Published in final edited form as:

IEEE Trans Inf Theory. 2008 ; 54(7): 3211–3216. doi:10.1109/TIT.2008.924668.

The Generalized Asymptotic Equipartition Property: Necessary and Sufficient Conditions

Matthew T. Harrison[Member, IEEE]

Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213 USA

(harrison@stat.cmu.edu)

Abstract

Suppose a string $X_1^n = (X_1, X_2, \dots, X_n)$ generated by a memoryless source $(X_n)_{n \geq 1}$ with distribution P is to be compressed with distortion no greater than $D \geq 0$, using a memoryless random codebook with distribution Q . The compression performance is determined by the “generalized asymptotic equipartition property” (AEP), which states that the probability of finding a D -close match between X_1^n and any given codeword Y_1^n , is approximately $2^{-nR(P, Q, D)}$, where the rate function $R(P, Q, D)$ can be expressed as an infimum of relative entropies. The main purpose here is to remove various restrictive assumptions on the validity of this result that have appeared in the recent literature. Necessary and sufficient conditions for the generalized AEP are provided in the general setting of abstract alphabets and unbounded distortion measures. All possible distortion levels $D \geq 0$ are considered; the source $(X_n)_{n \geq 1}$ can be stationary and ergodic; and the codebook distribution can have memory. Moreover, the behavior of the matching probability is precisely characterized, even when the generalized AEP is not valid. Natural characterizations of the rate function $R(P, Q, D)$ are established under equally general conditions.

Index Terms

Asymptotic equipartition property; data compression; large deviations; pattern-matching; random codebooks; rate-distortion theory

I. Introduction

Suppose a random string $X_1^n = (X_1, X_2, \dots, X_n)$ produced by a memoryless source $(X_n : n \geq 1)$ with distribution P on a source alphabet S , is to be compressed with distortion no more than some $D \geq 0$ with respect to a single-letter distortion measure $\rho(x, y)$.¹ The basic information-theoretic model for understanding the best performance that can be achieved is the study of random codebooks. If we generate memoryless random strings $Y_1^n = (Y_1, Y_2, \dots, Y_n)$ according to some distribution Q on the reproduction alphabet T , we would like to know how many such strings are needed so that, with high probability, we will be able to find at least one codeword Y_1^n that matches the source string X_1^n with distortion D or less. The crucial mathematical problem in answering this question is the evaluation of the probability that a given, typical X_1^n , will be D -close to a random Y_1^n . This probability can be expressed as

$$\text{Prob} \{Y_1^n \in B_n(X_1^n, D) \mid X_1^n\} = Q^n(B_n(X_1^n, D)) \quad (1)$$

where $B_n(X_1^n, D)$ denotes the “distortion ball” consisting of all reproduction strings that are within distortion D (or less) from X_1^n ; note that the matching probability in (1) is itself a random quantity, as it depends on the source string X_1^n .

The importance of evaluating (1) was already identified by Shannon in his classic study of rate-distortion theory [4], where he showed that, for the best codebook distribution $Q = Q^*$, we have

$$Q^n(B_n(X_1^n, D)) \approx 2^{-nR(P,D)} \quad (2)$$

where $R(P, D)$ is the rate-distortion function of the source.

The more general question of evaluating the matching probability (1) for distributions Q perhaps different from the optimal reproduction distribution Q^* , arises naturally in a variety of contexts, including problems in pattern-matching, mismatched codebooks, Lempel-Ziv compression, combinatorial optimization on random strings, and others; see, e.g., [5]–[13], and the review and references in [14]. In this case, Shannon’s estimate (2) is replaced by the so-called “generalized asymptotic equipartition property” (or generalized AEP), which states that

$$-n^{-1} \log Q^n(B_n(X_1^n, D)) \rightarrow R(P, Q, D) \quad \text{a.s.} \quad (3)$$

where “a.s.” stands for “almost surely” and refers to the random string X_1^n . The rate function $R(P, Q, D)$ is defined in a way that closely resembles the rate-distortion function definition,

$$R(P, Q, D) := \inf_W H(W \parallel P \times Q)$$

where $H(\cdot \parallel \cdot)$ denotes the relative entropy, and the infimum is over all (bivariate) probability distributions of random variables (U, V) with values on S and T , respectively, such that U has distribution P and the expected distortion $E[\rho(U, V)] \leq D$. (For a broad introduction to the generalized AEP, its applications and refinements, see [14] and the references therein.)

Although much is known about the generalized AEP and about $R(P, Q, D)$ [14], all known results are established under certain restrictive conditions. In particular, it is always assumed that $D_{\text{ave}}(P, Q) < \infty$ and that $D \neq D_{\min}(P, Q)$, where

$$\begin{aligned} D_{\text{ave}}(P, Q) &:= E_{(X,Y) \sim P \times Q} \rho(X, Y) \\ D_{\min}(P, Q) &:= \inf \{D : R(P, Q, D) < \infty\}. \end{aligned}$$

Often, the codebook distribution is required to be memoryless, and when it is not, it is further assumed that the distortion measure is bounded.

The main point of this paper is to remove these constraints, and to analyze which (if any) are essential for the validity of the generalized AEP. Our motivation is twofold. On one hand,

unnecessarily stringent conditions make the theoretical picture incomplete. On the other, there are applications which naturally require more general statements, such as universal lossy compression, where the source distribution is not known a priori [3].

