A Unified Framework for Problems on Guessing, Source Coding and Task Partitioning

M. Ashok Kumar*, Albert Sunny*, Ashish Thakre†, Ashisha Kumar†

Abstract

We study four problems namely, Campbell’s source coding problem, Arikan’s guessing problem, Huieihel et al.’s memoryless guessing problem, and Bunte and Lapidoth’s task partitioning problem. We observe a close relationship among these problems. In all these problems, the objective is to minimize moments of some functions of random variables, and Rényi entropy and Sundaresan’s divergence arise as optimal solutions. This motivates us to establish a connection among these four problems. In this paper, we study a more general problem and show that Rényi and Shannon entropies arise as its solution. We show that the problems on source coding, guessing and task partitioning are particular instances of this general optimization problem, and derive the lower bounds using this framework. We also refine some known results and present new results for mismatched version of these problems using a unified approach. We strongly feel that this generalization would, in addition to help in understanding the similarities and distinctiveness of these problems, also help to solve any new problem that falls in this framework.

Index Terms

Guessing, source coding, task partitioning, Rényi entropy, relative $\alpha$-Rényi entropy, Sundaresan’s divergence

I. INTRODUCTION

Information Theory is the mathematical theory of communication and was founded by Claude E. Shannon in his seminal paper [1]. The concept of entropy is most central to information theory. For a probability distribution $P = \{P(x), x \in \mathcal{X}\}$ with $P(x) > 0$, the entropy of $P$ is defined as

$$H(P) := \sum_{x \in \mathcal{X}} P(x) \log \frac{1}{P(x)}.$$  \hspace{1cm} (1)

Unless specifically stated, all the logarithms mentioned in the paper are to the base 2.

If $X$ is a random variable that follows $P$, the above quantity also refers to the entropy of $X$, and is denoted by $H(X)$. Entropy arises in various problems of information theory. In particular, Shannon showed that, in source coding, it is the expected code length (per letter) required to compress a source with alphabet set $\mathcal{X} = \{a_1, \ldots, a_M\}$ and probability

*Indian Institute of Technology Palakkad, Kerala, India Email: {ashokm, albert}@iitpkd.ac.in
†Indian Institute of Technology Indore, Madhya Pradesh, India Email: ashishblthakre10@gmail.com, akumar@iiti.ac.in
distribution $P$. If the compressor does not know the true distribution $P$, but assumes a distribution $Q$, then the optimal expected code length (per letter) is $H(P) + I(P, Q)$, where

$$I(P, Q) := \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)},$$

(2)

is known as the entropy of $P$ relative to $Q$ or the Kullback-Leibler divergence. $I(P, Q)$ can be interpreted as the penalty for not knowing the true distribution.

In his seminal paper, Shannon also argued that entropy can be regarded as a measure of uncertainty. According to him, “if there is such a measure, say $H(p_1, p_2, \ldots, p_n)$, it is reasonable to require of it the following properties."

1) (Continuity) $H$ is continuous in $p_i$’s.

2) (Monotonicity) If all the $p_i$ are equal, $p_i = 1/n$, then $H$ should be a monotonic increasing function of $n$. With equally likely events there is more choice, or uncertainty, when there are more possible events.

3) (Strong additivity) If a choice be broken down into many successive choices, the original $H$ should be the weighted sum of the individual values of $H$, that is,

$$H(p_1 p_{11}, \ldots, p_1 p_{1n}, \ldots, p_m p_{m1}, \ldots, p_m p_{mn}) = H(p_1, \ldots, p_m) + \sum_{i=1}^{m} p_i H(p_{i1}, p_{i2}, \ldots, p_{in}).$$

(3)

In terms of random variables, property 3 translates to, if $X$ and $Y$ are random variables, then $H(X, Y) = H(X) + H(Y|X)$, where $H(Y|X)$ is the conditional entropy of $Y$ given $X$. Indeed, Shannon proved that the only $H$ that satisfies the above three properties is the quantity given by Eqn. 1, Th. 2].

In 1961, Rényi [2] proved that, in the above, if one replaces property 3 by a weaker variant, namely 3’. (Additivity)

$$H(p_1 q_1, p_1 q_2, \ldots, p_1 q_n, \ldots, p_m q_1, p_m q_2, \ldots, p_m q_n) = H(p_1, \ldots, p_m) + H(q_1, \ldots, q_n),$$

(4)

that is, $H(X, Y) = H(X) + H(Y)$ if $X$ and $Y$ are independent,

then the quantities, now known as Rényi entropy of order $\alpha$, $H_{\alpha}(P) := \frac{1}{1-\alpha} \log \sum_{x \in \mathcal{X}} P(x)^{\alpha}$,

(5)

where $\alpha > 0$ and $\alpha \neq 1$, also satisfy properties 1, 2, and 3’. Refer Aczel and Daroczy [3] and the references therein for an extensive study of characterizations of information measures. By a simple application of the L’Hôpital rule, one can show that $\lim_{\alpha \to 1} H_{\alpha}(P) = H(P)$. Hence, Rényi entropy can be regarded as a generalization of the Shannon entropy.

In 1965, Campbell [4] gave an operational meaning to Rényi entropy. He showed that, instead of the expected code lengths, if one minimizes the cumulants of code lengths, then the optimal cumulant is the Rényi entropy. He also showed that the optimal cumulant can be achieved by encoding sufficiently long sequences of symbols. In 1988, Blumer and McEliece [5] (c.f. Sundaresan [6, Th. 8]) obtained the redundancy in cumulants in the mismatched case. Indeed, the optimal cumulant, assuming that the true underlying distribution is $Q$, when it is actually $P$, is $H_{\alpha}(P) + I_{\alpha}(P, Q)$, where

$$I_{\alpha}(P, Q) := \frac{\alpha}{1-\alpha} \log \left( \sum_{x \in \mathcal{X}} P(x) \left[ \sum_{x' \in \mathcal{X}} \left( \frac{Q(x')}{Q(x)} \right)^{\alpha} \right]^{\frac{1}{\alpha}} \right) - H_{\alpha}(P),$$

(6)
called $\alpha$-entropy of $P$ relative to $Q$ or the Sundaresan’s divergence $\mathbb{I}_\alpha$. Thus, $\mathbb{I}_\alpha(P,Q)$ is the penalty for not knowing the true distribution in the Campbell’s coding problem. $\mathbb{I}_\alpha(\cdot,\cdot)$ was also identified by Johnson and Vignat in the context of maximum Rényi entropy $\mathbb{I}_\alpha$. They called it relative $\alpha$-Rényi entropy. $\mathbb{I}_\alpha(\cdot,\cdot)$ for $\alpha > 1$ also arises in robust inference problems (see [9] and the references therein).

In 1994, Massey [10] studied a problem on guessing where one is interested in the expected number of guesses required to guess a random variable $X$ that assumes values from an infinite set, and found a lower bound in terms of Shannon entropy. Arikan [11] studied it for finite alphabet set and showed that Rényi entropy arises as the optimal solution for the moments of number of guesses. Later, Sundaresan [6] studied the redundancy in mismatched guessing. Indeed, he showed that the penalty in guessing according to a distribution $Q$ when the true distribution is $P$ is measured by $\mathbb{I}_\alpha(P,Q)$.

Bunte and Lapidoth [12] studied the problem of partitioning of tasks and showed that the Rényi entropy and Sundaresan’s divergence play a similar role in the optimal solution of the moments of the number of tasks performed.

Quite recently, Huieihel et al. [13] studied the memoryless guessing problem, which is a variant of Arikan’s guessing problem in which the guesses are i.i.d. (independent and identically distributed). They showed that Rényi entropy arises as the exact solution of the optimal factorial moments of the number of guesses.

On studying the aforementioned four problems, we could see a close relationship among them. In all these problems, the objective is to minimize moments or factorial moments of random variables, and the Rényi entropy and Sundaresan’s divergence arise as optimal solutions. This motivates us to find a connection between these four problems. Indeed, we solve all these problems by an unified approach.

A. Our Contribution

The main contributions of this paper are as follows

- Studying a more general problem and showing that Rényi and Shannon entropies as its solution.
- Casting problems on source coding, guessing and task partitioning as particular instances of this general problem, and deriving the lower bounds using this framework.
- A unified approach to derive bounds for the mismatched version of these problems in terms of Kullback-Leibler and Sunderesan divergences.
- Refinement of some known results for the mismatched version of these problems.
- New results on mismatched version of memoryless guessing and task partitioning.

The remainder of the paper is organized as follows. In Section II we describe all the problems and state the results (both known as well as new). In Section III we formulate a unified approach which allows us to derive information theoretic bounds for these problems. Proofs of the stated results are derived using the unified approach in Section IV. Finally, we summarize and conclude the paper in Section V.

II. Problem Statements and Known Results

In this section we discuss all the four problems in detail and present some known and new results.
A. Source Coding Problem

A source consists of an alphabet set \( \mathcal{X} = \{a_1, \ldots, a_M\} \) together with a probability distribution \( P \). A binary code \( C \) is a mapping from \( \mathcal{X} \) to the set of finite length binary strings. Let \( X \) be a random variable that assumes values from the set \( \mathcal{X} \) and follows a probability distribution \( P \). Let \( C(X) \) be the codeword assigned to \( X \). Let \( L(C(X)) \) be the length of codeword \( C(X) \). An obvious condition for the codewords is that they are uniquely decodable. Among uniquely decodable codes, the objective is to find the one that minimizes the expected code-length. That is,

\[
\text{Minimize } \mathbb{E}[L(C(X))]
\]

over all uniquely decodable codes \( C \). However, Kraft-McMillan proved the following.

