A Universal Decoder relative to a given Family of Metrics

Nir Elkayam and Meir Feder
Department of Electrical Engineering - Systems
Tel-Aviv University, Israel
Email: nirelkayam@post.tau.ac.il, meir@eng.tau.ac.il

Abstract

Given a family of decoding metrics, we describe a general universal decoder relative to this family in a random coding setting. The penalty for using this universal decoder is computed. The universal metric is constructed as follows. For each metric, a canonical metric is defined and conditions for the random coding prior to be normal are given. A sub-exponential set of canonical metrics can be merged into a single universal optimal metric.

I. INTRODUCTION

Metric based decoders are usually used to decode codes in digital communication. Typically, the metric tries to capture the most likely codeword that was transmitted. When the channel over which the code word was transmitted is known at the receiver, the Maximum Likelihood (ML) decoder is used for optimal performance (in terms of block error rate). When the channel is not known at the receiver, other solutions are required. Usually some sort of assumption is needed, for example, that the channel \( W_\theta(y|x) \) belongs to some family of channels indexed by \( \theta \in \Theta \). The family of channels can have some measure defined over it (Bayesian approach), e.g., fading channel, or the channel is one of the possible channels (deterministic approach), e.g., discrete memoryless channels (DMCs). The performance in the first case can be measured over the channel realization, and in the latter per specific channel. A Universal optimal decoder with respect to a given family of channels, means that the performance when using the universal decoder is not much worse than when the optimal decoder is used; specifically, the error exponent is not smaller than that of the optimal decoder.

In [1], Goppa offered the maximum mutual information (MMI) decoder, which decides in favor of the codeword having the maximum empirical mutual information with the channel output sequence. Other universal decoders have been suggested over the years for ISI channel [2], [3], Interference channels [4], finite state channels [5], [6], Individual channels [7], and [8]. In [9], Feder and Lapidoth provide a fairly general construction of a universal decoder for a given class of channels. Their construction has two key points. The first is the use of uniform prior, which means that all code words have the same probability, which allows the use of the merged decoder. The second is the definition of a separable family, which means that there is a countable set of channels that can be used as representative, in the sense that if the performance of the universal decoder good over these channels, then they would be good over the whole family of channels.

In [10], Merhav proved universality relative to a given set of decoding metric over arbitrary channels, where optimality means that the performance of the universal decoder is not much worse than the performance when any other metric from the family is used. However, the definition of the universal decoder in [10] assumes that there exists an equivalent of “types” that is a set of input and output sequences of the channel for which the metrics of the family provides the same value. The universal decoder that is based on ties in the metric thus does not seem to be the most general.

In this paper we generalize the construction in [10], which, in some sense also generalizes Feder and Lapidoths universal decoder, [9]. The quantity \( p_{e,m,x,y} \), which is the error probability using the metric \( m \), given that \( x \) was transmitted and \( y \) was received, is defined. This quantity plays an important role in the computation of the average error rate and can be used to compare the performance between different metrics. Based on \( p_{e,m,x,y} \) we will define for a metric \( m \) a “canonical metric” \( m(x, y) = -\log p_{e,m,x,y} \), which is equivalent to \( m \). The canonical metrics (or the associated error probabilities \( p_{e,m,x,y} \)) serve in turn to define a universal decoder relative to a set of metrics, called the Generalized Minimum Error Test. We derive the penalty of using this universal metric, and show when this metric is optimal (in the error exponent sense).

The definition of the canonical metric is related to the definition in [11] of empirical rate function. This rate function, which is a function of the transmitted word \( x \) and the received word \( y \), was defined in an attempt to remove the probabilistic assumption in communication and get a rate function that fits the individual sequences of input and output. Specifically, after we introduce the notion of normal prior, we show that for every metric \( m \) there exists a “tight” empirical rate function, i.e., there exists an equivalent metric \( m(x, y) = \log \left( \frac{P(x|y)}{Q(x)} \right) \), where \( Q(x) \) is the prior distribution and \( m(x, y) \approx -\log p_{e,m,x,y} \). Here \( \approx \) means equal up to a sub-exponential factor. A set of countable “tight” rate functions can be merged into a single universal decoding metric that is optimal with respect to all the decoding metrics in the set. This is the generalization of the merged decoder. When the family of metrics is separable, an optimal universal decoding exists for this family.
II. Notation

