)^2] \to [0, \infty) \) by
\[
I_1(\nu) = \begin{cases} H(\nu \parallel \nu_{11} \otimes K_X \times \nu_{21} \otimes K_Y), & \text{if } \nu \text{ is shift-invariant,} \\ \infty, & \text{otherwise.} \end{cases} \tag{2.1}
\]

**Theorem 2.1.** Suppose \( X \) and \( Y \) are critical, weakly irreducible Multitype Galton-Watson trees with transition kernels \( K_X \) and \( K_Y \). Assume \( \rho \) are bounded functions. Then,

(i) with \( \mathbb{P}_x \)-probability 1, conditional on the event \( \{X = x, V(T) = n\} \) the random variables \( \left\{ \rho^{(n)}(x,Y) \right\} \) satisfy an LDP with deterministic, convex rate-function
\[
I_\rho(z) := \inf_{\omega} \left\{ I_1(\omega) : \langle \rho, \omega \rangle = z \right\}.
\]

(ii) for all \( d \in (d_{\text{min}}, d_{\text{av}}) \), except possibly at \( d = d_{\text{min}}^\infty \),
\[
\lim_{n \to \infty} -\frac{1}{n} \log Q^n_x(B(X,D)) = R(\mathbb{P}_x, \mathbb{P}_y, d) \text{ almost surely,} \tag{2.2}
\]
where \( R(p,q,D) = \inf_{\nu} H(\nu \parallel p \times q) \).

### 2.2 Application [DST18]

**Mutations in mitochondrial DNA.** Mitochondria are organelles in cells carrying their own DNA. Like nuclear DNA, mtDNA is subject to mutations which may take the form of base substitutions, duplication or deletions. The population mtDNA is modelled by two-type process where the units are 1 (normals) and 0 (mutant), and the links are mother-child relations. A normal can give birth to either all normals or, if there is mutation, normals and mutants. Suppose the latter happens with probability or mutation rate \( \alpha \in [0, 1] \). Mutants can only give birth to mutants. A DNA molecule may also die without reproducing. We denote by \( \emptyset \) the event absence of offspring. Assume that the population is started from one normal ancestor. Suppose the offspring kernel \( K \) is given by
\[
K\{ (2, a_1, a_2) | 1 \} = \left( \frac{1}{2} \right) \prod_{k=1}^2 K_\alpha\{a_k | 1\},
\]
\[
K\{ (2, a_1, a_2) | 0 \} = \left( \frac{1}{2} \right) \prod_{k=1}^2 K_\alpha\{a_k | 0\},
\]
where \( K_\alpha\{\emptyset | 1\} = 0, K_\alpha\{0 | 1\} = \alpha, K_\alpha\{1 | 1\} = (1 - \alpha), K_\alpha\{0 | 0\} = 1 \) and \( K_\alpha\{\emptyset | 0\} = 0 \). Note that the matrix \( A \) given by
\[
A = \begin{pmatrix}
1 & 0 & 0 & 0 \\
\alpha & 1 - \alpha & 0 & 0 \\
\alpha & 0 & 1 - \alpha & 0 \\
\alpha^2 & \alpha(1 - \alpha) & \alpha(1 - \alpha) & (1 - \alpha)^2
\end{pmatrix}
\]
is weakly irreducible $4 \times 4$ matrix, see \[DMS03\], with largest eigen value 1 and the corresponding eigen vector given by

$$\pi = \begin{pmatrix} \pi(0,0) \\ \pi(0,1) \\ \pi(1,0) \\ \pi(1,1) \end{pmatrix} = \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix}$$

Therefore, Theorem \ref{thm:ldp} hold with the distortion-rate

$$R(P,Q,D) = \begin{cases} 0, & \text{if } D \geq \frac{3}{4}(1-\alpha) + \frac{1}{4}\alpha(1-\alpha)^3, \\ \infty, & \text{otherwise.} \end{cases} \tag{2.3}$$

3. LDP FOR TWO-DIMENSIONAL MULTITYPE GALTON-WATSON PROCESS

Given a probability measure $\mu : \mathcal{X} \times \hat{\mathcal{X}} \rightarrow [0,1]$ and transition kernel $K$ we define the two-dimensional multitype Galton Watson tree as follow:

- Assign the root $\eta$ type $(X(\eta), Y(\eta))$ independently according to $\mu$.
- Give any vertex $v$ with type $(a, b)$ offspring types and number of springs $C_{X,Y}(v)$ independent everything according to

$$K\{(C_{X,Y}(v) = (c_a, c_b)|| (a, b)\} = K_{x}\{c_a\}K_{y}\{c_b\}$$

We define the process-level empirical measure $\mathcal{L}_n$ induced by $X$ and $Y$ on $T \times \hat{T}$ by

$$\mathcal{L}_n(a_x, a_y) = \frac{1}{n} \sum_{v \in V} \delta((A_X(v), A_Y(v))(a_x, a_y), \text{ for } (a_x, a_y) \in \mathcal{M}[(\mathcal{X} \times \mathcal{X}^*_K)^2].$$

Note that we have

$$\mathcal{L}_n \otimes \phi^{-1}((x(v), y(v)), c_{x,y}(v)) = \frac{1}{n} \sum_{v \in V} \delta((A_X(v), A_Y(v))(x(v), y(v)), c_{x,y}(v))$$

$$= \frac{1}{n} \sum_{v \in V} \delta((x(v), y(v)), c_{x,y}(v)) := \hat{\mathcal{L}}_n((x(v), y(v)), c_{x,y}(v)),$$

where $\phi(A_X(v), A_Y(v)) = ((x(v), y(v)), c_{x,y}(v))$. The next Theorem which is the LDP for $\mathcal{L}_n$ of the process $X, Y$ is the main ingredient in the proof of the Lossy AEP.

**Theorem 3.1.** The sequence of empirical measures $\mathcal{L}_n$ satisfies a large deviation principle in the space of probability measures on $(\mathcal{X} \times \mathcal{X}^*_K)^2$ equipped with the topology of weak convergence, with convex, good rate-function $I_1$.

The proof of Theorem \ref{thm:ldp} above is dependent on the LDP for $\hat{\mathcal{L}}_n$ given below:

**Theorem 3.2.** The sequence of empirical measures $\hat{\mathcal{L}}_n$ satisfies a large deviation principle in the space of probability measures on $\mathcal{X}^2 \times \mathcal{X}^2$ equipped with the topology of weak convergence, with convex, good rate-function

$$I_2(\omega) = \begin{cases} H(\omega \| \omega_1 \otimes K_x \times K_y), & \text{if } \omega \text{ is shift-invariant,} \\ \infty, & \text{otherwise}, \end{cases} \tag{3.1}$$

where $\omega_1 \otimes K_x \times K_y((a, b), (c_a, c_b)) = \omega_1(a,b)K_x\{c_a\}K_y\{c_b\}$. 


4. Proof of Theorem 3.2, 3.1 and 3.2

4.1 Proof of Theorem 3.2

Corollary 4.1 ([DA16]). Let $Z$ be a weekly irreducible, critical multitype Galton-Watson tree with an offspring law $Q$ whose second moment is finite, conditioned to have exactly $n$ vertices. Then, for $n \to \infty$, the empirical offspring measure $M_Z$ satisfies an LDP in $P[Z \times Z^*]$ with speed $n$ and the convex, good rate function

$$\Phi_Q(\varpi) = \begin{cases} H(\varpi \| \varpi_1 \otimes Q) & \text{if } \varpi \text{ is weak shift-invariant,} \\ \infty & \text{otherwise.} \end{cases} \quad (4.1)$$

The proof of Theorem 3.2 follows from Corollary 4.1 by the contraction principle, see [DZ98] applied to the linear mapping given by $\mathcal{G}(M(x,y)) = \tilde{L}_n$, $z = (x, y)$. The rate function governing this LDP is given by

$$I_2(\omega) = \{ \Phi_Q(\varpi) : \varpi = \omega, Q = K \times K \}.$$  

We obtain the form of the rate function $I_2$ in (4.1) if we note that $P(\mathcal{X}^2 \times \mathcal{X}^{*^2}_k)$ where $Z = \mathcal{X}^2$ and $Z^* = \mathcal{X}^{*^2}_k$.

4.2 Proof of Theorem 3.1

Lemma 4.2. $\mathcal{L} \otimes \phi$ obeys an LDP on the space $P[\phi(\mathcal{X}^2 \times \mathcal{X}^{*^2}_k)]$ with good rate function $I_2$, where $\phi^{-1}((a, b), (c, d)) = (A_a, A_b)$ and $a_z = (z, c_z)$.