**Proposition 1.** \[14\] If \( C \) a uniquely decodable code, then

\[
\sum_{x \in \mathcal{X}} 2^{-L(C(x))} \leq 1. \tag{7}
\]

Conversely, given a length sequence that satisfies the above inequality, there exists a uniquely decodable code \( C \) with the given length sequence.

Thus, one can confine the search space of the expected code-length minimization to codes satisfying the Kraft-McMillan inequality.

**Theorem 1.** \[1\], Th. 5] Let \( X \) be a random variable on the source \( (\mathcal{X}, P) \). Let \( C \) be a uniquely decodable code. Then

\[
\mathbb{E}[L(C(X))] \geq H(P).
\]

Shannon also proved that the optimal code-length can be achieved by encoding long sequences of symbols. Indeed, he proved the following. Let \( \mathcal{X}^n \) denote the cartesian product of \( \mathcal{X} \) with itself \( n \) times; representing all possible \( n \)-length sequences of symbols from \( \mathcal{X} \). Let \( X^n := X_1, \ldots, X_n \) denote an i.i.d. sequence from \( \mathcal{X}^n \), and \( P_n \) denote the product distribution of \( X^n \), that is, \( P_n(X^n) = \prod_{i=1}^n P(X_i) \).

**Theorem 2.** \[1\], Th. 5] Let \( X^n \) be an i.i.d. sequence from \( \mathcal{X}^n \) following the distribution \( P_n \). Let \( C_n \) be a code such that \( L(C_n(X^n)) = \lceil -\log P_n(X^n) \rceil \). Then, \( C_n \) satisfies the Kraft-McMillan inequality and

\[
\lim_{n \to \infty} \frac{L(C_n(X^n))}{n} = H(P).
\]

Mismatch Case

If the compressor does not know the true distribution \( P \), but encodes according to the optimal length corresponding to some distribution \( Q \), then the minimum expected code-length is bounded above by

\[
\sum_{x \in \mathcal{X}} P(x) \left[ \log \left( \frac{1}{Q(x)} \right) \right] \leq \sum_{x \in \mathcal{X}} P(x) \log \left( \frac{1}{Q(x)} \right) + 1 = \sum_{x \in \mathcal{X}} P(x) \log \left( \frac{P(x)}{Q(x) P(x)} \right) + 1
\]

\[
= \sum_{x \in \mathcal{X}} P(x) \log \left( \frac{1}{P(x)} \right) + \sum_{x \in \mathcal{X}} P(x) \log \left( \frac{P(x)}{Q(x)} \right) + 1
\]

\[
= H(P) + I(P,Q) + 1.
\]
Thus, $I(P,Q)$ is the penalty for not knowing the true distribution. $I(P,Q)$ is a divergence function in the sense that $I(P,Q) \geq 0$ and $I(P,Q) = 0$ if and only if $P = Q$. A result similar to Theorem 2 can also be established for the mismatched case.

### B. Campbell Coding Problem

In Campbell’s coding problem, one minimizes the normalized cumulant of code lengths, that is,

$$\text{Minimize } \frac{1}{\rho} \log \mathbb{E}[2^{\rho L(C(X))}],$$

over all uniquely decodable codes $C$, and $\rho > 0$.

Minimization of cumulants arise in the probability of buffer overflow in coding with a buffer at the encoder [15]. Moments of random variable also help us understand various characteristics of the underlying probability distribution. With the help of moments, one can find the mean, variance, skewness, kurtosis, etc of a distribution. The underlying distribution can also be determined by its moment sequence. Campbell [4] gave a lower bound for the normalized cumulants in terms of Rényi entropy.

**Theorem 3.** [4, Lemma] Let $L(C(\cdot))$ be the length function of a uniquely decodable code $C$. Then

$$\frac{1}{\rho} \log \mathbb{E}[2^{\rho L(C(X))}] \geq H_{\alpha(\rho)}(P),$$

where $\alpha(\rho) = 1/(1 + \rho)$ and $H_{\alpha(\rho)}(P)$ is as in (5).

For the sake of convenience we suppress the dependence of $\rho$ on $\alpha$ and simply write $\alpha(\rho)$ as $\alpha$ in the remainder of this paper. Campbell also showed that, if we ignore the integer constraint of the length function, then

$$L(C(x)) = \log \frac{Z_{P,\alpha}}{P(x)^{\alpha}},$$

where $Z_{P,\alpha} := \sum_{x \in X} P(x)^{\alpha}$, satisfies (7) and achieves the lower bound in (8). Indeed, he showed that the lower bound in (8) can be achieved by encoding long sequences of symbols with code-lengths close to (9).

**Theorem 4.** [4, Theorem] If $C_n$ is a uniquely decodable code such that

$$L(C(x^n)) = \left\lfloor \log \frac{Z_{P,\alpha,n}}{P(x^n)^{\alpha}} \right\rfloor,$$

where $Z_{P,\alpha,n} := \sum_{x^n \in X^n} P(x^n)^{\alpha}$, then

$$\lim_{n \to \infty} \frac{1}{n\rho} \log \mathbb{E}[2^{\rho L(C_n(X^n))}] = H_\alpha(P).$$

**Mismatch Case**

Blumer and McEliece [5] and Sundaresan [6] studied the redundancy in the mismatched case of the Campbell’s problem. Sundaresan showed that the difference in the normalized cumulant from the minimum when encoding according to an arbitrary uniquely decodable code is measured by the $I_\alpha$-divergence up to a factor of 1. That can be generalized to the following.
Theorem 5. Let \( \rho \in (-1, 0) \cup (0, \infty) \). Let \( L : \mathcal{X} \to \mathbb{Z}_+ \) be an arbitrary length function that satisfies (7). Let
\[
R_c(P, L, \rho) := \frac{1}{\rho} \log E[2^{\rho L(X)}] - \min_K \frac{1}{\rho} \log E[2^{\rho K(X)}],
\]
where the expectation is with respect to \( P \) and the minimum is over all length functions \( K \) satisfying (7). Then, there exists a probability distribution \( Q_L \) such that
\[
I_{\alpha}(P, Q_L) - \log \eta - 1 \leq R_c(P, L, \rho) \leq I_{\alpha}(P, Q_L) - \log \eta,
\]
where \( \eta = \sum_x 2^{-L(x)} \).

It should be noted that the lower bound in (13) is arbitrarily large for sufficiently small \( \eta \). For example, for a source with two symbols, say \( x \) and \( y \), with code lengths \( L(x) = L(y) = 100 \), the left side of (13) becomes \( I_{\alpha}(P, Q_L) + 98 \). However, if \( 1/2 \leq \eta \leq 1 \), then (13) becomes
\[
|R_c(P, L, \rho) - I_{\alpha}(P, Q_L)| \leq 1,
\]
which is [6, Th. 8]. \( I_{\alpha}(P, Q_L) \) is, in a sense, the penalty when \( Q_L \) is not matched with the true distribution \( P \). \( I_{\alpha}(\cdot, \cdot) \) is also a divergence function in the sense that
1) \( I_{\alpha}(P, Q) \geq 0 \). 
2) \( I_{\alpha}(P, Q) = 0 \) if and only if \( P = Q \).

Moreover, as \( \alpha \to 1 \), \( I_{\alpha}(P, Q) \to I(P, Q) \) [6]. In view of this, a result analogous to Theorem 5 also holds for the Shannon source coding problem.

C. Arikan’s Guessing Problem

Let \( \mathcal{X} \) be a set of \( M \) objects. Bob thinks of an object \( X \) (a random variable) from \( \mathcal{X} \) according to a probability distribution \( P \). Alice guesses it by asking questions of the form “is \( X = x \)?”. The objective is to minimize the average number of guesses required for Alice to correctly guess \( X \). Suppose Alice uses a guessing strategy \( G \) for asking questions to Bob. By a guessing strategy, we mean a one-one map \( G : \mathcal{X} \to \{1, \ldots, M\} \), where \( G(x) \) is to be interpreted as the number of questions required by Alice to guess \( x \) correctly. Arikan studied the \( \rho \)-th moment of number of guesses and found upper and lower bounds in terms of Rényi entropy.

Theorem 6. Let \( \rho \in (-1, 0) \cup (0, \infty) \). Let \( X \) be a random guess according to the distribution \( P \), and let \( G \) be any guessing function. Then
\[
\frac{1}{\rho} \log \mathbb{E}[G(X)^\rho] \geq H_{\alpha}(P) - \log h_M, \]
where \( h_M = 1 + 1/2 + \cdots + 1/M \).

Arikan also showed that, if one follows the optimal guessing function \( G^* \), that is, guessing according to the decreasing order of \( P \)-probabilities, it does not require more than the Rényi entropy.

\[ \text{Arikan upper bounded } h_M \text{ by } 1 + \log M. \]
**Theorem 7.** Let \( \rho \in (-1, 0) \cup (0, \infty) \). If \( G^* \) is an optimal guessing function, then
\[
\frac{1}{\rho} \log \mathbb{E}[G^*(X)^\rho] \leq H_\alpha(P).
\]

Arikan also proved that Rényi entropy can be achieved by guessing long sequences of letters in an i.i.d. fashion.