This paper uses bold lower case letters (e.g., \( x \)) to denote a particular value of the corresponding random variable denoted in capital letters (e.g., \( X \)). Calligraphic fonts (e.g., \( \mathcal{X} \)) represent a set. Throughout this paper log will be defined to the base 2 unless otherwise indicated. \( P \{ A \} \) will denote the probability of the event \( A \). We will use the \( o(n) \) will be used to indicate a sequence of the form: \( n \cdot \lambda_n \) where \( \lambda_n \to 0 \) and \( o(1) \) is a sequence that goes to 0. A sequence of positive numbers \( \chi_n \) is said to be subexponential if \( \lim_{n \to \infty} \frac{1}{n} \log(\chi_n) = 0 \). With the \( o(n) \) notation this is just \( \chi_n = 2^{o(n)} \)

III. Preliminaries

A rate-\( R \) blocklength-\( n \), random code with prior \( Q_n(\mathbf{x}) \) is:

\[
\mathcal{C} = \{ \mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(2^{nR}) \} \subset \mathcal{X}^n,
\]

where each codeword \( \mathbf{x}(i) \) is drawn independently according to the distribution \( Q_n(\mathbf{x}) \). The encoder maps the input message \( k \), drawn uniformly over the set \( \{1, \ldots, 2^{nR}\} \) to the appropriate codeword, i.e., \( \mathbf{x}(k) \). A decoding metric is a function \( m : \mathcal{X}^n \times \mathcal{Y}^n \to \mathbb{R} \). The decoder associated with the decoding metric \( m \) is the function \( D_m : \mathcal{Y}^n \to \{1, \ldots, 2^{nR}\} \) defined by:

\[
D_m(\mathbf{y}) = \hat{i} = \arg \max_{1 \leq i \leq 2^n} (m(\mathbf{x}(i), \mathbf{y})).
\]

**Definition 1.** The average error probability associated with the decoder \( D_m \) over the channel \( W_n(\mathbf{y}|\mathbf{x}) \) is denoted by \( \hat{P}_{e,m,W}(R) \). The error probability captures the randomness in the codebook selection, the transmitted message, and the channel output.

**Remark 1.** When the decoding metric \( m(\mathbf{x}, \mathbf{y}) = W_n(\mathbf{y}|\mathbf{x}) \), the error probability is minimized and the decoder is called the Maximum Likelihood decoder. The mismatch case is when the metric is not matched to the channel and may result in performance degradation.

**Definition 2.** For a family of decoding metrics:

\[
\mathcal{M}_n = \{ m_\theta(\mathbf{x}, \mathbf{y}) : \theta \in \Theta_n, \mathbf{x} \in \mathcal{X}^n, \mathbf{y} \in \mathcal{Y}^n \}.
\]

Denote:

\[
\hat{P}_{e,\mathcal{M}_n,W_n}(R) = \min_{\theta \in \Theta_n} \hat{P}_{e,m_\theta,W_n}(R),
\]

the minimum error probability over the channel \( W_n(\mathbf{y}|\mathbf{x}) \) when the best metric matched for this channel is used.

**Definition 3.** A sequence of decoding metrics \( u_n \) (independent of \( \theta \)) and their associated decoders is called universal. The sequence of universal decoding metrics \( u_n \) (and their decoders) is optimal (in the error exponent sense) with respect to the family \( \mathcal{M}_n \) if:

\[
\lim_{n \to \infty} \frac{1}{n} \log \left( \frac{\hat{P}_{e,u_n,W_n(\mathbf{y}|\mathbf{x})}(R)}{\hat{P}_{e,\mathcal{M}_n,W_n(\mathbf{y}|\mathbf{x})}(R)} \right) \leq 0
\]

for any channel \( W(\mathbf{y}|\mathbf{x}) \).