Thus motivated, we give necessary and sufficient conditions for the generalized AEP in (3), and we precisely characterize the behavior of the matching probability in the pathological situations when the generalized AEP fails. Our results hold for *all* values of D , and they cover arbitrary abstract alphabets and distortion measures. We also allow the source to be stationary and ergodic, and the codebook distribution to have a certain amount of memory. We similarly extend the characterization of the rate function $R(P, Q, D)$ to the same level of generality. We show that it can *always* be written as a convex dual, and that a minimizer W in the definition of $R(P, Q, D)$ always exists (unless, of course, the infimum is taken over the empty set).

II. Notation and Assumptions

The source and reproduction alphabets S and T , along with their associated σ -algebras are assumed to be Borel spaces² and the distortion measure $\rho : S \times T \mapsto [0, \infty)$ is assumed to be product measurable. P and Q denote generic probability distributions on S and T , respectively, and let $X \sim P$ and $Y \sim Q$ be independent random variables (r.v.). Define³

$$\rho^Q(x) := \text{ess inf } \rho(x, Y).$$

If W is a probability distribution on $S \times T$, then we use W^S to denote the marginal distribution of W on S , and similarly for W^T . Define

$$R(P, Q, D) := \inf_{W \in W(P, D)} H(W \parallel P \times Q)$$

where the infimum is over the subset of probability distributions on $S \times T$ defined by

$$W(P, D) := \{W : W^S = P, E_{(U, V) \sim W} \rho(U, V) \leq D\}$$

and where $H(\mu \parallel \nu)$ denotes the relative entropy (in nats), namely,

$$H(\mu \parallel \nu) := \begin{cases} E_\mu \log \frac{d\mu}{d\nu}, & \text{if } \mu \ll \nu \\ \infty, & \text{otherwise.} \end{cases}$$

We use the convention that $\emptyset = +\infty$ and logarithms are base e . For independent r.v.s $X \sim P$ and $Y \sim Q$, define

$$\begin{aligned} \Lambda(P, Q, \lambda) &:= E \left[\log E[e^{\lambda \rho(X, Y)} | X] \right] \\ \Lambda^*(P, Q, D) &:= \sup_{\lambda \leq 0} [\lambda D - \Lambda(P, Q, \lambda)]. \end{aligned}$$

²Borel spaces include $\mathbb{R}^d$ as well as a large class of infinite-dimensional spaces, including Polish spaces. This assumption is made so that we can avoid certain pathologies while working with random sequences and conditional distributions [15].

³The essential infimum of a random variable η , is $\text{ess inf } \eta := \inf\{r : \text{Prob}\{\eta < r\} > 0\}$.

Consider now the product spaces S^n and T^n with the usual σ -algebras and with generic probability distributions P_n and Q_n , respectively. Let $X_1^n \sim P_n$ and $Y_1^n \sim Q_n$ be independent and let W_n be a probability distribution on $S^n \times T^n$. We use Q^n to denote the product distribution of Q on T^n . Define the sequence of single-letter distortion measures $(\rho_n : n \geq 1)$ by

$$\rho_n(x_1^n, y_1^n) := n^{-1} \sum_{k=1}^n \rho(x_k, y_k).$$

Since we are working with abstract alphabets, it makes sense to formally replace S, T, ρ, P, Q, W, X and Y with $S^n, T^n, \rho_n, P_n, Q_n, W_n, X_1^n$ and Y_1^n , respectively, in the above definitions of $\rho^n, W(P, D), R(P, Q, D), \Lambda(P, Q, D)$ and $\Lambda^*(P, Q, D)$ in order to define $\rho_n^{Q_n}, W_n(P_n, D), R_n(P_n, Q_n, D), \Lambda_n(P_n, Q_n, D)$ and $\Lambda_n^*(P_n, Q_n, D)$.

For each $x_1^n \in S^n$, let $P_{x_1^n}$ denote the empirical probability distribution of x_1^n on S and let $\delta_{x_1^n}$ denote the probability distribution on S^n that assigns probability 1 to the sequence x_1^n . Define

$$L_n(x_1^n, Q_n, D) := -n^{-1} \log Q_n(B_n(x_1^n, D))$$

where

$$B_n(x_1^n, D) := \{y_1^n \in T^n : \rho_n(x_1^n, y_1^n) \leq D\}$$

denotes the distortion ball of radius D at the point x_1^n .

Finally, consider the one-sided infinite sequence spaces S^∞ and T^∞ with the usual σ -algebras and with generic probability distributions $\mathbb{P}$ and $\mathbb{Q}$, respectively. Let $(X_n : n \geq 1)$ and $(Y_n : n \geq 1)$ be independent random sequences with distributions $\mathbb{P}$ and $\mathbb{Q}$, respectively. Let P_n and Q_n denote the marginal distributions of $\mathbb{P}$ and $\mathbb{Q}$ on S^n and T^n , respectively, so that $X_1^n \sim P_n$ and $Y_1^n \sim Q_n$ as before. Define

$$\begin{aligned} R_\infty(\mathbb{P}, \mathbb{Q}, D) &:= \limsup_{n \rightarrow \infty} n^{-1} R_n(P_n, Q_n, D) \\ \Lambda_\infty(\mathbb{P}, \mathbb{Q}, \lambda) &:= \limsup_{n \rightarrow \infty} n^{-1} \Lambda_n(P_n, Q_n, n\lambda) \\ \Lambda_\infty^*(\mathbb{P}, \mathbb{Q}, D) &:= \sup_{\lambda \leq 0} [\lambda D - \Lambda_\infty(\mathbb{P}, \mathbb{Q}, \lambda)]. \end{aligned}$$

It is straightforward to verify that when $\mathbb{P}$ is stationary with $P_1 = P$ and $\mathbb{Q}$ is memoryless with $Q_n = Q^n$, then $\Lambda_\infty(\mathbb{P}, \mathbb{Q}, \lambda) = \Lambda(P, Q, \lambda)$ and hence $\Lambda_\infty^*(\mathbb{P}, \mathbb{Q}, D) = \Lambda^*(P, Q, D)$.