**Theorem 8.** Let \( \rho \in (-1, 0) \cup (0, \infty) \). Let \( X_1, X_2, \ldots, X_n \) be a sequence of i.i.d. guesses. Let \( G^*_n(X_1, \ldots, X_n) \) be an optimal guessing function. Then,
\[
\lim_{n \to \infty} \frac{1}{n \rho} \log \mathbb{E}[G^*_n(X_1, X_2, \ldots, X_n)^\rho] = H_\alpha(X_1).
\]

**Remark** For the case when \( \rho > 0 \), Theorems 6, 7 and 8 were stated and proved in [11].

**Mismatch Case**

Suppose Alice does not know the true underlying probability distribution \( P \), and she guesses according to some guessing function \( G \). Sundaresan proved that the penalty in this mismatched guessing is again measured by the \( I_\alpha \)-divergence [6, Th. 6]. This can be refined to the following theorem.

**Theorem 9.** Let \( \rho \in (-1, 0) \cup (0, \infty) \). Let \( X \) be a random variable that assumes values from \( \mathcal{X} \) and follows the distribution \( P \). Let \( G \) be an arbitrary guessing function. Then, there exists a probability distribution \( Q_G \) on \( \mathcal{X} \) such that
\[
\frac{1}{\rho} \log \mathbb{E}[G(X)^\rho] = H_\alpha(P) + I_\alpha(P, Q_G) - \log h_M,
\]
where the expectation is with respect to \( P \).

**Theorem 10.** [6, Prop. 1] Let \( \rho \in (-1, 0) \cup (0, \infty) \). Let \( X \) be a random variable that assumes values from \( \mathcal{X} \) and follows the distribution \( P \). Let \( Q \) be another distribution. Let \( G_Q \) be the guessing strategy that guesses according to the decreasing order of \( Q \)-probabilities. Then
\[
\frac{1}{\rho} \log \mathbb{E}[G_Q(X)^\rho] \leq H_\alpha(P) + I_\alpha(P, Q),
\]
where expectation is with respect to \( P \).

The above two theorems can be combined to state the following, as in the Campbell’s coding problem.

**Theorem 11.** Let \( \rho \in (-1, 0) \cup (0, \infty) \). Let \( X \) be a random variable that assumes values from \( \mathcal{X} \) and follows the distribution \( P \). Let \( G \) be an arbitrary guessing function and \( G_P \) be the optimal guessing function for \( P \). Let
\[
R_g(P, G, \rho) := \frac{1}{\rho} \log \mathbb{E}[G(X)^\rho] - \frac{1}{\rho} \log \mathbb{E}[G_P(X)^\rho],
\]
where the expectation is with respect to \( P \). Then, there exists a probability distribution \( Q_G \) such that
\[
I_\alpha(P, Q_G) - \log h_M \leq R_g(P, G, \rho) \leq I_\alpha(P, Q_G)
\]

**Remark** For the case when \( \rho > 0 \), the above theorem was stated and proved in [6, Th. 6].
D. Memoryless Guessing

In Arikan’s guessing problem, the guesser Alice does not ask a question that was asked before. We now present a memoryless guessing where Alice does not keep track of her previous guesses; every-time she comes up with a guess independent of her previous guesses. Let $\hat{X}_1, \hat{X}_2, \ldots$ be a sequence of independent guesses according to a distribution $\hat{P}$. Let $P$ be the underlying distribution with respect to which Bob sets the guesses. Huieihel et al. [13] showed that, surprisingly, the optimal one for Alice, even if she knows the true underlying distribution, is not $P$, but $P(\alpha)$, the $\alpha$-scaled measure of $P$, defined by

$$P(\alpha)(x) = \frac{P(x)^\alpha}{Z_{P,\alpha}}.$$ 

The guessing function in this problem is defined as,

$$G(X, \hat{X}_1^\infty) := \inf\{K \geq 1 : \hat{X}_K = X\},$$

that is, the number of guesses until a success.

Unlike in Massey’s guessing problem, Huieihel et al. [13] minimized what are called factorial moments, defined for $\rho \in \mathbb{Z}_+$ as,

$$V_\rho(X, \hat{X}_1^\infty) = \frac{1}{\rho!} \prod_{l=0}^{\rho-1} (G(X, \hat{X}_1^\infty) + l).$$

Huieihel et al. [13] studied the following problem.

Minimize $\mathbb{E}[V_\rho(X, \hat{X}_1^\infty)]$, over all $\hat{P} \in \mathcal{P}$, where $\mathcal{P}$ is the probability simplex, that is, $\mathcal{P} = \{(P(x))_{x \in \mathcal{X}} : P(x) \geq 0, \sum_x P(x) = 1\}$. The optimal value, if attained, is denoted by $\mathbb{E}[V_\rho^*(X, \hat{X}_1^\infty)]$. Huieihel et al. [13] proved the following.

**Theorem 12.** [13, Th. 1] For any integer $\rho > 0$, we have

$$\frac{1}{\rho} \log \mathbb{E}[V_\rho^*(X, \hat{X}_1^\infty)] = H_\alpha(P)$$

and is attained by

$$\hat{P}_\rho^*(x) = \frac{P(x)^\alpha}{Z_{P,\alpha}}.$$

For sequence of random guesses $\hat{X}_n$, the above theorem can be stated in the following way. Let $\hat{X}_n = (\hat{X}_1, \ldots, \hat{X}_n)$ be a sequence of i.i.d. guesses from $\mathcal{X}$ with distribution $\hat{P}_n$ — the $n$-fold product distribution of $P$ on $\mathcal{X}$. Assume that $P_n$ is the true underlying distribution. Then

$$\lim_{n \to \infty} \frac{1}{n\rho} \log \mathbb{E}[V_\rho^*(X^n, (\hat{X}_1^n)^\infty)] = H_\alpha(P),$$

where the expectation is with respect to $P_n$.

It is now easy to obtain the result in the mismatched case.

**Theorem 13.** If the true underlying probability distribution is $P$ but Alice thinks that it is $Q$ and guesses according to its optimal one, namely $\hat{Q}_\rho^*(x) = Q(x)^\alpha/Z_{Q,\alpha}$, then

$$\frac{1}{\rho} \log \mathbb{E}[V_\rho^*(X, \hat{X}_1^\infty)] = H_\alpha(P) + I_\alpha(P, Q),$$

where the expectation is with respect to $P$. 

---

July 17, 2019  DRAFT
E. Tasks Partitioning Problem

Let $\mathcal{X}$ be a finite set of tasks. A task $X$ is randomly drawn from $\mathcal{X}$ according to a probability distribution $P$ and is performed. The distribution $P$ may correspond to the frequency of occurrences of the tasks. Suppose there are only $N$ keys to which these tasks are to be associated. Typically, $N < |\mathcal{X}|$. Due to the limited availability of keys, more than one task may be associated with a single key. When a task needs to be performed, the key associated with the task is pressed and all the tasks associated with the key will be performed. The objective in this problem is to minimize the number of redundant tasks performed. Usual coding techniques suggest us to assign tasks with higher probability to individual keys and leave the lower probability tasks unassigned to any keys. But for an individual all the tasks can be equally important. Just that some tasks may have more frequency of occurrences while others have lesser. If $N \geq |\mathcal{X}|$, then one can perform tasks without any redundancy. However, Bunte and Lapidoth [12] show that, even when $N < |\mathcal{X}|$, one can accomplish the tasks with much less redundancy on average, provided the underlying probability distribution is different from the uniform distribution. When $N < |\mathcal{X}|$ the only way is, to partition the set of tasks and assign keys in an intelligent way keeping the probability distribution in mind, so as to minimize the average number of redundant tasks performed.

Let $A = \{A_1, A_2, \ldots, A_N\}$ be a partition of $\mathcal{X}$ that corresponds to the assignment of tasks to keys. Let $A(x)$ be the cardinality of the subset containing $x$ in the partition. We shall call $A$ the partition function associated with $A$. Bunte and Lapidoth [12, Th. I.1] proved the following theorem for the case $\rho > 0$.

**Theorem 14.** Let $\rho \in (-1, 0) \cup (0, \infty)$. Let $X$ be a random task from $\mathcal{X}$ following distribution $P$. Then the following hold.

(a) For any partition of $\mathcal{X}$ of size $N$ with partition function $A$, we have

$$ \frac{1}{\rho} \log \mathbb{E}[A(X)^{\rho}] \geq H_\alpha(P) - \log N. $$

(b) (i) If $N > \log |\mathcal{X}| + 2$ and $\rho > 0$, then there exists a partition of $\mathcal{X}$ of size at most $N$ with partition function $A$ such that

$$ 1 \leq \mathbb{E}[A(X)^{\rho}] \leq 1 + 2^{\rho(H_\alpha(P) - \log \tilde{N})}, $$

where $\tilde{N} = \frac{N - \log |\mathcal{X}| - 2}{4}$.

(ii) If $\rho < 0$, then there exists a partition of $\mathcal{X}$ of size at most $N$ with partition function $A$ such that

$$ \frac{1}{2} 2^{\rho H_\alpha(P)} \leq \mathbb{E}[A(X)^{\rho}] \leq 1. $$

Bunte and Lapidoth also found conditions under which the redundancy can be made negligible on average while performing a long sequence of tasks.

**Theorem 15.** [12, Th. I.2] Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. tasks following distribution $P$. Then the following hold.