Consider the metric \( m(\mathbf{x}, \mathbf{y}) \). The pairwise error given that \( \mathbf{x} \) was transmitted and \( \mathbf{y} \) was received is given by:

\[
p_{e,m,x,y} = Q_n(m(\mathbf{X}, \mathbf{y}) \geq m(\mathbf{x}, \mathbf{y})).
\]

The probability given by (5) is for a single (random) code word to beat the transmitted word. Notice that in case of a tie, we are assuming that an error has occurred.

The term \( p_{e,m,x,y} \) can be used to compute the average error probability for any rate \( R \) as can be seen in the following lemma.

**Lemma 1.**

\[
\frac{1}{2} \leq \mathbb{E} \left( \frac{\hat{P}_{e,m,W}(R)}{\min(1, 2^{nR} \cdot p_{e,m,x,y})} \right) \leq 1
\]

where the expectation is with respect to \( Q(x)W(y|x) \).

This lemma is given in [10]. The upper bound follows from the union bound. The lower bound follows from a lemma proved by Shulman in [12] Lemma A.2], namely, that the union clipped to unity is tight up to factor 2 (for pairwise independent events).
Definition 4. We say that the metric $u_n$ dominate $m_n$ if for all $x, y$:

$$p_{e,u_n,x,y} \leq p_{e,m_n,x,y} \cdot 2^n \lambda_n$$

with $\lambda_n \to 0$.

The following lemma, which also appears in [10] follows from lemma (1):

Lemma 2.

$$\P_{e,u_n,W}(R) \leq 2 \cdot 2^n \lambda_n \cdot \P_{e,m_n,W}(R)$$

(8)

$$\P_{e,u_n,W}(R) \leq 2 \cdot \P_{e,m_n,W}(R + \lambda_n).$$

(9)

In particular, if metric $u_n$ dominates $m_n$, then the average error exponent when using the metric $u_n$ is not worse than using the metric $m_n$.

Obviously, an optimal universal decoder is a metric that dominates all the metrics in the family (3).

IV. The Generalized Minimum Error Test

In this section we propose our universal decoder, the Generalized Minimum Error Test (GMET). The decoder estimates the metric that brings to minimum the pairwise error.

Definition 5. For the given family of metrics (3), let:

$$2^{-n \cdot U_n(x,y)} \triangleq \min_{\theta \in \Theta_n} p_{e,m_\theta,x,y}.$$  

(10)

$U_n(x,y)$ is the GMET. Let:

$$K_n(\Theta_n) \triangleq \max_y \mathbb{E}_{Q_n(x)} \left(2^{-n \cdot U_n(x,y)}\right).$$

(11)

$K_n(\Theta_n)$ is the redundancy associated with the decoder.

The performance degradation of the proposed universal decoder relative to any metric $m_\theta(x,y)$ in the family can be measured through the redundancy as demonstrated in the following Theorem.

Theorem 1. For any $\theta \in \Theta_n$

$$p_{e,U_n,x,y} \leq p_{e,m_\theta,x,y} \cdot K_n(\Theta_n).$$

(12)

In particular, if $K_n(\Theta_n)$ is sub-exponential then the universal decoder $U_n(x,y)$ is optimal.

Proof:

$$p_{e,U_n,x,y} = Q_n \left(U_n(x,y) \geq U_n(x,y)\right)$$

$$= Q_n \left(2^{-n \cdot U_n(x,y)} \geq 2^{-n \cdot U_n(x,y)}\right)$$

$$(a) \leq \mathbb{E}_{Q_n(x)} \left(2^{-n \cdot U_n(x,y)}\right) \cdot 2^{-n \cdot U_n(x,y)}$$

$$(b) \leq p_{e,m_\theta,x,y} \cdot K_n(\Theta_n)$$

where (a) follows from Markov’s inequality and (b) from the definition of $K_n(\Theta_n)$ (11) and $U_n(x,y)$ (10). If $K_n(\Theta_n)$ is sub-exponential then $U_n(x,y)$ dominates every metric $m_\theta$. By Lemma (3) the universal decoder is optimal.