III. Memoryless Codebook Distributions

Our goal is provide necessary and sufficient conditions for the generalized AEP in (3) and for the characterization of the limit as a convex dual. In this section we restrict attention to the case where the code-book distribution is memoryless and where the source distribution is stationary and ergodic.

Under appropriate regularity conditions, it is well known in the literature that $R(P, Q, D)$ can be expressed as a convex dual [9]. (See [14] for a review and further references). Our first result is that the regularity conditions are not necessary. (Proofs of all results are collected in Sections V and VI.)

Proposition 1

$R(P, Q, D) = \Lambda^*(P, Q, D)$ for all D . If $W(P, D)$ is not empty, then this set contains a W such that $R(P, Q, D) = H(W||P \times Q)$.

We separate our results about the generalized AEP into two theorems. The first theorem asserts that the generalized AEP always holds along some subsequence (of n 's). Consequently, the generalized AEP can only fail if the limit does not exist.⁴ The second theorem gives necessary and sufficient conditions for the existence of the limit and also characterizes the pathology. Taken together, they provide necessary and sufficient conditions for the generalized AEP.

Theorem 2

Suppose $\mathbb{P}$ is stationary and ergodic with $P_1 = P$. Then

$$\liminf_{n \rightarrow \infty} L_n(X_1^n, Q^n, D) = R(P, Q, D) \quad \text{a.s.} \quad (4)$$

for all D . The result also holds with $L_n(X_1^n, Q^n, D)$ replaced by $R(P_{X_1^n}, Q, D)$.

Theorem 3

Suppose $\mathbb{P}$ is stationary and ergodic with $P_1 = P$. Then $\text{Prob}\{\lim_n L_n(X_1^n, Q^n, D) \text{ exists}\} < 1$ if and only if $0 < D = D_{\min}(P, Q) < \infty$ and $R(P, Q, D) < \infty$ and $\rho^Q(X_1)$ is not a.s. constant. Furthermore, in this pathological situation

$$\text{Prob}\{L_n(X_1^n, Q^n, D) = \infty \text{ infinitely often}\} > 0 \quad (5a)$$

$$\text{Prob}\{L_n(X_1^n, Q^n, D) < \infty \text{ infinitely often}\} = 1 \quad (5b)$$

$$\lim_{m \rightarrow \infty} L_{N_m}(X_1^{N_m}, Q^{N_m}, D) = R(P, Q, D) \quad \text{a.s.} \quad (5c)$$

where $(N_m : m \geq 1)$ is the (a.s.) infinite random subsequence of $(n \geq 1)$ for which $L_n(X_1^n, Q^n, D)$ is finite. All of the above also hold with $L_n(X_1^n, Q^n, D)$ replaced by $R(P_{X_1^n}, Q, D)$.

The proofs below show that $(N_m : m \geq 1)$ can also be (a.s.) characterized as the random subsequence for which

⁴We are considering limits in the extended sense: if a sequence diverges to $+\infty$ (or to $-\infty$), then we still say that the limit exists and identify the limit as $+\infty$ (or as $-\infty$).

$$n^{-1} \sum_{k=1}^n \rho^Q(X_k) \leq D. \quad (6)$$

Note that $D_{\min}(P, Q) = E\rho^Q(X)$, whenever the former is finite.

A simple example that illustrates the pathology is the following: Let $(X_n : n \geq 1)$ be the sequence 1, 0, 1, 0, ... with probability 1/2 and the sequence 0, 1, 0, 1, ... with probability 1/2, namely, the binary, stationary, periodic Markov chain (which is ergodic). Let Q be the point mass at 0, let $\rho(x, y) := |x - y|$ and let $D = 1/2$. Note that $\rho^Q(X_1) = X_1$ is not constant, that $D = D_{\min}(P, Q) = 1/2$ and that $R(P, Q, D) = 0$ is finite. In the case when $X_1 = 0$,

$L_n(X_1^n, Q^n, D) = 0$ for all n . In the case when $X_1 = 1$, however, $L_{2n}(X_1^{2n}, Q^{2n}, D) = 0$ and $L_{2n-1}(X_1^{2n-1}, Q^{2n-1}, D) = \infty$ for all n .

IV. Extensions to the Case With Memory

Although the source $(X_n : n \geq 1)$ can have memory, the generalized AEP stated thus far is restricted to the case where the reproduction distribution is memoryless, that is, L_n is evaluated with a product measure Q^n . We relax this assumption here and replace it with a strong mixing condition, namely

$$C^{-1}Q(A)Q(B) \leq Q(A \cap B) \leq CQ(A)Q(B) \quad (7)$$

for some fixed $1 \leq C < \infty$ and all $A \in \sigma(Y_1^n)$ and $B \in \sigma(Y_{n+1}^\infty)$ and all n . Examples include the cases where Q is memoryless ($C = 1$) and where Q is a hidden Markov model (HMM) whose underlying Markov chain has a finite state space with all (strictly) positive transition probabilities.