(i) If $\log N > H_\alpha(P)$, there exists partition $A_n$ of $\mathcal{X}^n$ of size at most $N^n$ with associated partition function $A_n$ such that

$$ \lim_{n \to \infty} \mathbb{E}[A_n(X^n)^{\rho}] = 1. $$
(ii) If $\log N < H_\alpha(P)$, then for partitions $A_n$ of $\mathcal{X}^n$ of size at most $N^n$ with associated partition function $A_n$,

$$\lim_{n \to \infty} \mathbb{E}[A_n(X^n)^\rho] = \infty.$$ 

Mismatch Case

Consider a scenario where one does not know the true underlying probability distribution $P$, but arbitrarily partitions $\mathcal{X}$. Then, the penalty due to such a partition can be measured by the $I_\alpha$-divergence as stated in the following theorem.

**Theorem 16.** Let $\rho \in (-1, 0) \cup (0, \infty)$. Let $A$ be a partition of $\mathcal{X}$ of size $N$ with partition function $A$. Then, there exists a probability distribution $Q_A$ on $\mathcal{X}$ such that

$$\frac{1}{\rho} \log \mathbb{E}[A(X)^\rho] = H_\alpha(P) + I_\alpha(P,Q_A) - \log N.$$ 

A converse result is the following.

**Theorem 17.** [12, Sec. IV] Let $\rho > 0$. Let $X$ be a random task from $\mathcal{X}$ following distribution $P$. Let $Q$ be another distribution on $\mathcal{X}$. If $N > \log |\mathcal{X}| + 2$, then there exists partition $A_Q$ (with an associated partition function $A_Q$) of $\mathcal{X}$ of size at most $N$ such that

$$\mathbb{E}[A_Q(X)^\rho] \leq 1 + 2^{\rho(H_\alpha(P) + I_\alpha(P,Q) - \log N)},$$

where $\tilde{N} = \frac{N - \log |\mathcal{X}| - 2}{4}$ and the expectation is with respect to $P$.

III. A Unified Approach

Our objective in this section is to formulate a more general problem to which the problems studied in the previous section are particular cases.

**Theorem 18.** Let $\rho \in (-1, 0) \cup (0, \infty)$. Let $\psi : \mathcal{X} \to [0, \infty)$ such that $\sum_{x \in \mathcal{X}} \psi(x) \leq b$ for some $b > 0$. Then

$$\frac{1}{\rho} \log \mathbb{E}[\psi(X)^{-\rho}] \geq H_\alpha(P) - \log b. \quad (14)$$

The lower bound is achieved when

$$\psi(x) = b \cdot P(x)^\alpha / Z_{P,\alpha} \quad \text{for} \ x \in \mathcal{X}. \quad (15)$$

To prove the above theorem, consider the following optimization problem

$$P_1 : \min_\phi \ sgn(\rho) \cdot \mathbb{E}[\phi(X)^{-\rho}]$$

subject to:

$$\sum_{x \in \mathcal{X}} \phi(x) \leq b \quad \text{and} \quad \phi(x) \geq 0 \quad \forall x \in \mathcal{X},$$

where $sgn(\rho)$ is the signum function of $\rho$. It is easy to check that objective function of problem $P_1$ is convex for all $\rho \in (-1, 0) \cup (0, \infty)$. Also, since the constraint set is compact, the minimum is indeed attained.

**Lemma 1.** Let $\phi^* = \{\phi^*(x), x \in \mathcal{X}\}$ be an optimal solution of problem $P_1$. Then, $\sum_{x \in \mathcal{X}} \phi^*(x) = b$. 

July 17, 2019 DRAFT
\textbf{Proof:} Suppose \( \sum_{x \in \mathcal{X}} \phi^*(x) < b \). Let \( \phi'(x) = \eta \phi^*(x) \) for \( x \in \mathcal{X} \), where \( \eta = b/\sum_{x \in \mathcal{X}} \phi^*(x) > 1 \). It is easy to see that \( \phi' \) satisfies the constraint of problem \( P_1 \). Further, we have

\[
\text{sgn}(\rho) \mathbb{E}[\phi'(x)^{-\rho}] = \text{sgn}(\rho) \eta^{-\rho} \mathbb{E}[\phi^*(x)^{-\rho}] < \text{sgn}(\rho) \mathbb{E}[\phi^*(x)^{-\rho}],
\]

where inequality (a) follows because \( \text{sgn}(\phi) \) if \( \rho \). These are also known as \( \text{lual variables} \). Multiplying (16) by \( \rho \) we get

\[
-|\rho| P(x) \phi^*(x)^{-(\rho+1)} + \lambda^* - \mu^*(x) = 0 \quad \forall x \in \mathcal{X},
\]

(16)

\[
\sum_{x \in \mathcal{X}} \phi^*(x) = b,
\]

(17)

\[
\phi^*(x) \geq 0 \quad \forall x \in \mathcal{X},
\]

(18)

\[
\mu^*(x) \geq 0 \quad \forall x \in \mathcal{X},
\]

(19)

\[
\text{and} \quad \mu^*(x) \phi^*(x) = 0 \quad \forall x \in \mathcal{X}.
\]

(20)

Multiplying (16) by \( \phi^*(x) \) and summing over all \( x \in \mathcal{X} \) we get

\[
\lambda^* = \frac{|\rho|}{b} \sum_{x \in \mathcal{X}} P(x) \phi^*(x)^{-\rho}.
\]

Substituting this in (16), we obtain

\[
\frac{|\rho|}{b} \sum_{x \in \mathcal{X}} P(x) \phi^*(x)^{-\rho} = |\rho| P(x) \phi^*(x)^{-(\rho+1)} + \mu^*(x) \quad \forall x \in \mathcal{X}.
\]

(21)

Since \( \mu^*(x) \geq 0 \), we have

\[
\sum_{x \in \mathcal{X}} P(x) \phi^*(x)^{-\rho} \geq b P(x) \phi^*(x)^{-(\rho+1)} \quad \forall x \in \mathcal{X}.
\]

(22)

Suppose \( \phi^*(\hat{x}) = 0 \) for some \( \hat{x} \in \mathcal{X} \). Then,

- for \( \rho \in (0, \infty) \), (22) says that the optimal value is infinite. This is not possible because we can always find a feasible point for problem \( P_2 \) where the objective function evaluates to a finite quantity.
• for $\rho \in (-1, 0)$, the left side of (22) is always finite, whereas the right side evaluates to $\infty$ — a contradiction.

Thus, we should have $\phi^*(x) > 0 \forall x \in \mathcal{X}$. This, in turn, implies that $\mu^*(x) = 0 \forall x \in \mathcal{X}$ (due to Eqn. (20)). Plugging $\mu^*(x) = 0$ in (21), we get

$$\phi^*(x) = \kappa P(x)^\alpha \quad \forall x \in \mathcal{X},$$

where $\kappa$ is the constant of proportionality. Summing over all $x \in \mathcal{X}$ and using the fact that $\sum_{x \in \mathcal{X}} \phi^*(x) = b$, we get $\kappa = b/Z_{P,\alpha}$. Thus, the optimal solution $\phi^*$ of problem $P_2$ has the following form

$$\phi^*(x) = \frac{b \cdot P(x)^\alpha}{Z_{P,\alpha}}$$

and

$$\mathbb{E}[^{\phi^*(X)^{-\rho}}] = b^{-\rho} \sum_{x \in \mathcal{X}} P(x) \cdot \left( \frac{Z_{P,\alpha}}{P(x)^\alpha} \right)^\rho = b^{-\rho} \left( \sum_{x \in \mathcal{X}} P(x)^\alpha \right)^{\frac{\alpha}{\rho}} \cdot \sum_{x \in \mathcal{X}} P(x)^\alpha = b^{-\rho} \left( \sum_{x \in \mathcal{X}} P(x)^\alpha \right)^{\frac{1}{\alpha}}.$$

Now, if $\psi : \mathcal{X} \rightarrow [0, \infty)$ such that $\sum_{x \in \mathcal{X}} \psi(x) \leq b$, then we have

$$\text{sgn}(\rho) \cdot \mathbb{E}[\psi(X)^{-\rho}] \geq \text{sgn}(\rho) \cdot \mathbb{E}[\phi^*(X)^{-\rho}] = \text{sgn}(\rho) \cdot b^{-\rho} \left( \sum_{x \in \mathcal{X}} P(x)^\alpha \right)^{\frac{1}{\alpha}}.$$

Taking log on both sides of the above inequality and multiplying throughout by $1/\rho$, we get (14). \hfill \square

**Theorem 19.** Let $\psi : \mathcal{X} \rightarrow [0, \infty)$ such that $\sum_{x \in \mathcal{X}} \psi(x) \leq b$, then

$$\mathbb{E}\left[ \log \frac{1}{\psi(X)} \right] \geq H(P) - \log b,$$

where $H(P)$ is the Shannon entropy. The lower bound is achieved when $\psi(x) = b \cdot P(x) \forall x \in \mathcal{X}$.