Theorem (1) provides a quite general construction of a universal decoder. However, bounding $K_n(\Theta_n)$ might not be an easy task.
A. Conditions for universal decoding

Obviously, for a family of a single metric there must be a universal decoder, i.e., that single metric or an equivalent metric. By analyzing this simple case we get a core understanding of when the proposed GMET decoder is optimal.

Definition 6. For a given metric \( m(x, y) \) let:

\[
\bar{m}(x, y) = -\frac{1}{n} \log (p_{c, m, x, y}). \tag{13}
\]

\( \bar{m}(x, y) \) is called the canonical metric associated with \( m(x, y) \).

The canonical metric \( \bar{m} \), which is equivalent to the original metric \( m \) as it induces the same order, is the proposed universal decoder for the family of a single metric \( m(x, y) \). The next definition captures the cases where this universal metric is optimal in the sense of Theorem 1.

Definition 7. 

1) The metric \( m(x, y) \) over the prior \( Q_n(x) \) is normal if

\[
\mathbb{E}_{Q_n(x)} \left( 2^{n \cdot \bar{m}(X, Y)} \right) \leq 2^{n \cdot \lambda_n}
\]

uniformly over \( y \), where \( \lambda_n \to 0 \).

2) The prior \( Q_n(x) \) is normal if each metric over \( Q_n(x) \) is normal (uniformly over all the metrics, i.e., the bound is independent of the metric).

Normal metrics are exactly those metrics that might admit an optimal universal decoder.

The following lemma provides a sufficient condition for a given discrete prior \( Q_n(x) \) to be normal. The condition is very mild and easy to check.

Lemma 3. Let \( \chi_n = -\log (\min_{x:Q_n(x)\neq 0} Q_n(x)) \). If \( \chi_n \) is sub-exponential, then \( Q_n(x) \) is normal.

Proof: The proof is given in the appendix (Proposition 6) where we also introduce the concept of “rate function” and relate it to the canonical metric.

B. The Merged decoder

If the family of metrics is finite, we can ”merge” all the metrics and get a bound on \( K_n(\Theta_n) \). If the family size grows sub-exponentially over the normal prior then the family admits a universal decoder.

Theorem 2. Suppose \( Q_n(x) \) is normal and \( \Theta_n \) is finite, then:

1) \( K_n(\Theta_n) \leq |\Theta_n| \cdot 2^{o(n)} \)

2) If \( |\Theta_n| \) grows sub-exponentially, then the family \( \Theta_n \) admits an optimal universal decoder.

Proof: For each \( y \):

\[
\mathbb{E}_{Q_n(x)} \left( 2^{n \cdot U_n(x, y)} \right) = \mathbb{E}_{Q_n(x)} \left( \frac{1}{\min_{\theta \in \Theta_n} p_{c, m, x, y}} \right)
\leq \sum_{\theta \in \Theta_n} \mathbb{E}_{Q_n(x)} \left( \frac{1}{p_{c, m, x, y}} \right)
\leq |\Theta_n| \cdot 2^{o(n)}
\]

(1) follows by taking max over all \( y \) and (2) from Theorem 1.

C. Separable family

Definition 8. The family of metrics relative to the prior \( Q_n(x) \) is separable if for each \( y \) there exists a sequence \( \mathcal{S}_y(n) \subset \Theta_n \) that is ”dense” in the family in the sense that for each \( x, \theta \in \Theta_n \) there exist \( \theta_y \in \mathcal{S}_y(n) \) such that:

\[
p_{c, m_{\theta_y}, x, y} \leq p_{c, m, x, y} \cdot 2^{o(n)} \tag{14}
\]

and \( |\mathcal{S}_y(n)| \) grows sub-exponentially. Of course, the bound in (14) is uniform over \( x, y \) and \( \theta \in \Theta_n \).