Proposition 4

If $\mathbb{P}$ and Q are stationary and Q satisfies (7), then

$$R_\infty(\mathbb{P}, Q, D) = \lim_{n \rightarrow \infty} n^{-1} R_n(P_n, Q_n, D) = \Lambda_\infty^*(\mathbb{P}, Q, D)$$

for all D .

The proof shows that limit supremum in the definition of Λ_∞ is actually a limit and Proposition 1 implies that $R_n(P_n, Q_n, D) = \Lambda_n^*(P_n, Q_n, D)$. For the special case when S and T are finite and Q is a finite state Markov chain, a formula for $R_\infty(\mathbb{P}, Q, D)$ not involving limits was identified in [7].

Theorem 5

If Q is stationary and satisfies (7), then Theorems 2 and 3 remain valid when $Q^n, Q^{N_m}, R(P, Q, D)$ and $R(P_{X_1^n}, Q, D)$ are replaced by $Q_n, Q_{N_m}, R_\infty(\mathbb{P}, Q, D)$ and $n^{-1} R_n(\delta_{X_1^n}, Q_n, D)$, respectively.

The mixing conditions here are strong enough to ensure that

$$D_{\min}(P, Q) = D_{\min}(\mathbb{P}, \mathbb{Q}) := \inf \{D: R_{\infty}(\mathbb{P}, \mathbb{Q}, D) < \infty\} \quad (8)$$

and that

$$\rho_n^{Q_n}(x_1^n) = n^{-1} \sum_{k=1}^n \rho^{Q_n}(x_k) \quad (9)$$

which is why the results for memory can still be in terms of $D_{\min}(P, Q)$ and ρ^Q . Extending Theorem 3 to situations where these do not hold seems difficult. The generalized AEP for $\mathbb{Q}$ with memory can also be found in [7], [12], [14], [16], in some cases under more general mixing conditions, but always for bounded distortion measure ρ and for $D \neq D_{\min}(\mathbb{P}, \mathbb{Q})$.

V. Proofs: Standard Cases

Owing to space constraints, the proofs presented here focus primarily on the case $D = D_{\min}(P, Q)$, which has received little or no attention in the literature. We also provide some justification for removing the standard assumption that $D_{\text{ave}}(P, Q) < \infty$. This is a moment condition on ρ and it is assumed in all previous treatments of the subject. More detailed proofs, including measurability issues, can be found in a longer version of this paper [2] and the technical report that preceded it [1].

When $\mathbb{Q}$ is stationary and satisfies (7), then

$$\left| \log Q_n(A_n) + \log Q_m(A_m) - \log Q_{n+m}(A_n \times A_m) \right| \leq \log C \quad (10)$$

and

$$\left| \Lambda_n(\delta_{x_1^n}, Q_n, n\lambda) + \Lambda_m(\delta_{x_{n+1}^{n+m}}, Q_m, m\lambda) - \Lambda_{n+m}(\delta_{x_1^{n+m}}, Q_{n+m}, (n+m)\lambda) \right| \leq \log C. \quad (11)$$

These two bounds make it relatively straightforward to extend the memoryless case to the case with memory. Replacing x with X in (11), taking expectations and using well-known facts about subadditive sequences (cf., [15, Lemma 10.21]) gives Proposition 4. Equation (10) combined with a blocking argument can be used to extend the memoryless generalized AEP to the case with memory. The details are tedious and analogous to existing arguments in the literature. We refer the interested reader to [2] and [1]. Henceforth, except for the special case $D = D_{\min}(P, Q)$ and our remarks about the lower bound below, we consider only the memoryless setting.

It is convenient in the proofs to temporarily redefine

$$D_{\min}(P, Q) := \inf \{D: \Lambda^*(P, Q, D) < \infty\}.$$

Proposition 1, once established, shows that $D_{\min}$ agrees with the definition in the text.

A. The Lower Bound and Replacing L_n With R_n

For each $\lambda \leq 0$, the exponential Chebyshev's inequality gives

$$\begin{aligned} nL_n(X_1^n, Q_n, D) &\geq \Lambda_n^*(\delta_{X_1^n}, Q_n, D) \\ &\geq n\lambda D - \Lambda_n(\delta_{X_1^n}, Q_n, n\lambda). \end{aligned}$$

Taking limits, using (11) to justify the sub-additive ergodic theorem [15, Th. 10.22], and then optimizing over $\lambda \leq 0$ gives

$$\liminf_{n \rightarrow \infty} L_n(X_1^n, Q_n, D) \geq \Lambda_\infty^*(\mathbb{P}, \mathbb{Q}, D) \quad \text{a.s.} \quad (12)$$

This holds for all D . Combined with the previous bound this also shows that whenever $L_n(X_1^n, Q_n, D)$ converges to $\Lambda_\infty^*(\mathbb{P}, \mathbb{Q}, D)$, $n^{-1}\Lambda_n^*(\delta_{X_1^n}, Q_n, D)$ converges to the same limit. Furthermore, using Lemma 7 below it is straightforward to show that Λ_n^* must be infinite when L_n is infinite. Using Propositions 1 and 4 to equate R_n and Λ_n^* , these remarks justify our claims that L_n can be replaced by $n^{-1}R_n$ in each of the theorems.