**Proof:** It is easy to see that Lemma 1 also holds if the objective function is $-\mathbb{E}[\log \psi(X)]$. Hence proceeding as in Theorem 18 with $\text{sgn}(\rho)\mathbb{E}[\psi(X)^{-\rho}]$ replaced by $-\mathbb{E}[\log \psi(X)]$, we get the optimal solution $\psi^*(x) = b \cdot P(x)$. This implies (23). \hfill \blacksquare

**Remark** In fact, for any convex and continuously differentiable function $f : \mathbb{R}^+ \rightarrow \mathbb{R}$, and $\psi : \mathcal{X} \rightarrow [0, \infty)$ such that $\sum_{x \in \mathcal{X}} \psi(x) \leq b$, we have

$$\mathbb{E}[f(\psi(X))] \geq \mathbb{E}\left[ f \left( f^{-1} \left( \frac{x^\alpha}{P(x)} \right) \right) \right],$$

where $f'$ is the derivative of $f$, $f^{-1}$ is the inverse of $f'$, and $\lambda^*$ is the solution of the equation

$$\mathbb{E}\left[ \frac{1}{P(x)} f^{-1} \left( \frac{\lambda^*}{P(x)} \right) \right] = b.$$

It is interesting to notice that, while the lower bound in Inequality (24) is dependant on function $f$, it is independent of function $\psi$. The lower bound in (24) is achieved when $\psi(x) = f^{-1} \left( \frac{\lambda^*}{P(x)} \right)$. Choosing $f(y)$ as $\text{sgn}(\rho) \cdot (y)^{-\rho}$ and $-\log(y)$ gives us the lower bound results on Rényi and Shannon entropies, respectively. The quantity on the right side of (24) is closely related to the $\phi$-entropies of Teboulle and Vajda [17].
We now extend Theorems 18 and 19 to sequences of random variables. Let $X^n$ be the set of all sequences of length $n$ of elements of $X$, and $P_n$ be the $n$-fold product distribution of $P$ on $X^n$, that is, for $x^n := x_1, \ldots, x_n \in X^n$, $P_n(x^n) = \prod_{i=1}^n P(x_i)$.

**Corollary 1.** Given any $n \geq 1$, if $\psi : X^n \to [0, \infty)$ is such that
\[
\sum_{x^n \in X^n} \psi_n(x^n) \leq b_n
\]
for some $b_n > 0$, then

1) For $\rho \in (-1, 0) \cup (0, \infty)$,
\[
\liminf_{n \to \infty} \frac{1}{n} \log \mathbb{E}[\psi(X^n)^{-\rho}] \geq H_\alpha(P) - \limsup_{n \to \infty} \frac{\log b_n}{n}.
\]

2) \[
\liminf_{n \to \infty} \frac{1}{n} \mathbb{E}\left[ \log \frac{1}{\psi(X^n)} \right] \geq H(P) - \limsup_{n \to \infty} \frac{\log b_n}{n},
\]

where the expectations are with respect to $P_n$.

**Proof:** It is easy to see that $H_\alpha(P_n) = nH_\alpha(P)$ and $H(P_n) = nH(P)$. Hence, applying Theorems 18 and 19, dividing throughout by $n$ and taking $\liminf n \to \infty$, the results follow.

---

**A. A General Framework for Mismatched Cases**

In this sub-section, we establish a unified approach for the case when there is a mismatch between the assumed and underlying distributions.

**Theorem 20.** Let $\rho \in (-1, 0) \cup (0, \infty)$ and $Q$ be a probability distribution on $X$. Suppose $\psi : X \to [0, \infty)$ is such that
\[
\psi(x)^{-1} \leq \left( b + \left( \frac{c \cdot Z_{Q,\alpha}}{Q(x)^\alpha} \right)^{|\rho|} \right)^{1/|\rho|}
\]
for some $c > 0$, and $b \geq 0$, then

1) $\mathbb{E}[\psi(X)^{-\rho}] \leq b + 2^\rho(H_\alpha(P) + I_{\alpha}(P,Q) + \log c)$ if $\rho > 0$,

2) $\mathbb{E}[\psi(X)^{-\rho}] \geq 2^\rho(H_\alpha(P) + I_{\alpha}(P,Q) + \log c^{-1/\rho} \log(1 + be^\rho))$ if $\rho < 0$.

where the expectation is with respect to $P$.

**Proof:**

**Proof of upper bound for the case $\rho > 0$:** We have
\[
\psi(x)^{-\rho} \leq b + \left( \frac{c \cdot Z_{Q,\alpha}}{Q(x)^\alpha} \right)^\rho.
\]

Hence
\[
\mathbb{E}[\psi(X)^{-\rho}] = \sum_{x \in X} P(x)\psi(x)^{-\rho} \leq b + c^\rho Z_{Q,\alpha}^\rho \sum_{x \in X} P(x)Q(x)^{-\alpha\rho}.
\]

But
\[
Z_{Q,\alpha}^\rho \sum_{x \in X} P(x)Q(x)^{\alpha - 1} = 2^{\rho(H_\alpha(P) + I_{\alpha}(P,Q))}. \quad \text{(25)}
\]
Hence (1) follows.

**Proof of lower bound for the case \( \rho < 0 \):** In this case, we have

\[
\psi(x)^{-\rho} \geq \left( b + \left( \frac{c \cdot Z_{Q,\alpha}}{Q(x)\alpha} \right)^{-\rho} \right)^{-1} \geq \left( b + c^{-\rho} \right) \left( \frac{Z_{Q,\alpha}}{Q(x)^\alpha} \right)^{-\rho},
\]

where inequality (a) follows because \( \frac{Z_{Q,\alpha}}{Q(x)^\alpha} \geq 1 \) and \(-\rho > 0\). Thus, we have

\[
\mathbb{E}[\psi(X)^{-\rho}] = \sum_{x \in \mathcal{X}} P(x) \psi(x)^{-\rho} \geq \left( 1 + b c^{-\rho} \right)^{-1} c^\rho Z_{Q,\alpha} \sum_{x \in \mathcal{X}} P(x) Q(x)^{-\rho},
\]

Hence, in view of (25), (2) follows.

**Theorem 21.** Let \( \rho \in (-1, 0) \cup (0, \infty) \) and \( Q \) be a probability distribution on \( \mathcal{X} \). If \( \psi : \mathcal{X} \to [0, \infty) \) is such that

\[
a Z_{Q,\alpha} \frac{\psi(x)}{Q(x)^\alpha} \leq \psi(x)^{-1}
\]

for some \( a > 0 \), then

\[
\text{sgn}(\rho) \cdot \mathbb{E}[\psi(X)^{-\rho}] \geq \text{sgn}(\rho) \cdot 2^{\rho (H_{\alpha}(P) + I_{\alpha}(P,Q) + \log a)},
\]

where the expectation is with respect to \( P \).

**Proof:** Since \( \psi(x)^{-1} \geq \frac{a Z_{Q,\alpha} \psi(x)}{Q(x)^\alpha} \), we have

\[
\text{sgn}(\rho) \cdot \mathbb{E}[\psi(X)^{-\rho}] = \text{sgn}(\rho) \cdot \sum_{x \in \mathcal{X}} P(x) \psi(x)^{-\rho}
\]

\[
\geq \text{sgn}(\rho) \cdot a^\rho Z_{Q,\alpha} \sum_{x \in \mathcal{X}} P(x) Q(x)^{-\rho}
\]

\[
= \text{sgn}(\rho) \cdot 2^{\rho (H_{\alpha}(P) + I_{\alpha}(P,Q) + \log a)}.
\]

**Corollary 2.** Let \( \rho > -1 \) and \( Q \) be a probability distribution on \( \mathcal{X} \). Suppose that \( \psi : \mathcal{X} \to [0, \infty) \) is such that

\[
\psi(x)^{-1} \leq \frac{c \cdot Z_{Q,\alpha}}{Q(x)^\alpha}
\]

for some \( c > 0 \). Then

1) \( \frac{1}{\rho} \log \mathbb{E}[\psi(X)^{-\rho}] \leq H_{\alpha}(P) + I_{\alpha}(P,Q) + \log c \) for \( \rho \neq 0 \),

2) \( \mathbb{E} \left[ \log \frac{\psi(X)}{\psi(\mathcal{X})} \right] \leq H(P) + I(P,Q) + \log c \) for \( \rho = 0 \),

where the expectation is with respect to \( P \).

**Proof:** If \( \rho \neq 0 \), setting \( b = 0 \) in Theorem 20, taking log, and dividing throughout by \( \rho \), the result follows. If \( \rho = 0 \), we have \( \alpha = 1 \), \( \psi(x)^{-1} \leq c Q(x)^{-1} \), and

\[
\mathbb{E} \left[ \log \frac{1}{\psi(X)} \right] = \sum_{x \in \mathcal{X}} P(x) \log(\psi(x)^{-1}) \leq \sum_{x \in \mathcal{X}} P(x) \log(c Q(x)^{-1})
\]

\[
= \log c + \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} - \sum_{x \in \mathcal{X}} P(x) \log P(x)
\]

\[
= H(P) + I(P,Q) + \log c.
\]
Corollary 3. Let \( \rho > -1 \) and \( Q \) be a probability distribution on \( \mathcal{X} \). Suppose \( \psi : \mathcal{X} \rightarrow [0, \infty) \) is such that

\[
\frac{aZ_{Q, \alpha}}{Q(x)^{\alpha}} \leq \psi(x)^{-1}
\]

for some \( a > 0 \). Then,

1) \( \frac{1}{\rho} \log \mathbb{E}[\psi(X)^{-\rho}] \geq \alpha_\rho(P) + I_\alpha(P, Q) + \log a \) for \( \rho \neq 0 \),
2) \( \mathbb{E}\left[\log \frac{1}{\psi(X)}\right] \geq H(P) + I(P, Q) + \log a \) for \( \rho = 0 \),

where the expectation is with respect to \( P \).