Theorem 3. If the family of metrics is separable, then \( U_n \) is an optimal universal decoder.
Proof: For each $y$:

$$
\mathbb{E}_{Q_n(x)} \left( 2^{n \cdot \tilde{U}_n(x, y)} \right) = \mathbb{E}_{Q_n(x)} \left( \frac{1}{\min_{\theta \in \Theta_n} p_{e, m_{\theta}, x, y}} \right) \leq \mathbb{E}_{Q_n(x)} \left( \frac{1}{\min_{\theta \in S_n(n)} p_{e, m_{\theta}, x, y}} \right) \cdot 2^{o(n)}
$$

(a) by the separability condition, (b) follows from the normality of the prior; see proposition 3 in the appendix.

V. EXAMPLES

A. Normal priors

Example 1. Let $Q_n(x^n)$ be an i.i.d. probability distribution function, namely, $Q_n(x^n) = \prod_{i=1}^n Q(x_i)$. Obviously $\min_{x} Q_n(x^n) = (\min_{x} Q(x))^n$ and $-\log (\min_{x} Q_n(x^n)) = n \cdot \log (\min_{x} Q(x))$, which is subexponential (actually, linear in $n$).

Example 2. Let $Q_n(x)$ be uniform over some set $B_n \subset \mathcal{X}^n$. If $|\mathcal{X}|$ is finite, then $|B_n| \leq |\mathcal{X}|^n$ and $-\log (|B_n|^{-1}) \leq -\log (|\mathcal{X}|^{-n}) = n \cdot \log (|\mathcal{X}|)$, which is subexponential. Hence, $Q_n(x)$ is normal.

B. DMC and Constant type metric

It is instructive to validate the fact that the family of metrics for DMCs, and more generally metrics, that are constant on types admits universal decoding over any normal prior. To see this notice that at first glance the metric family

\[
\Theta_n \rightarrow \mathbb{E}_{Q_n(x)} \left( \frac{1}{\min_{\theta \in \Theta_n} p_{e, m_{\theta}, x, y}} \right) \cdot 2^{o(n)}
\]

is infinite, but since only the ordering matters, the space is in fact finite. Moreover, the minimum in (10) occurs on some specific type

\[
|S_n(n)| \cdot 2^{o(n)} = 2^{o(n)}
\]

This class of metrics is interesting as it can be used to approximate several other interesting metrics, and there is a polynomial number of types, then the set of metrics that receives the minimum of each type dominates the minimization and we can merge only a polynomial number of metrics to get the universal metric, which then follows to be optimal.

C. Finite state metrics

In this section we define a family of metrics that can be calculated using a finite state machine. This family was used in \[5\], where it was proved that the conditional LZ can be used as a universal decoding metric for finite state channels. For each $n$, let $S_n$ be the state space with $|S_n|$ states. A state machine is defined by the next state function $g : \mathcal{X} \times \mathcal{Y} \times S_n \rightarrow S_n$, and $q : \mathcal{X} \times \mathcal{Y} \times S_n \rightarrow \mathbb{R}$ the output function. The first state is $s_0$. The metric $m(x^n, y^n)$ is computed as follows: Let $s_i = g(s_{i-1}, x_i, y_i)$ for $i = 1...n$, then: $m(x^n, y^n) = \sum_{i=1}^n q(s_i, x_i, y_i)$.

The number of next state functions is: $|S_n||S_n|^{(|\mathcal{X}| \cdot |\mathcal{Y}|)+1}$. Once the next state is given, only the type $(x, y)$ and the state determines the metric value. In particular, the number of metrics that minimizes the pairwise error in (10) can be bounded by $B_n = |S_n||S_n|^{(|\mathcal{X}| \cdot |\mathcal{Y}|)+1} \cdot (n+1)^{||\mathcal{X}| \cdot |\mathcal{Y}|}|S_n|$. Let $|S_n| = n^\alpha$. It is easy to see that $\lim_{n \rightarrow \infty} \frac{1}{n} \log (B_n) = 0$ for $\alpha < 1$ and $\lim_{n \rightarrow \infty} \frac{1}{n} \log (B_n) = \infty$ for $\alpha \geq 1$.