There is also a lower bound for Proposition 1.

$$H(W \parallel P \times Q) \geq \Lambda^*(P, Q, D) \text{ for all } W \in W(P, D). \quad (13)$$

This only makes sense if $W(P, D) \neq \emptyset$. A simple proof can be found in [14, Th. 2]. Note that the proof there does not make use of any regularity assumptions, other than the existence of a regular conditional distribution $W(y|x)$.

B. The Upper Bound: $D < D_{\min}(P, Q)$ or $D > D_{\text{ave}}(P, Q)$

The upper bound is

$$\limsup_{n \rightarrow \infty} L_n(X_1^n, Q_n, D) \leq \Lambda_\infty^*(\mathbb{P}, \mathbb{Q}, D) \quad \text{a.s.} \quad (14)$$

If $D < D_{\min}(P, Q)$ then the upper bound is trivial since the right side is $+\infty$. If $D > D_{\text{ave}}(P, Q)$ (the latter of which must then be finite), then the left side converges to 0 (Chebyshev's inequality and the ergodic theorem) and again the upper bound must hold.

For Proposition 1, the upper bound is

$$H(W \parallel P \times Q) \leq \Lambda^*(P, Q, D) \text{ for some } W \in W(P, D). \quad (15)$$

Again, this only makes sense if $W(P, D) \neq \emptyset$. If $D < D_{\min}(P, Q)$, then any $W \in W(P, D)$ trivially satisfies (15). If $D \geq D_{\text{ave}}(P, Q)$, then $W = P \times Q \in W(P, D)$ trivially satisfies (15).

C. The Upper Bound: $D_{\min}(P, Q) < D \leq D_{\text{ave}}(P, Q)$

This is the case considered in the literature, at least under the assumption that $D_{\text{ave}}(P, Q) < \infty$. Under appropriate regularity conditions, the upper bound (14) is an immediate consequence of the one-sided large deviations theorem in [17] applied to the random

variables $W_n := -n\rho_n(x_1^n, Y_1^n)$ and the constants $a_n := -D$ for a typical, fixed realization $(x_n : n \geq 1)$ of the source. See [14, Th. 1] for an overview of this technique using the Gärtner-

Ellis theorem [18, Th. V.6], which is a generalization of [17]. The required regularity conditions are placed on $\Lambda_\infty(\mathbf{P}, \mathbf{Q}, \lambda)$.

For the memoryless codebook case, $\Lambda_\infty(\mathbf{P}, \mathbf{Q}, \lambda) = \Lambda(\mathbf{P}, \mathbf{Q}, \lambda)$ and the regularity conditions needed in [17] are given by the following Lemma.

Lemma 6—If $D_{\min}(\mathbf{P}, \mathbf{Q}) < \infty$, then $\Lambda(\lambda) := \Lambda(\mathbf{P}, \mathbf{Q}, \lambda)$ is convex and continuously differentiable on $(-\infty, 0)$ with $\lim_{\lambda \downarrow -\infty} \Lambda'(\lambda) = E(\rho^Q(X)) = D_{\min}(\mathbf{P}, \mathbf{Q})$ and $\lim_{\lambda \uparrow 0} \Lambda'(\lambda) = D_{\text{ave}}(\mathbf{P}, \mathbf{Q})$ (which may be infinite). If $D_{\min}(\mathbf{P}, \mathbf{Q}) < D_{\text{ave}}(\mathbf{P}, \mathbf{Q})$, then $\Lambda(\lambda)$ is strictly convex on $(-\infty, 0)$.

Remarks—The regularity assumption $D_{\text{ave}}(\mathbf{P}, \mathbf{Q}) < \infty$ is used in the literature to establish these properties of Λ . It can be replaced by the much weaker assumption that $D_{\min}(\mathbf{P}, \mathbf{Q}) < \infty$. If Lemma 6 were known to be true for Λ_∞ , then the more general case with memory would also follow immediately from [17]. Establishing the strict convexity of Λ_∞ seems challenging, however, and we resorted to a blocking argument as mentioned above to derive the general case from the memoryless case.

Proof—Let Z be a real-valued, nonnegative random variable. Define $\Gamma(\lambda) := \log Ee^{\lambda Z}$. It is well-known [19, Lem. 2.2.5, Ex. 2.2.24], [9], [14] that Γ is nondecreasing, convex and C^∞ on $(-\infty, 0)$ with $\Gamma'(\lambda) \downarrow \text{ess inf } Z$ as $\lambda \downarrow -\infty$ and $\Gamma'(\lambda) \uparrow EZ$ (which may be infinite) as $\lambda \uparrow 0$. Furthermore, if $\text{ess inf } Z < EZ$ then Γ is strictly convex on $(-\infty, 0)$.

Now define $\Gamma(x, \lambda) := \log Ee^{\lambda \rho(x, Y)}$. Using $Z := \rho(x, Y)$, we see that $\Gamma(x, \lambda)$ has all the above properties of $\Gamma(\lambda)$ for each fixed x . Since $\Lambda(\lambda) = E\Gamma(X, \lambda)$, which must be finite on $(-\infty, 0]$ if $D_{\min} < \infty$, we need only show that these properties are preserved by expectation. Convexity and strict convexity are immediately preserved. Preserving properties of the derivative requires some justification. Moving the expectation inside the derivative is justified by convexity and the monotone convergence theorem for the left hand derivative (including $\lambda \uparrow 0$) and then finiteness and the dominated convergence theorem for the right hand derivative (including $\lambda \downarrow -\infty$). The fact that $D_{\min} = \lim_{\lambda \downarrow -\infty} \Lambda'(\lambda)$ is a well-known property of Fenchel–Legendre transforms.