Proof: If \( \rho \neq 0 \), the result follows from Theorem 22 taking \( \log, \) and dividing throughout by \( \rho \). If \( \rho = 0 \), we have \( \alpha = 1 \), \( \psi(x)^{-1} \geq aQ(x)^{-1} \), and

\[
\mathbb{E}\left[\log \frac{1}{\psi(X)}\right] = \sum_{x \in \mathcal{X}} P(x) \log(\psi(x)^{-1}) \geq \sum_{x \in \mathcal{X}} P(x) \log(aQ(x)^{-1}) = H(P) + I(P, Q) + \log a.
\]

Corollary 4. Let \( \rho > -1 \) and \( Q \) be a probability distribution on \( \mathcal{X} \). Let \( Q_n \) be the \( n \)-fold product distribution of \( Q \) on \( \mathcal{X}^n \). Suppose \( \psi_n : \mathcal{X}^n \rightarrow [0, \infty) \) is such that

\[
\psi_n(x^n)^{-1} \leq \frac{c_n \cdot Z_{Q, \alpha, n}}{Q_n(x^n)^{\alpha}}
\]

for some \( c_n > 0 \). Then,

1) \( \limsup_{n \rightarrow \infty} \frac{1}{n \rho} \log \mathbb{E}[\psi_n(X^n)^{-\rho}] \leq \alpha_\rho(P) + I_\alpha(P, Q) + \limsup_{n \rightarrow \infty} \frac{1}{n} \log c_n \) for \( \rho \neq 0 \),
2) \( \limsup_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}\left[\log \frac{1}{\psi_n(X^n)}\right] \leq H(P) + I(P, Q) + \limsup_{n \rightarrow \infty} \frac{1}{n} \log c_n \) for \( \rho = 0 \),

where the expectation is with respect to the product distribution \( P_n \) of \( P \) on \( \mathcal{X}^n \).

Proof: Using the fact that, for all \( \alpha \geq 0 \), \( \alpha_\rho(P_n) = n\alpha_\rho(P) \) and \( I_\alpha(P_n, Q_n) = nI_\alpha(P, Q) \) after an application of Corollary 2 then dividing throughout by \( n \) and taking limsup on both sides of the equation, the results follow.

Corollary 5. Let \( \rho > -1 \) and \( Q \) be a probability distribution on \( \mathcal{X} \). Let \( Q_n \) be the \( n \)-fold product distribution of \( Q \) on \( \mathcal{X}^n \). Suppose \( \psi_n : \mathcal{X}^n \rightarrow [0, \infty) \) is such that

\[
\frac{a_n Z_{Q, \alpha, n}}{Q_n(x^n)^{\alpha}} \leq \psi_n(x^n)^{-1},
\]

for some \( a_n > 0, n \geq 1 \). Then,

1) \( \liminf_{n \rightarrow \infty} \frac{1}{n \rho} \log \mathbb{E}[\psi(X^n)^{-\rho}] \geq \alpha_\rho(P) + I_\alpha(P, Q) + \liminf_{n \rightarrow \infty} \frac{1}{n} \log a_n \) if \( \rho \neq 0 \),
2) \( \liminf_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}\left[\log \frac{1}{\psi(X^n)}\right] \geq H(P) + I(P, Q) + \liminf_{n \rightarrow \infty} \frac{1}{n} \log a_n \) if \( \rho = 0 \),

where the expectation is with respect to the product distribution \( P_n \) of \( P \) over the set \( \mathcal{X}^n \).

Proof: Using the fact that, for all \( \alpha > 0 \), \( \alpha_\rho(P_n) = n\alpha_\rho(P) \) and \( I_\alpha(P_n, Q_n) = nI_\alpha(P, Q) \), after an application of Corollary 3 then dividing throughout by \( n \) and taking liminf on both sides of the equation, the results follow.
IV. PROOF OF THE RESULTS BY THE UNIFIED APPROACH

A. Source Coding Problem

Proof of Theorem 1 Choose \( \psi(x) = 2^{-L(C(x))} \), where \( L(C(x)) \) is the length of codeword \( C(x) \) assigned to alphabet \( x \). Since \( C \) is uniquely decodable, from Proposition 1 we have \( \sum_{x \in X} \psi(x) \leq 1 \). Now, an application of Theorem 19 with \( b = 1 \) tells us that we must have \( \mathbb{E}[L(C(X))] \geq H(P) \).

Proof of Theorem 2 Choose \( \psi_n(x^n) = 2^{-L_n(x^n)} \), where \( L_n(x^n) := L(C(x^n)) \) is the length of codeword \( C(x^n) \) assigned to sequence \( x^n \). Then, an application of Corollary 4 with \( c_n = 1 \) and \( Q = P \) gives us

\[
\liminf_{n \to \infty} \mathbb{E}[L_n(C(X))]/n \geq H(P).
\]

Further, we also have

\[
\psi_n(x^n)^{-1} = 2^{L(C(x^n))} = 2^{-[\log P_n(x^n)]} \leq 2^{-[\log P_n(x^n)]} = 1/P_n(x^n).
\]

Thus, an application of Corollary 4 with \( c_n = 1 \) and \( Q = P \) gives us

\[
\limsup_{n \to \infty} \mathbb{E}[L_n(C(X))]/n \leq H(P).
\]

B. Campbell Coding Problem

Proof of Theorem 3 Apply Theorem 18 with \( \psi(x) = 2^{-L(C(x))} \) and \( b = 1 \).

Proof of Theorem 4 Take \( \psi_n(x^n) = 2^{-L(C_n(x^n))} \). Then from (10) we have

\[
\frac{Z_{P,\alpha,n}}{P_n(x^n)\alpha} \leq \psi_n(x^n)^{-1} < 2 \cdot \frac{Z_{P,\alpha,n}}{P_n(x^n)\alpha}.
\]

Hence, (11) follows by applying Corollary 4 with \( c_n = 2 \) and Corollary 5 with \( a_n = 1 \) and \( Q_n = P_n \).

Proof of Theorem 5 Since length function \( K \) satisfies (7), an application of Theorem 18 with \( \psi(x) = 2^{-K(x)} \) gives us

\[
\frac{1}{\alpha} \log \mathbb{E}[2^{\rho K(X)}] \geq H_\alpha(P).
\]

Since \( K(x) = [\log(Z_{P,\alpha}/P(x)^\alpha)] \) satisfies (7) and \( 2^K(x) < 2 \cdot Z_{P,\alpha}/P(x)^\alpha \), from Corollary 2 we know that \( \frac{1}{\rho} \log \mathbb{E}[2^{\rho K(X)}] \leq H_\alpha(P) + 1 \), that is, the minimum in (12) is between \( H_\alpha(P) \) and \( H_\alpha(P) + 1 \). Hence,

\[
\frac{1}{\rho} \log \mathbb{E}[2^{\rho L(X)}] - H_\alpha(P) - 1 \leq R_c(P, L, \rho) \leq \frac{1}{\rho} \log \mathbb{E}[2^{\rho L(X)}] - H_\alpha(P).
\]

Let us now define a probability distribution \( Q_L \) as

\[
Q_L(x) = \frac{2^{-((1+\rho)L(x)}}{\sum_{x'} 2^{-((1+\rho)L(x')}}.
\]

Then,

\[
2^L(x) = Q_L(x)^{-\alpha} Z_{Q_L,\alpha} \cdot \frac{1}{\sum_{x'} 2^{-L(x')}} = Q_L(x)^{-\alpha} Z_{Q_L,\alpha} \cdot \frac{1}{\eta},
\]

where \( \eta = \sum_{x'} 2^{-L(x')} \). Applying Corollaries 2 and 3 with \( \psi(x) = 2^{-L(x)} \), we get

\[
\frac{1}{\rho} \log \mathbb{E}[2^{\rho L(X)}] = H_\alpha(P) + I_\alpha(P, Q_L) - \log \eta.
\]

Combining (26) and (27), we get the desired result.
Proof of Theorem 6 Let $G$ be any guessing function. Choose $\psi(x) = 1/G(x)$. Then, we have $\sum_{x \in \mathcal{X}} \psi(x) = \sum_{x \in \mathcal{X}} 1/G(x) = \sum_{i=1}^{M} 1/i = h_M$ (due to the bijective nature of function $G$). Now, an application of Theorem 18 with $b = h_M$, gives us

$$
\frac{1}{\rho} \log \mathbb{E}[G(X)^\rho] = \frac{1}{\rho} \log \mathbb{E}[\psi(X)^{-\rho}] \geq H_\alpha(P) - \log h_M.
$$

\hfill \blacksquare

Proof of Theorem 7 Let us rearrange the probabilities $\{P(x), x \in \mathcal{X}\}$ in non-increasing order, say $p_1 \geq p_2 \geq \cdots \geq p_M$. Then, the optimal guessing function $G^*$ is given by $G^*(x) = i$ if $P(x) = p_i$. Let us index the elements in the set $\mathcal{X}$ as $\{x_1, x_2, \ldots, x_M\}$, according to the decreasing order of their probabilities. Then, for $i \in \{1, \ldots, M\}$, we have

$$
\frac{Z_{P,\alpha}}{P(x_i)^\alpha} = \frac{\sum_{j=1}^{M} p_j^\alpha}{p_i^\alpha} \geq i = G^*(x_i).
$$