When $|S_n| = n$ there exists a next state function such that the state in every different time is different and in particular, given the sent and received words, we can construct a metric that gives maximal value to these two words and smaller metric value to other code words. This means that the minimum in (10) is always $Q(x)$ and the decoder is chosen according to the a priori information.

D. $k^{th}$ order Markov metrics

A $k^{th}$ order Markov metric is a special case of a finite state metric where the state contains the last $k$ samples of $x, y$. This class of metrics is interesting as it can be used to approximate several other interesting metrics, e.g., metrics of stationary ergodic channels and finite state channels with hidden channel state. It is interesting to examine the possible Markov order (as a function of $n$) that allows universal decoding. Denote the order by $k_n$. The number of states is then $|S_n| = (|\mathcal{X}| \cdot |\mathcal{Y}|)^{k_n}$. Universal decoding is possible as long as $|S_n| = n^\alpha, \alpha < 1$. Taking the log of both sides, we have: $k_n < \frac{\log(n)}{\log(|\mathcal{X}| \cdot |\mathcal{Y}|)}$.

VI. SUMMARY AND CONCLUSIONS

A general universal decoder for a family of decoding metrics is described. This decoder generalizes several known universal decoders. The decoder, termed Generalized Minimum Error Test, uses the minimum error principle. We would like to explore the usability of the minimum error principle in other areas, including universal source coding with side information. Some further research regarding the infinite (abstract) alphabet is needed. Also, extension of the results here to network cases and to channels with feedback is also investigated in on-going research.
APPENDIX A
RATE FUNCTIONS

In order to analyze the performance of a metric decoder, we would need somehow to "normalize" the metric in order to remove its redundancy, i.e., the performance depends only on the order induced by the metric and not the specific value of the metric. However, for rate function (which we will define here) the value of the metric provides a hint to the performance of decoding with this metric. Throughout we will refer to a distribution \( Q_n(x) \) as the prior and we’ll assume it to be fixed. Through the whole treatment above, the output sequence \( y \) always held fixed.

**Definition 9.**

1) A function \( R: \mathcal{Y} \to \mathbb{R} \cup \{-\infty\} \) is called **rate function** over the prior \( Q_n(x) \) if:

\[
Pr \{ R(X) \geq t \} = Q_n(R(X) \geq t) \leq 2^{-nt}. \tag{15}
\]

2) The function \( R \) is an **asymptotic rate function** if there exists a sequence \( \lambda_n \) such that \( \lambda_n \to 0 \) and

\[
Q_n(R(X) \geq t) \leq 2^{-n(t-\lambda_n)}. \tag{16}
\]

3) Given any function \( R \) (not necessarily a rate function) define:

\[
\Omega_R(x) = -\frac{1}{n} \log(Q_n(R(X) \geq R(x })). \tag{17}
\]

\( \Omega_R(x) \) is the **canonical rate function** associated with the order \( R \).

**Proposition 4.**
(a) The function \( \Omega_R(x) \) preserves the order induced by \( R \) on \( x \). That is, \( R(x_1) \leq R(x_2) \) implies \( \Omega_R(x_1) \leq \Omega_R(x_2) \). If \( Q_n(x_2) > 0 \), then \( R(x_1) < R(x_2) \) implies \( \Omega_R(x_1) < \Omega_R(x_2) \).
(b) \( \Omega_R(x) \) is a rate function.
(c) For any rate function \( R \) and any \( x \): \( R(x) \leq \Omega_R(x) \).

**Proof:**
(a) Let \( f(t) = Q_n(R(X) \geq t) \). \( f(t) \) is a decreasing function. \( \Omega_R(x) = -\frac{1}{n} \log(f(R(x ))) \) and (a) follows.

(b) Notice that since \( \Omega_R(x) \) preserves the order in \( R \) then \( Q_n(\Omega_R(X) \geq \Omega_R(x) ) = Q_n(R(X) \geq R(x )) ) = 2^{-n\Omega_R(x)} \) and the condition \( (15) \) is met with equality. For other \( ts \) that are not equal to any \( \Omega_R(x) \) the condition can be easily verified by noting that \( Q_n(\Omega_R(X) \geq t) = 2^{-n\Omega_R(x)} \) for some \( x \) such that \( t < \Omega_R(x) \).