Turning to Proposition 1, if $D_{\min}(\mathbf{P}, \mathbf{Q}) < D < D_{\text{ave}}(\mathbf{P}, \mathbf{Q})$, a proof of the upper bound (15) can be found in [14, Th. 2]. The achieving $W \in \mathcal{W}(\mathbf{P}, D)$ is identified as

$$\frac{dW}{d(P \times Q)}(x, y) := \frac{e^{\lambda_D \rho(x, y)}}{Ee^{\lambda_D \rho(x, Y)}}$$

where $\lambda_D < 0$ is uniquely chosen so that $\Lambda^*(\mathbf{P}, \mathbf{Q}, D) = \lambda_D D - \Lambda(\mathbf{P}, \mathbf{Q}, \lambda_D)$. Note that the properties of Λ given in Lemma 6 are sufficient regularity conditions for ensuring that λ_D exists and has the desired properties. See [19, Lemma 2.2.5] and [20, Th. 23.5, Corollary 23.5.1, Th. 25.1].

VI. Proofs: Nonstandard Cases

Here we consider the special case when $D = D_{\min} := D_{\min}(\mathbf{P}, \mathbf{Q}) < \infty$. Lemma 6 is applicable, so

$$D_{\min} = E(\rho^Q(X)).$$

Unlike most of the previous section, we explicitly treat the case of code-book distributions with memory. In particular, we assume that $\mathbb{Q}$ is stationary and satisfies (7). The bounds in (10) and (11) easily give (9) and

$$D_{\min} = \inf \{D : \Lambda_{\infty}^*(\mathbb{P}, \mathbb{Q}, D) < \infty\}$$

which means that $\Lambda^*(D_{\min}) := \Lambda^*(P, Q, D_{\min})$ is finite (infinite) exactly when $\Lambda_{\infty}^*(D_{\min}) := \Lambda_{\infty}^*(\mathbb{P}, \mathbb{Q}, D_{\min})$ is finite (infinite). The lower bounds in (12) and (13) are also valid. These lower bounds address the upper bounds in (14) and (15) in the case $\Lambda^*(D_{\min}) = \infty$. So in this section we will restrict attention to $\Lambda^*(D_{\min}) < \infty$. The proofs make use of the sets

$$A_n(x_1^n) := \{y_1^n \in T^n : \rho_n(x_1^n, y_1^n) = \rho_n^{Q_n}(x_1^n)\}.$$

We use $A(x) := A_1(x) = \{y : \rho(x, y) = \rho^Q(x)\}$ and we use the notation $1\{B\}$ to denote the indicator function of the event B .

Lemma 7

If $\Lambda^*(D_{\min}) < \infty$, then

$$\Lambda^*(D_{\min}) = -E[\log Q(A(X))] = H(W \| P \times Q)$$

for $W \in W(P, D_{\min})$ defined by

$$\frac{dW}{d(P \times Q)}(x, y) := \frac{1\{y \in A(x)\}}{Q(A(x))}.$$

Proof

Define

$$\tilde{\rho}(x, y) := \max\{\rho(x, y) - \rho^Q(x), 0\}$$

so that $\tilde{\rho}$ is a valid distortion measure and so that

$$Q(A(x)) = Q\left(\left\{y : \tilde{\rho}(x, y) = 0\right\}\right).$$

Let $\tilde{\Lambda}$ be defined analogously to Λ , except with $\tilde{\rho}$ instead of ρ . We have $\Lambda(\lambda) = \tilde{\Lambda}(\lambda) + \lambda D_{\min}$ so that

$$\begin{aligned}
& \Lambda^*(D_{\min}) \\
&= \sup_{\lambda \leq 0} \left[\lambda D_{\min} - \tilde{\Lambda}(\lambda) - \lambda D_{\min} \right] \\
&= \lim_{\lambda \downarrow -\infty} -\tilde{\Lambda}(\lambda) = - \lim_{\lambda \downarrow -\infty} E \left[\log E \left[e^{\lambda \tilde{\rho}(X,Y)} \middle| X \right] \right] \\
&= - E \left[\log E \left[\lim_{\lambda \downarrow -\infty} e^{\lambda \tilde{\rho}(X,Y)} \middle| X \right] \right] \\
&= - E \left[\log E \left[1 \left\{ \tilde{\rho}(X,Y) = 0 \right\} \middle| X \right] \right] \\
&= - E \left[\log Q \left(\left\{ y: \tilde{\rho}(X,y) = 0 \right\} \right) \right] = - E \left[\log Q(A(X)) \right].
\end{aligned}$$

We moved the limit inside the expectations using first the monotone convergence theorem and then the dominated convergence theorem. This also shows that the denominator in the definition of W is strictly positive P -a.s., so W is well-defined.

It is easy to see that $W \in W(P, D_{\min})$ and that

$$\begin{aligned}
& H(W \| P \times Q) \\
&= E \left[\frac{dW}{d(P \times Q)}(X, Y) \log \frac{dW}{d(P \times Q)}(X, Y) \right] \\
&= E \left[\frac{1\{Y \in A(X)\}}{Q(A(X))} \log 1\{Y \in A(X)\} \right] - E \left[\frac{1\{Y \in A(X)\}}{Q(A(X))} \log Q(A(X)) \right] \\
&= 0 - E \left[\log Q(A(X)) \right] = \Lambda^*(D_{\min})
\end{aligned}$$

which completes the proof.