That is, $G^*(x) \leq \frac{Z_{P,\alpha}}{P(x)^\alpha}$ fpr $x \in \mathcal{X}$. Now, an application of Corollary 2 with $Q = P$, $\psi(x) = 1/G^*(x)$, and $c = 1$, we get

$$
\frac{1}{\rho} \log \mathbb{E}[G^*(X)^\rho] = \frac{1}{\rho} \log \mathbb{E}[\psi(X)^{-\rho}] \leq H_\alpha(P) + I_\alpha(P, Q) + \log 1 = H_\alpha(P).
$$

\hfill \blacksquare

Proof of Theorem 8 Let $G_n^*$ be the optimal guessing function from the set of $n$-length sequences $\mathcal{X}^n$ to the set of natural numbers $\{1, 2, \ldots, M^n\}$. An application of Corollary 1 with $\psi(x^n) = 1/G_n^*(x_n)$ and $b_n = h_{M^n}$ gives us

$$
\liminf_{n \to \infty} \frac{1}{n \rho} \log \mathbb{E}[G_n^*(X^n)^\rho] \geq H_\alpha(P) - \limsup_{n \to \infty} \frac{\log h_{M^n}}{n}.
$$

We note that

$$
\limsup_{n \to \infty} \frac{\log h_{M^n}}{n} \leq \limsup_{n \to \infty} \frac{\log (2n \log M)}{n} = 0.
$$

Hence,

$$
\liminf_{n \to \infty} \frac{1}{n \rho} \log \mathbb{E}[G_n^*(X^n)^\rho] \geq H_\alpha(P) \tag{28}
$$

From the proof of the previous theorem, we know that $G^*(x^n) \leq \frac{Z_{P_n,\alpha}}{\gamma_n(x^n)^\alpha}$ $\forall x^n \in \mathcal{X}^n$. Now, an application of Corollary 4 with $Q_n = P_n$, $\psi_n(x) = 1/G_n^*(x)$, and $c_n = 1$, gives us

$$
\limsup_{n \to \infty} \frac{1}{n \rho} \log \mathbb{E}[G_n^*(X^n)^\rho] \leq H_\alpha(P). \tag{29}
$$

Combining (28) and (29), we get the desired result.

\hfill \blacksquare

Proof of Theorem 9 Let $G$ be a guessing function. Define a probability distribution $Q_G = \{Q_G(x), x \in \mathcal{X}\}$ as

$$
Q_G(x) = \frac{G(x)^{-(1+\rho)}}{\sum_{x' \in \mathcal{X}} G(x')^{-(1+\rho)}}.
$$
Then, we have
\[
\frac{Z_{Q,G}}{QG(x)^\alpha} = G(x) \sum_{x' \in \mathcal{X}} \frac{1}{G(x')} = G(x) \cdot h_M.
\]
That is, \( G(x) = \frac{Z_{Q,G}}{h_M \cdot QG(x)} \). Now, an application of Corollaries 2 and 3 with \( \psi(x) = 1/G(x) \), \( c = 1/h_M \) and \( a = 1/h_M \) yields the desired result.

**Proof of Theorem 10** Let us rearrange the probabilities \( \{Q(x), x \in \mathcal{X}\} \) in non-increasing order, say
\[
q_1 \geq q_2 \geq \cdots \geq q_M.
\]

Let \( G_Q \) be the guessing strategy that guesses according to the decreasing order of \( Q \)-probabilities, that is, \( G_Q(x) = i \) if \( Q(x) = q_i \). Then, as in the proof of Theorem 7, we have \( G_Q(x) \leq Z_{Q,G}/QG(x)^\alpha \) for \( x \in \mathcal{X} \). Hence, an application of Corollary 2 with \( \psi(x) = 1/G_Q(x) \) and \( c = 1 \) proves the result.

**Remark** We remark here that the above proof also covers the case when \( \rho \in (-1, 0) \), whereas the original proof of Arikan and Sundaresan dealt only the case \( \rho > 0 \).

**Proof of Theorem 11** Combining Theorems 7 and 9 we get
\[
R_g(P,G,\rho) \geq I(\alpha(Q,G)) - \log h_M.
\]
From Theorem 6, we know that
\[
\frac{1}{\rho} \log \mathbb{E}[G_P(X)^\rho] \geq H(\alpha(P)) - \log h_M.
\]
This along with Theorem 9 tell us that
\[
R_g(P,G,\rho) \leq I(\alpha(Q,G)).
\]

**D. Memoryless Guessing**

**Proof of Theorem 12** Consider
\[
\mathbb{E}[V_\rho(X, \hat{X}_1^\infty)] = \sum_{m=1}^{\infty} P\{G(X, \hat{X}_1^\infty) = m\} V_\rho(X, \hat{X}_1^\infty).
\]
Since
\[
P\{G(X, \hat{X}_1^\infty) = m\} = \sum_{x \in \mathcal{X}} P(x) \hat{P}(x)(1 - \hat{P}(x))^{m-1},
\]
we obtain
\[
\mathbb{E}[V_\rho(X, \hat{X}_1^\infty)] = \sum_{m=1}^{\infty} \sum_{x \in \mathcal{X}} P(x) \hat{P}(x)(1 - \hat{P}(x))^{m-1} \frac{1}{\rho} \prod_{l=0}^{\rho-1} (m + l)
\]
\[
= \sum_{x} P(x) \hat{P}(x)^{-\rho} = \mathbb{E}[\hat{P}(X)^{-\rho}],
\]
(30)
where (30) is by [13, p. 2]. Now, the result follows from Theorem 18 with \( \psi(x) = \hat{P}(x) \) and \( b = 1 \). Since \( \hat{P} \) is a probability distribution, we have \( \sum_{x \in \mathcal{X}} \hat{P}(x) = 1 \). Hence, by Theorem 18 \( \frac{1}{\rho} \log \mathbb{E}[V_\rho(X, \hat{X}_1^\infty)] = H_\alpha(P) \), attained by \( \hat{P}(x) = P(x)^\alpha/Z_{P,\alpha} \).

**Proof of Theorem 13** The result follows easily by taking \( \psi = \hat{Q}^*, c = 1, a = 1 \) in Corollaries 2 and 3.
E. Task Partitioning Problem

In this section, we establish all the results concerning the tasks partitioning problem discussed in Section II using the unified approach. In the following, Propositions 2 and 3 are the analogue of Kraft-McMillan theorem for partition functions.

**Proposition 2.** \[12, Prop. III.1\]: Let $A$ be the partition function associated with a partition of $\mathcal{X}$ of size $N$. Then,

$$
\sum_{x \in \mathcal{X}} \frac{1}{A(x)} = N.
$$

**Proposition 3.** \[12, Prop. III.2\] If $\lambda : \mathcal{X} \to \mathbb{N} \cup \{+\infty\}$ and $\sum_{x \in \mathcal{X}} 1/\lambda(x) = \mu$, then there exists a partition of $\mathcal{X}$ of size at most

$$
\min_{\alpha > 1} \alpha \mu + \log_\alpha |\mathcal{X}| + 2
$$

such that its partition function $A$ satisfies $A(x) \leq \min\{\lambda(x), |\mathcal{X}|\}$ for $x \in \mathcal{X}$.

**Corollary 6.** Let $N$ be such that $N > \log |\mathcal{X}| + 2$. If $\lambda : \mathcal{X} \to \mathbb{N} \cup \{+\infty\}$ is such that $\sum_{x \in \mathcal{X}} 1/\lambda(x) = (N - \log |\mathcal{X}| - 2)/2$, then there exists a partition of $\mathcal{X}$ of size at most $N$ with partition function $A$ such that $A(x) \leq \min\{\lambda(x), |\mathcal{X}|\}$ for $x \in \mathcal{X}$.

**Proof:** Let $|A|$ denote the size of partition $A$. Applying Proposition 3 with $\mu = (N - \log |\mathcal{X}| - 2)/2$, we see that there exists a partition $A$ of $\mathcal{X}$ with partition function $A$ such that

$$
|A| \leq \min_{\alpha > 1} \alpha \mu + \log_\alpha |\mathcal{X}| + 2 \leq 2\mu + \log |\mathcal{X}| + 2 \leq N
$$

and $A(x) \leq \min\{\lambda(x), |\mathcal{X}|\}$ for $x \in \mathcal{X}$. $\blacksquare$

**Proof of Theorem 14(a):** Follows from an application of Theorem 18 with $\psi(x) = 1/A(x)$ and $b = N$, and by Proposition 2.