(c) If \( R \) is a rate function then

\[
2^{-n\Omega_R(x)} = Q_n(\Omega_R(X) ) \geq \Omega_R(x) = Q_n(R(X) \geq R(x )) \leq 2^{-nR(x)}. \]

**Example 3.** Let \( R(x) = \frac{1}{n} \log \left( \frac{P_n(x)}{Q_n(x)} \right) \) where \( P_n(x) \) is a probability distribution, then \( R \) is a rate function. To see this, notice that by using Markov’s inequality

\[
Q_n(R(X) \geq t) = Q_n(2^{nR(X)} \geq 2^{-nt}) \leq \mathbb{E}_{Q_n(x)} \left( 2^{nR(X)} \right) \cdot 2^{-nt} = \mathbb{E}_{Q_n(x)} \left( \frac{P_n(x)}{Q_n(x)} \right) \cdot 2^{-nt} = \sum_x Q_n(x) \cdot \frac{P_n(x)}{Q_n(x)} \cdot 2^{-nt} = 2^{-nt}.
\]

**Definition 10.** We say that the function \( R: \mathcal{X} \to \mathbb{R} \cup \{-\infty\} \) is an **asymptotically tight rate function** over the prior \( Q_n(x) \) if there exists a sequence \( \lambda_n \) such that \( \lambda_n \to 0 \) and

\[
2^{-n(R(x)+\lambda_n)} \leq Q_n(R(X) \geq R(x)) \leq 2^{-n(R(x)-\lambda_n)}. \tag{18}
\]
Notice that for any function $R$, $\Omega_R(x)$ is a trivially asymptotically tight rate function (even not asymptotically). The following proposition gives sufficient conditions for any function $R$ to be an asymptotically tight rate function.

**Proposition 5.** For any function $R$ we have:

$$2^{-n - \Omega_R(x)} \leq \mathbb{E}_{Q_n(x)} \left( 2^{nR(x)} \right) \cdot 2^{-nR(x)}.$$  \hfill (19)

In particular, if

$$2^{-n(R(x) + \lambda_n)} \leq 2^{-n - \Omega_R(x)}$$  \hfill (20)

and

$$\mathbb{E}_{Q_n(x)} \left( 2^{nR(x)} \right) \leq 2^n \chi_n$$  \hfill (21)

with $\lambda_n, \chi_n \to 0$, then $R$ is an asymptotically tight rate function and $\mathbb{E}_{Q_n(x)} \left( 2^{n - \Omega_R(x)} \right)$ is also sub-exponential.

**Proof:** By Markov’s inequality:

$$2^{-n - \Omega_R(x)} = Q_n \left( R(X) \geq R(x) \right)$$

$$\leq \mathbb{E}_{Q_n(x)} \left( 2^{nR(x)} \right) \cdot 2^{-nR(x)}.$$

Combining this with (20), (21): $2^{-n(R(x) + \lambda_n)} \leq 2^{-n - \Omega_R(x)} \leq 2^{-n(R(x) - \lambda_n)}$, in other words, $R$ is asymptotically tight. Observing that (20) implies that $2^{-n(R(x) + \lambda_n)} \geq 2^{-n - \Omega_R(x)}$, and taking $\mathbb{E}_{Q_n(x)} \left( \cdot \right)$ of that we get $\mathbb{E} \left( 2^{n - \Omega_R(x)} \right) \leq 2^n \chi_n + \lambda_n$. \hfill \blacksquare

**Proposition 6.**

(a) Suppose that there exists some $B_n$ that bounds the rate function $R$ from above, i.e., $Pr \left\{ nR(X) \geq B_n \right\} = 0$. Then

$$\mathbb{E}_{Q_n(x)} \left( 2^{nR(X)} \right) \leq 1 + \ln(2) \cdot B_n.$$  \hfill (22)

(b) If $\mathcal{X}$ is discrete, $\chi_n = -\log \left( \min_{x:Q_n(x) \neq 0} Q_n(x) \right)$ is a bound on any rate function.