Lemma 7 gives the upper bound (15) and completes the proof of Proposition 1 for all possible values of D . We now address the upper bound (14) for the generalized AEP. Because of (9)

$$Q_{n+m}(A_{n+m}(x_1^{n+m})) = Q_{n+m}(A_n(x_1^n) \times A_m(x_{n+1}^{n+m}))$$

and we can use (10) in conjunction with the subadditive ergodic theorem [15, Th. 10.22] to conclude that

$$\begin{aligned}
& \lim_{n \rightarrow \infty} -\frac{1}{n} \log Q_n(A_n(X_1^n)) \\
& \stackrel{\text{a.s.}}{=} \lim_{n \rightarrow \infty} -\frac{1}{n} E[-\log Q_n(A_n(X_1^n))] \\
& = \lim_{n \rightarrow \infty} \frac{1}{n} \Lambda_n^*(P_n, Q_n, D_{\min}) = \Lambda_\infty^*(\mathbb{P}, \mathbb{Q}, D_{\min}).
\end{aligned} \tag{16}$$

The limit in the subadditive ergodic theorem is deterministic because it is shift invariant and the source is ergodic. The second equality comes from Lemma 7 and the third from Propositions 1 and 4. Note that if $\rho^Q(X)$ is a.s. constant, then $Q_n(A_n(X_1^n)) \stackrel{\text{a.s.}}{=} Q_n(B_n(X_1^n, D_{\min}))$ and (16) gives the upper bound (14).

Now suppose $\rho^Q(X_1)$ is not a.s. constant (and $D = D_{\min}$ and $\Lambda^*(D_{\min}) < \infty$). This is the only pathological situation where the upper bound does not hold. Our analysis makes use of recurrence properties for random walks with stationary and ergodic increments. What we need is summarized in the following lemma. The notation “i.o.” means “infinitely often.”

Lemma 8

Let $(U_n : n \geq 1)$ be a real-valued stationary and ergodic process and define $W_n := \sum_{k=1}^n U_k$ for $n \geq 1$. If $EU_1 = 0$ and $\text{Prob}\{U_1 \neq 0\} > 0$, then $\text{Prob}\{W_n > 0 \text{ i.o.}\} > 0$ and $\text{Prob}\{W_n \geq 0 \text{ i.o.}\} = 1$.

Proof

Define $W_0 := 0$ so that $(W_n : n \geq 0)$ is a random walk with stationary and ergodic increments [21]. [22] shows that $\{\liminf_n^{-1} W_n > 0\}$ and $\{W_n \rightarrow \infty\}$ differ by a null set. The ergodic theorem gives $\text{Prob}\{n^{-1} W_n \rightarrow 0\} = 1$, so $\text{Prob}\{W_n \rightarrow \infty\} = 0$. Similarly, by considering the process $-W_n$, we see that $\text{Prob}\{W_n \rightarrow -\infty\} = 0$.

Now $\{|W_n| \rightarrow \infty\}$ is invariant and must have probability 0 or 1. If it has probability 1, then since we cannot have $W_n \rightarrow \infty$ or $W_n \rightarrow -\infty$ we must have W_n oscillating between increasingly larger positive and negative values, which means $\text{Prob}\{W_n > 0 \text{ i.o.}\} = 1$ and completes the proof.

Suppose $\text{Prob}\{|W_n| \rightarrow \infty\} = 0$. Define

$$N(A) := \sum_{n \geq 0} 1\{W_n \in A\}, \quad A \subset \mathbb{R}$$

to be the number of times the random walk visits the set A . [21, Corollary 2.3.4] shows that either $N(J) < \infty$ a.s. for all bounded intervals J or $\{N(J)=0\} \cup \{N(J)=\infty\}$ has probability 1 for all intervals J (open or closed, bounded or unbounded, but not a single point). By assumption $|W_n| \not\rightarrow \infty$, so we can rule out the first possibility. Since $\text{Prob}\{W_0 = 0\} = 1$, we see that for any interval J containing $\{0\}$ we must have $\text{Prob}\{N(J) = \infty\} = 1$. In particular, taking $J := [0, \infty)$ shows that $\text{Prob}\{W_n \geq 0 \text{ i.o.}\} = 1$. Similarly, taking $J := (0, \infty)$ shows that

$$\begin{aligned} & \text{Prob}\{W_n > 0 \text{ i.o.}\} \\ &= \text{Prob}\{N(J) = \infty\} \\ &= \text{Prob}\{N(J) > 0\} \geq \text{Prob}\{U_1 > 0\} > 0. \end{aligned}$$

Returning to the main argument

$$\begin{aligned} & L_n(X_1^n, Q_n, D_{\min}) \\ & \geq -n^{-1} \log Q_n \left\{ y_1^n : n^{-1} \sum_{k=1}^n \rho^Q(X_k) \leq D_{\min} \right\} \\ & = \begin{cases} 0, & \text{if } \sum_{k=1}^n \rho^Q(X_k) \leq n D_{\min} \\ \infty, & \text{if } \sum_{k=1}^n \rho^Q(X_k) > n D_{\min} \end{cases} \\ & = \begin{cases} 0, & \text{if } W_n \leq 0 \\ \infty, & \text{if } W_n > 0 \end{cases} \end{aligned} \tag{17}$$

where $W_n := \sum_{k=1}^n (\rho^Q(X_k) - D_{\min})$. Lemma 8 shows that $\text{Prob}\{W_n > 0 \text{ i.o.}\} > 0$. This and (17) prove (5a).