Proof. Let $\Gamma \in P((\mathcal{X} \times \mathcal{X}^*_k)^2)$ and write $\Gamma_{\phi} = \{ \omega : \omega \otimes \phi \in \Gamma \}$. Note that if $A$ is closed (open) then $\Gamma_{\phi}$ is closed (open) since $\rho$ is bounded. Now suppose $F$ is closed subset of $P[(\mathcal{X} \times \mathcal{X}^*_k)^2]$ then we have

$$\lim_{n \to \infty} \frac{1}{n} \log P\{ \tilde{L} \otimes \phi \in F \} = \lim_{n \to \infty} \frac{1}{n} \log P\{ \tilde{L} \in F_{\phi} \} \leq - \inf_{\omega \in F_{\phi}} I_2(\omega) = - \inf_{\omega \otimes \phi \in F} I_2(\omega \otimes \phi).$$

Suppose $G$ is open subset of $P[(\mathcal{X} \times \mathcal{X}^*_k)^2]$ the we have also that

$$\lim_{n \to \infty} \frac{1}{n} \log P\{ \tilde{L} \otimes \phi \in G \} = \lim_{n \to \infty} \frac{1}{n} \log P\{ \tilde{L} \in G_{\phi} \} \geq - \inf_{\omega \in G_{\phi}} I_2(\omega) = - \inf_{\omega \otimes \phi \in G} I_2(\omega \otimes \phi).$$

By Lemma 4.2 and the contraction principle applied to the linear mapping $\hat{L}(\phi)(A_x, A_y) = \mathcal{L}(A_x, A_y)$, we have that $\mathcal{L}$ obeys a LDP on the space $P[\phi(\mathcal{X}^2 \times \mathcal{X}^{*^2}_k)] = P[(\mathcal{X} \times \mathcal{X}^*_k)^2]$ with rate function

$$I_1(\nu) = \{ I_2(\omega \otimes \phi) : \omega \otimes \phi = \nu \} = \begin{cases} H(\nu \| \nu_{1,1} \otimes K \times \nu_{2,1} \otimes K), & \text{if } \nu \text{ is shift-invariant,} \\ \infty & \text{otherwise.} \end{cases}$$

4.3 Proof of Theorem 2.1

(i) Notice $\rho^{(n)}(X, Y) = (\rho, \mathcal{L}^{(X,Y)}_n)$ and if $\Gamma$ is open (closed) subset of $\mathcal{M}((\mathcal{X} \times \mathcal{X}^*_k)^2)$ then

$$\Gamma_{\rho} := \{ \omega : (\rho, \omega) \in \Gamma \}$$

is also open (closed) set since $\rho$ is bounded function.
\[ \inf_{z \in l^2(\Gamma)} I_\rho(z) = - \inf_{\omega \in l^2(\Gamma)} I_1(\omega) \]

\[ \leq \lim_{n \to \infty} \inf \frac{1}{n} \log P\left\{ \rho^{(n)}(X, Y) \in \Gamma \big| X = x, V(T) = n \right\} \]

\[ \leq \lim_{n \to \infty} \inf \frac{1}{n} \log P\left\{ \rho^{(n)}(X, Y) \in \Gamma \big| X = x, V(T) = n \right\} \]

\[ \leq \limsup_{n \to \infty} \frac{1}{n} \log P\left\{ \rho^{(n)}(X, Y) \in \Gamma \big| X = x, V(T) = n \right\} \leq - \inf_{\omega \in \partial l^2(\Gamma)} I_1(\omega) \]

(ii) Observe that \( \rho \) are bounded, therefore by Varadhan’s Lemma and convex duality, we have

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

where

\[ \Lambda_\infty^*(t) := \lim_{n \to \infty} \frac{1}{n} \log \int e^{nt \langle \rho, \mathcal{L}_n^{(X,Y)} \rangle} dP_n(Y) \]

exists for \( \mathbb{P} \) almost everywhere \( x \). Using bounded convergence, we can show that

\[ \Lambda_\infty(t) = \lim_{n \to \infty} \Lambda_n(t) := \lim_{n \to \infty} \frac{1}{n} \int \log \left[ \int e^{nt \langle \rho, \mathcal{L}_n^{(X,Y)} \rangle} dP_n(Y) \right] dP_n(X) \]