**Proof:** (a) By [13] Lemma 5.5.1, for a non-negative R.V. $Z$ it holds that $\mathbb{E}(Z) = \int_0^\infty Pr \left\{ Z \geq z \right\} dz$.

$$\mathbb{E} \left( 2^{nR(X)} \right) = \int_0^\infty Pr \left\{ 2^{nR(X)} \geq t \right\} dt$$

$$= \int_0^\infty Pr \left\{ nR(X) \geq \log(t) \right\} dt$$

$$\leq 1 + \int_1^{2B_n} Pr \left\{ nR(X) \geq \log(t) \right\} dt$$

$$= 1 + \int_1^{2B_n} Pr \left\{ R(X) \geq \frac{\log(t)}{n} \right\} dt$$

$$\leq 1 + \int_1^{2B_n} 2^{-n \cdot \log(t)} dt$$

$$= 1 + \int_1^{2B_n} \frac{dt}{t}$$

$$= 1 + \ln(t) \bigg|_1^{2B_n} = 1 + \ln(2) \cdot B_n.$$  \hfill (1)

Inequality (1) follows because $Pr \left\{ \cdot \right\} \leq 1$ and $Pr \left\{ nR(X) \geq \log(2B_n) \right\} = 0$ and (2) because $R$ is a rate function.

(b) For any $x$ with $Q_n(x) > 0$ we have:

$$n \cdot R(x) \leq n \cdot \Omega_R(x)$$

$$= -\log \left( Q_n \left( R(X) \geq R(x) \right) \right)$$

$$\leq -\log \left( Q_n(x) \right)$$

$$\leq -\log \left( \min_{x:Q_n(x) \neq 0} Q_n(x) \right).$$  \hfill \blacksquare
REFERENCES

[1] V. D. Goppa, “Nonprobabilistic mutual information without memory,” Probl. Contr. Inform. Theory, vol. 4, pp. 97–102, 1975.
[2] A. Lapidoth and E. Telatar, “Gaussian isi channels and the generalized likelihood ratio test,” in Information Theory, 2000. Proceedings. IEEE International Symposium on. IEEE, 2000, p. 460.
[3] L. Farkas, “Blind decoding of linear gaussian channels with isi, capacity, error exponent, universality,” arXiv preprint arXiv:0801.0540, 2008.
[4] N. Merhav, “Universal decoding for memoryless gaussian channels with a deterministic interference,” Information Theory, IEEE Transactions on, vol. 39, no. 4, pp. 1261–1269, 1993.
[5] J. Ziv, “Universal decoding for finite-state channels,” Information Theory, IEEE Transactions on, vol. 31, no. 4, pp. 453–460, 1985.
[6] A. Lapidoth and J. Ziv, “On the universality of the l2-based decoding algorithm,” Information Theory, IEEE Transactions on, vol. 44, no. 5, pp. 1746–1755, 1998.
[7] Y. Lomnitz and M. Feder, “Universal communication over arbitrarily varying channels,” IEEE Transactions on Information Theory, vol. 59, no. 6, pp. 3720–3752, 2013.
[8] ——, “Universal communication - part i: Modulo additive channels,” IEEE Transactions on Information Theory, vol. 59, no. 9, pp. 5488–5510, 2013.
[9] M. Feder and A. Lapidoth, “Universal decoding for channels with memory,” Information Theory, IEEE Transactions on, vol. 44, no. 5, pp. 1726–1745, 1998.
[10] N. Merhav, “Universal decoding for arbitrary channels relative to a given class of decoding metrics,” IEEE Transactions on Information Theory, vol. 59, no. 9, pp. 5566–5576, 2013.
[11] Y. Lomnitz and M. Feder, “Communication over individual channels—a general framework,” arXiv preprint arXiv:1203.1406, 2012.
[12] N. Shulman, “Communication over an unknown channel via common broadcasting,” Ph.D. dissertation, Tel Aviv University, 2003.
[13] H. Koga et al., Information-spectrum methods in information theory. Springer, 2002, vol. 50.