**Proof of Theorem 14(b):** Let

$$
\lambda(x) = \lceil \beta \cdot P(x)^{-\alpha} \rceil, \quad \text{with } \beta = \frac{2 \cdot Z_{P,\alpha}}{N - \log |\mathcal{X}| - 2}.
$$

Then,

$$
\mu = \sum_{x \in \mathcal{X}} \frac{1}{\lambda(x)} \leq \sum_{x} \frac{1}{\beta \cdot P(x)^{-\alpha}} \leq \frac{N - \log |\mathcal{X}| - 2}{2}.
$$

Now an application of Corollary 6 tells us that there exists a partition of $\mathcal{X}$ of size at most $N$ with partition function $A$ such that $A(x) \leq \min\{\lambda(x), |\mathcal{X}|\}$ for $x \in \mathcal{X}$. Further, for any $\rho \neq 0$, we also have

$$
A(x) \leq \lambda(x) = \left( \lceil \beta \cdot P(x)^{-\alpha} \rceil \right)^{1/|\rho|} \leq \left( 1 + 2^{|\rho|} (\beta \cdot P(x)^{-\alpha})^{1/|\rho|} \right)^{1/|\rho|} = \left( 1 + \left( \frac{Z_{P,\alpha}}{NP(x)^\alpha} \right)^{|\rho|} \right)^{1/|\rho|},
$$

where $\tilde{N} = \frac{N - \log |\mathcal{X}| - 2}{4}$. Now, an application of Theorem 20 with $\psi(x) = 1/A(x)$, $b = 1$, $c = 1/\tilde{N}$, and $Q = P$ gives us

$$
\mathbb{E}[A(X)^\rho] \leq 1 + 2^\rho (H_\alpha(P) - \log \tilde{N}) \quad \text{if } \rho > 0
$$
and if \( \rho < 0 \), we have

\[
\mathbb{E}[A(X)^\rho] \geq 2^{\rho(H_a(P) - \log \tilde{N} - \rho^{-1} \log(1+N^{-\rho}))} = \left( \frac{\tilde{N}^{1-\rho}}{1+N^{1-\rho}} \right)^{2^\rho H_a(P)} \geq \frac{1}{2} \cdot 2^\rho H_a(P).
\]

Also, if \( \rho < 0 \), since \( A(X) \geq 1 \), we have \( A(X)^\rho \leq 1 \). Hence \( \mathbb{E}[A(X)^\rho] \leq 1 \). This completes the proof. \( \square \)

**Proof of Theorem 20**

**Part (i)** Proceeding as in the proof of Theorem 14(b), we can show that for any \( n \geq 1 \), there exists a partition \( A_n \) of \( \mathcal{X}^n \) of size \( N^n \) such that the associated partition function \( A_n(\cdot) \) satisfies

\[
\mathbb{E}[A_n(X^n)^\rho] \leq 1 + 2^{\rho(H_a(P) - \log \tilde{N})},
\]

where \( \tilde{N}_n = (N^n - n \log |\mathcal{X}| - 2)/2 \). We note that \( P_n \) is the \( n \)-fold product distribution of \( P \) on \( \mathcal{X}^n \). Thus, we have \( H_a(P_n) = n H_a(P) \). Now, if \( \log N > H_a(P) + \delta \), then there exists \( \delta > 0 \) such that \( \log N > H_a(P) + \delta \). Since \( \lim_{n \to \infty} \frac{\log \tilde{N}_n}{n} = \log N \), there exists a positive integer \( n_\delta \) such that \( \log N - \delta < \frac{\log \tilde{N}_n}{n} \) for \( n \geq n_\delta \). Thus, we have

\[
\mathbb{E}[A_n(X^n)^\rho] < 1 + 2^{n \rho(H_a(P) - \log N + \delta)} \quad \text{for} \quad n \geq n_\delta.
\]

Since \( \log N > H_a(P) + \delta \), we have \( \rho(H_a(P) - \log N + \delta) < 0 \). Consequently, \( \limsup_{n \to \infty} \mathbb{E}[A_n(X^n)^\rho] \leq 1 \). Since \( A_n(x^n) \geq 1 \) for all \( x^n \in \mathcal{X}^n \), we have \( \liminf_{n \to \infty} \mathbb{E}[A_n(X^n)^\rho] \geq 1 \). Thus, we conclude that \( \lim_{n \to \infty} \mathbb{E}[A_n(X^n)^\rho] = 1 \).

**Part (ii)** For each \( n \geq 1 \), let \( A_n \) be a partition of \( \mathcal{X}^n \) of size \( N^n \) with partition function \( A_n \). An application of Theorem 18 with \( \psi(\cdot) = 1/A_n(\cdot) \) and \( b = N^n \) gives us

\[
\mathbb{E}[A_n(X^n)^\rho] \geq 2^{\rho(H_a(P) - \log N^n)} = 2^{n \rho(H_a(P) - \log N)}.
\]

Now, since \( \log N < H_a(P) \) and \( \rho > 0 \), we have \( \lim_{n \to \infty} \mathbb{E}[A_n(X^n)^\rho] = \infty \). \( \square \)

**Proof of Theorem 16** Define a probability distribution \( Q_A = \{ Q_A(x), \ x \in \mathcal{X} \} \) as

\[
Q_A(x) = \frac{A(x)^{-1/(1+\rho)}}{\sum_{x' \in \mathcal{X}} A(x')^{-1/(1+\rho)}}.
\]

Then, we have

\[
\frac{Z_{Q_A,\alpha}}{Q_A(x)^\alpha} = A(x) \sum_{x' \in \mathcal{X}} \frac{1}{A(x')} = A(x) \cdot N,
\]

where the last equality follows due to Proposition 2. Rearranging the terms, we have \( A(x) = \frac{Z_{Q_A,\alpha}}{N Q_A(x)^\alpha} \). Now, an application of Corollaries 2 and 3 with \( \psi(x) = 1/A(x), \ c = 1/N \) and \( a = 1/N \) yields the desired result. \( \square \)

**Proof of Theorem 17** The proof is identical to that of Theorem 14(b), with the exception that we need to invoke Theorem 20 with \( \psi(x) = 1/A_Q(x), \ b = 1 \) and \( c = 1/\tilde{N} \). \( \square \)
V. SUMMARY AND CONCLUDING REMARKS

The purpose of this paper was to establish a close relationship between the problems on source coding, guessing, and task partitioning by unifying all the proofs of the results known in these problems. Indeed, we could establish the lower bound of these problems using an optimization problem and apply Campbell’s idea to achieve the lower bound in an asymptotic way. Our approach also enabled us to refine some of the known results as well as obtain some new results for the mismatched version of these problems. We were able to formulate a more general mathematical problem where the above mentioned problems are particular instances.

We feel that this generalization, in addition to help solving new problems that fall in this framework, would also provide new insights. For example, Arikan’s guessing problem can be interpreted as a particular case of a more general task partitioning problem, where tasks assigned to a particular key also have to be ordered.

The presented unified approach can be explored further in many ways. This includes, (i) Extension to general state-space: It would be interesting to see if the studied problems can be formulated and solved, for example, for countably infinite or continuous support sets. (ii) Applications: Arikan showed an application of the guessing problem in a sequential decoding problem [11]. Humblet showed that cumulants of code-lengths in Campbell’s problem has applications in minimizing the probability of buffer overflow in source coding problems [15]. We would like to see if other potential applications emerge from this unified study.

REFERENCES

[1] C. E. Shannon, “A mathematical theory of communication,” Bell System Technical Journal, vol. 27, no. 3, pp. 379–423, 1948.
[2] A. Rényi et al., “On measures of entropy and information,” in Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, 1961, pp. 547–561.
[3] D. Z. Aczél J., On Measures of Information and Their Characterizations, ser. Mathematical in Science and Engineering. Academic Press, 1975, vol. 115.
[4] L. Campbell, “A coding theorem and Rényi’s entropy,” Information and Control, vol. 8, no. 4, pp. 423–429, 1965.
[5] A. C. Blumer and R. J. McEliece, “The Rényi redundancy of generalized Huffman codes,” IEEE Transactions on Information Theory, vol. 34, no. 5, pp. 1242–1249, Sep. 1988.
[6] R. Sundaresan, “Guessing under source uncertainty,” IEEE Transactions on Information Theory, vol. 53, no. 1, pp. 269–287, Jan 2007.
[7] ——, “A measure of discrimination and its geometric properties,” in Proceedings IEEE International Symposium on Information Theory, June 2002, pp. 264–.
[8] O. Johnson and C. Vignat, “Some results concerning maximum Rényi entropy distributions,” Annales de l’Institut Henri Poincare (B) Probability and Statistics, vol. 43, no. 3, pp. 339–351, 2007.
[9] M. Ashok Kumar and R. Sundaresan, “Minimization problems based on relative $\alpha$-entropy ii: Reverse projection,” IEEE Transactions on Information Theory, vol. 61, no. 9, pp. 5081–5095, Sep. 2015.
[10] J. L. Massey, “Guessing and entropy,” in Proceedings of 1994 IEEE International Symposium on Information Theory, June 1994, pp. 204–.

[11] E. Arikan, “An inequality on guessing and its application to sequential decoding,” IEEE Transactions on Information Theory, vol. 42, no. 1, pp. 99–105, Jan 1996.

[12] C. Bunte and A. Lapidoth, “Encoding tasks and Rényi entropy,” IEEE Transactions on Information Theory, vol. 60, no. 9, pp. 5065–5076, Sep. 2014.

[13] W. Huleihel, S. Salamatian, and M. Médard, “Guessing with limited memory,” in 2017 IEEE International Symposium on Information Theory (ISIT), June 2017, pp. 2253–2257.

[14] T. Cover and J. Thomas, Elements of Information Theory. Wiley, 2012.

[15] P. Humblet, “Generalization of Huffman coding to minimize the probability of buffer overflow (corresp.),” IEEE Transactions on Information Theory, vol. 27, no. 2, pp. 230–232, March 1981.

[16] D. P. Bertsekas, Nonlinear Programming, 2nd ed. Belmont, MA: Athena Scientific, 2003.

[17] M. Teboulle and I. Vajda, “Convergence of best ϕ-entropy estimates,” IEEE Transactions on Information Theory, vol. 39, no. 1, pp. 297–301, Jan 1993.