Define the matrix \( A : \mathcal{X}^2 \times \mathcal{X}^2 \to \mathbb{R}_+ \cup \{0\} \) by

\[ A[\{a, \hat{a}, b, \hat{b}\}] = \sum_{(c, \hat{c}) \in \mathcal{X}^2} m(a, c) m(\hat{a}, \hat{c}) K_{x\{c | b\}} K_{y}\{\hat{c} | \hat{b}\} \]

As \( X, Y \) is critical and irreducible the matrix \( A \) is irreducible and the largest eigen value is 1. Therefore there exists a unique Perron-Frobenius eigen vector \( \pi \) (normalized to probability vector) corresponding to this largest eigen value (of the matrix \( A \)), see [DA10, Lemma 3.18], such that

\[ \frac{1}{n} \Lambda(nt) = \frac{1}{n} \sum_{j=1}^{n} \log \mathbb{E}_{Q_n} \left( e^{t \rho(\mathcal{A}_j \cdot \mathcal{A}_j(j))} \right) \to \langle \log e^{t \rho(\mathcal{A}_x \cdot \mathcal{A}_y) \rangle, \pi_1 \otimes K_x, \pi_2 \otimes K_y \rangle = d_{av} \]

where \( \pi_1 \) and \( \pi_2 \) are the first and second marginals of \( \pi \), respectively. Recall that by \( x \mathcal{D} p \) we mean \( x \) distributed as \( P \). Also let

\[ D_{\min}^{(n)} := \lim_{t \to -\infty} \frac{\Lambda_n(t)}{t} \]

so that \( \Lambda_n^*(d) = \infty \) for \( d < d_{\min}^{(n)} \), while \( \Lambda_n^*(D) < \infty \) for \( d > d_{\min}^{(n)} \). Observe that for \( n < \infty \) we have

\[ D_{\min}^{(n)}(d) = \mathbb{E}_{P_n} \left[ \text{essinf} \gamma D_{Q_n} \rho^{(n)}(X, Y) \right] \]

which converges to \( d_{\min} \). Using similar arguments as [DK02, Proposition 2] we obtain

\[ R_n(P_n, Q_n, d) = \sup_{t \in \mathbb{R}} \left( t d - \Lambda_n(t) \right) := \Lambda_n^*(d) \]

Now we observe from [DK02, Page 41] that the converge of \( \Lambda_n^*(\cdot) \to \Lambda_\infty^*(\cdot) \) is uniform on compact subsets of \( \mathbb{R} \). Moreover, \( \Lambda_n \) convex, continuous functions converge informally to \( \Lambda_\infty \) and hence we can invoke [See48, Theorem 5] to obtain

\[ \Lambda_n^*(d) = \lim_{\delta \to 0} \limsup_{n \to \infty} \inf_{|d - \delta| < \delta} \Lambda_n^*(\delta) \]

Using similar arguments as [DK02, Page 41] in the lines after equation (64) we have (2.3) which completes the proof.
REFERENCES

[CT91] T.M. Cover and J.A. Thomas. Elements of Information Theory. Wiley Series in Telecommunications, (1991).

[DA06] K. Doku-Amponsah. Large deviations and basic information theory for hierarchical and networked data structures. PhD Thesis, Bath (2006).

[DA10] K. Doku-Amponsah. Asymptotic equipartition properties for hierarchical and networked structures. ESAIM: PS 16 (2012): 114-138. DOI: 10.1051/ps/2010016.

[DA16] K. Doku-Amponsah. Large deviation Results for Critical Multitype Galton-Watson trees. https://arxiv.org/pdf/1009.3036.pdf

[DK02] A. Dembo and I. Kontoyiannis. Source Coding, Large deviations and Approximate Pattern. Invited paper in IEEE Transaction on information Theory, 48(6):1590-1615, June (2002).

[DMS03] A. Dembo, P. Mörters and S. Sheffield. Large deviations of Markov chains indexed by random trees. Ann. Inst. Henri Poincaré: Probab.et Stat.41, (2005) 971-996.

[DZ98] A. Dembo and O. Zeitouni. Large deviations techniques and applications. Springer, New York, (1998).

[Sce48] C.E. Shannon. (1948) A Mathematical Theory of Communication. Bell System Tech. J., 27:379-423,623-656.