Lemma 8 also shows that $\text{Prob}\{W_n \leq 0 \text{ i.o.}\} = 1$. Let $(N_m : m \geq 1)$ be the (a.s.) infinite, random subsequence of $(n \geq 1)$ such that $W_n \leq 0$. Note that

$$N_m^{-1} \sum_{k=1}^{N_m} \rho^Q(X_k) \leq D_{\min}$$

so

$$\begin{aligned} & L_{N_m}(X_1^{N_m}, Q_{N_m}, D_{\min}) \\ & \leq L_{N_m}(X_1^{N_m}, Q_{N_m}, N_m^{-1} \sum_{k=1}^{N_m} \rho^Q(X_k)) \\ & = -N_m^{-1} \log Q_{N_m}(A_{N_m}(X_1^{N_m})). \end{aligned} \quad (18)$$

Now, the final expression in (18) is a.s. finite because

$E[-\log Q_n(A_n(X_1^n))] = \Lambda_n^*(P_n, Q_n, D_{\min}) < \infty$. This proves (5b) and shows that $(N_m : m \geq 1)$ satisfies the claims of the theorem, including (6). Letting $m \rightarrow \infty$ in (18) and using (16) gives (5c), the upper bound along the sequence $(N_m : m \geq 1)$. Note that it also shows that the $\liminf_n L_n$ is a.s. $\Lambda_\infty^*(D_{\min})$ even in this pathological case, which proves Theorem 2 and its generalization in Theorem 5.

Acknowledgments

The author would like to thank I. Kontoyiannis, M. Madiman, and two anonymous reviewers for many useful comments and corrections, and I. Kontoyiannis for invaluable advice and for suggesting the problems that led to this correspondence.

This work was supported in part by a National Defense Science and Engineering Graduate Fellowship.

References

1. Harrison, M. Brown Univ, Div Appl Math. Providence, RI, APPTS #03-3: 2003. The First Order Asymptotics of Waiting Times Between Stationary Processes Under Nonstandard Conditions.
2. Harrison, MT. The Generalized Asymptotic Equipartition Property: Necessary and Sufficient Conditions 2007 [Online]. Available: <http://arxiv.org/abs/0711.2666>
3. Harrison, M.; Kontoyiannis, I. Proc 40th Ann Allerton Conf Commun Contr Comput. Allerton, IL: Oct. 2002 Maximum likelihood estimation for lossy data compression; p. 596-604.
4. Shannon, C. Coding theorems for a discrete source with a fidelity criterion. In: Slepian, D., editor. IRE Nat Conv Rec. Vol. 4. New York: IEEE; 1959. p. 142-163. Key Papers in the Develop. Inf. Theory
5. Zhang Z, Yang E-h, Wei VK. The redundancy of source coding with a fidelity criterion – Part I: Known statistics. IEEE Trans Inf Theory. Jan.1997 43:71–91.
6. Luczak T, Szpankowski W. A suboptimal lossy data compression based on approximate pattern matching. IEEE Trans Inf Theory. Sep.1997 43:1439–1451.
7. Yang, E-h; Kieffer, J. On the performance of data compression algorithms based upon string matching. IEEE Trans Inf Theory. Jan.1998 44:47–65.
8. Kontoyiannis I. An implementable lossy version of the lempel-ziv algorithm – Part I: Optimality for memoryless sources. IEEE Trans Inf Theory. Nov.1999 45:2293–2305.
9. Yang, E-h; Zhang, Z. On the redundancy of lossy source coding with abstract alphabets. IEEE Trans Inf Theory. May.1999 45:1092–1110.
10. Dembo A, Kontoyiannis I. The asymptotics of waiting times between stationary processes, allowing distortion. Ann Appl Prob. May.1999 9:413–429.
11. Yang, E-h; Zhang, Z. The shortest common superstring problem: Average case analysis for both exact and approximate matching. IEEE Trans Inf Theory. 1999; 45:1867–1886.

12. Chi Z. The first-order asymptotic of waiting times with distortion between stationary processes. *IEEE Trans Inf Theory*. Jan.2001 47:338–347.
13. Szpankowski, W. *Average Case Analysis of Algorithms on Sequences*. New York: Wiley; 2001.
14. Dembo A, Kontoyiannis I. Source coding, large deviations, and approximate pattern matching. *IEEE Trans Inf Theory*. Jun.2002 48:1590–1615.
15. Kallenberg, O. *Foundations of Modern Probability*. 2. New York: Springer; 2002.
16. Chi Z. Stochastic sub-additivity approach to the conditional large deviation principle. *Ann Prob*. 2001; 29(3):1303–1328.
17. Plachky D, Steinebach J. A theorem about probabilities of large deviations with an application to queuing theory. *Period Math Hung*. 1975; 6(4):343–345.
18. den Hollander, F. *Large Deviations*. Providence, RI: American Math Soc; 2000.
19. Dembo, A.; Zeitouni, O. *Large Deviations Techniques and Applications*. 2. New York: Springer; 1998.
20. Rockafellar, RT. *Convex Analysis*. Princeton, NJ: Princeton Univ Press; 1970.
21. Berbee, H. *Random Walks With Stationary Increments and Renewal Theory*. Vol. 112. Amsterdam, The Netherlands: Mathematisch Centrum; 1979. Mathematical Centre Tracts
22. Kesten H. Sums of stationary sequences cannot grow slower than linearly. *Proc AMS*. May; 1975 49(1):205–211.